Detection of Cardiovascular Disease by Machine Learning with Interpretability

Syed Md. Ahnaf Hasan
Dept. of Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
ahnafhasan335@gmail.com

***Abstract*—Heart diseases are one of the most prominent causes of death around the world. Early detection can lead to proper treatment which can save lives. The recent use of machine learning algorithms in diagnosing diseases has shown a lot of promise sparking its use in detecting cardiovascular disease as well. In this work, we have evaluated the performance of 4 machine learning models: Gaussian Naive Bayes, Random Forest, AdaBoost and GradBoost on a cardiovascular disease dataset. The results show that the Random Forest and GradBoost models obtain the best scores with 97.73% accuracy and 97.72% in F1 score. Finally, we have used LIME and SHAP XAI models to explain the predictions made by the model, to gain insight on the prediction making mechanism. It is seen that the explanations made by the XAI models can be used to verify the predictions made by the model, as they add transparency to the inner workings of the model. Hence, the use of machine learning models in diagnosing cardiovascular disease shows favourable prospects.**

In this work, we have analyzed the performances of four Machine Learning algorithms: Gaussian Naive Bayes, Random Forest, AdaBoost and GradBoost in detecting cardiovascular disease from a Kaggle dataset. Consequently, we have used two XAI models: LIME and SHAP to generate local and global explanations to the models' predictions respectively to add transparency.

***Index Terms*—Explainable AI, LIME, SHAP, Gaussian Naive Bayes, AdaBoost, Random Forest, GB**

I. Introduction

Cardiovascular disease is among the killer diseases that takes lives every year. World Health Organization (WHO) says that almost 31% of the deaths in world population happening due to heart disease [1]. Early diagnosis is therefore crucial in saving lives. Diagnosing diseases manually with clinical expertise can be time consuming and complex. Recent studies show the effectiveness of Machine Learning algorithms in automating the diagnosing processo of various diseases [2], [3].

II. Related Works

Recent research works have shown a lot of promise in using machine learning algorithms for diagnosing cardiovasular diseases. Hossent et al. used ML models such as Random Forest, Logistic Regression and Decision Tree on the UCI Cleveland dataset and found satisfactory results [4]. The Logistic regression model was found to be the best performing model with an accuracy of 92.10% in this research. Nadakinamani et al. compared the performance of various ML algorithms in such as REP Tree, M5P Tree, Random Tree, Naive Bayes, etc. in predicting cardiovascular disease. The Random Tree model was found to be the best performing model in the work with loawest MAE of 0.001 and lowest RMSE of 0.0231. In a recent work [5], a model using k-modes clustering with Huang starting was developled. They used GridSearchCV to hypertune the parameters of the developed model. Jiang et al. also analyzed various ML models in detecting cardiovascular disease among patients and found the XGBoost model showing very good results with an AUC score of 0.937 [6]. From the recent researches it is evident that Machine Learning algorithms show promising results in detecting cardiovascular disease. However, the model's predictions may not be inherently transparent. Paudel et al. shows that XAI models such as LIME should additionally be used with the Machine Learning models to provide transparency in the models' predictions [7] .

III. Methodology

The workflow of this research is shown in Fig. 1. At first the target variable is label encoded to convert the text terms to numerical values. Consequently the dataset is split into training and testing parts which constitute 80% and 20% of the whole dataset respectively. Later we train the machine learning models with the training split and later use the testing split to evaluate the performance of the models. The Machine Learning models used in this work are Gaussian Naive Bayes, Random Forest , AdaBoost and GradBoost. Finally, LIME and SHAP XAI models are used to interpret the predictions made by the models locally and globally.

A. Dataset

The Heart Disease Classification Dataset [7] available in Kaggle is used in this research. The distribution of the dataset is shown in Fig. 2. From the distribution plot it is observed that the dataset is slightly imbalanced with 810 instances for positive class and 509 instances for the negative class respectively. There are 8 feature variables available in the dataset.

B. Machine Learning Models

1) Gaussian Naive Bayes: Naive Bayes is a probabilistic algorithm that utilizes Bayes' theorem, assuming independence among features, which makes it particularly suitable for datasets with a moderate to large number of features.

