Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 14 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i14.51479

PAPER

Artificial Neural Networks with K-Fold Cross-Validation and Feature Selection for Early Heart Disease Prediction

Inssaf El Guabassi[1](✉), Zakaria Bousalem[2], Rim Marah[3], Abdellatif Haj[4]

[1]LAROSERI Laboratory, Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco

[2]Polydisciplinary Faculty, Sultan Moulay Slimane University, Beni Mellal, Morocco

[3]Faculty of Economics and Management, Sultan Moulay Slimane University, Beni Mellal, Morocco

[4]Faculty of Sciences and Technologies, Hassan 1st University, Settat, Morocco

elguabassi@gmail.com

ABSTRACT

The most common reason behind death all over the world is heart diseases. These conditions are to hit hardest in low- and middle-income nations, where 80% of premature heart attacks could be prevented. In this regard, early diagnosis also plays an important role in increasing patient health and survival rate from heart disease. The purpose of this study was to improve the forecasting power by means of feature selection techniques and then apply K-Fold cross validation in combination with high-performance ensemble machine learning (ML) methods (J48, Artificial Neural Networks (ANNs), Logistic Regression, Naive Bayes, K-Nearest Neighbors) by utilizing a dataset of 401,958 patients. Our experimental results demonstrate that ANNs achieve the highest accuracy at 91.48%. They also record the lowest Mean Absolute Error (MAE) of 0.13, highlighting their precision in predictions. Additionally, ANNs exhibit a low root Mean Squared Error (RMSE) of 0.26, further indicating their reliability in modeling.

KEYWORDS

diagnostic analytics, feature selection, healthcare system, heart diseases, K-fold cross-validation, machine learning

1 INTRODUCTION

Heart diseases are a group of common and serious conditions that affect the heart's structure and function [1]. They are serious conditions, most often which occur together and can greatly impact on someone's quality of life or even put their lives in danger. These diseases represent a major public health issue worldwide today, causing morbidity, disability, and death [2]. For instance, by 2030 the annual deaths from heart diseases are estimated to rise up to 22.2 million, if present trends continue [3]. In Morocco, heart diseases account for two out of five deaths (38%), making them the leading cause of death nationwide (World Health Organization, 2018). Some heart disease disorders can be inherited from parents, while others stem from lifestyle choices. Factors contributing to these conditions include diabetes,

El Guabassi, I., Bousalem, Z., Marah, R., Haj, A. (2024). Artificial Neural Networks with K-Fold Cross-Validation and Feature Selection for Early Heart Disease Prediction. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(14), pp. 102–115. https://doi.org/10.3991/ijoe.v20i14.51479

Article submitted 2024-07-30. Revision uploaded 2024-09-15. Final acceptance 2024-09-17.

smoking, high cholesterol levels, lack of physical activity, hypertension, and obesity [4]. Additionally, some environmental factors of physical or chemical origin are likely to be involved in the onset of cardiovascular diseases, including noise, carbon monoxide, and air pollution [5].

Heart diseases can manifest suddenly, such as in the case of a heart attack, or they can develop slowly over time, often without noticeable symptoms until they become severe. It is not possible to predict when a person will have a heart attack, so by concentrating on preventive measures one can try to lessen the risk factors through a healthy lifestyle that includes managing stress, non-smoking, balanced diet, and regular exercise [6]. Additionally, early screening and predictive methods can aid in identifying potential issues before they escalate, allowing healthcare providers to intervene effectively and in a timely manner. This proactive approach not only improves outcomes but also enhances overall cardiovascular health and well-being. In this context, machine learning can significantly aid in achieving these goals for heart diseases. ML could be used to analyze big data sets combining medical records, genetic information, lifestyle factors, and environmental exposures in identifying the patterns of developing heart diseases. With predictive modeling, ML may assist health professionals in early detection and risk assessment at a personal level, so that timely intervention is given with targeted preventive strategies. Besides that, with time and increase in time, ML algorithms could continuously learn themselves and improve their accuracy, probably revolutionizing the approach to prevent and manage heart diseases.

The challenge of utilizing ML for early prediction of heart diseases lies in developing accurate models capable of integrating and analyzing diverse data sources effectively. Key questions in this context include:

- How do we identify the relevant predictive factors?
- Which algorithm is optimal for the task?
- What criteria should guide the selection of the best algorithm?

Addressing these questions involves the complex tasks of identifying and prioritizing relevant predictors from intricate datasets, assessing various ML algorithms for their accuracy and suitability, and defining performance metrics including accuracy, sensitivity, and specificity in the light of computational efficiency, which will provide effective prediction of the targeted heart diseases at an early stage.

Thus, this study attempted to improve the early prediction of heart disease by using feature selection methods and K-fold cross-validation. These techniques are incorporated into modern ensemble classification algorithms with Artificial Neural Networks (ANNs), Naive Bayes (NBs), K-Nearest Neighbors (KNN), J48, and Linear Regression (LRs). The dataset for this study was obtained from the Centers for Disease Control and Prevention (CDC), which is a national public health institute of the United States. Predicated upon these techniques and data, the study strives to increase precision and reliability in prediction models in order to more effectively identify at-risk patients for heart disease before symptom occurrence.

The structure of the study is organized as follows: Section 2 highlights the significance of relevant approaches to heart disease. Section 3 outlines the methodology, new strategies, and techniques. Section 4 presents empirical findings, discussions, and detailed analysis. User interfaces are described in Section 5. Finally, Section 6 concludes with a summary of the key findings and suggests future directions.

2 RELATED WORK

Machine learning has advanced recently and has greatly impacted healthcare by helping to predict and diagnose heart disease. Several studies were conducted using different ML algorithms on datasets with useful patient data.

Table 1. Summary of research findings

Ref.	Dataset	Efficient Algorithm	Limitations	Accuracy
[7]	UCI	KNN	Consume significant secondary memory for data storage.	90.79%
[8]	UCI	KNN	Consume significant secondary memory for data storage.	88.52%
[9]	UCI	KNN	Consume significant secondary memory for data storage.	87%
[10]	American Heart Association dataset	Neural networks	Processing large datasets requires significant computational power.	89%
[11]	South African	J48	Processing large datasets necessitates high computational power and can be slow.	99%
[12]	Indian patients	SVM	High power usage, very slow with big data.	86.42%
[13]	Cleveland and Hungarian	Random Forest	Handling large datasets is slow and requires substantial computational power.	100%
[14]	UCI	Random Forest	Handling large datasets is slow and requires substantial computational power.	95.60%

As depicted in Table 1, the study by Shah et al. [7] used four algorithms: random forest, decision tree, NB, and KNN. They utilized a pre-existing dataset from the Cleveland database in the UCI repository, which includes data on heart disease patients. The dataset has 303 instances with 76 attributes, though only 14 were used. The KNN algorithm achieved the highest accuracy score of 90.789%. Similarly, Jindal et al. [8] conducted a study on heart disease prediction using supervised algorithms such as random forest, KNN, and logistic regression. They found that KNN was the most efficient algorithm, achieving an accuracy of 88.52% on the UCI dataset. Singh and Kumar [9] applied four supervised algorithms—decision tree, support vector machine (SVM), LR, and KNN—to predict heart disease using the UCI repository data for training and testing. Their study showed that KNN was the best algorithm, with an accuracy of 87%. Amin et al. [10] performed a study focusing on predicting heart disease with genetic neural networks considering various risk factors. This study achieved an accuracy of 89%. Masethe and Masethe [11] researched heart disease prediction using Bayes Net, CART, NB, REPTREE, and J48 algorithms. For a dataset from South Africa, the J48 algorithm reported an impressive accuracy rate of 99%. Ghumbre et al. [12] worked on heart disease diagnosis using Radial Basis Function (RBF) and Support Vector Machine (SVM) algorithms with an Indian patient dataset, achieving an accuracy of 86.42%. Ali et al. [13] evaluated three classification algorithms: random forests, decision tree, and KNN. The random forests approach achieved 100% sensitivity, specificity, and accuracy. Katarya and Meena [14] conducted a study in which they applied various machine learning algorithms to the UCI dataset. Their findings revealed that the Random Forest algorithm was the most effective, achieving an accuracy rate of 95.60%.

Research on predicting heart disease using ML faces limitations in data quality, algorithm variability, and potential biases. These factors can affect the reliability, generalizability, and clinical applicability of the findings.

3 METHODOLOGY

The dataset employed in this study is obtained from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) [15]. This open-source dataset, intended for academic and research purposes, comprises 401,958 instances and 279 features. The cases are categorized as either "Yes" or "No" for heart disease. It includes 292,422 individuals without a risk of heart disease, while 27,373 individuals are at risk of heart disease (see Figure 1).

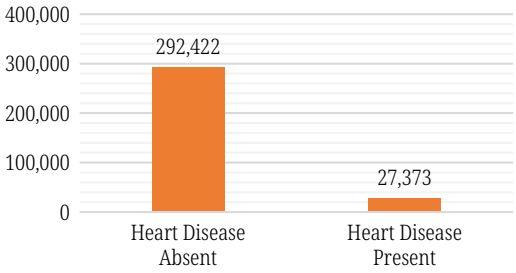


Fig. 1. Heart disease data

After choosing the dataset, the next crucial step is designing an architecture capable of meeting our expectations. As shown in Figure 2, the proposed system architecture comprises four fundamental phases: preprocessing, training, testing, and evaluation.

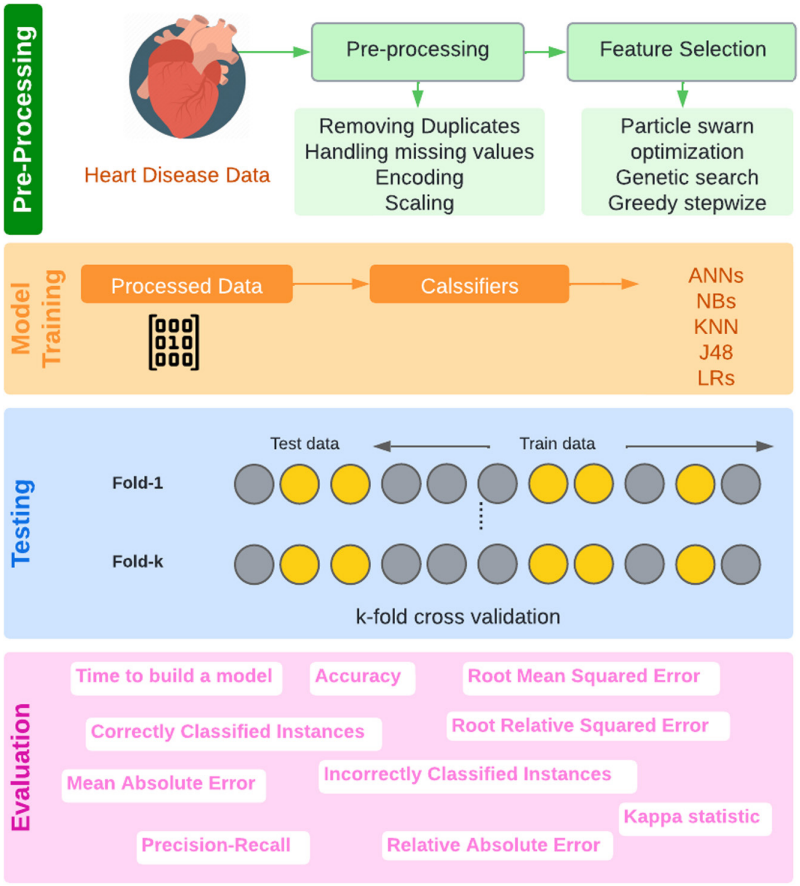


Fig. 2. Proposed operational workflow for heart disease classification

- **Preprocessing:** The preprocessing phase is the initial stage of any pipeline where raw data is cleaned and prepared for analysis. This involves cleaning operations such as the removal of duplicates, dealing with missing values, encoding categorical variables, and scaling numerical features. To ensure the preprocessing part was effective and the data was of good quality and would adapt well to the next steps, we applied feature selection techniques such as Particle Swarm Optimization, Genetic Search, and Greedy Stepwise due to their effectiveness in feature optimization and ability to explore complex search spaces. These methods were well-suited to our multivariate dataset, improving accuracy while reducing model complexity.
- **Training:** The training phase involves training the model by using the preprocessed data, which includes different algorithms for the same: ANN, KNN, NBs, LRs, and J48. The model learns from the input data, which identifies the patterns in the relationship between the input features and the target variable. This is an important stage because it determines how well the model can generalize and make correct predictions.
- **Testing:** The testing phase, as in Figure 3, is performed using K-fold cross-validation to scrutinize the model against an independent data set that has not been exposed to the model during the training process. We chose K = 10 for cross-validation as it offers a good balance between bias and variance, providing a more accurate estimate of model performance while helping to reduce the risk of overfitting.

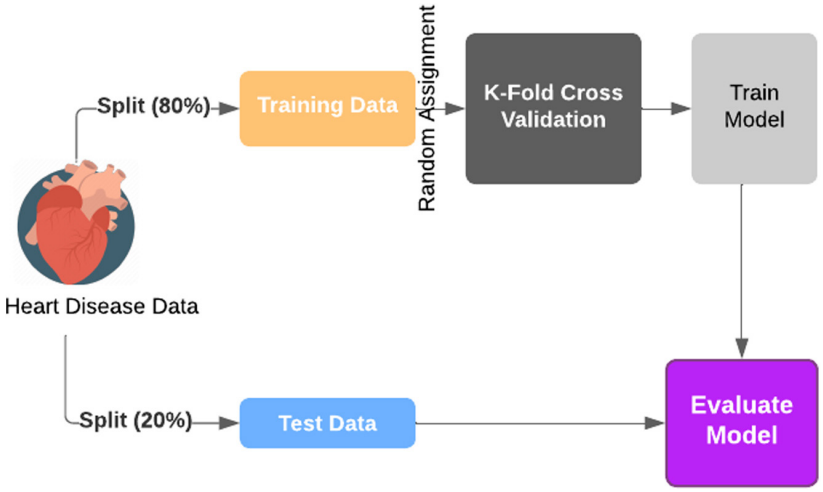


Fig. 3. K-fold cross-validation

The general algorithm of k-fold cross-validation is given in Algorithm 1.

Algorithm 1

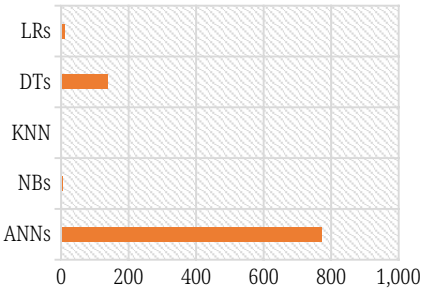
1. Shuffle the dataset randomly.
2. Split the dataset into k groups.
3. For each unique group:
   a. Treat the unique group as a test or validation dataset.
   b. Take the remaining groups as a training dataset.
   c. Fit a model on the training dataset and evaluate it on the test dataset.
   d. Keep the score of the evaluation and throw the model away.
4. Aggregate the recorded evaluation scores to summarize the model's overall performance.

- **Evaluation:** This is the last stage where the model's performance is thoroughly evaluated. This involves analyzing various performance metrics obtained during the testing phase and conducting cross-validation to ensure the model's robustness. The evaluation metrics considered for assessing the obtained model include:
  - Building time
  - Number of cases correctly classified
  - Number of cases incorrectly classified
  - Accuracy
  - Recall-Precision
  - Kappa statistic (KS)
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - Relative Absolute Error (RAE)
  - Root Relative Squared Error (RRSE)

4 RESULTS AND DISCUSSIONS

We now report a full evaluation of the developed model. The evaluation is based on different performance metrics that were obtained during testing and cross-validation in order to establish the robustness of the model. Such metrics provide insight not only with regard to effectiveness and accuracy but also toward understanding clearly what the strong features are and what things could be improved upon. The evaluation metrics in use should include time taken to build the model, correct and incorrect classified instances, accuracy, KS, MAE, RMSE, RAE, and RRSE. These metrics collectively offer a thorough evaluation of the model's predictive capabilities and overall performance.

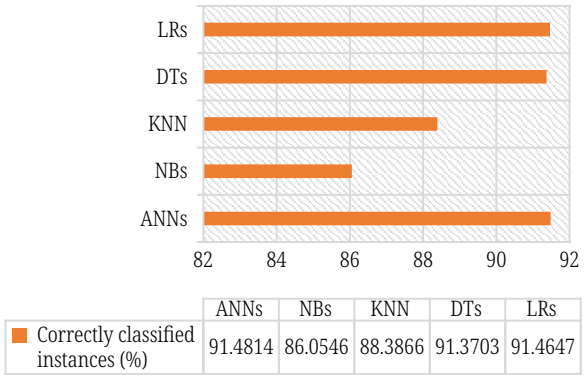
We first considered the metric model building time. This metric is important because it reveals the efficiency of an algorithm in terms of resources and time involved during computation. Results obtained are presented in Figure 4 in averages of seconds used to build a model for leading algorithms: ANN, KNN, NBs, LRs, and J48.



	ANNs	NBs	KNN	DTs	LRs
Time taken to build model (in Sec)	771.06	1.62	0.05	136.48	11.04

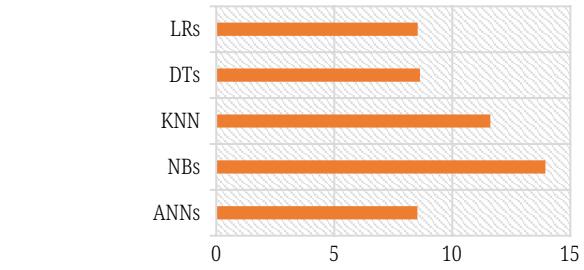
Fig. 4. Time taken to build model

The next metrics we examined are the correctly and incorrectly classified instances. These are essential for assessing our predictive model's performance. Correctly classified instances show how many predictions the model got right, whereas incorrectly classified instances reveal how many mistakes it made. Figures 5 and 6 visualize these results, respectively, offering a clear view of the model's accuracy and error rates across various algorithms.



	ANNs	NBs	KNN	DTs	LRs
Correctly classified instances (%)	91.4814	86.0546	88.3866	91.3703	91.4647

Fig. 5. Correctly classified instances



	ANNs	NBs	KNN	DTs	LRs
Incorrectly classified instances (%)	8.5186	13.9454	11.6134	8.6297	8.5353

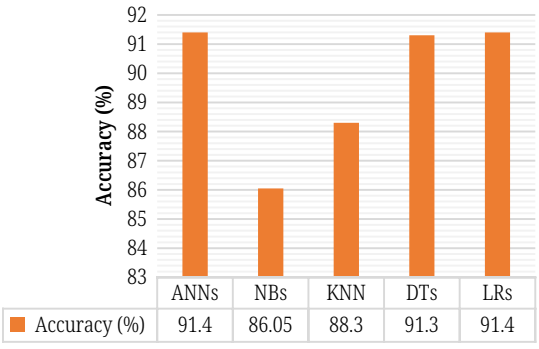
Fig. 6. Incorrectly classified instances

Another crucial metric we examined is the accuracy of the model, which provides an overall indication of how well the model is performing. The formula employed to calculate the accuracy is as follows (1):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Where True Positives (*TP*) are the number of cases where a model correctly predicts the positive class, True Negatives (*TN*) are the number of cases where a model correctly predicts the negative class, False Positives (*FP*) are the number of cases where a model incorrectly predicts the positive class, and False Negatives (*FN*) are the number of cases where a model incorrectly predicts the negative class.

