An AI-Based Model for the Prediction of a Newborn's Sickle Cell Disease Status

Souleymane Bosso Farota[1], Al Hassim Diallo[2], Mouhamadou Lamine Ba[1], Gaoussou Camara[1(✉)], and Ibrahima Diagne[2,3]

[1] LIMA, Université Alioune Diop, B.P. 30, Bambey, Senegal
gaoussou.camara@uadb.edu.sn
[2] Université Gaston Berger, B.P. 34, Saint-Louis, Senegal
[3] Centre de Recherche et de Prise en Charge Ambulatoire de la Drépanocytose, Université Gaston Berger, Dakar, Senegal

Abstract. Sickle cell disease remains a global public health problem. In Senegal, a neonatal screening and early follow-up program is conducted at the CERPAD. Such a program, started in April 2017, implements the strategy of systematic screening at birth and concerns children born in the maternity wards of the CHRSL as well as from the reference health center of the city of Saint-Louis. However, out of 18 257 newborns screened since the beginning of the program, only 49 (less than 0.5%) are pathological (SS, SC, etc.) which is extremely low compared to the cost in terms of human resources, working time and use of laboratory consumables. To mitigate the impacts of these limitations of the actual early detection and follow-up approach, we therefore propose in this paper a new approach to targeted screening based on artificial intelligence. We tested and compared the performances of five machine learning algorithms for the prediction of sickle cell status. The preliminary results are promising for the task of whether or not a given newborn has a potentially pathological profile, with the majority of the models showing a high prediction accuracy.

Keywords: AI · Machine learning · Predictive model · Neonatal screening · Targeted screening · Sickle cell · Senegal

1 Introduction

Sickle cell disease is a genetic disease with severe health implications on the daily life of the persons that suffer from it. It is estimated that each year over 300 000 babies with severe forms of these diseases are born worldwide, the majority in low and middle income countries.[1] Unfortunately, most of the children born with sickle cell disease in low-income or/and developing countries are still dying at an early age. For instance, without proper management of SS form, 50% of children die before the age of 5 [1]. As

[1] https://www.afro.who.int/health-topics/sickle-cell-disease.

A. D. Mambo et al. (Eds.): InterSol 2022, LNICST 449, pp. 96–104, 2022.
https://doi.org/10.1007/978-3-031-23116-2_7

a result, sickle cell disease remains nowadays a serious public health issue, particularly in Sub-Saharan Africa.

In Senegal, the Center of Research and Ambulatory Care for Sickle Cell Disease (CERPAD) is concerned about this state of affairs and has set itself the general objective of contributing to the fight against sickle cell disease in Senegal, mainly through fundamental and applied research programs. Its ambition is to collect and analyze epidemiological, clinical and socio-anthropological data in collaboration with other research teams at the University Gaston Berger by proposing an efficient model for neonatal screening and early management of sickle cell disease adapted to the public health system in Senegal. CERPAD has adopted the strategy of a systematic screening at birth since April 2017. Since then, out of 18,257 newborns screened, only 49 (less than 0.5%) are pathological (SS, SC, etc.). This is in contrast with the expensive cost of the screening program in terms of human resources, working time and use of laboratory consumables. Finding a sustainable funding of such a screening program in the medium and long term is very difficult, particularly in low-income countries. Therefore, in this paper, we propose a new targeted screening approach based on a machine learning model for sickle cell status prediction. The preliminary results, obtained by evaluating five standard machine learning algorithms on real data, are promising with the majority of the models showing a high prediction accuracy; accuracy value equals 100% and AUC close to 1.

The rest of the paper is organized as follows. Section 2 summarizes the state-of-the-art while Sect. 3 describes the dataset. The methodology will then be introduced in Sect. 4, followed by the presentation of our results and their discussion in Sect. 5. We conclude in Sect. 6.