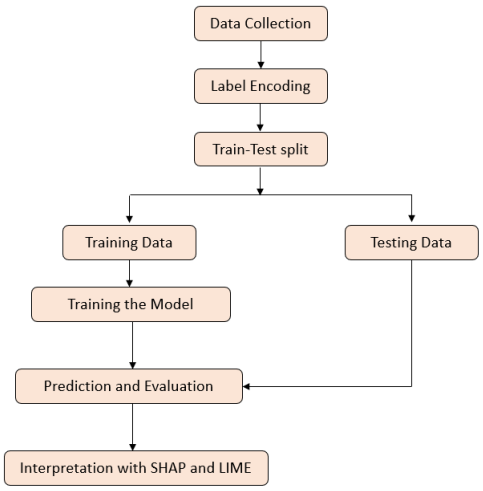

Fig. 1. Methodology

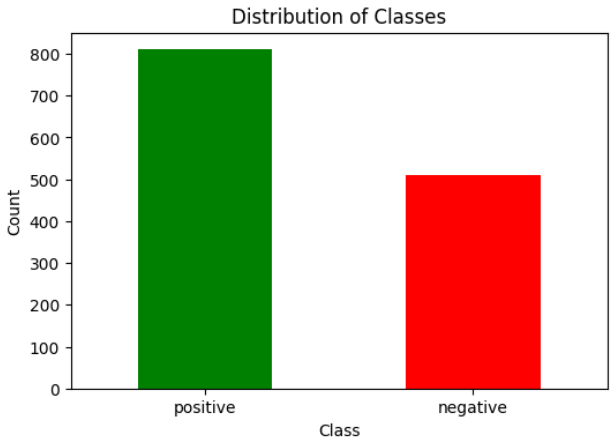

Fig. 2. Distribution of the dataset according to classes

Naive Bayes is one of the most popularly used algorithms for classification problems. It considers each of the features to be conditionally independent. Sadhu et el. used Gaussian Naive Bayes for Diabetes Risk Prediction [8]. The Gaussian Naive bayes uses a continuous distribution for the features to generate conditional probabilities which are used for calculating the class probability of an instance. It generates probabilities for each of the classes, later the class with the highest probability is considered as the predicted label.

2) Random Forest: A Random forest is an ensemble of decision trees [9]. A decision tree by itself may not make the most robust decision for a given problem. However, the combination of multiple decision trees' decisions can improve the generalization ability significantly. In this algorithm multiple decision trees make decisions on the bootstrapped datasets from the main dataset. Each of the predictions are later combined via bagging.

3) AdaBoost: Adaptive Boosting (AdaBoost) algorithm is the combination of multiple weak learners [10]. In this algorithm, multiple weak learners are used, each of which improve on the misclassified data of the previous weak learner. This significantly improves the robustness of the model.

4) GradBoost: Gradient Boost (GradBoost) is another boosting algorithm which combines multiple weak learners to improve the generalization ability of the overall model [11]. In this ensemble of weak learners the consecutive weak models are fit on the pseudo residuals of the former weak models instead of being fit on the actual labels.

C. Explainable AI (XAI)

Many of the machine learning models are not inherently interpretable. Therefore, their predictions are difficult to trust. In this regard, Explainable AI (XAI) models can provide use with explanations that show the decision making mechanism of the ML models used. LIME (Local Interpretable Model-Agnostic Explanations) is an XAI model which can generate perturbed instances around a local instance and generate explanations for the instance locally [12]. SHAP is another XAI model which can generate both local and global explanations with Shapley values [13]. Local explanations are with respect to a particular instance, while global explanations are generated taking all the instances into account.

D. Performance metrics

Four performance metrics: Accuracy, Precision, Recall and F1 score have been used to evaluated the trained Machine Learning models. Accuracy is the most common performance metric, which is the ratio of the total number of accurately detected instances to the total number of instances. This metric can be misleading if the dataset is imbalanced. Since, our dataset is imbalanced we additionally use precision, recall and f1 score. Precision tells us how many of the predicted instances for a class were actually correct. While recall tells us how many instances in a class were correctly identified. F1 score is the harmonic mean of precision and recall which provides a balance between the two metrics.

IV. Results and Analysis

A. Performance analysis of the Machine Learning models

The four Machine Learning models: Gaussian Naive Bayes, Random Forest, AdaBoost and GradBoost are trained using the train split of the dataset. Later, they are evaluated using the test split. The evaluation results of the models across the various performance metrics are shown in Table 1. It is observed that Random Forest and GradBoost algorithms obtain identical scores across accuracy, precision, recall and f1 score which are 97.73%, 97.73%, 97.73% and 97.72% respectively. The AdaBoost algorithm comes in second across all the metrics with competitive scores, while Gaussian Naive Bayes has worst performing scores, which are significantly lower than the other 3.