Figure 7 shows the results in accuracy where the different algorithms perform regarding predicting heart disease.



	ANNs	NBs	KNN	DTs	LRs
Accuracy (%)	91.4	86.05	88.3	91.3	91.4

Fig. 7. Accuracy of results

A number of key metrics that were measured over both the training and simulation phases covered: *Ks*, *MAE*, *RMSE*, *RAE*, and *RRSE*. Each of these different metrics provides insight into the model, showing different elements: model accuracy, error rates, and general performance in predicting heart disease. *Ks* is a measure for inter-rater agreement or reliability that adjusts for the agreement due to chance. It compares the observed agreement between two raters or methods with the agreement expected from chance alone. Equation (2) calculates the Kappa Statistic as:

$$Ks = \frac{p_0 - p_e}{1 - p_e} \tag{2}$$

Where p_0 is the observed agreement proportion and p_e is the expected agreement proportion by chance. *MAE* measures how much, on average, the predictions made differ from the actual values. It measures the average absolute difference between predicted and actual values over all observations. The formula that is used to calculate *MAE* is given below (3):

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|V_{predict} - V_{observ}\right| \tag{3}$$

N is the number of observations, $V_{predict}$ denotes the predicted value and V_{observ} denotes the observed value of the dependent factor.

RMSE was calculated taking the square root of the average of squared differences between predicted and actual values across all observations (cf. equation 4).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}[V_{predict} - V_{observ}]^2}{N}} \tag{4}$$

To assess the performance of our predictive model, we calculated the *RAE* and the *RRSE* using equations (5) and (6), respectively.

$$RAE = \frac{\sum_{i=1}^{n}\left|V_{predict} - V_{observ}\right|}{\sum_{i=1}^{n}\left|V_{predict} - \bar{V}_{observ}\right|} \tag{5}$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}[V_{predict} - V_{observ}]^2}{\sum_{i=1}^{N}[V_{predict} - \bar{V}_{observ}]^2}} \tag{6}$$

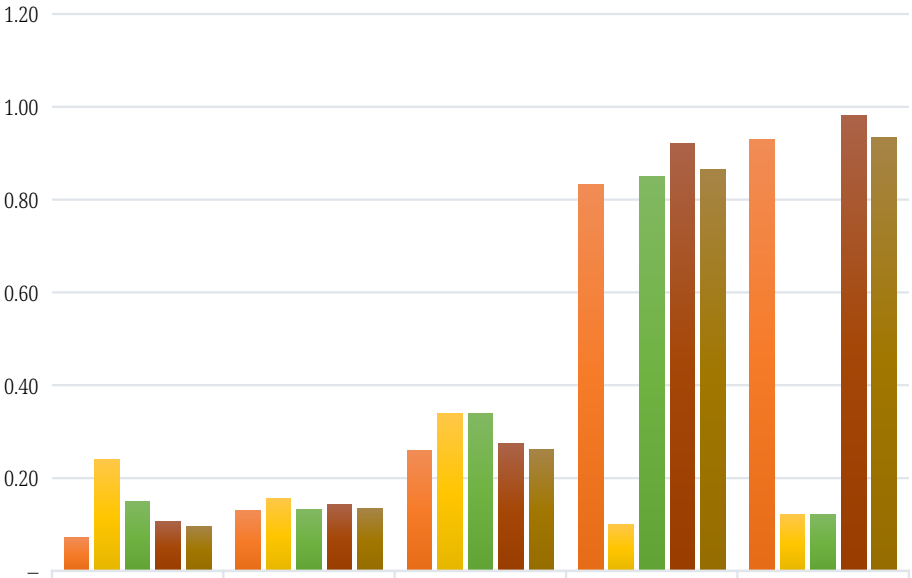
Where, $V_{predict}$ is the predicted value, V_{observ} is the observed value, and $\bar{V}_{observ}$ is the mean of the actual values.

The *RAE* is defined as the total absolute error of the sum of absolute error that will be measured through a simple predictor, whereas the *RRSE* is defined as the square root of square error that will be measured through the difference between a simple predictor and a particular predictor.

Shown in Figure 8 are the *Ks*, *MAE*, *RMSE*, *RAE*, and *RRSE* results.

To assess the effectiveness of our predictive models toward identifying a high-risk group for heart disease, we developed a table with major performance measures.

The measured indicators include True Positive Rate (*TP* Rate), False Positive Rate (*FP* Rate), Precision, Recall, F-Measure, and Class. Each measure looks at a different aspect of a model's performance; hence, from the results it is possible to have a great evaluation of the prediction capability. The details are presented in Table 2.



	Ks	*MAE*	*RMSE*	*RAE*	*RRSE*
ANNs	0.07	0.13	0.26	0.83	0.93
NBs	0.24	0.16	0.34	0.10	0.12
KNN	0.15	0.13	0.34	0.85	0.12
DTs	0.11	0.14	0.27	0.92	0.98
LRs	0.10	0.13	0.26	0.86	0.93

Fig. 8. Training and simulation results

Table 2. Accuracy measures

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
ANNs	0.996	0.953	0.918	0.996	0.955	No
	0.047	0.004	0.501	0.047	0.086	Yes
KNN	0.949	0.819	0.926	0.949	0.937	No
	0.181	0.051	0.25	0.181	0.21	Yes
NBs	0.905	0.621	0.94	0.905	0.922	No
	0.379	0.095	0.272	0.379	0.316	Yes
LRs	0.994	0.936	0.919	0.994	0.955	No
	0.064	0.006	0.493	0.064	0.114	Yes
J48	0.992	0.927	0.92	0.992	0.955	No
	0.073	0.008	0.46	0.073	0.126	Yes

We also set up a confusion matrix to really understand how our prediction models performed. The matrix provides important detailed information about the

classification accuracy of the model, showing how many *TP*, *TN*, *FP*, and *FN* are described. We present the confusion matrix in Table 3, which is very important to draw the conclusion of our models' effectiveness in identifying the risk factor for heart disease among individuals.

Table 3. Confusion matrix

	No	Yes	Class
ANNs	289455	1264	No
	25808	1270	Yes
KNN	275982	14737	No
	22170	4908	Yes
NBs	263226	27493	No
	16825	10253	Yes
LRs	288929	1790	No
	25335	1743	Yes
DTs(J48)	288389	2330	No
	25095	1983	Yes

In our study, we primarily aimed to improve early heart disease prediction by employing advanced feature selection techniques, K-fold cross-validation, and high-performance ensemble classification algorithms. We employed a range of models, including ANNs, KNNs, NBs, LRs, and the J48 decision tree, to assess their effectiveness in accurately classifying instances of heart disease.

Figure 4 illustrates the time required for model construction, with ANNs being the most time-intensive at 771.06 seconds, reflecting its complex training process. In contrast, KNN was the quickest at 0.05 seconds, followed by NBs at 1.62 seconds, LRs at 11.04 seconds, and DTs at 136.48 seconds. However, selecting the best algorithm requires evaluating additional performance metrics.

Figure 5 depicts the accuracy of these models, with ANNs achieving the highest at 91.48%, closely followed by LRs at 91.46% and DTs at 91.37%. KNN and NBs exhibited lower accuracies of 88.39% and 86.05%, respectively, suggesting they may be less suitable for this specific prediction task.

Figure 6 highlights ANNs' superior performance in minimizing classification errors, with the lowest percentage of incorrectly classified instances at 8.52%. Additionally, Figure 7 reaffirms ANNs' effectiveness with an accuracy of 91.4% in managing complex data.

Further analysis in Figure 8 reveals that both ANNs and KNN achieved the lowest MAE values at 0.13, underscoring their precision in predictions. ANNs also demonstrated a low RMSE of 0.26, along with LRs, indicating minimal deviations from actual values.

From Table 2, ANNs, LRs, NBs, and J48 demonstrated strong performance in identifying instances of the "No" class with high TP Rates and F-Measure.

However, all classifiers encountered challenges in predicting instances of the "Yes" class with lower Precision and Recall metrics, emphasizing the difficulty in accurately classifying these instances compared to those where the class is "No".

In conclusion, after comprehensive evaluation of all metrics, ANNs emerge as the most effective algorithm among those studied for predicting heart disease, showcasing superior accuracy and reliability in handling complex data and classification tasks.

5 USER INTERFACES

Figure 9 presents the HeartScan mobile application developed to predict the presence of heart diseases based on the data submitted by users. At first, when the application is opened, the user will be asked to fill in details such as Body Mass Index (BMI), status for diabetes, gender, age, asthma condition, kidney disease, and other relevant health information, as shown in Figure 10.

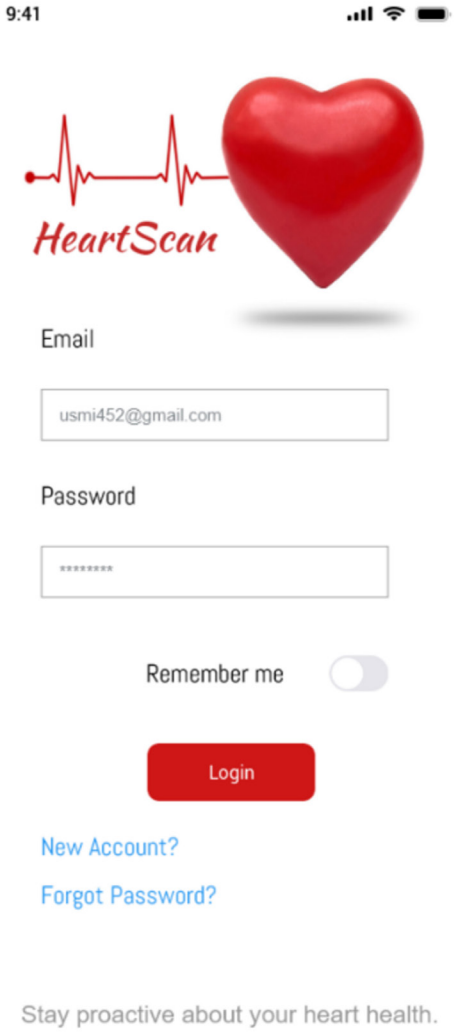


Fig. 9. HeartScan mobile application

The application ensures that all necessary fields are completed to provide an accurate assessment. Once all the required information is entered, users can submit their data for analysis.

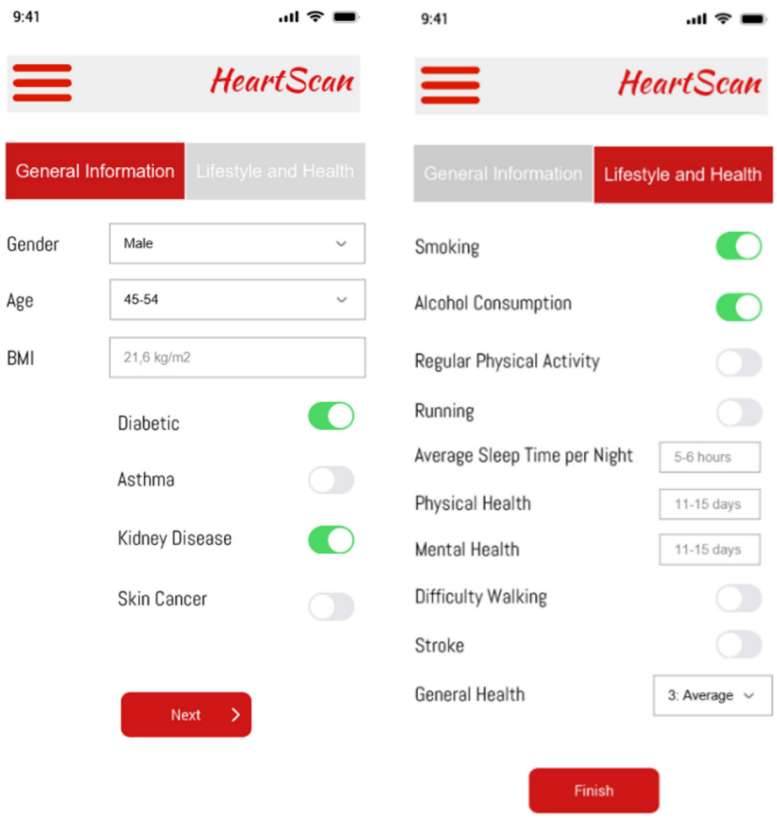


Fig. 10. User lifestyle and health information input screen

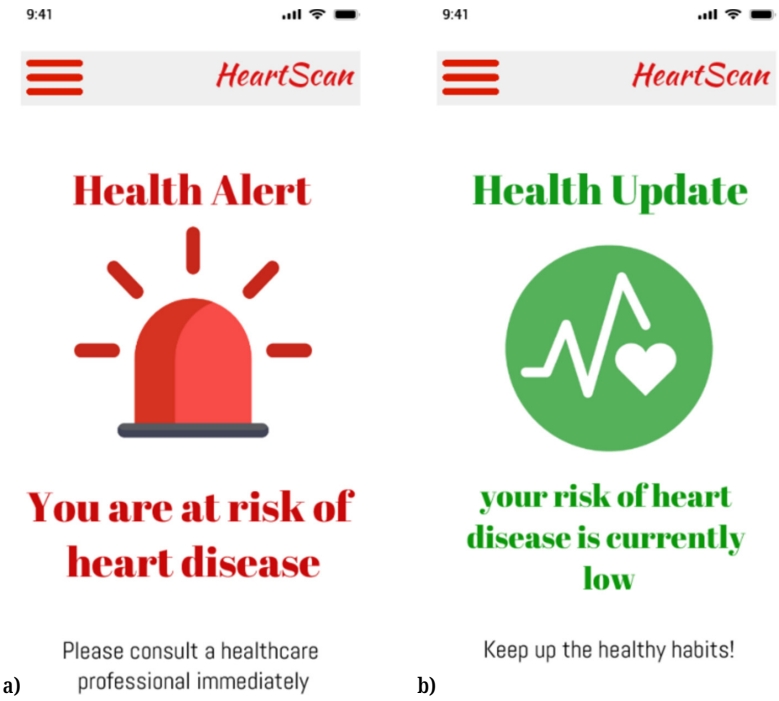


Fig. 11. Health prediction results

The result is displayed as a clear and concise message informing the user of their risk level for heart disease. If the application identifies a significant risk, it suggests seeking medical advice for further evaluation and management, as shown

in Figure 11a. Otherwise, if the user has no risk of heart disease, a message is displayed as shown in Figure 11b.

HeartScan's intuitive design is quite friendly and non-intimidating, making it an app for easy access and use by everyone, hence empowering them into proactive steps in monitoring their health. The predictive nature of the app can be channeled into early detection and prevention of diseases.

6 CONCLUSION

Heart diseases are a big concern in public health globally, contributing to the bulk of morbidity, disability, and mortality worldwide. Early detection of these diseases can be very important in improving outcomes and lessening their impact. Our study will work on developing predictive models by using techniques for feature selection and K-fold cross-validation with ML algorithms. In the experiments conducted, peak accuracy was found to be 91.48% using Artificial Neural Networks.

Here are three future works for this study:

- Investigate the applicability of the developed predictive models across diverse demographic and geographic populations to assess their generalizability.
- Explore the integration of new features or data sources to enhance the predictive accuracy of the models.
- Implement real-time or continuous monitoring systems based on the developed models to enable early detection and intervention for individuals at risk of heart disease.

7 REFERENCES

[1] A. B. Bhatt *et al.*, "Congenital heart disease in the older adult: A scientific statement from the American Heart Association," *Circulation*, vol. 131, no. 21, pp. 1884–1931, 2015. https://doi.org/10.1161/CIR.0000000000000204

[2] A. Domyati and Q. Memon, "Machine learning based improved heart disease detection with confidence," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 8, pp. 130–143, 2023. https://doi.org/10.3991/ijoe.v19i08.37417

[3] M. Asghari-Jafarabadi, K. Gholipour, R. Khodayari-Zarnaq, M. Azmin, and G. Alizadeh, "Estimation of myocardial infarction death in Iran: Artificial neural network," *BMC Cardiovascular Disorders*, vol. 22, 2022. https://doi.org/10.1186/s12872-022-02871-8

[4] A. El-Ibrahimi, O. Terrada, O. E. Gannour, B. Cherradi, A. E. Abbassi, and O. Bouattane, "Optimizing machine learning algorithms for heart disease classification and prediction," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 15, pp. 61–76, 2023. https://doi.org/10.3991/ijoe.v19i15.42653

[5] T. Münzel *et al.*, "Environmental risk factors and cardiovascular diseases: A comprehensive expert review," *Cardiovascular Research*, vol. 118, no. 14, pp. 2880–2902, 2022. https://doi.org/10.1093/cvr/cvab316

[6] N. Narisetty, A. Kalidindi, M. V. Bujaranpally, N. Arigela, and V. V. Ch, "Ameliorating heart diseases prediction using machine learning technique for optimal solution," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 16, pp. 156–165, 2023. https://doi.org/10.3991/ijoe.v19i16.42071

[7] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, 2020. https://doi.org/10.1007/s42979-020-00365-y

[8] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, p. 012072, 2021. https://doi.org/10.1088/1757-899X/1022/1/012072

[9] A. Singh and R. Kumar, "Heart disease prediction using machine learning algorithms," in *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, 2020, pp. 452–457. https://doi.org/10.1109/ICE348803.2020.9122958

[10] S. U. Amin, K. Agarwal, and R. Beg, "Genetic neural network-based data mining in prediction of heart disease using risk factors," in *2013 IEEE Conference on Information & Communication Technologies*, 2013, pp. 1227–1231. https://doi.org/10.1109/CICT.2013.6558288

[11] H. D. Masethe and M. A. Masethe, "Prediction of heart disease using classification algorithms," in *Proceedings of the World Congress on Engineering and Computer Science*, vol. 2, 2014, no. 1, pp. 25–29.

[12] S. Ghumbre, C. Patil, and A. Ghatol, "Heart disease diagnosis using support vector machine," in *International Conference on Computer Science and Information Technology (ICCSIT'),* Pattaya, 2011, pp. 84–88.

[13] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, p. 104672, 2021. https://doi.org/10.1016/j.compbiomed.2021.104672

[14] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study and analysis," *Health and Technology*, vol. 11, no. 1, pp. 87–97, 2021. https://doi.org/10.1007/s12553-020-00505-7

[15] K. Pytlak, "Indicators of Heart Disease (2022 UPDATE)," kaggle, 2022. [Online]. Available at: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

8 AUTHORS

Inssaf El Guabassi is with the LAROSERI Laboratory, Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco (E-mail: elguabassi@gmail.com).

Zakaria Bousalem is with the Polydisciplinary Faculty, Sultan Moulay Slimane University, Beni Mellal, Morocco.

Rim Marah is with the Faculty of Economics and Management, Sultan Moulay Slimane University, Beni Mellal, Morocco.

Abdellatif Haj is with the Faculty of Sciences and Technologies, Hassan 1st University, Settat, Morocco.