2 Review of the Literature

There are several research projects which have been developed for healthcare settings based on machine learning approaches and especially for sickle cell disease. These works could be classified according to the different phases of the evolution of the disease (crisis, complications, etc.) or the management process (screening, biological and radiological analysis, treatment, hospitalization). However, we offer a few examples of recent work that highlight the use of artificial intelligence (or machine learning) in this field. For instance, [2] developed a collection of 14 models with genetic risk score composed of different numbers of SNPs and used the ensemble of these models to predict HbF in patients with sickle cell anemia. The models were trained in 841 patients with sickle cell anemia and were tested in 3 independent cohorts. [3] developed a model to predict the severity of a patient's case, to determine the clinical complications of the disease, and to suggest the correct dosage of the treatment(s). They also presented similar work that attempted to estimate the severity of SCD in diagnosed patients to aid medical professionals in prescribing drugs. [4] propose Machine-learning algorithms for predicting hospital readmissions in sickle cell disease. In [5], image processing and machine learning techniques are used to automate the process of detection of sickle cells in microscopic images then classify the RBC into three shapes: circular, elongated (sickle cell) and other shape. The machine learning classifier random forest, logistic regression naïve baye and support vector machine were used in this research. [6] implements a powerful and efficient Multi-Layer Perceptron (MLP) classification algorithm that distinguishes Sickle

Cell Anemia (SCA) into three classes: Normal (N), Sickle Cells(S) and Thalassemia (T) in red blood cells. An Automated screening of sickle cells technique using a smartphone-based microscope and deep learning is proposed in [7]. An automated diagnosis model of sickle cell anemia using SVM classifier is proposed in [8] based on images.

To the best of our knowledge our work is the first in Sub-Saharan African that investigates the efficiency of machine learning approaches for the prediction of human sickle cell status at birth.

3 Materials

Our study concerns newborns screened and followed at the CERPAD. These newborns come from the maternity wards of the CHRSL and the reference health center of the city of Saint-Louis. Newborns are monitored during visits. They also receive emergency care for acute attacks or other complications related to the disease. Data management is done through an electronic patient record management system of the National Medical Information System for Senegal (SIMENS) [9, 10]. The system currently registers over 18,257 sickle cell patients screened.

Table 1. Description of the attributes of the Sickle Cell dataset

Attribute/variable	Type	Description
Weight	Real	The weight of the child at birth
Height	Real	The size of the child at birth
PC	Real	Cranial perimeter of the child at birth
Full term birth	Boolean	Birth at term of pregnancy
Premature	Boolean	If the child is premature or not
Number of WA	Integer	The number of weeks amenorrhea
Multiple pregnancy	Integer	The number of fetuses in pregnancy
Fetal distress	Boolean	Birth complications
Transfused	Boolean	Whether the mother received a blood transfusion
Sickle cell disease	Boolean	Pathological sickle cell status

In this study, a sickle cell disease dataset containing 5,732 individuals were extracted from the 18,257 records to design our expected model. There were several attributes in this dataset. Only 10 attributes were considered according to the domain experts' recommendation after many exchanges on features selection (see Table 1).

4 Methods

4.1 Overview of Our Approach

Machine Learning plays an important role in disease prediction [11]. Machine Learning algorithms are mainly divided into four categories: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning [12]. In this paper, we use supervised learning techniques to predict patient status. Our problem is to classify patients into a given class (sick or healthy). It is a priori difficult to know which of the existing classification algorithms is the best for our dataset. It is therefore necessary to test different algorithms, then to compare their performance and to deduce the best one for our case of study. The following are the procedural steps of the designed methodology applied in this research. The dataset presented above and including the attributes mentioned in Table 1 are provided as input to the different machine learning algorithms. The input dataset is divided into 80% for the training dataset and the remaining 20% for the test dataset. In this paper, we focus on five classifiers: AdaBoost, Logistic Regression (LR), Support Vector Machine (SVM), k-nearest neighbors (KNN) and Random Forest (RF), described here [13].

The objective of this study is to effectively predict whether the patient suffers from sickle cell disease or not. In this step, we will first define the evaluation measures that we will use to evaluate our models. The most important evaluation metric for this problem area is sensitivity, specificity, accuracy, F1 measure.

4.2 Experimental Performance Evaluation

We tested and compared the performance of the aforementioned machine learning algorithms (see Sect. 4.1) using our Sickle Cell real dataset presented in Sect. 3. For the performance evaluation purposes, we relied on the precision, recall (or sensitivity), specificity, F1-measure (also known as F1-score), accuracy, and AUC (Area Under the Curve) metrics. We sum up in Table 2 the definition, as well as the formula, of each of these performance evaluation metrics. Recall that these metrics are computed based on the confusion matrices obtained from the results of the testing phase of each machine learning algorithm on the test set.

The precision metric estimates the ability of the model to predict the correct classes of the individuals in the positive class. When it is necessary to determine the number of positive predictions, given predictions in both classes, that can be accurately predicted, recall is another useful evaluation measure, representing the proportion of positives successfully categorized. The F1-score provides a good balance between the precision and the recall when evaluating the performance of a classifier, particularly in the presence of unbalanced data.

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

Table 2. Description of the performance measures

Characteristic	Formula	Description
True Positive (TP)	no formula	Number of positive cases correctly screened
True Negative (TN)	no formula	Number of negative cases correctly detected
False Positive (FP)	no formula	Number of negative cases incorrectly screens as positive
False Negative (FN)	no formula	Number of cases that could not be detected
Precision	$\frac{TP}{TP+FP}$	Proportion of positive identifications that were actually correct
Recall	$\frac{TP}{TP+FN}$	Proportion of actual positives that were identified correctly
Specificity	$\frac{TN}{TN+FP}$	It measures the proportion of negatives correctly identified as such
Accuracy	$\frac{TN+TP}{TN+FN+TP+FP}$	It is the ratio of correct prediction given the total number of cases
F1-score	$2 * \frac{Precision*Recall}{Precision+Recall}$	Its value is equal to the harmonic mean of the precision and the recall

4.3 Evaluation Setting up

For the purpose of our comparative evaluation study, we used Scikit-Learn, a popular Python library that provides an implementation of the most popular existing machine learning algorithms. All the experiments have been performed on Jupyter notebook 6.4.5 and Orange3.

4.4 Segmentation of the Dataset

For fitting and evaluating each algorithm, the Sickle Cell Dataset has been divided into two parts as follows.

- A training set representing 80% of the entire dataset
- A test set representing 20% of the entire dataset

To avoid being biased we randomly select the individuals to include in the training and the testing set. A k-fold cross validation with $k = 10$ has been also performed during the training step.

4.5 Cross Validation Phase

To be safe against overfitting or underfitting, we introduced a cross validation step during the setting up of each model. To this end, we used Orange which is an open-source data visualization, machine learning, and data mining toolkit. It features a visual programming

front-end for exploratory data analysis and interactive data visualization, and can also be used as a Python library.

Figure 1 shows the k-fold cross validation pipeline with Orange using no over or under sampling. Our reasoning is that if we model a problem with 10% positive classes, we should not train the model with a 80:20 class distribution, as this will not reflect real life. Using Orange, we can balance the class distribution and perform k-fold cross-validation during the fitting of the learning model.

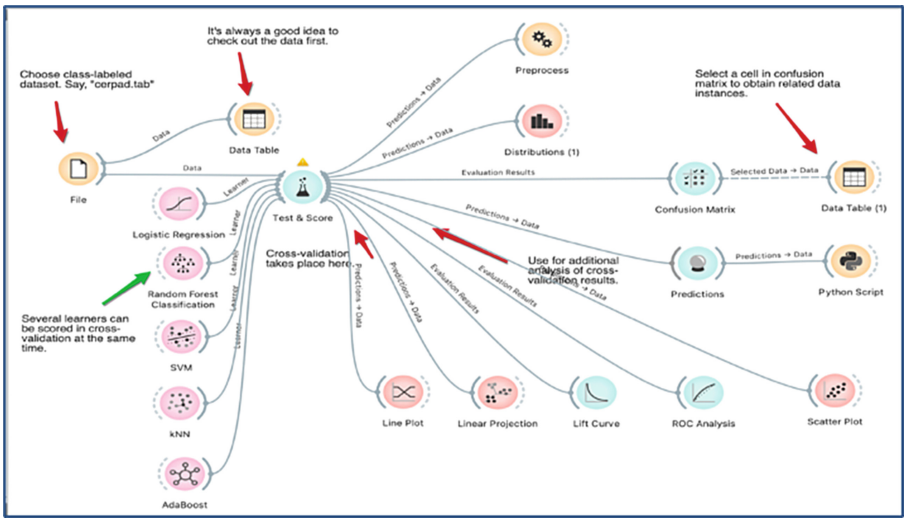

Fig. 1. K-fold cross-validation in Orange Tool

4.6 Selection of the Hyper-parameters of the Algorithms

A random stratification procedure has been used on the Sickle Cell data to produce 10 unique training datasets. The performance was good for each classification and the same training and test datasets were used to train and test the five classifiers under the same conditions and in the same environment, thus allowing the relative performance and consistency of the algorithms to be assessed for prediction. The conditions under which the classifiers were trained and tested were varied according to the evaluation of the criteria.

5 Results and Discussions

The performance measures of our five tested algorithms, depicted in Table 3 and Fig. 2, are obtained using the confusion matrix based on the results of the experiments conducted on the patient sickle cell dataset.

Results on Table 3 shows that all the models present very good performance regarding the different metrics. Indeed, excepting the precision of SVM and Random Forest which

is equal to 0.54 and 0.95 respectively, the values of the metrics for the algorithms are either to 0.99 or 1. In order to validate those results, we further conducted a k-fold cross validation of the models in order to mitigate the overfitting or underfitting problems that might bias the performances by providing no repeatable results.

Figure 2 gives the AUC values, *i.e.* the true positive rate versus the false positive rate, of the compared algorithms. While K-NN AUC value is close to 0.5, the other classifiers (SVM, LR, AdaBoost, RF) show very high AUC. This confirms the promising performances, depicted in Table 3, of SVM, LR, AdaBoost, and RF. The fact that the AUC of K-NN is very low could be explained by the unbalanced nature of the used Sickle Cell Data (the number of individuals that suffer from the pathological form of the disease is much lower than the number of individuals having either a non-pathological form of the disease or not the disease at all).

Table 3. Summary of the performance measures of the algorithms

	Precision	Recall	F1-score	Sensitivity	Accuracy
SVM	0.54	1.00	1.00	1.00	1.00
Logistic regression	0.95	0.99	0.99	0.99	0.95
K-NN	0.99	0.99	0.99	0.99	1.00
AdaBoost	0.99	1.00	1.00	1.00	1.00
Random Forest	0.95	0.99	1.00	1.00	1.00

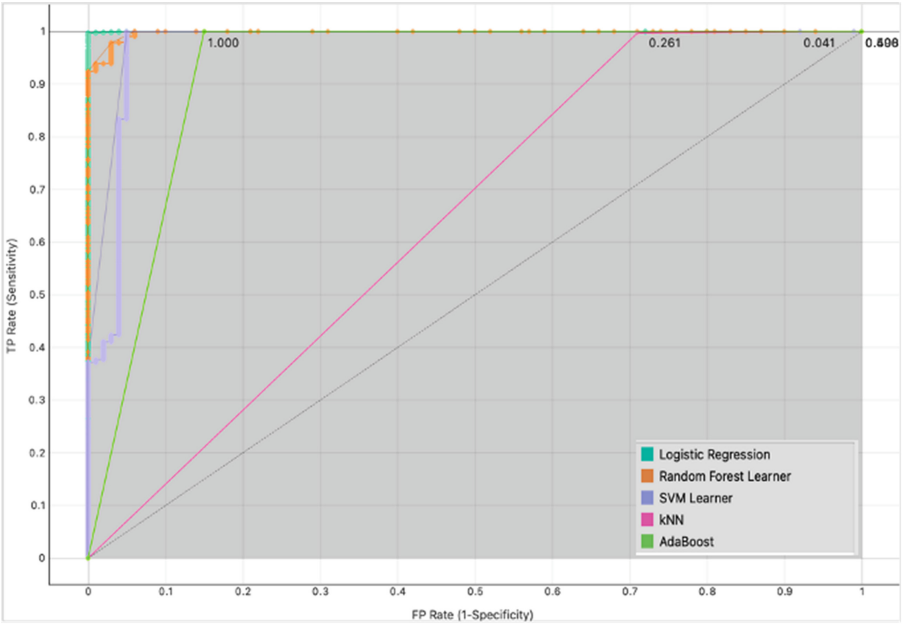

Fig. 2. AUC (Area Under the ROC Curve) of the compared algorithms

6 Conclusion

In this paper, we proposed an approach to neonatal screening for sickle cell disease targeting only children potentially carrying the sickle cell gene. For this purpose, we proposed a machine learning model built and trained on data collected in maternity wards on children at birth. Our model is based on the combination of five classification algorithms (AdaBoost, Logistic Regression, Support Vector Machine, k-nearest neighbors and Random Forest). The evaluation of our prediction model gives us an accuracy of 0.95 (LR) and 1 (the other classifiers). However, except K-NN all the tested classifiers have an AUC close to 1. We envision collecting and including data on parental sickle cell profiles in order to further boost the performances of our learning algorithm.

In the future, we plan to explore a probabilistic approach, e.g. by trying to infer the probability level of having potentially a pathological form of the disease, by considering our study as a regression problem. Furthermore, we also intend to extend our model to be able to predict directly the sickle cell profile (AA, AS, AC, SC, SS, etc.) of the newborns.

References

1. Thiam, L., et al.: Profils épidemiologiques, cliniques et hématologiques de la drépanocytose homozygote SS en phase inter critique chez l'enfant à Ziguinchor, Sénégal. Pan Afr. Med. J. **28**, 208 (2017). https://doi.org/10.11604/pamj.2017.28.208.14006
2. Milton, J.N., Gordeuk, V.R., Taylor, J.G., Gladwin, M.T., Steinberg, M.H., Sebastiani, P.: Prediction of fetal hemoglobin in sickle cell anemia using an ensemble of genetic risk prediction models. Circ. Cardiovasc. Genet. **7**, 110–115 (2014). https://doi.org/10.1161/CIRCGENETICS.113.000387
3. Alharbi, N.H., Bameer, R.O., Geddan, S.S., Alharbi, H.M.: Recent advances and machine learning techniques on sickle cell disease. Future Comput. Inform. J. **5**, 4(2020). https://doi.org/10.54623/fue.fcij.5.1.4
4. Patel, A., et al.: Machine-learning algorithms for predicting hospital re-admissions in sickle cell disease. Br. J. Haematol. **192**, 158–170 (2021). https://doi.org/10.1111/bjh.17107
5. Sen, B., Ganesh, A., Bhan, A., Dixit, S., Goyal, A.: Machine learning based Diagnosis and classification of Sickle Cell Anemia in Human RBC. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). pp. 753–758 (2021). https://doi.org/10.1109/ICICV50876.2021.9388610
6. Yeruva, S., Varalakshmi, M.S., Gowtham, B.P., Chandana, Y.H., Prasad, P.K.: Identification of Sickle Cell Anemia Using Deep Neural Networks. Emerg. Sci. J. **5**, 200–210 (2021). https://doi.org/10.28991/esj-2021-01270
7. de Haan, K., et al.: Automated screening of sickle cells using a smartphone-based microscope and deep learning. Npj Digit. Med. **3**, 1–9 (2020). https://doi.org/10.1038/s41746-020-0282-y
8. Wahed, F.F., Juliette, A.A., Sinthia, P., Mary, G.A.A.: Detection of sickle cell anemia using SVM classifier. In: AIP Conference Proceedings, vol. 2405, pp. 020006 (2022). https://doi.org/10.1063/5.0074138
9. Camara, G., Diallo, A.H., Lo, M., Tendeng, J.-N., Lo, S.: A national medical information system for Senegal: architecture and services. Stud. Health Technol. Inform. **228**, 43–47 (2016)
10. Diallo, A.H., et al.: Towards an information system for sickle cell neonatal screening in Senegal. Stud. Health Technol. Inform. **258**, 95–99 (2019)

11. Jayatilake, S.M.D.A.C., Ganegoda, G.U.: Involvement of machine learning tools in healthcare decision making. J. Healthc. Eng., 6679512 (2021). https://doi.org/10.1155/2021/6679512
12. Mohammed, M., Khan, M.B., Bashier, E.B.M.: Machine Learning: Algorithms and Applications. CRC Press, Boca Raton (2016). https://doi.org/10.1201/9781315371658
13. Sarker, I.H.: Machine learning: algorithms, real-world applications and research directions. SN Comput. Sci. **2**(3), 1–21 (2021). https://doi.org/10.1007/s42979-021-00592-x