Fig. 3 shows the confusion matrix of the Random Forest model on the test set. It is able to predict 161 and 97 positive and negative instances correctly, while it misclassifies the rest. The findings from all the confusion matrices of the four

models used are summarized in Table 2. As before, we see that the classification scenarios for the Random Forest and GradBoost models are similar. The Gaussian Naive Bayes though performs the worst in detecting True Positives, it is the most efficient in detecting True Negatives, as it only misclassifies one negative instance as positive.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Accuracy	Precision	Recall	F1
GNB	78.79%	85.86%	78.79%	78.92%
RF	97.73%	97.73%	97.73%	97.72%
Adaboost	96.21%	96.23%	96.21%	96.20%
GB	97.73%	97.73%	97.73%	97.72%

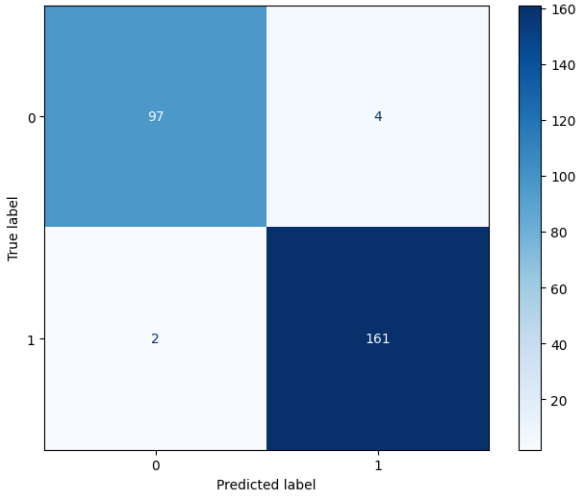

Fig. 3. Confusion Matrix of Random Forest

TABLE II
CONFUSION MATRIX RESULTS

Algorithm	TP	FP	FN	TN
GNB	108	1	55	100
RF	161	4	2	97
Adaboost	160	7	3	94
GB	161	4	2	97

B. Interpreting the predictions with XAI

1) Local Explanations with LIME: The LIME local explanations generated for the Random Forest and AdaBoost predictions on test instance-2 are depicted in Fig. 4 and Fig. 5. The LIME explanations rank the features used for predicting the particular instance. The ranked featuer are assigned weights for each of the classes, indicating which feature the model considered as a contribution for the prediction of a particular class. The probability values for each of the classes are displayed as well. The class with the highest probability value is the assigned class by the model. Though both of the models' predictions are same for test instance-2 which is negative, their probability values are different. Furthermore, their feature rank is different as well. Though both the models rank troponin and kcm as the first and second important features respectively, the random forest model ranks age as teh third important feature, which GradBoost considers age to be the least important feature. This LIME explanation can be used as a validation for the model's predictions, as we can see how model made its predictions.

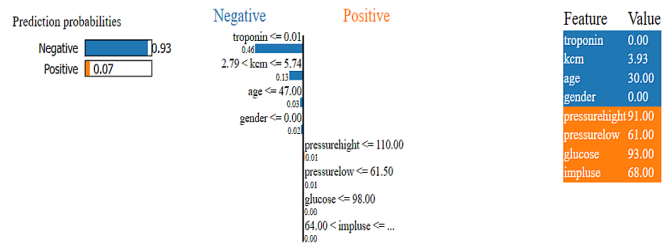

Fig. 4. LIME explanation for test instance-2 prediction by Random Forest

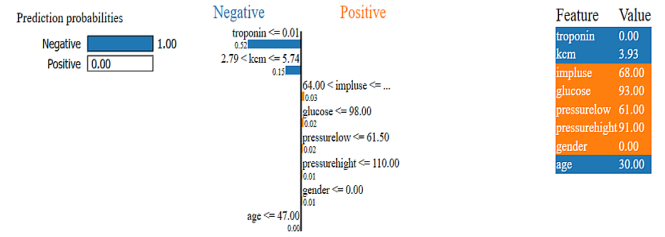

Fig. 5. LIME explanation for test instance-2 prediction by Gradient Boosting

2) Global Explanations with SHAP: Fig. 6 and Fig. 7 show the SHAP global explanations for Random Forest and GradBoost respectively in the form of Bee Swarm plot. Which feature value increase the probability of class 1 or decrease it for each of the models can be seen from the plots. It can be see that for both the models high values of troponin and kcm contribute to predicting the class as negative with the exception of a few outliers. Again, we see that the rest of the feature importance by both the models are not ranked similarly as age is ranked lowest for Gradboost, while it is ranked third by Random Forest.

