A Machine Learning Approach: Using Predictive Analytics to Identify and Analyze High Risks Patients with Heart Disease

Fadoua Khennou, Charif Fahim, Habiba Chaoui, and Nour El Houda Chaoui

***Abstract*—The risk of developing early heart disease is always prominent. In fact, according to the Central of Disease Control and Prevention (CDC), Cardiovascular disease accounts for nearly 801,000 deaths in the US.**

In this paper, we present a machine learning and decision based system for heart disease prediction.

We validate our approach on the Heart Disease Dataset gathered from the UCI machine learning repository.

Cleveland, Hungarian and Switzerland datasets are combined and trained with the use of SVM and Naïve Bayes algorithms.

Comparison and analysis with other models show that the accuracy of the proposed approach is 87% and 86% for SVM and Naïve Bayes respectively, which is much higher comparing to the existing approaches that use a small subset of data and no imputation technique.

***Index Terms*—Heart disease, machine learning, prediction, SVM, Naïve Bayes.**

I. Introduction

Health institutes are looking for effective ways to care for more patients in a shorter time, and to increase the quality of care through better coordination. These organizations also strive to improve the patient experience with fast, accurate, and non-invasive diagnostics. To reduce costs, they try to avoid unnecessary procedures and readmissions.

The importance of health informatics, or E-health as it is commonly known, has increased dramatically through the big data revolution. The discipline of medical informatics is involved in the way of presenting new and efficient methods of collection, storage and analysis of medical data, thus improve diagnosis and treatment of many patients [1].

Medical informatics also facilitates the appropriate management, analysis, and use of health-related data for more efficient health care delivery and service for clients and patients.

The philosophy of health informatics is to transfer greater control of health care to medical professionals and patients.

The healthcare informatics technologies are moving rapidly towards the analysis of electronic health record systems [2]. This actually lead to the generation of voluminous health-related data that can be an important asset in improving the overall quality of health care and the well-being of people, if used wisely.

The goal of this current work is to propose an approach to improve heart disease prediction using machine learning and data mining techniques.

Cardio-vascular diseases relate to the heart and vessels. They represent today in our society the main cause of deaths before cancers and road accidents [3].

This refers to conditions that involve a disorder of the irrigation of the heart by the coronary arteries. They can take different forms: angina pectoris or angina, heart failure, infarction. The symptoms depend on the typical specificity of this disease such as coronary heart disease, stroke, heart failure, hypertensive heart disease, cardiomyopathy, cardiac arrhythmia and congenital heart disease.

These diseases are often linked to the presence of certain factors: hypercholesterolemia, diabetes and obesity. The most common factor is the presence of high blood pressure.

In fact, some people are more likely than others to develop a heart disease, thus knowing the risk factors and consulting a doctor for early treatment is a good strategy to avoid any future complications.

The determination of the main risk factors, which lead to such illnesses can be of a great value for different patients. This allow to have a general profile assessment that can help in providing special emphasis on high risk patients. Proper treatment and care can then be granted to these profiles.

The use of machine learning and particularly predictive analytics [4], which is represented as a set of techniques based on statistical studies of past and present events to establish hypotheses about the future, gives researchers the ability to explore datasets, through various algorithms, and find correlations between different events. It will not only help in saving lives by providing early diagnosis of heart problems, but also save money by avoiding costly and invasive treatments.

In this paper, we propose an SVM and Naive Bayes based model to predict heart disease. In order to do that, we used three sets of data to have more input values, then we applied KNN algorithms to fill missing values.

II. Related Works

According to the WHO (World Health Organization) [5], cardiovascular diseases (CVDs) are the number one cause of deaths in the world. For that, identifying patients with high

Manuscript received August 2, 2019; revised October 11, 2019.

Fadoua Khennou and Nour El Houda Chaoui are with the Transmission and Treatment of Information Laboratory, Higher School of Technology, Sidi Mohamed Ben Abdellah University, Fez Morocco (e-mail: fadoua.khennou@usmba.ac.ma, houda.chaoui@usmba.ac.ma).

Charif Fahim is with the System Engineering Laboratory, ADSI Team, National School of Applied Sciences, Ibn Tofail University, Kenitra Morocco (e-mail: charif.fahim@gmail.com, mejhed90@gmail.com).

doi: 10.18178/ijmlc.2019.9.6.870

level risk for heart diseases will help to save lives.

