# Cotton Disease Detection on UAV Images: A Deep Learning-Based Approach with YOLOv7

Zakaria Kinda[1(✉)], Sadouanouan Malo[1], Thierry Roger Bayala[1], and Issa Wonni[2]

[1] Université Nazi Boni/LAMDI (Laboratory of Algebra, Discrete Mathematics and Computer Science), Bobo-Dioulasso, Burkina Faso
kindazakaria@yahoo.fr

[2] Institute of Environment and Agricultural Research of Burkina Faso (INERA), Dori, Burkina Faso

**Abstract.** Cotton is the most important agricultural product in Burkina Faso, and it is farmed by 25% of the country's population. Cotton diseases, on the other hand, are a big issue for this crop, contributing considerably to output losses. These disorders are detected manually, which increases the time required for therapy. Artificial intelligence techniques can help enhance cotton production by automatically recognizing these diseases. This research aims to detect cotton diseases using UAV photos obtained in the field. We used the YOLOv7 detection model fine-tuned on the tomato leaves dataset and subsequently applied to the cotton leaves dataset. On the cotton dataset, experimental results from the YOLOv7 model yielded mAP@0.50 (mean average precision), f1score, Precision, and Recall of 50.7%, 55.4%, 53.7%, and 57.4%, respectively.

**Keywords:** Deep Learning · UAV images · Cotton diseases · YOLOv7

## 1 Introduction

Agriculture is vital to the economy of Burkina Faso. There are different crop varieties, but cotton is the primary crop for the country's economic development. Burkina Faso is one of West Africa's top cotton growers. Cotton is cultivated by 25% of the country's population and accounts for 55% to 70% of total export revenues [1]. However, in recent years, we have seen a progressive reduction of 30% [2] in agricultural cotton production, which has had a significant influence on the country's economy. This decrease in output is the result of a multitude of issues inadequate rainfall distribution, inadequate input management, and, most importantly, disease is among them. Various types of illness damage the plant during its development cycle such as fungal and viral diseases, etc. To improve cotton yield, it is crucial to detect these diseases early. Phytopathologists in Burkina Faso, on the other hand, are conducting research to detect cotton diseases. In recent years, there has been an increase in the use of artificial intelligence in agriculture. This study is carried out in labs (microbiology, molecular biology, application laboratories,


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 234–249, 2025.
https://doi.org/10.1007/978-3-031-81573-7_19

9. Tayachi, M.: Sécurité des images par tatouage numérique et cryptographie dans les applications médicales. PHD thesis, school: Université de Bretagne occidentale-Brest; Université de Tunis El Manar (2021)
10. Zainol, Z., Teh, J.S., Alawida, M.: A new chaotic image watermarking scheme based on SVD and IWT. IEEE Access **8**, 43391–43406 (2020)
11. Alzahrani, A., Memon, N.A.: Blind and robust watermarking scheme in hybrid domain for copyright protection of medical images. IEEE Access **9**, 113714–113734 (2021)
12. Mohammed, A.A., Jebur, B.A., Younus, K.M.: Hybrid DCT-SVD based digital watermarking scheme with chaotic encryption for medical images. In: IOP Conference Series: Materials Science and Engineering, vol. 1152, no. 1, p. 012025. IOP Publishing (2021)
13. Fierro-Radilla, A., Nakano-Miyatake, M., Cedillo-Hernandez, M., Cleofas-Sanchez, L., Perez-Meana, H.: A robust image zero-watermarking using convolutional neural networks. In: 2019 7th International Workshop on Biometrics and Forensics (IWBF), pp. 1–5. IEEE (2019)

## 7 Conclusion

In this paper, we propose a new robust zero watermarking scheme for authenticating medical images. This scheme is based on the hyper-catadioptric system model and hyperbolic geometry. The main contributions of this new approach are as follows. Firstly, through simulations, we have determined the characteristics of the center and radius of the image plane. Then we propose to adapt our image plane to the structure of the database. We therefore propose to use a selection of points of interest well distributed over the whole image to ensure the integrity of the whole image. In addition, we propose to use an asynchronous stream encryption system to restore the image after an attack. Finally, we propose a new zero-watermarking scheme to secure and authenticate images in the distributed database. Implementing this model enabled us to observe transformations and adapt our watermarking scheme to the structure of the distributed database. The formal analysis of our schema reveals the chaotic aspect of the watermarking scheme and therefore its robustness. We made three major contributions. Using simulations, we determined the various characteristics of the hyper-catadioptric model. Then we related the hyperbolic structure of the distributed database to the image plane. Finally, we proposed a new watermarking scheme based on the hyper-catadioptric model and the hyperbolic structure. In the course of our work, we will implement the watermarking scheme proposed in this paper and make a comparative analysis of our solution with existing ones in terms of performance. We will exploit the regions of non interest to insert more data in the frequency domain.

## References

1. Singh, O.P., Singh, A.K., Srivastava, G., Kumar, N.: Image watermarking using soft computing techniques: a comprehensive survey. Multimedia Tools Appl. **80**(20), 30367–30398 (2021)
2. Mohanarathinam, A., Kamalraj, S., Prasanna Venkatesan, G.K.D., Ravi, R.V., Manikandababu, C.S.: Digital watermarking techniques for image security: a review. J. Ambient Intell. Human. Comput. **11**(8), 3221–3229 (2020)
3. Kamaruddin, N.S., Kamsin, A., Por, L.Y., Rahman, H.: A review of text watermarking: theory, methods, and applications. IEEE Access **6**, 8011–8028 (2018)
4. Tiken, C., Samli, R.: A comprehensive review about image encryption methods. Harran Üniversitesi Mühendislik Dergisi **7**(1), 27–49 (2022)
5. Geetha, S., Punithavathi, P., Infanteena, A.M., Sindhu, S.S.S.: A literature review on image encryption techniques. Int. J. Inf. Secur. Priv. (IJISP) **12**(3), 42–83 (2018)
6. Tiendrebeogo, T., Magoni, D.: Virtual and consistent hyperbolic tree: a new structure for distributed database management. In: Bouajjani, A., Fauconnier, H. (eds.) NETYS 2015. LNCS, vol. 9466, pp. 411–425. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26850-7_28
7. Comby, F., de Kerleau, C.C., Strauss, O.: Étalonnage de caméras catadioptriques hyperboloïdes. Traitement du Signal **22**(5), 419–431 (2005)
8. Cox, I.J., Doërr, G., Furon, T.: Watermarking is not cryptography. In: Shi, Y.Q., Jeon, B. (eds.) IWDW 2006. LNCS, vol. 4283, pp. 1–15. Springer, Heidelberg (2006). https://doi.org/10.1007/11922841_1

Then we give the elements of the set that allow us to check whether our transformation function is chaotic. These are the properties of the dynamics of regular systems, topological transitivity and sensitivity to initial conditions.

**Periodicity.** A point $p \in X$ is considered periodic of period k if k is a non-zero integer such that:

$$f^k(p) = p, and \forall h \in [0, k-1] f^h(p) \neq p; \tag{7}$$

We will note $Perk(f)$ the k-periodic set of points of $f$, and $Perk(f)$ the set of periodic points of any period.

According to Fig. 6 we have: $P \overset{f^1}{\longmapsto} P_m \overset{f^2}{\longmapsto} p \overset{f^3}{\longmapsto} P_m \overset{f^4}{\longmapsto} P$ From this we can say that f is periodical of period 4

**Regularity.** A discrete dynamical system (X, f) is said to be regular if all periodic points of f, called $Per(f)$, are dense in X. In a metric space (X, d), the dynamical system (X, f) is regular if and only if:

$$\forall x \in X, \forall \epsilon < 0, \exists p \in Per(f), d(x, p) < \epsilon \tag{8}$$

**Transitivity.** Indeed, in our system for any node of the hyperbolic tree that we take, there exists an image point that minimizes the distance from the node to the image point. Since, for any pair of openings U, V ⊂ X, there exists k > O such that :

$$f^k(U) \cap V \neq \emptyset \tag{9}$$

then f is topologically transitive.

**Dependence on Initial Conditions.** f has a sensitive dependence on initial conditions if there exists $\delta > 0$ such that, for all x ∈ X and for any neighborhood V of x, it exists y ∈ V and n ≥ O there such that:

$$d(f^n(x), f^n(y)) > \delta \tag{10}$$

$\delta$ is called the sensitivity constant of $f$.

**Chaotic System**

A function f: X → X is said to be chaotic on $X$ if:

1. (X, f) is regular,
2. f is topologically transitive,
3. f has a sensitive dependence on the initial conditions

Since our transformation function f is chaotic, then the system (X, f) is chaotic, and, according to Devaney, it is unpredictable because of its sensitive dependence on initial conditions. It cannot be decomposed or simplified into two non-interacting subsystems because of topological transitivity.

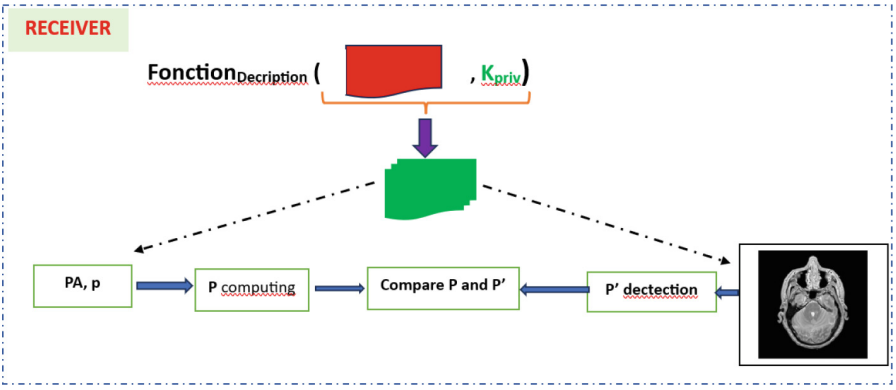



**Fig. 9.** .

## 6 Solution Analysis

### 6.1 Cryptographic System

In our approach the encryption technique is by the asynchronous stream cipher method. Indeed, since the function f of the dynamic key generation is parameterized by an OID which is a series of previously encoded numbers, our encryption algorithm can be considered as an asynchronous stream cipher. This type of encryption is also called self-synchronous stream encryption. The error distribution is limited to the size of the memory. Therefore, if ciphers in the cipher text are deleted or inserted, the receiver is able to resynchronize with the sender thanks to the memory. Concerning active attacks, if an active adversary modifies a part of the ciphers of the coded text, the receiver is able to detect it. It is for these qualities that we have adopted such an approach.

### 6.2 Formal Analysis

In this section, we present the robustness of our system using a hyperbolic tree topology based on chaos theory. In our context, the step space corresponds to the passage of the image from step i to step i+1 after modification of a pixel. The function f is associated with the composition of two projection functions which are respectively: the hyper-catadioptric projection and the stereographic projection. In particular, we consider discrete dynamical systems.

**Definition 1.** A discrete dynamical system is a pair (X, f) formed by:

- A non empty topological space (X, p), called the space of steps,
- A continuous function f: X → X, called successor function.

$$\begin{cases} x^0 \in X \\ n \in \mathbb{N}, x^{n+1} = f(x^n) \end{cases} \tag{6}$$

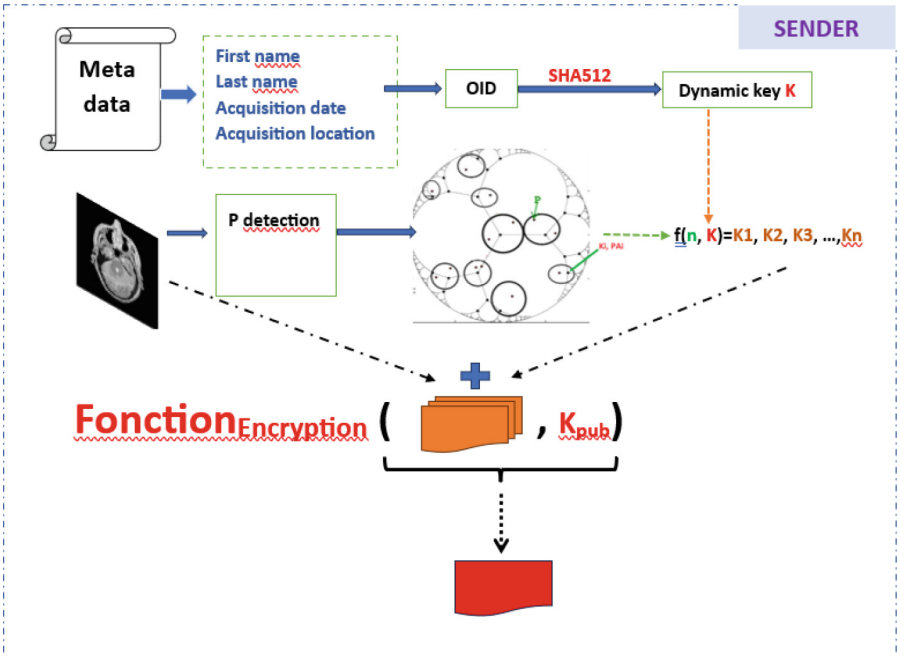



**Fig. 8.** Sender

6. **Step 6.** Determine the dynamic key $K$. We will generate K (512 bits) using a hash function (SHA512) parameterized by the image OID.
7. **Step 7.** Split the dynamic key $K$ into subkeys ($k_1$, $k_2$, $k_3$,..., $k_n$ ). This decomposition is based on the number of nodes returned in step 4.
8. **Step 8.** Associate each sub-key $k_i$ with the node of the tree and a set $N_i$. Subkeys represent public keys.
9. **Setup 9.** Encrypt the watermarked image and subkeys with the recipient's public key

**Receiver**

1. **Step 1.** Decrypt the message with your private key
2. **Step 2.** Search for the $k_i$ associated with your private key
3. **Step 3.** Determine for each sub-key $k_i$ of its server who stores it
4. **Step 4.** Determine the model parameters, the plane equation and the hyperbolic coordinates of each image point.
5. **Step 5.** Compute the inverse transform of each image point
6. **Step 6.** Determine new points of interest using the same detector
7. **Step 7.** Compare new detected points with computed points

On this figure we observe a convergence towards point $P_f$ On the image plane we will have a disk of center $P_0(u_0, v_0)$ and radius $P_0P_f$. So our first virtual address will be $(u_0, v_0)$. From this we can determine the address of its descendants until we reach the fixed tree depth.

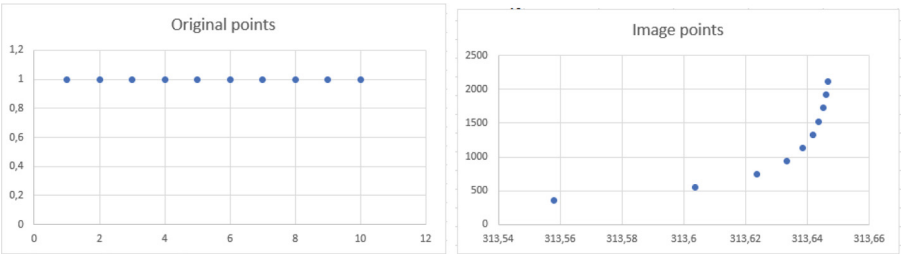



**Fig. 7.** Original points and Computing the image points

## 5 Solution

We're going to assimilate the image plane of the model to our hyperbolic space, which is the Poincaré disk. In this disk, we'll build our hyperbolic tree, which is the basic structure of our database. In our approach, each server stores the hyperbolic coordinates of the image points, the model parameters and the equation of the plane.

### 5.1 Our Scheme

Our schema is illustrated by these two Figs. 8 and 9

**Sender**

1. **Step 1.** Place the image in Euclidian space. The system must save this equation. The knowledge of this equation is necessary to reconstitute the points.
2. **Step 2.** Select pixels using an interest point. Selected pixels must have a good distribution on the image.
3. **Step 3.** Compute the transform of selected points
4. **Step 4.** Group image points to nearest node. Each node can save $N_i$ image points. The set of tuples ($N_1$, $N_2$, $N_3$..., $N_n$) represents our cryptographic signature. The system must return the number of nodes used.
5. **Step 5.** Generate the image $OID$ (Objet IDentifier). Since $OID$ is considered a private key, the system must ensure its uniqueness. It must be generated by a function that takes metadata information and image characteristics (entropy) as parameters. Any attempt to modify the image or metadata will produce a different $OID$.

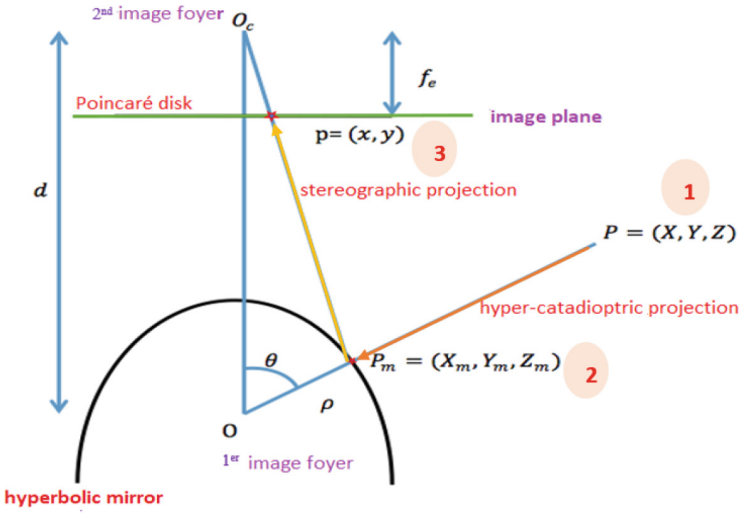



**Fig. 6.** Hyper-catadioptric and stereographic projections

$u_0$ and $v_0$ are the coordinates of the projection of the camera's optical axis. $k$ and $\alpha$ are parameters

The inverse transformation gives the direction to which the image point belongs. g* : $(u, v) \mapsto \overrightarrow{D}$

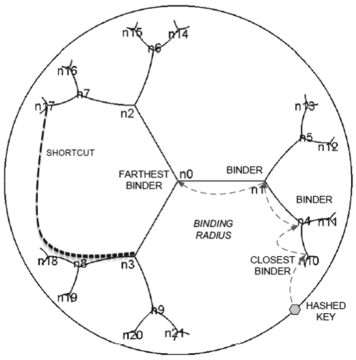
$$g^*(u,v) = \left(\frac{u}{\sqrt{u^2+v^2}}, \frac{v}{\sqrt{u^2+v^2}}, z\right) \tag{4}$$

$$z = r_i \frac{\alpha(k-1) \pm \sqrt{k(k-2)(\alpha^2+r_i^2)}}{r_i^2 k(k-2)\alpha^2} \tag{5}$$

The implementation of such a system requires the determination of intrinsic and extrinsic parameters. The determination of these parameters remains quite complex, hence the need to use calibration techniques. These are based on the knowledge of the pairs of points (3-D, 2-D). A mapping of these points and their projections will estimate the parameters ($u_0$ , $v_0$, $k$ and $\alpha$ ) of the model.

### 4.2 Model Simulation

We then estimate the parameters proposed in [7]. We will then have $u_0$ =160.49, $v_0$= 119.02, $k = 3.75$ and $\alpha = 183.45$ The aim of model simulation is to determine the boundary values and observe the various transformations. To visualize the transformation, we'll use one of the characteristics of hyperbolic space. Straight lines in Euclidean space are arcs in hyperbolic space. To achieve this, we'll consider aligned points in our Euclidean space and then determine the images using the $g$ function. After calculation, we obtain images represented by the Fig. 7.




**Algorithm 1:** Recursive building of our virtual hyperbolic tree.

```
1  Function NodeChildrenCoordComp (Node, q);
   Input  : Know the coordinates of every node: N
   Output: Computethecoordinatesofitschildren : N1...Np
2  step ← argcosh(1/sin(π/q));
3  angle ← 2π/q;
4  childCoords ← Node.Coords;
5  for i ← 1 to q do
6  |  ChildCoords.rotationLeft(angle);
7  |  ChildCoords.translation(step);
8  |  ChildCoords.rotationRight(π);
9  |  if ChildCoords ≠ Node.ParentCoords then
10 |  |  Node.TabChildCoords[i] = ChildCoords;
11 |  end
12 end
13 return ChildrenCoord;
```

**Fig. 5.** Hyperbolic database structure

necessary to set these two parameters. In Fig. 5 we have a tree of degree 3 . Starting from this figure, the first virtual server (node $n0$) will have address [0,0] (the center of the disk). It will then calculate the addresses of its three children ($n1$, $n2$ and $n3$). From these, each node will calculate the addresses of its two children until it reaches the depth of the tree. In [6], the authors proposed an algorithm for computing the coordinates of a node at each step. Based on this algorithm, we propose an algorithm (shown in Fig. 5) to build our hyperbolic tree.

## 4 Our Hyper-catadioptric Model

A hypercatadioptric system consists of a lens and a hyperbolic mirror.The camera's optical axis coincides with the mirror's axis of symmetry. Its mathematical model [7] will enable us to move from Euclidean space to hyperbolic space.

### 4.1 Hyper-catadioptric Model Ant Projection

The model of the hyper-catadioptric system is based on two types of projections [7]. A hyper-catadioptric projection is used to determine the $P_m(X_m, Y_m, Z_m)$ mirror projection of a $P(X, Y, Z)$ point in space. The second projection, called stereographic projection, determines the project $p(x, y)$ of the image plane of the point $P_m(X_m, Y_m, Z_m)$.

These two transformations can be resumed using the $g$ function:
$P(X, Y, Z) \mapsto p(x, y)$.

$$g(X, Y, Z) = (r_i \frac{X}{\sqrt{X^2 + Y^2}} + u_0, r_i \frac{Y}{\sqrt{X^2 + Y^2}} + v_0) \tag{3}$$

$r_c = (\sqrt{X^2 + Y^2}, Z)$; $r_i = \alpha r_c \frac{\pm\sqrt{k(k-2)(Z^2+r_c^2)} - Z(k-1)}{Z^2 - k(k-2)r_c^2}$

1. **Non-blind technique**: the sender must send both the watermarked image and the original image. The presence of the latter is essential for detecting or extracting the watermark.
2. **Blind technique**: the watermark is detected or extracted without the original image.
3. **Blind technique**: the watermark detection or extraction can be effected with or without the original image.

### 2.4 Zero Watermarking

In certain fields of application, these distortions are to be avoided. In such cases, zero watermarking techniques are used. Zero watermarking extracts important and unique characteristics (histogram, median, entropy, energy, co-occurrence matrix, ...) from the original image to build a signature (the watermark). These characteristics are extracted from the properties of the spatial [9] or frequency [10] domain. The signature must be stored securely. It can be used later to identify and authenticate the image.

In [9] the author first used static analysis of the host image to extract relevant characteristics. Then he extracted information from the patient before transforming it into a binary matrix. It uses a cumulative subtraction process on the host image to generate a matrix of the same size as the binary matrix. The new matrix is used to define Jacobian functions that generate a Jacobian matrix. The latter is used to construct a signature (key to be transmitted to the recipient). Comparisons show that their solution is better than existing methods for certain types of attack with good metrics. However, it is not as good for other types of attack with low metrics.

In [13] the authors proposed a zero-watermarking solution based on convolutional neural networks (CNNs). They first extracted features and from the trained CNN. The extracted features are and linked to the owner's brand to generate a master share. The main share is stored securely. When an image is subject to an ownership dispute, it is fed into the trained CNN as input and its inherent features are extracted. The extracted image features and the master share are used to retrieve the tattoo pattern. This approach is robust in the case of high distortion with 22 dB PSNR, but the processing time is long.

## 3 Database for Images

For transfer and archiving, we're going to use a distributed database based on [6]. This database is made up of virtual servers with virtual addresses that are hyperbolic coordinates. The database is chosen for its scalable, reliable and consistent structure. It is ideal for storing large amounts of data, and also supports queries on large data sets.

Setting up the structure of this distributed database is like incrementally building a hyperbolic tree in the Poincaré disk. The hyperbolic tree is characterized by two parameters: degree and depth. Before building the tree, it is

**Possible Attacks.** The watermarked image can be the object of several types of attack. These attacks aim to visualize or remove the watermark. In some cases, the attacker will add another watermark to make it difficult for the receiver to extract the watermark. Figure 4 summarizes the most common types of attack encountered in the literature.

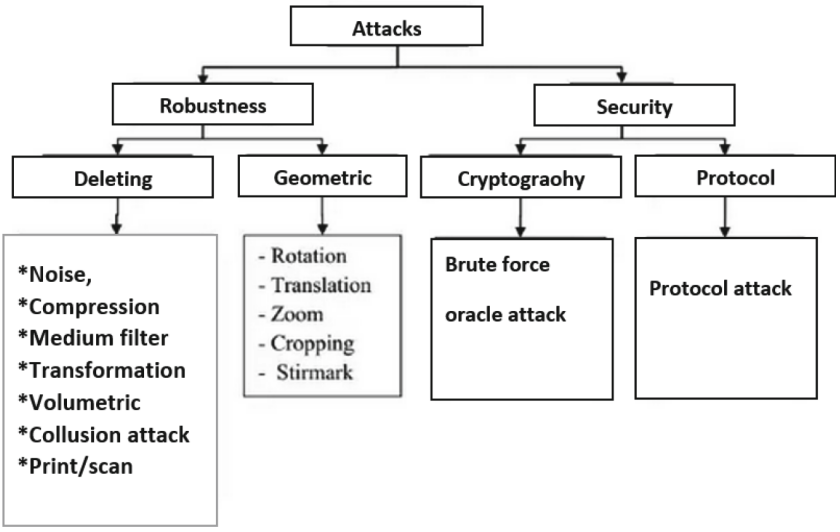



**Fig. 4.** Attaks

Most watermarking schemes encountered in the literature seek to be robust against geometric attacks and compression: this is the case for robust schemes. For fragile schemes, the solution is generally robust to compression but fragile to geometric attacks, so as to identify any attempts at modification.

**Watermark Detection or Extraction.** In some applications, the receiver just wants to detect the presence of a watermark. Detection is used to authenticate images. Other applications seek to extract the watermark as faithfully as possible. At this level we can have two possible watermarking schemes.

1. **Reversible watermarking**: this enables the original image to be restored, while removing the mark. It is used for authentication, as well as in military and medical applications.
2. **Irreversible watermarking**: permanently preserves the (often insignificant) changes to the original image when the watermark is inserted, even after the mark has been removed.

Once the receiver receives the watermarked image, the mark can be detected or extracted in one of three ways. These three ways can form a classification.

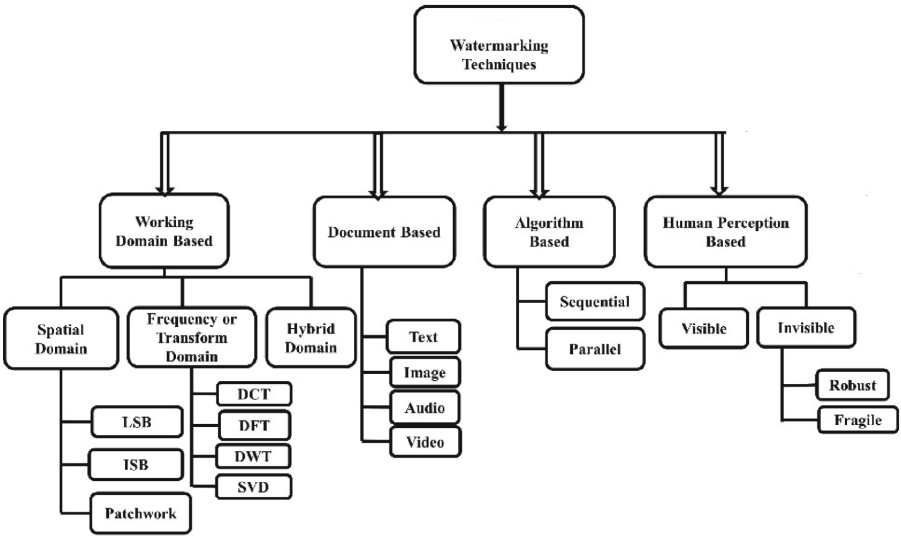



**Fig. 3.** watermarking process

**Watermark Insertion.** The watermark is inserted using a function that takes the original image $ImO$ and a secret key $K$ to produce a watermarked image $ImW$.

$$INSERTION(ImO, K) = ImW \tag{2}$$

Watermark insertion requires the choice of insertion domain (spatial and frequency).

1. In the **spatial domain**, pixels are coded on 8 bits (12 bits for DICOM images) and insertion can be performed on the Least Significant Bit (LSB) on selected pixels. Techniques in this domain (Fig. 3) have a good insertion capacity, but remain fragile in the face of geometric attacks.
2. In the **frequency domain**, an image can be decomposed into several components, with the watermark being inserted into one of the components. Unlike the previous domain, frequency-domain techniques (Fig. 3) have a low insertion capacity but are robust against geometric attacks.

The choice of domain and technique depends on the desired objective. Many schemes in the literature are based on a hybrid approach. In [11] the authors proposed a hybrid based on Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT) and Singular Value Decomposition (SVD). In [12] the authors proposed a scheme based on DCT and SVD. These techniques introduce distortion into the image. These distortions can be controlled by computing evaluation metrics, the main ones being : PSNR (Peak Signal to Noise Ratio), SSIM (Structural SIMilarity) and NC (Normalized Correlation).

Several techniques have been proposed in the literature for encrypting images [4,5]. These methods are not very effective. In fact, they provide a priori security, i.e. during transfer between sender and receiver. Once the image has been received and decrypted by the receiver, it is no longer protected.

### 2.2 Image Watermarking

Watermarking techniques are then used in association with cryptography to provide lasting protection. Image watermarking is a technique for visibly or invisibly inserting information (watermark) into an image (host). Digital watermarking uses a secret key for insertion and detection according to Fig. 2.

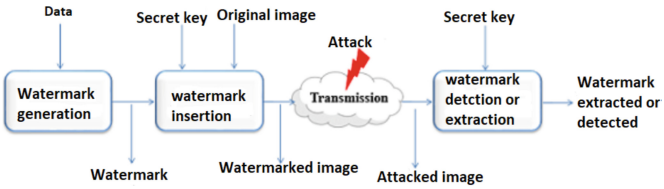



**Fig. 2.** watermarking process

Several watermarking algorithms have been proposed in the literature. These techniques can be classified according to Fig. 3.

Watermarking techniques can be used alongside cryptographic techniques to reinforce security. Watermarking techniques can be used alongside cryptographic techniques to enhance security. In [9], the various possibilities have been detailed.

### 2.3 The Different Step of Image Watermarking

**Watermark Generation.** The watermark generator takes a data item as a parameter in order to generate a mark according to the following formula:

$$GENERATOR(data) = watermark \tag{1}$$

In medical imaging, generators generally use meta-data. In [9], the author used the first letters of the surname and first name as data, and the generator is based on the Jocobian model. Some generators use chaotic functions to generate random sequences. In [10] the authors used a chaotic generator and the characteristics of the original image and the brand. They generated a unique secret key using some sub-band of the original image associated with the watermark. The key, numbers (randomly generated) and brand were used to calculate a matrix that would be sensitive to any modification of the key, and therefore to any modification of the brand or the original image. The difficulty with this approach is storing the key in a safe place. In [11] the authors have just used Arnold's transformation to encrypt a logo (watermark) using a key.

them, given their sensitive nature. Indeed, an altered image can cause a diagnostic or even therapeutic error. The information associated with these images must be protected in order to preserve medical confidentiality. Cryptography, which is generally used to secure documents on unsecured networks, is proving ineffective. Cryptography uses a secret key to encrypt a message and transfer it to a receiver. The receiver also uses a key to decrypt the message. The problem with cryptography is that it offers protection only at the point of transfer (a priori protection). To perpetuate this protection, digital watermarking is used as a complement to cryptography. The aim of watermarking is to conceal information (mark or watermark) in an image. Several watermarking techniques have been proposed in the literature [1–3].

These techniques are adapted to specific use cases and document types. To our knowledge, there is no universal watermarking scheme that adapts to any situation. In our context, our scheme must be adapted to a distributed database with a hyperbolic structure [6]. The hyperbolic structure is represented by a Poincaré disk in which a hyperbolic tree is constructed. Our approach aims to select pixels from a DICOM image and map them to server nodes in the database. To achieve this, we will adapt the structure of the given database to the image plane of the hyper-catadioptric system model. The advantage of this approach is that the schema introduces no distortion and is unpredictable. The approach must also be robust to compression and attack attacks.

Our article is structured as follows: first, we give an overview of image protection and present a database. Next, we study our hypercatadioptric model before proposing our new watermarking scheme. Moreover, we analyze our approach. We end with a conclusion and some perceptive remarks.

## 2 Image Protection

### 2.1 Image Encryption

Cryptography is the method used to secure exchanges in applications. It involves encrypting a document using a key, then transferring it to a recipient. The receiver uses a key to decrypt the document (Fig. 1).

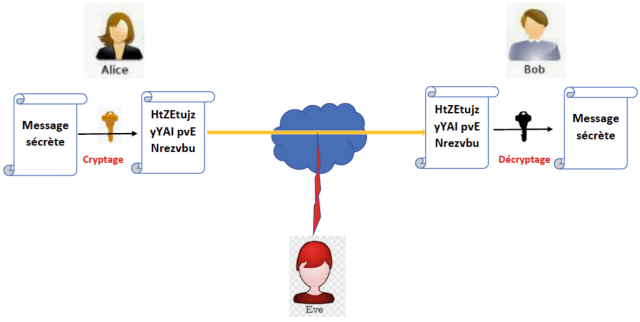



**Fig. 1.** Cryptography

# New Zero Watermarking Scheme Based on Hyper-catadioptric System Model and Hyperbolic Geometry

Boureima Koussoube[1(✉)], Moustapha Bikienga[2], Telesphore Tiendrebeogo[1], Kodjo Atiampo Armand[3], and Boureima Zerbo[4]

[1] Nazi BONI University, Bobo-Dioulasso, Burkina Faso
koussoubebrm@gmail.com
[2] Norbert ZONGO University, Koudougou, Burkina Faso
[3] Virtual University, Abidjan, Ivory Coast
armand.atiampo@uvci.edu.ci
[4] Thomas SANKARA University, Saaba, Burkina Faso

**Abstract.** In this paper we propose a new digital watermarking scheme for securing DICOM images in a distributed database. This new technique introduces no distortion to the images and will serve as a means of authenticating them. The database has a hyperbolic structure and its model is based on the Poincaré disk model, in which a hyperbolic tree is built. The coordinates of the tree nodes represent the virtual coordinates of the virtual servers. We assimilate the database structure to the image plane of a hyper-catadioptric system model. The image will be placed in a Euclidean space. Points will then be selected and computed according to the projection model of the hyper-catadioptric system. The set of image points computed constitutes our cryptographic signature. Each image point will be associated with the nearest node, and each server node will store image points and a model parameter, the plane equation and one of its public keys. Using its private key, the receiver can determine the image points and the various parameters to calculate the inverse transform of each image point for comparison. Formal analysis and simulations show that our approach is robust.

**Keywords:** Zero watermarking scheme · Hyper-catadioptric system model · Hyperbolic tree · Cryptographic signature

## 1 Introduction

The expansion of new information and communication technologies is making itself felt in the field of medical imaging. It is no longer limited to the production of medical images, but also to the manipulation, storage and transfer of images associated with some information. Since these images and information are transmitted over unsecured networks, it is essential to find safe ways of protecting


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 220–233, 2025.
https://doi.org/10.1007/978-3-031-81573-7_18

13. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI Mag. **24**(3) (2003). Art. n$^{o}$ 3, https://doi.org/10.1609/aimag.v24i3.1720
14. Hagberg, S.: "Bobo buveurs, Yarse colporteurs": Parenté à plaisanterie dans le débat public burkinabè ("Drinking Bobo, Trading Yarse". Joking Kinships in the Burkinabe Public Debate). Cah. DÉtudes Afr. **46**(184), 861–881 (2006)

Third, more work could be done to develop applications that make use of the ontology. This could involve developing more advanced event management systems, cultural heritage preservation tools, and tourism platforms that leverage the ontology to provide more rich and nuanced representations of culture.

Finally, more work could be done to evaluate the effectiveness of the ontology in supporting these applications. This could involve conducting user studies to assess the usability and usefulness of the applications and conducting empirical studies to assess the accuracy and completeness of the ontology.

## References

1. White, L.A.: Culture. Encyclopedia Britannica, 5 août 2022. https://www.britannica.com/topic/culture. consulté le 16 juillet 2023
2. Diallo, Y.: Joking relationships in Western Burkina Faso. Z. Für Ethnol. **131**(2), 183–196 (2006)
3. Hammond, P.B.: Mossi joking. Ethnology **3**(3), 259–267 (1964). https://doi.org/10.2307/3772882
4. Kaladzavi, G., Diallo, P.F., Kolyang, Lo, M.: OntoSOC: Sociocultural Knowledge Ontology (2015). https://airccse.org/journal/ijwest/papers/6215ijwest01.pdf
5. Heimbürger, A.: When Cultures Meet: Modelling Cross-Cultural Knowledge Spaces, vol. 166, p. 9
6. Berners-Lee, T., Hendler, J.: Publishing on the semantic web. Nature **410**, 1023–1024 (2001). https://doi.org/10.1038/35074206
7. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**(2), 199–220 (1993). https://doi.org/10.1006/knac.1993.1008
8. Hyvönen, E.: Cultural content creation. In: Hyvönen, E. (ed.) Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on Data, Semantics, and Knowledge, pp. 87–105. Springer, Cham (2012). https://doi.org/10.1007/978-3-031-79438-4_7
9. Bruseker, G., Carboni, N., Guillem, A.: Cultural heritage data management: the role of formal ontology and CIDOC CRM. In: Vincent, M., López-Menchero Bendicho, V., Ioannides, M., Levy, T. (eds.) Heritage and Archaeology in the Digital Age. Quantitative Methods in the Humanities and Social Sciences, pp. 93–131. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65370-9_6
10. Hyvönen, E.: Logic rules for cultural heritage. In: Hyvönen, E. (ed.) Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on Data, Semantics, and Knowledge, pp. 79–86. Springer, Cham (2012). https://doi.org/10.1007/978-3-031-79438-4_6
11. Hyvönen, E.: Cultural heritage on the semantic web. In: Hyvönen, E. (ed.) Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on Data, Semantics, and Knowledge, pp. 1–11. Springer, Cham (2012). https://doi.org/10.1007/978-3-031-79438-4_1
12. Carroll, J.J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., Wilkinson, K.: Jena: implementing the semantic web recommendations. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, in WWW Alt. '04, mai 2004, pp. 74–83. Association for Computing Machinery, New York, NY, USA (2004). https://doi.org/10.1145/1013367.1013381

```
SELECT ?event ?eventName ?eventLocation ?eventDate
WHERE {
  ?event rdf:type ex:Cultural_Event.
  ?event ex:takesPlaceIn ex:BoboDioulasso.
  ?event ex:eventName ?eventName.
  ?event ex:eventLocation ?eventLocation.
  ?event ex:eventDate ?eventDate.
}
```

This SPARQL query retrieves a list of cultural events happening in Bobo-Dioulasso, including details about their types, locations, and scheduled dates and times, based on the ontology. This data can be extremely useful for cultural event management and the promotion of cultural heritage in Bobo-Dioulasso a city of Burkina Faso.

## 7 Conclusion

In this paper, we have presented an ontology dedicated to modeling cultural concepts, with a particular focus on African cultures and the unique social practice of joking kinship. The ontology provides a comprehensive framework for representing, analyzing, and understanding culture, encompassing a variety of aspects such as art, religion, mythology, cuisine, history, education, economy, social structure, and locality.

Designed to integrate with Semantic Web technologies, the ontology facilitates the development of applications in the fields of cultural event management, cultural heritage preservation, and tourism. It also serves as a solid foundation for in-depth studies on specific cultural practices and phenomena, such as the system of joking kinship in Burkina Faso.

Beyond these practical applications, our work significantly contributes to the preservation of African cultural heritage. By codifying and making accessible the wealth of our traditions and history, the ontology acts as a guardian of our cultural legacy, ensuring its transmission to future generations. In a constantly evolving world, where cultures are under increasing pressure, this tool proves indispensable for maintaining our connection to our roots and strengthening cohesion within our communities. Ultimately, it represents a crucial step towards safeguarding our cultural identity and promoting a more united and resilient society.

## 8 Future Work

Looking forward, there are several directions for future work. First, the ontology could be expanded and enriched further to cover additional cultural practices and phenomena, both within Africa and in other parts of the world. This would involve conducting more fieldwork and ethnographic studies to gather detailed information about these practices and phenomena, and then incorporating this information into the ontology.

Second, more work could be done to develop and refine the rules and reasoning capabilities of the ontology. This would involve defining more specific rules for inferring new knowledge based on existing relationships and developing more sophisticated reasoning algorithms to process these rules.

In Fig. 4 we present an example of how the ontology is populated, along with a SPARQL query that can be used to retrieve communities that have a joking kinship relationship with the Peulh community.

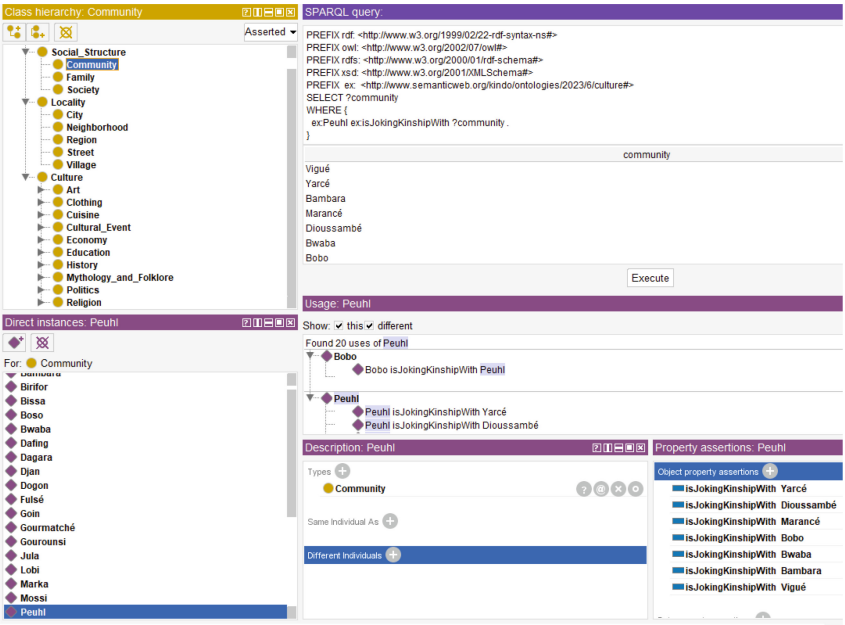



**Fig. 4.** Populating and query example in Protégé

This query will return all communities that have a joking kinship relationship with the Bobo community. The prefix ex: is used to denote the namespace of the ontology, and isJokingKinshipWith is the object property that represents the joking kinship relationship.

Use Case 2: Cultural Event Management

In African societies, cultural events serve as a vibrant showcase of the rich diversity and heritage, fostering community engagement and cultural preservation. Our ontology facilitates efficient organization, access, and dissemination of information related to these events. For example, to retrieve comprehensive details of all cultural events occurring in a specific location, such as BoboDioulasso, the following SPARQL query can be employed:

Language(?x) ^ isSpokenIn(?x, ?y) ^ Community(?y) - > FacilitatesCommunication(?x, ?y).

This rule stipulates that if a language is spoken within a community, it facilitates communication within that community.

Example 3: Mediation in Case of Conflict.

ConflictBetween(?x, ?y) ^ isJokingKinshipWith(?y, ?z) - > CanRequestMediation(?x, ?z).

This rule models the cultural practice whereby a community in conflict with another can request mediation from a third community linked by a joking kinship.

By integrating these SWRL rules into our ontology, we enhance its ability to model complex relationships and infer specific cultural knowledge, contributing to a richer understanding and more effective preservation of African cultural heritage.

## 6 Utilizing the Ontology: Use Cases and SPARQL Queries

In this section, we demonstrate the practical application of our ontology through two distinct use cases. These use cases illustrate how the ontology can be used to query and extract meaningful information related to African culture, particularly in the context of joking kinship and cultural events. We use SPARQL, a semantic query language for databases, to retrieve and manipulate data stored in our ontology.

Use Case 1: Joking Kinship Relationships Among Communities in Burkina Faso

In Burkina Faso, the practice of “joking kinship” or “parenté à plaisanterie” is a common cultural tradition among various communities. This tradition allows for friendly teasing and banter between members of different communities, fostering social cohesion and mutual respect.

Consider the communities of “Bobo”, “Peulh”, “Dafin”, “Dagara”, “Goulmanché”, “Mossi”, “Samo”, “Bissa”, “Yarcé”, “Yadga”, “Lobi”, “Birifor”, “Foulssé”, “Bawba”, “Tchèfo”, “Gourmantché”, and “Gourounsi”. These communities are diverse in their languages, customs, and traditions, yet they are bound by the joking kinship relationships. These relationships allow members of these communities to joke with each other, often in ways that would be considered disrespectful or offensive outside of this context. However, within the context of the joking kinship, these interactions are not only accepted but are also expected and appreciated.

For instance, a member of the “Bobo” community might tease a member of the “Peulh” community about a particular custom or tradition. The “Peulh” member would then respond with a similar jest about the “Bobo” community. This exchange, while seemingly contentious, actually serves to strengthen the bond between the two communities. It allows them to acknowledge and celebrate their differences, rather than allowing these differences to divide them.

This practice extends to all the mentioned communities, creating a complex web of relationships and interactions that serve to promote unity and understanding among the diverse communities of Burkina Faso. Through the joking kinship, these communities are able to maintain their unique identities while also fostering a sense of shared identity and mutual respect [14].

This ontology provides a robust and flexible framework for representing, analyzing, and understanding culture. It can be used to capture the richness and diversity of cultural practices, traditions, and artifacts, and to promote social cohesion and cultural preservation. The ontology is designed to be extensible, allowing for the addition of new classes and properties as new cultural phenomena are discovered and documented (Figs. 1, 2 and 3).


**Fig. 1.** Ontology implementation with protégé

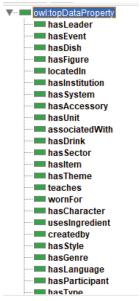



**Fig. 2.** Data properties

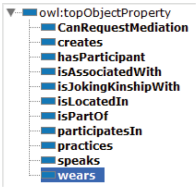



**Fig. 3.** Object properties

## 5 Rules and Reasoning Capabilities

Our ontology leverages SWRL to formulate rules that enhance reasoning capabilities and enable the inference of new knowledge from existing data. SWRL provides a formal syntax for expressing these rules, facilitating their integration and execution within the ontology. Below are illustrative examples:

Example 1: Joking Kinship.

Community(?x) ^ Community(?y) ^ isJokingKinshipWith(?x, ?y) - > PermitsJoking(?x, ?y).

This SWRL rule expresses that if two individuals belong to communities and are connected by a joking kinship, then they are allowed to tease each other.

Example 2: Languages and Communication.

**Table 1.** Class and data properties

| Class | Properties |
|---|---|
| Cultural Event | hasType, hasLanguage, takesPlaceIn, hasParticipant |
| Art | hasType, createdby, hasGenre, hasStyle |
| Religion | practicedAt, hasBelief, hasRitual, hasSymbol |
| Mythology and Folklore | hasType, associatedWith, hasCharacter, hasTheme |
| Cuisine | hasDish, hasDrink, usesIngredient |
| Clothing | hasItem, hasAccessory, associatedWith, wornFor |
| History | hasEvent, hasFigure, associatedWith |
| Education | hasSystem, hasInstitution, teaches, locatedIn |
| Economy | hasSector, associatedWith, locatedIn |
| Politics | hasSystem, hasLeader, associatedWith |
| Social Structure | hasUnit, associatedWith, locatedIn |
| Locality | hasPart |

**Table 2.** Object properties with description

| Object Property | Description |
|---|---|
| isPartOf | For hierarchical relationships |
| isAssociatedWith | For general associations |
| isLocatedIn | For spatial relationships |
| hasParticipant | For events |
| practices | For religious or cultural practices |
| speaks | For languages |
| creates | For artistic creation |
| wears | For clothing |
| participatesIn | For social structures |
| isJokingKinshipWith | For the joking kinship relationship in West African cultures |
| canRequestMediation | Represents the ability to request mediation or intervention in conflict resolution or dispute settlement |

These object properties can be used to establish relationships between different instances of the classes in the ontology. For example, the “isJokingKinshipWith” property can be used to establish a joking kinship relationship between two individuals in the “Community” class.

interconnected through a set of properties that express their relationships and attributes. As a domain ontology, it specifically focuses on the domain of culture and is designed to facilitate the analysis and understanding of culture.

- Culture: The overarching class that encompasses all aspects of culture. It is associated with various properties such as hasEvent, hasArt, hasReligion, hasMythologyAndFolklore, hasCuisine, hasClothing, hasHistory, hasEducation, hasEconomy, hasPolitics, hasSocialStructure, and hasLocality. These properties link the Culture class to the respective subclasses, creating a comprehensive network of cultural elements.
- Cultural Event: This class includes sub-classes like Festival, Exhibition, Concert, Dance. Each of these subclasses can be linked to the Culture class through the hasEvent property.
- Art: This class is divided into sub-classes such as Music, VisualArt, Literature, and PerformanceArt. The hasArt property connects these subclasses to the Culture class.
- Religion: This class has a sub-class Temple, which can be linked to the Culture class through the hasReligion property.
- Mythology and Folklore: This class includes sub-classes like Myths, Tales, and Legends, which can be linked to the Culture class through the hasMythologyAndFolklore property.
- Cuisine: This class has sub-classes such as TraditionalDish, TraditionalDrink, and Ingredients. The hasCuisine property connects these subclasses to the Culture class.
- Clothing: This class includes sub-classes like TraditionalClothing and Jewelry, which can be linked to the Culture class through the hasClothing property.
- History: This class has sub-classes like HistoricalEvent and HistoricalFigure, which can be linked to the Culture class through the hasHistory property.
- Education: This class includes sub-classes like TraditionalEducationalSystem and EducationalInstitution, which can be linked to the Culture class through the hasEducation property.
- Economy: This class has sub-classes like Trade, Agriculture, and Industry, which can be linked to the Culture class through the hasEconomy property.
- Politics: This class includes sub-classes like TraditionalPoliticalSystem and PoliticalLeader, which can be linked to the Culture class through the hasPolitics property.
- Social Structure: This class has sub-classes like Family, Community, and Society, which can be linked to the Culture class through the hasSocialStructure property.
- Locality: This class includes sub-classes like Region, City, Neighborhood, Street, Village, which can be linked to the Culture class through the hasLocality property (Tables 1 and 2).

The primary objective of this culture ontology is to provide a structured and interconnected representation of the various aspects of culture. It aims to facilitate the analysis and understanding of culture, and to promote social cohesion and cultural preservation. The competency questions this ontology aims to answer include: What are the different aspects of culture? How are these aspects related to each other? What are the specific characteristics of each aspect of culture?

communities. The Semantic Web is a collaborative effort led by the World Wide Web Consortium (W3C), which promotes common formats for data on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web dominated by unstructured and semi-structured documents into a "web of data" [6].

Ontologies play a crucial role in the Semantic Web, providing a shared and common understanding of a domain that can be communicated between people and machines. They are used for knowledge representation and are essential for tasks such as data integration, content-based indexing, database design, and information retrieval. Ontologies can be used to improve the accuracy of web search and enable machine-automated reasoning [7].

In the context of enhancing culture, Semantic Web tools can be used to create a rich, interconnected digital representation of cultural heritage. This can be achieved by using ontologies to represent complex relationships between different cultural artifacts, historical events, and cultural groups. For example, an ontology could be used to represent the relationships between different works of art, the artists who created them, the historical periods they belong to, and the cultural movements they are associated with. This would allow users to navigate through cultural heritage in a non-linear way, discovering new connections and gaining a deeper understanding of culture [8].

In addition to these general-purpose tools, there are also tools specifically designed for the cultural domain. For instance, the CIDOC Conceptual Reference Model (CRM) provides an ontology for cultural heritage information, which is widely used in museums, libraries, and other cultural institutions [13].

While existing ontologies like CIDOC CRM offer valuable frameworks for cultural heritage representation, our project required a bespoke ontology to accurately capture the unique facets of African cultural contexts. This is particularly true for the joking kinship system in Burkina Faso, a complex social structure not adequately addressed by CIDOC CRM. Our ontology goes beyond, providing detailed modeling of cultural events, languages, arts, and social structures, with a specific emphasis on their manifestation in African societies. By doing so, we ensure a comprehensive and culturally relevant tool, not just for representing joking kinship, but for a holistic analysis and preservation of African cultural heritage.

Furthermore, Semantic Web technologies can be used to connect cultural heritage data with other types of data, such as geographical data or social network data, providing new perspectives and insights. For example, it would be possible to explore the influence of a particular cultural movement on the works of art created in different cities, or to analyze the social network of a group of artists to understand their collaborations and influences.

## 4 Modelling the Ontology of Culture

The ontology of culture we have constructed here is a comprehensive domain ontology that encapsulates various aspects of culture, including events, art, religion, mythology, folklore, cuisine, clothing, history, education, economy, politics, social structure, and locality. Each of these aspects is represented as a class in the ontology, and they are

However, the rapid pace of modernization and globalization poses challenges, threatening to erode these traditional practices and potentially weakening the social fabric. Younger generations may find themselves disconnected from these cultural roots, underscoring the urgent need for tools that preserve and promote cultural knowledge and practices.

In response to this challenge, this paper delves into the ontology of culture, with a particular focus on African culture and the system of joking kinship in Burkina Faso. We present a comprehensive ontology that encapsulates various cultural dimensions, including art, religion, mythology, cuisine, and social structures. Our objective is to provide a robust framework for the representation, analysis, and understanding of culture, aiming to bolster social cohesion and preserve the rich cultural heritage that is integral to African societies.

By doing so, we contribute to the ongoing efforts to maintain cultural continuity, ensure the transmission of traditional knowledge to future generations, and uphold the social mechanisms that have sustained African communities for centuries.

## 2 Related Work

The study of culture and its representation in the digital space has been a topic of interest for many researchers. Here, we review some of the significant works related to our study.

OntoSOC: Sociocultural Knowledge Ontology by Guidedi Kaladzavi et al. [4] presents a sociocultural knowledge ontology modeling approach based on Engestrom Human Activity Theory (HAT). The ontology is designed to organize data, facilitate information retrieval, and introduce a semantic layer in the social web platform architecture. The authors envision the platform as a collective memory and Participative and Distributed Information System (PDIS) that allows communities to share and co-construct knowledge on permanent organized activities.

When Cultures Meet: Modelling Cross-Cultural Knowledge Spaces by Anneli Heimbürger [5] introduces an idea for constructing an information system, a cross-cultural knowledge space, which could support cross-cultural communication, collaborative learning experiences, and time-based project management functions. The system design is based on a cross-cultural ontology, and the system implementation on XML technologies. The author discusses the concept of time in a cultural context and the role of cultural competence in achieving project goals and promoting creativity and motivation through flexible leadership.

These works provide valuable insights into the representation of cultural knowledge in the digital space and the use of ontologies for structuring and retrieving this knowledge. Our study builds upon these works by focusing on the cultural experience, including the African culture, and proposing an ontology that captures various aspects of this experience.

## 3 Semantic Web Tool to Enhance Culture

The Semantic Web is an extension of the World Wide Web that enables people to share content beyond the boundaries of applications and websites. It has been described as a framework that allows data to be shared and reused across applications, enterprises, and


# Culture Ontology to Enhance Social Cohesion

Abdoul Azize Kindo[1(✉)], Gaoussou Camara[2], Sadouanouan Malo[1], Guidedi Kaladzavi[3], Théodore Marie Yves Tapsoba[1], and Kolyang[3]

[1] Ecole Supérieure d'Informatique, University of Nazi Boni, Bobo-Dioulasso, Burkina Faso
kindoazize@gmail.com
[2] Department of Mathematics, University of Alioune Diop of Bambey, Bambey, Senegal
[3] Department of Computer Science and Telecommunications, University of Maroua, Maroua, Cameroon

**Abstract.** Culture plays a pivotal role in shaping our understanding of the world, guiding our interactions, and influencing our behavior. In African societies, culture often lies at the heart of social cohesion and collective identity, providing a sense of belonging and a framework for social navigation. This paper presents an ontology of culture, with a focus on African culture, and specifically on the system of joking kinship in Burkina Faso. The ontology covers various aspects of culture including art, religion, mythology, cuisine, history, education, economy, social structure, and locality . The ontology is designed to provide a framework for representing, analyzing, and understanding culture, with the aim of promoting social cohesion and preserving cultural heritage. The ontology is enriched with properties and object properties, including a unique object property representing the system of joking kinship, a traditional social relationship that fosters tolerance, facilitates conflict resolution, and contributes to social harmony. The ontology is designed to be used in conjunction with Semantic Web technologies to support various applications related to cultural events and practices.

**Keywords:** Culture · ontology · Social Cohesion · Semantic Web · rules

## 1 Introduction

Culture, as a complex system of knowledge, beliefs, art, law, morals, customs, and capabilities acquired by individuals as members of society, plays an indispensable role in fostering social cohesion and shaping collective identity. This is particularly evident in African societies, where cultural practices and social structures provide a framework for interactions, influence behavior, and contribute to a sense of belonging [1].

A quintessential manifestation of this cultural influence on social cohesion is the system of joking kinship, prevalent in many African communities, including Burkina Faso. This unique social relationship, transcending blood ties, allows for open communication and jesting, fostering bonds, promoting tolerance, and facilitating conflict resolution. It is a vital mechanism for maintaining social harmony and strengthening communal ties [2, 3].


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 209–219, 2025.
https://doi.org/10.1007/978-3-031-81573-7_17

41. Singh, A., Jain, A., Biable, S.E.: Financial fraud detection approach based on firefly optimization algorithm and support vector machine. Appl. Comput. Intell. Soft Comput. **2022** (2022). https://doi.org/10.1155/2022/1468015
42. Singh, K., Best, P.: Anti-money laundering: using data visualization to identify suspicious activity. Int. J. Account. Inf. Syst. **34**, 100418 (2019). https://doi.org/10.1016/j.accinf.2019.06.001
43. Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C.B., Goldstein, T.: Saint: improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint arXiv:2106.01342 (2021). https://doi.org/10.48550/arXiv.2106.01342
44. Soni, V.D.: Role of artificial intelligence in combating cyber threats in banking. Int. Eng. J. Res. Dev. **4**(1), 7–7 (2019)
45. Tang, Q., et al.: Prediction of casing damage in unconsolidated sandstone reservoirs using machine learning algorithms. In: 2019 IEEE International Conference on Computation, Communication and Engineering (ICCCE), pp. 185–188. IEEE (2019). https://doi.org/10.1109/ICCCE48422.2019.9010785
46. Zhang, Y., Tong, J., Wang, Z., Gao, F.: Customer transaction fraud detection using xgboost model. In: 2020 International Conference on Computer Engineering and Application (ICCEA), pp. 554–558. IEEE (2020). https://doi.org/10.1109/ICCEA50009.2020.00122

26. Lopez-Rojas, E.A., Axelsson, S., Baca, D.: Analysis of fraud controls using the PaySim financial simulator. Int. J. Simul. Process Model. **13**(4), 377–386 (2018). https://doi.org/10.1504/IJSPM.2018.093756
27. Luengo, J., Fernández, A., García, S., Herrera, F.: Addressing data complexity for imbalanced data sets: analysis of smote-based oversampling and evolutionary undersampling. Soft. Comput. **15**, 1909–1936 (2011). https://doi.org/10.1007/s00500-010-0625-8
28. Luke, S., Cioffi-Revilla, C., Panait, L., Sullivan, K., Balan, G.: Mason: a multiagent simulation environment. Simulation **81**(7), 517–527 (2005). https://doi.org/10.1177/0037549705058073
29. Luke, S., et al.: The MASON simulation toolkit: past, present, and future. In: Davidsson, P., Verhagen, H. (eds.) MABS 2018. LNCS (LNAI), vol. 11463, pp. 75–86. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22270-3_6
30. MoMTSim, Inc: Mobile Money Transaction Simulator (2023). https://github.com/aiinfinancegroup/MoMTSim, version 0.1.0
31. Mudiri, J.L.: Fraud in mobile financial services. Rapport technique, MicroSave **30** (2013)
32. Muthali, A., et al.: Multi-agent reachability calibration with conformal prediction. arXiv preprint arXiv:2304.00432 (2023). https://doi.org/10.48550/arXiv.2304.00432
33. Narayan, A., Madhu Kumar, S., Chacko, A.M.: A review of financial fraud detection in e-commerce using machine learning. In: Intelligent Data Engineering and Analytics: Proceedings of the 10th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2022), pp. 237–248. Springer, Heidelberg (2023). https://doi.org/10.1007/978-981-19-7524-0_21
34. Nti, I.K., Somanathan, A.R.: A scalable rf-xgboost framework for financial fraud mitigation. IEEE Trans. Comput. Social Syst. (2022). https://doi.org/10.1109/TCSS.2022.3209827
35. Nunes, R.P.M., Bonacin, R., de Franco Rosa, F.: Methods for detecting fraud in civil and military service examinations: a systematic mapping. In: Latifi, S. (ed.) ITNG 2021 18th International Conference on Information Technology-New Generations. AISC, vol. 1346, pp. 203–208. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70416-2_26
36. Öztürk, M.S., Usul, H.: Detection of accounting frauds using the rule-based expert systems within the scope of forensic accounting. In: Contemporary Issues in Audit Management and Forensic Accounting, vol. 102, pp. 155–171. Emerald Publishing Limited (2020). https://doi.org/10.1108/S1569-375920200000102013
37. Park, J., Kwon, S., Jeong, S.P.: A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on smote and generative adversarial networks. J. Big Data **10**(1), 1–16 (2023). https://doi.org/10.1186/s40537-023-00715-6
38. Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020). https://doi.org/10.48550/arXiv.2010.16061
39. Raiter, O.: Applying supervised machine learning algorithms for fraud detection in anti-money laundering. J. Modern Issues Bus. Res. **1**(1), 14–26 (2021). https://doi.org/10.17613/2g0z-0814
40. Shahana, T., Lavanya, V., Bhat, A.R.: State of the art in financial statement fraud detection: a systematic review. Technol. Forecast. Soc. Chang. **192**, 122527 (2023). https://doi.org/10.1016/j.techfore.2023.122527

11. Chhabra Roy, N., Prabhakaran, S.: Internal-led cyber frauds in Indian banks: an effective machine learning-based defense system to fraud detection, prioritization and prevention. Aslib J. Inf. Manag. **75**(2), 246–296 (2023). https://doi.org/10.1108/AJIM-11-2021-0339
12. Chicco, D., Jurman, G.: An invitation to greater use of matthews correlation coefficient (mcc) in robotics and artificial intelligence. Front. Rob. AI, 78 (2022). https://doi.org/10.3389/frobt.2022.876814
13. Cordasco, G., Scarano, V., Spagnuolo, C.: Distributed mason: a scalable distributed multi-agent simulation environment. Simul. Model. Pract. Theory **89**, 15–34 (2018). https://doi.org/10.1016/j.simpat.2018.09.002
14. Danenas, P.: Intelligent financial fraud detection and analysis: a survey of recent patents. Recent Patents Comput. Sci. **8**(1), 13–23 (2015)
15. Dighe, D., Patil, S., Kokate, S.: Detection of credit card fraud transactions using machine learning algorithms and neural networks: a comparative study. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–6. IEEE (2018). https://doi.org/10.1109/ICCUBEA.2018.8697799
16. Fehler, M., Klügl, F., Puppe, F.: Techniques for analysis and calibration of multi-agent simulations. In: Gleizes, M.-P., Omicini, A., Zambonelli, F. (eds.) ESAW 2004. LNCS (LNAI), vol. 3451, pp. 305–321. Springer, Heidelberg (2005). https://doi.org/10.1007/11423355_22
17. Gelb, A., Mukherjee, A.: Digital technology in social assistance transfers for covid-19 relief: lessons from selected cases. CGD Policy Paper **181** (2020)
18. Kanobe, F., Bwalya, K.J.: Snags in mobile money in developing economies. Electron. J. Inf. Syst. Dev. Countries **88**(3), e12181 (2022). https://doi.org/10.1002/isd2.12181
19. Kosolwattana, T., Liu, C., Hu, R., Han, S., Chen, H., Lin, Y.: A self-inspected adaptive smote algorithm (sasmote) for highly imbalanced data classification in healthcare. BioData Mining **16**(1), 15 (2023). https://doi.org/10.1186/s13040-023-00330-4
20. Lawson-Lartego, L., Cohen, M.J.: 10 recommendations for African governments to ensure food security for poor and vulnerable populations during covid-19. Food Secur. **12**(4), 899–902 (2020). https://doi.org/10.1007/s12571-020-01062-7
21. Lin, J.: Backtracking search based hyper-heuristic for the flexible job-shop scheduling problem with fuzzy processing time. Eng. Appl. Artif. Intell. **77**, 186–196 (2019). https://doi.org/10.1016/j.engappai.2018.10.008
22. Lokanan, M.E., Sharma, K.: Fraud prediction using machine learning: the case of investment advisors in Canada. Mach. Learn. Appl. **8**, 100269 (2022). https://doi.org/10.1016/j.mlwa.2022.100269
23. Lopez-Rojas, E., Elmir, A., Axelsson, S.: PaySim: a financial mobile money simulator for fraud detection. In: 28th European Modeling and Simulation Symposium, EMSS, Larnaca, pp. 249–255. Dime University of Genoa (2016)
24. Lopez-Rojas, E.A., Barneaud, C.: Advantages of the PaySim simulator for improving financial fraud controls. In: Arai, K., Bhatia, R., Kapoor, S. (eds.) CompCom 2019. AISC, vol. 998, pp. 727–736. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22868-2_51
25. Lopez-Rojas, E.A.: Extending the retsim simulator for estimating the cost of fraud in the retail store domain. In: The 27th European Modeling and Simulation Symposium-EMSS, Bergeggi, Italy (2015)

Our future work shall focus on incorporating other machine learning models and the simulation of more fraudulent scenarios for the task of fraud detection as well as preserving privacy in the synthetic data.