CONCLUSION

In this project, we have tested the performances of four machine learning models for a cardiovascular disease dataset. The results show that the Random Forest and GradBoost models obtain the best performing scores across the various performance metrics for the test set. The addition of XAI models LIME and SHAP provide in-depth explanation to how the models made their predictions based on the features. The implementation of LIME and SHAP encourage the use of XAI models with ML algorithms in various tasks especially in the case of medical diagnosis, as false predictions made by the model can lead to dangerous consequences. This work can be

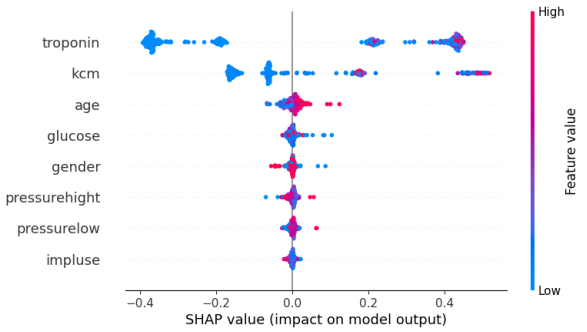

Fig. 6. SHAP Bee Swarm plot for predictions by Random Forest Model

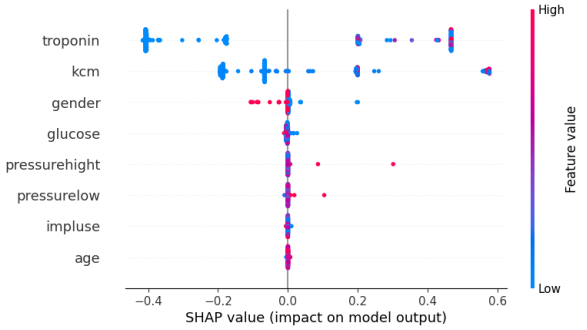

Fig. 7. SHAP Bee Swarm plot for predictions by Gradient Boosting Model

further improved by using larger and more diverse datasets, analyzing performances of other machine learning and deep learning models, and also by using other XAI models to gain more insights.

ACKNOWLEDGMENT

I would like to convey my gratitude to Annajiat Alim Rasel sir for his supervision and guidance on this project.

REFERENCES

[1] 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.

[2] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," in *Healthcare*, vol. 10, no. 3. MDPI, 2022, p. 541.

[3] N. Kumar, N. N. Das, D. Gupta, K. Gupta, and J. Bindra, "Efficient automated disease diagnosis using machine learning models," *Journal of healthcare engineering*, vol. 2021, 2021.

[4] M. A. Hossen, T. Tazin, S. Khan, E. Alam, H. A. Sojib, M. Moniruzzaman Khan, and A. Alsufyani, "Supervised machine learning-based cardiovascular disease analysis and prediction," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–10, 2021.

[5] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, p. 88, 2023.

[6] H. Jiang, H. Mao, H. Lu, P. Lin, W. Garry, H. Lu, G. Yang, T. H. Rainer, and X. Chen, "Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease," *International Journal of Medical Informatics*, vol. 145, p. 104326, 2021.

[7] P. Paudel, S. K. Karna, R. Saud, L. Regmi, T. B. Thapa, and M. Bhandari, "Unveiling key predictors for early heart attack detection using machine learning and explainable ai technique with lime," in *Proceedings of the 10th International Conference on Networking, Systems and Security*, 2023, pp. 69–78.

[8] A. Sadhu and A. Jadli, "Early-stage diabetes risk prediction: A comparative analysis of classification algorithms," *International Advanced Research Journal in Science, Engineering and Technology (IARJSET)*, vol. 8, no. 2, pp. 193–201, 2021.

[9] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[10] C. Ying, M. Qi-Guang, L. Jia-Chen, and G. Lin, "Advance and prospects of adaboost algorithm," *Acta Automatica Sinica*, vol. 39, no. 6, pp. 745–758, 2013.

[11] X. Zhu and J. Chen, "Risk prediction of p2p credit loans overdue based on gradient boosting machine model," in *2021 IEEE international conference on power, intelligent computing and systems (ICPICS)*. IEEE, 2021, pp. 212–216.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf