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

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.

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.

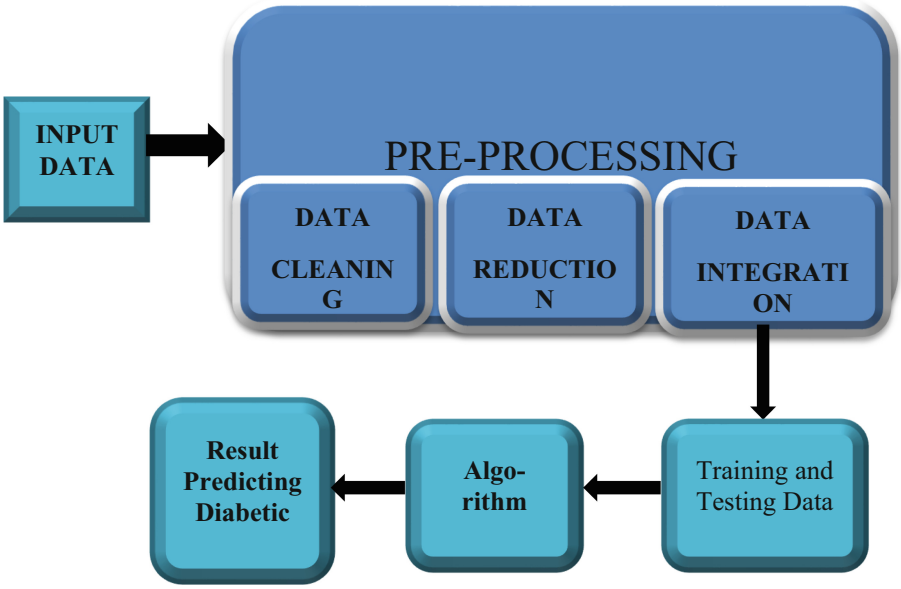

Fig. 1. System architecture

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,

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).

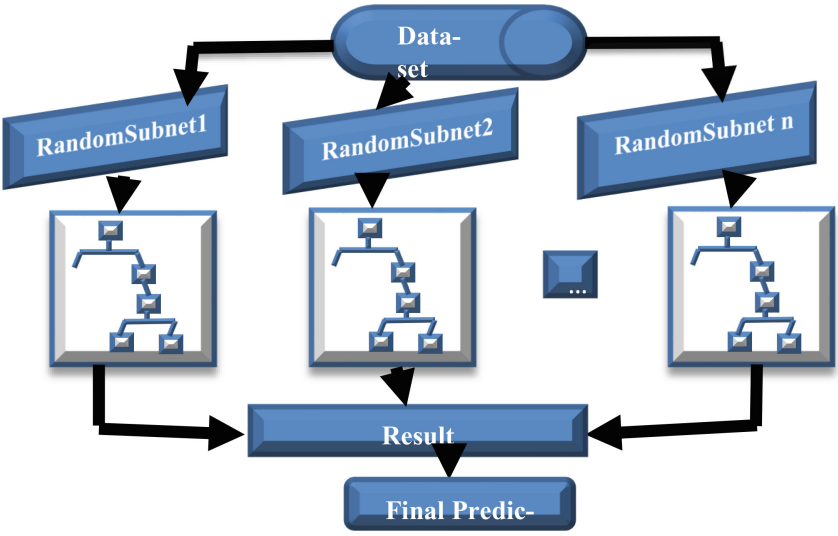

Fig. 2. Architecture of Random Forest

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. $X = X_1, X_2, X_3, \ldots.., X_n(where, X_2 - X_n = Distinctfeatures)$

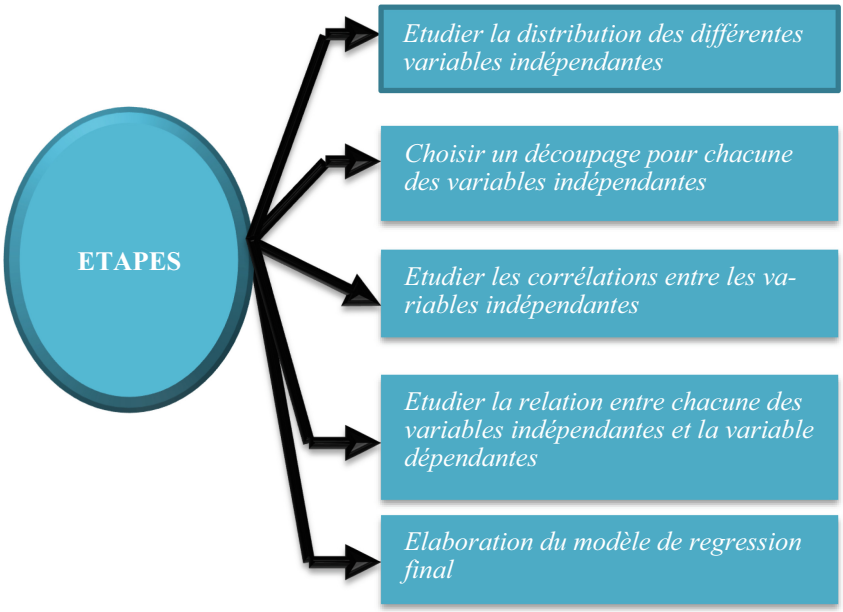

Fig. 3. Regression Logistic steps

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).

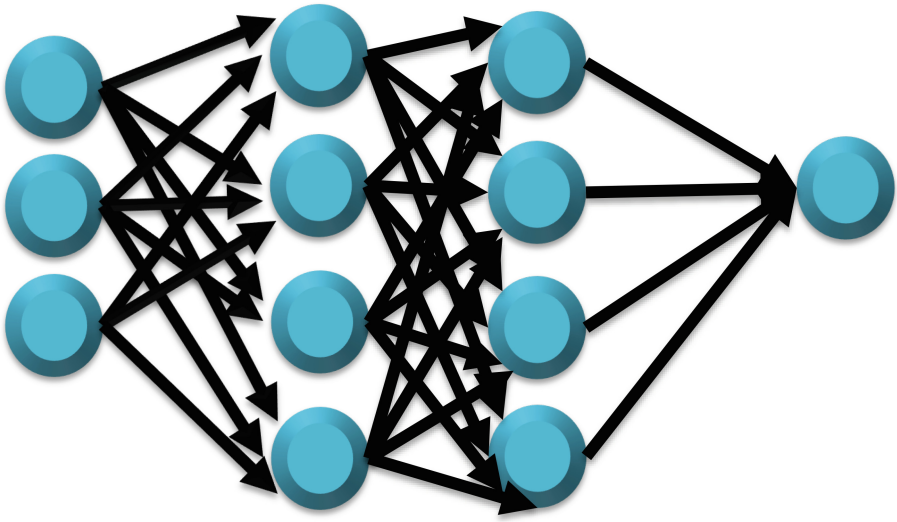

Fig. 4. Architecture of neural networks

*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.

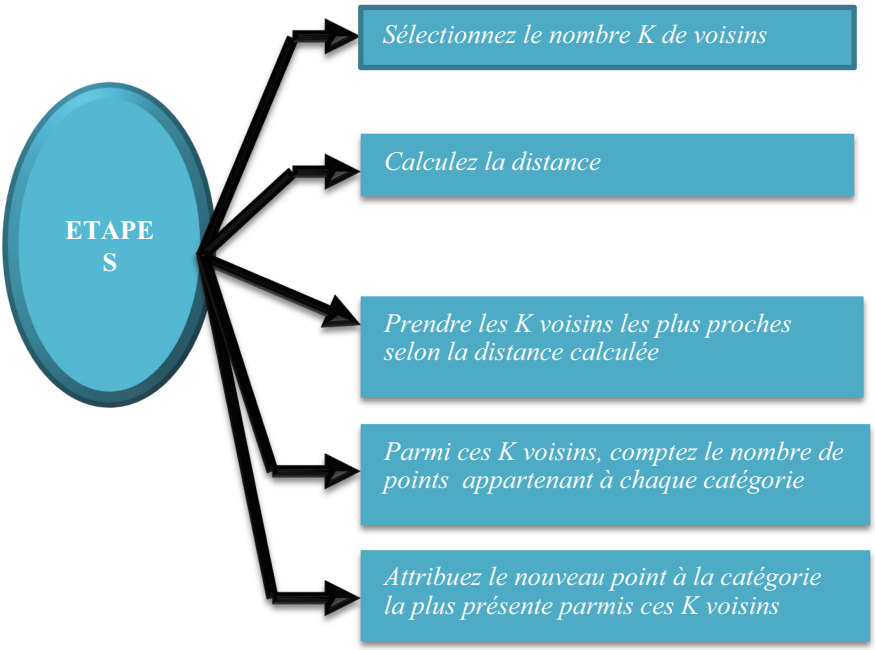

Fig. 5. K nearest neighbor's steps

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.

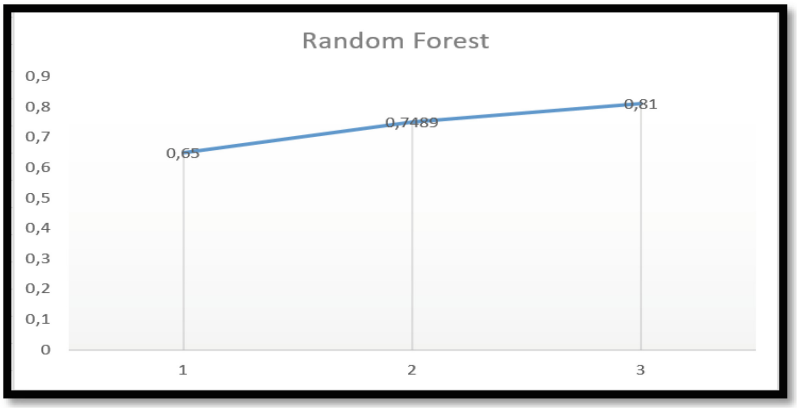

Fig. 6. Evolution of the tests with Random Forest

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.

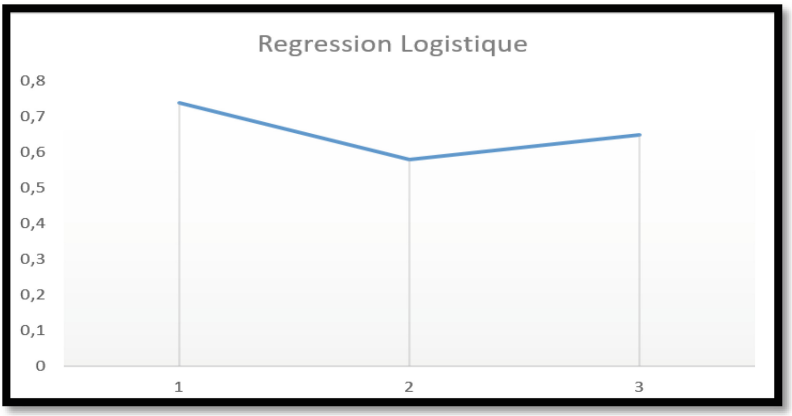

Fig. 7. Evolution of the tests with Logistic Regression

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.

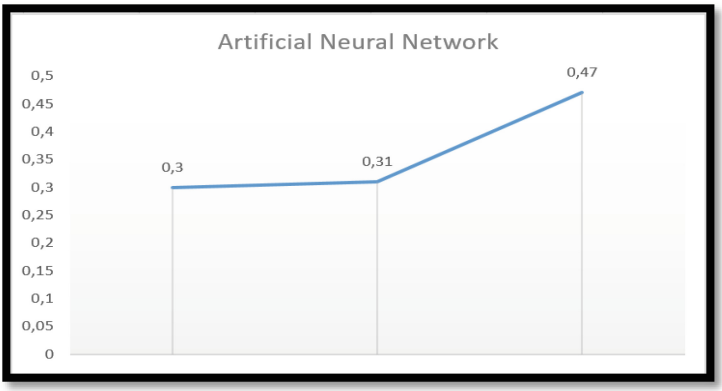

Fig. 8. Evolution of the tests with Artificial Neural Network

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.

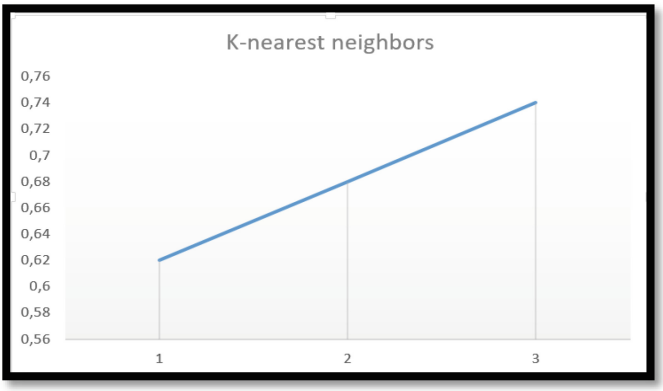

Fig. 9. Evolution of the tests with K-nearest neighbors

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.

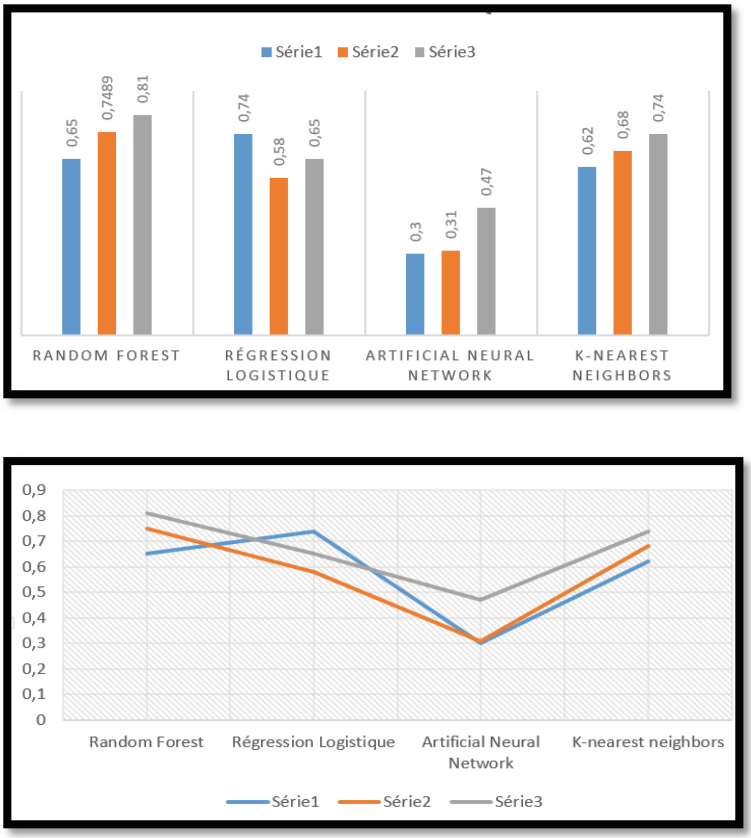

Fig. 10. Graphical representation of the results

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.

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