Predictive analytics provides tools and techniques such as machine learning algorithms to build powerful models that can help in predicting heart diseases at early stages using multiple risk factors.

Several research studies have been conducted to permit efficient prediction of patients likely to develop a heart disease.

In Ref. [6], the authors developed a Cardiopathy prediction system to help doctors evaluate a patients cardiopathy. To train their system they used a dataset with 13 input features (age, sex, pain sort, trestbps, cholesterol, abstinence glucose, resting ECG, easy lay pulse rate, exercise iatrogenic angina, old peak, slope, variety of vessels colored, and thal), then used KNN for classification.

While in the context of presenting a hybrid predictive model [7], the researchers proposed a framework called HCWV (Hybrid Classifier with Weighted Voting). This framework uses nine classifiers (SVM, Neural Network, Decision tree, generalized linear model, Lasso, Bayesian regularized Neural network, Classification and Regression Tree), and UCI database to predict heart disease. The accuracy of their model is 82.54%.

In Ref. [8], Jayshril S. Sonawane, D. R. Patil presented a predictive system for heart diseases, that uses multilayer perceptron. The network of this system is trained using backpropagation learning algorithm. The accuracy of their system was 97,5%. While their decision system has a very high accuracy, the need for analyzing their execution time performance through neural network, is very prominent.

In Ref. [9], the researchers depicted an approach based on a hybrid feature selection to predict heart disease. Their model uses Naïve Bayes and Random Forest for classification. The accuracy of their model is 84,1584% when using Naïve Bayes, and 84,1604% when using Random Forest.

The implementation of the Naïve Bayes and Decision Tree algorithms to predict heart diseases was studied in another research [10], the results of their model showed that Naïve Bayes is more powerful in term of accuracy than Decision Tree.

We summarize and describe in details all the previously analyzed approaches in Table I.

We conclude that approaches based on hybrid algorithms in the classification process using ANN, provide more accurate results than the ones using a single algorithm.

Indeed, as this problematic has been studied in the context of several research studies, the optimization of predictive models via the implementation of a single algorithm has reached its peak level.

We also notice that training models based on large databases, helps building models with higher accuracy than the one using small databases with a limited number of records.

The pre-analysis of the existing literature, as described in Fig. 1, helped us to adopt the best algorithms based on the complexity and the accuracy achieved in the previous papers. For that, we selected the three UCI datasets and we combined them into one database. Furthermore, we used KNN imputation technique to impute missing values and test over Naive Bayes and SVM.

TABLE I: COMPARISON OF RECENT PREDICTIVE STUDIES

Approach	Algorithm	Accuracy	Dataset
Prediction of Heart Diseases Using Data Mining Techniques.	Decision Tree Association rules K-NN ANN Naïve Bayes Hybrid Model	76% 55% 58% 85% 86% 96%	Unspecified
Feature Analysis of Coronary Artery Heart Disease Data Sets.	Decision tree Optimized decision tree	77,5% 78,56%	Cleveland and Hungarian datasets
Heart Disease Diagnoses using Artificial neural network.	ANN	88%	Cleveland dataset
Prediction of Heart Disease Severity with Hybrid Data Mining.	SVM + ANN + DT + generalized linear model + Lasso, regularized neural network Bayesian classification and regression tree	82,54%	Cleveland dataset
Prediction of Heart Disease Using Multilayer Perceptron Neural Network.	ANN + Random Forest	98%	Cleveland dataset
Prediction of Heart Disease Using Hybrid Technique For Selecting Features.	Naive Bayes Random Forest	84,1584 84,1604 %	Cleveland dataset
Smart heart disease prediction system using Improved K-Means and ID3 on Big Data.	K-Means	98,28%	(UCI) Unspecified dataset
Improving the Performance of Entropy Ensembles of Neural Networks (EENNS) on Classification of Heart Disease Prediction.	ANN	85,66%	Cleveland dataset

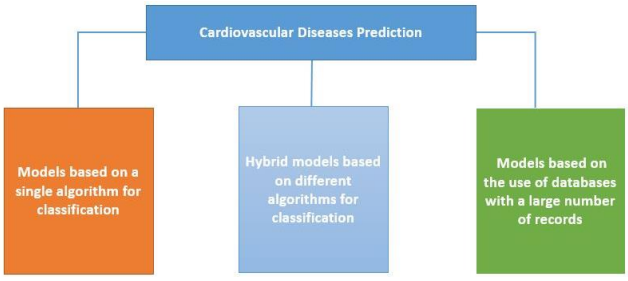

Fig. 1. Related works based on cardiovascular diseases prediction.