**Acknowledgements.** This research was made possible in part by the Digital Credit Observatory (DCO), a program of the Center for Effective Global Action (CEGA), with support from the Bill & Melinda Gates Foundation; JPMorgan Chase & Co.; and Google PhD Fellowship Program. Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by any funder or its affiliates. A number of financial institutions in the Sub-Saharan region provided expert opinions on the dynamics of the real financial ecosystem.

## References

1. Adewumi, A.O., Akinyelu, A.A.: A survey of machine-learning and nature-inspired based credit card fraud detection techniques. Int. J. Syst. Assur. Eng. Manag. **8**, 937–953 (2017). https://doi.org/10.1007/s13198-016-0551-y
2. Aftabi, S.Z., Ahmadi, A., Farzi, S.: Fraud detection in financial statements using data mining and gan models. Expert Syst. Appl. **227**, 120144 (2023). https://doi.org/10.1016/j.eswa.2023.120144
3. Alghofaili, Y., Albattah, A., Rassam, M.A.: A financial fraud detection model based on lstm deep learning technique. J. Appl. Secur. Res. **15**(4), 498–516 (2020). https://doi.org/10.1080/19361610.2020.1815491
4. Ali, A.A., Khedr, A.M., El-Bannany, M., Kanakkayil, S.: A powerful predicting model for financial statement fraud based on optimized xgboost ensemble learning technique. Appl. Sci. **13**(4), 2272 (2023). https://doi.org/10.3390/app13042272
5. Ali, G., Ally Dida, M., Elikana Sam, A.: Evaluation of key security issues associated with mobile money systems in Uganda. Information **11**(6), 309 (2020). https://doi.org/10.3390/info11060309
6. Apiors, E.K., Suzuki, A.: Effects of mobile money education on mobile money usage: Evidence from ghana. Eur. J. Dev. Res. 1–28 (2022). https://doi.org/10.1057/s41287-022-00529-x
7. Aslam, N., et al.: Anomaly detection using explainable random forest for the prediction of undesirable events in oil wells. Appl. Comput. Intell. Soft Comput. **2022** (2022). https://doi.org/10.1155/2022/1558381
8. Aswathi, M., Ghosh, A., Namboothiri, L.V.: Borda count versus majority voting for credit card fraud detection. In: Karuppusamy, P., Perikos, I., García Márquez, F.P. (eds.) Ubiquitous Intelligent Systems. SIST, vol. 243, pp. 319–330. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-3675-2_24
9. Bagga, S., Goyal, A., Gupta, N., Goyal, A.: Credit card fraud detection using pipeling and ensemble learning. Procedia Comput. Sci. **173**, 104–112 (2020). https://doi.org/10.1016/j.procs.2020.06.014
10. Botchey, F.E., Qin, Z., Hughes-Lartey, K.: Mobile money fraud prediction'a cross-case analysis on the efficiency of support vector machines, gradient boosted decision trees, and naïve bayes algorithms. Information **11**(8), 383 (2020). https://doi.org/10.3390/info11080383

sion and recall for fraudulent transactions but it has less effective performance than the boosting models. Even though the Logistic regression and Decision tree algorithms registered relatively lower MCC scores, they still showed remarkable performance for the task.

Financial institutions and service providers usually consider a number of factors for instance the computational costs and the complexity of a model before committing themselves to use it in real-world applications. However, most of them rely on model performance evaluation metrics that provide a balanced measure and a model that offers superior performance. The XGBoost would be a model of choice for production though it is more computationally intensive than Logistic regression or Decision trees. Security teams and managers shall be capable of making decisions on what model to adopt especially where specific requirements are involved, considering a simpler model if computational resources are a constraint.

## 6 Conclusions and Future Work

This study demonstrates that financial institutions can stay ahead of fraudsters by simulating unique fraudulent behaviours in mobile money services using the MoMTSim platform. Besides, the study shows that synthetic data that statistically resembles real data can be used for research in the absence of real data.

This work provides a comprehensive evaluation of common machine learning algorithms including Logistic regression, Gradient boosting, Decision trees, AdaBoost, XGBoost, and Random forest, in terms of their capacity to detect fraudulent mobile money financial transactions. By using Logistic regression as a baseline model, the study offers a benchmark against which the performance of more complex models can be compared. This provides a clear insight into the improvements possible with more sophisticated techniques that might be of interest to researchers, service providers or financial institutions. The study emphasizes the use of the MCC metric, a more balanced measure for classification problems as in the case of mobile money fraud, especially in scenarios with imbalanced classes including financial fraud detection. More so, the study identifies the XGBoost model as the most effective algorithm for this particular task, as it achieved the highest MCC (0.82), high precision and recall scores. The high performance of the model on structured data is attributed to its capabilities of parallel processing, and regularisation that integrates L1 and L2 regularisation to prevent overfitting. The flexibility of the model allows the definition of custom optimization objectives and evaluation criteria more easily, unlike the other algorithms. Also, the study highlights the importance of considering the computational costs associated with different models, which is crucial when considering the practical application of the algorithms. Ultimately, this work contributes to the growing body of evidence supporting the use of machine learning algorithms for detecting mobile money fraud in financial transactions, especially in the Sub-Saharan context.

**Table 2.** Model performance results

| Model | Precision | Recall | F1-score | MCC |
|---|---|---|---|---|
| XGBoost | 0.98 | 0.75 | 0.85 | 0.82 |
| Gradient Boosting | 0.97 | 0.74 | 0.84 | 0.81 |
| AdaBoost | 0.96 | 0.74 | 0.84 | 0.80 |
| Random Forest | 0.93 | 0.76 | 0.84 | 0.79 |
| Decision Tree | 0.81 | 0.82 | 0.81 | 0.75 |
| Logistic Regression | 0.68 | 0.86 | 0.76 | 0.67 |

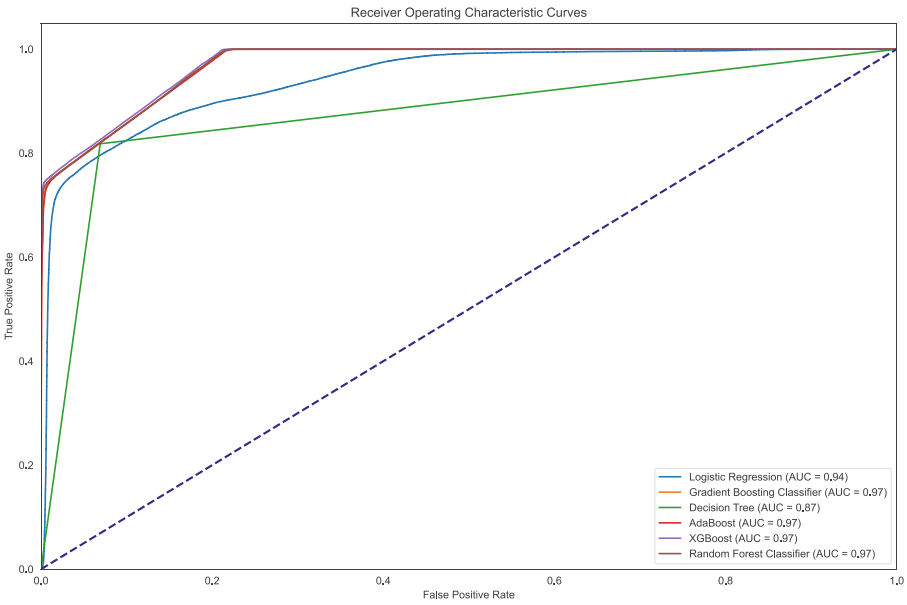



**Fig. 6.** ROC curves for the classification models used

Rule-based approaches underperform on either data mainly because of their inability to adapt to unanticipated scenarios in financial transactions.

The Gradient boosting classifier and AdaBoost also performed well, with MCCs of 0.81 and 0.80, respectively. In particular, the Gradient boosting classifier has a high MCC (0.81) and AUC of 0.97 implying overall good performance. The model has high precision which means the Gradient boosting classifier is good at identifying fraudulent mobile money transactions. AdaBoost showed similar performance to the Gradient boosting model, with a slightly lower MCC of 0.80. The Random forest classifier has an MCC of 0.79, which is good but still less than that one of XGBoost. It also has a high precision of 0.93 for fraudulent transactions, making it reliable in identifying fraudulent mobile money transactions. The Decision tree classifier has an MCC of 0.75, a balanced preci-

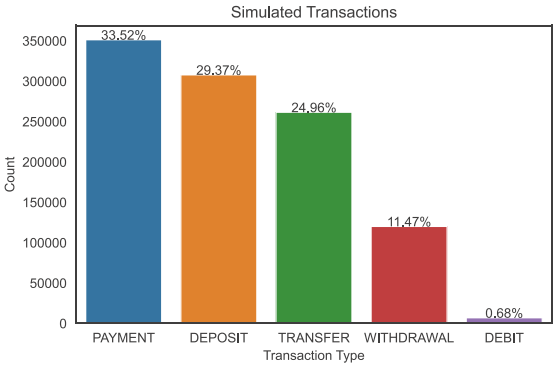



**Fig. 5.** Proportion of transactions simulated in MoMTSim_202306

**Model Performance.** The model performance results presented in this study are for the testing set and the Logistic regression model was used as the base model owing to its simplicity and interpretability. In the experiment, the Logistic regression model achieved an MCC of 0.67. The Logistic regression was able to detect fraudulent mobile money transactions with a high recall of 0.86, it was less precise, with a precision of 0.68 thus this performance served as the baseline. This implies that the model is good at identifying actual fraud cases, but it might also include more false positives, resulting in a lower precision. The model performances for all classifiers used in the experiment are shown in Table 2 and the ROC curves in Fig. 6. Other models including Gradient boosting, Decision trees, AdaBoost, XGBoost, and Random forest were then evaluated against the baseline model and they all demonstrated improvements with MCC scores ranging from 0.75 to 0.82. More specifically, the XGBoost model significantly outperformed the baseline model with an MCC of 0.82, revealing the effectiveness of more complex models in the realm of mobile money financial fraud detection. The XGBoost provides the best balance between precision and recall, among the models used. With the highest MCC (0.82), the highest precision of 0.98 for the fraudulent transactions and a higher AUC of 0.97, XGBoost offers a good balance, making it a preferred choice for the task compared to other models. By design, the XGBoost model supports parallel processing, making it quickly process large financial datasets and tree pruning allows for depth-first growth of trees and then pruning them, leading to more optimal trees, unlike Gradient boosting. Besides, it is flexible, allowing the definition of custom optimization objectives and evaluation criteria in order to fine-tune the model for specific needs. Moreover, it incorporates L1 (Lasso) and L2 (Ridge) regularization to prevent overfitting by minimizing the complexity of the model and distributing the weights more evenly across all the features [34,45,46]. Therefore, the XGBoost registers high performance on structured data compared to state-of-the-art deep learning models which often perform better on unstructured data.

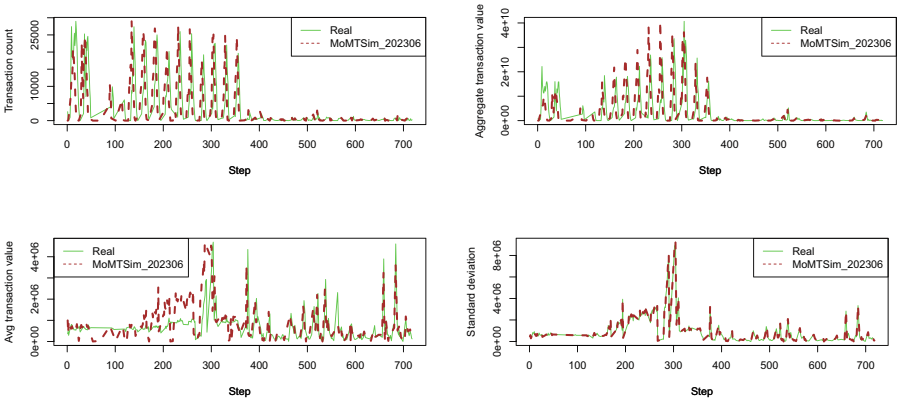



**Fig. 3.** Trends of transfer transactions.

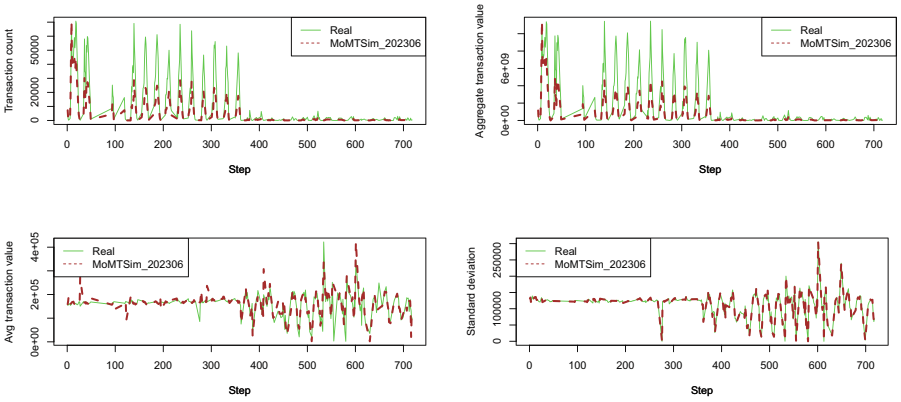



**Fig. 4.** Trends of deposit transactions

injected into MoMTSim mainly affected deposits, payments and transfers. Moreover, the fraudsters aim to carry out more transactions so as to increase their gains. A deposit is an entry point for hard cash into the mobile money system in the form of electronic money carried with the help of a mobile money merchant who facilitates the conversion. This implies that for other transactions to happen within the mobile money system, a deposit must have occurred initially. The debit transactions in the simulated data are the least frequent due to the clients avoiding high transaction charges associated with them. Usually, a client owning a bank account would prefer free deposits at the bank than a debit from their mobile money account which involves a service charge. Withdrawal is an exit point for converting electronic funds to hard cash ultimately a way to take money off the mobile money system. Usually, a number of clients engage in withdrawal transactions since cash is largely used in the Sub-Saharan region for daily purchases of goods and services.

other datasets registered relatively low total errors. The aggregated transactions for the real and synthetic datasets were visualised in order to compare the trends of transactions in the datasets. The transaction count, the aggregate transaction value, the average transaction value and the standard deviation for the real mobile money data and synthetic data were considered as shown in Figs. 2, 3 and 4. The trends in the payments, transfers and deposits were more of interest to show given that the fraud schemes discussed in Subsect. 1.1 mainly affected them. Other transaction types including debit and withdrawal that were present in the simulations showed statistical closeness even though their plots have not been included in this paper since they largely remained unaffected by the fraudulent activities. The uniform green line in all the plots indicates the trends of the real data while the dashed brown line indicates the trends of the synthetic data. Clearly, the trends are similar for both datasets and the small variations indicate that the datasets are not exactly the same. With this observation, MoMTSim generates synthetic datasets that statistically resemble the real data. Therefore, financial fraud classification using machine learning classifiers and rich synthetic data followed the assessment of the statistical closeness of the generated data to the real data.

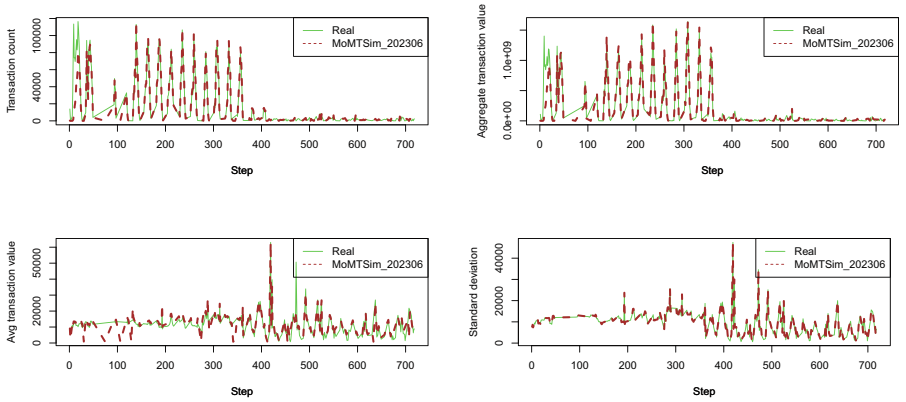



**Fig. 2.** Trends of payment transactions.

### 5.2 Mobile Money Fraud Classification Results

**Synthetic Transactions.** The simulated data containing 1,040,000 transactions had five transaction types composed of deposit, withdrawal, debit, payment and transfer. As shown in Fig. 5, 33.52% of the transactions are payments, followed by 29.37% deposits, 24.96% transfers, 11.47% withdrawals and 0.68% debits. In the simulated data, payment, deposit, and transfer transactions outnumber withdrawal and debit transactions. This is because the fraud schemes

two metrics (1),(2) and it is very useful especially when a financial institution is interested in a single model performance evaluation metric that combines recall and precision. Good financial fraud classification algorithms should have high scores for precision, recall and f1-score metrics [22]. The f1-score is expressed as

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

**Mathew Correlation Coefficient (MCC).** The MCC is a measure of the quality of binary classification. Unlike accuracy which deals with the proportion of correct predictions over all of the predictions, the MCC is a better measure and regarded as a balanced measure for classes with different sizes [12,38]. The capacity of the MCC to work well in scenarios where one class is more frequent than the other makes it a suitable metric for financial institutions to use [12,38]. The score is in the range of [-1,1], where a +1 corresponds to a perfect prediction while a −1 indicates an inverse prediction and a coefficient of 0 represents a random prediction [21]. The MCC is given by the relation

$$\text{MCC} = \frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{\sqrt{((\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN}))}}. \tag{4}$$

**Receiver Operating Characteristics (ROC).** Besides other common evaluation metrics, the ROC plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold values. The ROC curve that is closer to the left corner of the graph represents a good classification model while the one closer to the diagonal line represents a random model. The area under the ROC curve (AUC-ROC) was also used to compare the different fraud classification algorithms and its value ranges from 0 to 1. A value closer to 1 indicates good prediction while one closer to 0 is otherwise [10,22].

## 5 Results and Discussion

### 5.1 Conformity of Synthetic Datasets to Real Data

We executed MoMTSim [30] a number of times and several output files were written including synthetic mobile money transaction logs. Different simulations used different seed data as input. A complete simulation was executed for 720 steps, given that a step in the simulation platform represents one hour in the real world. This implies that a single simulation run represents one month (30 days) of transaction activity in the real ecosystem. The number of agents; mobile money merchants and clients were adjusted to output $1,040,000$ rows of transactions, sufficient for machine learning tasks. The resulting datasets were evaluated using the sum of squared errors (SSE) method and we obtained a dataset named MoMTSim_202306 to mean simulated for the month of June 2023 using MoMTSim. The dataset we picked for analysis had the least total error even though

was greater than two standard deviations above the mean transaction amount. Transactions that were abnormally large were more likely to be fraudulent and a feature was created. The dataset was normalized using the min-max scaler method since mobile money transactions do not follow the normal distribution. The dataset was split into a training set (70%) and a test set (30%), each containing legitimate and fraudulent transactions to ensure that training and testing were performed using distinct sets.

Fundamental metrics encompassing the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) form the basis for the model performance evaluation metrics that were used in the study. TP is concerned with an ML algorithm predicting that a transaction is fraudulent and the outcome is indeed fraud. On the other hand, FP involves the algorithm predicting a transaction to be fraudulent when actually it is not fraudulent. TN deals with a transaction predicted to be not fraudulent and there was no fraud while the FN which is the *hidden fraud* embraces the prediction of no fraud yet there was a fraudulent transaction. Common machine learning algorithms were adopted for financial fraud detection and their performances were evaluated to determine a consistent algorithm for the task [41].

### 4.4 Model Performance Evaluation

**Precision.** This is widely used and it is the ratio of correctly predicted positive observations to the total predicted positive observations given by the relation

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{1}$$

In practice, a balance needs to be established between the precision and recall for a financial fraud classification algorithm.

**Recall.** Sometimes referred to as sensitivity is the ratio of correctly predicted positive observations to all observations in the actual class, which is given by

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2}$$

Usually, a fraud classification algorithm with a higher recall but lower precision will correctly discern more of the fraudulent transactions and incorrectly predict more transactions to be fraudulent resulting in false positives. A classifier with a higher precision but lower recall will miss some fraudulent transactions but will not incorrectly predict too many transactions as fraudulent. Therefore, service providers and financial institutions aim to register a balance between the two metrics (1),(2) in order to achieve better results.

**F1-Score.** This is another common model performance evaluation metric concerned with the weighted average of precision and recall. F1-score balances the

Moreover, the simulated data used in our study contained 26.13% of fraudulent transactions making the synthetic data rich enough for fraud classification. The *transactionType* is categorical; deposit, withdrawal, transfer, payment and debit, and it was one-hot encoded in order to allow the machine learning models to utilise the feature as well as to improve model predictions.

**Table 1.** The independent features and the dependent variable in the synthetic mobile money transaction data

| Feature/Variable | Description | Measure |
|---|---|---|
| step | This maps a unit of time, a step in the simulation is an equivalent of one hour in the real world | Continuous |
| transactionType | Includes deposit, withdrawal, transfer, debit, and payment | Categorical |
| amount | Funds associated with a transaction type | Continuous |
| startingClient | Mobile money customer who initiates a transaction | Continuous |
| oldBalStartingClient | The starting balance of the client before initiating a transaction | Continuous |
| newBalStartingClient | The new balance of a client after initiating a transaction | Continuous |
| destinationClient | The recipient of funds after a transaction has taken place | Continuous |
| oldBalDestinationClient | The initial balance of the recipient client before a transaction is delivered | |
| newBalDestinationClient | The new balance of a recipient client after a transaction has taken place | Continuous |
| isFraud | The target variable, label 1 for fraud and 0 for a legitimate transaction | Categorical |

### 4.3 Feature Selection

Additional new features were created to improve model predictions. The balance differences between the old balance and the new balance for both the client starting the transaction and also for the recipient of a transaction were determined to form new features. This was aimed at identifying any significant changes in the account balance for the clients, potentially those associated with fraudulent transactions. The ratio of the transaction amount to the old balance of the starting client was also determined. A high transaction amount compared to the initial balance could be a signal of a fraudulent transaction. Similarly, the ratio between the transaction amount and the new balance of the starting client was determined. Also, large transactions were of interest, a transaction that

transactions by specifying higher probabilities of committing fraud. This allowed the generation of sufficient instances of fraudulent transactions in order to obtain rich synthetic datasets for fraud detection. Upon completion of all interactions in the simulation platform, diverse synthetic transaction files were written together with the parameter history and other log files as output.

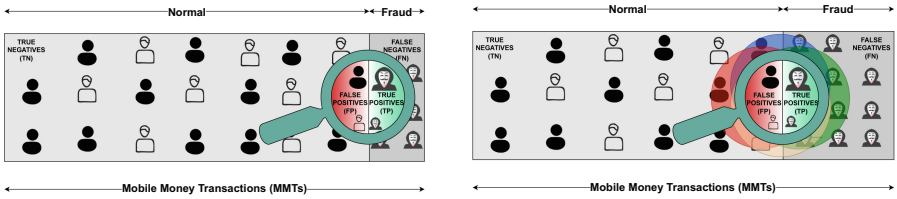
**Calibration and Validation of Simulations.** Calibration of simulation parameters was aimed at making sure that agents do not exhibit unusual behaviours [16,32]. Calibration also focused on avoiding the normative behaviour of agents since with agent-based modelling, various entities were modelled based on specific characteristics as observed from the real ecosystem. During this process, parameter sets leading to behaviours not present in the real ecosystem were removed. Documented practices guided by expert opinions and our understanding of the mobile money ecosystem were used to verify agent behaviours in simulations and statistical methods were used to assess the closeness of the synthetic data to the real data.

The conformity of the synthetic transaction datasets to the real data was measured using the sum of squared errors (SSE) method. We computed the difference between the real and synthetic data and the dataset with the least total error was selected for financial fraud detection using machine learning classifiers [26]. Other synthetic datasets that were not used for fraud classification still registered relatively low total errors implying they could as well be used for the same task.

### 4.2 Data Description, Cleaning and Preprocessing

With MoMTSim [30], we simulated 1,040,000 rows of mobile money transactions enriched using fraud schemes in Subsect. 1.1. The dataset contained 768,248 legitimate transactions while 271,752 were fraudulent transactions. The features in the synthetic data were based on those found in the real data with an addition of the target variable (label for fraud) in order to facilitate financial fraud detection using machine learning algorithms. The independent features and the dependent variable in the data are presented in Table 1. Unlike real data, incidences of missing data points in the synthetic data were eliminated during the design of the financial simulation platform. This implies that the traditional approaches for data cleaning do not apply to the resulting datasets which is unarguably one of the advantages of working with well-labeled synthetic datasets.

The identifiers for the client starting a transaction, *startingClient* and the recipient, *destinationClient* were removed prior to model building as they did not possess attributes that would affect fraud detection results. A common challenge with real financial data is the high-class imbalance, usually, researchers adopt for instance the SMOTE [8,27] to up-sample the minority class. However, the use of simulation addressed this challenge by generating sufficient instances of fraudulent transactions (see section 3) to enrich the data for financial fraud detection.




(a) Data with high-class imbalance. (b) Data with sufficient fraud.

**Fig. 1.** Mobile money transaction datasets for financial fraud detection

# 4 Methods

## 4.1 Financial Fraud Simulation Using MoMTSim

MoMTSim [30] is a multi-agent-based simulation (MABS) platform designed and calibrated using real mobile money transaction data to output diverse synthetic financial data. The core model in MoMTSim represents interactions of agents including clients and mobile money merchants based on probabilities extracted from the real data. A client has a profile, starting balance and other files containing properties of mobile money transactions; transaction types and aggregated transactions that are used during the simulations. Clients also participate in future transactions based on probabilities and their states are adjustable during a simulation. The entire simulation model is similar to a Markov process and agents carry common transactions that are present in the real mobile money ecosystem. The transaction types modelled include a deposit which is concerned with a client loading electronic money into their account via a mobile money merchant. A withdrawal is the opposite of a deposit, debit involves moving money from a mobile money account to a bank account. A transfer is concerned with the movement of electronic funds from one mobile money account to another account, while a payment includes the purchase of goods and services using electronic money in a mobile money account.

The object code implementation for the MoMTSim simulation platform uses a generic agent-based simulation toolkit MASON [13,28,29] which is fast enough and capable of handling large custom simulations. Besides MASON is multi-platform, supports parallelism and is capable of reinforcing computationally expensive simulations unlike NetLogo, Repast and AnyLogic [23,26]. A step in the simulation represents an hour in the real mobile money ecosystem(real world).

Fraud modelling in MoMTSim was carried out by defining specific fraud parameters for the fraud schemes discussed in Subsect. 1.1. Besides, transaction rules based on the fraudulent behaviours were defined in MoMTSim and a fraudulent client carries transactions in parallel with normal clients during a simulation. These transactions are contingent on probabilities of committing fraud, fraudster finding new victims and the chances of previous victims being at high risk for future fraud. We scheduled the fraudsters to fiercely carry out

level of expertise required and the desired level of accuracy and tolerance in a given regulatory environment.

Our study combines the simulation of contemporary fraudulent schemes in mobile money services with novel computational methods consisting of common ML classifiers for automated fraud detection. Besides, it addresses the challenge of class imbalance in real data through the generation of synthetic financial data that statistically resembles real transaction data.

## 3 The Challenge of Class Imbalance in Financial Data for Fraud Detection

Mobile money service providers and financial institutions face challenges with enriching their own data for effective fraud detection using machine learning techniques. At times, better algorithms need to be developed for efficient fraud detection with low false positives. The use of simulation to generate financial data with known fraud instances has been a major breakthrough for the financial industry. Simulation environments that are agent-based have yielded significant results in this domain [24,26]. For instance, Lopez-Rojas et al. [26] simulate a known financial crime pattern in the mobile money financial domain in order to generate fraudulent behaviour for fraud detection. Real financial datasets are highly imbalanced in nature and this makes detection of complex fraud patterns extremely difficult. Most studies in this domain rely on the use of the synthetic minority over-sampling technique (SMOTE) [8,27] to resolve the class imbalance in the real financial data with no consideration for the quality of the synthetic samples [4,19,37,40]. Simulation plays a crucial role in addressing this challenge given that the documented fraudulent behaviours in mobile money systems are accessible to the research community.

Figure 1a represents a typical real mobile money transaction data residing in the data warehouse of a service provider in the Sub-Saharan region. The rectangular block is a collection of genuine and very few fraudulent transactions. Clearly, the dataset exhibits a high-class imbalance for fraud detection using machine learning techniques. This challenge has made many of the service providers in the region use rule-based approaches that often result in many false positives even though they are easy to set up. Complex fraud patterns have been hardly studied and fraudsters are always adapting to the relatively straightforward control measures set by the service providers [26].

Figure 1b shows synthetic mobile money transaction data generated using agent-based modelling techniques with a sufficient amount of fraudulent transactions based on fraudulent behaviours in the real ecosystem. This approach ultimately solves the high-class imbalance problem in financial datasets as well as the obsoleteness of the historical data for studying new fraud scenarios. Simulation allows tuning of existing financial crime controls by modelling anticipated fraud activity. Besides, it ensures calibration of fraud control systems in order to enable them to adapt to emerging fraudulent behaviours, and the changing regulatory dynamics [26].

**Scenario-Based and Risk-Weighted Approaches.** This approach involves the formulation of scenarios that represent potential fraud patterns in the financial ecosystem. With this approach, more complex financial fraud patterns can be identified as compared to rule-based expert systems given that the scenarios serve the purpose of analyzing transactions and detecting suspicious activities fitting the patterns. Also, this approach can potentially uncover emerging fraud schemes [35].

Even though the scenario-based approach promotes pro-activeness in financial fraud detection, the development of accurate and relevant scenarios that suit the context necessitates significant domain expertise. Unknown fraud schemes can still be missed and updating scenarios is largely time-consuming [11,35].

The risk-weighted approach aims to assign a risk score to every transaction based on transaction amount, frequency, and location among other things and thus flag financial transactions associated with higher risk scores for further investigations. This allows for the prioritisation of resources based on the level of risk corresponding to a transaction which makes it more flexible than rule-based and scenario-based approaches. Oftentimes, financial institutions have specific risk appetite and tolerance levels, thus this approach can be customized for specific needs [44].

Depending on the accuracy of the risk model that has been put in place, the approach can still produce false positives or false negatives. Besides it requires expertise to develop and sustain risk-scoring models given that they can be complex, and further investigations are ultimately resource-intensive [14,43,44].

**Machine Learning (ML) Approach.** ML-based fraud detection is concerned with algorithms capable of learning from historical financial data in order to identify patterns and anomalies that are reflective of fraudulent activities. These algorithms include; Random forest [7,9,15], Logistic regression [15], XGBoost, Gradient boosting, AdaBoost and Decision trees [22]. They can learn human behaviour [10,42] and detect new and evolving fraud patterns in financial transactions. With the use of simulation, rich synthetic outputs can easily be used to train a number of ML algorithms for fraud detection. The researcher might not spend time for instance to clean the data since the simulations can be performed with contextual relevance to the task of financial fraud detection. In regard to other approaches, ML algorithms are more effective in dealing with large and complex datasets and they require minimal effort to maintain [3,33].

However, these algorithms may suffer over-fitting and thus isolate specific fraud patterns limiting their scope to detect different fraud schemes. Also, large amounts of data are required in model training and in the event of no historical data, the approach might not be feasible unless the financial institution can invest in synthetic data generation [39]. Financial institutions consider a number of factors before they commit themselves to a given approach for fraud detection. Some of these considerations include the volume of data that is to be processed, the cost of implementing a fraud detection measure and its maintenance, the

reduces the revenue of the service provider and large sums of money are lost when many mobile money merchants take part in it. Even though some service providers tried to put a rule-based approach of a time frame to isolate these transactions, a number of the practitioners quickly learned about the measure and adapted in terms of the schedules to commit fraud [5,31].

**Refund Fraud.** This scenario is commonly practised by mobile money clients (end-users of the service). It involves the fraudster making a payment for goods or services using their mobile money account. Then a refund or reversal is requested leading to a transfer transaction that is fraudulent. The fraudster keeps track of merchants that easily fall for this kind of fraud and aims to carry out as many transactions as possible including with potential new victims and eventually withdraws their gains out of the mobile money system [5,31].

## 2 Simulation and Fraud Detection Approaches

The use of simulation for fraud detection research has been presented by several studies [23–26]. The work in this paper expands on the capabilities of simulation using agent-based modelling techniques to develop models of current fraudulent tactics in mobile money services. The efforts in our study mainly focus on unique fraud schemes that are present in the Sub-Saharan context. Documented fraud scenarios by related studies [5,31] form the basis for modelling the unique fraud patterns in the real mobile money ecosystem.

### 2.1 Approaches for Financial Fraud Detection

**Deterministic, Rule-Based Approach.** Rule-based fraud detection is one of the common approaches used in low-resource settings. It is concerned with predefined transactional rules that are usually set by the service provider or financial institution in order to identify potentially fraudulent transactions [2,36]. It requires historical data that is usually available in financial institutions, expert knowledge and regulatory requirements often issued by designated regulatory authorities. This approach is widely used by financial institutions in Sub-Saharan Africa because of the ease of implementation, being relatively straightforward to understand and it can quickly identify basic fraud schemes. Moreover, this approach is often compliant with industry regulations and best practices that are at the forefront of the operations of financial institutions.

However, the rule-based approach is prone to producing many false positives in the event the rules are very strict or many false negatives when the rules are lenient. Too many false positives discourage the usage of financial services among customers. Financial institutions suffer from fraudsters who are usually very adaptive and since the rule-based approaches are manual, the systems can hardly adapt, and require expert knowledge to tune them. This, therefore, renders the rule-based approach very ineffective against evolving fraud schemes [1].

has been leveraged on many occasions for instance during the Covid-19 pandemic [17,20], to disburse relief cash to vulnerable communities. The service providers (telco operators) and financial institutions are at the centre of securing transactions happening on their platforms with guidance and regulation from the central banks. The service providers mostly rely on rule-based expert systems to detect incidences of financial fraud [6,18]. The challenge with that approach is the resulting high false positive rates due to the ineffectiveness of the rules on complex fraud patterns. Also with the dynamic nature of fraud in financial systems whereby fraudsters tend to be more adaptive than the service providers, controls need to be adjusted to suit this behaviour otherwise the race becomes unfair [5,24,26]. The changing patterns of fraud render historical data kept by the service providers obsolete for financial fraud detection even if the researcher is able to access the dataset. The financial records for mobile money transactions are very sensitive and often kept private denying the chance for outside researchers to participate in offering solutions to the fraud challenges. Besides, no diverse categorised fraud scenarios can be found which would inform the tuning of the existing financial fraud controls as well as the opportunity to develop better fraud detection techniques using computational methods [24].

Machine learning algorithms composed of Logistic regression, Random forest and Decision trees can be trained on data with labelled instances of financial fraud to detect future occurrences of the crime including complex fraud patterns. Such endeavour requires diverse, well-labelled data that is often difficult to obtain and at the same time, the data should be rich enough in terms of fraud cases for the intended tasks [24]. Owing to the intrinsically private nature of mobile money financial datasets and the class imbalances in the real datasets, this study generates diverse synthetic mobile money transaction datasets using a financial simulation platform [30]. MoMTSim is designed and calibrated based on real transaction data and its outputs are evaluated using the sum of squared errors (SSE) method by computing the difference between the real and synthetic data. With the agent-based modelling techniques used in its development, this study leverages simulation to model known fraudulent behaviours from the real ecosystem to enrich the synthetic datasets by defining specific fraud parameters in the model. Using the rich synthetic transaction datasets, this study performs financial fraud detection using common machine learning algorithms and evaluates the efficacy of the models in identifying unique fraudulent patterns in mobile money transactions.

### 1.1 Unique Fraudulent Behaviours in Mobile Money Transactions

**Split Deposit Fraud.** Split deposit fraud involves the mobile money merchant who acts as an intermediary between the customer and the mobile money system, facilitating the conversion of hard cash into electronic money and vice-versa. In this scenario, a dishonest mobile money merchant splits cash deposits into the client's account in the form of small chucks to enable earning of higher commission because of the many deposits made. These transactions happen in short time intervals involving a particular mobile money account. This fraudulent activity

# Financial Fraud Detection Using Rich Mobile Money Transaction Datasets

Denish Azamuke(✉), Marriette Katarahweire, and Engineer Bainomugisha

Makerere University, Pool Road, Kampala, Uganda
denishazamuke@gmail.com, baino@mak.ac.ug
https://cs.mak.ac.ug/

**Abstract.** In an era marked by the rise of digital transactions, mobile money platforms continue to experience rampant fraud and thus effective fraud detection approaches are key for maintaining the integrity of financial systems, especially in the Sub-Saharan region. This study simulates known fraudulent scenarios found in mobile money platforms in Sub-Saharan Africa using a multi-agent-based simulation platform called MoMTSim. MoMTSim generates rich synthetic mobile money transaction datasets that are statistically close to the real mobile money transaction data. The study examines common classification models including Logistic regression, Gradient boosting, Decision trees, AdaBoost, XGBoost, and Random forest for financial fraud detection. The models were evaluated using several performance metrics including Precision, Recall, F1-score, AUC-ROC, and notably, the Matthews correlation coefficient (MCC), which is particularly effective for imbalanced classes common in financial data. The results demonstrate that all tested models are capable of identifying fraudulent transactions, with varying degrees of success. The XGBoost model stood out with the highest MCC (0.82) and AUC of 0.97, indicating superior overall performance. Meanwhile, the Logistic regression model served as a benchmark with an MCC of 0.67, revealing the performance enhancements offered by more complex models. However, the study also underscores the importance of considering the computational costs associated with more complex models. The findings affirm the potential of machine learning algorithms for fraud detection and provide valuable insights into model selection based on performance and computational requirements.

**Keywords:** Mobile money transactions · Simulation · Agent-based modelling · Fraud detection · Machine learning

## 1 Introduction

The challenge of increasing financial fraud in mobile money transactions in the Sub-Saharan region has numerous consequences on the economy and existing programmes that aim to promote financial inclusion. Mobile money technology


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 190–208, 2025.
https://doi.org/10.1007/978-3-031-81573-7_16

## 5 Conclusion and Future Projects

Mango detection can be carried out using the YOLOv5 algorithm. This algorithm performed well, producing a high accuracy value of 98% for ripe mangoes and 98% for unripe mangoes. This research can be improved by applying other detection algorithms to compare the best performances in mango detection. In future work, the program could perform detection based on ripening percentage, where it is up to growers to define their harvest percentage. The program could also specify mango quality. Our detection visualization interface could be made more interactive for growers, who could apply the detection or comment on it, so that we can adjust it to their needs, all automated by a robotic arm to make mango picking safer and faster.

## References

1. Yolo v5 model architecture. https://iq.opengenus.org/yolov5/
2. Amrutkar, A.R., Jaisingpure, H.B., Bhujade, P.A.: Ripening and quality detection of mango using arduino (2018)
3. New-workspace b3mpu: mango dataset (2022). https://universe.roboflow.com/new-workspace-b3mpu/mango-yxaa7. Accesses 14 June 2023
4. Basri, H., Syarif, I., Sukaridhoto, S.: Faster r-cnn implementation method for multi-fruit detection using tensorflow platform. In: 2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), pp. 337–340. IEEE (2018)
5. Boureima Barry, S.B.B.: Fiche sectorielle: Mangue du burkina faso (2023). https://www.apexb.bf/assets/pages/Fiche. Accés le 20 Mai 2023
6. New-workspace c1vsu: mango dataset dataset (2022). https://universe.roboflow.com/new-workspace-c1vsu/mango-dataset. Accessed 14 June 2023
7. Girshick, R.: Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448 (2015). https://doi.org/10.1109/ICCV.2015.169
8. KGP: Mango object detection dataset (2023). https://universe.roboflow.com/kgp-w7w3l/mango-object-detection-r1c0y. Accessed 14 June 2023
9. Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C.: Deepfruits: a fruit detection system using deep neural networks. Sensors **16**(8), 1222 (2016)
10. Work: Mango2 dataset (2023). https://universe.roboflow.com/work-54ewq/mango2-7fjpw. Accessed 01 June 2023
11. Yusro, M.M., Ali, R., Hitam, M.S.: Comparison of faster r-cnn and yolov5 for overlapping objects recognition. Baghdad Sci. J. **20**(3), 0893–0893 (2022)

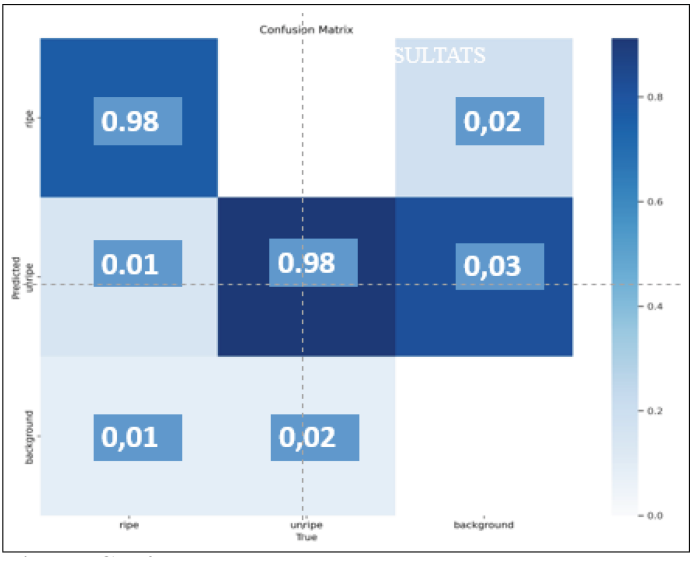



**Fig. 6.** Confusion matrix

mangoes. At the end of our study, we found that our ripe mango detection algorithm has several strong points compared with existing work. Firstly, we obtained a detection accuracy of 98% for ripe mangoes, which is a promising result. However, we recognize that the detection accuracy for unripe mangoes is 98%, indicating potential room for improvement to minimize false detections. By focusing our efforts on fine-tuning the hyperparameters and increasing the size of the training dataset, we could improve the overall accuracy of our model.

### 4.2 Discussion

One of the strengths of our algorithm is its ability to detect and classify ripe and unripe mangoes. This specificity is crucial for farmers, enabling them to determine the optimum harvesting time and ensure the quality of their produce. What's more, our model is adapted to the different types of mango varieties in BURKINA FASO, making it more versatile and applicable in different agricultural mango orchards. Another advantage of our model is that it does not require significant computing resources to apply detection. Thanks to the use of the YOLOv5s algorithm, our system is designed to be fast and efficient, enabling it to be deployed on devices with limited computing capacity. Compared with existing work, our approach offers a more flexible and efficient solution for the detection of ripe mangoes. By exploiting the capabilities of deep learning, our model can adapt to a variety of conditions, such as changes in brightness and variations in mango shapes and colors, and performance can be continually improved.

General Object Detector will have a backbone for pre-training it and a head to predict classes and bounding boxes. The Backbones can be running on GPU or CPU platforms. The Head can be either one-stage (e.g., YOLO, SSD, RetinaNet) for Dense prediction or two-stage (e.g., Faster R-CNN) for the Sparse prediction object detector. Recent Object detectors have some layers (Neck) to collect feature maps, and it is between the backbone and the Head.

## 4 Results and Discussion

### 4.1 Results

Users can upload their own images or videos, submit them to the mango detection algorithm and visualize the results in a clear and comprehensible way. Thanks to Streamlit, we were able to quickly develop a high-performance web application, while offering an attractive and easy-to-use user interface for mango detection (Fig. 5).

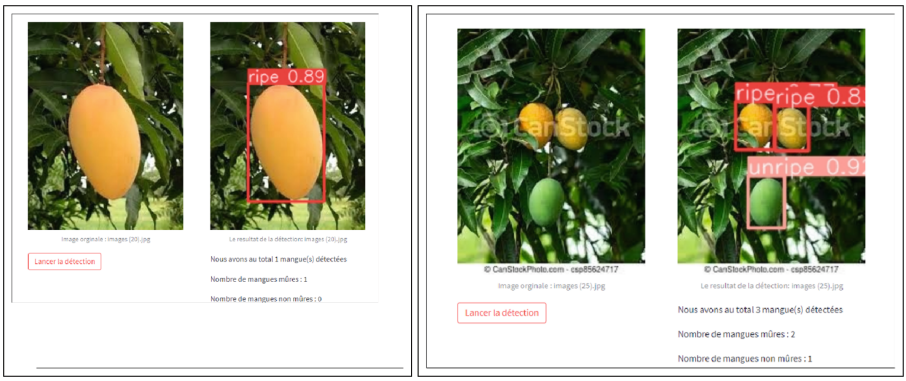



**Fig. 5.** The program detects ripe and unripe mangoes

**Evaluating the algorithm's performance** through the confusion matrix, we obtained a detection accuracy of 98% for ripe mangoes and 98% for unripe mangoes. At the end of our study, we found that our ripe mango detection algorithm has several strong points compared with existing work. Firstly, we obtained a detection accuracy of 98% for ripe mangoes, which is a promising result. However, we recognize that the detection accuracy for unripe mangoes is 98%, indicating potential room for improvement to minimize false detections. By focusing our efforts on fine-tuning the hyperparameters and increasing the size of the training dataset, we could improve the overall accuracy of our model (Fig. 6).

**Evaluating the algorithm's performance through the confusion matrix**, we obtained a detection accuracy of 98% for ripe mangoes and 98% for unripe

language; Google-Colab Colab or Collaboratory was used to train our model, thanks to free access to the GPU, which speeds up training time. As for data preparation, Roboflow and LabelImg were used for images. Finally, Pytorch was integrated into YOLOv5 to implement our program, not forgetting Streamlit, which was used to design our interface for visualizing the detections made by the model (Fig. 3).

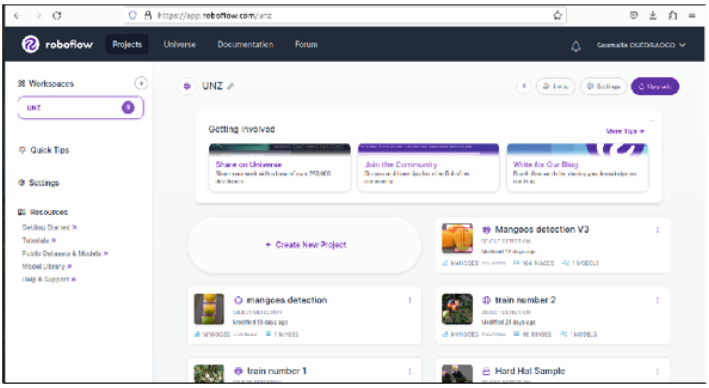



**Fig. 3.** We've created a project called UNZ on roboflow, which contains our datasets of mango images used to train our model.

**Architecture of YOLO.** Yolov5 improved the performance and its architecture, basednon high-level Object detection architecture (Fig. 4):

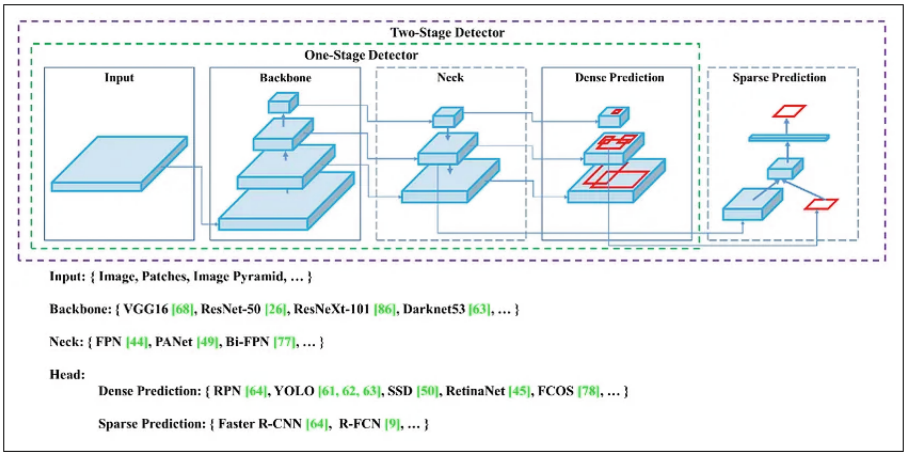



**Fig. 4.** Architecture of YOLO

| Model | size (pixels) | mAP$^{val}$ 0.5:0.95 | mAP$^{val}$ 0.5 | Speed CPU b1 (ms) | Speed V100 b1 (ms) | Speed V100 b32 (ms) | params (M) | FLOPs @640 (B) |
|---|---|---|---|---|---|---|---|---|
| YOLOv5n | 640 | 28.0 | 45.7 | **45** | **6.3** | **0.6** | **1.9** | **4.5** |
| YOLOv5s | 640 | 37.4 | 56.8 | 98 | 6.4 | 0.9 | 7.2 | 16.5 |
| YOLOv5m | 640 | 45.4 | 64.1 | 224 | 8.2 | 1.7 | 21.2 | 49.0 |
| YOLOv5l | 640 | 49.0 | 67.3 | 430 | 10.1 | 2.7 | 46.5 | 109.1 |
| YOLOv5x | 640 | 50.7 | 68.9 | 766 | 12.1 | 4.8 | 86.7 | 205.7 |
| YOLOv5n6 | 1280 | 36.0 | 54.4 | 153 | 8.1 | 2.1 | 3.2 | 4.6 |
| YOLOv5s6 | 1280 | 44.8 | 63.7 | 385 | 8.2 | 3.6 | 12.6 | 16.8 |
| YOLOv5m6 | 1280 | 51.3 | 69.3 | 887 | 11.1 | 6.8 | 35.7 | 50.0 |
| YOLOv5l6 | 1280 | 53.7 | 71.3 | 1784 | 15.8 | 10.5 | 76.8 | 111.4 |
| YOLOv5x6 + TTA | 1280 1536 | 55.0 **55.8** | 72.7 **72.7** | 3136 - | 26.2 - | 19.4 - | 140.7 - | 209.8 - |

**Fig. 2.** All the YOLOv5 models [1]

**Accuracy.** This metric measures the percentage of relevant detection results. This can be determined using the following equation: Accuracy = TP/(TP/FP) where TP (True Positive) represents the number of correctly detected objects in a given class. FP (False Positive) is when the model incorrectly identifies a region of the image as a positive object, when in reality there is no object of that class in that region.

**Recall.** This metric measures the percentage of total results correctly classified. It is determined using the following formula: R = TP/(TP+FN) where FN (False Negative) is when the model fails to detect a positive object in an image.

### 3.3 Model Deployment

The trained model was retrieved on my local machine by downloading a file summarizing the weights of the neurons, in other words the "educated" network. Thanks to this "educated" weight, we were able to perform mango detection locally on a personal machine. The advantage of this approach is that object detection can be performed more quickly and lightly on a personal machine, without the need for the heavy infrastructure required for initial model training. Using this pre-trained weight, we designed a visualization interface with the streamlit framework, enabling mango detection in an image, a video, or using the webcam.

### 3.4 Development Tools

The complete system requires different types of software, tools and frameworks for its implementation and deployment. Python was used as the development

Central-West Region of BURKINA FASO. In addition, we also exploited several image banks available on the Roboflow platform, which offers free public image datasets of mangoes [3,6,8,10]. Images were also collected from sites such as Depositphotos and Alamy. After collecting the images, we annotated them using two tools: Roboflow's online tool and labelImg for local annotation. The annotations were made in the YOLO format, corresponding to the YOLO algorithm. We used two classes of objects for annotation: "ripe" for ripe mangoes and "unripe" for unripe mangoes. Once the images were correctly annotated, we performed pre-processing on them. We resized the images to $640 \times 640$ to match the model's input size, and we normalized the images to enable effective model training. Our dataset consists of 500 mango images, with 80% of the images used for training, 10% for validation, and the remaining 10% for testing. To organize the dataset and specify the paths to the training, testing, and validation image folders, we created a .yaml configuration file. This file defines the root directory of the dataset and the relative paths to the folders containing the training, testing, and validation images (Fig. 1).

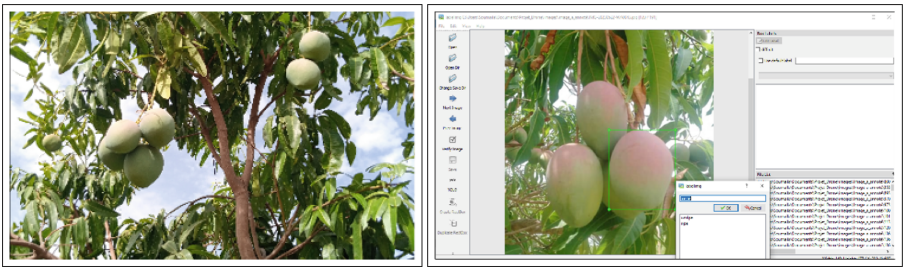


**Fig. 1.** On the left are the images taken with REO and on the right the annotations made with LabelImg

### 3.2 Model Selection and Training

Version 5 of the YOLO algorithm has several sub-versions, including 'n', 's', 'm', 'l' and 'x'. We decided to use the YOLOv5s version as it is designed to be faster for object detection in latency-critical applications. Once the model has been chosen, before moving on to training, it's necessary to set the hyperparameters that have an impact on model performance. These include learning rate, batch size set at 640, number of epochs and many other parameters. After fine-tuning the hyperparameters, the model is then trained with our custom dataset prepared on google colab thanks to free access to the GPU. Training is carried out by successive iterations on the training images, and the model adjusts its weights to improve detection performance. We evaluated the model's performance on validation data after training, using various measures including precision, recall and f1-score to ensure the model's ability to generalize to new data (Fig. 2).

Arduino IDE which has the role of sending the detection result remotely via the GSM module. The process begins by capturing the image with a camera, which is then sent to MATLAB for further processing. MATLAB uses the HSV color space algorithm, which provides pixel information in the form of digital HSV values. This numerical value is compared with ideal sample values pre-stored in the database using MATLAB. Once the color of an image is known according to the different stages of ripeness of the mango, MATLAB sends the unique code of this color to the Arduino. If the mango is ripe, then we have quality detection; the model simply checks the brownish color threshold and makes a decision. The proposed method was tested on around 200 samples, of which 193 samples were accurately identified, representing a percentage of 96.5%. The weaknesses of this study are that the Arduino IDE can handle the tests, but will be limited to the actual deployment by its processing and storage capacity. Accuracy can also be improved.

**Akshay Ramesh et al.** [2] chose to identify the different ripening stages of climacteric fruits such as mango using the Arduino IDE, which has the role of sending the detection result remotely via the GSM module. The process begins by capturing the image with a camera, which is then sent to MATLAB for further processing. MATLAB uses the HSV color space algorithm, which provides pixel information in the form of digital HSV values. This numerical value is compared with ideal sample values pre-registered in the database using MATLAB. Once the color of an image is known according to the different stages of ripeness of the mango, MATLAB sends the unique code of this color to the Arduino. If the mango is ripe, quality is detected, and the model simply checks the brownish color threshold and makes a decision. The proposed method has been tested on around 200 samples, 193 of which have been identified with precision, i.e. a percentage of 96.5%. The weaknesses of this work are that the Arduino IDE can handle the tests, but will be limited to the actual deployment by its processing and storage capacity. Accuracy can also be improved.

## 3 Methodology

To achieve the announced downstream objective, we will address several key aspects. First, we will perform an in-depth analysis of the object detection features and specifications. Next, we will proceed to the design of the detection system, defining the steps and components necessary to achieve optimal performance. We will then implement our program using YOLOv5, adjusting parameters and performing experiments to improve the results. Finally, we will evaluate the performance of our program using metrics such as precision, recall, and F1 score.

### 3.1 Data Preparation

In the course of our study, we collected images of mangoes at several sites, notably in mango orchards in REO, the capital of the Sanguie Province in the

visual observation and touch, are often subjective, slow and can lead to errors of judgment. By integrating advanced machine learning and computer vision techniques, it is possible to develop a more efficient and accurate automated and objective ripe mango detection system [5]. The objective of this study is to develop a robust and efficient ripe mango detection program capable of accurately locating ripe mangoes on trees and counting the total number using computer vision. Specifically, our study aims to ensure a better quality of the fruits offered on the market and to contribute to reducing post-harvest losses while optimizing logistics operations in the mango industry. This is why we focus on analyzing, designing and implementing a mango detection program using the You Only Look Once algorithm version 5 (YOLOv5). YOLOv5 is widely used for real-time object detection. Its lightweight architecture and high accuracy make it an ideal choice for our ripe mango detection application.

## 2 Similar Works

Object detection processes using artificial intelligence techniques have been practiced around the world. Several methods have been proposed in the past for the detection of different fruits, including:

**R-CNN, Faster R-CNN, Fast R-CNN.** Susoven jana et al. [4] proposed a deep learning method using faster R-CNN for the classification of a multi-fruit set, namely mango and pitaya fruits. The dataset used is a farmer's real catch at harvest time, which is divided into two (2) classes: mango and pitaya. In this research, the MobileNet model on the TensorFlow platform was used. The proposed method achieved good results. The accuracy score reached around 99%. There's also Ross Girshick, Microsoft Research [7] who proposes a Faster R-CNN algorithm for object detection. This algorithm trains the very deep VGG16 network 9 times faster than R-CNN, is 213 times faster at test time and achieves a higher mAP on PASCAL VOC 2012. Compared with SPPnet, Fast R-CNN forms VGG16 3 times faster, is tested 10 times faster and is more accurate. Following the same logic, Inkyu et al. [9] used a fruit detection approach using deep convolution neural networks. It applies the transfer learning method on previous work that led to the development of a state-of-the-art object detector called Faster Region-based CNN (Faster R-CNN). The performance of the implemented detector was evaluated over several fruits, and an F1 score of 83% was obtained, which is slightly higher than the results of previous work, which was 80%. The weakness of this work lies in the choice of detection method. Indeed, it should be noted that in an identical test environment, YOLOv5 performs better than the Faster R-CNN algorithm. This was demonstrated in a comparative study between the Faster R-CNN algorithm and YOLOv5 [11]. Beyond the choice, many results show a low detection score (83%).

**Another Technique** has been developed by Akshay Ramesh et al. [2] to identify the different ripening stages of climacteric fruits such as mango, using the

# Analysis, Design and Implementation of a Ripe Mango Detection Program in Burkina Faso

Moustapha Bikienga[1], Roland Manegaouindé Tougma[1,2(✉)], and Soumaïla Ouedraogo[1]

[1] Norbert ZONGO University, Avce Maurice Yameogo BP376, Koudougou, Burkina Faso
manegarodrol@gmail.com

[2] Joseph KI ZERBO University, CFX2 7R6 Ouagadougou, Burkina Faso
https://www.ujkz.bf/

**Abstract.** The analysis, concept, and implementation of a computer program capable of detecting ripe mangoes are at the core of this study. Traditional methods are often faced with calibration errors. Recently, deep learning has shown promising performance in visually guided agricultural applications. Faced with these constraints, it is necessary to establish an automatic system for robust and efficient detection of mangoes in orchards. In this study, a fast implementation system of a mango detector, distinguishing between ripe and unripe mangoes based on deep learning using the YOLOv5 algorithm, was developed. From a simple photo, the algorithm detects and counts the number of mangoes on a tree. This artificial intelligence system (deep neural network) was trained on a dataset of over 500 annotated mango images. Experimental results show that the algorithm achieves 98% precision, 98% recall, and an F1-score of 98%. This satisfactory precision in mango detection offers significant advantages in terms of efficiency and accuracy compared to traditional methods. However, it should be noted that our system has certain limitations. Nevertheless, our study demonstrates promising results in the field of ripe mango detection.

**Keywords:** Deep learning · YOLO algorithm · Mangoes · Performance metrics · Computer vision

## 1 Introduction

In BURKINA FASO, the mango is one of the six (6) so-called promising sectors identified as having strong potential for the diversification of exports and represents between 11 and 18% of West African mango production [5]. The mango constitutes about half of the national fruit production in volume [5]. It is also a very important economic, social and climatic issue in BURKINA FASO. However, the detection of fruit maturity is an essential task in agriculture and the food industry. However, traditional methods of detecting fruit maturity, such as


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 181–189, 2025.
https://doi.org/10.1007/978-3-031-81573-7_15


**Fig. 4.** Evolution of Hate Speech Comments Based on Monthly Granularity

## 5 Conclusion

The objective of this study was to gather and analyze data on the security situation in Burkina Faso. We collected data using web scraping techniques, which were subsequently employed to train transformer models for sentiment analysis and hate speech detection. The results from our visualizations reveal an increase in negative sentiments and hate speech during periods marked by terrorist attacks. These efforts lay the groundwork for a system that could contribute to decision-making regarding Burkina Faso's security situation.

Future directions for our work include collecting data from heterogeneous sources such as other social media platforms. Additionally, we plan to identify spatial named entities within the collected data to visualize the relationship between these named entities and the various analyses derived from the data we've gathered.

## References

1. Lin, T., Wang, Y., Liu, X., et al.: A survey of transformers. arXiv preprint arXiv:2106.04554 (2021)
2. https://reports.unocha.org/fr/country/burkina-faso
3. Demange, J.: Four sentiments with FlauBERT. Hugging Face repository (2021). https://huggingface.co/DemangeJeremy/4-sentiments-with-flaubert
4. Aluru, S.S., Mathew, B., Saha, P., Mukherjee, A.: Deep Learning Models for Multilingual Hate Speech Detection (2020)
5. https://fr.wikipedia.org/wiki/Massacre_de_Solhan

– Hate
– No_Hate

## 4 Results

### 4.1 Sentiment Analysis

By using the *4-sentiments-with-flaubert* [3] model, we have successfully determined the sentiments of internet users through their comments on the security situation in Burkina Faso. Figure 3 depicts the evolution of different sentiments on a monthly basis. The curve labeled *negative_sentiment* (in red) generally stands above the other curves. Notably, this curve exhibits significant spikes indicating a very high negative sentiment, which usually arises after significant attacks: the first in December 2016 with the attack on Nassoumbou, in November 2019 when the country suffered high human losses due to the attack on Semafo de Boungou, in June 2019 with the attack on the village of Solhan [5], and in August 2022 with the attack on Nohao near the city of Bittou. We also observe a reduction in negative comments starting from September 2022.

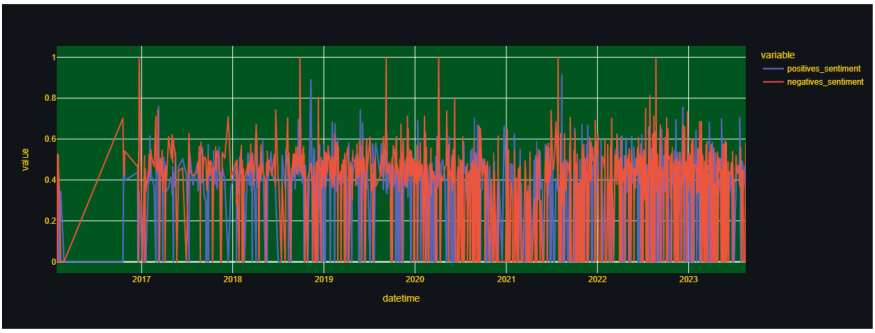



**Fig. 3.** Sentiment Evolution Curve Based on Monthly Granularity

### 4.2 Detection of Hate Speech

Analyzing comments using the *hubert-mono-french* [4] model enabled us to create Fig. 4. This figure presents the evolution of the number of comments containing hate speech on a monthly basis. It is evident that hate speech is present in user comments, but overall, non-hateful messages dominate.

**Table 1.** List of Article Keys

| Key | **Description** |
|---|---|
| *article_type* | Article type (report, article, etc.) |
| *article_title* | Title of the article |
| *published_date* | Publication date of the article |
| *origin* | Source of the article |
| *url* | Article access URL |
| *content* | Article content |
| *comments* | List of comments on the article |
| *comments_number* | Number of comments |

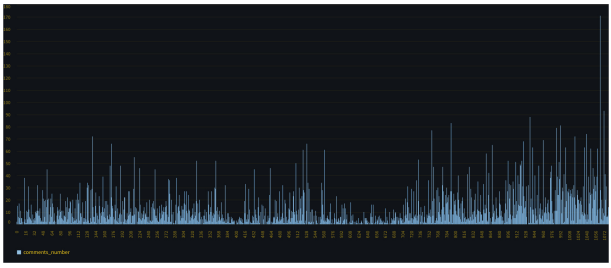


**Fig. 2.** Distribution of the Number of Comments per Article

- NEGATIVE: A negative sentiment is an unfavorable or unpleasant feeling. For example, fear following a terrorist attack; a citizen might experience negative sentiments due to fear for their own safety or that of their loved ones.
- MIXED: A mixed sentiment combines both positive and negative feelings. A citizen might feel both sadness for the attack victims and anger towards the perpetrators, or both fear for personal safety and hope for the country's recovery.
- OBJECTIVE: A sentiment can be considered objective if it's based on facts and evidence rather than personal opinions or beliefs. Objective sentiments might stem from understanding the reasons behind an attack, relying on facts like conflict history, attackers' motivations, etc., rather than personal preferences or beliefs. However, it's essential to note that even in this case, individuals might still hold subjective feelings about the situation.

### 3.3 Hate Speech Detection

In this analysis, our goal is to detect hate speech in comments. For this purpose, we employ *hubert-mono-french*, a model specialized in this task. It classifies text as either hateful or not hateful, using the labels:

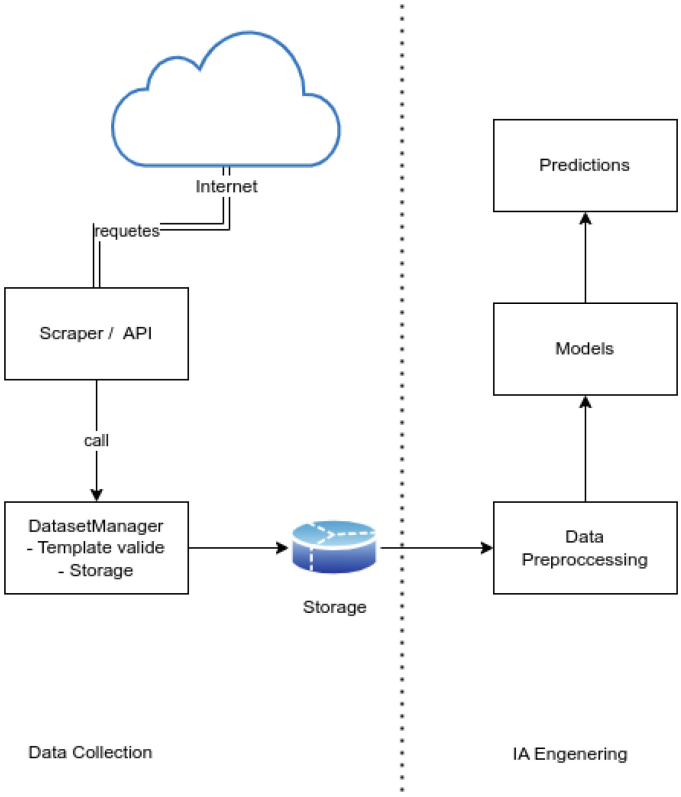



**Fig. 1.** Architecture

Data preprocessing involves converting all text to lowercase, removing non-alphanumeric characters, and eliminating accents from accented letters. The preprocessed data is stored in *json* format within a list. The essential details are listed in Table 1.

We have collected 2120 articles with a total of 29560 comments. The distribution of the number of comments per article is illustrated in Fig. 2.

### 3.2 Sentiment Analysis

We utilize *4-sentiments-with-flaubert* [3], a pre-trained model on French language data, designed for sentiment analysis. Given input text, it classifies the text into four categories:

– POSITIVE: A positive sentiment is a favorable or pleasant feeling. For instance, an internet user believing that the Burkinabe authorities are on the right path to defeat terrorism.

Artificial intelligence offers numerous methods and techniques for data collection and analysis. It enables sentiment analysis, hate speech detection, named entity extraction, and more. Hence, artificial intelligence could contribute to addressing the security challenge in Burkina Faso. However, applying AI methods requires a substantial amount of data. To address this, we focus our research questions on the following points: how to acquire a sufficient quantity of data for AI methods application? What relevant analyses can be conducted based on the data to enhance decision-making?

The objective of our research is to gather data on Burkina Faso's security situation and analyze it using artificial intelligence's methods and techniques. In this work, we make the following contributions: establishing a textual database or corpus on the security situation, proposing an AI-based methodological approach for data analysis, and suggesting a dashboard for visualizing analysis results.

This article comprises five sections, with this introduction being the first. The second section presents our methodological approach. Section 3 demonstrates the application of our approach. In Sect. 4, we present the various obtained results and their interpretation. Section 5 encompasses a conclusion and the prospects of our work.

## 2 Methodological Approach

In this work, we propose an approach divided into two main parts: the construction of a dataset and the analysis of this data using artificial intelligence models. The first part involves collecting, preprocessing, and storing textual data related to Burkina Faso's security situation. As for the second part, it entails applying AI models to this dataset to conduct various analyses. These two parts are interconnected to ensure a reliable analysis based on up-to-date information. Figure 1 provides an overview of the architecture of our approach. Our work is available on GitHub[1].

## 3 Methodological Application

### 3.1 Construction of the Dataset

For our initial work, we are using the website *lefaso.net*[2] as our data source. On this platform, we collect publications (or articles) related to the current security situation in Burkina Faso. These articles are gathered and displayed in a paginated manner under a section titled *"Terrorist Attacks"*. To achieve this, we have implemented a web scraping module to retrieve these articles as well as the comments made by users on these articles. Consequently, we have acquired textual data that we preprocess to retain important information and the appropriate format.

[1] https://github.com/abdoulfataoh/security-situation-analysis.
[2] https://www.lefaso.net/.

# Artificial Intelligence for the Analysis of the Security Situation in Burkina Faso

Abdoul Fataoh Kaboré(✉), Maïmouna Ouattara, Rodrique Kafando, Aminata Sabané, Abdoul Kader Kaboré, and Tegawendé F. Bissyandé

Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement (CITADEL), Ouagadougou, Burkina Faso
abdoulfataoh@gmail.com
https://citadel.bf

**Abstract.** In the face of the insecurity caused by terrorism that Burkina Faso has been experiencing since 2015, the population doesn't hesitate to express their feelings. The various reactions of the population are expressed through comments on different social platforms, thereby creating a significant amount of data. Analyzing these opinions can provide assistance in decision-making related to security. This analysis can be accomplished through techniques and methods offered by artificial intelligence (AI). In this article, we introduce a web scraping tool to gather data for our research. Subsequently, we employ sentiment analysis and hate speech detection models based on transformers [1]. Through this research, our contributions are as follows: establishing a textual database or corpus related to the security situation, proposing a methodological approach based on AI for analyzing this data, and suggesting a dashboard for visualizing the analysis results.

**Keywords:** security situation · web scraping · artificial intelligence · transformers

## 1 Introduction

Since August 2015, Burkina Faso has been the target of terrorist attacks, affecting both the Defense and Security Forces (FDS) and civilian populations. This situation of insecurity has widespread repercussions across the country. Socially, the situation report from OCHA [2] counts, as of April 30, 2022, 1,520,012 internally displaced persons, of which 59.13% are children, and 4,258 closed schools. The population expresses various opinions and sentiments about the country's security situation. Nowadays, with the proliferation of Information and Communication Technologies (ICT), a significant portion of these reactions is channeled through social media. Analyzing these opinions could aid decision-making related to security.

CITADEL-UVBF.


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 175–180, 2025.
https://doi.org/10.1007/978-3-031-81573-7_14

# Responsible Artificial Intelligence for Sustainable Development in Africa (workshop)

7. Amani, D., Ahmed, K., Richard, C.S.S.: A new automatic ontology construction method based on web dat (2021)
8. Samaa, E., Yoon, V., Manoj, A.: An automatic ontology generation framework with an organizational. In: Proceedings of the 53rd Hawaii International Conference on System Sciences (2020)
9. Gruber, T.: A translation approach to portable ontology specifications knowledge (1993)
10. James Malone, H.P.: Ontological models for information retrieval of product-service: trends and open issues (2016)
11. Jiofa, T.: conception d'une architecture multi-agent de marketplace b2b (2021)
12. Lorhard, J.: Gdoas Scholastica (1606)
13. Mbathe, P.: Modele de recherche d'information semantique dans un corpus de documents textuels (2021)
14. na, M.C.R.P., Tovar-Vidal, B.: Review faculty of computer science, Puebla, Mexico (2023)
15. Ndie, T.D.: An entity-based black-box specification approach for modeling wireless community network services (2019)
16. Joel, C.O., Toni, F.: Automatic product ontology extraction from textual reviews (2021)
17. Navarro-Almanza, R., Juarez-Ramirez, R., Licea, G.: Automated ontology extraction from unstructured texts using deep learning. In: Intuitionistic and Type-2 Fuzzy Logic Enhancements in Neural and Optimization Algorithms: Theory and Applications, pp. 727–755 (2020)
18. Salhi, K.: La fusion des ontologies (2018)
19. yowyob. https://www.yowyob.com
20. Wu, Z., Palmer, M.: Verb semantics and lexical selection (1994)

The 500 search sentences constructed was applied to 2 previous solutions and the accuracy, precision, recall and the F1 score for each of the solutions was calculated. A comparison of these three solutions are presented in Table 2.

From the comparison above, it is seen that the results from CAOGen is best. These results are best because text2onto produces general concepts while CAOGen produces concepts specific to the company. With yowyob's search engine functioning with CAOGen, yowyob customers are able to find more relevant materials and this eases their trade process. Also the recommendation ability of the system has permitted customers to discover new services and their different forms and varieties.

## 6 Conclusion

This paper was aimed at proposing an approach for the automatic construction of ontologies. To automatically construct ontologies, this approach name Company Automatic Ontology Generator (CAOGen) uses a data mining technique (clustering) to regroup concepts based on their similarity and the targeted objective of the company to produce a knowledge tree from which an ontology is built. The whole process is done following Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. CAOGen was applied on the search engine of the e-commerce platform search.yowyob.com to automatically construct an ontology for their catalogue of service with 5531 services. The main aims of the resulting ontology was to make semantic search and to ensure recommendation of services. The resulting ontology had an accuracy of 0.994, a precision of 0.992, a recall of 1.0 and a F1 score of 0.996 for semantic search and an accuracy of 0.888, a precision of 0.990, a recall of 0.849 and a F1 score of 0.914 for recommendation.

The main limit of this research work is its time complexity for the construction of ontologies. Actually it has a time complexity of $O(n^3)$ and takes a long time to execute. A possible remedy to this could be putting in place a distributed architecture for creating the ontology binary tree take care of the possible synchronization issue.

## References

1. Andreia, D., Maria, J.: Simple method for ontology automatic extraction. Int. J. Adv. Comput. Sci. Appl. **3**(12) (2012)
2. Ashraf, A., Fayez, A.: Semantic similarity measures between words: a brief survey (2018)
3. Benouaret, I.: Un système de recommandation contextuel et composite pour la visite personnalisée de sites culturels (2017)
4. CAOGen: Yowyob search engine using caogen. https://api-services.yowyob.com/caogen
5. Chronopoulos, S.D.: Study of KPIs which affect the outcome in Rugby Union fixtures (2019)
6. Cimiano, P., VÄolker, J.: Text2onto a framework for ontology learning and data-driven (2014)

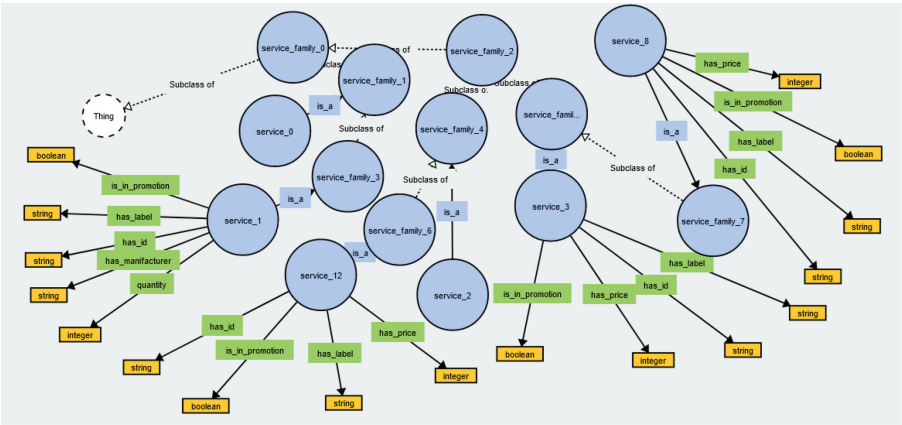



**Fig. 2.** An extract of Yowyob's service ontology

## 5 Results Analysis and Discussions

### Impact of CAOGen on Semantic Search and Recommendation Results

From the 500 search sentences, CAOGen returned 351 True positive results, 146 True negative results, 3 false positive results and no false negative results. From this, an accuracy of 0.994, precision of 0.992, recall of 1.0 and a F1 score of 0.996 for semantic search and an accuracy of 0.888, precision of 0.990, recall of 0.849 and a F1 score of 0.914 for recommendation results were obtained. This precision shows that majority of the prediction made by the system were correct and the high accuracy shows that the systems did very few errors.

**Table 2.** Comparison of search and recommendation results.

| Search | Old Yowyob search | Text2onto | CAOGen |
|---|---|---|---|
| Accuracy | 0.398 | 0.398 | 0.994 |
| Precision | 1.0 | 0.6 | 0.992 |
| Recall | 0.142 | 0.427 | 1.0 |
| F1-score | 0.249 | 0.499 | 0.996 |
| Recommendation | Old Yowyob search | Text2onto | CAOGen |
| Accuracy | 0.398 | 0.398 | 0.888 |
| Precision | 1.0 | 0.6 | 0.990 |
| Recall | 0.142 | 0.427 | 0.849 |
| F1-score | 0.249 | 0.499 | 0.914 |

**Updating a Node to the Ontology:** CAOGen is completely dynamic. That is, nodes can also be modified. To modify a node, CAOGen simply searches for the node, deletes it and inserts a new one with the new properties. It is done this way in order to make sure the new position of the node matches with its new properties.

## 4 Experimentation

The experiementation was done on the catalogue of service of yowyob enterprise.

To construct an ontology of catalogue of service for yowyob, CAOGen followed the following steps:

- Step 1. Business understanding Yowyob [19] is an online e-commerce platform that permits its users to buy and sell goods and services. Just like a product, a service is something that can be sold, but is intangible [15]. The Objectives of yowyob was to ensure semantic correctness in search results; and to be capable to recommend new services to its users. From that, we decided to construct an ontology of catalogue of service for yowyob.
- Step 2. Data understanding Yowyob had 5531 services with each service having 19 attributes. These attributes are mainly numeric, string or Boolean. Majority of their attributes were empty or NULL. Because of that, there was the need to augment the data with other resources from internet.
- Step 3. Data preparation Once yowyob's data was understood, CAOGen performed three actions on the data:
  - Data correction. CAOGen started by replacing all empty values by NULL for better management,, removed unnecessary white spaces and special characters.
  - Feature extraction. At this stage, CAOGen extracted the most pertinent characteristics of each service.
  - Data augmentation. CAOGen augmented the data by searching for more information about each service of yowyob from wordnet and Wikipedia.
- Step 4. Clustering To construct the ontology, CAOGen calculated the similarity matrix and applied hierarchical clustering algorithms on the data. Then the results were used to create an ontology containing 11062 classes and 5531 individuals. Unique names were given to each of the classes and individuals created. A screen shot of part of the ontology produced is shown in Fig. 2 below.
- Step 5. Ontology deployment After creating the ontology, a search engine was implemented to make the ontology available to the public [4] rom which the resulting ontology can be queried.

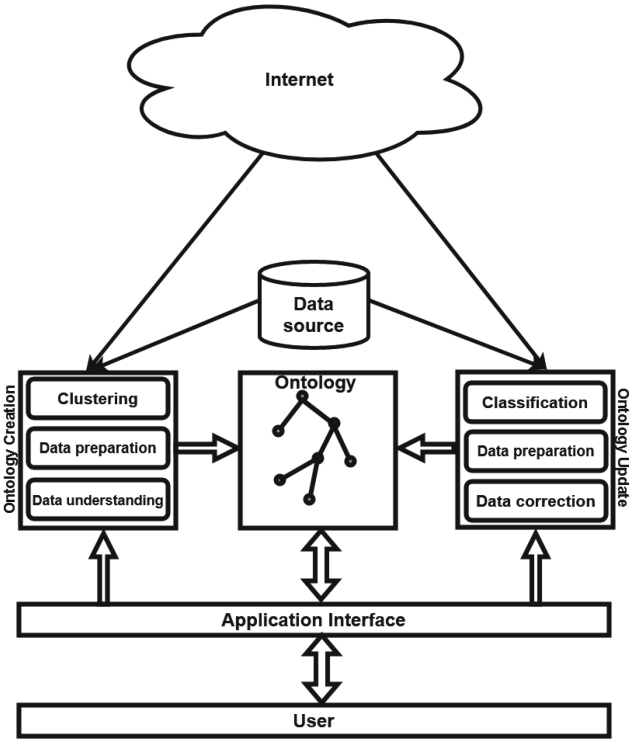



**Fig. 1.** CAOGen Architechture

**Insertion of a New Concept in the Ontology:** When the ontology is created, the company might have new concepts and wants to add it into the ontology. CAOGen has some modules for that. To do that, it loops through all the individuals in the ontology and calculate their similarities with the new concepts. Then, it inserts it where appropriate depending on their similarity scores.

**Searching for a Concepts:** One of the most important aspect of ontologies is the ability of carrying out semantic search. CAOGen give the ability to companies to directly query the produced ontology in order to make search. To do that, CAOGen takes in a description text for the search, and returns a list of individuals matching the input text.

**Deleting a Node from the Ontology:** Companies' visions change with time and hence they have changing objectives. A company's objective can change and it decides to change the content of its ontology by deleting nodes, CAOGen permit to do that. This is simply done by searching for the node and destroying it.

## 3 Proposed Solution

### 3.1 CAOGen: Company Automatic Ontology Constructor

The method presented here is based on a CRISP-DM technique developed by IBM. The stages involved can be summarized as:

1. Business understanding: This stage consists of identifying the objective of the ontology to be created.
2. Data understanding: After defining the objective of the ontology, the initial data is been collected and the data types (structured or non-structured data), the data sources, the data file format, the data encoding and data quantity are determined.
3. Data Preparation: The preparation stage consists of three main phases: data correction, feature extraction and data augmentation.
4. Modelling (construction of the ontology):. The prepared data is been passed into a constructor that will produce an ontology. The ontology produced is prepared for deployment and hosting.
5. Evaluation: This is done by selecting a number of concepts and querying the ontology for the similar concepts and complement concepts and calculating the Accuracy, Precision, Recall and F1 score.
6. Deployment: The company can download the ontology and deploy it on its own servers or can decide to leave it on CAOGen server and can query it when and where needed through the CAOGen's API.

### 3.2 CAOGen's Principle of Ontology Construction and Management

Figure 1 is CAOGen's architecture showing the different stages involve in the process of ontology creation and update as well as the interaction of the system with the final users through an API.

**Ontology Creation:** The construction of an ontology is a problem of classification where a set of concepts are given to a model for it to classify them based on their similarity and their proximity. [18] The process of ontology creation can be summarized as:

1. Calculate the distance matrix
2. Use agglomerative clustering on the concepts to produce a binary tree of concepts
3. Create an empty ontology
4. Loop through all the nodes on the tree and if the node is not a leaf create a class on the ontology, if the node is a leaf create a class an and individual attached to it
5. loop through the arcs and connected the corresponding classes together.

features that distinguish it from other works. First, by representing the learned knowledge at a meta level within a Probabilistic Ontology Model (POM). Second, the system calculates a confidence for each learned object allows to design sophisticated visualizations of the POM. Third, it avoids processing the whole corpus from scratch each time it changes, only selectively updating the POM according to the corpus changes. The authors [8] present a framework based on three main stages to automatically construct ontologies. These stages are the generation phase; the Refinement phase; and the Mapping phase. This approach is nice but a reference ontology is required from an expert of the domain. A semi-automatic ontology construction method based on machine learning algorithms to facilitate the reading and ease updating of financial data and to guarantee their understanding was also proposed by [7]. Their method consists of three modules which are: Pre-processing module, learning module, and the Visualization module. This method has two main draw backs which are the fact that it's just semi-automatic and the fact that it's just for a specific domain (financial domain). The authors [16] propose a methodology in five steps for automatically extracting ontologies, in the form of mnemonics from product reviews. First, attach labels to the various facts found in the review ; then the entity Extraction; after that is the aspect Extraction; after that, they proceed with the synonym extraction; the last stage is the ontology extraction. This method faces a major disadvantage that, the system must first function for some period of time in order to have enough user review text before the ontology can be build.

**Table 1.** Comparative table of previous works

| | Andreia, D. [1] | Benouaret, I [6] | E. Samaa, Y.V. [8] | (R. Navarro-Almanza [17] | D. Amani, K. [7] | Joel, C.O. [16] |
|---|---|---|---|---|---|---|
| Type of text (unstructured/ structured) | Unstructured | Unstructured | Unstructured | Unstructured | Unstructured | Unstructured |
| Insertion of new element | No | Yes | No | No | No | No |
| Type of end users | Generic | Generic | Generic | Generic | Users of financial domain | Generic |
| ontology precision | Low precision | Low precision | Low precision | Low precision | Low precision | High precision |
| language of input text | English | English and Spanish | Not mentioned | | | English |
| Domain of application | Any domain | Any domain | Any domain | Any domain | Finances | Products |
| Level of automation | Automatic | Automatic | Semi-automatic. | Automatic | Semi-automatic | Automatic |
| Can ensure recommendation | | | | | | Yes |
| Source of input | Generic text document | Generic text document | Generic text document | Wikipedia and WordNet | Generic | From products' review |

Globally, after comparing related works as seen in Table 1, it is notice that these works are either domain specific, semi-automatic or produces ontologies that can't be used by a specific company for a specific purpose. Following that, this paper proposes to build a custom ontology using data mining techniques from a company's specific concepts and using wordnet and wikipedia to augment them.

**Similarity and Proximity:** Similarity measurement give a numerical value that shows the extent to which a concept resemble the other. On the other hand, proximity measurement gives a numerical value that shows the extent to which a concept is related to the other [2]. That is, we can say that similarity measurements can be used to recommend substitutes of concepts while proximity can be used to recommend complements of concepts.

**Semantic Distance Measurement Methods.** There exist 3 semantic measurement methods:

*Structural Measurements* [3]. Example: the Wu and Palmer similarity measure.

$$sim_{wu-palmer} = \frac{2 * dept(lcs(s1, s1))}{dept(s1) + dept(s2)}[20] \tag{3}$$

where dept(s) is the number of nodes from the root of the graph to the node s and lcs(s1,s2) (least common subsumer of s1 and s2) is the closest parent node from which s1 and s2 originates.

*Intentional measurements.* Examples: the cosine measures and the The Pearson correlation coefficient.

$$\cos \vartheta = \frac{\overrightarrow{a} \cdot \overrightarrow{b}}{|\overrightarrow{a}| \left|\overrightarrow{b}\right|} \tag{4}$$

$$sim_{pearson} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{5}$$

*Hybrid Measurements Method.* Structural measurements can ensure similarity but they cannot ensure proximity. On the other hand, intentional measurements can ensure proximity between concepts but cannot guarantee similarity between them. Therefore, a hybrid system is proposed in CAOGen that combines both structural and intentional measurement techniques so as to ensure both similarity and proximity.

### 2.1 Related Works

With years, different methods have been developed to automatically construct ontologies. Some of these methods include the one proposed by

Andreira and Maria [1] who proposed a simple method for the construction of ontologies from text documents in which they use the frequency of terms or concept in the document to create properties axioms and restrictions by using correlation between the concepts and wordnet technologies to create the ontology. Their proposal is not quite interesting because it doesn't have a wide range of application since it goes from generic documents. Another of this system named Text2onto [6] also carry out similar functions. Text2onto has three main

form of a graph and used to make complex predictions in that domain [13]. A such knowledge graph is called an ontology. Ontologies can be used across many fields such as artificial intelligence, Recommendation Systems, Search engines and many others. The construction of ontologies is a classification process in which concepts in the domain and all the relationship between these concepts need to identified, axioms among these concepts need to be set and the properties of these concepts also have to be define [10]. Manual classification process is done by an expert in the domain while automatic classification of concepts can be done with the use of more advanced techniques such as Machine learning and data mining.

The main objective of this work is to propose a system that is capable of automatically classifying concepts of a company to construct ontologies. In other words, it proposes a system that is capable of: collecting and preparing a set of data of a company; automatically classifying the concepts to produce a knowledge tree; and construct an ontology from the knowledge tree generated.

The rest of this paper, shall first present the state of art (Sect. 2), then it will present an approach to automatically construct ontologies and its management (Sect. 3); An application of this approach on the enterprise yowyob will exposed in Sect. 4; finally, the results obtained will be discuss (Sect. 5). At the end of this paper a general conclusion and some future work is exposed.

## 2 State of the Art

The classification of concepts to construct an ontology is a difficult problem [14]. To define this problem, let's consider a set E of concepts to be classified with the number of elements of E $> 0$. Then can we find a partition of the set E in K sub-classes? That is, Can we automatically construct an ontology by finding among all the partitions $P_k$ of a set E, a partition $P^*$ that maximizes the value of the classification, $W(P_i)$? So, we are searching for a partition $P^*$ such that

$$W(P^*) = \underset{P \in P_k}{max} \; \Sigma_{i=1}^{k} sim(C_i) \tag{1}$$

**Ontologies:** The most quoted definition of an ontology is the following one by [9]: An ontology is a specification of a conceptualization. By conceptualisation, he refers to an abstract view of real-world entities and their relations to be represented [5]. An ontology can be seen as a 6-uplet where its components are: Concepts, Attributes, Relationships, Functions, Individuals or Instances and Axioms [11].

$$Ontology = \{C, A, R, F, I, X\} \tag{2}$$

where C= CONCEPTS, A= ATTRIBUTES, R= RELATIONSHIPS, F= FUNCTIONS, I= INDIVIDUALS, X= AXIOMS.

# CAOGen: An Automatic Ontology Constructor Based on Data Mining Techniques

Thomas Djotio Ndie[1(✉)], Bernabé Batchakui[2], Cyril Deyou Ngounou[3], and Karl Jonas[4]

[1] National Polytechnic School of Yaounde (ENSPY), University of Yaounde, Yaounde I, Cameroon
tdjotio@gmail.com
[2] National Polytechnic School of Yaounde (ENSPY), University of Yaounde I, Yaounde, Cameroon
bbatchakui@gmail.com
[3] Computer Science Department, University of Yaounde I, Yaounde, Cameroon
deyoucyril35@gmail.com
[4] Bonn-Rhein-Sieg University of Applied Sciences, Rheinbach, Germany
karl.jonas@h-brs.de

**Abstract.** The construction of ontologies is a classification process in which concepts in a the domain and the relationships between these concepts need to be identified. The classification of concepts to construct an ontology is a difficult problem. The holistic objective is to propose a system that is capable of automatically classifying concepts of a company to construct ontologies. To achieve this, researchers proposed some solutions among which Norms2Onto, Text2onto and APOET (Automatic Product Ontology Extraction from Textual) but majority of them are either domain specific or construct ontologies with generic data which makes the resulting ontology not precise. This paper proposes a system named CAOGen (Company Automatic ontology Generator) that applies CRISPDM (Cross Industry Standard Process for Data Mining) methodology to construct ontologies by using data mining techniques to automatically classify concepts of a specific company to produce its ontology while using wordnet and wikipeadia to augment them. The validation of this work is done through the construction of an ontology of catalogue of service for an Enterprise in Cameroon named yowyob (yowyob.com). After evaluation of the ontology, the system had an accuracy of 0.994, a precision of 0.992, a recall of 1.0 and a F1 score of 0.996 for semantic search and an accuracy of 0.888, a precision of 0.990, a recall of 0.849 and a F1 score of 0.914 for recommendation.

**Keywords:** Automatic construction of ontologies · Data mining · Classification of concepts

## 1 Introduction

A set of information about a domain can be used to develop knowledge in that domain [12]. This knowledge which may be very simple can be combined in the


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 163–172, 2025.
https://doi.org/10.1007/978-3-031-81573-7_13

10. Kelleher, J.D., Tierney, B., Tierney, B.: Data Science An Introduction. CRC Press (2018)
11. Powers, D.M.: Evaluation: from precision, recall, and F-measure to ROC, informedness, markedness, and correlation. J. Mach. Learn. Technol. **2**(1), 37–63 (2011)
12. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychol. Methods **7**, 147 (2002)
13. Ford, B.L., et al.: Missing data procedures: a comparative study. Sampling Studies Section, Sample Surveys Research Branch, Statistica (1976)
14. Wayman, J.C.: Multiple imputation for missing data: what is it and how can I use it. Annual Meeting of the American Educational Research Association, Chicago, IL, vol. 2, p. 16 (2003)
15. Ridzuan, F., Zainon, W.M.N.W.: Diagnostic analysis for outlier detection in big data analytics. Procedia Comput. Sci. **197**, 685–692 (2022)
16. Gleason, T.C., Staelin, R.: A proposal for handling missing data. Psychometrika **40**, 229–252 (1975)
17. Acuna, E., Rodriguez, C.: A meta analysis study of outlier detection methods in classification. Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, vol. 1, p. 25 (2004)
18. Liu, H., Motoda, H.: Data reduction via instance selection. In: Instance Selection and Construction for Data Mining, pp. 3–20. Springer (2001)
19. Chen, J., Shao, J.: Nearest neighbor imputation for survey data. J. Official Stat. **16**, 113 (2000)
20. Sinsomboonthong, S.: Efficiency comparison in prediction of normalization with data mining classification. Diabetes **768**, 231 (2021)
21. Livingston, F.: Implementation of Breiman's random forest machine learning algorithm, ECE591Q Machine Learning Journal Paper, pp. 1–13 (2005)

processing method without normalization or encoding ($f1\_score1$), and finally by applying our complete data processing method ($f1score2$). The complete data preprocessing technique has been utilized to enhance the classification models performance, which is measured using the $F1\_score$. Techniques like normalization and coding have been employed to enhance the quality of the input data [15]. A comparative analysis of the model's performance before and after the data preprocessing can assist in determining the effectiveness of this method. However, this technique may not be suitable for all types of data. Nonetheless, by applying data preprocessing techniques, the performance of classification models can be significantly improved (Fig. 2).

## 5 Conclusion

In this paper, we propose a framework that addresses quality data issues and aims to generate high-quality datasets for machine learning algorithms. The framework employs an architectural approach that combines different data preparation techniques. These techniques include handling missing data, encoding, and normalization. The results of our implementation show satisfactory results.

While the results obtained from the Iris Dataset are encouraging, it would be valuable to assess the performance and the robustness of the proposed data pre-processing framework on a wider range of datasets, under different scenarios, including noisy data, imbalanced datasets, and varying levels of data quality.

## References

1. Garcia, S., Luengo, J., Herrera, F.: Data Preprocessing in Data Mining, pp. 59–105. Springer (2015)
2. Potdar, K., Pardawala, T.S., Pai, C.D.: A comparative study of categorical variable encoding techniques for neural network classifiers. Int. J. Comput. Appl. **175** (2017)
3. Luengo, J., García, S., Herrera, F.: On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl. Inf. Syst. **32**, 77–108 (2012)
4. Song, Q., Shepperd, M.: A new imputation method for small software project data sets. J. Syst. Softw. **80**, 51–62 (2007)
5. Pandey, A., Jain, A.: Comparative analysis of KNN algorithm using various normalization techniques. Int. J. Comput. Netw. Inf. Secur. **9**, 36 (2017)
6. Cousineau, D., Chartier, S.: Outliers detection and treatment: a review. Int. J. Psychol. Res. **3**, 58–67 (2010)
7. Motoda, H., Liu, H.: Feature selection, extraction, and construction. Commun. IICM (Institute of Information and Computing Machinery, Taiwan) **5**, 67–72 (2002)
8. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. **22**(2), 85–126 (2004)
9. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)

Also, for quality checks, we set the threshold value at 80%. It should be noted that other performance measures can also be used depending on the specifics of the data and the situation.

Table 1, shows the results ($f1\_score2$) obtained after testing our dataset with different algorithms.

**Table 1.** Model Evaluation Results.

| algorithms | $f1\_score0$ (%) | $f1\_score1$ (%) | $f1\_score2$(%) |
|---|---|---|---|
| Support Vector Machine (SVM) | 61.33% | 90.85% | 93.55% |
| Random Forest Classifier | 73.66% | 91.53% | 95.58% |
| XG Boost | 75.55% | 91.35% | 95.58% |
| Naive Bayes | 76.65% | 91.53% | 95.58% |
| K-Nearest-Neighbors Classifier | 72.65% | 92.43% | 94.03% |

Let us consider:

- $f1\_score0$ Results before applying our processing method,
- $f1\_score1$ Results after applying our processing method without normalization and encoding,
- $f1\_score2$ Results after applying our processing method with normalization,

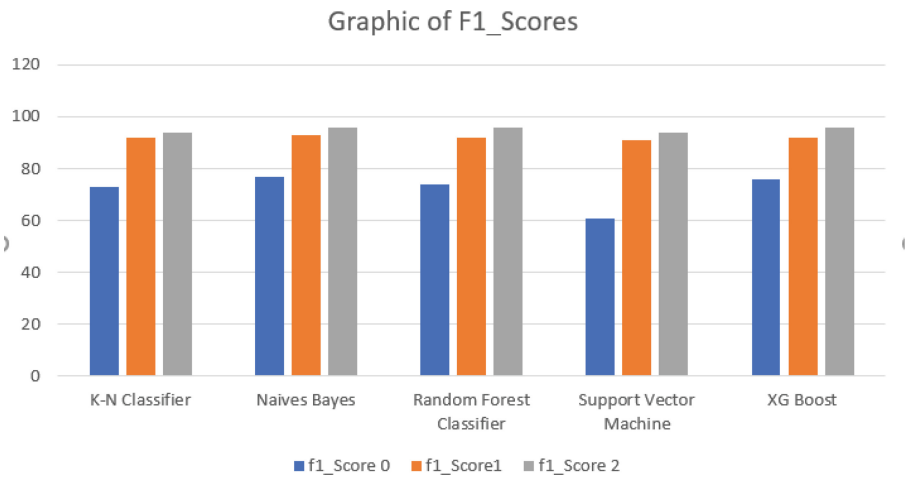



**Fig. 2.** Graphic of $F1\_scores$ (%)

Table 1 shows firstly the results obtained without applying our data processing method ($f1\_score0$), secondly the results obtained by applying our data

# 4 Experimentation of Approach

## 4.1 Programming Language and Framework

To facilitate and accelerate the implementation of our solution, we chose the Python programming language and the framework Streamlit. Python is a programming language that respects object-oriented programming paradigms. Consequently, as a Python-based development framework, streamlit was chosen. Streamlit is an open-source platform that enables the creation of data apps with Python, using Python scripting APIs, widgets, and instant deployment. The platform integrates various libraries such as Scikit Learn, OpenCV, vega-Lite, PyTorch, Numpy, Seaborn, Deck GL, Tensorflow, Python, Matplotlib, and Pandas, which can be helpful for machine learning engineers to create Python-based applications.

## 4.2 Execution Environment

Our work setup consists of a laptop running Ubuntu Operating System with the following specifications.

- Processor: Intel(R) Core(TM) i5-8259U CPU @ 2.30GHz, 4 cores and 8 logic processors
- RAM: 8 Gb
- Graphics: Intel(R) Iris(R) Plus Graphics 655

## 4.3 Presentation of the Dataset

We have decided to use the Iris flower dataset to test our data processing approach. This dataset is a multivariate set that was made famous by the British statistician and biologist. It comprises 50 samples from each of the three Iris species (Iris setosa, Iris virginica, and Iris versicolor). Four features were measured from each sample, including the length and width of the sepals and petals in centimeters. You can access the Iris dataset through the provided link https://www.kaggle.com/datasets/saurabh00007/iriscsv/download?datasetVersionNumber=1.

## 4.4 Results and Discussion

This section presents the results of fitting our models to the dataset. We measured the performance using the $F1\_score$. The $F1\_score$ considers precision and recall, two important metrics for classification models. Precision measures the number of true positives divided by the total number of positive predictions, while recall measures the number of true positives divided by the total number of true positives and false negatives. The $F1\_score$ is a more reliable measure than accuracy because it takes into account both precision and recall [11].

**Algorithm 1.** Algorithm used by the High-Quality Data Preparation framework

1: **Input**: $D$: Original dataset with $C_n$ columns
2: ML:Machine learning algorithm
3: nval: number of folds of the cross-validation
4: $\sigma$: Threshold
5: **Output** : $QD$ : Quality data
6: **Begin**
7: $D \leftarrow MissingData(D)$
8: $D \leftarrow Outliers(D)$
9: **for all** column $C$ in $D$ **do**
10: **if** $C$ is numeric data **then**
11: $C \leftarrow Normalizer(C)$
12: **else**
13: $C \leftarrow Encoder(C)$
14: **end if**
15: **end for**
16: $F1\_score \leftarrow TrainML(ML, D, nval)$
17: **if** $F1\_score \geq \sigma$ **then**
18: $QD \leftarrow D$
19: **else**
20: Review your dataset
21: **end if**
22: **End**

The Algorithm 1 illustrates the algorithm used by the High-Quality Data Preparation framework for preparing high-quality data. Lines 1 to 4 specify the input data, which includes the dataset, quality threshold, and Machine Learning algorithms. Line 5 indicates that the expected output is high-quality data. In line 7, we first address missing data, and in line 8, we utilize the previously processed dataset with missing data to handle outliers [20]. From line 9 to line 15, we iterate through each column of the dataset, normalizing numerical data (line 11) and encoding categorical data (line 13). We now have a prepared dataset ready to be used in the learning process. In line 15, the function train is used to train the model [21]. We use the Cross-validation to evaluate the model. The cross-validation model randomly divides the training data into *nval* folds ($nval = 10$). In each iteration of the dataset, the cross-validation model uses one fold as the validation dataset. It uses the remaining $nval - 1$ folds to train a model. Each of the *nval* models is tested against the data from all the other samples. In line 16, the function TrainML is used to train the model. We consider 70% of the prepared dataset as the training data, leaving 30% for testing purposes. The $F1\_score$ (line 16) metric is used to assess the performance of the model and to decide whether the data are of high quality or not. the function TrainML repeats the training of the model *nval* times. In lines 16 to 20, if $F1\_score$ is greater than or equal to the threshold specified we return high-quality data else we prompt the user to review their dataset.

missing values with the mean of non-missing values of the same variable. This technique is widely used in the scientific community for its simplicity and robustness. For the treatment of outliers, the interquartile range technique is proposed. This technique detects outliers by calculating the difference between the first and third quartiles of a variable. Values outside this range are considered outliers and replaced by the variable's median value [8,12]. It should be noted that these techniques are not the only ones available for handling missing and outlier data.

- **step 2:** Data normalization and encoding are also two important steps in data preparation. In our approach, two techniques are proposed for normalizing and encoding data. The process begins by checking whether categorical data are present for each feature in the dataset. When the selected feature contains categorical data, the hot coding technique is proposed to transform the data. This technique involves creating binary variables for each category of the categorical variable. This technique is widely used in the scientific community to encode categorical data due to its simplicity and its ability to avoid unnecessary weighting of categories. When the selected feature does not contain categorical data, normalization is performed. For normalization, the min-max normalization technique is proposed. This technique involves transforming the data to fall within a specific range of values, usually between 0 and 1. This technique is widely used in the scientific community to normalize data due to its simplicity and robustness [9]. After normalization and encoding, the data is transformed and ready to be used for analysis and modeling. It should be noted that these techniques are not the only ones available for data normalization and encoding.

$$X_{normalized} = \frac{\chi - min(\chi)}{max(\chi) - min(\chi)} \tag{1}$$

where: min and max are the minimum and the maximum value of $\chi$

- **step 3:** Data quality testing is also an essential step in data preparation. In this paper, an approach is proposed to check that data have been properly cleaned. For this, models based on machine learning algorithms such as support vector machines (SVM), Naive Bayes, K-Nearest-Neighbors Classifier XG Boost Classifier, and random forests are used. This evaluation is carried out using the input data transformed to form these models. Model performance is then measured against parameters such as the $F1\_Score$ [10,17]. In this approach, if the $F1\_score$ is greater than or equal to a defined threshold, the data are considered to be of high quality. However, if the $F1\_score$ is below the defined threshold, the data quality testing process starts again.

The following Algorithm 1 is used to describe our methodology.

According to the research conducted by *Motoda and al.* in [7], feature selection is a crucial process that aims to eliminate some of the initial features and keep only the most relevant ones. This can be achieved by using a specific criterion to optimally reduce the feature space. On the other hand, feature extraction refers to the process of generating new features that can be utilized independently or in conjunction with others.

Inspired by these methods, we propose an architectural method that gathers data preprocessing methods to produce quality data. In the next section, our method which is the combination of the different methods contained in the above-mentioned works is described.

## 3 Approach to Data Preprocessing

The Fig. 1, below presents our method of data preprocessing which is based on the different techniques that have been defined in the existing work.

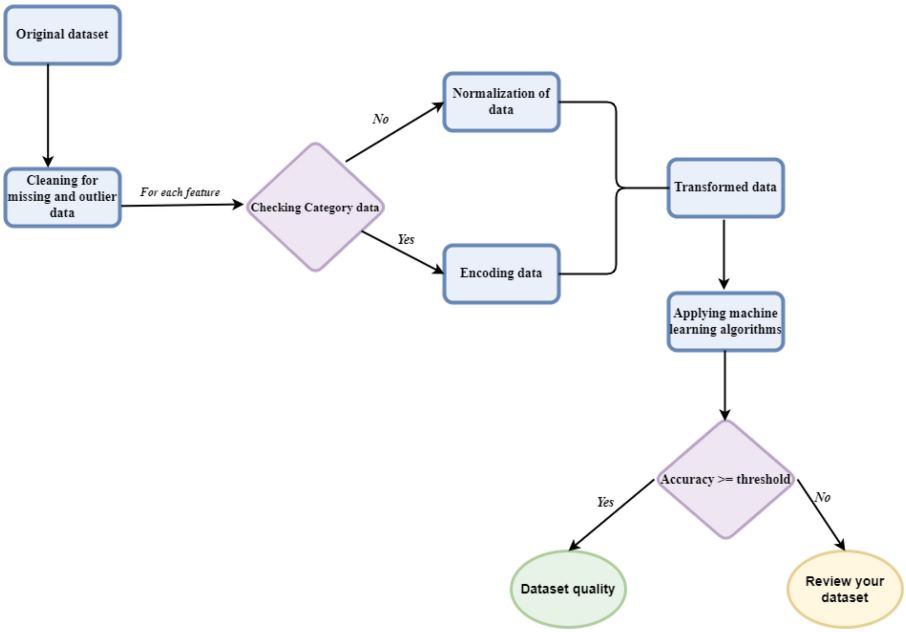



**Fig. 1.** approach of data preprocessing

– **step 1:** Handling missing and outlier data is a crucial step in preparing data for analysis and modeling. In our approach, two techniques are proposed for handling missing and outlier data. For the treatment of missing data, the mean imputation technique is proposed. This technique involves replacing

anomalies [19]. Therefore, it is imperative to set up a data preprocessing method that integrates the essence of data preprocessing techniques, allowing all those who wish to obtain quality data from the raw data.

In this paper, we introduce an architectural approach that combines various data preprocessing techniques to generate high-quality data for machine learning algorithms. Data quality refers to the condition of data based on factors such as accuracy, completeness, consistency, and reliability. Measuring the level of data quality can help identify potential errors that need to be corrected. The data pre-processing process intervenes in the correction of these errors.

Thus, this paper is structured as follows. Section 2 discusses related work on preprocessing techniques. Section 3 presents our proposed preprocessing approach. Section 4 presents the experimentation of our approach we will end with a conclusion and perspectives in Sect. 5.

## 2 Related Work

There are many scientific studies on data preparation techniques based on machine-learning algorithms. For each technique, a lot of work has been done to improve data preprocessing. *Potdar and al.* [2] have compared seven different techniques that can be used for encoding categorical variables. The main objective of this study was to classify a dataset using neural networks. To achieve this, the authors made use of a second-hand car dataset. Based on the prediction results, the sum Coding and Backward Difference Coding techniques have provided an accuracy of 95%. Therefore, it can be inferred that these two techniques are most suitable for making predictions that involve a categorical dataset.

According to [3], when analyzing data, we often face a loss of information due to missing values. To tackle this problem, several techniques have been developed, such as imputation techniques, which help in substituting the missing data. The authors have presented several methods for imputing the missing data, and analysis of the processed data using these techniques has proven to improve the accuracy of the model. In the analysis of data, we find losses of information due to the presence of missing values. To remove these missing data, various techniques have been explored, such as the use of imputation techniques, and the authors have presented various methods for imputing missing data. Analysis of the data processed using imputation techniques helps to improve model accuracy [3] and [4,13,14,16].

*Pandey and al.* [5] describe two normalization techniques that they implemented. The authors use these techniques to create a global classification of IRIS data and measure the accuracy of their approach using the cross-validation method and the R programming language. Specifically, the article focuses on two normalization approaches: Z-Score normalization and Min-Max normalization. *Cousineau and al.* [6] discuss different methods for detecting possible outliers. These techniques can be divided into two categories: those that work with univariate data and those that work with multivariate data. The work will evaluate each case and provide appropriate recommendations.

# Towards a Framework for the Preparation of High Quality Data for Use by Machine Learning Algorithms

Rasidatou Nabi[1(✉)], Yaya Traoré[1], and Julie Thiombiano[2]

[1] University of Joseph KI-ZERBO, Ouagadougou, Burkina Faso
rasidatou.nabi@ujkz.bf
[2] University of Nazi BONI, Bobo-Dioulasso, Burkina Faso

**Abstract.** Nowadays, companies and organizations have access to various data collection tools that enable them to amass vast amounts of data, which can be stored in databases. This data can be leveraged by machine learning algorithms to extract valuable information for decision-makers. However, this raw data is often of poor quality, containing errors such as missing data and outliers, requiring the intervention of technicians and domain specialists to prepare the data to ensure the $F1\_Score$ of the analysis. This article proposes a framework for preparing high-quality data for machine learning algorithms, as manually identifying reliable data from a large pool can be challenging and time-consuming. Our approach is an architectural method that combines data preparation techniques to generate dataset quality.

**Keywords:** Data processing · Quality data · Missing data · Encoding · Normalization

## 1 Introduction

Presently, many data collection tools have been developed, which allow companies and organizations to collect large amounts of data and store them in databases. Machine learning can analyze data to provide actionable insights for decision-makers. However, the raw data collected is often of poor quality because it can contain inaccurate data, missing data, outliers, and so on. Therefore, a data processing phase is necessary.

Data processing refers to any activity aimed at improving the quality, usability, accessibility, or portability of data. The ultimate goal of data preparation is to enable people and analytical systems to have clean, usable data converted into useful information [1]. Data processing involves analyzing raw data to produce quality results. This includes cleaning missing data and outliers, encoding categorical data, normalizing data, and reducing data through attribute selection.

Several techniques are used to process the data depending on the anomalies that may exist and the types of data. For this purpose, many contributions have been made, but each contribution does not take into account the treatment of all


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 154–162, 2025.
https://doi.org/10.1007/978-3-031-81573-7_12

7. Flocon-Cholet, J.: Classification audio sous contrainte de faible latence. Theses, Université de Rennes (2016). https://theses.hal.science/tel-01395495
8. Isard, A., McKelvie, D., Mengel, A., Møller, M.B.: The MATE workbench annotation tool, a technical description
9. Jiang, H., Wu, X., Xie, X., Wu, J.: Audio public opinion analysis model based on heterogeneous neural network. In: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 449–453. https://doi.org/10.1109/ICCECE51280.2021.9342052
10. Kipp, M.: Anvil-a generic annotation tool for multimodal dialogue. In: Seventh European Conference on Speech Communication and Technology. Citeseer (2001)
11. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE **13**(5), e0196391. https://doi.org/10.1371/journal.pone.0196391. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391
12. Luitel, S., Anwar, M.: Audio sentiment analysis using spectrogram and bag-of-visual- words. In: 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI), pp. 200–205 (2022). https://doi.org/10.1109/IRI54793.2022.00052
13. Ouedraogo, I., Some, B.M.J., Benedikter, R., Diallo, G.: Mobile technology as a health literacy enabler in African rural areas: a literature review. preprint, In Review (2021). https://doi.org/10.21203/rs.3.rs-243773/v1. https://www.researchsquare.com/article/rs-243773/v1
14. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964. ISSN: 2379-190X
15. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8 (2013). https://doi.org/10.1109/FG.2013.6553805
16. Sloetjes, H., Wittenburg, P.: Annotation by category - ELAN and ISO DCR
17. Yin, Y., Hanes, D.W., Skiena, S., Clouston, S.A.P.: Quantifying healthy aging in older veterans using computational audio analysis. J. Gerontol. Ser. A glad154 (2023). https://doi.org/10.1093/gerona/glad154
18. Zheng, Y., Peng, J.E.: ELAN (EUDICO linguistic annotator). RELC J. **53**(2), 469–474 (2022). https://doi.org/10.1177/00336882221089052

application works and its labelling approach. The results of this experimental phase satisfy the objectives set for the introduction of this tool.

## 5 Conclusion

PLAVIDA, the audio data annotation platform, has been presented: from its architecture to the structure of the final annotated data, passing through the labelling logic. It consists of a frontend implemented by an Android application which uses an API to access the backend managed by the python language. PLAVIDA is an easy, intuitive platform for annotating audio and video data in African languages. It has a convivial interface that is easy to use by all categories of people (literate and illiterate). It also allows annotated audio and video to be available in 3 data formats: CSV, XML and JSON. PLAVIDA would be useful to researchers who wish to conduct studies on audio classification by allowing them to annotate large corpora of data in their language of choice. The basic functionality of the application, which is to annotate audio and video, is operational. The application is easy to install on smartphones and robust. However, there are many functionalities that could be the subject of future work on the application. The introduction of languages and audio data by users needs to be developed, while ensuring the reliability and security of these data. The ability to annotate data offline should be integrated into the application to make it easier to use. Audio and video segmentation functionality needs to be added to the application backend. We are working to integrate these functionality into the application in the near future.

## References

1. Bagher Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2236–2246. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1208. https://aclanthology.org/P18-1208
2. Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., Liberman, M.: ATLAS: A flexible and extensible architecture for linguistic annotation. http://arxiv.org/abs/cs/0007022
3. Bird, S., Liberman, M.: A formal framework for linguistic annotation. http://arxiv.org/abs/cs/9903003
4. Boukabous, M., Azizi, M.: Multimodal sentiment analysis using audio and text for crime detection. In: 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp. 1–5 (2022). https://doi.org/10.1109/IRASET52964.2022.9738175
5. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008)
6. Etienne, C.: Apprentissage profond appliqué á la reconnaissance des émotions dans la voix. https://theses.hal.science/tel-02479126

annotated data in one of the main languages spoken in Burkina Faso. From the 100 audio files, 75 were annotated 5 times on 5 through the application. 3 were annotated 3 times on 5 and the 2 others did not receive any annotation. The 75 files annotated 5 times were evaluated following the process described through Sect. 2.2. The annotated data was successfully exported in CSV format, JSON format and XML format. Note that there were audio files for which, on the 5 possible annotations for an audio, two different emotions were each attributed twice and a third different emotion. These audio files express emotions that are very similar. For example, relief and satisfaction in a speech, hubris and pride in a speech, or shame and embarrassment in a speech, and so on.

## 4 Discussion

The choice of 5 annotators per sound is motivated by the nature of the audio and video data that will be introduced into the application for annotation. The authors, who limited themselves to 3 annotations [6], recorded the audio and video data themselves in an environment that was already very well prepared. These sentences are written in such a way that they already express the emotions, i.e. the environment is very clear and the quality of the data is high. However, we believe that the domain of expression of the data used by these authors is very restricted and that their emotion recognition models could show limitations when we will use them on data collected in a popular domains such as radios and telephone calls. We will provide the ability to annotate data recorded in popular domains such as speech recorded during radio and television broadcasts. Given that the identification of emotions in these types of audio and video will not be as simple for an annotator as in existing work, we have therefore chosen 5 annotations per sound to allow better labelling of this data.

The other aspect which also justifies the choice of 5 annotators is the target of the annotators. Indeed, most of the existing works target some annotators who are only natives of the language to be annotated and who would have a good experience (high age). But given the multitude of African languages and the large amount of data that will be collected from radio and television, this approach does not make it easy to obtain very large quantities of annotated data. We are going to make it possible to annotate the data for a larger number of people, such as pupils, students, shopkeepers, farmers and many others. To ensure that the data is properly annotated, we need to go beyond 3 annotations per sound. In addition, the annotation made by an illiterate profile should be prioritised in the case of a tie in the 5 annotations, because an illiterate profile annotator has a better experience in his language since he only speaks this language on a daily basis.

In order to make PLAVIDA easier to use and more accessible to annotators, we decided to use Android technology, which is widely used in Africa and is currently growing [13].

For a test phase, we used 100 audio files. This number of files may seem small for a full evaluation of the application, but it gives an idea about how the

**Table 1.** Description of final table content

| Column name | Description |
|---|---|
| Id_audio | corresponds to the audio identifier |
| audio_language | corresponds to the language in which the audio was spoken |
| audio_reference | corresponds to the name used to identify an audio in the list of audio files. It is unique for each audio |
| emotion_name | corresponds to the emotion attributed to an audio |
| emoji | Corresponds to the code associated with an emotion in csv format |


(a) JSON format (b) XML format

(c) CSV format

**Fig. 6.** Overview of the three possible data export formats

### 3.3 Experimentation

At the time we are writing this paper, the platform has been used by 20 students (literate profile) from the Nazi BONI University and 20 people (illiterate profile) to annotate the audio in three languages: Moore, Dioula and Fulfulde. A total of 100 audio files of 10 min were entered into the platform. These audio data was collected from YouTube and segmented in order to facilitate the identification of an emotion. For efficiency reasons, each annotator could annotate a maximum of 10 audios during the test. We have not yet done a large-scale evaluation of the software functioning in the context of a large-scale annotation project, which is essential to fully demonstrate the validity of our approach. However, local audio is currently being collected from local radio stations in order to create a corpus of

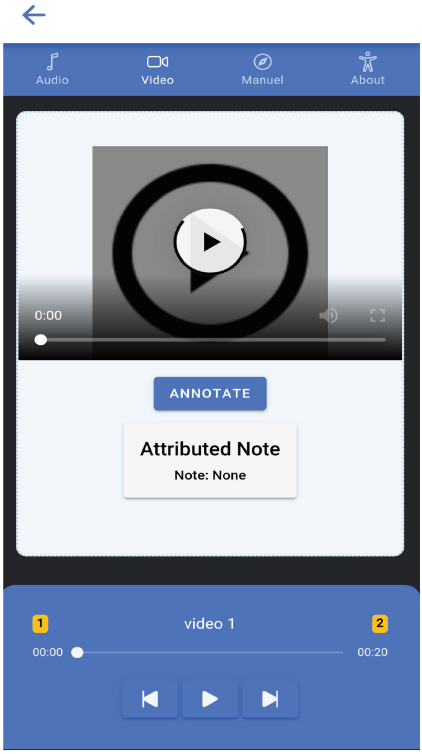



**Fig. 5.** Video annotation page

– The third point describes multimodal annotation. It presents a description of the three modalities: sound, gesture and visual. It also explains how to perform multimodal annotation.

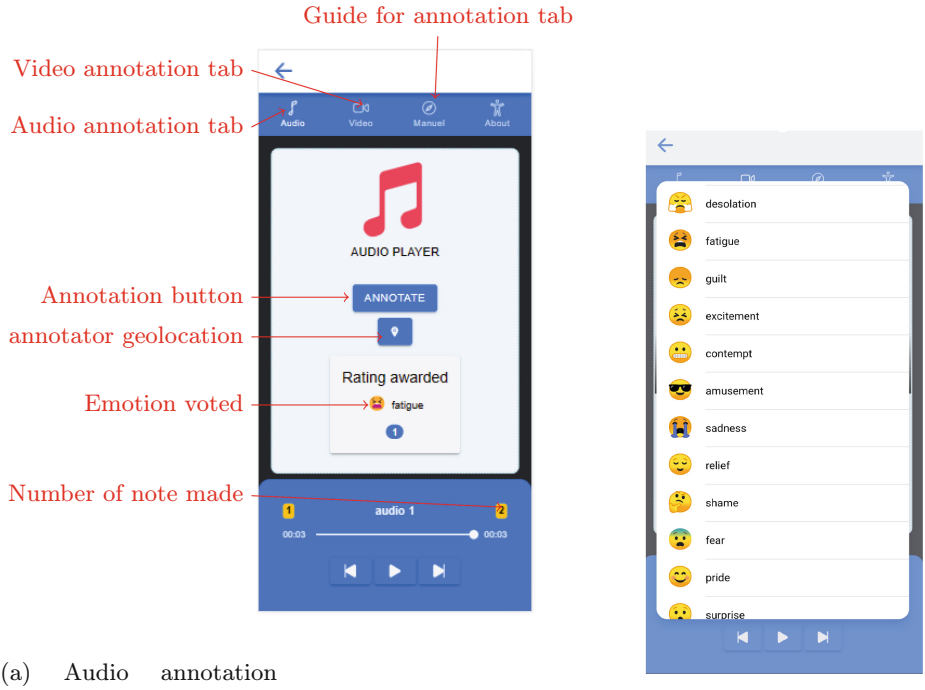
These presentations are intended to guide annotators who have no knowledge of emotions and annotations. They are not intended to influence any annotator. We recommend annotators to assign emotions to sound and video freely and according to their feelings.

### 3.2 Description of Final Data Structure

As the aim is to have annotated data available in several languages to facilitate machine learning work, we propose 3 data formats in which anyone wishing to work with their data can export the final table for Machine Learning tasks. These 3 formats are CSV, XML and JSON. These are data formats the most widely used as input to Machine Learning algorithms. The final data contain the informations shown in the Table 1.

Figure 6a shows the json format of the exported data, Fig. 6b shows the xml format and Fig. 6c the csv format.

annotate audio files in their mother tongue. Figure 4a is the annotation page and Fig. 4b is the list of emotions used to annotate a sound.




(a) Audio annotation page

(b) Emotion list

**Fig. 4.** Audio annotation page and motions list

**Video Annotation:** This is the part that concerns video annotation. In this part, the annotator must take into account three modalities: sound, visual and gestures. A list of videos also appears. On the basis of these three modalities, the annotator assigns an emotion to the video. A list of emotions is also available at this level with emoticons to enable an 'illiterate' person to annotate a video. For more information on multimodal annotation, the annotator should refer to the "annotation guide" section. The Fig. 5 represents the video annotation page.

**Annotation Guide:** The Annotation Guide screen presents three points:

- The first point explains the emotions and a description of the emotions that appear in the audio and video annotation pages. The aim of this section is to provide a clear understanding of certain emotions that are not well known by certain annotators. It also to highlight the nuances between similar emotions. This section also shows the correspondence between the names of the emotions and the emoticons.
- The second point explains how to annotate a sound and how the labelling will be done after the different annotations;

# 3 Results

## 3.1 User Interfaces

It consists of four (04) parts:

**Registration and Login Interface:** The annotator registration section provides the annotator's qualifications in the language they are annotating. This section also makes it possible to specify the annotator's profile: "literate" profile and "illiterate" profile. This is an important factor in the reliability of his annotation. In the case of a discrepancy on an annotation, we will consider the annotation made by the "illiterate" profile, who is a native speaker of the language. This is because it is assumed that this person uses the same language daily. To annotate a sound, the user must log in. Logging in allows us to retrieve the annotator's identifiers and find out who has annotated which sound. Figure 3a is the account creation page and Fig. 3b is the login page.


(a) Sign Up page (b) Sign In page

**Fig. 3.** Account creation and login page

**Audio Annotation Interface:** This section plays audio files randomly. It also allows users to listen to a sound and assign it an emotion using an emotion button. This button displays a list of emotions with their corresponding emoticons. The presence of the emoticons allows an 'illiterate' user to annotate the sounds without being able to read the name of the emotion. This interface also shows the number of annotations already made for a sound. A sound is only visible on this interface when its maximum annotation has not been reached. We have set the maximum annotation at 5. Some works have considered a maximum annotation of 3 annotators [6]. For reasons of annotation quality, annotators can only

account to the detriment of the annotation from a user with a literate profile. If the annotations are made by users with the same profile, it is difficult to distinguish between them. We cancel the annotations for this sound and add it back to the list of audio files to be annotated. This will allow other users to annotate this sound. This process is described by the diagram of the Fig. 2.

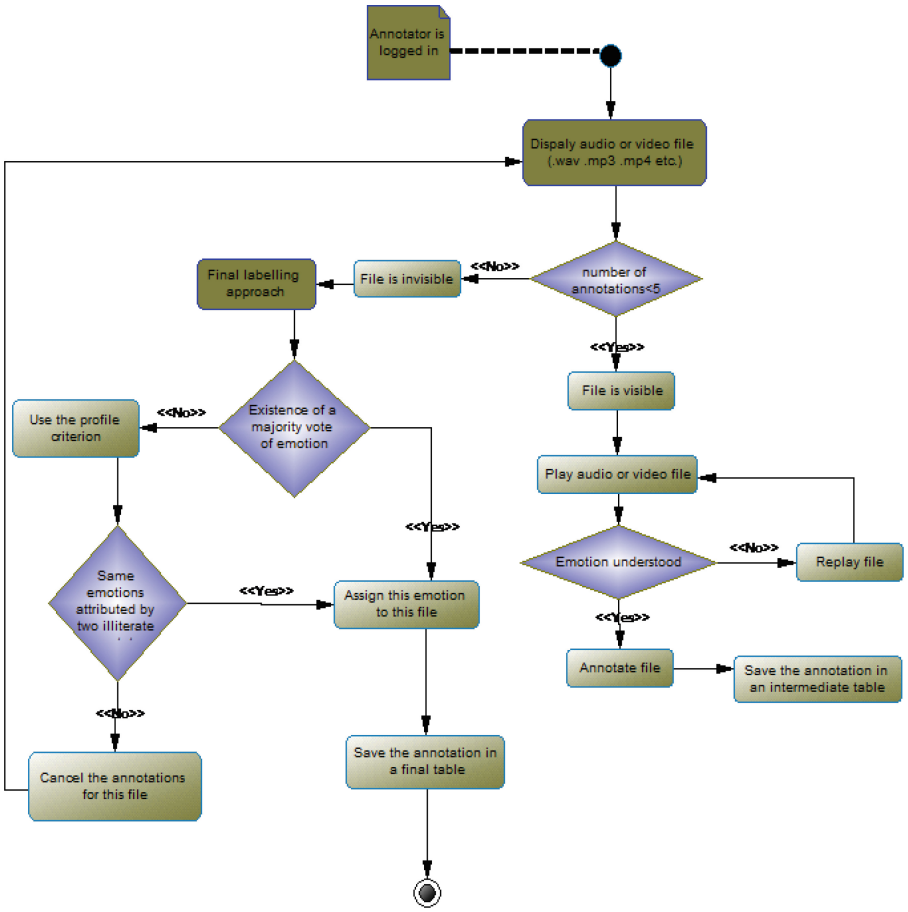



**Fig. 2.** Labelling logical description diagram

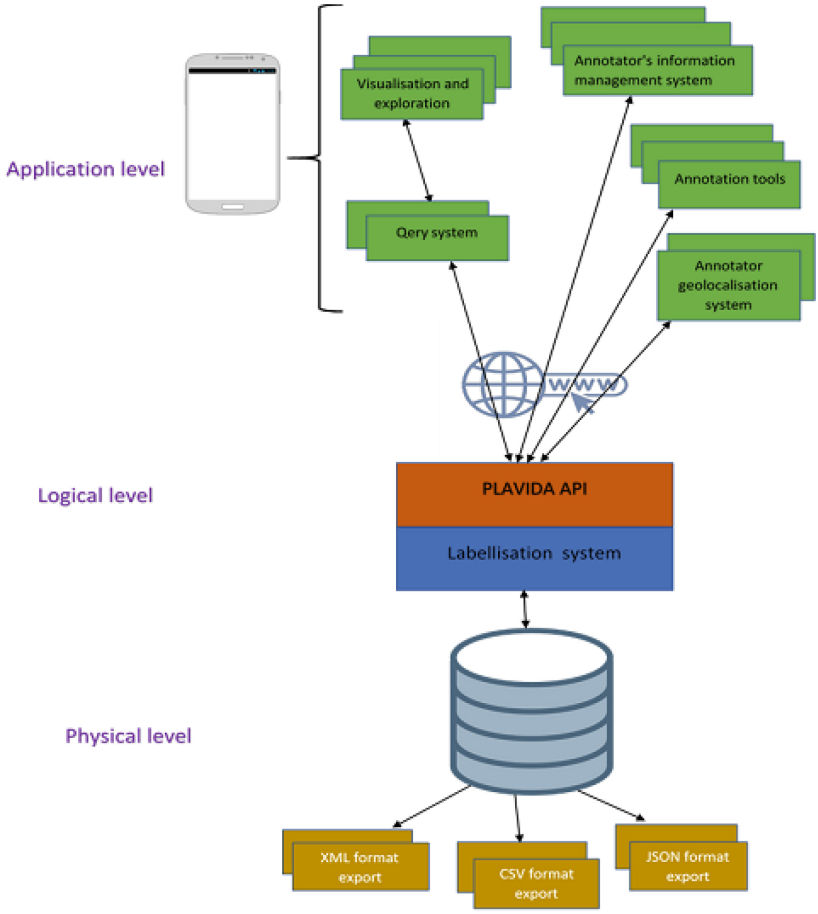



**Fig. 1.** PLAVIDA Architecture

evaluate that sound and its annotation is stored in a final table. This final table indicates the final emotion associated to this sound. The evaluation of a sound is based on two criteria, one of which is a priority and the other a secondary criterion. The priority criterion consists in choosing an emotion that has received a majority vote. In others words, the number of attribution of this emotion to this sound, as it was the case in the studies of [5]. From a total of 5 annotations for a sound, if at least 3 annotators attribute the same emotion to this sound, then we consider this sound expresses this emotion, whatever the profile of the annotators. But if there is an equality in the annotation of a sound with different emotions, for example on the 5 annotations if two emotions appear each twice in the annotation of a sound, we use the second criterion to decide and choose the emotion that seems to be better adapted. In this case, we use the annotator's profile. The annotation from a user with an illiterate profile will be taken into

## 2 Methodology

To implement this tool, a state of the art was first carried out. The aim of the state of the art was to identify existing tools for audio and video data annotation and to see the applicability of these tools for annotating audio and video in African languages. The existing tools did not take into account African languages and could not be well used to annotate these languages.

Then, we have defined the architecture of the tool we want to implement. This architecture has enabled us to understand the general structure of the system, the relationships between the elements that make up the application and the different functionalities that need to be developed. This structure is the result of a series of strategic decisions we made during the analysis and design phase.

### 2.1 PLAVIDA Architecture

The architecture of our solution is shown in Fig. 1 and consists of three levels: the application level, the logical level and the physical level.

**The Application Level**
It relays requests from the user (annotator, administrator) to the logical layer, and in return presents the information returned by the processing of this layer. To facilitate access to PLAVIDA, we have opted to use it through a smartphone. We have therefore implemented this layer through an Android application.

**The Logical Level**
The logical level includes annotation logic and an API for accessing the database. The annotation logic consists of the majority vote, the user profile and the language to which the annotator belongs. The API defines a set of procedures that describe the operations that the application performs on the data in response to user requests made through the presentation layer. These operations include creating, modifying, searching and storing annotations.

**The Physical Level**
It is the part that manages access to PLAVIDA data. The storage strategy we are using is an RDBMS (Relational Database Management System). The data can therefore be accessed via SQL queries formulated by the user.

Finally, in order to have audio and video data very well annotated, and based on the work of [6], we have defined the following labelling approach:

### 2.2 Labelling Approach

We count the number of annotations per sound. Every time a file is annotated, the annotation is recorded in a table with informations about the sound and the annotator. In this table, we have the following informations: the identifier of the audio, the name of the audio, the name of the emotion assigned to it by each annotator, the profile of the annotator and the number of annotations made to this sound. Once a sound has reached the maximum number of annotations, we

with different annotation schemes), platform-independent, based on XML and has an intuitive graphical user interface. For project integration, Anvil allows voice transcripts to be imported and data to be exported in text and table formats for future statistical processing. Annotated data is stored in a single XML file. ATLAS [2]: Architecture and Tools for Linguistic Analysis Systems was born out of the need for applications covering corpus construction, evaluation infrastructure and multimodal visualisation. The main objective of ATLAS is to provide powerful abstractions of annotation tools and formats in order to maximise flexibility and extensibility. ATLAS was inspired by annotation graphs [3], a graph model for linear signals annotation (such as text and speech) indexed by intervals, for which efficient database storage and query techniques are applicable. Common Voice[1] is a web-based platform for collecting voices. It is open to the public and is fed by the voices of volunteer contributors from all over the world. For a language to be included on common voice, a certain number of sentences must be available. The phrases are read by the contributors to create the dataset. ANNEMO (ANNotating EMOtions)[2] is a web-based open-source tool for annotating affective and social behaviours from audiovisual data. In this tool, the annotator must connect to a web-based annotation interface using a unique identifier. The interface is divided vertically into two parts: a scrolling list of audiovisual recordings is presented on the left-hand side, while the video and annotation cursor are displayed one below the other on the right-hand side of the window. All annotation data is automatically saved on a server in the form of log files. This tool has been used to annotate the RECOLA database [15].

As a criticism of these works, first of all, each of these applications was created to annotate an audio database in a specific language, mainly in English, German and French and do not include African languages. Secondly these applications do not allow a sound to be annotated by a large number of annotators, which would not remove any doubt about the annotation of certain sounds that express very similar emotions. Thirdly accessibility to these applications poses a real problem for anyone wanting to set up a database in an African language. Common Voice, to which several languages can be added does not allow emotions annotation; it only allows voice collection. Then none of these tools provides annotated data in both CSV, XML and JSON formats. At last The labelling approach we want to adopt is not taken into account by any of the existing tools, so we cannot use them. It is to overcome all these limitations of existing platforms that we have set up PLAVIDA This paper first presents a methodology in which we present the architecture of PLAVIDA tool and the labelling approach that we have followed. Secondly we present the results by highlighting application's user interfaces, the final structure of the data annotated through the application and the experimentation of the application. Then, we discuss the results and finally we present the conclusion.

[1] https://commonvoice.mozilla.org/fr.
[2] https://diuf.unifr.ch/main/diva/recola/annemo.html.

learning algorithms for the analysis and prediction of a specific aspect in audio. The most important limiting factor in the study of audio classification is the lack of labelled databases. Indeed, the community lacks audio databases:

- of large-sized;
- incorporating a large range of annotations;
- in other languages than English;
- and, most importantly, freely available to all.

To obtain labelled audio data in African languages, we have created an application which allows to annotate a voice with the emotions expressed in it. We called this application PLAVIDA (PLatform for Audio and VIdeo Data Annotation). We have created the application by integrating 3 African languages for the time being: Moore, Dioula and Fulfulde, which are the main languages spoken in Burkina Faso and for which we have voice data available from local radio stations. But the application is designed to be able to add several other languages and to annotate voice data in these languages. The aim of setting up this application is to be able to create speech corpora in various African languages in order to encourage research on audio data in these languages. PLAVIDA would be useful to researchers who wish to conduct studies on audio classification by allowing them to annotate large corpora of data in their language of choice.

Numerous studies have been carried out on audio data analysis [4,9,12,17]. To carry out this work properly, labelled data is required [1,11,14]. Data labelling must follow a rigorous procedure to ensure that each label associated with each piece of data is consistent. To achieve this, some applications have been created to facilitate labelling. Most recent annotation applications are based on Java, use XML for file exchange and have an object-oriented design [10]. MATE [8] is a tool that aims to simplify the tasks of annotating, displaying and querying speech or text corpora. It is designed to help humans to create linguistic resources and to facilitate the use of data by different groups, providing a tool that can be used with many different annotation schemas. Any annotation schema that can be converted to XML can be used with the tool. The tool is written entirely in Java, which means it is platform-independent. The software provides a number of predefined style sheets for use with particular annotation schemes. But its main strength is the ability to write new stylesheets for existing or new schemas. ELAN [16] is an annotation tool for audio and video recordings. With ELAN, a user can add an unlimited number of textual annotations to audio and/or video recordings. An annotation can be a phrase, word, glossary, comment, translation or description of any feature observed in the media. An annotation can be linked to a media item or refer to other existing annotations. Annotation content consists of Unicode text and annotation documents are stored in XML format. EUDICO [18] is an effort to enable multi-user annotation of a centralised corpus via a web interface. The tool should enable multimodal video annotation. EUDICO is based on an existing tool called Media Tagger which is used in various research institutes but requires a special hardware or software configuration. Anvil [10] (Annotation of Video and Language) is a tool for annotating audiovisual content incorporating multimodal dialogues. Anvil is highly generic (usable

# PLAVIDA, an Annotation Tool for Audio and Video in African Languages

Go Issa Traoré(✉), Borlli Michel Jonas Some, Ousmane Ouédraogo, and Lucien Kalmogo

Université Nazi BONI, Bobo Dioulasso, Burkina Faso
goissatraore@yahoo.fr, sborlli@gmail.com, oueo5587@gmail.com, lucienkalmogo21@gmail.com

**Abstract.** PLAVIDA, PLatform for Audio and VIdeo Data Annotation is a platform designed to facilitate audio and video data annotation. To perform sound classification tasks with Machine Learning algorithms, we need annotated data on these sounds. It is on the basis of this annotated data that these algorithms will learn to make classifications. However, the community lacks labelled audio data on African languages. PLAVIDA will allow researchers the opportunity to create a multimedia labelled databases which can be used as input in Artificial Intelligence models. This could boost research around audio classification in several African languages. We have used python and Android IONIC/Angular technology to develop this tool. The innovation in PLAVIDA, is the possibility given to illiterate people to be able to interact with, when we want to labelle sound or video in African local languages. The tool can be then used both by literate and illiterate people. The type of labelling we are faced on concern the emotional perception people can have when listening or watching a media. It incorporates an annotation logic based primarily on the maximum rate of the same emotional perception over all. In the case where there is no majority vote, the user profile criterion is used. The data annotated using this application can be exported in XML, CSV or JSON format. These types of format are the data formats used to create Artificial Intelligence models.

**Keywords:** Annotation tool · Audio and video data annotation · Database · African languages

## 1 Introduction

Audio classification is the process of listening to and analysing audio recordings using computer tools. Otherwise known as sound classification, it is at the heart of various modern Artificial Intelligence technologies, including automatic speech recognition, virtual assistants, text-to-speech applications and so on. It is also used in radio stream analysis, video archiving, audio coding, music classification, auditory scene recognition [7], etc. This work requires well-prepared data based on precise prediction objectives. This allows to apply the appropriate machine


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 141–153, 2025.
https://doi.org/10.1007/978-3-031-81573-7_11

# Ontology, Data Preparation

37. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., et al.: CamemBERT: a Tasty French Language Model, In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219, Online. Association for Computational Linguistics (2020)
38. Lefaso.net: https://lefaso.net/ 28.08.2022
39. Burkina 24: https://burkina24.com/) 28.08.2022
40. Netafrique: https://netafrique.net/) 28.08.2022
41. Doccano: https://github.com/doccano/doccano
42. Chinchor, N., Sundheim, B.: MUC-5 Evaluation Metrics, In Fifth Message Understanding Conference (MUC-5) In: Proceedings of a Conference Held in Baltimore, Maryland (1993)

17. Yadav, V., Bethard, S.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning models, In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics (2018)
18. Bikel, D.M., Miller, S., Schwartz, R., Nymble, R.W.: A high-performance learning name-finder. In: Proceedings of the fifth conference on Applied natural language processing, pp.194–201. Association for Computational Linguistics (1997)
19. Masayuki, A., Matsumoto, Y.: Japanese Named Entity Extraction with Redundant Morphological Analysis, In: Proc. Human Language Technology conference – North American chapter of the Association for Computational Linguistics (2003)
20. Satoshi, S.: Nyu: Description of the Japanese NE System Used For Met-2. In: Proc. Message Understanding Conference (1998)
21. Carreras, X., Marquez, L., Padr'o, L.: Named entity extraction using adaboost. Proc. 6th Conf. Nat. Lang. Learn. 31:1–4 (2002)
22. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**, 2493–2537 (2011)
23. Kim, Y., Jernite, Y., Sontag, D., Rush, A.: "Character-aware neural language models. AAAI **30**(1), 2741–2749 (2016)
24. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics, (2019)
25. Huang, Z., Xu, W., Yu, K.: "idirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991, (2015)
26. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354, (2016)
27. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labelling. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics (2018)
28. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781, (2013). https://doi.org/10.48550/arXiv.1301.3781
29. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information, arXiv preprint arXiv:1607.04606, (2016). https://doi.org/10.48550/arXiv.1607.04606
30. Peters, M., et al.: Deep contextualized word representations, In: 6th International Conference on Learning Representations (2018)
31. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, (2017)
32. Wisam, Q., et al.: An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. pp. 200–204. (2019). https://doi.org/10.1109/IEC47844.2019.8950616
33. Nothman, J., et al.: Learning multilingual named entity recognition from Wikipedia, Artificial Intelligence 194 (2013).https://doi.org/10.1016/j.artint.2012.03.006
34. Yuan, Y., Jiang, Y., Tu, K.: "Bidirectional Transition-Based Dependency Parsing. Proc AAAI Conf Artif. Intell. **33**(01), 7434–7441 (2019). https://doi.org/10.1609/aaai.v33i01.33017434
35. Sang, E.F.T.K., Meulder, F.D.: Introduction to the CoNLL-2003 shared task: Language independent named entity recognition, In: Proceedings of the seventh conference on Natural language (2003)
36. Zhang, Q.P.Y., Bolton, Y.J., Manning, C.D.: Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, ArXiv, abs/2003.07082, (2020)

their performance is limited in the case of ORG entity. Nevertheless, the Stanza model performs slightly better than Flair. Future work will focus on depth analysis of the architecture of each model under study in order to propose a new model able to identify the 8 entity types with high prediction performance, low memory resource consumption, and high execution speed.

## References


1. Rau, L.F.: Extracting company names from text, In: Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application, pp. 29–32. (1991). https://doi.org/10.1109/CAIA.1991.120841
2. Grishman, R., Sundheim, B.: Message understandingconference-6: A brief history. In: COLING Volume 1: The 16th International Conference on Computational Linguistics, volume 1 (1996)
3. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes **30**, 3–26 (2007). https://doi.org/10.1075/li.30.1.03nad
4. Erik, F., Kim, T.S.: Introduction to the conll-2002 shared task: language-independent named entity recognition. Proc 6th Conf. Nat. Lang. Learn. **31**, 1–4 (2002)
5. Kim, T.S., et al.: Introduction to the CONLL-2003 Shared Task: Language-Independent Named Entity Recognition, In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL (2003)
6. Chen, H.H., Lee, J.C.: Identification and Classification of Proper Nouns in Chinese Texts". In: Proc International Conference on Computational Linguistics (1996)
7. Shihong, Y., Bai, S., Wu, P.: Description of the Kent Ridge Digital Labs System Used for MUC-7. In: Proc. Message Understanding Conference (1998)
8. Georgios, P., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D.: Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. In: Proc. Conference of Association for Computational Linguistics (2001)
9. Thierry, P.: The Multilingual Named Entity Recognition Framework. In: Proc. Conference on European chapter of the Association for Computational Linguistics, (2003)
10. Fei, H.: Multilingual Named Entity Extraction and Translation from Text and Speech. Carnegie Mellon University, Pittsburgh (2005). Ph. D. Thesis
11. Etaiwi, W., Awajan, A., Suleiman, D.: Statistical arabic name entity recognition approaches: A survey. Procedia Computer Science **113**, 57–64 (2017)
12. Satoshi, S., Nobata, C.: Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy, In: Proc. Conference on Language Resources and Evaluation (2004)
13. Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: Introduction to the bio-entity recognition task at jnlpba, In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications, pp. 70–75. Association for Computational Linguistics (2004)
14. Habibi, M., Weber, L., Neves, M., Wiegandt, D.L., Leser, U.: "Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics **33**(14), i37–i48 (2017)
15. Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: "Overview of biocreative: critical assessment of information extraction for biology. BMC Bioinformatics **6**, S1 (2005)
16. Burr, S.: Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets, In: Proc. Conference on Computational Linguistics. Joint Workshop on Natural Language Processing in Biomedicine and its Applications (2004)

**Table 14.** Non-exhaustive list of names of state agencies in Burkina Faso

| Organisation' names (ORG) |
|---|
| Commission interministérielle pour le processus d'apurement du passif du foncier urbain |
| Initiative des journalistes africains pour la coopération et le développement |
| syndicat national des administrateurs civils, des secrétaires et adjoints administratifs du Burkina Faso |

Apart from Camembert NER model, it seems that models that integrate Contextual word embeddings provide better results than those which used standard word embeddings like Spacy NER models (large and medium version). Table 15 describes models along with their size and overall performance on "Strict" axe. Stanza and Flair, the best in terms of performance display size higher than 1 Go. Concerning the Spacy models, we note that its medium version has a similar performance to the large one. In terms of ratio between performance and model size, we can say that the medium version is the best. As described in Section, one of the requirements of future model is that it must use less resources and therefore the model size must be small as possible. In our future work, we will deeply investigate on each architecture in order to choose the best one that fulfills our all requirements.

**Table 15.** Summary of the models according to their size and their overall performance in relation to the "Strict" axis.

| Models | Size (in Mo) | Strict |
|---|---|---|
| spacy_fr_core_news_lg | 15 | 0,59 |
| spacy_fr_core_news_md | 43 | 0,58 |
| spacy_fr_core_news_sm | 545 | 0,51 |
| Flair | 1300 | 0,65 |
| Stanza | 1120 | 0,69 |
| Camembert NER | 430 | 0,16 |

## 4 Conclusion

We have presented in this paper a comparative study of the performance of NER models on a dataset that we built from Burkina Faso online media. This study presents the first results of our ongoing work consisting in building a new NER model for the French language taking into account 8 types of entities, namely PER, LOC, ORG, MISC, PER-type, DATE, MONEY, EVENT. However, only 3 entities are evaluated in the present work (PER, LOC, ORG). The comparative study highlighted that Stanza and Flair models have the best performance in terms of the ability to identify PER and LOC entities. However,

**Table 12.** Performance evaluation results of NER models according to LOC entity type

| | Strict | Exact | Partial | Type |
|---|---|---|---|---|
| spacy_fr_core_news_lg | **0,70** | **0,72** | **0,70** | **0,81** |
| spacy_fr_core_news_md | **0,68** | **0,70** | **0,68** | **0,80** |
| spacy_fr_core_news_sm | 0,62 | 0,65 | 0,62 | 0,74 |
| flair_pred_ner | **0,70** | **0,72** | **0,70** | **0,82** |
| stanza_pred_ner | **0,70** | **0,74** | **0,70** | **0,81** |
| camembert_pred_ner | *0,06* | *0,08* | *0,46* | **0,81** |

**Table 13.** Performance evaluation results of NER models according to ORG entity type

| | Strict | Exact | Partial | Type |
|---|---|---|---|---|
| spacy_fr_core_news_lg | 0,41 | 0,53 | 0,41 | 0,54 |
| spacy_fr_core_news_md | 0,40 | 0,52 | 0,40 | 0,53 |
| spacy_fr_core_news_sm | 0,34 | 0,44 | 0,34 | 0,47 |
| flair_pred_ner | **0,51** | **0,56** | **0,51** | **0,66** |
| stanza_pred_ner | **0,61** | **0,67** | **0,61** | **0,70** |
| camembert_pred_ner | *0,29* | *0,32* | *0,57* | **0,73** |

To summarize, except for the Camembert NER models and the small version of Spacy, all models correctly identify people's names (type "PER") and places (type "LOC") from texts. The Stanza and Flair models perform the best, although Stanza is slightly better than Flair. The major problem remains in the ability to correctly identify the names of organizations (type "ORG") where performance is weakest compared to the previous types. We think that this is related to the names of the state agencies in Burkina Faso which are particularly long as can be seen in Table 14. In addition, the nomenclatures of these names do not always follow the standard norms where the first letters of compound names must begin with a capital letter. These different points explain why models were not able to reliably catch the full names of entities. For instance, when considering the entity "syndicat national des administrateurs civils, des secrétaires et adjoints administratifs du Burkina Faso", all models only identified the entity "Burkina Faso" as a location (LOC). This is a mistake since this entity stands for the well-known trade union organization of Burkina Faso.

Moreover, concerning the Camembert NER model, even if it globally performs the best on the identification of the types of entities (axis "Type"), however, its performances are surprisingly the worst. However, it is a model obtained by fine-tuning the Camembert model which has proven itself. Even if it is difficult to explain the exact reason, we think that the fine-tuning procedure was not sufficient.

similar results ($\geq 60\%$ on overall axes). These results corroborate those obtained by P. Qi et al. 2020 where the performances of the Stanza, Flair, and Spacy models were compared to their ability to correctly identify the types of entities (axis "Type"), in particular in the French language. On the other hand, the Camembert NER model, even if it presents the best in the identification of the types of entities ("Type" axe), it is clear that it strangely produces the worst results on the first 3 axes ("Strict", "Exact" and "Partial"). In other words, this model is not able to correctly identify the full names of the entities.

**Table 10.** Overall performance along with the 4 evaluation axes

| | Strict | Exact | Partial | Type |
|---|---|---|---|---|
| spacy_fr_core_news_lg | 0,59 | 0,65 | 0,73 | 0,70 |
| spacy_fr_core_news_md | 0,58 | 0,64 | 0,72 | 0,70 |
| spacy_fr_core_news_sm | 0,51 | 0,58 | 0,67 | 0,63 |
| flair_pred_ner | **0,65** | **0,68** | **0,76** | **0,77** |
| stanza_pred_ner | **0,69** | **0,74** | **0,79** | **0,79** |
| camembert_pred_ner | *0,16* | *0,17* | *0,51* | **0,80** |

Tables 11 to 13 highlight the performance of models along with each entity types PER, LOC, and ORG. Concerning the entity type PER, we note that the Stanza and Flair models outperform all other models with performance greater than or equal to 82% on the 4 axes even if the Camembert NER model has a performance equal to that of Stanza on the "Type" axe (cf. Table 11). As for entity type LOC, illustrated in Table 12, Stanza, Flair, the medium and large versions of Spacy present almost similar performances (above 70%) on the 4 axes. However, we have noticed that the performance of all models is lower for entity type ORG (cf. Table 13) compared the two first entity types. Nevertheless, the Stanza and Flair models present the best performances.

**Table 11.** Performance evaluation results of NER models according to PER entity type

| | Strict | Exact | Partial | Type |
|---|---|---|---|---|
| spacy_fr_core_news_lg | 0,73 | 0,76 | 0,73 | 0,82 |
| spacy_fr_core_news_md | 0,71 | 0,75 | 0,71 | 0,81 |
| spacy_fr_core_news_sm | 0,61 | 0,67 | 0,61 | 0,69 |
| flair_pred_ner | **0,82** | **0,84** | **0,82** | **0,89** |
| stanza_pred_ner | **0,85** | **0,87** | **0,85** | **0,92** |
| camembert_pred_ner | *0,08* | *0,08* | *0,51* | **0,92** |

- **Incorrect (INC):** the system prediction and the golden annotation don't match (Scenario 6);
- **Partial (PAR):** the predicted entity name is partially correct (Scenario 4 &5);
- **Missing (MIS):** the golden annotation is not captured by a system (Scenario 3);
- **Spurius (SPU):** system produces a response that doesn't exist in the golden annotation (Scenario 2);

From these 5 metrics, Precision and Recall are computed for each axe as follows:

- **Exact match** (i.e., strict and exact axes):

$$Precision = \frac{COR}{COR + INC + PAR + SPU} \tag{2}$$

$$Recall = \frac{COR}{COR + INC + PAR + MIS} \tag{3}$$

- **Partial match** (i.e., partial and type):

$$Precision = \frac{COR + 0.5 \times PAR}{COR + INC + PAR + SPU} \tag{4}$$

$$Recall = \frac{COR + 0.5 \times PAR}{COR + INC + PAR + MIS} \tag{5}$$

In our study, we adopted this last approach since it cover all scenario we described above and also integrate the exact match metrics.

### 3.3 Results and Discussion

In order to qualitatively assess NER model's performance, we have downloaded pre-trained models as python packages. It is important to mention that no models have been fine-tuned. We have pre-trained models "as is". In addition, to assess NER models' performance, we've used a python package named "nervaluate" (https://pypi.org/project/nervaluate/) which contains all metrics described above.

Besides, let's remind that the models under study provide 4 entity types PER, LOC, ORG, and MISC. To reliably evaluate their performance, we consider the three first entity types PER, LOC, and ORG. The type MISC has not been taken into account to avoid confusion. Indeed, a given NER model can consider an EVENT as MISC yet EVENT is out of our evaluation.

Table 10 presents the overall performance of different models along with the 4 axes where the F1-score was used. On all axes, the Stanza and Flair models perform the best. They correctly identify the types of entities (79% and 77% respectively) and relatively well the full names of entities. However, their performance in exactly identifying both the full name and type (axe "Strict") is relatively fair (69% and 65% respectively). Nevertheless, compared to Flair, Stanza is the one with the highest score on the 4 axes. As for the Spacy models, they present intermediate performances compared to the Stanza and Flair models. However, we can see that its medium and large versions provide almost

**Table 8.** Example of scenario 5

| Golden standard | | NER prediction model | |
|---|---|---|---|
| Entity text | Entity type | Entity text | Entity type |
| Lamoussa | PER | Lamoussa | PER |
| habite | -- | habite | -- |
| à | -- | à Ouagadougou | LOC |
| Ouagadougou | LOC | | |

**Table 9.** Example of scenario 6

| Golden standard | | NER prediction model | |
|---|---|---|---|
| Entity text | Entity type | Entity text | Entity type |
| Lamoussa | PER | Lamoussa | ORG |
| habite | -- | habite | -- |
| à | -- | à Ouagadougou | LOC |
| Ouagadougou | LOC | | |

- 6. NER system gets the boundaries and entity type wrong.

References [4, 5] proposed the exact match metrics. Which considers that a prediction is correct only when the predicted label for the entity's full name is matched to exactly the same words as the gold label of that entity. So, we think that this approach only considers the first three scenarios which are viewed as standard classification problems. [2] proposed an evaluation based on 2 axes:

– the ability of the systems to catch the correct entity type. In other words, a correct entity type is validated if an entity is assigned the correct type, regardless of its text boundaries
– the reliability to find the exact entity's name regardless of its entity type.

More recently, [42] proposed an evaluation that is viewed as an extension of the [2] approach. They proposed four axes of evaluation:

– **Strict**: entity full name and type match exactly
– **Exact:** exact boundary match over the entity's full name, regardless of the type;
– **Partial**: partial boundary match over the entity's full name, regardless of the type;
– **Type**: entity type exactly match, regardless of the entity's full name;

Then for each axe, Precision, Recall and F1 score are computed based on MUC-5 evaluation metrics [43]. In fact, MUC-5 introduced 5 metrics to tackle all errors that occur in the different scenario describe above:

– **Correct (COR):** Both entity name and type are the same (Scenario 1);

**Table 5.** Example of scenario 2. NER model identifies the verb "habite" as a location (LOC)

| Golden standard | | NER prediction model | |
|---|---|---|---|
| Entity text | Entity type | Entity text | Entity type |
| Lamoussa | PER | Lamoussa | PER |
| habite | -- | habite | LOC |
| à | -- | à | -- |
| Ouagadougou | LOC | Ouagadougou | LOC |

- 3. NER system misses an entity

**Table 6.** Example of scenario 3. NER model omits "Ouagadougou" as entity

| Golden standard | | NER prediction model | |
|---|---|---|---|
| Entity text | Entity type | Entity text | Entity type |
| Lamoussa | PER | Lamoussa | PER |
| habite | -- | habite | -- |
| à | -- | à | -- |
| Ouagadougou | LOC | Ouagadougou | -- |

- 4. NER system assigns wrong entity type (Table 7)

**Table 7.** Example of scenario 4. NER model identify "Ouagadougou" as entity but the type is wrong

| Golden standard | | NER prediction model | |
|---|---|---|---|
| Entity text | Entity type | Entity text | Entity type |
| Lamoussa | PER | Lamoussa | PER |
| habite | -- | habite | -- |
| à | -- | à | -- |
| Ouagadougou | LOC | Ouagadougou | ORG |

- 5. NER model gets the boundaries of the entity's name wrong. In our example (see Table 8), NER model identifies "à Ouagadougou" as entity whereas in the golden annotation, the right entity is "Ouagadougou".

**Table 3.** Example of labelled text provided by Doccano

| N° | Full name | Entity-type | Start position | End position |
|---|---|---|---|---|
| 1 | international burkinabè | PER-TYPE | 2 | 25 |
| 2 | Charles Kaboré | PER | 26 | 40 |
| 3 | Angleterre | LOC | 66 | 76 |
| 4 | mercato | EVENT | 94 | 106 |
| 5 | décembre-janvier | DATE | 103 | 119 |
| 6 | Crystal Palace | ORG | 177 | 191 |
| 7 | West Bromwich Albion | ORG | 195 | 215 |
| 8 | Burkinabè | PER-TYPE | 125 | 135 |

L'international burkinabè Charles Kaboré pourrait se retrouver en Angleterre lors du prochain mercato (décembre-janvier). Le Burkinabè est pisté par des clubs anglais notamment Crystal Palace et West Bromwich Albion.

But, before exposing these methods, it is important to understand the different scenarios involved during the evaluation. Let's take the following example: "Lamoussa habite à Ouagadougou" (Table 5).

So, when comparing the golden annotations with the output of a NER system, 6 different scenarios might occur (Table 6):

- 1. Entity full name and type match

**Table 4.** Example of scenario 1

| Golden annotations | | NER prediction model | |
|---|---|---|---|
| Entity text | Entity type | Entity text | Entity type |
| Lamoussa | PER | Lamoussa | PER |
| habite | -- | habite | -- |
| à | -- | à | -- |
| Ouagadougou | LOC | Ouagadougou | LOC |

- 2. NER system hypothesized an entity (i.e., false alarm is detected).

more easily the functions of a person thanks to the type "PER-TYPE". Table 2 details all the types of entities considered in our study.

Note that most of these entities are included in other languages like English as is the case in the Spacy NER model.

To achieve the objective of our study, our first step is to evaluate the performance of existing models on our dataset in order to choose the best NER architecture. Of course, this evaluation will be done only on the basis of the 4 entity types they provide.

In this section, we will detail the dataset that we produced from these 8 types of entities. Then, we will describe the performance evaluation methods used to assess the models under study.

### 3.1 Dataset Description

The data was randomly extracted between 2017 and 2022 from the most well-known online media of Burkina Faso, namely "Lefaso.net" [38], "Burkina 24" [39], and "Netafrique" [40]. While some media such as Lefaso.net provides RSS (Rich Site Summary) feeds in order to easily extract articles, for other sources, articles have been manually extracted. For each article, we have extracted its title, summary, and content. In all, our dataset is composed of 2860 texts. In order to be able to evaluate the different models, we need previously labeled data. For this, we used 16 people well impregnated in the news in Burkina. Among them, 12 have labeled the data while the other 4 people are the validators of the work carried out. This ensures the reliability of the work. As a working tool, we used the recent version of Doccano [41]. In fact, Doccano is an open-source text annotation tool for humans that provides annotation features for text classification, sequence labeling, and sequence-to-sequence tasks. Then, for each text, the annotator must highlight entities' full names and types according to the 8 types described above. Finally, the resulting dataset is known as "golden annotations". Table 3 shows an example of labeled text. For each identified entity, its full name, entity type, and its position in the text are provided.

### 3.2 Presentation of NER Evaluation Methods

Several criteria have been proposed to assess NER systems performance. The most typical way to evaluate NER systems is to use standard metrics precision, recall, and f1 score. In fact, the precision metric is the percentage of correct named entities found by the NER system. Concerning recall, it stands for the percentage of the named entities in the golden annotations that are retrieved by the NER system. Finally, the F1 score is defined as follows (Table 4):

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{1}$$

### 2.4 Camembert NER

Camembert-Ner is a NER model that was fine-tuned from CamemBERT [37] and was trained on the Wikiner dataset. Note that the CamemBERT is a French language model trained over 130 GB unlabeled dataset. It is based on advanced Context word embeddings as described in [24] Unfortunately, we did not find any paper about "Camembert-Ner". But the model is available on the HuggingFace website.

## 3 NER Models Evaluation

**Table 2.** List of entity types considered in our study

| Entity types | Description | Examples |
|---|---|---|
| PER | People name | "Salif Diallo", "Dr Sanfo" |
| LOC | Natural and human-made landmarks, structures, geopolitical features, … | "centre Muraz", "Boucle du mouhoun", |
| ORG | Companies, political groups, sport clubs, government bodies, public organization, … | "Coris Bank International", "Ministère de l'agriculture", … |
| MISC | Miscellaneous entity | "Coronavirus", "Paludisme", |
| PER-Type | Job types or roles held by a person | "ministre de la fonction publique", "acteur burkinabé" |
| DATE | Absolute or relative date | "15 Juillet 2020", "Mai 2011", |
| MONEY | Monetary values, including unit | " 1 millions de Francs CFA", "52 000 FCFA", |
| EVENT | historical, social and naturally occurring events | "FESPACO 2021", "Coupe du monde 2022" |

As mentioned in the introduction, our goal is to provide a user-friendly NER model for detecting entities from Burkina Faso media information. The model must be able to use fewer resources and be fast in terms of execution time. In addition, in the literature, in most of the models proposed for the French language, only the entity types of PER, ORG and LOC, and MISC are detected. This does not allow qualitative analysis of the entities. For this, our goal is to propose a model allowing us to detect in addition to the 4 entities proposed in the literature, 4 other entities "PER-TYPE" for person job types or roles, "MONEY", "DATE" and "EVENT". We think it is important to take them into account in our future model because of the additional information they provide, especially the "PER-TYPE" and "EVENT" types. Most of the time, this information is masked in the existing models (considered as "MISC"). Let's take the case of the "PER-TYPE" type. It will allow us to distinguish the proper name of a person from his function or his rank, the latter being a variable data in time. Thus, on a given dataset, it is possible to determine

it supports more than 66 languages and 76 pretrained pipelines for 23 languages. In particular, for the French language, it provides mainly 3 language models that differ not only in their size (model's name is suffixed by "sm", "md" and "lg" for small, medium, and large versions respectively) but also from the size of words vector as described in Table 1. All these models include a NER model.

**Table 1.** List of French models provided in Spacy

| French model | Size (MB) | Vector |
|---|---|---|
| fr_core_news_sm | 15 | 0 unique vectors |
| fr_core_news_md | 43 | 20k unique vectors (300 dimensions) |
| fr_core_news_lg | 545 | 500k unique vectors (300 dimensions) |

As we can see in Table 1, medium and large versions use pre-trained word embeddings. Concerning the small, word encoding stage is performed by using a slight version of Bag of Word fully detailed in [32]. Concerning the NER model, it has been pre-trained from Wikiner data [33]. It works as follows: once the text has been tokenized and then transformed into vector representation using the pre-trained word embeddings. Finally, a neural network named Bidirectional Transition-Based Dependency Parsing [34], a variant of Bi-LSTM is used to predict entities.

### 2.2 Flair

Ref [27] proposed FLAIR which stands for an open-source NLP framework that provides about twenty pre-trained models for NER, post-tagging, and text classification tasks. Its particularity is built on "contextual word embeddings" described above. The NER model has been pre-trained on the Conll2003 dataset [35] and is made up of mainly two sub-models: To perform NER tasks, each text is passed as a sequence of characters to a bidirectional character-level neural language model in order to create contextual embeddings representation. The resulting vectors are used as the input of the BiLSTM-CRF stage to make the prediction.

### 2.3 Stanza

Stanza [36] is an open-source python NLP toolkit that supports 66 languages including French. Like Flair, Stanza NER component used "contextual word embeddings". It was firstly training a forward and a backward character level LSTM language model, and at tagging time we concatenate the representations at the end of each word position from both language models with word embeddings, and feed the result into a standard one-layer Bi-LSTM sequence tagger with a conditional random field (CRF)-based decoder. In a real-time application, it can be run on CPUs or GPUs.

MISC. In addition, as Spacy provides 3 variants of NER models, we have considered 6 models. Besides, some platforms like Google Natural Language API, Microsoft Azure Cognitive Services provide service for NER identification. But they are out of our study since they are proprietary models and to our knowledge, no paper talks about their model architecture.

For the remainder of this paper, we present an overview of NER models considered in this study in Sect. 2. Then, we describe datasets and evaluation methods used to assess NER models in Sect. 3. Finally, we discuss the models' performance in Sect. 4.

## 2 Overview of Ner Model's

Most of the time, NLP tasks consider NER as sequence labeling problems. Text is treated as a sequence of words to be labeled with linguistic tags. Generally, the proposed architecture is made up of 3 stages:

- **Pre-processing**: it consists in tokenizing text as a list of words (also called tokens), lemmatization (assigning the base forms of words. For example, the lemma of "was" is "be", and the lemma of "rats" is "rat"), post-tagging (assigning word types to tokens, like verb or name) etc.;
- **Word encoding**: in this stage, each token is converted into n-dimensional vector;
- **Prediction modeling**: at this last stage, the encoded text is used as input of a neural network architecture (CNN or Bi-LSTM, and a subsequent conditional random field (CRF) decoding layer [25, 26] in order to make the final prediction.

It is important to note that the word encoding stage stands for the core of NER models. The well-known approach used is word embeddings which is a kind of word representation that allows words with similar meaning to have a similar representation. Three types of word embeddings have been proposed as mentioned in [27]:

- **Standard word embeddings** like Word2Vec [28] and FastText [29] which are pre-trained over a very large collection of unlabeled data and provide a static vector for each token;
- **Character-level features** proposed in Lample et al., 2016, which are not pre-trained, but trained on task data to capture task-specific subword features.
- **Contextualized word embeddings** [27, 30] that catch word semantics in his context. In other words, these models produce different vectors for the same word depending on its contextual usage.

If the preprocessing stage can be more or less similar for all NER models, they mainly differ from the word representation approach. We present an overview of each model considered in our study.

### 2.1 Spacy

Spacy is an open-source framework for advanced Natural Language Processing (NLP) written in Python and Cython proposed by [31]. It proposes a pipeline made up of multiple components like POS tagging, Lemmatization, text classification, NER, etc. In addition,

question answering, summarization, automatic translation, etc. One of the first works in this field was presented by [1]. Their work consisted in extracting and identifying company names from text based on heuristics and handcrafted rules. The well-known Sixth Message Understanding Conference (MUC-6) [2] has played an important role in this field. Indeed, it was in this meeting that the term "Name Entity" was defined for the first time which consisted in identifying all people's names, organization, and geographic location from the text. These tasks were known as "Named Entity Recognition and Classification (NERC)". A survey of different research on NERC has been detailed in [3]. Furthermore, most of the studies in the NERC domain were initially carried out in the English language [4], but have been extended to other specific languages like German [5], Chinese [6, 7], French [8, 9] and Arabic [10, 11]. These studies have mainly been focused on extracting 4 kinds of entities: PER (Proper noun of Person, e.g., Blaise Compaoré), LOC (Location, e.g., Burkina Faso), ORG (organization, e.g. Coris Bank International) and MISC (miscellaneous that include all other types of entities, e.g. Covid19). But, according to the objective of the study, a specific entity type can be defined. [12] proposed 200 types of entity which includes many fine subcategories like religion, animal, product, event, etc. In addition, many researchers have proposed to extract an entity's name in specific domains such as biomedical [13–16].

In order to automatically identify entities, several NER systems have been proposed. These systems can be split into 3 groups. The first one known as "knowledge-based systems" were based on lexical resources and domain-specific knowledge. In other words, it consists in listing all entities pertained to a specific and then they are automatically identified using a simple algorithm. However, as mentioned in [17], these approaches work well when the lexicon is exhaustive. Otherwise, the system remains unstable and unmaintainable mainly when new entities have to be taken into account. The second group uses a machine learning approach which builds a model based on training data, in order to make predictions and then replace human curated rules. Techniques such as Hidden Markov Models (HMM) [18], Support Vector Machines (SVM) [19] and Decision Trees algorithms [20] and AdaBoost classifiers [21] were commonly used to build NER systems. The last one also called "Feature-inferring neural network systems" use advanced machine learning techniques known as neural network algorithms. These methods build a model that simulates human brain neurons. Techniques like Recurrent Neural Networks (RNN), including bidirectional Long Short-Term Memory (bi-LSTM) [22] Convolution Neural Network (CNN) [23] and Bidirectional Encoder Representations from Transformers (BERT) [24] have been considered. In terms of performance, [17] show that Neural Network NER systems outperform the other approaches.

Our study consists in building a NER system that will automatically identify an entity's name from Burkina Faso news. In this context, our first step was obviously to evaluate existing NER models in the context of Burkina Faso. This paper presents the first result of this work. Since the most spoken language.

In Burkina Faso is French, then we only consider NER models that include the French language. Thus, 4 open sources framework that include French NER models have been taken into account namely Spacy, Flair, Stanza, and Camembert. Note that the particularity of these frameworks is that they are multilanguage systems based on Neural Network algorithms and identify 4 entity types namely PER, LOC, ORG and

# Comparative Study of Name Entity Recognition Models in Burkina Faso Context

Sibiri Tiemounou[1(✉)], Wend Yam Serge Boris Ouédraogo[1,2], Moumouni Djibo[2], Yaya Traoré[3], Ali Maïga[4], Souleymane Zio[1], and François Zougmoré[2]

[1] Institut du Génie Informatique et Télécommunications, École Polytechnique de Ouagadougou, Ouagadougou, Burkina Faso
sibiri.tiemounou@gmail.com

[2] Laboratoire Matériaux et Environnement, Université Joseph KI-ZERBO, Ouagadougou, Burkina Faso

[3] Laboratoire de Mathématiques et Informatique, Université Joseph KI-ZERBO, Ouagadougou, Burkina Faso

[4] FDI MATELEC, Paris, France

**Abstract.** Name Entity Recognition (NER) is an important core component of Natural Language Processing (NLP) systems for identifying entities like person names, locations, and organizations. Many NER models have been proposed in the literature those whose architecture are based on deep neural networks are the most efficient. Burkina Faso, a West African country, is a French-speaking country with its culture and its specificities in the use of the French language. Our work consists in building a user-friendly NER model that reliably identifies 8 entity types from Burkina Faso media information where the French language is the most spoken. To achieve this goal, we firstly build a dataset that has been labeled along these entity types. Then, in order to choose the best architecture, we assessed 6 multilanguage NER models namely Spacy (with its three pre-trained models), Flair, Stanza, and Camembert NER. This paper presents the performance evaluation of existing French NER models when applying to media news data of Burkina Faso. They have been assessed along 3 common entity types Person (PER), LOCation (LOC), and ORGanization (ORG). Results show that Stanza and Flair outperform all models under study with a percentage greater than 70%. They reliably identify a person's name and location entities. However, their performance is relatively fair in correctly extracting organization entities due to the Burkina Faso context.

**Keywords:** Name Entity Recognition · Natural language processing · Neural Network · NER models

## 1 Introduction

Name Entity Recognition (NER) is a subfield of Natural Language Processing (NLP) that consists in identifying named entities like person, location, organization, etc. based on text analysis. It stands for the core components of many NLP applications such as


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 123–138, 2025.
https://doi.org/10.1007/978-3-031-81573-7_10

14. Olenik, S., Lee, H.S., Güder, F.: The future of near-field communication-based wireless sensing. Nat. Rev. Mater. **6**, 286–288 (2021)
15. Hossain, M.I., Markendahl, J.I.: Comparison of LPWAN Technologies: Cost Structure and Scalability. Wireless Pers. Commun. **121**, 887–903 (2021)
16. Dangi, R., Lalwani, P., Choudhary, G., You, I., Pau, G.: Study and Investigation on 5G Technology: A Systematic Review. Sensors **22**(1), 26 (2022)
17. Akasiadis, C., Pitsilis, V., Spyropoulos, C.D.: A Multi-Protocol IoT Platform Based on Open-Source Frameworks. Sensors **19**(19), 4217 (2019)
18. Tariq, M.A., Khan, M., Raza Khan, M.T., Kim, D.: Enhancements and Challenges in CoAP—A Survey. Sensors **20**(21), 6391 (2020)
19. Biswajeeban, M., Attila, K.: The Use of MQTT in M2M and IoT Systems: A Survey. IEEE Access **8**, 201071–201086 (2020)
20. Ioana, A., Korodi, A.: DDS and OPC UA Protocol Coexistence Solution in Real-Time and Industry 4.0 Context Using Non-Ideal Infrastructure. Sensors **21**(22), 7760 (2021)
21. Li, F., Wang, C.: Artificial intelligence and edge computing for teaching quality evaluation based on 5G-enabled wireless communication technology. J. Cloud Comput. **12**(1), 45 (2023)
22. Merenda, M., Porcaro, C., Iero, D.: Edge Machine Learning for AI-Enabled IoT Devices: A Review. Sensors **20**(9), 2533 (2020)
23. Xu, Z., Liu, W., Huang, J., Yang, C., Lu, J., Tan, H.: Artificial intelligence for securing IoT services in edge computing: a survey. Security and communication networks **2020**, 1–13 (2020)
24. Adam, B., Kaveh, M.: The rise of artificial intelligence in healthcare applications. In: Adam, B., Kaveh, M. (eds.) Artificial Intelligence in Healthcare, pp 25–60. Academic Press, Cambridge (2020)
25. Bajwa, J., Munir, U., Nori, A., Williams, B.: Artificial intelligence in healthcare: transforming the practice of medicine. Future healthcare journal **8**(2), 188–194 (2021)
26. Kamruzzaman, M.M., Alrashdi, I., Alqazzaz, A.: New opportunities, challenges, and applications of Edge-AI for connected healthcare in Internet of Medical Things for smart cities. J. Healthcare Eng. **2022**(1), 2950699 (2022)
27. Motti, V.G., Berkovsky, S.: Healthcare Privacy. In: Knijnenburg, B.P., Page, X., Wisniewski, P., Lipford, H.R., Proferes, N., Romano, J. (eds.) Modern Socio-Technical Perspectives on Privacy, pp 203–231. Springer, Cham (2022)
28. Wang, C., Zhang, J., Lassi, N., Zhang, X.: Privacy protection in using artificial intelligence for healthcare: Chinese regulation in comparative perspective. In: Healthcare, vol. 10, no. 10, p. 1878. MDPI (2022)
29. Monteiro, A., Santos, S., Gonçalves, P.: Precision agriculture for crop and livestock farming—brief review. Animals **11**(8), 2345 (2021)
30. Debauche, O., Mahmoudi, S., Elmoulat, M., Mahmoudi, S.A., Manneback, P., Le-beau, F.: Edge AI-IoT pivot irrigation, plant diseases, and pests identification. Procedia Computer Science **177**, 40–48 (2020)
31. Neethirajan, S.: The role of sensors, big data and machine learning in modern animal farming. Sensing and Bio-Sensing Research **29**, 100367 (2020)
32. Sun, L., Jiang, X., Ren, H., Guo, Y.: Edge-cloud computing and artificial intelligence in internet of medical things: architecture, technology and application. IEEE Access **8**, 101079–101092 (2020)

## 6 Conclusion

In the ever-evolving landscape of technology, the fusion of Edge AI and Internet of Things (IoT) has emerged as a transformative force, revolutionizing the way intelligent systems are designed, deployed, and experienced. Throughout this research paper, we have delved into the intricate realms of these cutting-edge technologies, exploring their architectures, applications, and future perspectives.

In conclusion, the synergy between Edge AI and IoT is reshaping the landscape of intelligent systems, bringing us closer to a future where machines are not just smart but also contextually aware and responsive. Embracing the full potential of these technologies demands continuous research, innovation, and collaboration.

The fusion of Edge AI and IoT has given rise to sophisticated architectures that bridge the gap between the physical and digital worlds. Edge computing, with its decentralized approach, enables real-time data processing and analysis at the edge devices, reducing latency and enhancing efficiency. Incorporating Edge AI technologies into agriculture holds the promise of transforming the industry into a more sustainable, efficient, and resilient sector.

## References

1. Rahmani, A.M., Bayramov, S., Kiani Kalejahi, B.: Internet of things applications: opportunities and threats. Wireless Pers. Commun. **122**(1), 451–476 (2022)
2. Pérez, J., Díaz, J., Berrocal, J., López-Viana, R., González-Prieto, Á.: Edge computing: A grounded theory study. Computing **104**(12), 2711–2747 (2022)
3. Sheikh, H., Prins, C., Schrijvers, E.: Artificial Intelligence: Definition and Background. In: Sheikh, H., Prins, C., Schrijvers, E. (eds.) Mission AI: The New System Technology, pp. 15–41. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-21448-6_2
4. Singh, R., Gill, S.S.: Edge AI: a survey. Internet of Things Cyber-Phys. Syst. **3**, 71–92 (2023)
5. Yu, W., et al.: A survey on the edge computing for the Internet of Things. IEEE Access **6**, 6900–6919 (2017)
6. Kang, K.D.: A Review of Efficient Real-Time Decision Making in the Internet of Things. Technologies **10**(1), 12 (2022)
7. Hamdan, S., Ayyash, M., Almajali, S.: Edge-computing architectures for Internet of Things applications: a survey. Sensors **20**(22), 6441 (2020)
8. Lina, L., Ruisheng, S., Bai, W., Lei, Z., Ning, J.: A Universal Complex Event Processing Mechanism Based on Edge Computing for Internet of Things Real-Time Monitoring. IEEE Access **7**, 101865–101878 (2019)
9. IEEE Innovation at work: https://innovationatwork.ieee.org/real-life-edge-computing-use-cases/, last accessed 2023/10/13
10. Filali, A., Abouaomar, A., Cherkaoui, S., Kobbane, A., Guizani, M.: Multi-access-edge computing: a survey. IEEE Access **8**, 197017–197046 (2020)
11. Haque, K.F., Ahmed, A., Kumar, Y.: Comprehensive Performance Analysis of Zigbee Communication: An Experimental Approach with XBee S2C Module. Sensors **22**(9), 3245 (2022)
12. Chendong, L., Yilin, Z., Huanyu, Z.: A Comprehensive Study of Bluetooth Low Energy. J. Phys: Conf. Ser. **2093**(1), 012021 (2021)
13. Pahlavan, K., Krishnamurthy, P.: Evolution and Impact of Wi-Fi Technology and Applications: A Historical Perspective. Int. J. Wireless Inf. Networks **28**, 3–19 (2021)

Overall, Edge-IA and IoT will revolutionize the agricultural sector. By enabling real-time agricultural data processing and decision-making at the source of collection, Edge-IA and IoT will help optimize crop yields, improve livestock management and promote sustainable farming practices. With a growing global population to feed, Edge-IA and IoT will be key to meeting agricultural challenges and ensuring a sustainable future for all.

## 5 Challenges and Future Direction

Edge computing is made up of peer-to-peer systems, wireless networks, etc. and the adoption of a complete system is required to manage the system as a whole. The security and confidentiality of a distributed system is a real challenge.

During data processing at the edge, user data can be processed, putting their sensitive information at risk. Authentication of gateway nodes is one of the security issues in edge computing. Each edge node is managed differently. It is therefore difficult to implement a common security method everywhere.

Another challenge is the security of data sharing and transmission processes. Multiple devices need to work together to perform many tasks, and data needs to be shared securely. Data storage is provided by various third-party suppliers, and their storage devices spread over several sites also increase the risk of attack, for two reasons. Firstly, the data is separated into several slices and stored in different places. It is therefore easy to store incorrect data or to lose it. This makes it difficult to guarantee data integrity. Secondly, data downloaded to storage can be modified by malicious users, leading to confidentiality problems and data leaks.

It would be preferable to use local pre-processing, which could mask private information and reduce the amount of data transmitted.

In some cases, encryption could be used to protect privacy. However, before learning or inference tasks can be performed, the encrypted data must be decrypted. This requires an increase in the amount of essential computation. To meet this challenge, the authors [32] propose the use of homomorphic encryption, which allows the training or inference task to be performed directly on encrypted data.

Looking forward, the future of Edge AI and IoT appears incredibly promising. The rapid advancements in hardware, such as powerful yet energy-efficient microprocessors, are driving the development of more complex AI models that can be deployed at the edge. Additionally, the proliferation of 5G networks will significantly enhance the capabilities of Edge AI and IoT, enabling seamless connectivity and faster data transmission.

Moving on the future, it is crucial to address the challenges associated with these technologies, such as security vulnerabilities, data privacy concerns, and ethical considerations. Collaboration between academia, industry, and policymakers will be vital in creating robust frameworks and standards that ensure the responsible development and deployment of Edge AI and IoT solutions.

As the amount of data generated by healthcare facilities increases, so does the risk of data breaches and unauthorized access. Edge AI is emerging as a potential solution to this problem, offering the potential to guarantee data confidentiality and security [26].

Edge intelligence is ideal for healthcare facilities because it enables data to be analyzed without leaving the facility's secure network. With data maintained on site, the risk of data breaches and unauthorized access can be reduced. In addition, it enables healthcare organizations to ensure that they comply with the requirements of various privacy laws, such as HIPAA (Health Insurance Portability and Accountability Act) [27]. Advanced AI can be beneficial in identifying and reporting privacy breaches [28]. This allows healthcare facilities to take steps or actions before a breach occurs.

The potential for advanced intelligence to improve data security in the healthcare sector is clear. As the healthcare sector continues to evolve, advanced AI can become an indispensable tool for protecting patient data.

### 4.3 Edge-AI and IoT for Intelligent Agriculture

Edge-AI and IoT-based smart farming refers to the use of advanced AI and the Internet of Things (IoT) to improve efficiency and productivity in the agricultural industry.

Today, the use of AI techniques is playing an important role in the development of precision agriculture [29]. Predictive crop models can be designed. They take into account different production factors such as: climate; soil condition; seed condition; water supply; disease risk, etc. By taking all these factors into account when designing the model, it is easier to adjust any of the elements to optimize yield.

In irrigation management [30], for example, IoT can be used to monitor and control crop irrigation in real time. Sensors could be placed in fields to measure soil moisture, temperature and other key factors, helping to determine the optimum time for irrigation and avoid wasting water. Advanced computing combined with artificial intelligence will enable the data collected by the sensors to be processed and analyzed locally in real time. Personalized irrigation strategies can then be recommended for each plot based on soil and crop characteristics. This enables efficient and sustainable farming practices to be put in place.

Another benefit of advanced AI in agriculture is its ability to improve animal welfare [31]. Using wearable sensors, farmers will be able to monitor the health and well-being of their livestock in real time. Edge-IA devices will analyze the data from these sensors to detect early signs of disease. This will allow farmers to intervene before problems arise or worsen. This will improve animal welfare, reduce the use of antibiotics and increase productivity.

The ability to optimize crop yields and reduce waste is another key benefit of using advanced AI in agriculture [29]. Indeed, the analysis of data from soil sensors, weather stations and satellite images by Edge-IA devices will be able to provide farmers with real-time information about the health and needs of their crops. This information can then be used to make informed decisions about irrigation, fertilization and pest control, so that crops receive the precise amount of resources they need to thrive. This can lead to increased crop yields, reduced resource consumption and minimized environmental impact.

- **The Intelligent-Edge layer**: this layer lies between the IoT device layer and the cloud layer. It consists of edge-intelligent devices for the collection, storage, artificial intelligence-based processing and real-time analysis of data from intelligent devices and its transmission to the cloud using wireless transmission technologies such as 5G, Sigfox, LoRa-WAN, NB-IoT [21, 22] etc. This layer also plays a role in securing data transmission. Indeed, features such as tamper detection, encryption and others can be implemented at the edge-intelligent device level, preventing malicious attacks against IoT devices and securing user data to the cloud [23].
- **The Cloud layer**: this layer comprises servers dedicated to storage and data processing. It therefore hosts the data from the edge-intelligent layer, for in-depth processing. The Cloud layer gives IoT application programmers the means to work with heterogeneous objects without having to rely on a specific hardware platform. It is also responsible for service management (Fig. 2).

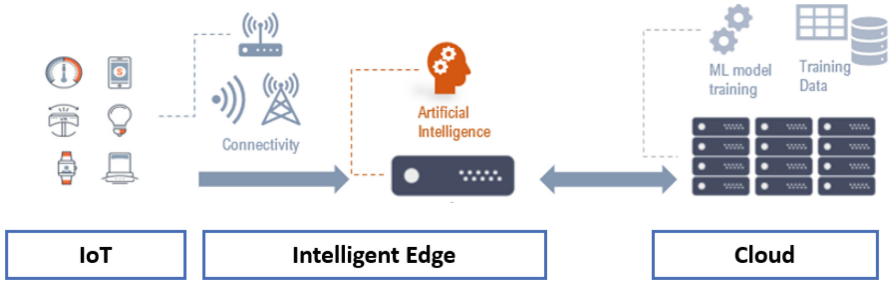



**Fig. 2.** Intelligent edge architecture.

### 4.2 Edge-AI and IoT for Connected Health

When healthcare professionals are faced with life-and-death decisions, the quality of the information available to them is critical. More accurate information and real-time access mean better decisions. Thanks to the power of the Internet of Things (IoT) and advanced AI, connected healthcare systems will be able to deliver real-time data to providers and enable more efficient care delivery.

Edge intelligence allows data to be processed and analyzed in real time without having to send it to a central server such as the cloud [4]. This can be very useful in healthcare facilities, where data needs to be analyzed quickly and securely.

One of the benefits of advanced AI in medical data is its ability to detect anomalies [24]. This can help healthcare staff to quickly identify potential health problems, take appropriate action and make more informed decisions about patient care.

The use of advanced AI can also improve the efficiency of healthcare systems. Real-time analysis of data can enable healthcare providers to reduce waiting times and improve patient outcomes [25]. Edge intelligence can also be used to automate the scheduling of appointments or the management of medical records.

# 4 Advanced Edge-AI for Convergent Services and Applications

## 4.1 Edge-AI Concepts and Opportunities

**Concepts:** intelligent Edge refers to the combination of Edge Computing and Artificial Intelligence techniques. In simple terms, the Intelligent Edge will enable AI-based processing algorithms to be run at the heart of the devices that are part of the Edge Computing network, using a reasonable internet connection. Everything happens in the device: collection, storage and AI-based processing. This results in extremely short response times of a few milliseconds. The information arrives in real time with unparalleled accuracy, since the algorithms will have refined the raw data. Even today, the majority of heavy computing is done in the cloud and requires large computing capacities. The Intelligent Edge will move some of the computational processing flow directly into the devices, to reduce the use of the cloud significantly. With this type of device, all data can be stored before being sent to a remote site for further analysis, if required. These intelligent devices, thanks to the availability of sensor data in the field, can interact autonomously without interaction with the central site.

**Opportunities.** Edge AI has huge potential. Here is a non-exhaustive list of the advantages that can be derived from this technology:

- Ease of management: Edge-AI enabled devices are easy to use because the objects are completely autonomous.
- Reduced latency: The AI will not be subject to congestion problems because it is as close as possible to the objects.
- Deployment "in the field": Edge-AI lends itself by definition very well to IoT uses, but also in all mobile devices, starting with autonomous cars.
- Reduced costs: Bandwidth and data exchange costs are reduced because the amount of data transmitted is minimal.
- Security and confidentiality: Edge-AI makes it possible to filter the data to be uploaded to the cloud if necessary. Only the desired information is transmitted, after being anonymized for example. Data confidentiality and the protection of privacy are two major issues to which Edge-AI can provide a solution, since this approach completely reverses the paradigm compared with AI in the Cloud, which requires systematic transmission of all the data before it is analyzed by the AI.

The following Fig. 2 shows an example of a three-layer intelligent systems architecture assisted by IoT and Edge-IA. The layers of the architecture are identical to those of the Edge computing architecture with the exception of the edge-intelligent layer where we will find the artificial intelligence algorithms. Here, IoT devices represent the end users of the architecture.

- **The IoT layer**: this is the first layer and includes sensors, RFID devices and actuators for collecting information such as temperature, humidity, location, air quality and movement. Once the information has been collected, it is transmitted to the Edge layer via a secure wireless connection. This layer requires the use of standardized plug-and-play mechanisms for configuring heterogeneous objects. IoT mega-data is created at this layer.

and transmission power. These objects are connected by networking and communication protocols, allowing them to communicate and cooperate with each other to share information. These protocols represent the backbone of the IoT, they specify the data exchange format, data encoding, object addressing schemes as well as data packet routing [10]. Sequence control, flow control and retransmission of lost packets are also part of their functionality. In the Edge-IoT system, several protocols of different types are used. However, the best technology for our IoT use cases means that we need to accurately weigh the criteria in terms of range, bandwidth, QoS, security, power consumption, and network management.

The Tables 1 and 2 provide an overview of the most commonly used networking and messaging protocols.

**Table 1.** Overview of the most commonly networking protocols

| Protocols | Frequency | Data rate | Range | Topology | Service |
|---|---|---|---|---|---|
| **Zigbee** [11] | 2,4 GHz | 250 kbps | 10-100 m | Mesh, Star | Home automation |
| **BLE** [12] | 2,4 GHz | 1 Mbps | 100 m | Star, Bus | Small appliances |
| **WIFI** [13] | 5 GHz | 1 Gb | 100 m | Star, Mesh | Home |
| **NFC** [14] | 13,56 MHz | 420 kbps | 20 cm | P2P | Payment system |
| **LoRaWAN** [15] | 125 kHz | 50 kbps | 5-20 km | Star | Smart city |
| **5G** [16] | 30-300 GHz | 50 Gbps | Several kms | Star | Virtual reality applications |

**Table 2.** Overview of the most commonly messaging protocols

| Application protocols | Restful | Trans. Layer protocol | Architecture | Security | QoS |
|---|---|---|---|---|---|
| **AMQP** [17] | No | TCP | Pub-Sub | SSL | Yes |
| **CoAP** [18] | Yes | UDP | Pub-Sub Req-Res | DTLS | Yes |
| **MQTT** [19] | No | TCP | Pub-Sub | SSL | Yes |
| **DDS** [20] | Yes | UDP/ TCP | Pub-Sub Req-Res | DTLS, TLS, DDS Security | Yes |

affect transmission time are reduced. Then considerably, the network performance is improved to meet the latency sensitive IoT applications.

**Computation**: in the Edge-IoT architecture, the computation follows the Edge cloud combination mode. The computation will be offloaded to the Edge servers thus relieving the cloud servers. Therefore, the speed and efficiency of the network in terms of resource utilization will be guaranteed.

**Storage**: the number of devices connected to the IoT is increasing and producing huge amounts of data. Previously, to be processed, all this data was at centralized cloud servers. This severely affected the user experience. The migration of all data collected by IoT devices to Edge storage servers solves this problem. Edge computing uses load balancing technology across different nodes, which would improve the quality of service.

The following Fig. 1 shows an example of a three-layer intelligent systems architecture assisted by IoT and Edge computing. However, note that the Edge-IoT architecture can change depending on the problem it wants to solve for an IoT application [7].

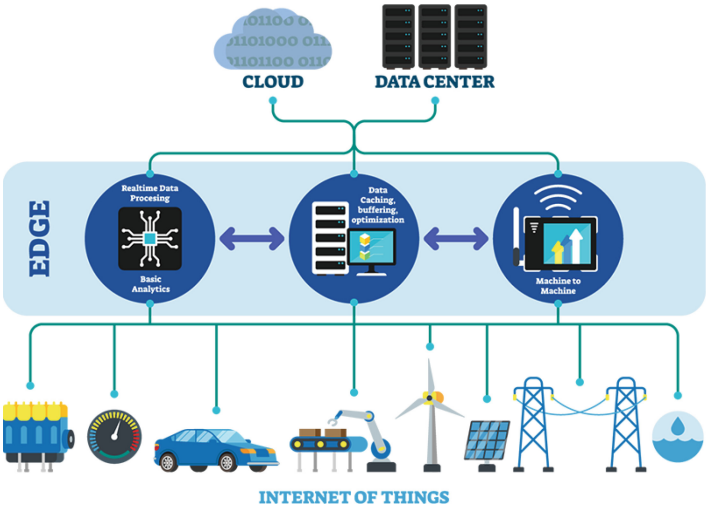



**Fig. 1.** IoT-Edge architecture from [9]

However, in-depth studies of this new computing model have shown that traditional non-IA techniques have limitations in improving the performance of Edge computing. With the successful application of AI in various fields, Edge computing researchers are beginning to turn their attention to AI, in particular machine learning, a branch of AI that has gained popularity over the last few decades.

### 3.2 Edge-IoT Protocols for Intelligent Systems

The Edge-IoT system is a heterogeneous network that includes different devices that may have different communication, data processing and storage platforms, networks

Processing the data from IoT monitoring is much more difficult. Redesigning the monitoring system's architecture will enable real-time data processing and responsiveness.

To enable real-time performance, the IoT monitoring system must continuously evaluate data as it comes in and improve its insight and decision-making capabilities to support sophisticated business logic [6].

Many research works have shown that edge computing can improve the performance of IoT networks and provide real-time services. They highlight the benefits of edge computing for various intelligent applications.

Authors in [5] discuses edge computing for Internet of Things (IoT) applications. It examines how edge computing can improve the performance of IoT networks, particularly latency, by improving computing and storage capabilities, and providing real-time services to users. The paper presents a comprehensive review of edge computing architectures for IoT, classifying them by architecture and comparing their performance in terms of network latency, compute capacity, and storage space. The authors also examine the security issues associated with these approaches, as well as the advantages and disadvantages of edge computing for IoT. In conclusion, the paper highlights the potential benefits of edge computing for various smart applications, including smart grids, smart cities, and smart transportation.

Paper [7] discusses the challenges of storing and processing the enormous amount of data generated by IoT devices, as their resources are limited. It explains that while cloud computing can be used to handle some of these resource issues, it can also introduce latency in time-critical IoT applications. This paper also highlights the limitations of Edge-computing architectures (ECAs)-IoT and proposes solutions to overcome these limitations. It outlines various IoT implementations in the edge computing domain and suggests four distinct scenarios for using IoT implementations with ECAs-IoT. Additionally, the document discusses the concepts of edge computing, implementation strategies, and edge intelligence. It provides a taxonomy for IoT applications based on different criteria and compares fog computing to cloud computing.

The universal complex event processing (CEP) mechanism for IoT real-time monitoring is proposed in [8]. It Proposes a formalized hierarchical complex event model including raw event, simple event and complex event, which reduces the complexity of event modeling. The approach allows for flexible, complicated events to be defined by programming using sophisticated time and space semantics. The proposed system is installed at the network edge between cloud-based apps Two application case studies of IoT monitoring for fruit transportation and city road manhole cover status are implemented based on the CEP system to highlight our proposed complex event model.

## 3 Edge-IoT Architecture and Protocols for Intelligent Systems

### 3.1 Edge-IoT Architecture for Intelligent Systems

Edge computing addresses the requirements of IoT applications on three categories, such as communication, computation and storage [5].

**Communication**: with the assistance of Edge computing, the performance in terms of bandwidth utilization, latency, device power consumption and packet data overhead that

- Description of the concept and architecture of Edge-IA technology by focusing on its applications in connected health and smart agriculture.
- Discussion of the security challenges and future directions in Edge-AI on IoT Environment.

The remainder of this paper is organized as follows: In Sect. 2, we introduce the motivation for using Edge computing and summarise related work on edge computing for real-time Internet of things applications. In Sect. 3, we present an example architecture for Edge computing-assisted intelligent IoT systems and integration protocols. Then, in Sect. 4, we discuss the concept and opportunities of Edge-IA technology and also present a architecture and application areas in connected health and smart agriculture. In Sect. 5, we discuss security challenges and future directions. Finally, in Sect. 6 we conclude by showing the benefits that Edge-IA brings to intelligent systems.

## 2 Background

### 2.1 Motivation for Using Edge Computing

The growth of internet of Things has led to increased demand for data processing and analysis, which introduced the development of new computing models, such as edge computing, which focuses on bringing processing and data closer together, storing end-user data to reduce latency, improve performance, and relieve computing demand on centralized data centers. Data processing and analysis at the network edge helps improve the performance of IoT networks and provide real-time services to users.

Traditional cloud computing architectures, while powerful and versatile, are not always equipped to meet the demands of real-time IoT applications. The latency associated with data transit to centralized cloud servers (as well as bandwidth limitations) can hinder the very essence of IoT: immediate, contextual responsiveness. To overcome these limitations, a paradigm shift is underway, which brings computation closer to the data source and therefore to the edge of the network [5]. Edge Computing is the transformative solution to the real-time challenges posed by IoT. Indeed, Edge-based IoT deployments must be managed, efficiently, effectively, and automatically. IoT applications that involve endpoints, edge computing and fog are complex to manage.

### 2.2 Edge Computing for Real-Time Internet of Things Applications

The IoT real-time monitoring market is expanding quickly in many IoT innovative applications, including smart logistics, smart farms, environmental monitoring, intelligent transportation, and smart power grids; that to the rapid development of sensors, GPS position sensors, RFID tags and readers, smart objects, and other IoT sensing technologies.

By the year 2025, 50 billion devices will be online, according to new research. These Internet of Things (IoT) sensing devices are data generators that continuously provide fresh perception data, offering enormous potential for real-time monitoring and intelligent application across numerous industries.

improves efficiency, and enables data-driven decision-making across various sectors, including home automation, healthcare, transportation, agriculture, and industry [1]. IoT represents a significant shift in how we interact with our surroundings, offering new possibilities for convenience, productivity, and innovation while also raising challenges related to data security, privacy, and scalability.

Edge computing is a distributed computing paradigm that brings computational processing and data storage closer to the source of data generation, typically at the"edge" of a network or IoT devices, rather than relying on centralized cloud data centers. This approach reduces latency, enhances real-time processing, and conserves bandwidth by processing data locally or in nearby edge servers [2]. Edge computing is essential for applications requiring low latency, such as autonomous vehicles, industrial automation, and IoT devices, enabling faster responses and improved efficiency. It complements cloud computing by decentralizing computing resources, allowing for more efficient and responsive data processing in various domains. We can result that this reduces latency, enhances data privacy, and conserves bandwidth by processing and analyzing data locally or in nearby edge nodes.

Artificial Intelligence (AI) is a multidisciplinary field of computer science focused on creating machines and computer systems that can perform tasks that typically require human intelligence. For this, AI encompasses machine learning and deep learning techniques that enable computers to perform tasks that typically require human intelligence, such as learning from data, making decisions, recognizing patterns, and natural language understanding [3].

Edge AI (or intelligent Edge) appeared in a desire to integrate artificial intelligence as close as possible to sensors or connected objects (IoT). The advantages are multiple: absence of prerequisites for permanent internet connectivity, data confidentiality, reduction of latency. To achieve the expected results, Edge AI makes it possible to move part of the computing flow directly into the connected objects, and therefore to minimize the use of the cloud for the tasks related to the processing.

In essence, Edge AI empowers devices and sensors to become"smart" by running AI algorithms locally, enabling quicker decision-making and reducing the need for constant internet connectivity. It finds applications in various fields, including autonomous vehicles, smart cities, healthcare, and industrial automation, where low latency and real-time processing are critical [4]. Edge AI is at the forefront of innovation, addressing the demands of our increasingly connected and data-driven world while raising challenges related to resource constraints, security, and scalability.

In this paper, we provide a holistic view of the convergence of Edge AI and IoT, offering insights into its architectural foundations, diverse applications, ongoing challenges, and promising future directions. The keys contributions of this paper are listed as bellow:

- Overview of IoT, Edge Computing, and Artificial Intelligence (AI) paradigms. Particular.
- Summary of related work on the motivations for using Edge computing and Edge-IoT systems for real-time monitoring.
- Overview on networking and communication protocols for Edge-IoT systems.


# Edge- AI and Internet of Things for Intelligent Systems: Architectures, Applications and Future Perspectives

Fatou Diop(✉), Babacar Mbaye Faye, and Ibrahima Niang

Cheikh Anta Diop University, Dakar, Senegal
{fatou110.diop,babacarmbaye.faye,ibrahima1.niang}@ucad.edu.sn

**Abstract.** In an era where connectivity is ubiquitous and data flows ceaselessly, Edge IoT (Internet of Things) systems have emerged as a pivotal technological paradigm poised to revolutionize the way we interact with the digital world. As the proliferation of IoT devices continues unabated, the traditional cloud-centric model for data processing and decision-making faces significant challenges. Edge IoT systems, by shifting computational intelligence closer to the data source, offer a compelling solution to these challenges, promising greater efficiency, lower latency, and enhanced real-time decision-making capabilities. On the other hand, the convergence of edge computing, artificial intelligence (AI), and the Internet of Things (IoT) has ushered in a new era of intelligent systems with transformative potential across various domains. This includes the development of edge computing infrastructure, AI models optimized for edge deployment, and IoT device architectures that support real-time data processing and analysis. This paper proposes a comprehensive review on Edge computing, AI, and IoT integration for intelligent systems. We focus on the concept of Edge-AI and its potential applications in real-time processing for Internet of Things. The paper introduces Edge-IA technology, explores commonly used networking and messaging protocols in IoT, and discusses the opportunities and challenges of Edge-IA and IoT in various industries, including agriculture and healthcare.

**Keywords:** Internet of things · Edge computing · Artificial intelligence · Edge-AI

## 1 Introduction

In recent years, the Internet of Things (IoT) has become one of the most important technologies of the 21st century. IoT is a revolutionary concept that involves connecting everyday objects and devices to the internet, allowing them to collect and exchange data, communicate with each other, and be remotely controlled. IoT transforms ordinary items, such as appliances, vehicles, wearables, and industrial machines, into "smart" devices capable of sensing, transmitting, and acting upon data. This interconnected network of devices and sensors provides real-time information, enhances automation,


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 111–122, 2025.
https://doi.org/10.1007/978-3-031-81573-7_9

9. Liu, N., Dong, X.: Research on the reform and security of college examination. Journal of Shenyang Architectural University (2008). [Citation Time(s):1]
10. Wang, L., Liu, H.-Y.: Analysis on the Restrict of College Examination Reform. Higher Engineering Education Research (2006)
11. Han, S.-J.: The exploration of evaluation mode in the high professional university. Educ. Prof. **29**, 166–167 (2011)
12. Li, X.-Q., Shu, H.: The Exploration and Practice of Examination Model Reformed by Stratified Teaching in Higher Vocational (2011)
13. Mathematics Curriculum. Journal of Chifeng University, vol. 9, pp. 252–253
14. Niu, W.-X.: Quality education calls for the examination reform. Educ. Naval Acad. **3**, 73–74 (2000)
15. Zhao, J.-Z.: The research on the present situation of the examination reform in higher vocational colleges. Chin. Exam. **7**, 40–44 (2010). [Citation Time(s):1]
16. Al-hawari, F., Althawbih, H., Alshawabkeh, M., Abu Nawas, O.: Integrated and secure web based examination management system. Computer Applications in Engineering Education (2019)
17. Khan, T.: Examination Scheduling System with Proposed Algorithm, s. d., 1
18. Omole, C.: Modifying Assessment Modes in the Science Classroom as a Solution to Examination Malpractice, s. d., 1
19. Jamaluddin, N.F., Aizam, N.A.H.: Timetabling Communities' Demands for an Effective Examination Timetabling Using Integer Linear Programming, s. d., 1
20. Ajith, R., Bijlani, K.: Serious Game as a Performance Assessment Tool That Reduces Examination Anxiety, s. d., 1

## 4 Results

We have a bank of questions that we can enrich over time with our two methods. These results are presented in a CSV file, and we plan to set up a platform to automate the populating of this dataset.

## 5 Conclusion

In this paper, we have recalled the context and the goal of automating evaluations. Subsequently, we used NLP to give a proposal for automatic question generation by proposing two features(negation, and duplication).

On related works such as [1,2] and [3], we have multiple choice generations question that are based on wikipedia topics and courses respectively. In all cases, it is a generation of questions on knowledge bases. To improve these rich scientific works, we propose a reformulation of the questions already existing in a question bank by presenting two functions. One modifies the question by taking its negation, the other duplicates the question by modifying a main keyword.

In perspective, we are looking:

- The quality of the questions produced by our two functions.
- The generation of distractors from the questions
- The implementation of the application in a general way.

discuss other ways of generating questions such as information retrieval.

## References

1. Graesser, A.C., Wisher, R.A.: Question generation as a learning multiplier in distributed learning environments. US Army Research Institute for the Behavioral et Social Sciences, Alexandria VA (cf. p. 1, 21, 23, 85, 122) (2001)
2. Raynaud, T.: Génération de questions á choix multiples thématiques á partir de bases de connaissances. https://tel.archives-ouvertes.fr/tel-02901501/document
3. Nwafor, C.A., Onyenwe, I.E.: An automated multiplechoice question generation using natural language processing techniques. https://aircconline.com/ijnlc/V10N2/10221ijnlc01.pdf
4. uvs(DISI) Université Virtuelle du Sénégal[en ligne] (modifié en mars 2023). https://www.uvs.sn/luniversite-2/luniversite-en-chiffres/. Accessed 21 June 2022
5. DATASCIENTEST [en ligne] (Created on 22/07/2020). https://datascientest.com/introduction-au-nlp-natural-language-processing. Accessed 11 Feb 2022
6. spacy[en ligne] (created since 2016). https://spacy.io/usage/spacy101#whats-spacy. Accessed 21 Oct 2022
7. Kuchling, A.M.: (Documentation Python 3.11.2) (modified on March 30, 2023). https://docs.python.org/fr/3/howto/regex.html. Accessed 30 Oct 2022
8. waset. Conférence internationale sur le système de gestion des examens Riga, Lettonie du 18 au 19 juin 2020. https://jemsenegal.wordpress.com/2021/04/06/jems-un-modele-de-systeme-demanagement-des-examens-pour-les-universites-du-senegal/

```
[ ]  1 import re
     2 def affirmative_en_negatives(x):
     3     verbe=trouve_le_verbe(x)
     4     print(verbe)
     5     phraseNeg=re.sub(verbe, ' ne '+verbe+' pas ' ,x)
     6     return phraseNeg
     7
```

```
[ ]  1 phrase=input('Donner une phrase: ')
```

```
Donner une phrase: La balise main spécifie le contenu principal d'un document.
```

```
[ ]  1 trouve_le_verbe(phrase)
```

```
'spécifie'
```

```
[ ]  1 affirmative_en_negatives(phrase)
```

```
spécifie
'La balise main  ne spécifie pas  le contenu principal d'un document.'
```

**Fig. 3.** Affirmative to negative

```
1 import re
2 qu=input('Une question')
3 listMots=[]
4 while nbq in(range(3)):
5   nbq=int(input('Combien de question voulez vous reproduire?'))
6
7 for j in range(nbq):
8   m=input('saisir un mot')
9   listMots.append(m)
10 print(listMots)
11 mot_a_remplacer=input('Donnez le mot cles a remplacer')
12 listqu=[qu]
13 i=0
14 for mot in listMots:
15   quNouveau=re.sub(mot_a_remplacer,mot,qu)
16   listqu.append(quNouveau)
17   i=i+1
18 listqu
```

**Fig. 4.** Duplicate question

```
Une questionQuel attribut HTML est utilisé pour ajouter un style dans un élément?
saisir un motlien
saisir un motjeu de caractere
saisir un motcouleur de font
['lien', 'jeu de caractere', 'couleur de font']
Donnez le mot cles a remplacerstyle
['Quel attribut HTML est utilisé pour ajouter un style dans un élément?',
 'Quel attribut HTML est utilisé pour ajouter un lien dans un élément?',
 'Quel attribut HTML est utilisé pour ajouter un jeu de caractere dans un élément?',
 'Quel attribut HTML est utilisé pour ajouter un couleur de font dans un élément?']
```

**Fig. 5.** results of the duplication

Detecter le verbe dans la phrase

```
def trouve_le_verbe(phrase):
  ph=nlp(phrase)
  listToken=[]
  for token in ph:
   listToken.append([token.text,token.tag_])
   for m in listToken:
     if m[1]=='AUX' or m[1]=='VERB':
       break
  vb=m[0]
  return vb
```

**Fig. 2.** Verb detection

"ne ... pas" (in french) with the sub function. The sub() function searches for all substrings where the RE matches and substitutes them with a different string (Fig. 3).

### 3.3 Question Generation by Duplicating the Question

The execution of this code gives us the following results:

For this part we used regular expressions with the sub function which is explained above.

1. The user has to give a keyword to replace.
2. Then we build the dictionary of words that we will use
3. From this dictionary, we build a list to have a set of questions.

For more details on regular expressions in Python, please consult the documentation [7]. Try to add more coments about the methods and the performance of the models (Fig. 4).

This algorithm produces the following results (Fig. 5).

### 3.2 NLP Approaches Based on Word-Embedding

As in [4], NLP, Natural Language Processing, is a discipline that focuses on the understanding, manipulation and generation of natural language by machines. Thus, NLP is really at the interface between computer science and linguistics. It is about the ability of the machine to interact directly with humans.

#### 3.2.1 Description of Our Two Algorithms Stated Above

Preprocessing In NLP, the first step is often preprocessing, which includes the tokenization and text cleaning steps. We are content here with a minimalist preprocessing: removal of punctuation and stop words (for visualization and vectorization methods based on counts). We propose to use the spaCy library which allows better automation in the form of preprocessing pipelines. spaCy is a free open source library for advanced natural language processing (NLP) in Python. If we work with a lot of text, we might want to know more about this. For example, what is it? What do the words mean in their context? Who does what to whom? What companies and products are mentioned? Which texts are similar? spaCy is designed specifically for production use and helps we create applications that process and "understand" large volumes of text. It can be used to create systems for information extraction or natural language understanding, or to preprocess text for deep learning. For more information on spacy see [5].

#### 3.2.2 Question Generation Using Negation

Can we reproduce other questions from this topic? It is possible to use the negation or remove the negation.

Example: the question (The imain; tag specifies the main content of a document. True or False and the correct answer is True) One could be modified as follows: The main tag does not specify the main content of a document. True or False and the correct answer is False Likewise for the question (Which of the tags does not belong to HTML 5? Here all the answers are good) In this case the modification is as follows: Which of the tags belong to HTML 5 ? and the answer is no right answer.

To make a question negative, we propose two steps: **Verb detection:** To write this algorithm, we first installed the spacy library in order to be able to use the various useful functions to find the verb in the sentence and import the 'fr_core_news_sm' language (Fig. 2).

Once the verb is found, we will define an "affirmative_en_negative" function or we use regular expressions (re) with the sub function.

Regular expressions (denoted RE or regex patterns in this document) are essentially a small, highly specialized programming language embedded in Python and whose manipulation is made possible by the use of the re module. Using this small language, you define rules to specify a match against a desired set of strings; these strings can be sentences, email addresses, TeX commands or whatever. You can also use RE to modify a string or split it in different ways. Like our case in the "affirmative_en_negative" function, we enclose the verb in

- testinvite It is an e-examination platform that offers the following solutions:
  - the creation of exams,
  - the design of examination processes,
  - the presentation of online exams,
  - report generation
  - analysis of examination results,
  - the pricing structure.

  In their documentation, it specifies that there is no payment but certain sections such as sending exam invitations to candidates by the Test Invite electronic messaging system, starting a candidate's exam, Video or photo surveillance for exam security are charged. Head to testinvite for more information.

We have many systems that offer interesting solutions for exams, But we propose automatic methods for populating a question bank using NLP.

| | A | B | C | |
|---|---|---|---|---|
| | No | QUESTIONS | REPONSE | DISTRACTEURS(OPTIONNEL) |
| 1 | 1 | Quel attribut HTML est utilisé pour ajouter un style dans un élément? | Style | |
| 2 | 2 | Quel attribut HTML est utilisé pour ajouter un lien dans un élément? | href | |
| 3 | 3 | Quel attribut HTML est utilisé pour ajouter une couleur de fond dans un élément? | background | |
| 4 | 4 | C'est quoi un élément sémantique ? | Un élément qui a une signification. | |
| 5 | 5 | Lesquelles des balises n'appartiennent pas à HTML 5? | <br><font><br><strike><br><s><br><u> | |

**Fig. 1.** bank of question

## 3 Matériel et Méthodes

In this study, we have taken the case of the UN-CHK, more precisely the multimedia stream with the Web Technology1-HTML/CSS module. We use two methods to populate the question bank:

- Use the negation of a question.
- Duplicate a question by modifying a keyword.

### 3.1 Presentation of the Dataset Web Technology-Initiation in HTML/CSS

This dataset has three parts: (Fig. 1)

- The question column proposed by teachers or tutors
- The answer column of each question
- The distractors column which is optional. The user can give it but we can propose an algorithm that calculates it from the proposed question.

We propose two techniques for question generation:

1. Use the negation of a question.
2. Duplicate a question by changing a keyword

Before this design idea using NLPs, we had set up a prototype whose goal was to bank exam subjects. This project is based on a database model and is UN-CHK, Senegal developed with the php language using the MVC model. This MVC model does not allow us to enrich the bank of questions in an automatic way, so it will be necessary to use other models of information retrieval and/or machine learning, more precisely the NLP models. The paper is organized as follow: We present our subject bank dataset in the next section. Then we explain the NLP approach we used to automatic question generation in section three. We describe the two algorithms we used to generate questions automatically in the fourth section and we finish by a general conclusion in the last section.

## 2 Bibliographic Review

The comprehensive design of computerized testing systems has been explored by many researchers in different institutes around the world such as those in [9–20]. For example in [9–16], the effective reform of college examination management is discussed Based on the concept of scientific development of examination management. In [16], the analysis, design, development, integration, deployment and security of a web-based exam management system which was developed in-house at the German Jordanian University (UGJ) was addressed. An algorithm for the exam scheduling system is recalled in [17]. An algorithm for the exam scheduling system is recalled in [17]. In relation to malpractices, [18] then identified modifications that could help relieve the student from exam-related stress and thus increase the student's effort towards effective learning and discourage malpractices in the long run. term. The problem of university examination timetable was explained in [19] using topics from Universiti Malaysia Terengganu (UMT). Still for the simplification of exams, [20] proposes the "Serious games" as tools to carry out evaluations by reducing the exam anxiety encountered by learners.

However, we have plenty of EMS that exist today and which offer important solutions for e-exams with competitive prices. We will detail them on the following lines:

- testwe
  testwe is a complete e-Exam solution: online platform for teaching staff and administration, offline software for learners In their white paper entitled "THE 5 IMPACTS OF THE E-EXAM" they describe how e-exams can help institutions compared to paper-based exams.
- Managexam By bringing together teaching and student administration around the same digital environment, Managexam facilitates the organization and management of evaluations while guaranteeing a high standard of educational integrity at each stage. Thanks to its simple and flexible tool Managaexam allows the design of:
  - Classic subjects
  - Evaluation activity enriched with video, audio recordings or large documents with customization options such as random drawing of questions or even variability of data between candidates

  For more details, go to

# Question Design Using NLP

Maty Sene Kane[1(✉)], Alassane Diop[1], Zinflou Arnaud[2], and El Hadji Mamadou Nguer[3]

[1] Université Numérique Cheikh Hamidou Kane (formerly UVS), Thiès, Senegal
maty1.sene@unchk.edu.sn, alassane.diop@unchk.edu.sn
[2] Hydroquebec-CANADA, Montreal, Canada
zinflou.arnaud@hydroquebec.com
[3] Université Numérique Cheikh Hamidou Kane (formerly UVS), Dakar, Diamniadio, Senegal
elhadjimamadou.nguer@unchk.edu.sn

**Abstract.** Cheikh Hamidou KANE Digital University (ex UVS) has more than 60,000 students enrolled in 46 tracks containing more than 1,000 courses, the assessment of which requires the preparation of nearly 2,000 exam topics per year. Since its opening in 2013, it is estimated that nearly 11,000 topics have been prepared, and this number is growing (see [4]). In addition to these numbers, on the one hand we have the lack of an examination management system (EMS). On the other hand, we see enormous difficulties in receiving assessment topics at the teacher level, given the large number of topics a teacher has to design. We propose a framework called JEMS (Jolof Examination Management System) that allows for declarative expression and evaluation of topics close to their design. In order to be able to generate questions automatically, we have made use of other models such as natural language processing based models. This framework aims to automate evaluation questions from an existing question bank for a CE (Constituent Element). Automating evaluations in this context poses a number of scientific challenges that constitute the contributions of the presented work:

- The implementation of a file template in csv format.
- Question models that identify entities in CEs and generate statements against the natural language processing model

**Keywords:** NLP · generate questions automatically · examination management system

## 1 Introduction

Computer-based exams have grown in popularity in recent years as they support e-learning and make it easier to manage (i.e. design, set up, schedule and score, publish results) of exams. However, adopting an appropriate system that meets the needs of a certain institution can be difficult due to several concerns related to integration with existing systems, security, customization, ease of use and cost.


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 103–110, 2025.
https://doi.org/10.1007/978-3-031-81573-7_8

## References


1. Kasemthaweesab, P., Kurutach, W.: Association analysis of Diabetes Mellitus (DM) with complication states based on association rules, In: Proc. 7th IEEE Conference on Industrial Electronics and Applications, pp. 1453–1457. (2012)
2. http://www.who.int/mediacentre/factsheets/fs312/en/
3. Zarkogianni, K., et al.: A review of emerging technologies for the management of diabetes mellitus. IEEE Trans. Biomed. Eng. **62**(12), 2735–2749 (2015)
4. Collins, G.S., Mallett, S., Omar, O., Yu, L.M.: Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. BMC Med. **9**(1), 103 (2011)
5. Shankar Acharya, D.O., Samanta, S., Vidyarthi, A.S.: Computational intelligence in early diabetes diagnosis: A review. The Review of Diabetic Studies: RDS **7**(4), 252 (2010)
6. Fikirte Girma Wolde, M., Sumitra, M.: Prediction of Diabetes using Data Mining Techniques, Dept of Computer Science and Engineering, In: Proceedings of the 2nd International conference on Trends in Electronics and Informatics (ICOEI) (2018)
7. Terry Jacob, M., Elizabeth, S.: Analysis Supersived Learning Techniques for Cost Effective Diaease Prediction using on-Clinical Parameters, IIITM-KTechno park, Trivndrum, (2018)
8. Butwall, M., Kumar, S.: A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier. International Journal of Computer Applications **120**, 0975–8887 (2015)
9. Dewangan, A.K., Agrawal, P.: Classification of Diabetes Mellitus Using Machine Learning Techniques. International Journal of Engineering and Applied Sciences, **2** (2015)
10. Devi, M.R., Shyla, J.M.: Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus. International Journal of Applied Engineering Research **11**, 727–730 (2016)
11. Hina, S., Shaikh, A., Abul Sattar, S.: Analyzing Diabetes Datasets using Data Mining. Journal of Basic & Applied Sciences **13**, 466–471 (2017)
12. Mujumdar, A., Vaidehi, V.: Prédiction du diabète à l'aide d'algorithmes d'apprentissage automatique Conférence international sur les tendances récentes en informatique avancée, 2019, ICRTAC (2019)
13. Yuvaraj, N., Sri Preetha, K.R.: Prédiction du diabète dans les systèmes de santé à l'aide d'algorithmes d'apprentissage automatique sur le cluster. Hadoop Calcul de cluster **22**, 1–9 (2017)
14. Sisodia, D.: DS Sisodia Prédiction du diabète à l'aide d'algorithmes de classification Process. Comput. Sci. **132**, 1578–1585 (2018)
15. Erveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K.: Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science **82**, 115–121 (2016). https://doi.org/10.1016/j.procs.2016.04.016
16. Nai-Arun, N., Sittidech, P.: Ensemble Learning Model for Diabetes Classification. Advanced Materials Research 931–932, 1427–1431 (2014). https://doi.org/10.4028/www.scientific.net/AMR.931-932.1427

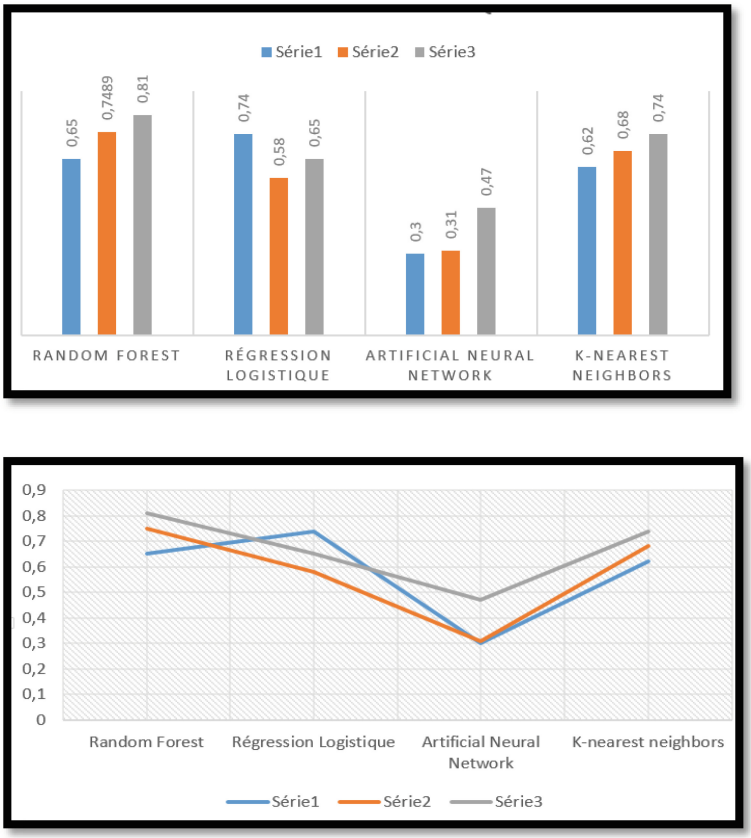



**Fig. 10.** Graphical representation of the results

## 5 Conclusion

Diabetes, acknowledged as one of the most perilous and persistent ailments leading to elevated blood sugar levels, has increasingly permeated every aspect of daily life, impacting families on a profound level. Recognizing the pervasive influence of this chronic disease, there is a heightened awareness of the critical medical challenge posed by the early detection of diabetes. In our study, we sought to compare three machine learning algorithms: Random Forest, Logistic Regression, and Neural Networks. The experimental findings, derived from the Kaggle dataset, reveal Random Forest's superiority in terms of heightened accuracy compared to the other algorithms. Subsequent endeavors will involve replicating the same experiments on additional diabetes databases or diverse datasets to validate and enhance the obtained results, with the ultimate aim of refining the algorithms for improved accuracy.

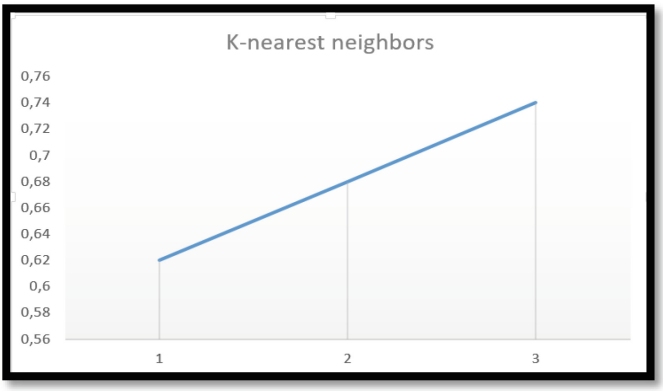



**Fig. 9.** Evolution of the tests with K-nearest neighbors

**Table 5.** Results obtained.

| Model | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| Random Forest | 0,65 | 0,7489 | 0,81 |
| Logistic Regression | 0,74 | 0,58 | 0,65 |
| Artificial Neural Network | 0,30 | 0,31 | 0,47 |
| K nearest neighbors | 0,62 | 0,68 | 0,74 |

The graphs depict the performance of four models: Random Forest, Logistic Regression, K Nearest Neighbors, and Artificial Neural Network based on accuracy rates across varying iterations. Notably, the Random Forest model outperforms the Logistic Regression model, with the exception of iteration 1. Its peak accuracy is observed in iteration 3, reaching 81%. Conversely, the Artificial Neural Network (ANN) exhibits consistently modest results across different iterations, while the K Nearest Neighbors (KNN) algorithm demonstrates improvement with each iteration (Fig. 10).

In direct comparison, the Random Forest, Logistic Regression, and K Nearest Neighbors models outshine the Artificial Neural Network model. This underscores the notion that ANN, being a deep learning algorithm, may find greater utility in other data types, such as image processing. Early identification of individuals at a high risk of diabetes is a pivotal challenge in the healthcare domain. Within this study, the Random Forest model is pitted against the Logistic Regression model, the Artificial Neural Network model, and the K Nearest Neighbors model. The findings suggest that machine learning approaches effectively leverage extensive data sourced from electronic medical records for predicting diabetes risk.

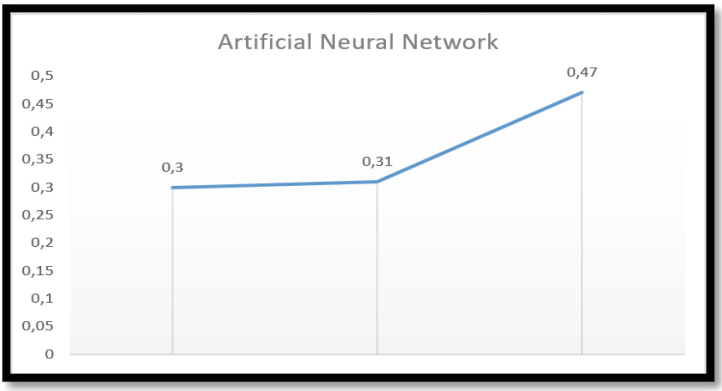



**Fig. 8.** Evolution of the tests with Artificial Neural Network

In the figure, we can clearly see that the best result is captured in the last test, this is justified by the fact that the more the tests increase, the more it gives satisfactory results. The graph represents the accuracy rate for the Artificial Neural Network model according to the number of iterations. It can be seen that the best iteration is iteration 3 with an accuracy rate equal to 47%.

### 4.4 K-nearest Neighbors

In the same proposed work, we also applied the K nearest neighbor's algorithm on the same machine learning and test sample data while varying the K in steps of two. We obtained results for different tests. The results obtained from the experiments performed are reported in the Table 4 below:

**Table 4.** Results of the different tests with k-nearest neighbors

| Model | Test1 K = 1 | Test 2 K = 3 | Test 3 K = 5 |
|---|---|---|---|
| K-nearest neighbors KNN | 0,62 | 0,68 | 0,74 |

From the results shown in the table we can see that the performance results vary between 0.62 and 0.74 for all precision measurements. We also notice that as the k parameter increases, the rate of precision measurements also increases. This is more explicit in the Fig. 9 below.

In the figure, we can clearly see that the best result is captured in the last test. This again shows that the performance is proportional to the variation of K. The graph represents the accuracy rate for the K-nearest neighbor's model as a function of the values of K. It is illustrated that the best iteration is the iteration with k = 5 with an accuracy rate equal to 74%. Thus, a summary Table 5 of the different algorithms and their results are designed with the graph of illustration regrouping the four models.

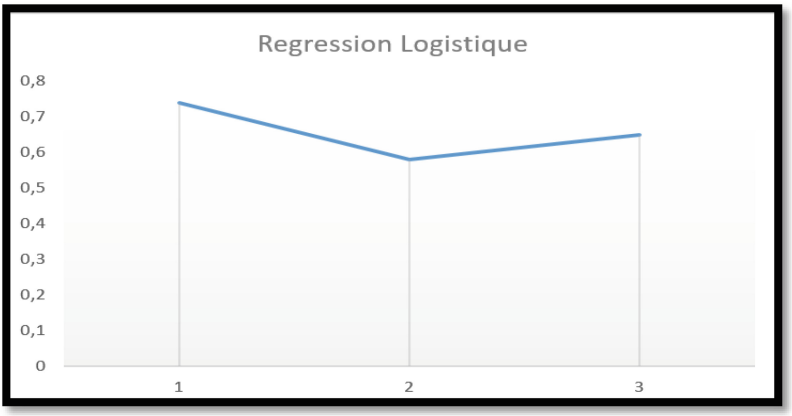



**Fig. 7.** Evolution of the tests with Logistic Regression

In the figure, we can clearly see that the best result is captured in the first test. This divergence in performance between the training and the test phase is known as overlearning. The graph represents the accuracy rate for the logistic regression model as a function of the number of iterations. It is shown that the best iteration is iteration 1 with an accuracy rate equal to 74%.

### 4.3 Artificial Neural Network

In the context of our proposed research, we implemented the artificial neural network algorithm using the same dataset employed for machine learning and testing. Diverse tests were conducted, and it's worth mentioning that, for this learning approach, a sigmoid activation function was utilized for the output layer, while both the input layer and the hidden layer employed a relu activation function. The outcomes of the experiments are detailed in the Table 3 provided below:

**Table 3.** Results of the different tests with artificial neural network

| Model | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| Artificial Neural Network (ANN) | 0,30 | 0,31 | 0,47 |

From the results shown in the table, we can see that the performance results vary between 0.30 and 0.47 for the precise measurements. We also notice that as the number of tests increases, there is sometimes a slight increase. We can deduce that with the low accuracy rates of ANN that this algorithm which is a deep learning algorithm is more interesting on other types of data. This is more explicit in the Fig. 8 below.

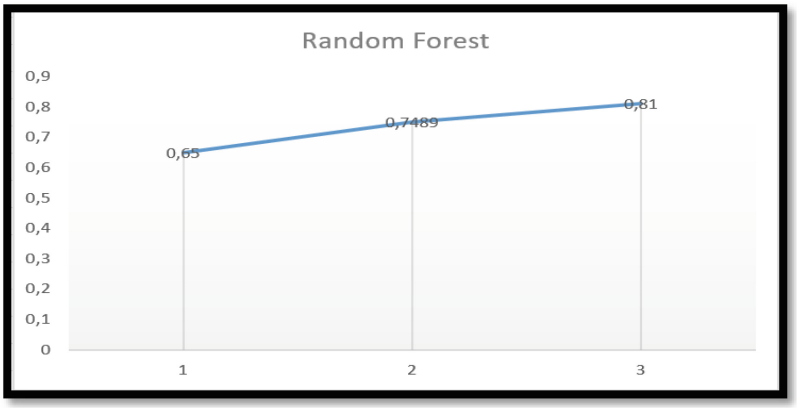



**Fig. 6.** Evolution of the tests with Random Forest

In the figure, we can clearly see that the best result is captured in the last test, this is justified by the fact that the more the tests increase, the more the model is trained, the more efficient it is and the better the results. The graph represents the accuracy rate for the Random Forest model as a function of the number of iterations. It is illustrated that the best iteration is the iteration 3 with an accuracy rate equal to 81%.

### 4.2 Logistic Regression

Subsequently, in the proposed research, we employed the logistic regression algorithm on the identical dataset used for machine learning and testing. Outcomes for various tests were acquired, and the results from the conducted experiments are presented in the Table 2 below:

**Table 2.** Results of the different tests with Logistic regression

| Model | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| Logistic Regression (RL) | 0,74 | 0,58 | 0,65 |

From the results shown in the table we can see that the performance results vary between 0.58 and 0.74 for the precise measurements. We also notice that as the tests multiply, we sometimes see an increase and also a decrease as the tests multiply. This is more explicit in the Fig. 7 below.

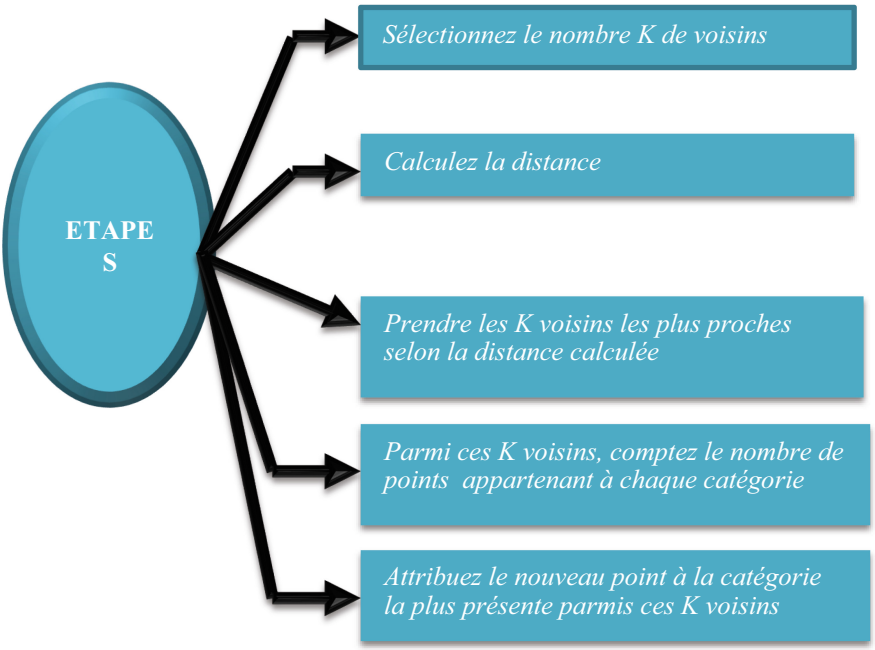



**Fig. 5.** K nearest neighbor's steps

### 4.1 Random Forest (RF)

In our proposed study, we implemented the Random Forest algorithm in the realm of machine learning and tested it with sample data. Various tests were conducted, and the outcomes of these experiments are documented in the Table 1 provided below:

**Table 1.** Results of the different tests with random Forest

| Model | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| Random Forest (RF) | 0,65 | 0,7489 | 0,81 |

From the results shown in the table, we can see that the performance results vary between 0.65 and 0.81 for the precise measurements. We also notice that the more the tests are multiplied, the more we note that the precision tends towards 100 which shows a good control at the training level. This is more explicit in the Fig. 6 below.

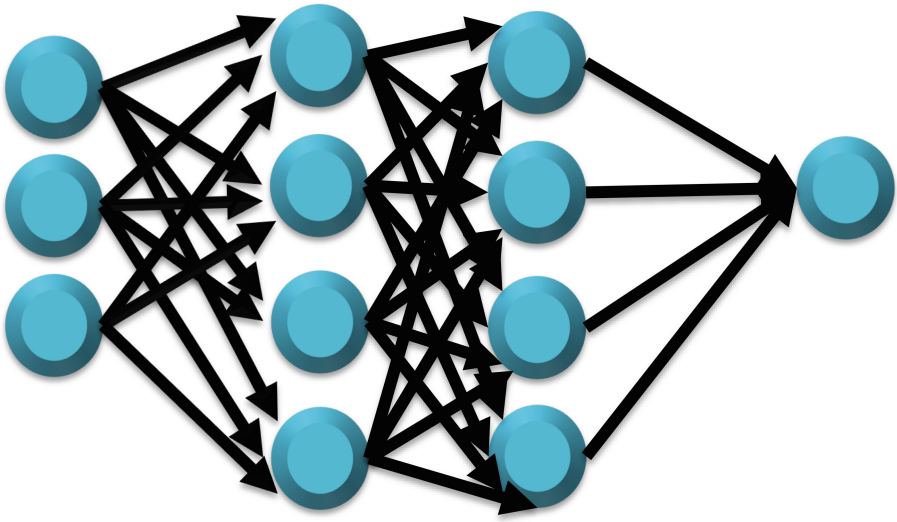


**Fig. 4.** Architecture of neural networks

*K Nearest Neighbors (KNN) stands out as one of the uncomplicated supervised learning methods employed to address classification and regression challenges. Its functionality revolves around the classification of new data points, determined by their similarity to neighboring data points. KNN is characterized as an algorithm devoid of assumptions about the data structure and distribution, rendering it a non-parametric algorithm. It is also referred to as a lazy learner algorithm, as it refrains from immediate learning from the training set. Instead, it stores the dataset and, during classification, takes action based on the dataset. KNN operates by classifying or predicting outcomes based on a fixed number (K) of data points in close proximity to input points. Essentially, for a selected value of K, an input point is classified or anticipated to belong to the same class as the nearest K neighboring points (Fig. 5).

## 4 Experimentation and Results

This section elucidates the methodology employed in this research article, outlining the acquisition of datasets and features. Additionally, it delves into the algorithms utilized and their corresponding evaluation criteria. Following the preprocessing steps that involved addressing null values and eliminating missing data, the predictive model was initially crafted using three algorithms: logistic regression, Random Forest, and neural networks. The Kaggle dataset under consideration encompassed individuals aged between 45 and 84, totaling over 700 participants. The impact of machine learning on diabetic prediction was tested comprehensively in this demographic. Google Colab served as the chosen working environment due to its ease of management and lack of requisite tool installations. The ensuing section presents the accuracy results derived from the examination of four distinct algorithms.

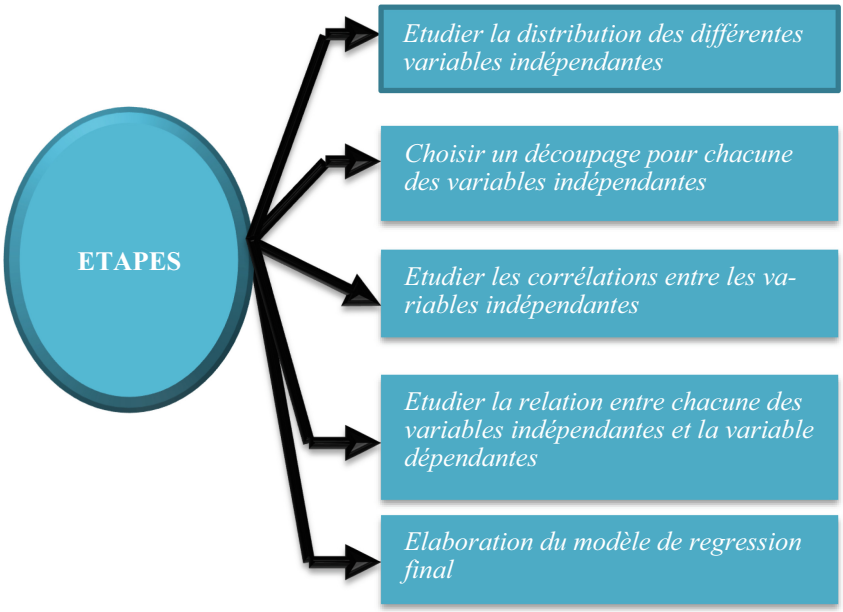



**Fig. 3.** Regression Logistic steps

Features are multiplied by random weights, $W_1, W_2, W_3, \ldots\ldots, W_n$ and added with bias values , $b = b_1, b_2, \ldots.., b_n$. The resulting values are then input into a non-linear activation function, with various types of activation functions possible.

Activation functions can include (2), (3), (4), (5), which are some examples of activation functions (Fig. 3).

$$Sigmoidfonction : \sigma(z) ora(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

$$Tanhfunction : a(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \tag{3}$$

$$RectifiedLinearUnit(RELU) : a(z) = \max(0, z) \tag{4}$$

$$LeakyRELU : a(z) = \max(0.001 * z, z) \tag{5}$$

Here is a synopsis of the proposed neural network model relevant to our study. The architecture comprises three layers, consisting of an input layer, two hidden layers, and an output layer. It's noteworthy that we have chosen to incorporate two hidden layers, but there is flexibility to create more or fewer layers as needed. The algorithm operates in the following manner: the outputs from the input layers serve as inputs to the first hidden layer, the outputs of which become inputs for the second hidden layer, and ultimately, the outputs of the second hidden layer become inputs for the output layer. The output layer produces the final predictive result (Fig. 4).

its predictive accuracy typically surpasses that of a single decision tree. As a rule, the greater the number of trees in the forest, the heightened robustness the forest exhibits (Fig. 2).

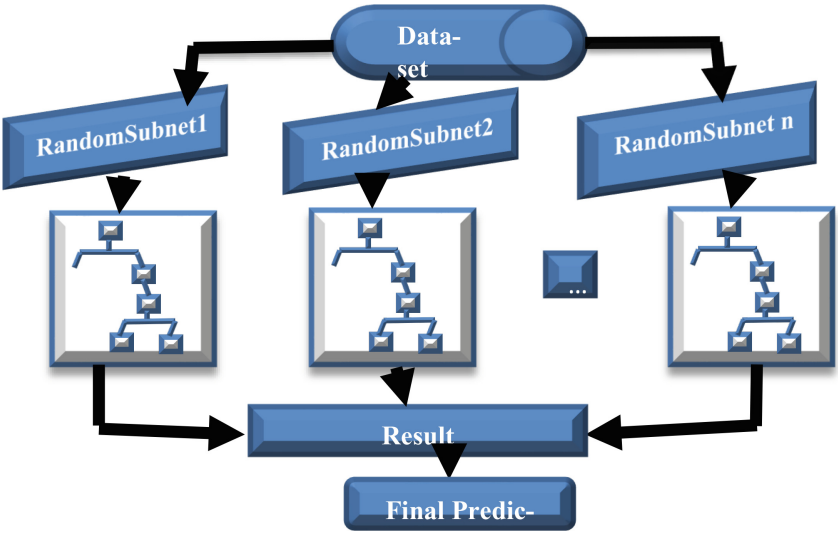



**Fig. 2.** Architecture of Random Forest

Presented below is the suggested framework for this algorithm aimed at predicting whether a patient is diabetic or not. The dataset used for consideration is comprised of diabetic data. Initially, 100 samples (Random Subnet) are extracted from the dataset, and an individual decision tree (Tree) is constructed for each sample. Each decision tree produces an output as a result. Ultimately, the final outcome is determined based on the collective voting of all the results generated by the various trees.

*Logistic Regression (LR): Logistic regression (LR) serves as a discriminant model contingent on the dataset's quality, characterized by the features:

$$\begin{aligned} X &= X_1,\ X_2,\ X_3, \ldots, X_n(where, X_2 - X_n W_1,\ W_2, W_3,\ \ldots, W_n, biais\ b \\ &= b_1, b_2,\ \ldots, b_n = Distinct\ features\ poidset\ Cours\ C \\ &= C_1,\ C_2,\ \ldots, C_n \end{aligned} .$$

The equation for posterior estimation is expressed as follows:

*Artificial Neural Network (ANN) stands as a fundamental machine learning technique, forming the backbone of various deep learning algorithms. The training of the ANN model is accomplished using raw data, and in comparison to alternative classifiers, it possesses an extensive array of tuning parameters, contributing to its intricate structure. Optimizing the error in neural network instances requires a substantial amount of time compared to other techniques. To address this, instances of the neural network algorithm are trained on the graphics processing unit using CUDA programming. Each individual neural node within the ANN is trained with a set of features. $X = X_1, X_2, X_3, \ldots.., X_n(where, X_2 - X_n = Distinctfeatures)$

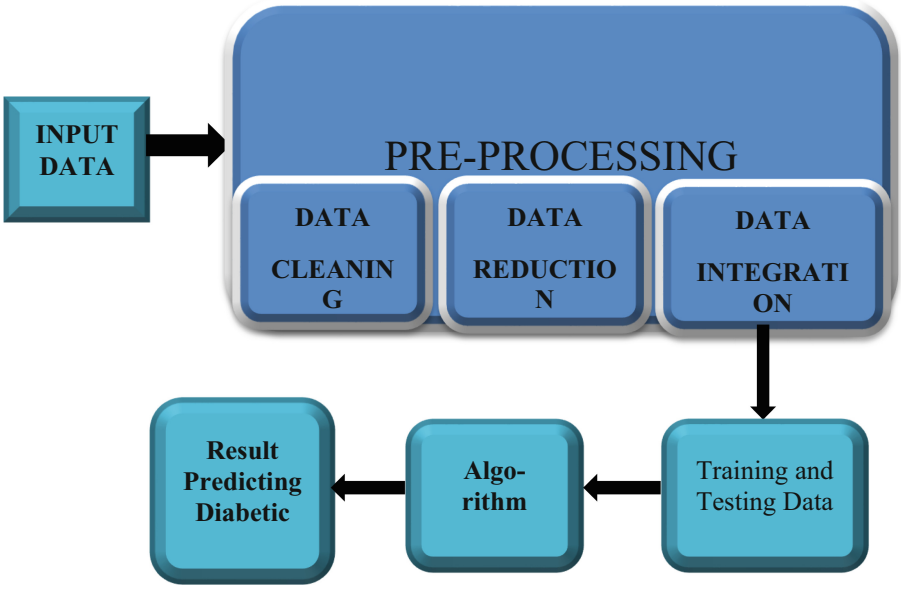



**Fig. 1.** System architecture

### 3.2 Data Pre-processing

Data pre-processing stands out as a pivotal stage in the data discovery methodology. Health information commonly exhibits missing or inconsistent data.

DATA CLEANING involves the strategic identification and correction, or removal, of inaccurate records. It encompasses the identification of incomplete, incorrect, inaccurate, or irrelevant information, with subsequent actions involving substitution, modification, or deletion of coarse data. Information cleaning is executed interactively using data merchandising tools or through scripted processes.

DATA INTEGRATION is the process of retrieving and consolidating heterogeneous data into a unified format and structure. This integration facilitates the utilization of diverse types of data, such as information sets, documents, and tables, for personal or business processes and functions.

DATA REDUCTION entails transforming numerical or alphabetical data, derived from empirical observation or experimentation, into a refined, organized, and simplified form. The core objective is to condense vast amounts of data into meaningful components.

### 3.3 Algorithms Used

*As implied by its name, the Random Forest algorithm constructs a forest comprised of numerous decision trees. This supervised classification algorithm is esteemed for its rapid execution. A multitude of decision trees converges to create a Random Forest, and predictions are made by averaging the outcomes of each individual tree. Notably,

Dewangan & Agrawal [9] adopt a hybrid classification model by combining various classification methods, such as C4.5, Random Forest, and Multilayer Perceptron. Trained on a diabetes dataset from the UCI repository, their work focuses on the detection and classification of diabetes mellitus.

Devi & Shyla [10] delve into early prediction of diabetes using diverse machine learning techniques, including Naïve Bayes, Multilayer Perceptron, Random Forest, Random Tree, and Modified J48. Their analysis, based on the Indian PIMA dataset, reveals that the modified J48 classifier achieves the highest accuracy among the considered techniques.

In another research endeavor [11], the comparison of different classification algorithms, including Naïve Bayes, Multi-Layer Perceptron, J48, Random Forest, and regression, is undertaken to extract intelligent results from diabetic patient data.

Aishwarya and Vaidehi [12] employ a multitude of machine learning algorithms, such as Support Vector Machines, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-NN, Gaussian Naïve Bayes, Bagging Algorithm, and Gradient Boost Classifier. Their study, using the Indian PIMA and another diabetes dataset, reveals a logistic regression accuracy of 96%.

Tejas and Pramila, in a different approach, focus on two algorithms, logistic regression, and SVM, for diabetes prediction. Their preprocessing of data yields optimal results, with SVM exhibiting superior accuracy at 79%.

Yuvaraj and Sripreethaa [13] construct a diabetes prediction model using Random Forest, Decision Tree, and Naïve Bayes algorithms within Hadoop-based clusters. The application of preprocessing techniques results in a notable 94% accuracy with the Random Forest algorithm.

Deepti and Dilip [14] explore the Decision Tree, SVM, and Naive Bayes algorithms, employing a ten-way cross-validation for enhanced performance. Naïve Bayes emerges with the highest accuracy at 76.30%, using the Pima Indian Diabetes dataset.

In a study by Sajida et al. [15], the role of Adaboost and bagging ensemble machine learning methods using J48 as a base for classifying diabetes mellitus is discussed. The experiment demonstrates that the Adaboost ensemble machine learning technique outperforms the J48 decision tree in classifying patients as diabetic or non-diabetic based on diabetes risk factors.

## 3 Proposed Architecture

Datasets are sourced from Kaggle. During the second phase, the data undergoes preprocessing, encompassing cleaning, integration, and processing. Employing machine learning algorithms enhances accuracy in our findings (Fig. 1).

### 3.1 Patient Database

Data was gathered and systematically analyzed to construct a robust model. The dataset comprises pertinent and valuable information acquired through a thorough questioning process. This information is categorized meticulously, with a primary focus and further subdivision into narrowed categories.

include heart disease, stroke, kidney disease, and mortality [1]. According to the World Health Organization (WHO), the global prevalence of diabetes in adults over 18 years of age was 8.5% in 2014 [2]. Moreover, WHO predicts that by 2030, diabetes will become the seventh leading cause of death [2]. Predicting diabetes has become a crucial focus in health research, and contemporary computer models play a significant role in aiding decision-making and supporting self-management of the disease [3].

The increasing importance of machine learning in healthcare is evident, as these techniques consistently deliver high-performance accuracy results, simultaneously reducing human error in decision-making processes. Consequently, they alleviate the strain on healthcare resources [4]. Ideally, the further development of models incorporating prior knowledge would enhance the accuracy of diabetes prediction [5]. Access to health data from a patient's health records has the potential to extract meaningful information and unveil hidden knowledge.

This study aims to conduct a comparative evaluation of the performance of machine learning-based models for predicting diabetes. The prediction approaches were applied to datasets sourced from Kaggle. To the best of our knowledge, this study represents one of the most significant efforts to date in the realm of diabetes prediction. The paper is structured as follows: Sect. 1 provides a comprehensive overview of works utilizing learning techniques in the health domain, with a specific focus on applications related to diabetes. In Sect. 2, the architecture of the system, based on different algorithms, is explained. Section 3 presents the obtained results, offering a comparative analysis of the outcomes derived from the various algorithms employed.

## 2 State of the Art

Machine learning techniques have found application across various domains, with the medical field being no exception. Notably, several studies leverage machine learning classifiers to address medical challenges, with a particular emphasis on the chronic and complex nature of diabetes, a condition that has captured global research attention.

One noteworthy study by Fikirte Girma et al., titled "Prediction of diabetes using data mining techniques" [6], focuses on predicting diabetes operations through the application of the Back propagation rule. The findings indicate that Back propagation demonstrates superior accuracy in polygenic prediction compared to SVM, J48, and Naïve Bayes formulas.

In another study, Terry Jacob et al. present "Analysis of Supervised Learning Techniques for Cost Effective Disease Prediction" [7], which explores cost-effective disease prediction using non-clinical parameters. The study employs various algorithms, with Naive Bayes yielding an 80.37% accuracy, and REP trees achieving a maximum accuracy of 77%, especially when considering Logistic regression.

A methodology proposed by Butwall & Kumar [8] relies on the Random Forest (RF) classifier to understand the behavior of diabetes in alignment with specific lifestyle parameters, including physical activity and emotional states, particularly in elderly diabetics. The research, conducted on the Indian Pima diabetic database from the UCI Machine Learning Lab, reveals the effectiveness of RF in diagnosing diabetes mellitus based on provided attribute values.

# Integration of Artificial Intelligence with Diabetic Data for Increasingly Personalized Medicine

Madiop Diouf[1,3(✉)], Thierno Amadou Diallo[2], Elhadji Ndiaye Diallo[3], Birahime Diouf[3], and Ibra Dioum[3]

[1] USSEIN University, LITA ESP/UCAD, Kaolack BP 55, Dakar, Senegal
madiop.diouf@ussei.edu.sn
[2] UASZ University, LI3 Ziguinchor, Senegal
t.diallo@univ-zig.sn
[3] LITA ESP/UCAD, Dakar, Senegal

**Abstract.** Diabetes is considered the most deadly and chronic disease that causes an increase in glucose. Polygenic disease is one in which the exocrine gland does not produce the hypoglycemic agent and according to the International Federation of Polygenic Diseases 382 million people live with polygenic disease in the world. By 2035, this number will double to 592 million. Diabetes mellitus or simply the disease can be a disease due to increased blood glucose levels. Many difficulties can arise if diabetes is not treated and not identified by the doctor. Thus, artificial intelligence (AI), which has become the new term we hear every day in recent years, generally defines the ability of a machine to act on its own and which is not explicitly programmed to reproduce actions or functions that are generally those of human beings. Today, we find it in our computing machines, social networks, transportation and in the medical sector etc. Therefore, machine learning is one of the disciplines of artificial intelligence that seeks to find a way to create computer programs that automatically improve with experience. In this work, we will focus on the use of machine learning algorithms for the prediction of diabetes, which is a dysfunction of the blood sugar regulation system, in order to reduce the risks of complications of this chronic disease on the health of the patient. To achieve this goal, we used machine learning algorithms such as Random Forest RF, Logistic Regression RL, K-nearest neighbors KNN and Neural Networks ANN and the data were extracted from Kaggle which is a web platform owned by Google that operates as a community for data scientists and developers. The performance of the classifiers was compared based on the accuracy rate.

**Keywords:** ANN · RF · RL · KNN · IA

## 1 Introduction

Diabetes is a prevalent and serious medical condition with widespread global impact. The severity of the disease is attributed to complications that arise when individuals either neglect to screen for diabetes or fail to receive appropriate care. Common complications


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 89–102, 2025.
https://doi.org/10.1007/978-3-031-81573-7_7

15. Schäfer, M., Nadi, S., Eghbali, A., Tip, F.: Adaptive test generation using a large language model, arXiv preprint arXiv:2302.06527 (2023)
16. Xia, C.S., Wei, Y., Zhang, L.: Automated program repair in the era of large pre-trained language models, In: Proceedings of the 45th International Conference on Software Engineering (ICSE 2023). Association for Computing Machinery (2023)
17. Xia, C.S., Zhang, L.: Conversational automated program repair, arXiv preprint arXiv:2301.13246 (2023)
18. Sobania, D., Briesch, M., Hanna, C., Petke, J.: An analysis of the automatic bug fixing performance of chatgpt, arXiv preprint arXiv:2301.08653 (2023)
19. Feng, S., Chen, C.: Prompting is all your need: Automated android bug replay with largelanguage models, arXiv preprint arXiv:2306.01987 (2023)
20. Sobreira, V., Durieux, T., Madeiral, F., Monperrus, M., Maia, M.A.: Dissection of a bug dataset: Anatomy of 395 patches from defects4j, In: Proceedings of SANER (2018)

reports into formal test case specifications. The promising results, in terms of *executability* (i.e., the test case is syntactically correct), and {*validity*} (i.e., the test case actually makes the program fail), suggest that ChatGPT can " *understand*" the semantics of bug reports. This finding is essential as it opens new research directions with large language models, towards automating test case generation with a human in the loop (for writing bug reports).

**Acknowledgments.** This work was conducted as part of the Artificial Intelligence for Development in Africa (AI4D Africa) program, with the financial support of Canada's International Development Research Centre (IDRC) and the Swedish International Development Cooperation Agency (Sida).

## References

1. Bissyandé, T.F., et al.: IEEE 24th international symposium on software reliability engineering (ISSRE). IEEE **2013**, 188–197 (2013)
2. Lopez, A.: Statistical machine translation. ACM Computing Surveys (CSUR) **40**, 1–49 (2008)
3. Stahlberg, F.: Neural machine translation: A review. Journal of Artificial Intelligence Research **69**, 343–418 (2020)
4. Allamanis, M., Barr, E.T., Devanbu, P., Sutton, C.: A survey of machine learning for big code and naturalness. ACM Computing Surveys (CSUR) **51**, 1–37 (2018)
5. Hu, X., Li, G., Xia, X., Lo, D. Jin, Z.: Deep code comment generation, In: Proceedings of the 26th conference on program comprehension, pp. 200–210. (2018)
6. Goues, C.L., Pradel, M., Roychoudhury, A.: Automated program repair. Commun. ACM **62**, 56–65 (2019)
7. Monperrus, M.: Automatic software repair: A bibliography. ACM Computing Surveys (CSUR) **51**, 1–24 (2018)
8. Gulwani, S., Polozov, O., Singh, R., et al.: Program synthesis, Foundations and Trends®. Programming Languages **4**, 1–119 (2017)
9. Le, X.-B.D., Bao, L., Lo, D., Xia, X., Li, S., Pasareanu, C.: On reliability of patch correctness assessment, In: 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), IEEE, pp. 524–535. (2019)
10. Xiong, Y., Liu, X., Zeng, M., Zhang, L., Huang, G.: Identifying patch correctness in test-based program repair, In: Proceedings of the 40th international conference on software engineering, pp. 789–799. (2018)
11. Anand, S., et al.: An orchestrated survey of methodologies for automated software test case generation. J. Syst. Softw. **86**, 1978–2001 (2013)
12. Taneja, K., Xie, T.: Diffgen: Automated regression unit-test generation, In: 2008 23rd IEEE/ACM International Conference on Automated Software Engineering, IEEE, pp. 407–410. (2008)
13. Thummalapenta, S., Xie, T., Tillmann, N., De Halleux, J., Schulte, W.: Mseqgen: Object-oriented unit-test generation via mining source code, In: Proceedings of the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering, pp. 193–202. (2009)
14. Fraser, G., Arcuri, A.: Evosuite: On the challenges of test case generation in the real world. IEEE sixth international conference on software testing, verification and validation IEEE **2013**, 362–369 (2013)

section will cover some limitations and suggest some research directions for future work to increase the amount of executable and valid test cases.

### 5.1 Threats to Validity

In this study, ChatGPT, an openly available LLMs known to have been trained on public data is exploited to generate test cases for bugs of the Defects4J dataset. As Defects4J is a standard benchmark that is publicly available, it is likely that parts of the dataset have been part of the model's training data, this is a threat to the validity of the results as it reflects a data leakage problem. To address this concern, we performed manual investigations of the generated test cases to ensure their difference with the original tests, confirming ChatGPT's capability of understanding the semantics of the bug report. Still, in further iterations of this work we will assess ChatGPT's performance on newly reported faults. Additionally, due to the randomness of ChatGPT, it is essential to verify that ChatGPT is correctly replying to the given prompt before starting the experiments.

### 5.2 Limitations & Future Work

In this study the *Executability* only reflects if a generated test case is directly executable or not. This doesn't reflect the amount of effort for a human to make it executable. After manually reviewing the generated test cases, it has been shown that most can be made executable through the modification of one or two lines of code. The most common issues are usually: missing imports, duplicate function names, or the use of a deprecated function. Those limitations could systematically be fixed in future work (e.g. with prompt engineering) significantly increasing the amount of executable test cases.

For future iterations of this study we suggest to investigate the potential of ***fine-tuning LLMs*** to translate the informal bug reports into formal test cases with a higher rate of executable and valid test cases.

In our experiments ***bug reports*** where collected and directly used as prompt to demonstrate the feasibility and applicability of our idea to address real software faults reported by the users. Nevertheless, pre-processing the textual data might be beneficial to keep ChatGPT focused on the main context of the bug report, therefore increasing the executability and validity of the generated test case.

To fix a bug, the first step is to reproduce it with a bug-triggering test case which has been proven feasible in our study. To reach to directly fix a newly reported bug, further research should be made on how we can improve and automatically deploy APR tools to not just automatically address the bug but fix it.

## 6 Conclusion

Large language models have recently gained substantial popularity, thanks to the public release of ChatGPT, whose potential as a disruptive technology has been largely advertised. The literature has empirically studied various capabilities of the inner language model for various natural language processing tasks. In this work, we investigate the feasibility of leveraging the inner language of the GPT model to translate informal bug

of some Math project versions is that the test suite is compiling, but their execution is never-ending. Therefore, there were 10 Math bug reports for which we could not determine the *validity* of the generated test cases, which explains the lower validity results for this project. Overall, we got valid test cases for 30% of the bug reports, which is really promising. Among the executable test cases, we observed that 59% were valid. This shows that once the initial *executability* challenge is passed, the tests generated by ChatGPT are in fact valid, highlighting its understanding. Meaning that ChatGPT was able to understand the semantics of the user written bug report and translate them into a bug-triggering test case. These results highly motivate further research in this domain.

**Table 2.** Generation performance for ChatGPT: percentage of bug reports where we successfully generated at least one test case.

| | | Percentage of generation success | | |
|---|---|---|---|---|
| Project | # of bug reports | Overall executability | Overall validity | Validity among executable |
| Chart | 6 | 33% | 17% | 50% |
| Cli | 30 | 53% | 37% | 69% |
| Closure | 127 | 46% | 28% | 59% |
| Lang | 60 | 60% | 43% | 72% |
| Math | 100 | 43% | 15% | 35% |
| Time | 19 | 84% | 68% | 81% |
| Total | 342 | 50% | 30% | 59% |

Further manual investigations also highlighted that *executability* and *validity* can in most cases be fixed with minor modifications (e.g., adding relevant imports or changing duplicated function names).

**Findings**: The experimental results based on ChatGPT APIs show that a large language model (LLM) can take bug reports as inputs and produce test cases that are executable in 50% of cases. Beyond *executability*, about 30% of the bugs could be reproduced with valid test cases. Specifically, over half (59%) of the executable test cases were valid test cases.

These results, which are based on an off-the-shelf LLM as-a-service, show promises for automated test case generation, beyond unit testing, leveraging complex information from user-reported bugs.

## 5 Discussion

Overall, our empirical study validates the hypothesis that ChatGPT can "understand" bug reports: given a bug report, it can extract its semantics and translate them into formal test cases. However, some challenges that should be addressed in the future remain. This

The instruction is unique for all queries to ChatGPT and is as simple as follows: "{*write a Java test case for the following bug report:*}". For the bug report, our feasibility study considers that no pre-processing should be applied on the bug report, and the information should not include follow-up comments or attachments. For every prompt, we request the ChatGPT five (5) times and assess the different generated test cases. In practice, before running the generated test cases, the ChatGPT outputs are parsed to clean them from natural language texts (e.g., explanations) that would lead to compilation failures. Afterwards, the test cases are systematically included in the test suite, which is fully executed by the Defects4J test pipeline. Execution results are then logged, allowing us to compute the metrics of executability and validity.

## 4 Results

Figure 1 provides an illustrative example of a bug report (from the CLI project) and the associated formal test cases (ground truth in Defects4J and generated from ChatGPT). As we can see in this example, ChatGPT is able to reproduce a formal test case from a bug report, which can enable various software automation tasks, such as spectrum-based fault localization, patch validation in program repair, and more generally automated software testing.


**Fig. 1.** Example of bug report and associated test cases - Cli 17 from Defects4J

On the Defects4J dataset, we compute the proportion of bug reports for which ChatGPT is able to successfully generate test cases.

Table 2 summarizes the metrics. On average, executable test cases were obtained for 50% of the bug reports across all projects. The validity of the generated test cases varies greatly from one project to another, which can be explained by the different source and format of user-written bug reports. Unfortunately, a commonly known issue

## 3 Experimental Setup

In this section we overview the settings under which we assess the capability of a Large Language Model to translate informal bug reports from real software projects into formal test case specifications that reproduce the buggy behavior. In particular, we present the benchmark, the metrics as well as the experimental design.

### 3.1 Benchmark

We consider the **Defects4** repository [20], which includes real-world faults from various Java software development projects as enumerated in Table 1. We collect the bug reports associated to these faults. One must mention that not all bug reports were available, and some bugs referred to the same bug report, in that case it was only considered once to avoid bias in the results because of duplicates. We considered Defects4J due to its wide adoption in the software testing and software research communities.

**Table 1.** Java Projects from Defects4J used in the study.

| | Projects | | | | | |
|---|---|---|---|---|---|---|
| | Chart | Cli | Closure | Lang | Math | Time |
| #Faults | 26 | 39 | 174 | 64 | 106 | 29 |
| #Faults associated to Bug Reports | 6 | 30 | 127 | 60 | 100 | 19 |

### 3.2 Metrics

As introduced, the goal of the study is to assess whether ChatGPT can generate test cases from bug reports. Therefore, our evaluation is focused on measuring the quality of the generated test cases. We thus consider two main metrics:

- ***Executability***: a ChatGPT-generated test case may not even be syntactically correct to be compiled and executed. While often, the generated test case can be made executable after manually implementing small edits (e.g., adding relevant imports), we conservatively consider that executability is a binary metric, and is automatically computed once ChatGPT outputs are yielded (with no manual changes added).
- ***Validity***: an executable test case may or may not fail on the target buggy program. We follow the convention of patch validation in program repair and consider the generated test case to be valid only when it, indeed, fails on the buggy program. Otherwise it is considered as invalid.

### 3.3 Experimental Design

We rely on the ChatGPT API (version 3.5) for our experiments. We construct the prompt by concatenating two pieces of information: the instruction and the bug report.

With recent advances in natural language processing techniques, such as the advent of large language models (LLMs), a wide range of tasks have seen machine learning achieve, or even exceed, human performance. Machine translation [2, 3], in particular, has been a very active field where several case studies have been explored beyond language translation. For example, in software engineering, several research directions have investigated the feasibility of leveraging natural language inputs for producing programming artefacts and vice-versa. Some milestones have been recorded in the literature in code summarization [4, 5], program repair [6, 7], and even program synthesis [8]. Nevertheless, bug reports have scarcely been explored. Yet, automating bug reproduction via analysis of bug reports holds tremendous value. In this work, we propose to *study the feasibility of exploiting an LLM for reproducing bugs*. We focus on ChatGPT, which has recently received much attention and presents the advantage that its model has been trained on a large corpus of natural language text as well as source code of software programs.

But *can ChatGPT understand bug reports*? We consider the management of bug reports as an example case where machine learning can be helpful while keeping the human in the loop. "Understanding bug reports" suggests the eventual possibility of reproducing the reported bug. Our prompt is therefore focused on requesting ChatGPT to exploit a bug report's textual content (in natural language) and generate a formal test case (in a programming language). We assume that if ChatGPT can generate a test case that not only is executable but also fails on the associated buggy program version, then ChatGPT may have "understood" (in the sense of "*captured the semantics of unwanted execution behavior reported by the user*"). This assumption is, obviously, an over-approximation of the relevance of the generated test case since the generated failing test case may be based on random inputs that are irrelevant to the reported bug. Nevertheless, it would constitute a first milestone towards automatic test case generation based on user inputs, which reflects realistic and complex user experience.

## 2 Related Work

To address a bug, the first step is to understand it and reproduce it. To demonstrate that there is a bug, the developer has to write a bug-triggering test case since the original test suites are usually scarce and incomplete [9, 10]. Several techniques [11–14] have been developed to help developers with this very time-consuming process. However, these techniques mostly rely on formal specifications.

Recently, some work on test case generation using large language models (LLM) was done by [15] but their TestPilot still requires the functions signature and implementation as prompt. The use of ChatGPT to directly enhance Automated Program Repair (APR) techniques [16–18] highlights even more the potential of this new LLM. Feng et al. [19] have investigated ChatGPT's ability to help developers reproduce the bug while extracting important steps to reproduce the bug from the bug report. However they did not perform any test case generation. Additionally, they still require a human to actually reproduce the bug. In contrast, in our study, we are aiming at using the unprocessed human written bug report as direct input for test case generation, enabling automatic bug reproduction with only a human-in-the-loop to report the bug but not to address it.


# Leveraging Conversational AI for Accelerating User-Driven Software Testing

Aminata Sabané[1,2], Laura Plein[3], and Tegawendé F. Bissyandé[1,2,3](✉)

[1] CITADEL Interdisciplinary Excellence Centre in Artificial Intelligence for Development, Ouagadougou, Burkina Faso
aminata.sabane@ujkz.bf, tegawende.bissyande@citadel.bf
[2] Université Joseph Ki-Zerbo, Ouagadougou, Burkina Faso
[3] SnT, Université du Luxembourg, Luxembourg, Luxembourg
laura.plein@men.lu

**Abstract.** This work addresses a research challenge in automating the translation of natural language inputs into programming language specifications. We consider the case of bug reports, which are informally written by users, and that must be specifying into executable test cases for reproducing the bug on the target software. Software bugs are indeed largely reported in natural language by users. Yet, we lack reliable tools to automatically address reported bugs (i.e., enabling their analysis, reproduction, and bug fixing). We therefore build on the recent promises brought by ChatGPT for various tasks, including in software engineering, and establish the following research question: *What if Conversational Artificial Intelligence (AI) models could be used to explore the semantics of bug reports as well as to automate their reproduction?* We evaluate the capabilities of ChatGPT, a state-of-the-art conversational AI, i.e., chatbot, using the popular Defects4J benchmark with its associated bug reports. The results reveal that ChatGPT can generate executable test cases that could trigger 50% of the bugs reported in natural language. These results are promising not only for the research community, but also for practitioners.

**Keywords:** ChatGPT · Debugging · Translation · Test cases · Bug reports

## 1 Introduction

Software users are expected to provide feedback on their experience in running programs. Such feedback often leads to various improvements by developers responding to feature requests and bug reports. In this respect, development platforms, such as GitHub, offer tool support for collecting reports and continuously monitoring how developers address them. Unfortunately, various studies have shown that bug reports are under-exploited [1]. Recurrently, indeed, researchers and practitioners point to the general quality of such reports: developers put much effort to "understand" and reproduce the potential bugs that are reported; researchers struggle to build tools for automatically capturing the semantics of the natural language text and transforming them into actionable inputs for existing (testing) frameworks.


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 81–88, 2025.
https://doi.org/10.1007/978-3-031-81573-7_6

# Artificial Intelligence

11. Hahn, C.: A Domain Specific Modeling Language for Multiagent Systems, In: présenté à 7th Int. Conf. on Autonomous Agents and Multiagent Systems, Estoril, Portugal, pp. 233–240. (2008)
12. Santos, F., Nunes, I., Bazzan, A.L.C.: Supporting the Development of Agent-Based Simulations: A DSL for Environment Modeling, In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), pp. 170–179. (2017). https://doi.org/10.1109/COMPSAC.2017.224
13. Phung, S., et al.: An Agent-Based Modeling and Simulation Environment for Dynamic Biological Systems, In: Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP), The Steering Committee of The World Congress in Computer Science, Computer p. 349. (2012)
14. Pavón, J., Gómez-Sanz, J.: Agent oriented software engineering with INGENIAS, In: International Central and Eastern European Conference on Multi-Agent Systems, Springer, pp. 394–403 (2003)
15. Gago, I.S.B., Werneck, V.M., Costa, R.M.: Modeling an educational multi-agent system in maSE. In: International Conference on Active Media Technology, Springer, pp. 335–346 (2009)
16. Wooldridge, M., Jennings, N.R., Kinny, D.: The Gaia methodology for agent-oriented analysis and design. Auton. Agents Multi-Agent Syst. **3**, 285–312 (2000)

The overall proposition takes account of the model modelling and implementation, model simulations, data management and visualization in relation with the agroecosystem domain. The overall proposition allows for agroecosystem experts to easily design agroecosystem models.

In terms of the way forward, a programming language will be proposed for agroecosystem-oriented model development. In addition, the detail design and implementation of the independent platform will be done. Finally, a methodology will be proposed as a guideline for the design and simulation of agroecosystem models.

## References


1. Jahel, C. : Analyse des dynamiques des agroécosystèmes par modélisation spatialisée et utilisation d'images satellitaires, cas d'étude de l'ouest du Burkina Faso. PhD Thesis, AgroParisTech (2016)
2. Kremmydas, D., Athanasiadis, I.N., Rozakis, S.: A review of Agent Based Modeling for agricultural policy evaluation. Agric. Syst. **164**, 95–106 (2018). https://doi.org/10.1016/j.agsy.2018.03.010
3. Belem, M., Saqalli, M.: Development of an integrated generic model for multi-scale assessment of the impacts of agro-ecosystems on major ecosystem services in West Africa. J. Environ. Manage. **202**, 117–125 (2017). https://doi.org/10.1016/j.jenvman.2017.07.018
4. Belem, M., Bazile, D., Coulibaly, H.: Simulating the Impacts of Climate Variability and Change on Crop Varietal Diversity in Mali (West-Africa) Using Agent-Based Modeling Approach. Journal of Artificial Societies and Social Simulation **21**(2), 8 (2018). https://doi.org/10.18564/jasss.3690
5. Grignard, A., Taillandier, P., Gaudou, B., Vo, D.A., Huynh, N.Q., Drogoul, A.: GAMA 1.6: Advancing the Art of Complex Agent-Based Modeling and Simulation, In: PRIMA 2013: Principles and Practice of Multi-Agent Systems, Boella, G., Elkind, E., Savarimuthu, B.T.R., Dignum, F., Purvis, M.K. (eds.), Lecture Notes in Computer Science. pp. 117-131. Berlin, Heidelberg: Springer, (2013). https://doi.org/10.1007/978-3-642-44927-7_9.
6. Bousquet, F., Bakam, I., Proton, H., Le Page, C.: Cormas: Common-pool resources and multi-agent systems. In: Tasks and Methods in Applied Artificial Intelligence, Pasqual Del Pobil, A., Mira, J., Ali, M., (eds.), in Lecture Notes in Computer Science, vol. 1416, pp. 826–837. Berlin, Heidelberg: Springer Berlin Heidelberg (1998). https://doi.org/10.1007/3-540-64574-8_469
7. Lytinen, S.L., Railsback, S.F.: The Evolution of Agent-based Simulation Platforms: A Review of NetLogo 5.0 and ReLogo
8. Czarnecki, K., O'Donnell, J.T., Striegnitz, J., Taha, W.: DSL Implementation in MetaOCaml, Template Haskell, and C++ , In: Domain-Specific Program Generation: International Seminar, Dagstuhl Castle, Germany, March 23-28, 2003. Revised Papers, Lengauer, C., Batory, D., Consel, C., Odersky, M., (eds.), in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, p. 51-72. (2004). https://doi.org/10.1007/978-3-540-25935-0_4.
9. Gérard, S., Dumoulin, C., Tessier, P., Selic, B.: 19 Papyrus: A UML2 Tool for Domain-Specific Language Modeling, In: Model-Based Engineering of Embedded Real-Time Systems, Springer, Berlin, Heidelberg, pp. 361–368. (2010). https://doi.org/10.1007/978-3-642-16277-0_19
10. Nassar, M., et al. : Vers un profil UML pour la conception de composants multivues, Rev. Sci. Technol. Inf.-Sér. Obj. Logiciel Bases Données Réseaux, 11(4), (2005)

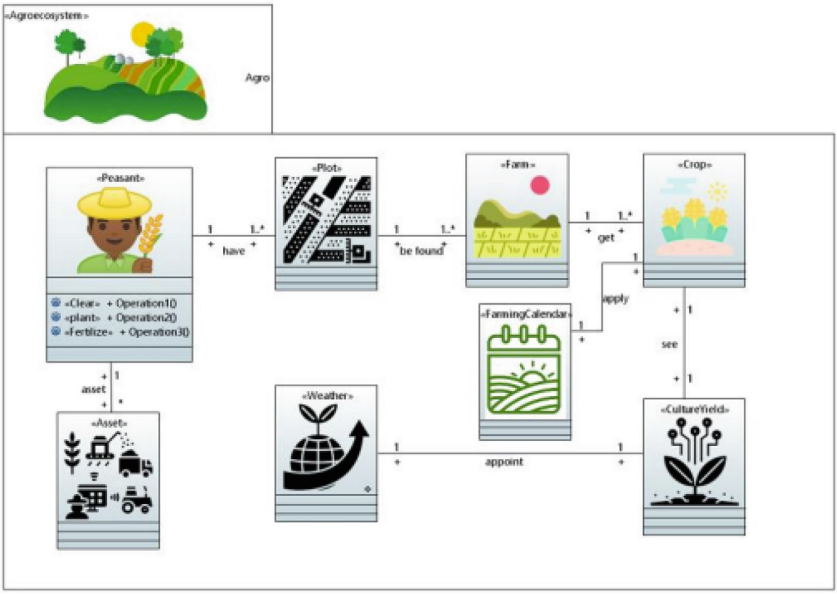



**Fig. 4.** Application of the graphical notation language in the modelling of an agroecosystem

None of these propositions take account of agroecosystem modelling and simulation. They do not propose read-to-use concepts in the agroecosystem domain. In addition, although these previous studies propose a high level abstract syntax, they do not propose any methodology for MAS and ABM modelling and development.

A range of methodologies [14–16] propose guidelines for MAS and ABM design. Furthermore, they propose an abstract syntax for multi-agents and ABM development. However, these methodologies do not propose read-to-use concepts for agroecosystem modelling and simulation, making difficult their use by an agroecosystem expert.

This work concerns the development of DSL for agroecosystem-oriented models creation and simulation. The main aim is to develop read-to-use concepts for modelling and simulating agroecosystems. The abstract syntax and graphical notation language developed allows for an agroecosystem expert to design easily agroecosystem models. At this current stage, however, account is not taken of the code generation and simulation. This aspect will be delved into in future work.

## 5 Conclusion

The main objective of this study was to define a basis for the development of an independent platform for modelling and simulating agroecosystem-oriented models. Originally, we aimed to propose a meta-model, a UML profile, a domain specific modelling language and a general architecture. To this end, we used an approach based on DSL and agent-based modelling.

**Table 1.** Graphical notation of the different concepts

| Stereotypes (concepts) | Meta-class | Graphical notation |
|---|---|---|
| **Household** | Agent | |
| **Agroecosystem** | Package | |
| **Farm** | Class | |
| **Plot** | Cell | |
| **Cropping System** | Class | |
| **Crop** | Class | |
| **Farming Calendar** | Agent | |
| **Climate Agent** | Class | |
| **Weather** | Class | |
| **Farmer** | Role | |
| **Breeder** | Role | |
| ••• | ... | ... |

## 4 Discussion

The development of DSL in the multi-agents system and agent-based modelling is not new. A range of works have already been carried out in this domain. In this regard, [11] proposed a domain specific modelling language for multi-agent systems that provides a clear syntax and semantics to define agent-based systems in a graphical visualized manner. However, this DSL has been proposed for general purpose in the multi-agent system domain. The use of this DSL is difficult for non-expert multi-agents, particularly in a specific domain such as agroecosystem. [12] proposed a DSL to target the agent-based modelling and simulation domain with a focus on modelling the simulated environment. In order to allow biologists to create and simulate models easily, [13] proposed GRANITE (Genetic Regulatory Analysis of Networks Investigational Tools Environment), for large-scale model simulation in the domain of biology.

### 3.2 UML Profiles

In order to allow for graphical modelling and code generation, a UML profile has been built respecting the meta-model. Each sub-domain of the meta-model is then represented as a UML package. ***Simulation, Agent, Agroecosystem, Persistence*** and ***Visualization*** packages are considered as stereotypes extending the UML package. Each concept of the meta-model has later been considered as a stereotype. In this way, the concept of ***Model, Simulation, Scenario, Parameter, Output*** extends the UML Class. In the agent package, the concept of ***Agent, Role, Capability, Resource*** and ***Location*** extend the UML class. As for the agro-ecosystem package, ***Household***, ***Climate*** and ***Animal*** extend the ***Agent*** stereotype. In the same package, ***Farmer***, ***Breeder*** and ***Forestry*** extend the ***Role*** concept. Finally, ***Clear***, ***Plant***, ***Cultivate***, ***Transport***, ***Stock***, ***Sell***, ***Buy*** extend the ***Capability*** concept.

### 3.3 Graphical Notation Language for Agroecosystem Modelling

Based on the proposed profile, a graphical notation language for agroecosystem modelling has been proposed. The graphical language built in this study concerns only the agroecosystem package and the related UML profile. To each main concept of the package, a visual representation has been associated. Table 1 provides the description of some visual representations used in this study, while Fig. 4 provides an example of the use of graphical language.

### 3.4 Architecture of the Platform

A multi-layer architecture has been proposed for the platform development. This architecture describes the overall structure and functionalities of the platform. The main layers of the architecture are:

**View layer:** provides interfaces to manage interactions between users and the frameworks for models development and simulation and to visualize the simulation outputs.

**Modelling layer:** provides tools and utilities for graphical modelling and model implementation using a DSL. it contains as principal components the DSML and a DSL. it should also allow for code generation.

**Simulation layer:** provides tools and utilities for model simulation. it should implement the simulation, agent and agroecosystem sub-domains of the meta-model.

**Utilities layer**: provides tools and utilities for data management and data visualization. This layer should help to manage interaction with different formats of data supports (database, GIS data, CSV, txt, JSON, XML, etc.) and to visualize data (map, chart, file, video, etc.). It implements the persistence and visualization sub-domains of the meta-model.

***Plot*** is a kind of *Cell* characterized by a land cover type. In addition, a ***Household*** applies a ***Cropping System*** characterized by a set of cultivated ***Crops***. A ***Crop*** represents a crop species characterized by its name and yields under different climate conditions and a ***Farming Calendar***. A ***Farming Calendar*** is a set of farming operations performed at a specific period of the year.

The ***Climate*** agent computes the climate condition. It is characterized by different climate conditions and their probability distribution.

In this system, the ***Household*** agent plays the following roles: ***Farmer***, ***Breeder*** and ***Forestry*** **role**. In terms of ***Capabilities***, a ***Household*** can, among other things, ***Clear***, ***Plant***, ***Cultivate***, ***Transport***, ***Stock***, ***Sell***, ***Buy***, etc. The ***Climate*** agent plays the ***Weather*** role.

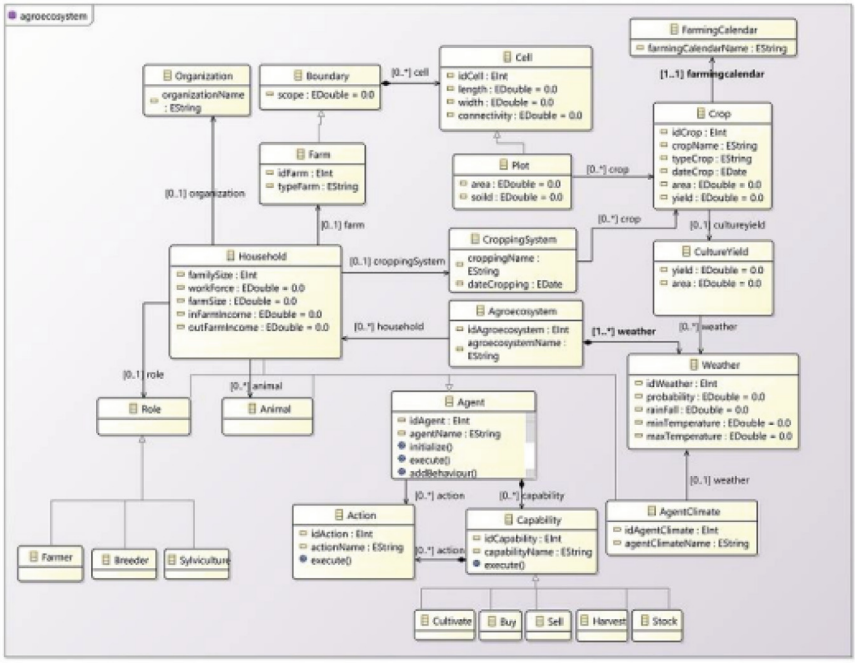



**Fig. 3.** The meta-model of agro-ecosystem sub-domain

**"Persistence" Sub-Domain**

The «Persistence» sub-domain covers the collection, storage and archiving of data, as for example, the interactions between models and data sources, such as ontologies, DSLs, GIS and Big Data.

**"Visualization" Sub-Domain**

The «Visualization" sub-domain is responsible for simulation results visualization.

**"Simulation" Sub-domain**

The « Simulation» sub-domain describes the core concepts for a general description of simulation models. Then, a **Model** is characterized by its name, a set of ***Agents***, ***Scenarios***, a set of ***Output*** variables. A ***Model*** is used to achieve Simulation. A Simulation is characterized by a range of ***Scenarios***. A ***Scenario*** is characterized by a set of parameters, each parameter being used to initialize a ***Variable***. An ***Output*** variable is related to a mode of ***Visualization***.

**"Agent" Sub-domain**

The agent sub-domain provides the description of the core concepts for representing a multi-agents system (Fig. 2). In this sub-domain, An ***Agent*** is characterized by a set of ***Roles***, ***Capabilities***, ***Acquaintances*** and a ***Location*** in *the* ***Environment***. A ***Role*** defines the behavior of an agent in the system and a ***Capability*** defines a task that an agent can achieve. An ***Environment*** is a cellular automaton characterized by a set of ***Cells***.

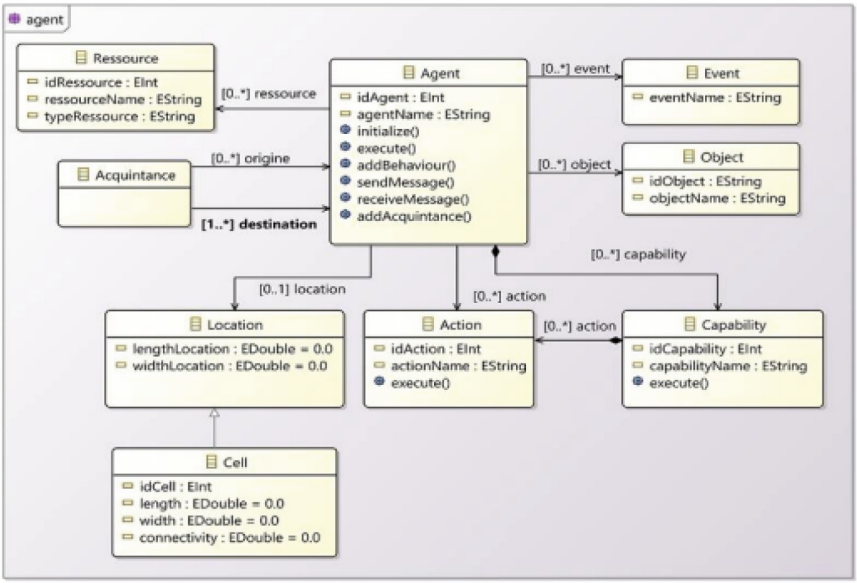



**Fig. 2.** The meta model of agent sub-domain

**"Agroecosystem" Sub-Domain**

The agroecosystem sub-domain describes the core concepts required to represent an agroecosystem based on agent-based point of view (Fig. 3). The agrosystem sub-domain is an extension of the agent sub-domain.

An ***Agroecosystem*** is a model characterized by three types of agent: ***Household***, ***Climate*** and ***Animal*** agents. A ***Household*** has a ***Farm*** characterized by a set of ***Plots***. A

### 2.5 Fifth Step: DSML Development

Using the meta-model and UML profile previously developped, a DSML has been developped. The DSML proposes a graphical notation language for agro-ecosystem oriented model design.

### 2.6 Sixth Step: Architecture Analysis

Finally, a multi-layer architecture is proposed for the development of the independent platform. It proposes an overall description of the framework structure and behavior.

## 3 Results

As regards the main results, this study allowed to develop an abstract syntax language, a UML profile, a graphical notation language and to propose an architecture for the development of an independent platform. These results are described in the following section.

### 3.1 Abstract Syntax Language

For a better representation of the model, our system has been organized under five sub-domains: “Persistence”, “Simulation”, “Agent”, “Agroecosystem” and “Visualization” sub-domains (Fig. 1). For each sub-domain, the main concepts have been identified and a meta-model developed.

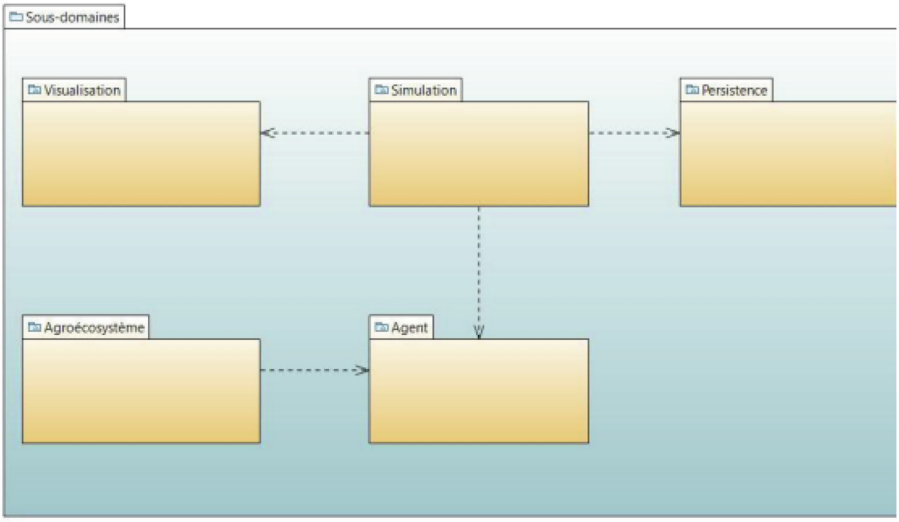



**Fig. 1.** The different parts of the meta-model

taken at how an independent platform could be put in place for agroecosystem-oriented models development and simulation. Through this platform, we aim to automatize the development and simulation of models. In addition, this platform should make it possible to interact with external data sources (GIS, database, CSV file, Excel file, etc.) and to visualize simulation results. Specifically, the platform should include:

- a graphical language for agroecosystem-oriented model design;
- a domain specific language as programming language;
- tools for visualizing simulation results;
- and tools for models simulations.

### 2.2 Second Step: Domain Analysis

The main objective of this step is to understand the domain under study through the identification of the main concepts of study. Drawing from the study specification, the literature review and expert judgment, the domain analysis step helped to identify the main concepts. These have been identified to take account of the model simulation and visualization, agroecosystem models development using an agent-based modelling approach.

### 2.3 Thirst Step: Development of the Meta-model

Based on the main concepts previously identified, a meta-model has been developped as follows: To start with, we defined the abstract syntax of our language through the delineation of meta-models, using domain concepts (i.e. domain vocabulary) and the relationships between these concepts. To this end, we used the papyrus tool [9].

### 2.4 Fourth Step: Creating the UML Profile

Based on the meta-model, a UML profile has been created. A papyrus tool [8] was used at this stage. Firstly, we created i) stereotypes using the domain concepts, ii) the list of tagged values and iii) constraints that can be expressed in OCL (Object Constraint Language) on stereotypes and tagged values.

The UML profile is an extension of UML meta-model for a specific domain. In fact, UML offers a number of writing rules or standardized graphical representations, as well as common mechanisms or concepts applicable to all diagrams. Certain elements, such as stereotypes, are specifically designed to ensure the adaptation and evolution of the notation, mainly to take into account the particularities of the different situations to be modelled. Compared to the standard UML formalism, developers often wish to add extra characteristics to consider certain specificities of their application domain [10]. But UML does not meet all needs, because there are few tools for analysing specifications and not enough concepts from certain domains. To meet this need, UML features an extensibility mechanism based on stereotypes, tagged values and OCL constraints. This extension mechanism makes it possible to particularize the UML meta-model to accommodate specific modelling needs [10].

interacting at different scales of description. In an ABM, the agent has its own decision model. The agent can perceive its environment and acts on it, which in turn may impact its decision-making process. These ABM characteristics are relevant to represent the agroecosystem. For example, [3] used agent-based model to represent the impact of cropping systems on soil carbon sequestration and household income. [4] used multi-agents system to represent the impacts of climate change and variation on crop diversity management by the farmers.

However, although agro-ecosystem modelling is becoming more and more important, a specific modelling framework is missing not only for modelling agroecosystems but also for their simulation. Currently, modellers use general modelling and simulation platforms, such as Gama [5], Cormas [6], NetLogo [7], etc. that non-modellers and non-agent experts find difficult to use.

The objective of this study is to propose an independent platform for agent-based modelling and simulation of agroecosystems. Specifically, this study, the first step in developing the general framework, aims to propose a meta-model of agroecosystem, a graphical notation language for agroecosystem-oriented modelling and to design the architecture of the platform.

In this study, we used an approach based on domain-specific language (DSL) and agent-based modelling. A DSL is a language offering expression power focused on a particular domain, such as a specific class of applications or an aspect of the application [8]. A DSL is a programming language for a specific domain. It is opposite to classical programming language, such as Java, C++ and C#, used for a range of domains. In this study, we are using DSL in order to propose a new programming language for agroecosystem-oriented model development. The objective is to overcome the difficulties related to the development of a model using general programming language and non-dedicated programming language.

As a result, a meta-model is proposed as the DSL abstract syntax. Thereafter, the language has been concreted by proposing a graphical notation language. Finally, a multi-layer architecture is proposed. The overall proposition takes account of the model design, simulations and their visualization.

This paper is organized as follows: The second section describes the general approach and tools used in the study. The last section describes the main results arrived at in the study. Specifically, this section describes the meta-model for agro-ecosystem modelling, the UML profile and the DSML. Finally, the results are discussed.

## 2 Materials and Method

Our methodology involves six steps, including the study specification, the domain analysis, the UML profile development, the DSML development and the architecture design.

### 2.1 First Step: Specification

The first step aims is to define the scope of our study. Specifically, the objective was to define the goal of the platform and the intended use of the platform. Then, a look is

# A Domain Specific Language (DSL) for Agroecosystems Modelling and Simulation

Jean-Armand Yanogo[1,2(✉)], Mahamadou Belem[1,2], Toundé Mesmin Dandjinou[1,2], Saïd Cham's Nour Ougda[1,2], and Theodore Marie Yves Tapsoba[1,2]

[1] Université Nazi Boni, Bobo-Dioulasso, Burkina Faso
Jeanarmand_yanogo@yahoo.fr
[2] Laboratoire d'Algèbre, de Mathématiques Discrètes et d'Informatique (LAMDI), Université Nazi Boni, Bobo-Dioulasso, Burkina Faso

**Abstract.** Modelling agroecosystems is a complex process that implies understanding the interactions between the different elements of the system. However, although agroecosystem modelling is becoming more and more important, a specific modelling framework is missing not only for the modelling of agroecosystems but also for their simulation. Currently, modellers use general modelling and simulation platforms that non-modellers find difficult to apply. Consequently, a specific framework for agroecosystem modelling and simulation is required. This study intends to propose a basis for the development of an independent platform for the creation and simulation of agroecosystem-oriented models. Specifically, the objectives of this paper are to achieve the requirement analysis and the domain analysis, to propose a domain specific modelling language and to design the platform architecture. Using a model-driven engineering and an agent-based modelling approach, a meta-model has been proposed as the abstract syntax of the language. Thereafter, the language has been concreted by proposing a graphical notation language. Finally, a multi-layer architecture has been proposed. The overall proposition takes account of the model development, simulations and their visualization. The general framework will be developed as part of the next steps.

**Keywords:** agroecosystem · simulation · agent-based model · meta-model · domain specific language · domain specific modelling language

## 1 Introduction

Agroecosystems are cultivated ecosystems composed of abiotic and biotic elements that interact with each other in an agricultural, pastoral and forestry-type space, modified by man for the purpose of food production [1]. Agroecosystems are characterized by a dynamic and structural complexity coming from the interaction between the ecological and socio-economic processes in which they are integrated. So modelling agroecosystems is a complex process that implies understanding the interactions between the different elements of the system. Agent-based models (ABMs) are highly relevant for representing and modelling agro-ecosystems [2]. ABMs comprise a set of agents


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 68–78, 2025.
https://doi.org/10.1007/978-3-031-81573-7_5

15. Jiang, H., Liang, Y.: Online path planning of autonomous Drones for bearing-only stand-off multi-target following in threat environment. IEEE Access **6**, 22531–22544 (2018)
16. Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J.: Multi agent cooperation and competition with deep reinforcement learning. PLoS ONE **12**, 6379–6390 (2017)
17. Wu, C., et al.: DRONE autonomous target search based on deep reinforcement learning in complex disaster scene. IEEE **7**, 117227–117245 (2019)
18. Hidayatno, A., Destyanto, A.R., Hulu, C.A.: Industry 4.0 technology implementa-tion impact to industrial sustainable energy in Indonesia: A model conceptualization. Energy Procedia **156**, 227–233 (2019)
19. Singh, V., Misra, A.K.: Detection of plant leaf diseases using image segmentation and soft computing techniques. Information processing in Agriculture **4**(1), 41–49 (2017)
20. Suganthi, J., Suganthi, V., Giridharan, S.: Detection and Prevention Mechanism of Snakes and Insects Biting from Farmers using IOT Monitoring System. Open Access Quarterly International Journal **2**(1), 298–30 (2018)
21. Vaca-Castano, G., Driggers, R., Furxhi, O., Arvidson, C., Mazzotti, F.: Multispectral camera design and algorithms for python snake detection in the Florida Everglades. In Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imagery XXV **10986**, 272–279 (2019)
22. Implications for herpetology and global health: Durso A.M., Moorthy G.K., Mohanty S.P., Bolon I., Salathé M., and Ruiz de Castañeda R. Supervised learning computer vision benchmark for snake species identification from pho-tographs. Frontiers in Artificial Intelligence **4**, 582–110 (2021)
23. Bandala, A.A., Dadios, E.P., Vicerra, R.R.P., Lim, L.A.G.: Swarming algorithm for unmanned aerial vehicle (drone) quadrotors–swarm behavior for aggregation, foraging, for-mation, and tracking. Journal of Advanced Computational Intelligence and Intelligent Infor-matics **18**(5), 745–751 (2014)
24. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. Journal of internet services and applications **1**(1), 7–18 (2010)
25. Jenkins, B.: Watching the watchmen: Drone privacy and the need for oversight. Ky. LJ **102**, 161 (2013)
26. Rahman, A., Jin, J., Wong, Y.W., Lam, K.S.: November. Development of a cloud-enhanced investigative mobile robot. In 2016 International Conference on Advanced Mechatronic Systems (ICAMechS), pp. 104-109 (2016)
27. Simelane, P.T., Kogeda, O.P., Lall, M.: A cloud computing augmenting agricultural activities in marginalized rural areas: A preliminary study. In: 2015 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), pp. 119-124 (2015)
28. Li, C., Sun, X., Cai, J.: Intelligent mobile drone system based on real-time object detection. Journal of Artificial Intelligence **1**(1), 1 (2019)
29. Fao, I., W UNICEF:. The state of food security and nutrition in the world. Rome, Italy: Food and Agriculture Organization of the United Nations. (2017)
30. WORLD HEALTH ORGANISATION: Snakebite Envenoming. Available at https://www.who.int/news-room/fact-sheets/detail/snakebite-envenoming. (2019)

## 9 Conclusion

In this paper, we have managed to identify common snakes that are mostly found in MRAs and what type of agriculture is mostly practiced. We have also try to discover mechanisms mostly used to prevent snakebites, killing of snakes, and reasons for killing them in MRAs. We, therefore, outline that the proposed cloud computing model that augments the use of ICT to improve agriculture as an activity in MRAs will be of great help since people living in MRAs seem to be neglected and suffering from snakebites and envenoming which leads to them to dissert farms. Farmers' disserting farms lead to low yields or food security issues in MRAs. The proposed cloud architecture would use Drones to track and detect snakes in farms. Farmers could observe everything on their farms via phones connected with drones. This will improve food security, safe farming, and balance biodiversity.

## References

1. Nugroho, A.D.: Agricultural market information in developing countries: A literature re-view. Agricultural Economics **67**(11), 468–477 (2021)
2. Citroni, R., Di Paolo, F., Livreri, P.: A novel energy harvester for powering small UAVs: Performance analysis, model validation and flight results. Sensors **19**(8), 1771 (2019)
3. Shaikh, F.B., Haider, S.: Security threats in cloud computing. Internet Technology and Secured Transactions (ICITST), In: 2011 International Conference for (2011)
4. Kamei, K.: Cloud networked robotics. IEEE Network **26**(3), 28–34 (2012)
5. World Health Organisation: Snakebite Envenoming. (2019) Available at https://www.who.int/news-room/fact-sheets/detail/snakebite-envenoming, last accessed 2021/17/05
6. Naik, S., Khuntdar, B.K., Mohanta, M.P., Mondal, S.: A clinico-epidemiological study of snakebite among children in a rural medical college from eastern India. Interna-tional Journal of Pediatrics and Neonatology **1**(1), 11–14 (2020)
7. Joshi, N.P.: Ecological and ethnobotanical values of weeds found in the spring rice fields in Chitwan. Nepal. Ethnobotany Research and Applications **22**, 1–19 (2021)
8. Cruz, L.S., Vargas, R., Lopes, A.A.: Snakebite envenomation and death in the develop-ing world. Ethnicity & disease **19**(1), 42 (2009)
9. Wood, D., Sartorius, B., Hift, R.: Classifying snakebite in South Africa: Validating a scoring system. South African Medical Journal **107**(1), 46–51 (2017)
10. Wood, D., Sartorius, B., Hift, R.: Estimating the burden of snakebite on public hospi-tals in KwaZulu Natal. Wilderness & Environmental Medicine **27**(1), 53–61 (2016)
11. Sharma, R., Kamble, S.S., Gunasekaran, A.: Big GIS analytics framework for ag-riculture supply chains: A literature review identifying the current trends and future perspec-tives. Computers and Electronics in Agriculture **155**, 103–120 (2018)
12. Ju, C., Son, H.I.: Multiple UAV systems for agricultural applications: Control, implementation, and evaluation. Electronics **7**(9), 162 (2018)
13. KEHOE B., KAHN G., MAHLER J. Autonomous multilateral debridement with the raven surgical robot. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1432-1439. IEEE (2014)
14. Castiblanco, C., Rodriguez, J., Mondragon, I., Parra, C., Colorado, J.: Air drones for explosive landmines detection. In: ROBOT2013: First Iberian Robotics Conference, Springer. Colombia:Bogota, pp. 107-114. (2014)

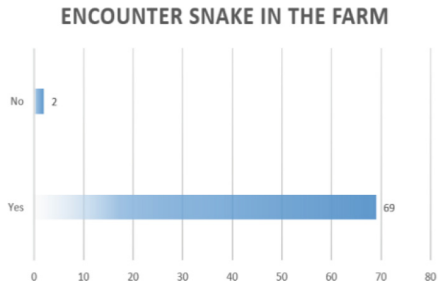



**Fig. 7.** Farmers who encountered snakes in the farm

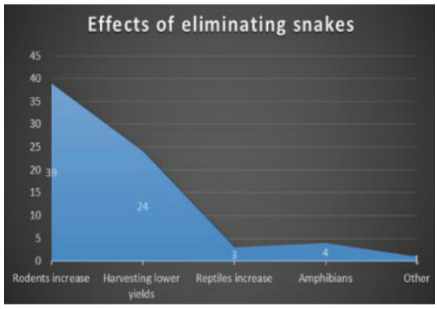



**Fig. 8.** Effects of eliminating snakes

In Fig. 7 97% of the farmers agree that they encounter snakes on their farms. Farmers eliminate or remove 73% of these snakes, and 67% of the snakes fight back, which may result in hospitalization or fatality. Farmers being killed by snakes lead to a reduction in productivity (since some productive and knowledgeable farmers die from snake bites), and it also leads to the loss of livestock. With more frequent snake bites in MRAs, farmers surrender farms to snakes leading to fewer farms to grace in. Some-times farmers may use fire to clear such bushy areas leading to the loss of animals, trees, ecosystems, etc. Snakes also get killed on farms in rural areas, which creates an imbalance in biodiversity. While these snakes are eliminated, 54% of the rodents impact their agricultural products, as presented in Fig. 8.

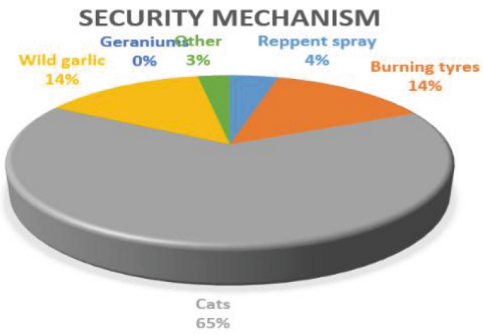



**Fig. 9.** Security mechanisms to curb snakes

As shown in Fig. 9 farmers use different strategies to curb snakes in their farms, however the above-presented security mechanism is not able to the handle death rate of both farm workers and snakes on farms. The most common strategy or method used by farmers in farms to get rid of snakes is cats, which is at 65%.

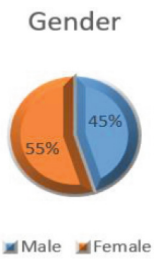



**Fig. 2.** Gender

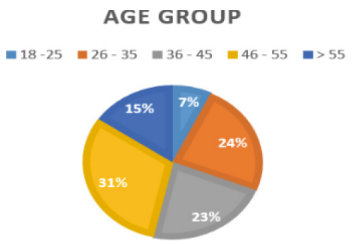



**Fig. 3.** Age group

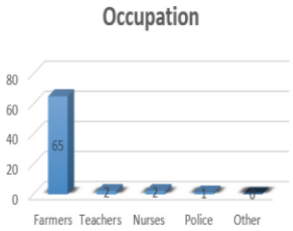



**Fig. 4.** Occupation

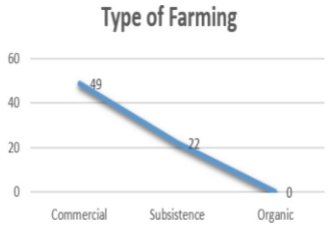



**Fig. 5.** Type of farming

Data on gender, age bracket, occupation (see Fig. 4), and type of farming (see Fig. 5) were collected. This was the way of finding out which gender dominates MRAs and their age bracket, whether they are employed or not, and also how many people they are supporting. Data on how farming is practiced in MRAs were collected, as the tools used and the type of farming. A total number of 71 farmers were given questionnaires and participated. In our data, as shown in Fig. 2 we discovered that out of 71 participants, 55% were male and 45% female. Figure 3 also shows that it was mostly farmers aged 46-55 at 31% and followed by youth at 24% aged 26-35, which proves the scarcity of jobs in MRAs.

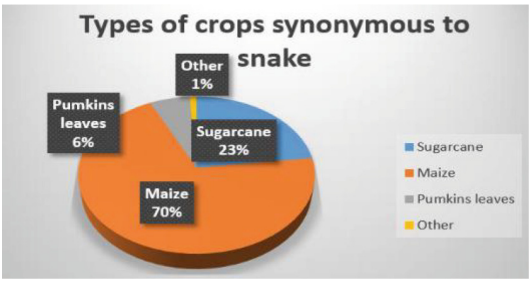



**Fig. 6.** Types of crops synonymous to snakes

In farms, we also discovered that 86% of crops are synonymous with a snake, as presented in Fig. 6. Most farmers have planted maize with 70% and 23% of its sugarcane, as reflected in Fig. 6.

In Fig. 1, the bottom of the architecture shows different farms with different kinds of snakes. We have farms from different locations in the Zululand district. A swarm of drones will fly around the phone, controlled by the farmer's phone. Farmers have two options either they download or upload information using the different types of phones they are using. Since the drones are going to be tracking and detecting snakes in farms. Using the mobile phone, tracked and detected information is uploaded from the phone to the cloud wirelessly. And we have a firewall to prevent all trespassers from tamper-ing with the system. Tracked and detected information is then transmitted to the central access point connected to the server providing mobile network services. The server is a database that stores the user's information about that particular object that is detected. In the cloud, everything will be processed then cloud controllers process the request to provide mobile users (farmers) with the corresponding cloud services. When the farmer opens his/her Swarm of Cloud-based Unmanned Arial Vehicle system on the phone, he/she will be able to see the real-time navigation of drones as they track and detect snakes in the farm.

## 7 Methodology

The most comprehensive investigation used questionnaires to learn about how farming is practiced in Zululand, which mobile devices are most commonly used in MRAs, current safety measures used against snakes, the most common snake affecting MRAs, methods they use to live, detect and handle snake bites. We created the paper-based questionnaire, and since we were conducting the study in MRAs, we went there physically to clarify more questions. The questions were both structured and unstructured questions; it was close-ended questionnaires. We interviewed both farmers and farm workers. The case study was conducted in the Zululand district. The sample was non-random. We were questioning the people on a willing basis.

## 8 Data Analysis and Results

The main objective of this paper was to identify what are common snakes that are mostly found in MRAs and what type of agriculture is mostly practiced. Mechanisms are mostly used to prevent snakebites, killing of snakes, and reasons for killing them in MRAs. We collected data through close-ended questionnaires given to farmers in the Pongola Zululand district in the KwaZulu-Natal province of South Africa. Farmers lack skills to improve their security and safety due to a lack of technological tools to help them improve their farming to conclude our study properly. It was necessary to analyse the data so that we could correctly test our suggestions as well as answer our research questions and present the results of the study to our readers in an understandable and convincing form. A total number of 71 farmers were given questionnaires and participated. Data on gender, age bracket, occupation, number of people in the household were collected, crops and types that are synonymous with snakes, frequent encounters with the snake and What they do with the snake or does it defend itself, and reasons for killing snakes, common things that are bitten by a snake(s) in farms and security mechanisms currently in place to curb snakes. The data that answers the above-men-tioned points are also presented graphically below.

to have far-reaching effects across today's society, trans-forming our lives and how we do business. The agricultural industry seems to have embraced drone technology with open arms, using these advanced tools to transform modern farming. The advent of drones, better known as drones, has proved to be beneficial for the overall development of the human race – both technologically and strategically [28].

Drones-based agricultural robotics, in particular, has attracted immense interest as is evident in 10-year sales forecasts of the technology [25]. In Agriculture, from crop monitoring to planting, livestock management, crop spraying, irriga-tion mapping, and more. Agricultural drones help to achieve and improve what's known as precision agriculture. This approach to farming management is based on observing, measuring, and acting based on real-time crop and livestock data. It erases the need for guesswork in modern farming and allows farmers to maximize their yields and run more efficient organizations, all while enhancing crop production.

There are multiple uses for agricultural drones, including Scouting land and crops, checking for weeds and spot-treating plants, monitoring overall crop health and managing livestock, and monitoring for health issues. Drones have propulsion systems, infrared cameras, global positioning and navigation sys-tems, programmable controllers, and automated flight planning. Plus, with cus-tom-made data processing software, any collected information can instantly be used towards better management decisions. This will make it easier for research-ers to track and detect snakes on farms.

In recent years, the cost of agriculture drones has rapidly declined, which has not only led to the explosion of drone use cases in agriculture but has made it a no-brainer investment for modern farmers [11].

Multi-rotor Drones are the cheapest, easiest to make and to operate of all the drone types mentioned above (based on aerial platforms). Single-pilot livestock management and observation are made possible by including cameras on multi-rotor drones.

## 6 Proposed Architecture

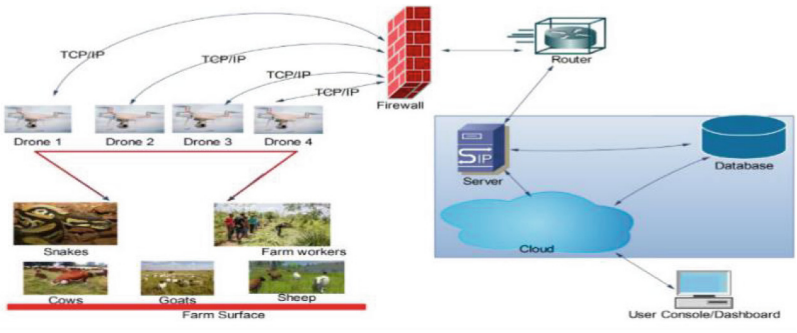



**Fig. 1.** Proposed system architecture

to accurately track snakes while Faster R-CNN would be on image detection since it's difficult to spot green snakes in long green grass and sugarcane.

## 3 Purpose, Objective, and Significance of the paper

The purpose of this study was to ascertain the penetration and use of drones as means of detecting and tracking snakes in the farming community of rural Zu-luland region, KwaZulu-Natal. The objective was to develop a model of cloud-based swarm Robotics for clustered drones to accurately detect and track sta-tionary and motion-based snakes in an agricultural environment or farm. The model platform could assist the local farming community from snakebites and envenoming. In addition, the model will help curb snake killings and enhance biodiversity in rural areas. The expected contribution of this research is devel-oping a cloud-based swarm DRONE model that facilitates the intelligent detec-tion and tracking of stationary and motion-based snakes in farms using PSO and a Faster R-CNN algorithm. This will improve safe farming and biodiversity con-servation on and also improve food security in developing countries.

## 4 Cloud Robotics

Cloud computing is a computing style that provides power referenced with In-formation Technology (IT) as a service [24]. Cloud computing is the type of computing that allows access to information and computer resources anywhere as long as the network connection is available. Most organizations and individ-uals have migrated their tasks, data, and applications to the cloud. Such include but are not limited to Amazon, where we find Dropbox, Twitter, Instagram, Quora, etc. All these are applications used by millions of people that use the Cloud.

Cloud Robotics is a topic that has garnered much attention from the research community, especially with the proliferation of cloud infrastructure, improved communications technologies, and commoditized Robotics. Cloud Robotics is any automation system that relies on either data or code from a network to sup-port its operation [13]. Cloud Robotics can also be described as any automation system that relies on either data or code from a network to support its operation. Drones-based agricultural Robotics, in particular, have attracted immense inter-est, as is evident in 10-year sales forecasts of the technology [25]. The benefits of Cloud Robotics include shared knowledge databases that disparate robots have access to, the ability to offload processing-intensive tasks to cloud infra-structure, and robot access to skill/behaviour databases [26].

In this paper, we incorporate cloud and drones, and agricultural workers will be able to gather and store data, automate redundant processes, and improve effi-ciency. Since Drones will be tracking and detecting snakes on farms, the data recorded or snake's data tracked will be stored in the cloud for future use.

## 5 Drones in Agriculture

ICT in agriculture has attracted a lot of attention and researchers for the past few years all over the globe, as it seems to have outstanding benefits compared to old ways of practicing farming [27]. Drones technology is a phenomenal inno-vation that continues

and processing system to detect Burmese pythons on the wild. They have leveraged IMEC sensor array technology to collect data on the Vis-NIR regions of the spectrum of several pythons to determine a set of bands that can help with the detection. They further allude that Hyperspectral measurements and band selection algorithms show that an optimal solution involves the combination of bands in both visual and NIR. Due to technological fabrication issues is more difficult and expensive to combine very broad-spectrum bands in a single mul-tispectral focal plane array. For that reason, they compare the Vis multispectral sensor vs. the NIR multispectral sensor showing that NIR multispectral sensor has a lead on the task. Consequently, their analysis was concentrated on the NIR multispectral collection to find an algorithm that performs best in the task of python discrimination. They first trained a random forest algorithm since it con-siders only spectral information. Later, they studied the use of deep learning to improve the detection of pythons in the wild. They discovered that deep learning algorithms need lots of labeled data to converge. Their collection included less than 100 images. Therefore, they decided to use a pre-trained model on RGB data and take advantage of transfer learning by fine-tuning the network using their data. Accordingly, they selected three bands that are used in substitution of the original RGB bands in the deep network. They conclude that future research-ers need to cover aspects of studying different deep network architectures, in-clude more spectral bands on the algorithm; perhaps a new data collection that supports these networks is required, and finally, actual hardware implementation of the camera.

Some researchers use a combination of algorithms to achieve the task of classifying snakes [22]. In their study, they investigate human perception and the se-lection of words in describing a snake based on their visual view. The descrip-tions are presented in unstructured text, and the NLP processing involves pre-processing, feature extraction, and classification. Four machine learning algo-rithms (naïve Bayes, k-Nearest Neighbour, Support Vector Machine, and Deci-sion Trees J48) were used during training and classification.

Generally, all existing works related to swarm intelligence were derived from the group or social behaviour of animals or insects [23]. In their study, they exhibit the compatibility of applying swarm algorithm on DRONE quadrotors for aerial surveillance, search, and reconnaissance operations through flight for-mations and reconfiguration by abiding swarming patterns and behaviour.

They conclude that the increased number of robots can yield higher accuracy. This directly implies that the centroid can be controlled accurately by increasing the number of robots because the resolution of the swarm increases proportion-ally with the number of swarm members. The foraging behaviour experiment revealed that the time it takes to hit or reach the desired position decreases with the increase in swarm members. They suggest that for future purposes, aggregation can be implemented, including other robot types i.e., underwater or land-based robots. Optimized searching algorithms in three-dimensional domains can be derived from the foraging behaviour. Lastly, they suggest that all of this be-haviour can be mixed with other algorithms, such as pheromone and CNN algorithms, to introduce multi-tasking for the swarm. Adopting their suggestion, our study seeks to combine Swarm and Faster R-CNN algorithm to track and detect snakes in agricultural fields in MRAs. Swarm algorithm will guide could-based Drones

faster automation, digitization, and big data collection that manufacturers and industries must align with. Efficiency im-provement, reducing costs, and maintaining quality can be more comfortable than before with the help of these components. So this is what motivated us to work on such an experiment where the effective procedure of automated SCDRONE tracking and detection of snakes in farms can somehow help protect farmers and farm workers from dangerous snakes and also protect the very same snakes from being killed by farmers for biodiversity purposes. This will help farmers to produce better and support the integration of industry 4.0.

Many algorithms have been used to try to track and detect plants and animals on farms. The algorithms take images, make segmentation to extract features from them, and use the features, and the machine classifies which disease the plant has [19]. In their study, they develop the design of how machine learning can be used in automatically detecting plant diseases by seeing the plant leaves. Their objective was to construct a system that takes images as input, and after precise testing, it gives the disease name in the output. To implement their proposed method, they collected data manually and used a faster R-CNN algorithm and some necessary tools. Their implementation process consists of several seg-ments and pre-processing, which is described below: Dataset Collection, Ten-sorflow in disease detection, Labelling of the images, and Algorithm used.

Their expected result is obtained with some computational efforts where the efficiency of the proposed algorithm can be shown, and the classification of leaf disease can be specified. They managed to achieve an accuracy value, and that is 67.34%. Their study is similar to our study because we track and detect mo-tion-based objects (snakes) while they are detecting stationary plant diseases. We both opt to use the Faster R-CNN algorithm for image detection because it's much faster. Another reason we opt for faster R-CNN is its ability to detect objects in real-time compared to R-CNN and the Fast R-CNN algorithm. Our model allows a swarm of cloud-based Drones to track and detect snakes in real-time in MRAs, which is why we also adopt PSO for guiding the Drones' path. We will improve from the 67.34% accuracy they achieved even though we use different objects.

Many algorithms have been used to track and detect snakes, to be precise. The surveillance and tracking of the snakes are difficult due to their size and the nature of their movement [20]. In their study, they proposed a system that seeks to identify snakes and small dangerous insects to the farmers and improve productivity using a classification algorithm. Their classification algorithm and feature extraction describe the unique features of snakes and small dangerous insects. Then they produce the buzzer in the current location and display the location use of GPS. In their method, the detection of motion in the video frame and identification of the objects in the area of motion using features extraction, which describes the unique features of snakes and small dangerous insects. They use the platform Raspberry PI, which they say is the sequence of credit card– sized single-board computers established in the United Kingdom by the Raspberry PI Foundation. The position of the snakes, once detected, is tracked in order to calculate the distance of snakes with areas of farmers in the agricultural land.

Given the magnitude of the snake problems, better detection tools are needed to help to find the snakes [21]. In their study, they study the development of a camera

integrate and evaluate a set of low-cost technologies that allow the detection of explosive landmines autonomously and without com-promising the mission. DRONE was equipped with cameras that enable visual feedback of the terrain in real-time. By capturing several sequences of images, visual algorithms for landmine detection are applied. The detection process was performed using the still images captured by the bottom camera of the drone. Outdoor tests were conducted using tuna cans as landmine-like objects and un-der different sunlight profiles and wind speeds. Some objects were randomly placed on the surface (fully visible), while others were partially buried.

The difference between their study and the current study is that we are not only tracking and detecting still objects, but we are also tracking moving objects in the form of snakes. Our Drones will be capable of tracking and detecting sta-tionary and motion-based snakes. Our model is also incorporated with the cloud to store this information and classify the kind of snakes detected, which will help as awareness people of MRAs to be aware of such snakes and relocate those snakes to areas of safety before they are killed.

Artificial Intelligence (AI) is a vital technology for the future of drone systems to improve their independent performance (Yadav and Gaur, 2014). Drones should be able to perform cloud-based tasks autonomously and have abilities to perform self-determination of tasks.

Future drones should be able to autonomously plan flight paths based on their respective missions and corresponding constraints [15]. In our study, Drones fly around the farm, tracking and detecting snakes. When the constraints changes, Drones should be able to autonomously adjust the flight path.

One of the characteristics of intelligent Drones in the future is the ability to efficiently perform complex tasks through independent cooperation [16]. AI has played an increasingly important role in the field of automated control of drones [17]. In their study, they prove that deep reinforcement learning can be success-fully applied to an ancient puzzle game Nokia Snake after further processing. A game with four directions of movement. Through deep intensive learning and training, the Snake (or self-learning Snake) learns to find the target path autonomously, and the average score on the Snake Game exceeds the average score on the human level.

Therefore, their proposed Snake algorithm to be able to find the target path autonomously is an attempt and key technology designing of autonomous search and rescue personnel and material dispatching drones. They apply the reinforce-ment learning method to the process of simulating the autonomous exploration of the target by the drone in the game environment of the Snake Game. The snake body is used to represent the drone, and the hotspot is used to represent the target to be searched.

In their study, a single drone was used for testing, and results were achieved. In our study, through a combination of PSO and Faster R-CNN algorithm, we used a swarm of drones to track and detect snakes on a farm. The swarm of cloud-based Drones will be utilized to track and detect snakes.

In the recent era, the swift development in digital technology has started a new evaluation in the industrial revolution called Industry 4.0 generally [18]. This revolution is all about introducing modern technologies to connect the compo-nents with the industries and support sustainability as well. It brings new and augmented algorithms to promote

The study is similar to ours because we seek to utilize a swarm of Drones in agriculture to detect and track motion-based snakes on farms. The difference is that we don't only discuss Drones. Our study develops a cloud-based model to help farmers detect and track motion-based snakes on farms. The Swarm of Cloud-based Unmanned Aerial Vehicles (SCDRONE) tracks and detects mo-tion-based snakes that can pose a danger to farm workers.

According to [12], 80% of the commercial market for Drones is expected to be occupied by agricultural Drones in the future. They further outlined that by in-troducing Drones to traditional agriculture, working hours and labour require-ments have been significantly reduced, and the efficiency of agricultural works has improved significantly. However, they've seen that using a single drone it's a drawback because it uses the battery as its main source of power. In their study, they propose a multi-drone system that will make it possible to carry out coop-erative works simultaneously to curb the inefficiency of a single drone in terms of time and energy. In their study, they develop a multi-drone system for agri-culture using the distributed swarm control algorithm and evaluate the system's performance, which is also similar to our study because we target the swarm of Drones.

The difference is that we don't only check the functionality of the swarm of Drones in terms of performance and efficiency to execute tasks. Our study de-velops a cloud-based model to help farmers detect and track motion-based snakes in farms using the Swarm algorithm and the Faster R-CNN algorithm. Our study is not only beneficial to farmers. It even helps to protect and save snakes that farmers and farm workers kill. The Swarm of Cloud-based Un-manned Aerial Vehicle (SCDRONE) tracks and detects any motion-based snakes that can pose a danger to farm workers. Performance and efficiency in executing tasks will be observed and improved, cloud and drone interaction will be checked, and path planning will also be monitored.

Cloud computing can be defined as utilizing the internet to provide technology-enabled services to people and organizations [3]. Cloud Robotics, to be specific, is a topic that has gathered much attention from the research community glob-ally, especially with the increase in cloud infrastructure, improved communica-tions technologies, and commoditized Robotics. Robotic services are systems, devices, and robots with three functions: Sensation, actuation, and control [4].

There have been numerous studies in the area of Drones used in agriculture, wildlife tracking and conservation, Cloud Robotics, drone-based image pro-cessing, and drone-based path planning.

Cloud Robot and Automation systems a system that relies on either data or code from a network to support their operation, i.e., where not all sensing, computa-tion, and memory are integrated into a single standalone system [13]. In their definition, the researchers are trying to include future systems and many existing systems that involve network teleportation or networked groups of mobile ro-bots, such as Drones or warehouse robots, as well as advanced assembly lines, processing plants, and home automation systems.

Drones have been used to detect explosives landmines [14]. The researchers outlined that the military has been the first to deploy machines to overcome the risks involved when a human carries out the landmine detection process. The goal of their study was to

agriculture. In Sect. 6, we explore snakes. In Sect. 7, we present related work. In Sect. 8, we present the proposed archi-tecture. In Sect. 9, we discuss the methodology followed by findings and data analysis. Lastly, in Sect. 10, we present the conclusion.

## 2 Related Work

Snakebite is a neglected tropical disease and one of the major causes of mortality in developing countries [6]. They further state that deaths due to snakebites are 2.8% of total deaths. Most cases are during monsoon 55% and in rural areas 93%. Snake bites are well-known veterinary emergencies in many parts of the world, especially in rural areas [7]. They further allude that, Snake-bite is an environmental and climatic hazard. It results in the death or chronic disability of many animals and people, especially those involved in farming.

Papua New Guinea has one of the world's highest incidence rates of snake bites [8]. Papua New Guinea records 561.9 cases per 100 000 population. They fur-ther allude that in Africa, the annual incidence rate of snake bites in the Benue Valley of northeastern Nigeria is 497 per 100,000 population, with a case-fatal-ity ratio of 12.2%. In northern Africa, the species that causes most bites and deaths belong to the family Viperidae, Echis sp (saw-scaled vipers).

In the case of our study, which is MRAs of Southern Africa, there are some 38 venomous snake species in South Africa (SA), of which approximately half are dangerous to humans [9]. They further allude that the highest incidence of snakebite in South Africa is in the rural northeastern coastal belt of KwaZulu-Natal. Small local studies have suggested an annual incidence of snakebite in northeastern parts of the province of 28–96.5 per 100 000. Prevalent species that cause problems in KwaZulu-Natal include Mozambique spitting cobra (Naja mossambica) and puff adder (Bitis arietans), an elapid and viperid respectively, both of which have potent cytotoxic venom. The black mamba (Dendroaspis polylepis) and various cobra (Naja) species are elapids possessing potent neurotoxic venom and muscle weakness. The boomslang (Dispholidus typus), a col-ubrid with a haemorrhagic venom, can cause potentially fatal bleeding.

The case study [10] presented that the subtropical low-lying northeast of KZN accounted for the majority of snakebites, in keeping with other studies showing hot, humid climates in low-lying rural areas to be hotspots for snakebite for snakes (puff adder, boomslang, and Mozambique spitting cobra). The 3 districts representing this region (uMkhanyakude, Zululand, and uThungulu) are all un-derdeveloped and have primarily rural subsistence populations of KZN.

Drones allow farmers to observe their fields from the sky, which can reveal many issues on the farm, common among which is irrigation-related problems, soil variations, and fungal and pest infestations [11]. In their study, researchers discuss the different types of Drones, and their application in pest control, crop irrigation, health monitoring, animal mustering, geo-fencing, and other agricul-tural-related activities. The author further shares the advantages and potential benefits of Drones in agriculture. The study does manage to present four major types of Drones. Though the multi-rotor drones, with their ability to hover on the spot and take off and land vertically, seem suited for agriculture, their limited flight time is a major concern.

Cloud computing can be defined as utilizing the internet to provide technology-enabled services to people and organizations [3]. Cloud computing has emerged over the past years and has become the most common and helpful source of information between partners and people who are needy and who want to share certain information. If agriculture were practiced with the latest technologies in MRAs, we wouldn't face challenges such as famine, poverty, unsafe farming and wildlife killings, crime, and rural-to-urban migration.

Robotic services are systems, devices, and robots with three functions: Sensa-tion, actuation, and control [4]. Providing cloud-based robotic services in agri-culture will curb many challenges faced by farmers and farm workers in devel-oping countries. About 5.4 million snake bites occur yearly, resulting in 1.8 to 2.7 million cases of envenoming [5]. They further say there are between 81 410 and 137 880 deaths and around three times as many amputations and other per-manent disabilities each year. In their findings, it is also discovered that most of these occur in Africa, Asia, and Latin America. In Africa, WHO researchers found that there are an estimated 435 000 to 580 000 snake bites annually that need treatment. 70% of this envenoming affects farmers (both young and old) in poor rural communities in low- and middle-income countries.

Snakebite is a neglected tropical disease and one of the major causes of mortality in developing countries [6]. They further state that deaths due to snakebites are 2.8% of total deaths, and most cases are during monsoon 55%, and from rural areas, 93%.

MRAs farmers and farm workers are not safe due to dangerous snakes that roam around farms. Maintaining a high level of biodiversity is important to all life on earth, including humans, and snakes are an important part of that biodiversity. Snakes make up a significant proportion of the middle-order predators that keep our natural ecosystems working. Killing them also creates an imbalance in the ecosystem. A cloud-based Robotics model that harnesses a cluster of drones to detect and track stationary and motion-based snakes that pose a danger to farm-ers and farm workers in the context of day navigation for better safety of both farmers and snakes can be a solution.

In this paper, we therefore present findings of a preliminary study of how farmers from MRAs conduct farming amid snakes, the type of technologies they are using to detect snakes, what common snakes are mostly found in MRAs, and what type of agriculture is mostly practiced, mechanism mostly used to prevent snakebites, killing of snakes and reasons killing them in MRAs, among other demographic information. We collected data through close-ended question-naires given to farmers in the Zululand district in the KwaZulu-Natal province of South Africa. Due to the high likelihood of illiteracy among MRAs, we were also physically present to explain the questions in the questionnaire. We com-piled and analyzed the data after it was all collected.

The plan and implementation of the proposed system will be based on the find-ings of this preliminary study, which provides a solution to the problem of de-tecting and tracking stationary and motion-based snakes in farms in MRAs to curb the envenoming and deaths of farmers and farm workers. The system will improve safe farming, maintain a high level of biodiversity and enhance yields in MRAs.

The rest of this paper is organized as follows: In Sect. 2, we present the pur-pose, objective, and significance of the paper. In Sect. 3, Cloud robotics and their importance are explored. In Sect. 4, explore more about Drones. In Sect. 5, we discuss Drones in

# A Cloud-Based Drones' Model for Detection and Tracking of Stationary and Motion-Based Snakes in Farms in Marginalized Rural Areas: A Preliminary Study

Phumlani T. Simelane(✉) and Okuthe P. Kogeda

University of the Free State, Bloemfontein 039, South Africa
{2019872668,kogedapo}@ufs.ac.za

**Abstract.** Marginalized Rural Areas (MRAs) practice farming as their main source of food, employment, and income, yet in most cases, they lack the basic resources and skills to improve their yields and farming techniques. Due to the lack of information and resources, farmers experience snake bites. Others even get killed without help due to the long distance they travel to obtain help from healthcare facilities. Farmers being killed by snakes lead to a reduction in productivity (since some productive & knowledgeable farmers die), and it also leads to the loss of livestock. With more frequent snake bites in MRAs, farmers surrender farms to snakes leading to fewer farms to grace in. Hence there is a need to design and develop a swarm cloud-based drone model for tracking and detection of sta-tionary motion-based snakes in farms for better safety of both farmers and snakes. We collected data through close-ended questionnaires given to farmers from the Zululand district in South Africa. The preliminary study results show that 97% of farmers encounter snakes on their farms, and 70% of the crops are synonymous with snakes. Farmers eliminate 73% of these snakes, and 67% of the snakes fight back, which may result in hospitalization or fatality. While these snakes are elim-inated, 54% of the rodents impact their agricultural products. The most common strategy or method used by farmers in farms to get rid of snakes is cats, which is at 65%.

**Keywords:** Drones · Tracking · Detection · CNN · PSO · Snakes · MRAs · Agriculture and food security

## 1 Introduction

Information and Communication Technology (ICT) in agriculture has attracted a lot of attention and research over the past few years all over the globe, for it seems to have outstanding benefits compared to traditional farming. ICT has played a major role in collecting and sharing timely and accurate information on markets, weather, inputs, and prices in developing countries [1]. The use of drones in agriculture and smart farming is very effective because drones can give farmers a bird's eye view of their fields while remaining close to the terrain and so providing more precise evaluations [2].


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 54–67, 2025.
https://doi.org/10.1007/978-3-031-81573-7_4

77. Jones, A., Vidalis, S.: Rethinking digital forensics. Annals of Emerging Technologies in Computing **3**(2), 41–53 (2019). https://doi.org/10.33166/AETiC.2019.02.005
78. Adedayo, O.M.: Big data and digital forensics, Rethinking Digital Forensics, In: 2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF). (2016)
79. Omer, M.A., Yazdeen, A.A., Malallah, H.S., Abdulrahman, L.M.: A Survey on Cloud Security: Concepts, Types, Limitations, and Challenges. Journal of Applied Science and Technology Trends **3**(2), 101–111 (2022). https://doi.org/10.38094/jastt301137
80. Apau, R., Koranteng, F.N.: An overview of the digital forensic investigation infrastructure of Ghana. Forensic Science International: Synergy **2**, 299–309 (2020). https://doi.org/10.1016/j.fsisyn.2020.10.002

56. Soltani, S., Seno, S.A.H.: A survey on digital evidence collection and analysis, In: 7th International Conference on Computer and Knowledge Engineering, Iran: IEEE, pp. 1–7. (2017)
57. Lee, S., Kim, H., Lee, S., Lim, J.: Digital evidence collection process in integrity and memory information gathering, (2005)
58. Karagiannis, C., Vergidis, K.: Digital evidence and cloud forensics: Contemporary legal challenges and the power of disposal. Information (Switzerland) **12**(5), 181 (2021). https://doi.org/10.3390/info12050181
59. Nikkel, B.J.: Improving evidence acquisition from live network sources. Digit. Investig. **3**(2), 89–96 (2006). https://doi.org/10.1016/j.diin.2006.05.002
60. Ferguson, R.I., Renaud, K., Wilford, S., Irons, A.: PRECEPT: a framework for ethical digital forensics investigations. J. Intellect. Cap. **21**(2), 257–290 (2020). https://doi.org/10.1108/JIC-05-2019-0097
61. Allie, R.: Judgement of Hanna Cornelius: Case No: CC 04/2018. Cape Town: Western Cape High Court, pp. 1–69. (2018)
62. Masipa, J.: Oscar Pistorius Murder Charge: Case No: CC113/2013, pp. 1–26. (2016)
63. Wilson, S.D.J.: Tshegofatso Pule Murder: Case No: SS36/2021, pp. 1–10. (2022)
64. Desai, S.: Henri Christo Van Breda Murder Case: No SS17/16, pp. 1–270. (2018)
65. Makgoba, M.W.: The Report into the Circumstances Surrounding the Deaths of Mentally Ill Patients: Gauteng Province. (2017)
66. Khan, A., Wiil, U.K., Memon, N.: "Digital forensics and crime investigation: Legal issues in prosecution at national level", in *5th International Workshop on Systematic Approaches to Digital Forensic Engineering*. SADFE **2010**, 133–140 (2010). https://doi.org/10.1109/SADFE.2010.8
67. van Beek, H.M.A., van den Bos, J., Boztas, A., van Eijk, E.J., Schramp, R., Ugen, M.: Digital forensics as a service: Stepping up the game. Forensic Science International: Digital Investigation **35**, 301021 (2020). https://doi.org/10.1016/j.fsidi.2020.301021
68. Jarrett, A., Choo, K.R.: The impact of automation and artificial intelligence on digital forensics. WIREs Forensic Science **3**(6), 1–5 (2021). https://doi.org/10.1002/wfs2.1418
69. Marshall, K., Rea, A.: Legal challenges in cloud forensics. In: 27th Annual Americas Conference on Information Systems, AMCIS 2021 (2021)
70. Akinbi, A.O.: Digital forensics challenges and readiness for 6G Internet of Things (IoT) networks. WIREs Forensic Science (2023). https://doi.org/10.1002/wfs2.1496
71. Choi, M., EL Azzaoui, A., Kumar Singh, S., Mohammed Salim, M., Reward Jeremiah, S., Hyuk Park, J.: The Future of Metaverse: Security Issues, Requirements, and Solutions. Human-centric Computing and Information Sciences, vol. 12 (2022)
72. Sun, Y., Tian, Z., Li, M., Zhu, C., Guizani, N.: Automated Attack and Defense Framework toward 5G Security. IEEE Netw **34**(5), 247–253 (2020). https://doi.org/10.1109/MNET.011.1900635
73. Pooyandeh, M., Han, K.J., Sohn, I.: Cybersecurity in the AI-Based Metaverse: A Survey. Applied Sciences (Switzerland) **12**(24), 129993 (2022). https://doi.org/10.3390/app122412993
74. Batista, D., et al.: Exploring Blockchain Technology for Chain of Custody Control in Physical Evidence: A Systematic Literature Review. Journal of Risk and Financial Management **16**(8), 360 (2023). https://doi.org/10.3390/jrfm16080360
75. Daryabar, F., Dehghantanha, A., Choo, K.K.R.: Cloud storage forensics: MEGA as a case study. Australian Journal of Forensic Sciences **49**(3), 344–357 (2017). https://doi.org/10.1080/00450618.2016.1153714
76. Martini, B., Choo, K.K.R.: Cloud storage forensics: OwnCloud as a case study. Digit Investig **10**(4), 287–299 (2013). https://doi.org/10.1016/j.diin.2013.08.005

38. Mabuto, E.K., Venter, H.S.: State of the art of Digital Forensic Techniques. Information Security for South Africa (ISSA) **2011**, 1–7 (2011)
39. Ahmed Ali, S., Memon, S., Sahito, F.: Challenges and solutions in cloud forensics, In: ACM International Conference Proceeding Series, Association for Computing Machinery, pp. 6–10. (2018). https://doi.org/10.1145/3264560.3264565
40. Mousa, A.N., Ithnin, N., Almolhis, N., Zainal, A.: A Consumer-Oriented Cloud Forensic Process Model, In: 2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC 2019), pp. 1–6. (2019)
41. Akter, O., Akther, A., Uddin, M.A., Manowarul Islam, M.: Cloud Forensics: Challenges and Blockchain Based Solutions. International Journal of Wireless and Microwave Technologies **10**(5), 1–12 (2020). https://doi.org/10.5815/ijwmt.2020.05.01
42. Montasari, R., Hill, R.: Next-Generation Digital Forensics: Challenges and Future Paradigms, In: 2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3), pp. 1–8. (2019)
43. Baig, Z.A., et al.: Future challenges for smart cities: Cyber-security and digital forensics. Digital Investigation **22**, 3–13 (2017). https://doi.org/10.1016/j.diin.2017.06.015
44. Montasari, R.: An overview of cloud forensics strategy: Capabilities, challenges, and opportunities. Strategic Engineering for Cloud Computing and Big Data Analytics (2017). https://doi.org/10.1007/978-3-319-52491-7_11
45. Stoyanova, M., Nikoloudakis, Y., Panagiotakis, S., Pallis, E., Markakis, E.K.: A Survey on the Internet of Things (IoT) Forensics: Challenges, Approaches, and Open Issues. IEEE Communications Surveys and Tutorials **22**(2), 1191–1221 (2020). https://doi.org/10.1109/COMST.2019.2962586
46. Herman, M., et al.: NIST cloud computing forensic science challenges, Gaithersburg. Maryland (2020). https://doi.org/10.6028/NIST.IR.8006
47. Isaac Abiodun, O., Alawida, M., Esther Omolara, A., Alabdulatif, A.: Data provenance for cloud forensic investigations, security, challenges, solutions and future perspectives: A survey. Journal of King Saud University - Computer and Information Sciences **34**(10), 10217–10245 (2022). https://doi.org/10.1016/j.jksuci.2022.10.018
48. Jain, P., Mahalkari, A.: Review of Cloud Forensics: Challenges, Solutions and Comparative Analysis. Int J Comput Appl **178**(34), 28–34 (2019). https://doi.org/10.5120/ijca2019919220
49. Sharma, P., Arora, D., Sakthivel, T.: UML-based process model for mobile cloud forensic application framework - a preliminary study. International Journal of Electronic Security and Digital Forensics **12**(3), 262 (2020). https://doi.org/10.1504/IJESDF.2020.108296
50. Kent, K., Chevalier, S., Grance, T., Dang, H.: Guide to Integrating Forensic Techniques into Incident Response. Gaitherburg, Maryland (2006)
51. Ninawe, P., Ardhapurkar, S.: Design and Implementation of Cloud Based Mobile Forensic Tool. In: ICIIECS'15 : DRDO sponsored 2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems : 19th and 20th March 2015 : proceedings (2015)
52. Kazim, A., Almaeeni, F., Al Ali, S.: "Memory Forensics: Recovering Chat Messages and Encryption Master Key," In: 2019 10th International Conference on Information and Communication Systems (ICICS) : 11–13 June, 2019, Jordan University of Science and Technology, Irbid, Jordan, pp. 1–7. (2019)
53. Guttman, B., White, D.R., Walraven, T.: Digital Evidence Preservation, (2022). https://doi.org/10.6028/NIST.IR.8387
54. Chow, K.P., et al.: Digital Evidence Search Kit. IEEE, Taipei Taiwan (2005)
55. Silvarajoo, V.R., Yun Lim, S., Daud, P.: Digital Evidence Case Management Tool for Collaborative Digital Forensics Investigation, In: 2021 3rd International Cyber Resilience Conference, CRC 2021, Institute of Electrical and Electronics Engineers Inc. (2021). https://doi.org/10.1109/CRC50527.2021.9392497

21. Jansen, W., Ayers, R.: Guidelines on Cell Phone Forensics Recommendations of the National Institute of Standards and Technology. Nist Special Publication, **800**(101) (2007)
22. Siddaway, A.P., Wood, A.M., Hedges, L.V.: How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. Annu. Rev. Psychol. **70**, 747–770 (2019). https://doi.org/10.1146/annurev-psych-010418
23. Khan, K.S., Kunz, R., Kleijnen, J., Antes, G.: Five steps to conducting a systematic review (2003) [Online]. Available: http://www.ncbi.nlm.nib.gov/entrez/query/
24. Okoli, C.: A Guide to Conducting a Standalone Systematic Literature Review. Communications of the Association for Information Systems, **37**, 1–33 (2015) [Online]. Available: http://aisel.aisnet.org/cais/vol37/iss1/43
25. Oosterwyk, G., Brown, I., Geeling, S.: A Synthesis of Literature Review Guidelines from Information Systems Journals. Kalpa Publications in Computing **12**, 250–260 (2019)
26. Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., Linkman, S.: Systematic literature reviews in software engineering - A systematic literature review. Information and Software Technology **51**(1), 7–15 (2009). https://doi.org/10.1016/j.infsof.2008.09.009
27. Schryen, G.: Writing qualitative is literature reviews—Guidelines for synthesis, interpretation, and guidance of research. Commun. Assoc. Inf. Syst. **37**, 286–325 (2015). https://doi.org/10.17705/1cais.03712
28. Yadav, D., Mishra, M., Prakash, S.: Mobile forensics challenges and admissibility of electronic evidences in India, In: Proceedings - 5th International Conference on Computational Intelligence and Communication Networks, CICN 2013, pp. 237–242. (2013). https://doi.org/10.1109/CICN.2013.57
29. Zawoad, S., Hasan, R.: Digital Forensics in the Age of Big Data: Challenges, Approaches, and Opportunities, In: 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, IEEE, pp. 1320–1325. (2015). https://doi.org/10.1109/HPCC-CSS-ICESS.2015.305
30. Casino, F., et al.: Research Trends, Challenges, and Emerging Topics in Digital Forensics: A Review of Reviews. IEEE Access Institute of Electrical and Electronics Engineers Inc **10**, 25464–25493 (2022). https://doi.org/10.1109/ACCESS.2022.3154059
31. Barmpatsalou, K., Cruz, T., Monteiro, E., Simoes, P.: Current and future trends in mobile device forensics: A survey. ACM Comput Surv **51**(3), 1–31 (2018). https://doi.org/10.1145/3177847
32. Said, H., Yousif, A., Humaid, H.: IPhone forensics techniques and crime investigation. In: The 2011 International Conference and Workshop on Current Trends in Information Technology (CTIT 11), Dubai, United Arab Emirates, pp. 120–125 (2011). https://doi.org/10.1109/CTIT.2011.6107946
33. Kang, S.H., Park, K.Y., Kim, J.: Cost effective data wiping methods for mobile phone. Multimed Tools Appl **71**(2), 643–655 (2014). https://doi.org/10.1007/s11042-013-1603-9
34. Pandey, A.K., et al.: Current Challenges of Digital Forensics in Cyber Security. (2020). https://doi.org/10.4018/978-1-7998-1558-7.ch003
35. Almehmadi, T., Batarfi, O.: Impact of Android Phone Rooting on User Data Integrity in Mobile Forensics, In: 2nd International Conference on Computer Applications & Information Security (ICCAIS' 2019) : 01–03 May, 2019 Riyadh, Kingdom of Saudi Arabia, pp. 1–6. (2019)
36. Chanajitt, R., Viriyasitavat, W., Choo, K.K.R.: Forensic analysis and security assessment of Android m-banking apps. Australian Journal of Forensic Sciences **50**(1), 3–19 (2018). https://doi.org/10.1080/00450618.2016.1182589
37. Janarthanan, T., Bagheri, M., Zargari, S.: IoT Forensics: An Overview of the Current Issues and Challenges. Advanced Sciences and Technologies for Security Applications (2021). https://doi.org/10.1007/978-3-030-60425-7_10

2. Neware, R., Khan, A.: Cloud Computing Digital Forensic challenges, In: Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018), pp. 1–3. (2018)
3. Mohay, G., Technical Challenges and Directions for Digital Forensics. (2005). [Online]. Available: www.e-evidence.info/cellular.html
4. Barmpatsalou, K., Cruz, T., Monteiro, E., Simoes, P.: Mobile Forensic Data Analysis: Suspicious Pattern Detection in Mobile Evidence. IEEE Access **6**, 59705–59727 (2018). https://doi.org/10.1109/ACCESS.2018.2875068
5. Chen, S., Hao, X., Luo, M.: Research of mobile forensic software system based on windows mobile. International Conference on Wireless Networks and Information Systems, WNIS **2009**, 366–369 (2009). https://doi.org/10.1109/WNIS.2009.32
6. Ayers, R., Brothers, S., Jansen, W.: Guidelines on mobile device forensics, Gaithersburg. Maryland (2014). https://doi.org/10.6028/NIST.SP.800-101r1
7. Hummert, C., Pawlaszcyk, D.: Mobile Forensics – The File Format Handbook. Springer International Publishing (2022). https://doi.org/10.1007/978-3-030-98467-0
8. Burrows, C., Zadeh, P.B.: A Mobile Forensic Investigation into Steganography. [Online]. Available: http://www.ericsson.com/res/docs/2015/ericsson-
9. Yusof, M.N., Mahmod, R., Abdullah, M.T., Dehghantanha.: Mobile Forensic Data Acquisition in Firefox OS, In: Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), 2014 Third International Conference on : date April 29 2014-May 1 2014, pp. 691–5. (2014)
10. Humphries, G., Nordvik, R., Manifavas, H., Cobley, P., Sorell, M.: Law enforcement educational challenges for mobile forensics. Forensic Science International: Digital Investigation **38**, 301129 (2021). https://doi.org/10.1016/j.fsidi.2021.301129
11. Ghafarian, A.: Foreniscs Analysis of Cloud Computing Services, In: Science and Information Conference 2015, pp. 1–5. [Online]. (2015) Available: www.conference.thesai.org
12. Sonia Akter, S., Shahriar Rahman, M.: Cloud Forensic: Issues, Challenges and Solution Models, *ArXiv*, pp. 2–23. (2023)
13. Lim, S.Y., Johan, A., Daud, P., Ismail, N.A.: Dropbox forensics: Forensic analysis of a cloud storage service. International Journal of Engineering Trends and Technology **1**, 45–49 (2020). https://doi.org/10.14445/22315381/CATI3P207
14. Choo, K.-K.R., Esposito, C., Castiglione, A.: Evidence and Forensics in the Cloud: Challenges and Future Research Directions. IEEE Cloud Computing **4**(3), 1–6 (2017)
15. Sibiya, G., Venter, H.S., Fogwill, T.: Digital Forensics in the Cloud: The State of the Art, In: 2015 IST-Africa Conference : 06–08 May 2015, Lilongwe, Malawi, Malawi: IEEE (2015)
16. Shah, J.J., Malik, L.G.: Cloud forensics: Issues and challenges. In: International Conference on Emerging Trends in Engineering and Technology, ICETET, IEEE Computer Society, pp. 138–139. (2013). https://doi.org/10.1109/ICETET.2013.44
17. Dhake, B., Limaye, H., Motwani, D.: Cloud Forensics: Threat Assessment and Proposed Mitigations. In: 2022 International Conference for Advancement in Technology, ICONAT 2022, Institute of Electrical and Electronics Engineers Inc, (2022). https://doi.org/10.1109/ICONAT53423.2022.9725922
18. Lutta, P., Sedky, M., Hassan, M., Jayawickrama, U., Bakhtiari Bastaki, B.: The complexity of internet of things forensics: A state-of-the-art review. Forensic Science International: Digital Investigation **38**, 301210 (2021). https://doi.org/10.1016/j.fsidi.2021.301210
19. Fernandes, R., Colaco, R.M.: A New Era of Digital Forensics in the form of Cloud Forensics: A Review July, 2020, In: Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020), pp. 1–6. (2020)
20. Ayers, R., Brothers, S., Jansen, W.: Guidelines on Cell Phone Forensics Guidelines on Mobile Device Forensics. Archived NIST Technical Series Publication Archived Publication, vol. 1 (2007)

must adopt sophisticated techniques to bypass encryption and navigate authorization mechanisms [57].

The integration of blockchain and cryptographic technologies may redefine data integrity and chain of custody practices, enhancing forensic evidence reliability [74] However, these innovations introduce complexities in data retrieval and analysis, necessitating continuous skill advancement in cloud architecture, network protocols, and cryptography to navigate distributed data storage and security [43].

### 6.3 Anticipated Challenges & Opportunities

The evolution of mobile and cloud forensics presents both challenges and opportunities. The complexity of data structures and encryption mechanisms demands refinement of data acquisition and decryption techniques. The increasing use of ephemeral messaging and transient data in both environments requires novel approaches for evidence capture and preservation [75, 76].

Though privacy-focused legislation and public awareness may limit data access, this presents potential for creating forensic investigation tactics that respect individual rights [60]. As the lines between mobile and cloud environments become less distinct, forensic professionals can employ integrated tools and procedures to produce an extensive digital narrative [77, 78].

Standardized procedures, innovative technologies, and educational materials will be made available through collaboration between academics, business, and law enforcement [79], enhancing the discipline's effectiveness and encouraging sound answers to new challenges [80].

## 7 Conclusion

To summarize, the field of digital forensics investigations, particularly in the domains of mobile and cloud forensics, is fraught with complexities. The constantly changing mobile device landscape, with its variety of operating systems and applications, necessitates ongoing adaptability and technical know-how. Furthermore, the dynamic and decentralized nature of cloud environments presents challenges that call for novel extraction, analysis, and preservation techniques for digital evidence. These difficulties are made more difficult by the quick speed of technical innovation, privacy concerns, and legal constraints. Collaboration between experts, the creation of cutting-edge tools, and a profound knowledge of both technological and legal issues will be vital in overcoming these obstacles and guaranteeing the integrity of digital forensic investigations in the face of the expanding digital landscape.

## References

1. Venter, H.S.: Mobile Forensics using the Harmonised Digital Forensic Investigation Process, Information security for South Africa (ISSA), pp. 1–10. Johannesburg South Africa, IEEE, Sandton (2014)

### 5.5 Enhancing Forensic Tools & Techniques

Mobile and cloud forensics face challenges in navigating encrypted data and robust security measures. Professionals use techniques like brute-force attacks and cryptographic analysis to decrypt data, utilizing cryptographic expertise and methodologies [52]. Therefore, forensic tools and techniques have become crucial for mobile and cloud forensics evolution. Advancements in machine learning, artificial intelligence, and big data analytics improve efficiency and accuracy [68]. Furthermore, collaborative partnerships between experts and developers infuse innovation into investigative methodologies, enabling forensic professionals to navigate complex scenarios while maintaining evidentiary integrity [68].

### 5.6 Improving Legal & Regulatory Frameworks

Mobile and cloud forensics require legal and privacy awareness, compliance with regulations, and ethical practices. Acquiring permissions, consent, and data protection laws ensures admissibility and reliability of forensic evidence in legal proceedings [60].

Clear guidelines for evidence acquisition, admissibility, and chain of custody enhance investigative processes, promoting equitable application and public trust [69]. Improved legal and regulatory frameworks are crucial for mobile and cloud forensics. Additionally, collaborative efforts between experts, policymakers, and practitioners align legal standards with technological advancements, safeguarding privacy rights and ethical considerations [60].

## 6 Future Trends in Mobile & Cloud Forensics

### 6.1 Mobile Technology Advancements

The development of mobile technology, such as fifth generation (5G) connectivity, virtual reality (VR), augmented reality (AR), and pervasive computing also known as the internet of things (IoT) devices, will have a huge impact on the future of mobile forensics [70].

The acquisition, analysis, and interpretation of a larger variety of data formats will provide issues for forensic practitioners as mobile devices become increasingly integrated into daily life and company processes. To do this, new approaches [71], improved encryption techniques, and multimedia fusion must be developed [72].

Data analysis and pattern identification will be fundamentally influenced by artificial intelligence and machine learning, necessitating the adaptation and adoption of these technological breakthroughs by forensic professionals [73]. To successfully traverse the changing mobile world after this paradigm change, interdisciplinary collaboration and ongoing education are required.

### 6.2 Cloud Computing & Security Innovations

Cloud forensics is closely linked to the evolution of cloud computing and security innovations. As digital evidence spreads across complex infrastructures, investigators

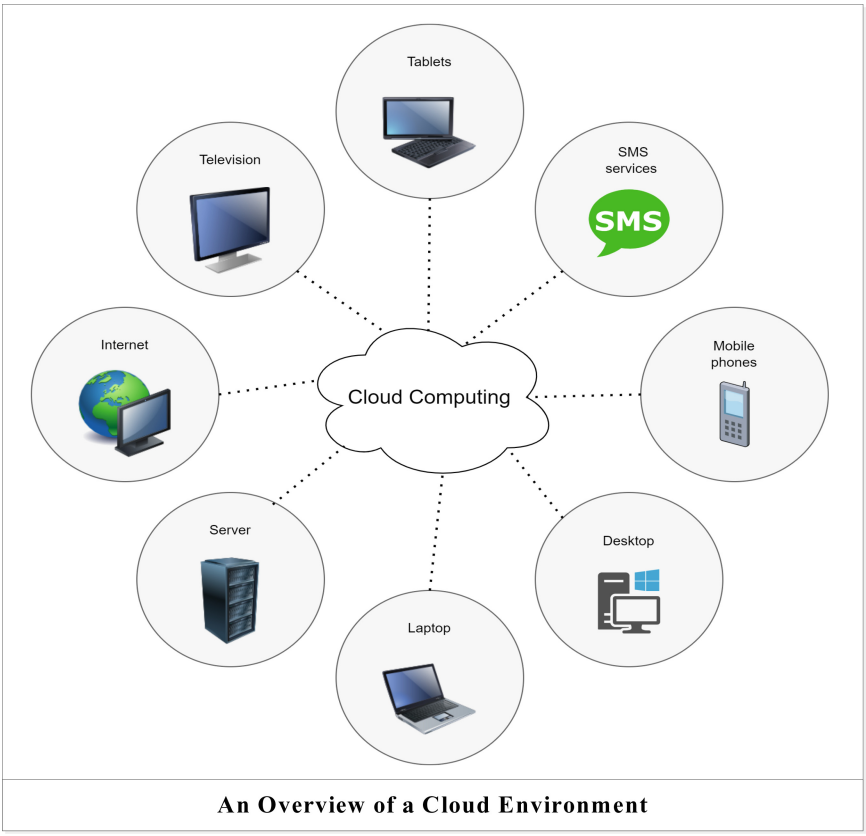



**Fig. 2.** An overview of a cloud environment setup.

### 5.3 Information Sharing & Collaboration

Collaborative engagement and information sharing are crucial for mobile and cloud forensics advancement. These platforms enable multidisciplinary expertise to fuse, fostering innovative solutions and enhancing the field's capacity to tackle complex scenarios [30]. Furthermore, collaboration plays a crucial role in enabling lawful data access, expediting the investigation process, and preserving data integrity, privacy regulations, and critical evidence acquisition [53-59].

### 5.4 Continuous Training & Skills Development

Continuous training and skill refinement are crucial for professionals in mobile and cloud forensics to remain competent and adapt to evolving landscapes [20]. This culture fosters adaptability and optimal investigative efficacy.

### 4.2 Case Study 2: Cloud Forensics

In the case of Marli van Breda Murder, the accused (Henri Christo Van Brend) was indicted for the murder of his family members, namely his parents, brother, and sister, and for an attempted murder for trying to kill his sister. He was also charged for defeating and obstructing the administration of justice [64].

Cloud forensics was conducted for the purpose of evaluating acquired digital forensic evidence from several devices, including smartphones and laptops. To establish time-frames and acquire evidence essential to the trial, cloud forensic experts were involved in the data extraction and analysis from multiple cloud-based services, emails, and other digital sources. The accused was then charged with three life sentences in prison for count one to three- and fifteen-years imprisonment for count four- and twelve-years imprisonment for count five.

Another instance was of the horrible death of mentally ill patients who were moved from one hospital facility to the other, where there were no resources for them. This case of Life Esidimeni tragedy resulted in the deaths of the moved mentally ill patients [65]. Cloud forensic was done on all the available digital records to uncover the decision-making process that led to the transfers of the patients. A commission of enquiry was established, and evidence was presented even though justice has still not been served [65].

## 5 Discussion and Best Practices

### 5.1 Forensic Acquisition & Imaging Methods

Forensic acquisition and imaging are crucial in mobile and cloud forensics, ensuring data integrity and maintaining the evidentiary chain of custody [49]. Traditional methods, like physical and logical acquisitions, are complemented by advanced techniques like live acquisitions [31]. Acquiring methods must consider device characteristics, proprietary software, security measures, and adhere to established best practices [50]. Data can also be shared across multiple platforms via the cloud environment. An overview of a general cloud environment setup is shown in Fig. 2.

### 5.2 Mobile Forensics Standardization

A foundational element of meticulous investigative procedures is the development of standardized methods in both mobile and cloud forensics. The processes for gathering, analyzing, and interpreting data are standardized to ensure consistency, dependability, and uniformity. The reliability and admissibility of evidence in court processes are increased when forensic experts follow established protocols because they are more effective at navigating complexity [66, 67]. A reliable and organized approach to digital investigations is fostered by the collaborative establishment of complete standards that consider legal, ethical, and technological factors.

**Table 2.** (*continued*)

| Challenge | Description of Challenge |
|---|---|
| Challenges of Live Forensics in Cloud Environments | Conducting live forensics within cloud environments introduces distinct challenges due to the complexity of virtualized systems and the potential for unintended data alterations. Effectively addressing the challenges of live cloud forensics requires specialized expertise in virtual machine forensics and the implementation of non-intrusive methodologies to acquire real-time data without disrupting cloud services [48] |

### 4.1 Case Study 1: Mobile Forensics

There are several instances of successful prosecution in the South African courts, that emanated from mobile forensics. The famous and emotional Hannah Cornelius Case which involves kidnaping, rape, murder and robbery was cracked using the evidence from a mobile device. The accused persons (Vernon Junaid Witbooi, Geraldo Parsons, Eben Van Niekerk, and Nashville Julius) were all indicted and charged with robbery with aggravated circumstances, kidnapping, attempted murder, rape, and murder [61]. When digital forensics was conducted on the devices, mobile forensics investigation retrieved the movement patterns of the suspects using Cellular tower information, GPS locations, and communication records from the mobile phones of the suspects and victims. All the accused persons were convicted of their indictment.

Another widely televised case was of the Paralympian Oscar Pistorius who shot and killed his then girlfriend River Steenkamp on the eve of valentine in 2013. The accused was initially convicted of capable homicide, but the appeal turned the charge into murder. Mobile forensic was performed on his mobile devices to reconstruct the events that may have led to the shooting. The suspect was successfully convicted and sentenced to thirteen (13) years in prison [62].

In the case of Tshegofatso Pule, the boyfriend of the deceased was charged with premeditated murder of his 8-month pregnant girlfriend. His mobile devices were seized for the purpose of digital forensic investigations. The mobile forensic revealed that there were calls, WhatsApp and text messages communication between the boyfriend and the hit man. The hitman ultimately turned himself a state witness and got a lesser charge, and the boyfriend was sentenced to life imprisonment [63].

Another example is of a South African national soccer team captain, Senzo Meyiwa, who was shot and killed in an alleged house robbery. In this case, mobile records from the suspects deduced that they communicated, and the suspects know each other. The case is still on trial, but mobile forensics has played a crucial role in uncovering clues about the hidden communications between the suspects and the late soccer star, then girlfriend.

**Table 2.** (*continued*)

| Challenge | Description of Challenge |
|---|---|
| Cross-Border Data Requests and Legal Considerations | Cloud forensics involves cross-border data storage, requiring strict adherence to protocols and legal issues. Collaboration with foreign organizations and cloud service providers ensures evidence acquisition, navigating legal systems and data privacy laws [44] |
| Data Fragmentation in Distributed Cloud Storage | Data fragmentation results from the distributed storage architecture used in cloud environments, where data is scattered across various physical locations [16]. It is extremely difficult to meticulously piece together disparate facts to create a consistent evidentiary narrative. To facilitate the seamless integration of fragmented data into the investigative process, forensic investigators must show proficiency in data reconstruction techniques and correlation approaches [10] |
| Cloud Service Misconfigurations and Security Incidents | Misconfigurations and security incidents can jeopardize the integrity and confidentiality of data in cloud settings. A thorough knowledge of cloud infrastructure, networking, and security procedures is required to recognize and analyze[45] cloud service misconfigurations and security breaches. To preserve evidence and minimize potential harm, security issues must be promptly detected and remedied [2] |
| Dynamic IP Addressing and Tracking Network Traffic | In cloud systems, IP addresses are assigned dynamically, which makes it difficult to follow network activity and attribute it to certain sources during forensic investigations [46]]. Tracing dynamic IP addresses and identifying communication patterns within the cloud infrastructure require knowledge of network forensics and the use of advanced network analysis tools [46] |
| Preservation of Metadata in the Cloud | A significant problem in cloud forensics is the preservation of the metadata linked to cloud-stored data. When conducting investigations, metadata is essential for proving the legitimacy, provenance, and contextual significance of the data. To ensure the admissibility and dependability of gathered evidence, careful preservation measures must be put in place to prevent unintentional metadata alteration or loss [47] |

(*continued*)

**Table 2.** (*continued*)

| Challenge | Description of Challenge |
|---|---|
| Log Collection, Retention, and Analysis | Critical elements of cloud forensics include the gathering and examination of logs from cloud infrastructures. To extract crucial information about system activities, user interactions, and potential security incidents from the collection and retention of logs from various sources, advanced log analysis techniques are required [41]. To extract valuable insights from log data, forensic investigators must have a thorough understanding of log gathering procedures and analytical tools [41] |
| Collaboration & Cooperation with Cloud Service Providers | An essential component of cloud forensics is effective coordination and cooperation with cloud service providers. Clear communication channels and the development of cooperative partnerships with cloud providers are required to obtain the essential data access and cooperation. This kind of cooperation makes it easier to acquire data legally and securely, which improves the effectiveness and precision of forensic investigations [41] |
| Coping with Rapid Data Growth in Cloud Environments | The rapid growth of cloud data volumes presents challenges in handling, processing, and analysing vast amounts of information during forensic investigations [41]. Scalable forensic methodologies and efficient data handling techniques are needed to expedite investigations without compromising accuracy. Continuous research and development are necessary to ensure timely and efficient retrieval of relevant evidence [42] |
| Handling Transient Data in the Cloud | Due to the dynamic generation, modification, and deletion of virtual resources, forensic practitioners encounter difficulties when trying to capture and preserve volatile material in cloud systems. For effective analysis and preservation without interrupting cloud services, live forensic expertise is essential [43] |

(*continued*)

### 3.2 Cloud Forensics: Major Challenges

## 4 Real-World Instances

There are some real-world instances where mobile and cloud forensics helped the courts in terms of successful convictions in South Africa. These real-world examples discussed in the next subsection highlight the value of digital evidence in contemporary court cases and the difficulties forensic professionals confront when working with developing technology and digital platforms.

**Table 2.** Tables of Cloud Forensics Challenges

| Challenge | Description of Challenge |
|---|---|
| Data Location and Jurisdiction Complexities | Due to geographical dispersion and the involvement of service providers, cloud forensics encounters difficulties in locating data and identifying its jurisdictional consequences. It takes specialized legal knowledge, global collaboration, and cutting-edge technical approaches to resolve these problems [18] |
| Multi-Tenancy and Data Isolation | Strong data isolation methods are needed in multi-tenant cloud infrastructures to protect the integrity and confidentiality of forensic investigations. For effective data management and customer-specific data security, it is essential to properly identify and separate customer-specific data [19] |
| Data Encryption and Decryption | Forensic investigators trying to access and decrypt cloud-stored data face significant difficulties due to the widespread usage of strong data encryption algorithms by cloud service providers. To overcome the obstacles preventing data access, the encryption processes used to strengthen cloud data privacy call for sophisticated cryptographic knowledge and cutting-edge deciphering approaches [39] |
| Responsibility and Security Model in the Cloud | Cloud forensics faces challenges in delineating responsibilities and security measures between service providers and customers [40]. The shared responsibility model requires understanding security responsibilities and identifying responsible parties in case of breaches or data compromises |

(*continued*)

**Table 1.** (*continued*)

| Challenge | Description of Challenge |
|---|---|
| Data Fragmentation and Reconstruction | Data fragmentation, in which relevant data may be scattered across several physical sectors or file locations, is a common occurrence in mobile devices [7]. Data carving competence and precise linkage of scattered fragments are required to ensure the coherent reconstruction of fragmented data and to rebuild a coherent evidence picture [34]. To meet this challenge, data reconstruction must be approached with care and thoroughness, assuring the accuracy and validity of the evidence that is recovered |
| Overwriting and Data Corruption Risks | When mobile devices are handled incorrectly during forensic investigations, there is a chance that important evidence's integrity could be jeopardized by accidental data overwriting or corruption. This problem highlights the importance of following best practices to assure data preservation and the accuracy of investigative results. It also highlights the necessity of using strict forensic procedures and preservation methods to prevent against data tampering [35] |
| Technological Advancements and Updates | The field of mobile forensics continues to face difficulties due to the constant evolution of mobile technology, which is characterized by regular operating system updates and hardware improvements. To ensure that forensic procedures and tools are compatible with the modern mobile landscape, constant research and development activities are required due to the quick rate of technological advancement. As a result, to continue being effective in their research, mobile forensic practitioners need to keep up with all current technology developments [36] |
| Legal and Privacy Issues | Legal and privacy considerations of many different types have a significant impact on how mobile forensic investigations are conducted [37]. In the context of gathering digital evidence, it is crucial to ensure strict adherence to legal requirements, obtain any appropriate warrants, and protect the privacy rights of everyone concerned [38]. The complex interplay between technology advancements and legal restrictions emphasizes the necessity of competent legal representation and in-depth familiarity with relevant legislation to negotiate this treacherous terrain with caution |

**Table 1.** (*continued*)

| Challenge | Description of Challenge |
|---|---|
| Locked and Password-Protected Devices | The ubiquity of password-protected and locked devices is a significant challenge for mobile forensic investigators. Data extraction procedures become more complex because of the strict access constraints that protect device contents, especially when dealing with reluctant or unresponsive device owners [31, 32]. To overcome these difficulties, specialist forensic methods must be used to get beyond the authentication obstacles that prevent the collection of relevant evidence |
| Cloud Services and Data Synchronization | The prevalence of cloud services and data synchronization in modern mobile forensics creates complex scenarios where crucial digital evidence may be stored in distant cloud repositories rather than only on the physical device. The use of cloud-based data synchronization and storage adds layers of complexity to the investigation process because it necessitates abiding by legal requirements and working cooperatively with cloud service providers to gain access to, and secure crucial data kept in off-site locations [31] |
| Third-Party Applications and Data Access | In their efforts to obtain and understand application-specific data, forensic practitioners face a difficult issue due to the widespread use of third-party applications within mobile ecosystems. To successfully recover crucial evidence, it is necessary to have a thorough understanding of various application frameworks. Third-party programs may differ in their encryption and data storage procedures [31, 32]. Continuous research and flexibility in response to shifting mobile application paradigms are necessary to ensure the seamless integration of various third-party applications into the forensic workflow |
| Recovery of Deleted Data | Recovering deleted data from mobile devices is a difficult task since contemporary smartphone architectures include tools like TRIM or similar technologies that hasten the deletion of idle storage areas [24]. These data-wiping methods necessitate forensic methodologies and specialist knowledge to identify and recreate any remaining traces of wiped data, assuring a thorough and accurate inquiry [33] |

(*continued*)

## 3 Mobile and Cloud Forensics Challenges

The challenges brought about by mobile and cloud technologies in relation to forensic investigation processes have resulted in research evolving to cover these domains. The key challenges identified through literature are depicted in Table 1 and Table 2 below.

### 3.1 Mobile Forensics: Major Challenges

**Table 1.** Tables of Mobile Forensics Challenges

| Challenge | Description of Challenge |
|---|---|
| Device Diversity & Fragmentation | Due to the massive diversity of mobile devices, which are characterized by a wide range of unique hardware configurations, operating systems, and proprietary applications, the field of mobile forensics faces a difficult task [28–31]. The development of forensic procedures and technologies capable of accommodating this broad variety of device types is required due to the significant fragmentation within the mobile ecosystem [31]. Considering this steadily growing device diversity, determining universal applicability becomes challenging, necessitating the need for adaptable forensic techniques to efficiently handle the range of mobile devices found throughout investigations |
| Encryption & Security Measures | The use of smartphones has grown tremendously, and this has increased the importance placed on using strong encryption and security measures to protect critical data. Mobile forensic practitioners have a significant barrier when trying to access and decipher encrypted data without the necessary passcodes or biometric credentials [31, 32]. To overcome the obstacles to data decryption, cutting-edge cryptographic knowledge and novel methodology must be developed. The deployment of complex encryption mechanisms to safeguard user information manifests as a barrier to conventional forensic procedures [31, 32] |

(*continued*)

evidence, these disciplines demand the union of technical proficiency, legal grasp, and methodological rigor [19].

### 1.3 Research Objectives

The major goal is to explore the many complex problems that arise in the fields of mobile and cloud forensics. This study attempts to provide a thorough understanding of the challenges that forensic practitioners face by methodically investigating the complexities involved in the extraction, processing, and interpretation of digital evidence from mobile devices and cloud environments.

The study attempts to uncover developing trends, cutting-edge approaches, and prospective solutions to address these difficulties through a thorough review of the existing situation. The study also aims to highlight the significance of standard operating procedures, interdisciplinary cooperation, and ongoing skill improvement as the foundations of efficient mobile and cloud forensics. This research study intends to develop digital investigative methods and improve the integrity of forensic results by offering insight on the challenges and opportunities within these disciplines.

## 2 Methodology

This study followed a systematic review protocol which involved a comprehensive search for relevant literature on cloud and mobile forensics challenges [22]. Several guidelines have been developed for conducting systematic reviews [23–26]. However, to achieve its main goal of synthesizing and interpreting previous research on cloud and mobile forensics challenges, this study followed the systematic review approach by [27] The stages involved in this synthesis are as follows: Framing, search and assessment, synthesis, and interpretation.

### 2.1 Framing

Digital forensics investigations have grown extremely complex due to the evolving and extensive use of mobile and cloud technologies. This has resulted in forensic practitioners encountering new challenges during the different phases of the investigation process. This study uses existing literature to highlight these challenges. It further provides real-world instances where the use of mobile and cloud forensics was able to solve cases.

### 2.2 Search and Assessment

The online databases used in this review were IEEE Xplore, SpringerLink, Google Scholar, and Science Direct. The search strings used were: “mobile forensics” OR “mobile forensics challenges” OR “challenges in mobile forensics”, “cloud forensics” OR “cloud forensics challenges” OR “challenges in cloud forensics”. From the results obtained, paper abstracts were read to identify relevant papers and discard irrelevant ones. Once all the irrelevant papers were filter out, full text reading of the remaining papers was done.

and mined from digital devices, to ensure that there is trust from the side of the mobile and cloud system users. A general overview of a digital forensic investigations and the types of forensic investigations is depicted in Fig. 1.

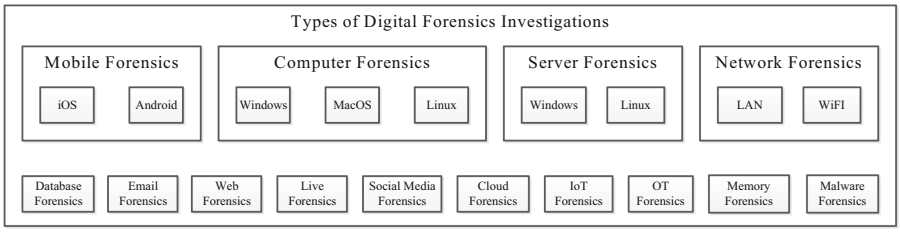



**Fig. 1.** Overview of the types of Digital Forensic Investigations

### 1.1 Importance of Mobile & Cloud Forensics in Modern Investigations

The scope and complexity of digital investigations have been redefined by the growing use of mobile devices and the migration of data to cloud systems. The digital traces left in cloud storage and on mobile devices provide priceless insights into user behavior, data transactions, and communication patterns.

These observations have been helpful in a variety of situations, including criminal investigations, civil lawsuits, cybersecurity events, business compliance audits, and more. In an increasingly digital environment, mobile and cloud forensics serve as accountability pillars by making it possible to find hidden evidence, follow data trails, and clarify the chronology of events [18]. The significance of these forensic disciplines cannot be overstated, as forensic disciplines assist in [19] building confidence, maintaining data integrity, and applying justice in technology environments that continue to expand [19].

### 1.2 Overview of Mobile Forensics & Cloud Forensics

A variety of approaches are included in mobile forensics that are targeted at retrieving, analyzing, and deciphering data from mobile devices. The variety of information included on these devices, from text messages and call logs to application usage history and geolocation information [6, 20, 21], offers a profound understanding of user behavior.

Conversely, cloud forensics tackles the intricate challenges posed by virtualized environments, involving the acquisition and analysis of data stored remotely in cloud service providers' infrastructure. This includes unearthing evidence from virtual machines, deciphering encrypted data, and discerning the implications of data fragmentation across distributed cloud storage [19].

In addition, cloud forensics addresses the complex issues brought on by virtualized settings and entails the capture and analysis of data kept remotely in the infrastructure of cloud service providers. To ensure the precise and admissible presentation of


# Digital Forensics Investigations: Major Challenges in Mobile and Cloud Forensics

Tanita Singano, Norman Nelufule(✉), Boitumelo Nkwe, Kele Masemola, Daniel Shadung, Zamo Ngubane, Ntombizodwa Thwala, and Japhtalina Mokoena

Defence and Security Cluster, Information and Cybersecurity Centre (ICSC), Council for Scientific and Industrial Research (CSIR), Pretoria 0184, Brummeria, South Africa
nnelufule@csir.co.za

**Abstract.** Due to the extensive use of mobile devices and cloud computing, digital forensics investigations have grown increasingly complex. The main challenges that forensic investigators in the fields of mobile and cloud forensics encounter are discussed in this study. The proliferation of various devices, operating systems, and applications creates challenges for data capture, extraction, and interpretation in the context of mobile forensics. Additionally, the process is made more difficult by the usage of privacy and encryption tools. The extraction and analysis of digital evidence from distributed, frequently encrypted cloud systems is the focus of cloud forensics, on the other hand. Significant barriers include questions of jurisdiction, data ownership, and secure access. This presentation examines the evolving field of digital forensics with a particular emphasis on the complex problems that mobile and cloud technologies present.

**Keywords:** Digital Forensic · Digital Evidence · Mobile Forensics · Cloud Forensics

## 1 Introduction

The increased use of mobile devices and the general use of cloud-based services have transformed how individuals and organizations interact with and store digital data in the modern cyber environment. The important domains of mobile and cloud forensics, which entail the methodical extraction, analysis, and interpretation of digital evidence from mobile devices and cloud settings, were created due to this paradigm shift [1–3]. While cloud forensics focuses on the investigation of data stored inside distant cloud infrastructures, mobile forensics deals with the investigation of digital artifacts present within smartphones, tablets, and wearable devices [4–17].

To meet the expectations of the legal, corporate, and law enforcement sectors, the convergence of various disciplines is a crucial aspect of contemporary digital investigations. The complexities of these forensic disciplines are dynamically expanding, bringing previously unknown challenges and opportunities, as the border between mobile and cloud technology continues to blur. It is important to ensure that digital evidence can be traced


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 35–53, 2025.
https://doi.org/10.1007/978-3-031-81573-7_3

For where the clients will be deployed Raspberry PIs are advantageous, because they are 24.5% cheaper than the cheapest x86-64 client that was investigated. This provides an extra seat for every four Cloudgate seats that are deployed, in a rural area where accessibility is more important than speed that extra seat makes a big difference. Also when considering tens or hundreds of deployments those 24.5% savings start to add up.

## References


1. Andrew, M.: Pinet end of life announcement (2020). http://pinet.org.uk/blog/2020/10/27/PiNet-end-of-life.html. Accessed 01 Oct 2021
2. Debian Community. dnsmasq (2020). https://wiki.debian.org/dnsmasq. Accessed 02 Oct 2021
3. Hollingworth, G.: The raspberry pi piserver tool (2018). https://www.raspberrypi.org/blog/piserver/. Accessed 27 June 2021
4. iPXE Community. ipxe (2021). https://ipxe.org/. Accessed 14 May 2021
5. Janina, A.: Teaching with Raspberry PIs and PiNet (2017). https://www.raspberrypi.org/blog/teaching-pinet/. Accessed 27 June 2021
6. Jing, J., Helal, A.S., Elmagarmid, A.: Client-server computing in mobile environments. ACM Comput. Surv. (CSUR) **31**(2), 117–157 (1999)
7. kernel.org. Squashfs 4.0 filesystem. https://www.kernel.org/doc/html/latest/filesystems/squashfs.html. Accessed 01 Oct 2021
8. Lee, D., Won, Y.: Booting linux faster. In: 2012 3rd IEEE International Conference on Network Infrastructure and Digital Content, pp. 665–668. IEEE (2012)
9. LTSP. Linux terminal server project, about. https://ltsp.org/. Accessed 14 May 2021
10. Lucy, H.: Raspberry pi 4 vs raspberry pi 3b+ (2020). https://magpi.raspberrypi.org/articles/raspberry-pi-4-vs-raspberry-pi-3b-plus. Accessed 03 Oct 2021
11. Maga, D., Hiebel, M., Knermann, C.: Comparison of two ICT solutions: desktop PC versus thin client computing. Int. J. Life Cycle Assess. **18**(4), 861–871 (2013)
12. Raspberry Pi Foundation. What is a Raspberry Pi? https://www.raspberrypi.org/help/what-%20is-a-raspberry-pi/. Accessed 13 May 2021
13. Rodeh, O., Bacik, J., Mason, C.: Btrfs: the linux b-tree filesystem. ACM Trans. Storage (TOS) **9**(3), 1–32 (2013)
14. Siebörger, I., Terzoli, A., Hodgkinson-Williams, C.: LTSP DNS round robin clusters: green technology access enablers for telecommunication services in marginalised communities. In: Proceedings of the Southern African Telecommunication and Networks Conference (SATNAC), East London Convention Centre, pp. 393–398 (2011)
15. Terzoli, A., Siebörger, I., Tsietsi, M., Gumbo, S.: Digital inclusion: a model for e-infrastructure and e-services in developing countries. In: International Conference on e-Infrastructure and e-Services for Developing Countries, pp. 85–98. Springer (2017)
16. Thinyane, M., Dalvit, L., Terzoli, A., Clayton, P.: The internet in rural communities: unrestricted and contextualized. ICT Africa **13**, 15–25 (2008)
17. Uimonen, P., Hellström, J.: ICT4D donor agencies and networks. In: The International Encyclopedia of Digital Communication and Society, pp. 1–9 (2015)
18. W3.org. WebDriver, W3C Working Draft (2021). https://www.w3.org/TR/webdriver/. Accessed 21 Sept 2021

### 6.3 Cost of Workstation

As illustrated in Table 3 a Raspberry Pi 4B workstation is 24.5% cheaper than the Cloudgate client this might not look like a significant difference. However for every four Cloudgate workstations that are bought that would have provided six Raspberry Pi 4B workstations. These numbers add up quickly when considering tens or even hundreds of deployments in a lab or community centre.

### 6.4 Network Booting and File Management

In the duration of the research project the usage of network booting shaved off several hours in the management of the different clients, installing software, transferring files between devices, booting the systems, etc.

The Raspberry PIs came with version 2 of the Python programming language, but the Selenium automation suite works only with version 3. Therefore version 3 had to be installed on both setups, and network booting cut the installation time in half. Because the Raspberry Pi clients boot from the same image this was installed once and made available to both clients, this goes for all other software that were installed. Network booting also cut all the installation times to a quarter of the time it would have taken for the x86-64 clients. For the x86-64 clients the required programs had to be installed only in the LTSP server this made them available to all x86-64 clients.

File management is seamless, after downloading, creating or updating a file on one machine, and it does not matter whether it is a Raspberry Pi or x86-64 client. The files are automatically made available to all the other devices that login as the same user that has the file saved. Network booting therefore saves the administrator of these clients a lot of time. Considering the fact that in rural areas skills to manage the clients is limited this is a useful feature. The next section concludes the paper.

## 7 Conclusions

The main objective of this research project was to investigate cost-effective computing infrastructure that can be used in schools or community centres using Raspberry PIs. As a solution to this a LAN that can boot both Raspberry PIs and x86-64 clients was created. The Linux Terminal Server Project was used to implement this research project and support for Raspberry PIs was manually added.

In general Raspberry Pi clients performed poorly compared to x86-64 clients, this is because the Intel processors are well designed for desktop applications. However Raspberry Pi clients are still usable as desktop computers. Given that lightweight applications and the right OS are used they do not crash or significantly slow down. With the Raspberry Pi 4B providing a smoother experience that is comparable to x86-64 clients.

**Table 3.** Price comparison of Clients

| Shop | Intel NUC | Cloudgate | Mecer | Pi 3B+ | Pi 4B |
|---|---|---|---|---|---|
| TakeaLot | R6020.00 | | | R3649.00 | R4549.00 |
| Makro | R5889.00 | | | | |
| Cloudgate | | R5404.00 | | | |
| TechHut | | | R6550.00 | | |
| BidorBuy | | | R7500.00 | | |
| PiShop | | | | R3360.00 | R4080.00 |

## 6 Findings and Discussion

This section presents the findings from the different tests that were run on the two types of clients and general discussion about all the factors that contributed in the results. Because the deployments will make use of Raspberry Pi4B clients, only the Pi4B client is mentioned, and compared to the cheapest x86-64 client, Cloudgate.

### 6.1 Difference Between ARM and X86-64

Raspberry Pi clients use an ARM processor which is designed for lightweight applications. On the other hand x86-64 clients are built for heavyweight applications. This is well illustrated in Fig. 4, where the CPU benchmark test overloads the clients with complex calculations that heavily utilise the CPU. The Raspberry Pi clients perform poorly as compared to x86-64 clients. This is because of the difference in their architecture. Figure 3 illustrates the temperature of the clients when they are overloaded but not doing a lot of CPU heavy calculations and this performance difference is not seen here.

### 6.2 Testing Chromium and LibreOffice

As seen on Table 1 Raspberry Pi clients performed poorly compared to x86-64 clients with the Pi 4B being 6.6 times worse than the Cloudgate client. This 6X difference is however not seen in the LibreOffice tests. The comparison of these two clients on the LibreOffice Writer test is 4.6X in favor of the Cloudgate client. For LibreOffice Calc it is also 4.6X in favor of the Cloudgate client.

The ratio of 6.6X difference is also not observed in the manual tests of Chromium, where the Cloudgate client is 2.5-5X faster depending on the website that is visited. Raspberry Pi clients performed worse at loading the "RUConnected" website than other websites. This could be because of the different versions in Chromium and its drivers that were used in the different clients. Latest versions were used for each architecture as a way to provide optimum performance.

running on the background on the client or a poor cooling system. The former is very unlikely as all running processes were stopped.

### 5.4 Sysbench

The CPU performance of the x86-64 clients is more than 60x better than that of the Raspberry Pi clients as illustrated on Fig. 4, and the difference comes from the two architectures being built for different types of applications. This difference in performance is discussed in detail in Sect. 6.

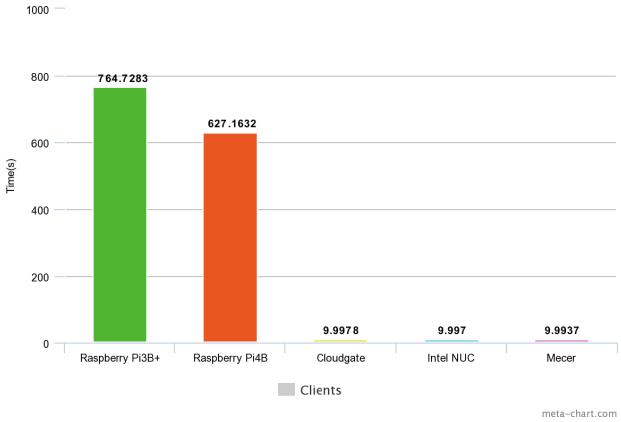



**Fig. 4.** CPU performance of clients

### 5.5 Cost of Clients

This subsection explores the different prices of the clients that were used. The prices are for a full workstation which includes a monitor (19") at the price of R2000.00 and a keyboard + mouse combo at R150.00. These are shown in Table 3.

To ensure that prices are fair and markup on the different clients does not influence the comparisons, we considered the prices from vendors that specialise in each client be used. Vendors that specialise in Intel NUC and Mecer clients were not found, two vendors for each were used as an indication of the range of their cost. This does not affect the conclusions that were made from the cost comparisons. The clients that are compared are the Cloudgate and Raspberry Pi 4B. Vendors that specialise in these two clients were found cloudgate.co.za and pishop.co.za respectively.

**LibreOffice Calc.** For this test the Mecer client finished the execution of the LibreOffice Calc program in an average of 1.4 s. This again represented the fastest execution time compared to all the tested clients. The execution time of the Raspberry Pi 3B+ Rev 1.3 was 22.2 s which represented the worst execution time amongst the bunch. Same as with LibreOffice Writer and Chromium tests most of the recorded time represents the opening of LibreOffice Calc. The actual computation for the x86-64 clients is less than two seconds and less than 5 s for the PIs.

As seen in the results for the LibreOffice Writer and Calc tests (Table 2) there is a 48% and 24.8% difference respectively between the Raspberry Pi 3B+ and Raspberry Pi 4B test results. No conclusive reason was reached for this difference in performance as it is not observed in the Chromium tests and both clients use the same OS. Further investigation needs to be made to identify what the reason for this difference is.

### 5.3 Stress Test User Interface (s-tui)

Because of the Pi4Bs temperature issue that was experienced at the beginning of implementation stress tests were executed on the clients. This was done to test the performance of the clients under 100% utilisation and see if any temperature issues arise, especially on the Raspberry Pi clients. To mitigate any unintended damage on the Raspberry Pi clients the stress tests were only run for a maximum of 10 min. After the 10 min the utilisation percentage of the Pi4B was 99.3% and that of Pi3B+ was 98.9% as illustrated in Fig. 3.

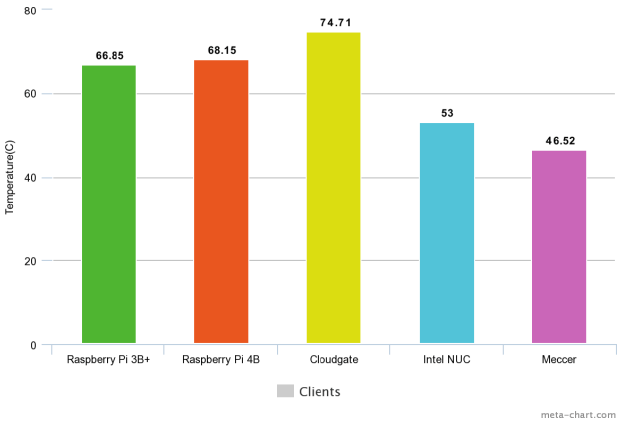



**Fig. 3.** Stress test results

The performance of the Cloudgate client was not expected with a recorded temperature of 74.71 °C. This represented the highest temperature after 10 min of 100% utilisation. This could be due to other, heavy, processes that might be

### 5.1 Chromium Tests

The Chromium web browser was chosen because firstly it comes with both installations of Raspberry Pi OS and that of Ubuntu OS, and secondly no drivers were found for Raspberry Pi for Firefox. The FireFox web browser was used by Ingrid Sieborger [14] in her paper to test her LTSP Round Robin cluster to be used in a rural area.

Table 1 shows the comparison of the clients for the Chromium test.

**Table 1.** Chromium test comparisons

| Clients | Intel NUC | Cloudgate | Mecer | Pi 3B+ | Pi 4B |
|---|---|---|---|---|---|
| Time(s) | 13 | 12.9 | 11.7 | 87.5 | 85.9 |

As can be seen from Table 1, the clients performed comparatively as expected with the Mecer client and out performing the rest with an execution time of 11.7 s. The Raspberry Pi 3B+ had the slowest execution time in the bunch on average taking 87.5 s. Some of the challenges that occurred with this test were the following error messages `NoSuchElementException` and `NoSuchWindowException`. Explicit waits were used to wait for the DOM to load all of the required elements, these waits make up the bulk of the noted times.

### 5.2 LibreOffice Tests

LibreOffice was chosen because it is widely used in the Linux ecosystem and for the fact that it is open source. One of the inhibitors of deployments to rural communities is lack of funds. Therefore the usage of free and open source products is essential to reducing costs.

**Table 2.** LibreOffice Writer and Calc test comparisons

| LibreOffice | Intel NUC | Cloudgate | Mecer | Pi 3B+ | Pi 4B |
|---|---|---|---|---|---|
| Writer(s) | 2.8 | 2.8 | 1.6 | 24.8 | 12.9 |
| Calc(s) | 3.6 | 3.6 | 1.4 | 22.2 | 16.7 |

**LibreOffice Writer.** As was the case with the Chromium browser tests, the Mecer clients out performed all the other clients as expected. However the difference between the execution times of the different x86-64 clients is minuscule, which is 1.3 and 1.1 s on the Intel NUC and Cloudgate clients respectively. The Raspberry Pi 3B+ Rev 1.3 had the worst execution time of 12.9 s. In the LibreOffice tests the actual computation time was instant in x86-64 clients being under a second and for Raspberry PIs under 6 s. Most of the recorded time is made up of the time it takes to open the LibreOffice program.

**Chromium Test Script.** The Selenium framework was used to create a test script for the chromium browser. Webdrivers had to be manually installed for both systems, but a ChromeDriver was sufficient for the x86-64 clients and for the Raspberry PIs, the Chromium WebDriver was used. A WebDriver is a remote control interface that enables the control of user agents. A WebDriver provides a protocol that can be used in multiple programming languages as a way for programs to remotely control the behaviour of a web browser [18].

The script simulates a user using the Chromium browser, where firstly a timer is started. The web browser is then opened in incognito mode and loads all the required elements. It then opens google.com then on the search bar types "RUConnected[1]". It then opens the Rhodes University's RUConnected website and inserts login details in the presented login form, in this case my login details are used. It then navigates to the chosen module page and downloads all past exam question papers. After the files are downloaded the timer stops.

**LibreOffice Test Script.** The LibreOffice test programs were also implemented in Python using an OpenOffice Python package named uno. The Python-UNO package allows the usage of the standard OpenOffice.org API from within a Python script, to develop uno programs.

A timer is started, then the LibreOffice Writer script opens LibreOffice, with an open port to connect to later, using a simple Bash script. On a different thread the script connects to LibreOffice and writes a Lorem Ipsum letter. After the letter is done the timer is stopped and LibreOffice is closed.

Similar to the LibreOffice Writer script, the Calc script starts a timer and opens LibreOffice, with an open port to connect to later, using a simple Bash script. On a different thread the script connects to LibreOffice and opens a spreadsheet it then opens a csv file with student names and marks. The csv file is used to populate two columns of the spreadsheet, the timer is then stopped and LibreOffice closed. The results to these tests and for benchmarking are discussed in the following section.

## 5 Test Results

To test the performance of the different clients, three Python scripts were created to simulate user interaction as detailed in the previous section. Benchmark testing tools were also used to measure the performance of the CPU of the clients. Because the setting of the project research is in a rural community one of the important factors to be compared was the cost of the hardware, which will be presented at the end of this section. All tests were executed 20 times on each client and the average was used to compare results.

[1] RUConnected is the eLearning platform of Rhodes University.

mentioned in the related work section, this directory contains data for servers. In this case the PC is an LTSP server.

Operations that are not wanted for booting the Raspberry PIs are manually deactivated on the file system. These include:

- `dphysswapfile`: this script manages the PIs swap file. A swap file allows Linux to use some of the disk space as RAM. When the device starts running out of RAM, the swap space is used to swap some data from the RAM to the disk space. This frees up the RAM to do more urgent and possibly important operations. When the RAM finishes with the prioritised operations, it swaps back the content from the disk. This functionality is not needed for network booting, the PIs also don't have a disk, therefore the server will take care of all the disk related requirements.
- `resize2fs_once`: this script resizes the root file system to fill the partition of the available disk. Because the Pi does not have a disk this script may have undesired results.
- `raspiconfig`: this is the Raspberry Pi configuration tool, this tool provides the user access to a number of capabilities. These include changing the password of the default/root user, enabling or disabling SSH, screen blanking, etc. Because of the type of setup this computers will be used we don't want learners to have access to such abilities. There will be a dedicated admin user that will have some of these capabilities.

On the other hand, NFS is used to enable the Raspberry Pi Clients to access the file system of the server as they don't have an SD card to locally store data.

### 4.2 Preparing LTSP Server

LTSP was installed and support for x86-64 clients was configured. The installation did not recognize Raspberry Pi clients, this was solved by using Organisational Unique Identifier(OUI). OUIs refer to the first three octets of a MAC address. The Raspberry Pi Foundation has three OUIs these were all listed on the server to be identified as Raspberry Pi clients.

After the server was able to identify Raspberry Pi clients the boot directory as described in Sect. 4.1 was linked to the TFTP server. This is transferred to all clients that are connected to the network and are identified as Raspberry PIs.

### 4.3 Testing Scripts

To simulate the usage of programs that are normally used on deployments in rural communities Python scripts were created. The scripts were written for the Chromium browser as it came with the installations of both Raspberry Pi OS and Ubuntu 20.04. Scripts for automating the testing of LibreOffice Writer and Calc were created as they are often used in such deployments.

**LibreOffice Python Uno Package.** There are three ways to automate LibreOffice, the first one is to control LibreOffice externally, meaning you execute LibreOffice first and run a script that will interact with that instance of LibreOffice. The second one is to run the script internally, meaning before you run open LibreOffice add the script to LibreOffice and execute it. The third way is to add an extension to LibreOffice, this is ideal for adding extra functionality to LibreOffice. The last one is what is used to add extra buttons or other fields.

The first method was chosen for our project, i.e. opening LibreOffice and connecting using a Python script. This method is ideal for the project because the time it takes the clients to open LibreOffice needs to be noted.

**Stress Test and Benchmarking.** The Stress Test Terminal (s-tui) was used to run stress tests on the client machine. Sysbench was used for evaluating the clients CPU performance.

## 4 Implementation

The server and each client have their own peripherals, this is for making the environment as close to the real world as possible. This also makes it easy to test all the clients simultaneously, and to configure and debug them.

### 4.1 Preparing Raspberry Pi Client Image

As illustrated in Fig. 2 the setup contains two Raspberry Pi clients. Because the server is an x86-64 machine this makes the OS that it is running incompatible with the Raspberry PIs as they are ARM based machines. Therefore the OS for the Raspberry PIs had to be manually processed and added to LTSP before these clients can be able to boot from the network.

The first step to achieve this was to choose the appropriate OS for the PIs, the first choice was to use an Ubuntu OS for Raspberry PIs. This was an ideal choice as this would make the user interface for all the clients both x86-64 and PIs the same. This would also make the tests fair for both architectures as they are running the same OS. The OS worked fine on the Raspberry Pi 4 Model B Rev 1.1, however, it was slower compared to when it was running the Raspberry Pi OS. The Ubuntu operating system made the Raspberry Pi 3 Model B Plus Rev 1.3 client unusable as it was constantly crashing. This problem was expected as Ubuntu is designed for computers with a minimum RAM of 4GB and the Raspberry Pi 3 Model B Plus Rev 1.3 that was used only has 2GB of RAM.

After the operating system was chosen the next step was to prepare the ISO file. The Raspbian OS ISO file is manually prepared for network booting, this was done by firstly creating a live loop device. A loop device is like a virtual USB or disk drive. But instead of mapping its data blocks to a physical device it maps it to a regular file in the file system. The boot directory and the rest of the Raspbian OS file system are then copied to the /srv/ltsp/ directory. As

- **Ethernet Cables**: Five cables were used, one to connect the switch to the internet the other five to connect the clients to the switch.
- **5X Peripherals**: Four sets of peripherals were used to interface with the clients and the last one was for the server.

### 3.3 LTSP

As mentioned before, LTSP is a software that uses several tools to enable network booting of multiple Local Area Network(LAN) clients from a single LTSP server [9]. A server has a Linux OS and the clients boot from an "identical" copy of that OS. This makes maintaining tens or hundreds of clients as easy as maintaining a single computer. Some of the tools LTSP uses to enable network booting are:

- **iPXE**: which is a leading open source network booting firmware [4]
- **dnsmasq**: which is a tool that mainly provides two services DNS forwarding and DHCP, it is well suited to providing these services to a small network. Dnsmasq supports both static and dynamic DHCP IP-address leasing and TFTP for network booting of diskless machines [2].
- **NFS**: which is a mechanism for storing files on a network. It allows client computers in turn users to access files over a network, these can then be used and as if they were stored locally.
- **Secure Shell Protocol(SSH) or Lightweight Directory Access Protocol(LPDAP)** these are used to authenticate and authorize users.
- **mksquashfs**: this is a Linux tool that is used to compress files and directories. [7].

### 3.4 Testing Tools

The testing of the clients was carried out in two different phases, firstly testing the execution of commonly used programs. Secondly benchmarking the hardware performance of the clients. The tools that were used to test the common programs are discussed first.

**Selenium.** Selenium is an open-source software that provides an automated testing framework that is used to test if websites function as expected in various browsers and platforms. The Selenium software is a suite of software systems that cater to different use cases. In this project the Webdriver tool was used to test the performance of the clients in opening, navigating and downloading files using the Chromium Web browser.

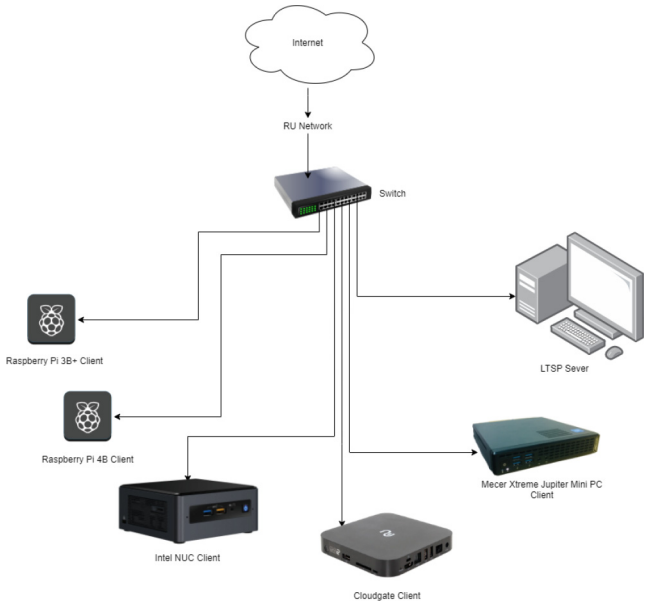



**Fig. 2.** Hardware setup

- **Server**: The server is running Ubuntu 20.04.1 using Linux kernel version 5.11.0, all the other x86-64 computers are running the same OS. The server has an Intel(R) Core(TM) i7-870 CPU with a base frequency of 2.93GHz. The CPU has 4 cores and 8 threads, the L1d, L1i, L2 and L3 caches are 128KiB, 128KiB, 1MiB and 8MiB respectively.
- **Raspberry Pi 4B**: The Raspberry PIs are running the Raspberry Pi OS and it is using kernel version 5.10.0, both PIs use the same OS. The Pi 4B has an ARM Cortex-A72 64-bit quad core processor with a frequency of 1.5GHz and an on-board 802.11ac WiFi. It supports full gigabit Ethernet(throughput not limited) and has 4GB of RAM.
- **Raspberry Pi 3B+**: This Pi has an ARM Cortex-A53 64-bit quad core processor with a 1.4GHz frequency and also has an on board 802.11ac WiFi. It supports gigabit Ethernet(throughput is limited to ca. 300Mbit/s).
- **Mecer Xtreme Jupiter Mini PC**: This client has an Intel(R) Celeron(R) CPU G3930 at a frequency of 2.90GHz. The CPU has 2 cores per socket, 1 thread per core and 4GiB SODIMM DDR4 Synchronous 2133 MHz system memory.
- **Intel NUC**: The Intel NUC has an Intel(R) Celeron(R) J4005 CPU at a frequency of 2.00GHz. The CPU has 2 cores per socket and 1 thread per core and has 4GiB system memory.
- **Cloudgate**: The Cloudgate client has an Intel(R) Celeron(R) J4115 CPU at a frequency of 1.80GHz. The CPU has 4 cores per socket and one thread per socket. This client also has an 8GiB system memory.

### 3.1 Methodology

The research we are presenting here was experimental meaning a number of experiments were carried out to determine whether Raspberry PIs can be used as computing infrastructure in schools/community centres. The following experiments were carried out to determine this:

Booting Raspberry Pi clients over a network using LTSP, because this is important when considering the setting where these clients will be deployed. The skills to maintain and manage the clients are limited, and network booting reduces the time the administrator spends on each deployment site as they only have to manage only one device.

Testing the clients was carried out automatically and manually, automatic tests were used to find the computational consistency of the clients. And manual test were used to see how the clients will perform in real time with a human across the screen.

### 3.2 Hardware

The hardware infrastructure design of the project is illustrated in Fig. 2, the server and all the clients are connected to the Switch with Ethernet cables. The three x86-64 clients were chosen because those are the type of computers that are used in such deployments.

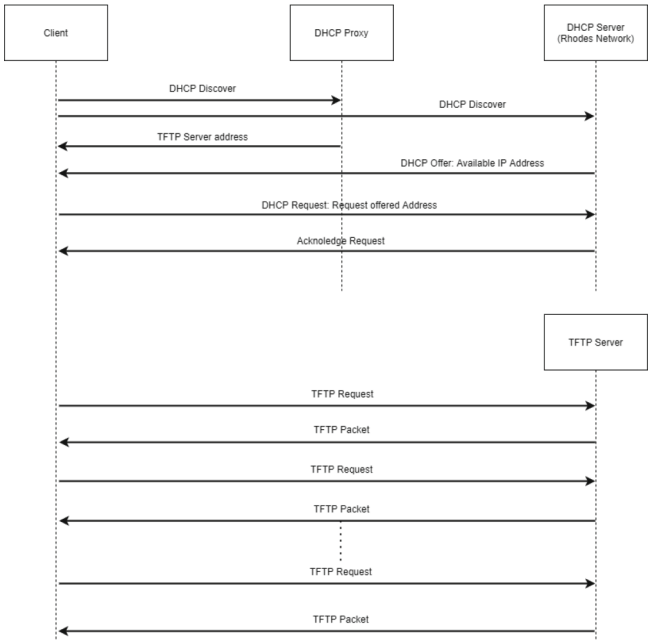



**Fig. 1.** Client, DHCP and TFTP Server interactio

Figure 1 illustrates the process that happens in the implemented system. The DHCP Proxy and TFTP server are hosted in the same computer as the LTSP server. Because of Rhodes University regulations the clients get their IP addresses from the Rhodes University/Hamilton building networks DHCP server.

### 2.6 Related Work

There are various open source Linux projects developed, for community use, that utilize network booting with easy to use disk-less computers. The two most popular ones are presented here as they relate to our project.

**PiNet.** PiNet is an open source and free project that was developed by Andrew Mulholland. It was developed alongside teachers from all over the world. It enables teachers to manage a whole classroom of Raspberry PIs from a single computer making administration and maintenance tasks very easy [5]. This project was phased out in October 2020 and someone else is yet to pick it up.

PiNet is based on LTSP5, and this is a problem because LTSP has been redesigned from scratch and does not support prior versions. For the project to use the later versions of LTSP it has to also be redesigned from scratch. Though this is less important and easy to fix it is worth mentioning, PiNet uses Raspbian Stretch therefore does not support the latest version of Raspberry PIs, Pi4 [1].

**PiServer.** PiServer was developed by the Raspberry Pi team, similar to PiNet it enables users to manage Raspberry PIs from a single x86-based server(central computer) running the x86 version of Raspbian OS [3]. PiServer enables users to network boot generic Pi clusters and is targeted to a larger audience than only schools, this the main target audience for PiNet.

PiServer does not use LTSP to boot it's clients, it does however use the same tools to enable network booting. These include DHCP for leasing IP addresses to clients, Lightweight Directory Access Protocol(LDAP) for authentication and authorization and NFS for sharing the servers file system.

Both of these projects boot over a network and this is what enables them to manage the Raspberry PIs(clients) from a single personal computer(server). The clients share a single file system and users can login any of the client devices and find their data there. However both projects only support network booting Raspberry Pi clients and have user interfaces for configuration. This prevents users from connecting other diskless computers they might have and abstracts the operations that happen under the hood making it tricky to customize. The project we are presenting in this paper solves these problems, some of the methods that went into the designs and testing of the system are discussed next.

## 3 Methodology and Design

This chapter looks into the design of the system we developed. The methods that were used to carry out the research, test the clients and programs that were used in the tests are discussed in this section.

### 2.3 Raspberry Pi

The Raspberry Pi is the name of a series of inexpensive credit-card sized computers that can be used with common peripheral devices that are normally used with desktop computers like a monitor, standard keyboard and mouse. It can do everything that you would expect a desktop computer to do from browsing the Internet to playing video games [12].

The latest version of the Raspberry Pi is the fourth-generation Raspberry Pi computer series, it has a 1.5GHz clock speed processor, with RAM that is up to 8GB, a gigabit Ethernet adapter, 2.4GHz and 5.0GHz IEEE 802.11ac wireless, Bluetooth 5.0, 2*USB 2 and 2*USB 3 ports [10]. These specifications make the Raspberry Pi ideal for creating a cost effective computing infrastructure, which can also boot over a network.

### 2.4 Booting

Booting refers to a sequence of events/operations that happen before the computer is ready to display content. The events are called a boot sequence, and each computer has a boot sequence. The Linux boot process is completed by copying the kernel binary image to the secondary storage on the RAM disk, and then loading it into the main memory, and finally running the kernel image. [8]. This involves a number of programs working together to achieve the final results.

These programs include the Master Boot Loader(MBR), where its main function is to identify where the OS is located and load it into the RAM. It also contains information about the GRUB. GRUB stands for GRand Unified Boot Loader, it is the most common boot loader for Linux systems. The GRUB splash screen is typically the first thing that appears when a computer is turned on, this can be used to select a kernel image. The GRUB file also starts the `init` program, this is always the first program to be run. The kernel then loads a temporary root file system using `initrd` which stands for Initial Ram Disk. At this point the system executes run level programs, afterwards the real root file system can be mounted.

### 2.5 Network Booting

Network booting is similar to normal booting and the key difference is that the computer gets the image from a server in the network. This enables a computer to load an operating system directly from the network without any attached local storage devices like a SSD, HDD, USB, SD card, etc. This is made possible with the help of technologies like the iPXE. Preboot Execution Environment(PXE) is a client-server interface that allows computers in a network to be booted from the server before deploying the obtained PC image, and iPXE is an open-source implementation of PXE. The other technologies that are typically used are the Dynamic Host Configuration Protocol(DHCP) server for assigning IP addresses. The Trivial File Transfer Protocol(TFTP) server to store and transfer the boot program when requested, and the NFS to enable client computers to access files in the server computer.

**Fat/Thick Clients.** Thick clients on the other hand do the bulk of their processing locally. Unlike thin clients the type of communication that they have with the server has to do with storage. This includes requesting or updating archival information on the server. Thick clients have better processing power than thin clients, because all processing is done locally they tend to be a little bit faster too. One of the major advantages of thick clients is the fact that they can be used independently, without the server and perform exactly the same way they did when connected to the server. This is one of the key reasons thick clients are used in the implementation of the project we are presenting in this paper. That is, to be able to use them independently in a case where the server fails.

### 2.2 Linux File System

File systems are designed to structure the storage of non-volatile information/data, this is done through providing a name space and metadata structure [13]. Desktop computers need to be able to store data on a hard disk or a similar type of memory, like a USB drive. Firstly, it is non-volatile meaning it does not require a constant power in order to retain the stored information. Unlike Random Access Memory(RAM) in which after power is off the stored content is deleted, in disk memory data does not get deleted. Secondly, disk storage is inexpensive compared to RAM memory.

A file system also needs an Application Programming Interface(API) that enables either the system or users to manipulate its objects like files or directories with restricted access. Some of the most popular tasks include creating or deleting a file or directory, and moving or copying them [13]. The API does this by using algorithms which efficiently determine where files are stored and how to get them. Because it is free and open source, we opted to use the Linux File System.

**Linux Directory Structure.** The directories in Linux are structured in a tree like hierarchy with root at the top of the tree.

Some of the noteworthy directories in the Linux file system are listed and their functions are explained below:

- **root**: is the home directory of the superuser this is typically the administrator of the system. This user has unlimited privileges.
- **boot**: This is where the files that are needed to startup the computer are stored. These includes the grub, bootloader and kernel.
- **mnt**: This is a temporary mount-point for regular file systems.
- **src**: This directory contains data for servers. For example this is where HTML(/src/http) or TFTP(/srv/tftp or /srv/www/) files are stored when a web or TFTP server from a Linux system.

The family of Linux Operating System(OS) are used in a variety of settings, one of which is in mini computers. Raspberry Pi OS which was created by the Raspberry Pi Foundation to be used on their Raspberry Pi mini computers is one of such applications.

most likely worsen as it will get harder for them to participate in the digitized environment.

The main objective of this study is to implement the support of Raspberry Pi clients on LTSP which can be used as workstations in either rural schools or community centres. When providing computing infrastructure to marginalised communities there are problems that one comes across, one of which being lack of funds [17]. Therefore it is always better to opt for open-source and affordable software and hardware alternatives. That is why Raspberry PIs and the LTSP were used in this project. Before starting the research project a study of the background and other projects that are related to it was undertaken. The next section discusses the background and related works.

## 2 Background and Related Work

In this section all the necessary concepts that were required to successfully achieve the objectives of the research project are reviewed and briefly discussed. The first topic to get discussed is the Client/Server model:

### 2.1 Client/Server Model

The client-server model is a distributed model in which the server provides a service, resource and computational power etc. to the clients. The client relies on sending requests to the server in order to gain access to services that it requires. Depending on the type of client it is, the client depends on the server to do some operations. Two of the most popular types of clients for the client/server model are thin and fat/thick clients. There are numerous similarities between thin and thick clients in both cases the client sends a request and receives responses from the server. The server, in both cases, acts as a middleman.

Because of its easiness and advantage in protecting data with access control and security policies, the client/server model has been has been well adopted and is used in diverse settings. Some of the most popular examples that this model is used include: Network File System (NFS), File Transfer Protocol (FTP), and Hypertext Transfer Protocol (HTTP) [6]. As mentioned above, clients can be either thin or thick and the below sections explain their difference:

**Thin Client.** A thin client is designed such that all, with the exception of controlling peripherals, processing is done on the server. The client functions as a stripped down terminal to the server and it requires constant communication with the server to do all computation [11]. These types of clients provide a seamless experience for the user and enable them to interact with the client as though they are working on the server computer. Thin clients do not have a disk, are inexpensive and they are typically unusable/unreliable without the server.

# Investigating Cost-Effective Computing Infrastructure for Schools/Community Centres Using Raspberry PIs

Live Tembiso, Zelalem Shibeshi(✉), and Alfredo Terzoli

Department of Computer Science, Rhodes University, Grahamstown, South Africa
{z.shibeshi,a.terzoli}@ru.ac.za

**Abstract.** Computing infrastructure plays a significant role in our lives. This has been made even more apparent by the Corona virus pandemic. However a large number of people still do not have access to any form of end user computing infrastructure. This paper looks at the reliability and feasibility of using Raspberry PIs as a form of computing infrastructure for communities that would otherwise not have access to computers. The Raspberry PIs are compared to other mini computers that use x86-64 processors. In the implementation of the project an LTSP (Linux Terminal Server Project) network that supports both x86-64 and Raspberry Pi clients was created in order to test both systems in the same network. LTSP makes it easy to boot LAN clients from a single image. Because LTSP does not support Raspberry Pi clients this support was added in manually. Three programs that are commonly used in deployments of this nature were tested, LibreOffice Writer, LibreOffice Calc and the Chromium browser, using custom Python scripts. To test and compare the performance of the different clients sysbench and s-tui benchmark tests were used to benchmark the CPU and I/O performance of the clients. The results show that the raw computational power of the x86-64 clients is 60x better than that of the Raspberry Pi clients. The usage of the different types of clients is comparable with the x86-64 clients being 2.5-6x faster than the Raspberry Pi clients.

**Keywords:** LTSP · Raspberry Pi · Linux file system · ICT4D

## 1 Introduction

Like many in developing nations, a large number of South Africans still don't have access to the Internet due to various reasons [15]. This is a difficult truth to come to terms with especially as we are living in the information age and connectivity is an essential component of access to information. Some of the key reasons for the lack of connectivity are cost of computing infrastructure and scarcity of skills [16]. If this problem persists to hinder marginalized communities, which are mainly poor and have below-average education, their situations will


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 18–34, 2025.
https://doi.org/10.1007/978-3-031-81573-7_2

wireless network virtualization depends on specific access technologies, and the wireless network contains many more access technologies compared to wired network virtualization and each access technology has its own unique characteristics, which makes convergence, sharing and abstraction difficult to achieve [11,17]. How to define a VM migration method taking into account the constraints in a wireless network environment?

## References


1. Alain, F.: Qu'est-ce-que la virtualisation? (2019). https://www.piloter.org/techno/support/virtualisation.htm. Accessed 13 July 2019
2. Benbrahim, S.E.: Migrations en temps réel des machines virtuelles interdépendantes. Ph.D. thesis, École Polytechnique de Montréal (2016)
3. Benedicte, B.: Migration informatique: le guide pour reussir (2021). https://blog.hubspot.fr/marketing/migration-informatique. Mis en ligne le 29 Novembre 2021. Accessed 11 Dec 2021
4. Chowdhury, N.M.K., Boutaba, R.: A survey of network virtualization. Comput. Netw. **54**(5), 862–876 (2010)
5. Quelle est la configuration requise pour windows xp? (2022). https://frameboxxindore.com/fr/windows/what-are-the-minimum-requirements-for-windows-xp.html. Accessed 12 May 2022
6. La migration informatique, qu'est-ce que c'est? (2021). https://www.redhat.com/fr/topics/automation/what-is-it-migration. Mis en ligne le 04 Fevrier 2021. Accessed 11 Dec 2021
7. Qu'entend-on par migration informatique? (2021). https://www.groupesl.com/2021/08/02/migration-informatique. Mis en ligne le 02 Aout 2021. Accessed 11 Dec 2021
8. Jacques, L.: Introduction aux systèmes informatiques: architectures, composants, mise en œuvre, pp. 1–2. Dunod (2017)
9. Keshavamurthy, U., Guruprasad, H.: VM migration: a survey. Global J. Eng. Sci. Res. (2015)
10. Kherbache, V.: Ordonnancement des migrations à chaud de machines virtuelles. Ph.D. thesis, Université Côte d'Azur (2016)
11. Liang, C., Yu, F.R.: Wireless network virtualization: a survey, some research issues and challenges. IEEE Commun. Surv. Tutor. **17**(1), 358–380 (2014)
12. Passante, B.: Quelle est la difference entre la bande wifi 2.4 ghz et la 5 ghz ? (2022). https://sogetel.com/aide/internet/quelle-est-la-difference-entre-la-bande-wi-fi-2-4-ghz-et-la-5-ghz. Accessed 12 May 2022
13. Pham, T.S.: Autonomous management of quality of service in virtual networks. Ph.D. thesis, Université de Technologie de Compiègne (2014)
14. Popek, G.J., Goldberg, R.P.: Formal requirements for virtualizable third generation architectures. Commun. ACM **17**(7), 412–421 (1974)
15. Reseau, I.: Qu'est-ce-qu'un réseau informatique ? (2019). https://primabord.eduscol.education.fr/qu-est-ce-qu-un-reseau-informatique. Mis en ligne le 07 juin 2016. Accessed 13 July 2019
16. Venkatesha, S., Sadhu, S., Kintali, S.: Survey of virtual machine migration techniques. Memory (2009)
17. Wang, X., Krishnamurthy, P., Tipper, D.: Wireless network virtualization. In: 2013 International Conference on Computing, Networking and Communications (ICNC), pp. 818–822. IEEE (2013)

$$v \begin{cases} c_v : 64MB \\ d_v : 1Go = 1000\ MB \\ P_v : 233\,MHz \\ k_v : 0no - devices \\ l_v : 10software \\ r_v : 10(LAN) \\ b_v : 0 \end{cases}$$

$c_v + d_v = 64 + 1000 = 1064$

$p_v + k_v + l_v + r_v + b_v = 233 + 0 + 10 + 10 + 0 = 273$

We have: $c_v + d_v > p_v + k_v + l_v + r_v + b_v$. What's more, today's computers are v e r y powerful and their RAM and hard disk capacities can reach the order of gigabytes (for RAM) or terabytes (for hard disks).

## 8 Conclusion and Perspectives

This paper looks at the migration of virtual machines, which is fast becoming a must in data centres, as server hosting capacities are becoming more and more elevated. We began our discussion with an overview of IT migration. Here, we presented a brief overview of the IT migration environment and its motivations. We then discussed the notion of machine migration. In particular, we have formally defined the notions of machine and machine migration. We then discussed the concept of virtual machine migration. Specifically, we have defined the formal characterisation of a virtual machine and the formal language formulation of the virtual machine migration process. We then showed that the virtual machine migration problem is an NP-hard problem. Finally, we proposed an optimal algorithm for solving this problem, and discussed our algorithm.

In summary, the migration of virtual machines is used on a daily basis to improve application performance, reduce power consumption and increase the efficiency of the system. This is a wide-ranging field, with a number of research questions and challenges coming to the fore. It is quite broad and a number of research questions and challenges are coming to the fore. Several improvements can be made to this work in future projects.

A first perspective to this work is to find a mathematical model allowing to evaluate the total migration time of a VM in accordance with our proposed migration approach. Then compare our migration approach with those in the literature.

Another perspective would be to extend our method to the simultaneous (grouped) migration of parallel and interdependent virtual machines [2].

Migrating a VM from one physical host to another in a multi-hop network must take into account different path parameters such as: bandwidth, number of nodes and distance. How can we implement a migration method that defines the optimal migration path in order to reduce the transfer time for migration packets?

Virtualisation, irrespective of wired or wireless networks, can be considered as a process dividing the entire network system [11,17]. However, the distinctive properties of the wireless environment, in terms of time-varying channels, attenuation, mobility, broadcast, etc., make the problem more complicated. In addition,

$D(d_v, d_2) = D(1000, 4000) = 1$ (because $1000 < 4000$)
We have: $\eta(v) < \eta(m_1)$
The vector associated with the virtual machine has the coordinates:
$vm = (64, 1000, 233, 0, 4, 100, 0, 128, 2000, 512, 2, 4, 10, 0)$
Let $vm^{'}$ be the vector representing the data of the virtual machine $vm$ running at a given time. it can have for coordinates:
$vm^{'} = (32, 512, 233, 0, 2, 100, 0, 128, 2000, 512, 2, 4, 10, 0)$
The total weight of bag $m_2$ is:
$W = 256 + 4000 + 1000 + 2 + 8 + 10 + 0 = 5276$
Given that the various migration conditions are verified, we can carry out the migration of the $vm$ through the algorithm by passing the parameters as follows: $MigrationVM(32, 512, 233, 0, 2, 100, 0, 128, 2000, 512, 2, 4, 10, 0)$.

## 7 Discussion

In practice, the migration of a running virtual machine from a physical machine to another can be summarized in the following steps:

(1) saving the context in which the VM is run;
(2) creation of the VM on the target machine;
(3) copying VM data from the source machine to the destination machine;
(4) destruction of the VM in the source machine;
(5) restoring the execution context of the VM.

The data transfer time of a running VM relies mainly on the data of the vector v, since it is the vector that effectively represents the data transfer time of a running VM and information about the hosting physical host (vector $m$) need not be moved, since the VM will be hosted on the destination host with a similar architecture.

We can further reduce the migration data for the vector $vm$ to 02 These are the main elements: **the amount of RAM used by the VM, the amount of hard disk used by the VM**, because these two parameters have a very large amount of data to transfer and are constantly modified during the VM's operation, which is not the case for the others: $c_v + d_v > p_v + k_v + l_v + r_v + b_v$. In addition, when the VM's execution context is saved, information such as the virtual processor frequency, the number of devices used by the VM, the number of software applications, the number of network connections and the virtual bandwidth is saved and can be configured on the target host responsible for hosting the migrating VM.

*Proof.* Let's take the minimum characteristics of a personal computer defined in Sect. 3 as being the values of $v$ when the VM is running. Taking randomly the other values of $v$ we have:

**Algorithm 2:** MigrationSend

**Enter**: $V[]$ an array of real numbers, of length 14, such that $V = (c, d, p, k, l, r, b, c^{'}, d^{'}, p^{'}, k^{'}, l^{'}, r^{'}, b^{'})$.
**Output**: A stack of reals of length 14.

```
start
    V'[]A stack of reals of length 14;
    for i from 7 to 1 do
        Stack(V', V[i]);
        Stack(V', V[i + 7]);
    return (V');
```

**Algorithm 3:** MigrationReceive

**Enter**: $P[]$ a stack of real numbers of length 14.
**Output**: An array of real numbers, length 14.

```
start
    W[] an array of real numbers of length 14;
    for i from 1 to 7 do
        W[i] = Unstack(P);
        W[i + 7] = Unstack(P);
    return (W);
```

**NB:** The $Stack()$ and $Unstack()$ functions used in these algorithms are operations in the algorithmic data structure known as the "stack".

### 6.2 Illustration De Notre Solution

To illustrate our approach, let's consider the example of the migration of a VM below:
Let $m_1$ and $m_2$ be two physical machines with the following characteristics:

$$m_1 \begin{cases} c_1 : 128\ MB \\ d_1 : 2\ Go = 2000\ MB \\ P_1 : 512\ MHz \\ k_1 : 2\ devices \\ l_1 : 10\ softwares \\ r_1 : 10(LAN) \\ b_1 : 0 \end{cases} \qquad m_2 \begin{cases} c_2 : 256\ MB \\ d_2 : 4\ Go = 4000\ MB \\ P_2 : 1\ GHz = 1000\ MHz \\ k_2 : 2\ devices \\ l_1 : 8\ software \\ r_2 : 10(LAN) \\ b_2 : 0 \end{cases}$$

Let $v$ be a virtual machine of the machine $m_1$ with the following characteristics:

$$v \begin{cases} c_v : 64MB \\ d_v : 1Go = 1000\ MB \\ P_v : 233\,MHz \\ k_v : 0(no - devices) \\ l_v : 4softwares \\ r_v : 100(WAN) \\ b_v : 0 \end{cases}$$

this transfer is: **the algorithm by priority** because the data to be transferred during migration does not require an order of priority, for example data from the central memory is the first data to be transferred, after data from the hard disk then data from the hard disk.

### 6.1 Description of the Algorithm of Our Solution

To lay the foundations for this part of our work, we make the following assumptions:

- $m_1$ and $m_2$ are personal computers;
- $m_1$ and $m_2$ belong to a local unicast network;
- $m_1$ and $m_2$ are switched on and meet the conditions described in Sect. 2.1;
- the virtual machine $vm$ is created;
- the virtual machine $vm$ is running;
- priority is assigned to $vm$ components during migration as follows: 1 for $c_v$, 2 for $c_1$, 3 for $d_v$, 4 for $d_1$, 5 for $p_v$, 6 for $p_1$, 7 for $k_v$, 8 for $k_1$, 9 for $l_v$, 10 for $l_1$, 11 for $r_v$, 12 for $r_1$, 13 for $b_v$, 14 for $b_1$.

The $vm$ virtual machine will be migrated as follows:

1. for each component of the $vm$ vector currently running, their value must be stacked in decreasing order of priority, i.e. from number 14 to number 1.
2. the migration traffic containing the stack is sent to the destination machine $m_2$ (see Algorithm 2);
3. Once the stack arrives in the $m_2$ machine, it will be unstacked according to the defined priority order (see Algorithm 3).

The Algorithm 1 of complexity $O(7)$ is the main algorithm used to perform the migration, the Algorithm 2 of complexity $O(7)$ is used to send the migration packet and the Algorithm 3 of complexity $O(7)$ is used to receive the migration packet.

**Algorithm 1:** Migration algorithm (MigrationVM)

**Enter**: $T[]$ an array of reals, of length 14, representing the migrated $vm$, such that $T = (c_v, d_v, p_v, k_v, l_v, r_v, b_v, c_1, d_1, p_1, k_1, l_1, r_1, b_1)$.
**Output**: An array of reals, of length 14, representing the $vm$ migrated.

```
start
    V[] a stack of reals of length 14;
    T'[] an array of reals of length 14;
    V = migrationSend(T);
    T' = migrationReceive(V);
    return (T');
```

We want to migrate the virtual machine $vm$ from a machine $m_1$ to $m_2$ taking into account the execution parameters of each machine and that of the virtual machine. The problem is to find the 14-tuplets $(c_v, d_v, p_v, k_v, l_v, r_v, b_v, c_1, d_1, p_1, k_1, l_1, r_1, b_1)$ such that the following conditions are met:

- $D(d_v, d_2) = 1$;
- $\eta(v) \leq \eta(m_1)$;
- $V(vm) = vm^{'}$ whwere $vm^{'} = (v, m_2)$.

This problem can be reduced to the problem of the rucksack where:

- $vm = (c_v, d_v, p_v, k_v, l_v, r_v, b_v, c_1, d_1, p_1, k_1, l_1, r_1, b_1)$ is the vector of objects to be transported;
- $m_2$ the bag where we are going to place the objects;
- the total weight of the bag is: $W = c_2 + d_2 + p_2 + k_2 + l_2 + r_2 + b_2$;
- the content of each component of the vector $vm$ represents the weight of each object to be transported;
- the object values are the data for the VM currently running;
- the rucksack $m_2$ must be filled in such a way that the transfer time of the object values is minimal while respecting the weight constraint. In other words: $\sum(a_i) \leq W$, where the $a_i, i \in [1; 14]$ are the weight of each object to be transported.

**Conclusion:** The VM migration problem is an NP-hard problem because the rucksack problem is an NP-hard problem.

## 6 Virtual Machine Migration Approach

Let $m_1$ and $m_2$ two physical machines located respectively in execution environments $A$ and $B$ such that : $m_1 = (c_1, d_1, p_1, k_1, l_1, r_1, b_1)$ and $m_2 = (c_2, d_2, p_2, k_2, l_2, r_2, b_2)$, with $(c_i, d_i, p_i, k_i, l_i, r_i, b_i)_{1 \leq i \leq 2} \in \mathbb{R}_+^7$ and $c_i$ represents the capacity of the central memory, $d_i$ the capacity of the hard disk, $p_i$ frequency processors, $k_i$ the number of devices, $l_i$ the number of software, $r_i$ the type of network to which the machine belongs, $b_i$ the capacity of the bandwidth.

Let $vm$ be a virtual machine of machine $m$ such that: $vm = (v, m_1)$, where $v = (c_v, d_v, p_v, k_v, l_v, r_v, b_v)$ represents the vector associated with the states of the virtual main memory, the capacity of the virtual hard disk, the frequency of the virtual processors, the number of devices connected to the VM, the number of software applications in the VM, the number of network connections in the VM and the virtual bandwidth.

Let $V : E \rightarrow F$, be the virtual machine migration function such that $V$ is subjective and $E, F \in \mathbb{R}_+^{14}$ respectively represent the source environment and the VM's target environment.

We want to migrate the virtual machine $vm$ from a machine m running in the physical machine $m_1$ is minimal. The optimal algorithm to carry out

# 5 Virtual Machine Migration Problem

## 5.1 Description of the Problem

In a network virtualization environment, the migration of equipment must be managed efficiently. This migration is not limited to just the migration that of virtual machines from one virtual network to another, but also for the transfer of virtual routers [4].

When a host or a physical router encounters a breakdown, has a problem of maintenance, computing power, energy saving, it is necessary to start the migration of the VMs if ex running in this host in order to guarantee a good quality of service to users. The major problem of the migration of VMs resides in the need to stop the VM throughout the duration of its move. The applications and services that run on VMs are generally critical and it is therefore difficult to consider stopping them even for a short period [10]. The main issue in migration VMs is therefore to minimize this downtime in order to give the user as much illusion as possible that the machine has never stopped. Thus, transferring the migration traffic of a VM from one failing host to another while ensuring minimal downtime of that VM is an issue that requires attention [10].

Furthermore, for VM migration to be possible, the target machine must have sufficient memory space to host the migrating VM. Given that a VM can be connected to several networks then the migration must keep its various connections active in order to ensure continuity of service.

In addition, the migration of a VM from one physical host to another in a multi-hop network must take into account the various path parameters, namely: bandwidth, number of nodes and distance. In such a case, what will be the optimal migration path allowing us to reduce the transfer time of migration packets?
The research problem is the optimization of the migration time of the virtual machine in the networks.

## 5.2 Mathematical Formulation

Given $m_1$ and $m_2$ two physical machines located respectively in execution environments $A$ and $B$ such that: $m_1 = (c_1, d_1, p_1, k_1, l_1, r_1, b_1)$ and $m_2 = (c_2, d_2, p_2, k_2, l_2, r_2, b_2)$, with $(c_i, d_i, p_i, k_i, l_i, r_i, b_i)_{1 \leq i \leq 2} \in \mathbb{R}_+^7$ and $c_i$ represents the capacity of the central memory, $d_i$ the capacity of the hard disk, $p_i$ frequency processors, $k_i$ the number of devices, $l_i$ the number of software, $r_i$ the type of network to which the machine belongs, $b_i$ the capacity of the bandwidth.

Let $vm$ be a virtual machine of machine $m_1$ such that: $vm = (v, m_1)$, where $v = (c_v, d_v, p_v, k_v, l_v, r_v, b_v)$ represents the vector associated with the states of the virtual main memory, the capacity of the virtual hard disk, the frequency of the virtual processors, the number of devices connected to the VM, the number of software applications in the VM, the number of network connections in the VM and the virtual bandwidth.

Let $V : E \rightarrow F$, be the virtual machine migration function such that: $E, F \in \mathbb{R}_+^{14}$, respectively represent the source environment and the VM's target environment.

A VMs operating system environment can be migrated from one physical machine to another, provided that there are sufficient similarities between the system architectures of these host machines [16].

Given $m_1$ and $m_2$ two physical machines such that: $m_1 = (c_1, d_1, p_1, k_1, l_1, r_1, b_1)$ and $m_2 = (c_2, d_2, p_2, k_2, l_2, r_2, b_2)$. We can formally define the migration of a VM from an initial environment $A$ to a final environment $B$ by a function $V : A \to B$. $A$ and $B$ sont des espaces vectoriels avec $A = \mathbb{R}_+^{14}$ and $B = \mathbb{R}_+^{14}$.

We make the following assumptions:

1. $m_1, m_2 \in \mathbb{R}_+^7$ and verify the assumptions made in Sect. 3.1;
2. $c_1 \leq c_2$, $d_1 \leq d_2$, $p_1 \leq p_2$ and $b_1 = b_2$;
3. $D(d_v, d_2) = 1$, where $D$ is a Boolean function used to say whether the hard disk $d_2$ contains enough space to host the VM (1=Yes et 0=No);
4. $\eta(v) \leq \eta(m)$;
5. $V$ a subjective function;
6. $\forall vm_1 \in A, V(vm_1) = vm_2$ avec $vm_2 \in B$;
7. $V(vm_1) = V(v, m_1) = (V(v), V(m_1)) = (v, m_2) = vm_2$

**Let's show that $V$ is a subjective function:**

Given $y \in B$, let's find $x = (v_1, m_1) \in A$ such that: $y = V(x)$,
with $v_1 = (c_1, d_1, p_1, k_1, l_1, r_1, b_1) \in \mathbb{R}_+^7$ and $m_1 = (c_0, d_0, p_0, k_0, l_0, r_0, b_0) \in \mathbb{R}_+^7$.
$y \in B \Rightarrow y = (v, m)$, where $v = (c_v, d_v, p_v, k_v, l_v, r_v, b_v) \in \mathbb{R}_+^7$ and $m = (c, d, p, k, l, r, b) \in \mathbb{R}_+^7$.
$\exists c_2, d_2, p_2, k_2, l_2, r_2, b_2$ such that: $c_2 Rc, d_2 Rd, p_2 Rp, k_2 Rk, l_2 Rl, r_2 Rr$ and $b_2 Rb$, where $R$ is the order relation in $\mathbb{R}_+^*$.
$\exists c_3, d_3, p_3, k_3, l_3, r_3, b_3$ such that: $c_3 Rc_v, d_3 Rd_v, p_3 Rp_v, k_3 Rk_v, l_3 Rl_v, r_3 Rr_v$ and $b_3 Rb_v$, où $R$ is the order relation in $\mathbb{R}_+^*$.
As $c \leq c_0$ we can take $c_0 = c$ because $c \leq c$.
As $d \leq d_0$ we can take $d_0 = d$ because $d \leq d$.
As $p \leq p_0$ we can take $p_0 = p$ because $p \leq p$.
As $b \leq b_0$ we can take $b_0 = b$ because $b \leq b$.
We can also take $p_0 = p_2, k_0 = k_2, l_0 = l_2, c_1 = c_3, d_1 = d_3, p_1 = p_3, k_1 = k_3, l_1 = l_3, r_1 = r_3, b_1 = b_3$.
So for $x = (v_1, m_1)$ with $v_1 = (c_3, d_3, p_3, k_3, l_3, r_3, b_3)$ and $m_1 = (c, d, p, k_2, l_2, r_2, b)$, we have bel and well $V(x) = y$.

*Property 4. We talk about virtual wire migration when:* $r_1 = 1$ *or* $r_1 = 10$ *or* $r_1 = 11$ *or* $r_1 = 100$.

*Property 5. We talk about virtual wireless migration when:* $r_1 = 101$ *or* $r_1 = 110$ *or* $r_1 = 111$ *or* $r_1 = 1000$.

Popek et Goldberg [14] define a virtual machine as an environment created by a virtual machine monitor (VMM) or hypervisor. A VMM is the software layer providing virtualization[1]. It virtualizes all the resources of a physical machine, thus defining and supporting the running of several virtual machines. There are several VMMs, including Vmware workstation, Xen and virtual box [13].
An essential feature of a virtual machine is that the software running in it is limited to the resources and abstractions provided by the VM [16].

### 4.2 Formal Definition of a Virtual Machine

Let $m$ a machine such that $m = (c, d, p, k, l, r, b)$, where $c, d, p, k, l, r, b$ represent respectively: the capacity of the main memory, the capacity of the hard disk, the frequency of the processors, the number of connected devices, the number of software programs, the type of network and the capacity of the bandwidth.

We assume that a VM belongs to a physical machine $m$ and is defined by a state vector: $vm = (v, m)$, with $v = (c_v, d_v, p_v, k_v, l_v, r_v, b_v)$, representing the vector associated with the states of the virtual main memory, the capacity of the virtual hard disk, the frequency of the virtual processors, the number of devices connected to the VM, the number of software applications in the VM, the number of VM software, the number of VM network connections and the virtual bandwidth if the VM is a virtual router.

We make the following assumptions:

1. $vm \neq 0$;
2. we will consider that VM is active and that it contains data currently being executed;
3. $c_v, d_v, p_v, k_v, l_v, r_v, b_v$ verify the assumptions of Sect. 3.1;
4. $c_v < d_v$
5. $d_v < d$;
6. $c_v < c$;
7. $p_v < p$;
8. $k_v \leq k$;
9. the norm of the vector $vm$ is defined as follows:

$$\eta(vm) = \sqrt{c_v^2 + d_v^2 + p_v^2 + k_v^2 + l_v^2 + r_v^2 + b_v^2 + \eta(m)^2}.$$

### 4.3 Migration of Virtual Machines

Virtual machine migration is the ability to move the operating system instance from one physical machine to another [16].

[1] The virtualization is the set of techniques making it possible to dissociate the characteristics physical characteristics of a hardware or software system user-oriented applications [1].

### 3.3 Machine Migrations

Given that a machine is a component of a computer system, we can define machine migration as the movement of a machine from one operating environment to another.

We can formally define migration as a function $M : A \rightarrow B$, where $A$ denotes the initial environment and $B$ the final environment. $A$ and $B$ are vector spaces.

We make the following assumptions:

1. $A, B \in \mathbb{R}_+^7$;
2. we will consider $m$ machine that is active;
3. $M$ est is a subjective function;
4. $\forall m \in A, M(m) = m^{'}$ with $m^{'} \in B$;
5. $M(m) = M(c, d, p, k, l, r, b) = (M(c), M(d), M(p), M(k), M(l), M(r), M(b))$ $= (c^{'}, d^{'}, p^{'}, k^{'}, l^{'}, r^{'}, b^{'}) = m^{'}$;
6. $d \leq d^{'}$.

**Let's show that $M$ is a subjective function:**
Given $y \in B$, let's find $x = (c_1, d_1, p_1, k_1, l_1, r_1, b_1) \in B$ such that: $y = M(x)$, with $c_1, d_1, p_1, k_1, l_1, r_1, b_1 \in \mathbb{R}_+^*$.
$y \in B \Rightarrow y = (c, d, p, k, l, r, d)$, with $c, d, p, k, l, r, b \in \mathbb{R}_+^*$ $\exists c_0, d_0, p_0, k_0, l_0, r_0, b_0$ such that : $c_0 R c, d_0 R d, p_0 R p, k_0 R k, l_0 R l, r_0 R r$ and $b_0 R b$, where $R$ is the order relation in $\mathbb{R}_+^*$.
We can take $c_1 = c_0, p_1 = p_0, k_1 = k_0, l_1 = l_0, r_1 = r_0, b_1 = b_0$.
Furthermore, since $d \leq d_1$ we can take $d_1 = d$ because $d \leq d$.
So for $x = (c_0, d, p_0, k_0, l_0, r_0, b_0)$, so we get $M(x) = y$.

*Property 1. We talk about software migration when the following condition is met:* $c^{'} = c$ *and* $d^{'} = d$ *and* $p^{'} = p$ *and* $k^{'} = k$ *and* $r^{'} = r$ *and* $l^{'} \neq l$ *and* $b^{'} \neq 0$.

*Property 2. We talk about data migration when one of the following cases is true:*

- $c^{'} \neq c$ *and* $d^{'} \neq d$ *and* $p^{'} \neq p$ *and* $k^{'} \neq k$ *and* $r^{'} \neq r$ *and* $l^{'} = l$ *and* $b^{'} \neq 0$.
- $c^{'} \neq c$ *and* $d^{'} \neq d$ *and* $p^{'} \neq p$ *and* $k^{'} \neq k$ *and* $r^{'} = 0$ *and* $l^{'} = l$ *et* $b^{'} \neq 0$.
- $c^{'} \neq c$ *and* $d^{'} \neq d$ *and* $p^{'} \neq p$ *and* $k^{'} \neq k$ *and* $r^{'} \neq r$ *and* $l^{'} \neq l$ *and* $b^{'} \neq 0$.
- $c^{'} \neq c$ *and* $d^{'} \neq d$ *and* $p^{'} \neq p$ *and* $k^{'} \neq k$ *and* $r^{'} \neq 0$ *and* $l^{'} \neq l$ *and* $b^{'} \neq 0$.

*Property 3. We talk about migration to the cloud when owner 1 or owner 2 is checked.*

## 4 Virtual Machine Migration Concept

### 4.1 Notion of Virtual Machine

A virtual machine (VM) is a software implementation of a physical machine that runs programs like a real machine [16].

5. the possible values of $r$ are: $r = 0$ the machine is not connected to any network, $r = 1$ the network type is PAN, $r = 10$ the network type is LAN, $r = 11$ the network type is MAN, $r = 100$ the network type is WAN, $r = 101$ the network type is WPAN, $r = 110$ the network type is WLAN, $r = 111$ the network type is WMAN, $r = 1000$ the network type is WWAN.
6. $p, b \in \mathbb{R}_+$ with $p \geq 233\ MHz$;
7. the possible values of $b$ are [12]: 2,4 $GHz$ and 5 $GHz$;
8. $k \neq 0 \Rightarrow b = 0$;
9. $c < d$;
10. the norm of the vector $m$ is defined by $\eta$ as follows:

$$\eta(m) = \sqrt{c^2 + d^2 + p^2 + k^2 + l^2 + r^2 + b^2}.$$

*Remark 1.* If $b = 0$ then the machine is a personal computer, otherwise it is a router.

*Remark 2.* If $k = 0$ then the PC has no connected devices (basic devices are not included).

*Remark 3.* When the processor is multi-core, the frequency of the machine is the maximum of the frequencies of all the cores.

### 3.2 Computer Network

A computer network is a set of computer elements (computers, routers, printers, etc.) connected to each other [15]. A computer network makes it possible to share data, documents, applications and printers [15].

Depending on the context, the term network may refer to the architecture, the prototype or the IT infrastructure. With the evolution of the Internet, several networking technologies have emerged to overcome the difficulties encountered in deploying the computer network and to meet the growing needs of businesses. Network virtualization was thus born. A network environment supports network virtualization if it allows the coexistence of several virtual networks on the same physical infrastructure [4].

A virtual network is a set of virtual equipment (computers, routers, etc.) and virtual links interconnected with each other. Virtual equipment and virtual links are created thanks to a software layer called the Virtual Machine Monitor (VMM) or hypervisor [9,16]. It virtualizes all the resources of a physical machine, thus defining and supporting the running of several virtual machines.

IT migration must therefore take account of the network technology used and the characteristics of the network.

### 2.3 Advantages of Migration

IT migration has several advantages, including [9]:

- improving system performance: the aim is to increase the performance of an application or business by adding new functions to the system;
- energy savings: the migration is carried out in such a way as to support applications running on a minimum number of servers. This is done to maximize the use of resources and for energy management. This is done to maximize resource utilization and for energy management;
- ease of maintenance: migration simplifies maintenance tasks, helping to reduce downtime due to system maintenance. Migration can also be used when you want to replace one computer system with another, where the information relating to the latter is migrated to another environment and returned after the new system has been installed;
- fault tolerance: migration is carried out in the event of a system fault, to increase the availability of the services provided by the system.

## 3 Machine Migration Concept

### 3.1 Notion of Machine

A machine can defined as an electronic and programmable device capable of automatically and rationally processing information. A machine here is either a personal computer or a router. The elements constituting the operating context of a machine are characterized by:

- the state of its central memory;
- the state of its hard disk;
- the state of its processors;
- the state of its connected devices (basic devices are not counted);
- the state of its softwares (installed applications and system software);
- the state of its active network connections;
- the state of its bandwidth if the machine is a router.

We can formally define a machine as a state vector: $m = (c, d, p, k, l, r, b)$ where $c, d, p, k, l, r, b$ représentent capacity of main memory, hard disk capacity, processor frequency, number of connected devices, number of software applications, type of network and bandwidth capacity.

We make the following assumptions:

1. $m \neq 0$;
2. we take as the minimum values for a personal computer those verifying the minimum characteristics required for the Windows XP system [5];
3. $c, d, k, l, r \in \mathbb{N}$ with $l \neq 0$;
4. $c \geq 64\ MB$, $d \geq 2\ GB$, et $k, l,\ r$ are binary numbers;

- internally: this refers to any modification of system variables and requirements aimed at improving the existing IT system. Example: upgrading an operating system or an application;
- externally: this involves the replacement of a computer system or the abandonment of one infrastructure in favor of another. It may also involve moving the system from one physical location to another.

To facilitate these migrations, it may be useful to implement careful planning and an infrastructure automation strategy [6].

### 2.1 Migration Conditions

As a general rule, before starting any IT migration, you need to make sure that:

- the storage space of the final environment is greater than or equal to the storage space of the initial environment;
- the architecture of the target environment is fairly close to or at least has the same characteristics as the source environment;
- the applications are compatible with the operating system.

### 2.2 Types of Migration

Depending on the project, IT migration may involve one or more types of move. There are therefore 03 main types of IT migration [3,6,7]:

- data (or storage) migration: this involves moving data from one type of storage system to another. This process is often carried out as part of an upgrade aimed at increasing storage capacity, improving performance, reducing costs, reducing footprint or adding new capacity. During migration, data must be moved between two database engines. The challenge of any database migration is to implement it without affecting the data language or reading protocol. A database migration is successful when the tools implemented manage to modify the data without altering the structure of the database;
- Software migration: this can involve either the operating system or application software. Software migration involves moving software from one computer system to another. It can be time-consuming and involve a number of risks, including downtime, incompatibility between applications and the loss of customized settings;
- migration to the cloud: this involves moving IT systems from traditional onsite data centers to cloud environments, or from one cloud environment to another.

process that can take several forms [3]: a transfer of data from one storage space to another, a transformation from one data format to another, or a conversion to make raw data usable on a particular type of system.

Salah-Eddine [2] affirms in his thesis work that a few years ago, the migration of virtual machines was not done in real time, but only after a complete stop of the virtual machines. This could be explained by the lack of automatic tools for real- time migration of virtual machines. However, in response to the rapid growth in the number of virtual machines and virtual networks, automated tools for real-time migration of virtual machines have become essential. In addition, given the current economic context, virtual machine networking service providers have an interest in carrying out rigorous migrations in order to retain their customers by offering them continuous services that are better than their competitors, thus enabling them to gain more market share [2]. Virtual networks and virtual machines are becoming increasingly popular, and the number of users is growing all the time. Consequently, providers of these services have no choice but to increase their investments in order to respond effectively to the growing needs of their customers [2]. Despite the potential vision of virtual machine migration, several important research challenges have been addressed and remain to be tackled.

For some years, several scientific studies have focused on the informal study of the virtual machine migration problem. In this article, we present a formal definition of the virtual machine migration problem. Our contributions are summarized in three points:

- formally define the migration environment for a machine;
- formally define the migration environment for a virtual machine;
- show that virtual machine migration is a rucksack problem

The rest of this article is organized as follows. In Sect. 2, we present a general overview of IT migration. Section 3 describes the concept of machine migration. Section 4 describes the migration of a virtual machine. The problem of migrating a virtual machine is presented in Sect. 5. Our virtual machine migration method is presented in Sect. 6. Our method is discussed in Sect. 7, and Sect. 8 concludes the paper.

## 2 General Overview of IT Migration

IT migration can be defined as the passage of an IT system from an initial execution environment to a final environment.

A computer system is a set of hardware and software computing and telecommunications resources whose purpose is to collect, process, store, route and present data [8].

The execution environment can be defined as the set of elements required for a computer system to function properly. Examples of environments include: a room, a server, a machine, software, a network, the cloud.

IT migration projects typically have many company-specific variables and requirements [6]. IT migration can be done at two levels:

# Proposal for a Formal Definition of the Virtual Machine Migration Problem

Thomas Djotio Ndie[1(✉)], Joel Casimir Tagne[1], and Karl Jonas[2]

[1] University of Yaounde I, Yaoundé, Cameroon
tdjotio@gmail.com, joelcasimirt@gmail.com
[2] Bonn-Rhein-Sieg University of Applied Sciences, Rheinbach, Germany
karl.jonas@h-brs.de

**Abstract.** Thanks to the development of New Information and Communication Technologies (NICT), computer tools are increasingly in demand in almost all sectors of activity to meet to the needs of people, property, companies... Thus, IT migration is therefore a process that can affect all companies that host data relating to their customers, suppliers, partners or even general statistics.

In this article, we consider the virtual machine migration problem as a rucksack problem. As server hosting capacities have become increasingly 'elevated', the need for consolidation and load balancing has led to a strong interest in virtual machine migration.

We have used the following approach: we begin by presenting the computer system migration environment. Next, we propose a formal definition of the virtual machine migration environment. We then show, using the formal language, that virtual machine migration is an NP-hard problem of the rucksack type. We then propose the MigrationVM algorithm of complexity $O(7)$ as an optimal solution to the problem of migration virtual machines in networks. We have discussed our algorithm and we believe that virtual machine migration data can be reduced to two main elements: the amount of RAM used by the virtual machine and the amount of hard disk used by the virtual machine.

**Keywords:** Virtual machine · Network · Service · Migration · Migration time

## 1 Introduction

Thanks to the development of New Information and Communication Technologies (NICT), computer tools are increasingly in demand in almost all sectors of activity to meet the needs people, property, companies... Thus, IT migration is therefore a process that can affect all companies that store customer data, relating to their customers, suppliers, partners or even general statistics [3]. Migrating data may be necessary when opting for a larger storage space or when implementing a new database management strategy. IT migration is a data transfer


A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 3–17, 2025.
https://doi.org/10.1007/978-3-031-81573-7_1

# Systems and Cloud Computing

## Cybersecurity and Privacy

## Wireless Networks



## E-health

# Contents – Part I

# Contents – Part II

| | |
|---|---|
| Nikola Djuric | University of Novi Sad, Serbia |
| Olasupo Ajayi | University of the Western Cape, South Africa |
| Ousmane Sadio | Universite Cheikh Anta Diop, Senegal |

| | |
|---|---|
| Didier Bassole | Université Joseph Ki-Zerbo, Burkina Faso |
| Marcellin Atemkeng | Rhodes University, South Africa |
| Nikodemus Angula | Namibia University of Science and Technology, Namibia |
| Malo Sadouanouan | Nazi Boni University, Burkina Faso |
| Idy Diop | Cheikh Anta Diop University, École Supérieure Polytechnique, Senegal |
| Idris Ahmada Rai | State University of Zanzibar, Tanzania |
| Abdi Talib Abdalla | University of Dar es Salaam, Tanzania |
| Christelle Scharf | Pace University, USA |
| Yahya Hamad Sheikh | State University of Zanzibar, Tanzania |
| Bernd Westphal | German Aerospace Center, Germany |
| Antero Järvi | University of Turku, Finland |
| Rania Abdulhaleem Mokhtar | Taif University, Saudi Arabia |
| Abubakar Bakar Diwani | State University of Zanzibar, Tanzania |
| Abdellah Boulouz | Ibn Zohr University, Morocco |
| Alemnew Sheferaw Asrese | Aalto University, Finland |
| Ali Idarous Adnan | State University of Zanzibar, Tanzania |
| Amar Kumar Seeam | Middlesex University London, UK |
| Andre Muhirwa | University of Rwanda, Rwanda |
| Antoine Bagula | University of the Western Cape, South Africa |
| Avinash Mungur | University of Mauritius, Mauritius |
| Ben Daniel | University of Otago, New Zealand |
| Clarel Catherine | University of Technology, Mauritius |
| David Baume | University of London, UK |
| Desogi Eisa Abdelrahman | Sudan University of Science and Technology, Sudan |
| Enock Mbewe | University of Cape Town, South Africa |
| Eugene C. Ezin | University of Abomey Calavi, Republic of Benin |
| Ibrahima Gaye | Alioune Diop University of Bambey, Senegal |
| Joel Mtebe | University of Dar es Salaam, Tanzania |
| Khamis Abdul-latif Khamis | State University of Zanzibar, Tanzania |
| Lulu Mahai | University of Dar es Salaam, Tanzania |
| Maryam M. Khamis | State University of Zanzibar, Tanzania |
| Maryam Jaffar Ismail | State University of Zanzibar, Tanzania |
| Mbuyu Sumbwanyambe | University Of South Africa, South Africa |
| Mercy Mbise | University of Dar es Salaam, Tanzania |
| Michael Gallagher | University of Edinburgh, UK |
| Mikko Ruohonen | Tampere University, Finland |
| Mjumo Mzyece | University of the Witwatersrand, South Africa |
| Natasha Zlobinsky | Meraka Institute, CSIR, South Africa |
| Nawaz Mohamudally | University of Technology, Mauritius |

**Workshops Chair and Co-chairs**

| | |
|---|---|
| Malo Sadouanouan | Université Nazi Boni, Burkina Faso |
| Tounwendyam Frédéric Ouedraogo | Université Norbert Zongo, Burkina Faso |

**Publicity and Social Media Chair**

| | |
|---|---|
| Telesphore Tiendrebeogo | Université Nazi Boni, Burkina Faso |

**Publications Chair**

| | |
|---|---|
| Tiguiane Yélémou | Université Nazi Boni, Burkina Faso |

**Web Chair**

| | |
|---|---|
| Pasteur Poda | Université Nazi Boni, Burkina Faso |

## Technical Program Committee

| | |
|---|---|
| Rashid A. Saeed | Taif University, Saudi Arabia |
| Sere Abdoulaye | Université Nazi Boni, Burkina Faso |
| Sallam Osman Fageeri | University of Nizwa, Oman |
| Bharat S. Chaudhari | MIT World Peace University, India |
| Osama Rayis | Africa Technology City, Sudan |
| Idowu Diyaolu | Obafemi Awolowo University, Nigeria |
| Tounwendyam Frédéric Ouedraogo | Université Norbert Zongo, Burkina Faso |
| Kennedy Ronoh | Strathmore University, Kenya |
| Pragasen Mudali | University of Zululand, South Africa |
| Mahmoud Abdulwahab Alawi | Karume Institute of Science and Technology, Tanzania |
| Namatovu Hasifah Kasujja | Makerere University Kampala, Uganda |
| Emmanuel Eilu | Uganda Christian University, Uganda |
| Abdi T. Abdalla | University of Dar es Salaam, Tanzania |
| Kalum Priyanath Udagepola | Scientific Research Development Institute of Technology, Australia |
| Ayodeji O. Oluwatope | Obafemi Awolowo University, Ile-Ife, Nigeria |
| Arsène Sabas | Canadian Nuclear Safety Commission, Canada; Institut de Mathématiques et de Sciences Physiques, Benin |

# Organization

## Steering Committee

| | |
|---|---|
| Rashid A. Saeed | Taif University, Saudi Arabia |

## Organizing Committee

### General Chair

| | |
|---|---|
| Rashid A. Saeed | Taif University, Saudi Arabia |

### General Co-chairs

| | |
|---|---|
| Oumarou Sie | Université Aube Nouvelle, Burkina Faso |
| Théodore Marie Yves Tapsoba | Université Nazi Boni, Burkina Faso |

### TPC Chair and Co-chairs

| | |
|---|---|
| Sere Abdoulaye | Université Nazi Boni, Burkina Faso |
| Yahya Hamad Sheikh | State University of Zanzibar, Tanzania |
| Abubakar Bakar Diwani | State University of Zanzibar, Tanzania |
| Abdi Talib Abdalla | University of Dar es Salaam, Tanzania |
| Borlli Michel Jonas Somé | Université Nazi Boni, Burkina Faso |

### Sponsorship and Exhibit Chair

| | |
|---|---|
| Seydou Golo Barro | Université Nazi Boni, Burkina Faso |

### Local Chair

| | |
|---|---|
| Mesmin Dandjinou | Université Nazi Boni, Burkina Faso |

the foundation for a digitally empowered and sustainable future.We eagerly anticipate your continued engagement and participation in future editions of AFRICOMM, as we collectively strive to explore ideas, innovations, and partnerships for the advancement of e-Infrastructure and e-Services.

Rashid A. Saeed
Abdoulaye Sere

# Preface

It is with immense gratitude and excitement that we extend a warm invitation to all participants, researchers, professionals, and enthusiasts who joined us at the 15th International Conference on e-Infrastructure and e-Services for Developing Countries (AFRICOMM 2023). This distinguished conference, organized by the European Alliance for Innovation (EAI), took place from November 23–25, 2023, at the esteemed Sissiman Hotel in Burkina Faso.

The success of AFRICOMM 2023 in Burkina Faso was made possible through the collaborative efforts of numerous individuals, and we would like to extend our heartfelt thanks to all involved. A special thanks is extended to EAI for their instrumental role in the conference organization. Their dedication to advancing research and innovation has been crucial in bringing together diverse perspectives and expertise to address the unique challenges and opportunities in the development of e-Infrastructure and e-Services within the African context.

Success would also not have been possible without the dedication and hard work of the local committee, the technical program committee, authors, and reviewers. Their commitment to ensuring the highest standards in research and innovation greatly contributed to the rich and diverse program of the conference.

Special thanks to Sidi Mohamed Galiam Ouedraogo, the Director General of Electronic Communications (DGCE/MTDPCE/Burkina Faso), for his visionary talk to the young African researchers, are also extended to Alain Mille from Universitaires Sans Frontières, for his thought-provoking keynote address on "Intelligence Artificielle en Afrique : pour quel développement?" (Artificial Intelligence in Africa: for which development?). His insights into the role of artificial intelligence in the African context sparked meaningful discussions and added significant value to the conference.

We express our sincere appreciation to Pascal Urien, a distinguished professor at Télécom Paris, for his captivating keynote presentation on "Building Trust with secure elements and open technologies: crypto device use cases." His expertise shed light on critical aspects of cybersecurity, contributing to the broader conversation on digital trust and security. The conference, set against the backdrop of Burkina Faso's rich cultural heritage, provided a dynamic forum for researchers, academics, industry experts, and policymakers to share insights, discuss challenges, and propose innovative solutions for the development of e-Infrastructure and e-Services in developing countries.

Africa, with its rapidly evolving landscape in Information and Communication Technologies (ICT) and Telecommunications, is in great need of events such as AFRICOMM 2023. These conferences play a crucial role in bridging the gap between technological advancements and the unique challenges faced by the continent. They serve as platforms for collaboration, knowledge exchange, and the exploration of solutions tailored to the specific needs of African nations. As we reflect on the conference's success, we extend our gratitude to the organizing committee, sponsors, and all those who played a pivotal role in making this conference a reality. Your dedication and contributions have laid

*Editors*
Abdoulaye Sere
Nazi BONI University
Bobo-Dioulasso, Burkina Faso

Oumarou Sie
New Dawn University
Ouagadougou, Burkina Faso

Rashid A. Saeed
Taif University
Taif, Saudi Arabia

ISSN 1867-8211 ISSN 1867-822X (electronic)
Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering
ISBN 978-3-031-81572-0 ISBN 978-3-031-81573-7 (eBook)
https://doi.org/10.1007/978-3-031-81573-7



This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Abdoulaye Sere · Oumarou Sie · Rashid A. Saeed
Editors

# Towards new e-Infrastructure and e-Services for Developing Countries

15th International Conference, AFRICOMM 2023
Bobo-Dioulasso, Burkina Faso, November 23–25, 2023
Proceedings, Part II

Springer

The LNICST series publishes ICST's conferences, symposia and workshops.

LNICST reports state-of-the-art results in areas related to the scope of the Institute. The type of material published includes

- Proceedings (published in time for the respective event)
- Other edited monographs (such as project reports or invited volumes)

LNICST topics span the following areas:

- General Computer Science
- E-Economy
- E-Medicine
- Knowledge Management
- Multimedia
- Operations, Management and Policy
- Social Informatics
- Systems

# Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 588

Abdoulaye Sere
Oumarou Sie
Rashid A. Saeed (Eds.)


588

# Towards new e-Infrastructure and e-Services for Developing Countries

LNICST

15th International Conference, AFRICOMM 2023
Bobo-Dioulasso, Burkina Faso, November 23–25, 2023
Proceedings, Part II

## Part 2

EAI

Springer