In this research study, we only focus on adopting non-complex algorithms for the sake of the execution time performances, for this reason we will not shed light on ANN.

III. MATERIALS AND METHODS

A. Data Source

Through this research study, we used the heart diseases datasets available in UCI Machine Learning Repository [11], namely Cleveland, Hungarian and Switzerland sources.

These datasets are the most commonly used databases by

machine learning researchers, they contain 76 attributes, but only 14 of them are referred by all published studies.

The output field, which has varying values from 0 (absence) to 4, denotes the presence or the absence of the disease. Studies on the three databases have focused on distinguishing absence (value 0) from presence (values range from 1 to 4). In order to achieve better accuracy, we have combined the three data sets into one database to train our model.

13 of the features listed in Table II were used as inputs data for our model. The attribute "Num" is the output (predicted) data for the model, its value varies between 0 and 4; 0 means the absence of the heart disease, and when it is greater than 0 this means the presence of the disease.

To simplify our prediction, we only used 0 for the absence of the disease and 1 when the heart disease is present.

TABLE II: FEATURES DESCRIPTION

Clinical Features	Description
Age	Age
Ca	Number of major vessels (0-3) colored by flourosopy
Chol (mg/dl)	Serum cholesterol
Cp	Chest pain type
Exang	Exercise induced angina
Fbs	Fasting blood sugar
Oldpeak	ST depression induced by exercise relative to rest
Restecg	Resting electrocardiographic results
Sex	Gender
Slope	The slope of the peak exercise ST segmen
Thal	3=normal ; 6 = fixed defect; 7=reversible defect
Thalach	Resting electrocardiographic results
Trestbps(mmHg)	Maximum heart rate achieved
Num	Diagnosis of heart disease

The Data sets that we used contain some attributes with missing values, to deal with that we used imputation technique that we will present in the Results section.

B. Algorithms

There are many algorithms available for machine learning and they are classified into two broad categories, depending on the nature of learning.

The process of selecting an automatic learning algorithm involves matching the characteristics of the data to be learned with the biases of the existing approaches.

The machine learning algorithms can be divided into two main groups: supervised learning algorithms that are used to build predictive models and unsupervised learning algorithms that are used to construct descriptive models.

These techniques are profoundly different from traditional IT approaches. Indeed, in machine learning processes, the result depends on a learning itself dependent on data from which the particular behavior is extracted to generalize it to a class of problems.

- **Naïve Bayes**

The naive Bayesian algorithm uses a simplified version of the Bayes formula to decide to which class belongs to a new instance. The Naïve Bayes classifier focuses on the rule, which explains that the presence or absence of a disease depends on a feature itself. It assumes that features are independent of each other [7].

The Bayesian rule, presented in (1), is used to calculate the class probability for a given dataset:

$$P(C_k / X) = P(C_k) * \frac{P(C_k / X)}{P(X)} \quad (1)$$

where X is an instance that needs to be classified and C_k is respective class. $P(C_k / X)$ represents the probability of vector X belongs to class C_k.

Since we have a data set with 13 independent feature and one dependent feature which is the predicted value (Num), Naive Bayes can be used for our classification, because this algorithm doesn't require that variables should be dependent. In addition to that, it is high speed even when applied to large databases in comparison to other algorithms' performance.

- **SVM**

The basic concept of SVM is based on statistical learning theory. The SVM algorithm attempts to find a separating hyperplane in the feature space, and data from different classes reside on different sides of that hyperplane. Then we calculate the margin, which is the minimum of distance between points in the hyperplane.

Support vectors are data points closer to the hyperplane that influence the position and orientation of the hyperplane. By using these support vectors, we maximize the margin of the classifier. Supporting vectors are the points that help build the SVM.

To separate the two classes of data points, many hyperplanes can be chosen.

Our goal is to find a plan with the maximum margin, that is, the maximum distance between the data points of the two classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified more precisely.

Kernel functions are used for nonlinear mapping of training samples in high dimensional space.

Different kernel functions, such as polynomial, Gaussian and sigmoid are used for mapping and maximizing the separation between data points.

In the literature, it is mentioned that the performance evaluation of Support Vector Machine has shown a high percentage in terms of accuracy and precision in various medical problem.

In our work, we used the SVM algorithm for classification to generate two classes. Class of patients who do not have cardiovascular disease, and the other class of patients with cardiovascular disease.

- **K-nearest neighbor**

K nearest neighbor (KNN) is a similarity-based algorithm which means that it classifies by measuring how two or more objects are similar or related.

KNN algorithm also called as lazy learning, because it doesn't need to build a model to predict or to classify data, it's a similarity based algorithm, which means that it estimates how likely a data point is to be a member of one group or another depending on what group the data points nearest to it are in.

KNN can be used in statistical estimation and pattern recognition and in classification problems. In our case we used KNN algorithm to impute missing data values in our dataset.

The missing value can be approximated by the values of the point that are closest to it, based on other features by finding the k nearest neighbors and then, the most common value among all neighbors is taken then mean value is calculated.

C. Used Platform: Python

Python is an interpreted programming language that can be used in multiple programming forms, it supports object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library [12].

In our model, we used Python because it provides free and open libraries for data mining and data analysis. Especially a library called SciKit-Learn providing a wide sort of algorithms for classification, regression and clustering.

D. Machine Learning: Predictive Analytics

Predictive analysis provides estimations of the probability of a future outcome. It is important to remember that no statistical algorithm can predict the future with 100% accuracy.

Indeed, predictive analytics reveals patterns and relationships between data, and makes possible predictions of what will happen in the future; with a high probability; based on historical databases, as well as trends that are emerging. There are methods that can trace trends and future developments, thanks to historical data.

Statistical, mathematical and linguistic methods create real added value and profit for customers thanks to the large amounts of Big Data. Predictive analytics is a key challenge for digital businesses, which have integrated customer experience management as a success factor.

Already today predictive analytics is established as an integral part of marketing, customer analytics, budgeting or customer relationship management and the healthcare sector is no exception.

Predictive analytics in this area is likely to provide effective results by improving the quality of service. It has indeed, the future to transform the health care sector.

Whereas we find predictive analytics very useful in a lot of real world use cases, several challenges may hinder the proper benefit that we can gain from these methods and their appropriate learning. We can list these research challenges as follows:

- No sufficient data to test and train the model.
- Having to deal with a lot of missing values.
- Data streaming from distributed and varied sources.
- Feature selection and construction process.
- Standardization and integration of the model into a real word application.
- Data security challenges in regards to the collection phase.

As you can notice, although we have the necessary tools and methods to make the right predictions, it is important to consider these different issues that can actually impact the results of the proposed models.

On the one hand, the main challenge is related to the volume of data through which the prediction is implemented.

The more data we have, the more we will be able to gain profit from its hidden values. Along with that, this can impact directly the deployment of the model into a real world application.

On the other hand, as we are dealing with the healthcare data, we have to process sensitive patients' records, which can be very hard to acquire and collect. And if we have the ability to do so, we will have to deal with security compliance rules, which can be very challenging based on the unstructured and distributed nature of these data.

All in all, we can say that this is a double-edged weapon that we need to handle strictly to respond to the actual needs.

IV. Proposed Model

The main focus of the proposed model is to use two algorithms in each step of the training process. The following steps in Fig. 2 and Fig. 3 describe the main workflow of the prediction learning process.

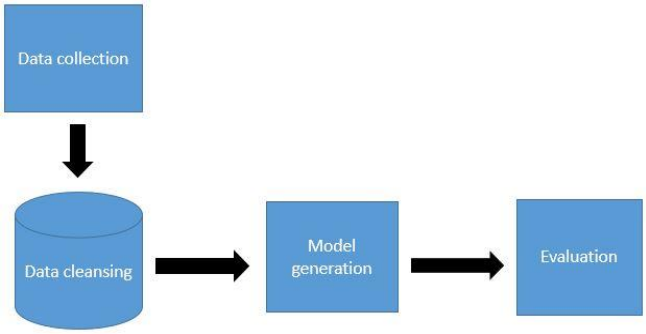

Fig. 2. Predictive learning process.

Data Collection: The process of collecting and transforming data into a suitable format for analysis.

Data cleansing: After collecting data, it should be explored and revised in order to correct errors and handle missing values, because machine learning is based on the quality of the data.

Model generation: This step involves choosing the appropriate algorithm to structure the models.

Model evaluation: This consists on analyzing how effective the algorithm is and how much it has been derived from historical data by using a set of test data. Afterwards, a comparison study is intended so as to measure the accuracy of the proposed approach in regards to other related works.

- STEP 1: The first step in the proposed model is the collection of data.

We downloaded the Cleveland, Hungarian and Switzerland heart diseases datasets from the UCI Machine learning repository. Then we combined all the datasets into one CSV file.

- STEP 2: We proceed by data cleaning with the use of the imputation of missing values attributes, using KNN algorithm.
- STEP 3: We divided data into training set and test set (70% for training and 30% for testing)
- STEP 4: We used SVM algorithm to implement the classification.
- STEP 5: We calculated the accuracy of the model using SVM.
- STEP 6: We implement the classification using Naïve Bayes.
- STEP 7: We calculate the accuracy of the model using

Naïve Bayes.

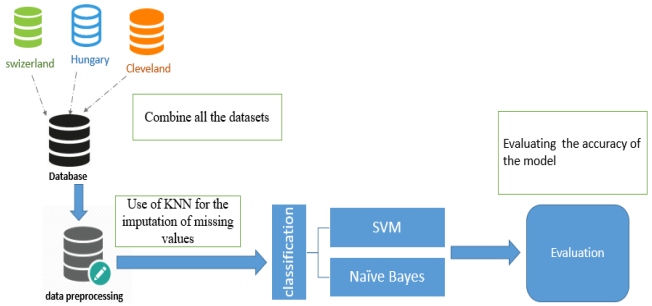

Fig. 3. Clinical decision predictive model.

In order to have an idea of our features selection and how each of the 13 features are distributed to discriminate between the two predictive cases (when num=1 and num=0), we generated the plots presented in Fig. 4.

Fig. 3 shows histograms and bar graphs of the 14 attributes in the heart diseases data set, for patients with heart disease (red) and without heart disease (Green).

According to these graphs we can conclude that people with heart diseases are mostly older and are male having higher blood pressure, higher cholesterol levels and lower heart rate than the Thallium stress test.

V. Results and Experiments

We used two classifiers (SVM, Naïve Bayes) and heart diseases database to train the model.

We evaluated the model by measuring its accuracy, which is described in (2), when using SVM and Naïve Bayes.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (2)$$

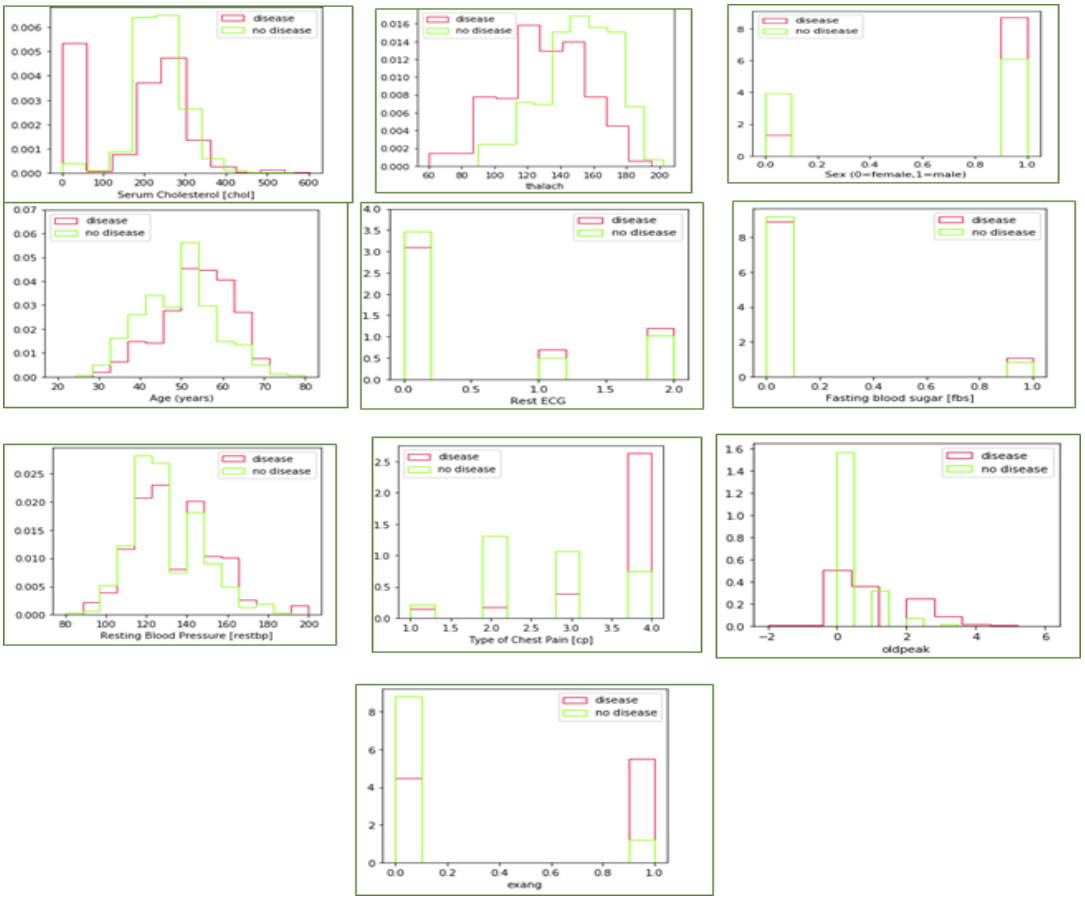

Fig. 4. Features distribution.

Table III shows the result when using SVM and Naïve Bayes.

TABLE III: Results

Algorithm	Accuracy
SVM	87%
Naïve Bayes	86%

We compared the results with other approaches that used the same database and the same algorithms for classification (SVM and Naïve Bayes).

In Ref. [13], the authors refer that accuracy of SVM is 75.4% and for Naïve Bayes is 85.4% when using just one of the UCI databases and without using missing value imputation.

TABLE IV: Comparing to Recent Studies

Algorithm	Accuracy
SVM [13]	75.4%
Naïve Bayes [7]	85.4%
HCWV [7]	82.54%
Using SVM in our model	**87%**
Using Naïve Bayes in our Model	**86%**

In Ref. [7], the authors build a model named HCWV (Hybrid Classifier with Weighted Voting) that uses 9 classifiers; SVM, Neural Network, Decision tree, generalized linear model, Lasso, Bayesian regularized Neural network, Classification and Regression Tree for training and they used

one of the UCI database. The accuracy for this model (HCWV) was 82.54%.

Table III and Fig. 5 show the comparison of the achieved accuracy comparing to other recent models that used the same database.

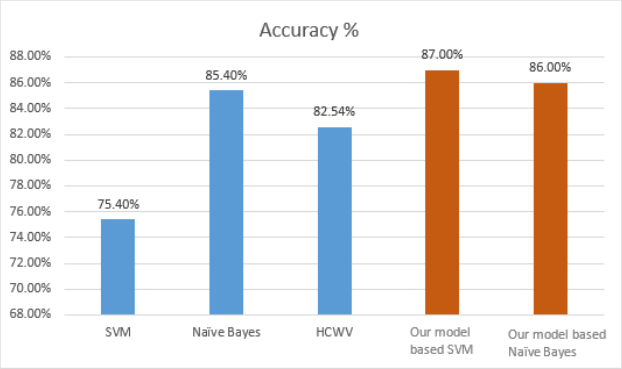

Fig. 5. Accuracy evaluation diagram.

According to the results in table IV, we can see that the proposed model achieved higher and better accuracy than other classification models that uses SVM or Naïve Bayes, with just one of the three UCI databases and without imputing missing values.

Imputing missing values and using larger data than other models helped to obtain quite promising results in classifying the possible heart disease patient with an accuracy of 87% for SVM and 86% for Naïve Bayes.

VI. Conclusion

The purpose of our proposed technique is to achieve more accurate prediction of heart diseases using SVM and Naïve Bayes.

The presented approach consists of combining multiple datasets to have more input data to train the model, then using imputation techniques with KNN algorithm to fill missing values.

As a perspective, we intend to implement an optimized approach to increase the accuracy while training the model on a big dataset. This can help us further to implement a mobile decision system for heart disease prediction with accurate results and precise risk factors.

References

[1] F. Khennou, Y. I. Khamlichi, and N. E. H. Chaoui, "Designing a health data management system based hadoop-agent," in *Proc. 2016 4th IEEE International Colloquium on Information Science and Technology (CIST)*, 2016, pp. 71-76.

[2] F. Khennou, Y. I. Khamlichi, and N. E. H. Chaoui, "Improving the use of big data analytics within electronic health records: A case study based OpenEHR," *Procedia Computer Science*, vol. 127, pp. 60-68, 2018.

[3] M. Kirmani, "Cardiovascular disease prediction using data mining techniques," *Oriental Journal of Computer Science and Technology*, vol. 10, pp. 520-528, 2017.

[4] S. J. Al'Aref, K. Anchouche, G. Singh *et al.*, "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging," *European Heart Journal*, 2018.

[5] WHO | Cardiovascular diseases (CVDs). [Online]. Available: http://www.who.int/cardiovascular_diseases/en/

[6] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, "Heart disease prediction using data mining techniques," in *Proc. 2017 International Conference on Intelligent Computing and Control (I2C2)*, 2017, pp. 1-8.

[7] M. Saini, N. Baliyan, and V. Bassi, "Prediction of heart disease severity with hybrid data mining," in *Proc. 2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, 2017, pp. 1-6.

[8] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network," in *Proc. 2014 International Conference on Information Communication and Embedded Systems (ICICES)*, 2014, pp. 1-6.

[9] F. Babič, J. Olejár, Z. Vantová, and J. Paralič, "Predictive and descriptive analysis for heart disease diagnosis," in *Proc. 2017 Federated Conference on Computer Science and Information Systems*, 2017, pp. 155-163.

[10] S. Maheswari and R. Pitchai, "Heart Disease prediction system using decision tree and naive bayes algorithm," *Current Medical Imaging Reviews*, vol. 14, 2018.

[11] *UCI Machine Learning Repository: Heart Disease Data Set*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease

[12] *Welcome to Python.org*. [Online]. Available: https://www.python.org/

[13] H. A. Esfahani and M. Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier," in *Proc. 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, 2017, pp. 1011-1014.

Fadoua Khennou received her MSc degree in information systems security from the National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco. She is a part time teacher for the MSc degree "Architectures and App. Sys. of Information" in the National School of Applied Sciences, Fez Morocco, and a PhD candidate in the "Transmission and Treatment of Information" Laboratory, Higher School of Technology, Sidi Mohamed Ben Abdellah University, Fez, Morocco.

Her current research interests include big data, e-health systems, SQL, hadoop, database management, OpenEHR, NoSQL, health interoperability, distributed computing and machine learning.

Charif Fahim is a software engineer. He has a bachelor degree in software engineering from Ibn Zohr University, Agadir Morocco. He has a master degree in information systems security from the National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco.

As part of his master thesis project, he worked on machine learning and specifically predictive analytic algorithms with python programming.

Habiba Chaoui is a full professor in computer science at the National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco.

She is responsible for the research MSc degree "Information Systems Security" specialty "Security of Systems and Computer Networks".

She is the head of the research team "Data analysis and Information Security" in the National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco.

Her research interests include information systems security, intrusion detection and prevention systems, database management systems, big data and cloud computing security.

Nour El houda Chaoui is a full professor in computer science at the National School of Applied Sciences, Sidi Mohammed Ben Abdallah University, Fez, Morocco.

She is also a project manager at Sidi Mohamed Ben Abdellah University and took part of different programs as the "Emergency Program" that was conducted from March 2009 to December 2012 at Sidi Mohamed Ben Abdellah University, Fez, Morocco.

Her research interests include advanced information systems, database systems, big data, machine learning, medical information systems and telemedicine.