Comparing Four Machine Learning Models and an Ensemble for Predicting Heart Attacks from Routine Clinical Features

Oliver Olejar
Western University
oolejar@uwo.ca

Micky Huynh
Western University
mhuynh52@uwo.ca

Jinil Desai
Western University
jdesai43@uwo.ca

***Abstract*—Cardiovascular disease (CVD) is the leading cause of death globally. An estimated one-third of deaths in 2021 stemmed from CVD. Many factors contributing to cardiovascular disease are preventable, with 80% of premature heart attacks and strokes being avoidable with the right interventions and warning signs. Main risk factors for cardiovascular disease are already being collected routinely, and the amount of medical and health data continues to grow. A large collection of information allows for the opportunity to employ machine learning led analysis as a potential intervention method for early detection. This paper examines the performance of four models (XGBoost, TabPFN, Neural Network, and Random Forest Classifier) and an ensemble model comprising all four, to predict heart attacks from key patient risk factors. The ensemble model uses a weighted voting system to determine the final classification. Their performance was analyzed using accuracy, precision, recall, F1-score, AUROC, AUPRC, and cross-validation. Based on these 6 performance metrics, the ensemble model was shown to be marginally better than XGBoost and TabPFN model, on the basis of AUPRC by approximately 1%, with an accuracy of approximately 98%. This shows promising evidence that machine learning powered analysis can serve as an intervention method for early detection of heart attacks, as well as provide valuable insights for clinicians.**

I. Introduction

Cardiovascular disease (CVD) is the foremost cause of death globally, which comprises one-third of all mortality, or 20.5 million deaths in 2021 [1]. CVD has many risk factors, meaning it is preventable with the right interventions. Prevention is predicted to be possible for 80% of premature CVD, such as strokes and heart attacks [1]. Risk factors for CVD include physical activity levels, sodium intake, alcohol consumption, tobacco smoking, obesity, raised blood pressure, diabetes, lipids, and air pollution [1]. Information and statistics on CVD risk factors, such as the aforementioned predictors, are already being collected regularly for patients. About 30% of the world's data is generated solely by the healthcare industry [2]. Vast amounts of patient records and metrics enable the application of machine learning driven analysis. This paper presents a comparison of four machine learning models trained to predict heart attacks. Included in this comparison of the four models (XGBoost, TabPFN, Neural Network, and RandomForestClassifier) is an ensemble model comprising all four models. The performance of all of these models are evaluated using accuracy, precision, recall, F1-score, AUROC, and AUPRC.

This study aims to identify the most effective predictive models for accurately determining the presence of cardiovascular disease in patients. Model performance is primarily evaluated based on predictive accuracy to determine the most reliable approach. The findings of this research have the potential to support clinical decision-making by providing medical practitioners with data-driven tools to aid in early diagnosis. Ultimately, this work seeks to improve patient outcomes by enabling earlier and more accurate detection and prevention of cardiovascular disease.

II. Related Work

The paper by DafniRose et al. [3] utilizes six different machine learning models to predict cardiac disease and compares the results. These models include Random Forest, Linear Regression, Support Vector Machine, K-Nearest Neighbour Classifier, Gaussian Naive Bayes, and a Gradient Boosting Classifier. To select the features for training, a chi-squared test was used. Then, cross-validation was applied to choose the number of features from the top scores. They evaluated the models using accuracy, precision, recall, and F1-score. Excluding the k-fold validation results, the Gaussian Naive Bayes model performed the best for predicting cardiovascular disease. The results with k-fold validation show that the Gaussian Naive Bayes model once again had the highest performance. This paper misses the opportunity to examine AUROC and AUPRC, which can reveal performance issues that the four types of metrics analyzed may miss.

This paper [4] develops a hybrid ensemble model to predict heart attacks using three machine learning and deep learning methods. These include SVM (Support Vector Machine), ANN (Artificial Neural Network), and LSTM (Long Short-Term Memory network): a type of recurrent neural network designed to learn from sequential or time-series data. SVM (Support Vector Machine): a supervised classifier good for structured data and clear class boundaries. ANN (Artificial Neural Network): a general nonlinear model that can capture complex feature relationships. LSTM (Long Short-Term Memory network): a type of recurrent neural network designed to learn from sequential or time-series data. LSTM is designed

for sequential data, for example, ECG signals or continuous monitoring over time. If the dataset is tabular (one row per patient, static medical attributes), LSTM's sequential modelling capabilities are wasted, and it can even overfit. SVM scales poorly as data grows; kernel computations become slow and memory-intensive. It is also sensitive to hyperparameter tuning and data normalization. Combining three very different model types (SVM, ANN, LSTM) adds complexity but not necessarily better accuracy unless there's true complementarity. Without careful blending (stacking, weighted averaging, calibration), the ensemble can produce unstable predictions.

The paper introduced by B.N. Nava-Martinez et al. [5] evaluates twelve models, six supervised learns and six ensemble approaches, achieving accuracies in the range 90.2% and 98.9% on three independent clinical datasets. The ensemble methods are: stacked models built on XGBoost, Random Forest, Logistic Regression, Decision Trees, Gradient Boosting, and a separate soft-voting ensemble. Their approach has high accuracy, but is computationally heavy, which may prove to be impractical for a hospital with limited resources. The paper relies on static datasets, which raises questions about how the model will interact with real-time data that is often incomplete or noisy. Further adding to this issue, the models being evaluated on dataset three, a small dataset with 299 samples, show weaker correlations and inconsistent patterns. It also implies underfitting on several of the models. The study also required lengthy, computationally intensive preprocessing, suggesting that the raw data contain substantial noise, skew, and variance, which introduces another limitation: hospitals may not have access to this level of preprocessing.

III. Methodology

A. Dataset

The dataset used in this study was sourced from the Kaggle platform and contains clinical records collected at Zheen Hospital in Erbil, Iraq, between January 2019 and May 2019 [6]. The dataset includes patient-level attributes such as age, gender, heart rate, systolic and diastolic blood pressure, blood glucose level, creatine kinase-MB (CK-MB), and troponin. The target variable is a binary outcome indicating the presence or absence of cardiovascular disease. Although the dataset was found to be relatively complete, standard data cleaning and preprocessing steps were performed and are described in subsequent sections. This dataset provides a structured foundation for developing predictive models for cardiovascular disease risk.

B. Exploratory Data Analysis

The features with the strongest positive correlation with the Result are the cardiac biomarkers, meaning that as the values of these features increase, the probability or severity of the Result also tends to increase. Troponin (r = 0.229): This is the second strongest positive correlation. Troponin is a protein released into the blood when the heart muscle is damaged. A positive correlation of 0.229 suggests it is the single best predictor among these features for the given

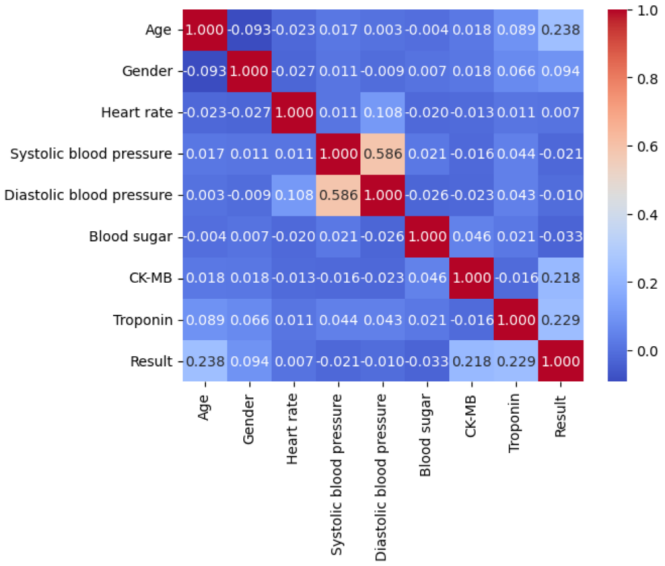

Fig. 1. Correlation Matrix.

Result. CK-MB (r = 0.218): This is the third strongest positive correlation. Creatine Kinase Myocardial Band (CK-MB) is another enzyme marker for heart muscle injury. Its correlation of 0.218 is very close to Troponin's, confirming its predictive value for the Result. Age (r=0.238): This is the strongest positive correlation. Age plays a huge role in heart health and efficiency. A positive correlation of 0.238 indicates that it is one of the best predictors for heart diseases. The mean of the features of the dataset is as follows: Mean Age: 56.192 Mean Heart Rate: 78.337 Mean Systolic Blood Pressure: 127.171 Mean Diastolic Blood Pressure: 72.269 Mean Blood Sugar Levels: 146.634 Mean CK-MB: 15.274 Mean Troponin: 0.361 With 449 females and 870 males, the dataset roughly has 34% females and 66% males. There were 0 NaN values in the dataset.

C. Preprocessing and Model Training

After preprocessing, the dataset is split into training and testing sets. Feature scaling will be applied to numerical attributes when required by the model's architecture. The split will be as follows: the training set is used to fit model parameters, and the test set is a holdout set used strictly for final evaluation. A fixed random seed (seed = 42) was used to ensure reproducibility. An 80/20 split is used for training and testing, respectively. The following four models are trained and tested: XGBoost, TabPFN, Neural Network (NN), and Random Forest Classifier. We used K-fold cross-validation on the training set to tune the hyperparameters for our models. After tuning and selecting the best hyperparameters, each model is retrained on the same training set. We used bootstrapping (B = 200) to only analyze the stability and variance of XGBoost and Random Forest models. This will allow us to estimate the variance and derive a 95% confidence interval for the two models mentioned, thus allowing the two models to be com-

pared for ensemble voting. Cross-validation → retraining with tuned hyperparameters → bootstrapping to assess stability of XGBoost and RandomForest → evaluation on test set. Our neural network will require us to tune hyperparameters such as the number of hidden layers, the number of neurons per layer, activation functions, and the learning rate. As mentioned in the preprocessing section, dataset standardization is required for this model. Cross-validation is performed (k=5 for NN, k=3 for Random Forest and XGBoost), and afterwards the model is trained on the training set. The pipeline is similar to XGBoost and RandomForest, omitting the bootstrapping. Cross-validation → Configuration of hyperparameters → Evaluation on test set. Purposed pipeline: preprocessing → 80/20 dataset split → k-fold cross-validation → retraining with tuned hyperparameters → bootstrapping (only XGBoost and RandomForest) → evaluation on test set.

IV. System Architecture

The preprocessed data and the training details discussed are applied to four models and an ensemble model. The four models chosen are XGBoost, TabPFN, Neural Network (NN), RandomForestClassifier. The ensemble model comprises all four of these machine learning models (Fig. 4). A voting mechanism with weights was assigned to each model, in order to get an overall result from the ensemble model. A threshold is applied to output a binary classification (1 = heart attack, 0 = no heart attack). Random Forest and XGBoost were chosen as tree-based models are known to be good with tabular data. TabPFN selected as it's a transformer model designed specifically for tabular data. NN is added to introduce a non-linear baseline. All models are combined into an ensemble soft-voting method. The performance of these 5 models is compared using accuracy, precision, recall, F1-score, AUROC, and AUPRC, under cross-validation.

1) Neural Network Hyperparameters And Architecture: Our non-linear baseline, the feed-forward neural network, on the eight features of the dataset. The architecture is as follows: three hidden layers with widths of 64, 32, and 16 neurons, respectively, with Batch Normalization, ReLu activation, and Dropout (p = 0.3) regularization, with a sigmoid activation in the output layer for binary classification. The hyperparameters were tuned using 5-fold stratified cross-validation. We use AUROC as the selection criterion and early stopping based on validation loss. The best configuration is retrained on the full training set and evaluated on the test set. The hyperparameter space of the NN is as follows: Hidden layer widths, dropout, learning rate, weight decay (for L2 regularization), and the number of epochs.

2) Random Forest Classifier Hyperparameters and Tuning: The selected hyperparameters were designed to balance model complexity, robustness and computational cost. The n_estimators values (300, 600) ensure a sufficiently large ensemble to stabilize predictions and reduce variance. The criterion option (gini and entropy) allow evaluation of different split quality measures. The max_depth values (3, 5, 7) constrain tree growth to limit overfitting while preserving the ability to model nonlinear relationships. The min_samples_split (2, 5, 10) and min_samples_leaf(1, 2, 4) parameters regularize tree structure by preventing overly specific splits. The max_features options (sqrt, 0.5) control feature subsampling, encouraging diversity among trees and improving generalization.

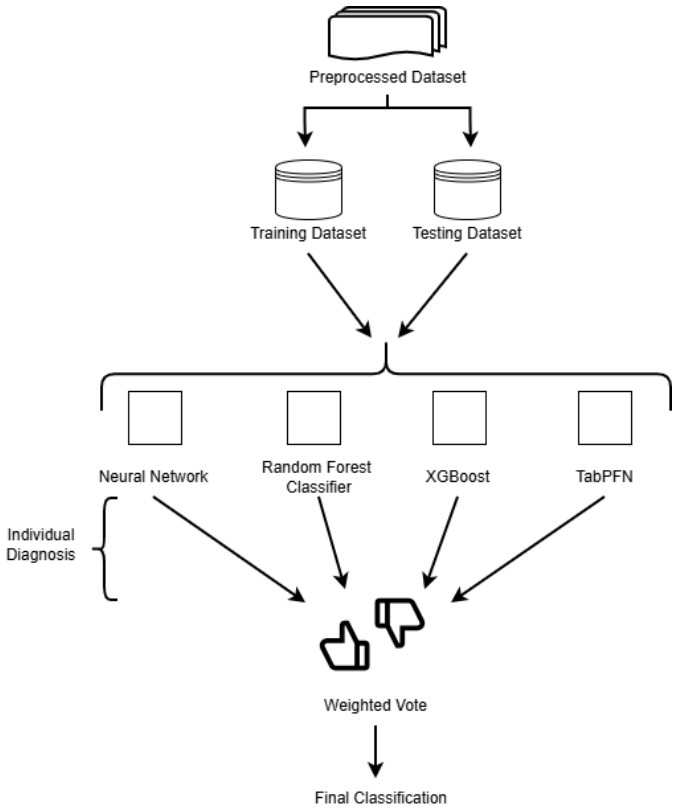

Fig. 2. Ensemble Model Architecture.

3) XGBoost Hyperparameters: The chosen hyperparameters were designed to balance model complexity, performance and computational efficiency. The n_estimators values (50, 100, 200) explore varying ensemble sizes to capture both simple and complex patterns. The max_depth values (3, 5, 7) control tree complexity to reduce overfitting while enabling meaningful feature interactions. The learning_rate values(0.01, 0.1, 0.3) allow a trade-off between stable learning and fast convergence. These ranges are widely used in practice and provide effective coverage of the model's key tuning parameters.

4) Random Forest and XGBoost Tuning: We tuned both models' hyperparameters using a 3-fold cross-validation grid search, evaluating multiple combinations and selecting the best configuration based on validation performance.

V. Results

A. Evaluation Metrics

Table 1 shows the values of these evaluation metrics for each model.

The scores on these metrics for the NN are lower than those for other models used. This is expected, as tree-based models

typically perform better on tabular data, whereas NNs require a deeper dataset than what was used and excel at mapping non-linear relationships. The tree-based models (XGBoost, Random Forest) and TabPFN excel at tubular data.

All models exhibit high levels of accuracy and precision, meaning that false positives and false negatives are rare. High recall scores ($\sim$ 88.9% to $\sim$ 98%) indicate that the models rarely miss any true positives. Strong F1-scores indicate that false positives and negatives are relatively low occurrences. All AUROC scores above 95% show that each of our models is excellent at distinguishing between patients with and without cardiovascular disease across all thresholds. AUPRC values are consistently very high ($\sim$ 98% to $\sim$ 99%). To complement F1-scores, we also used AUPRC. The primary difference between the two is that F1 is dependent on thresholds, whereas AUPRC isn't very dependent on thresholds. High AUPRC scores imply that our models correctly classify samples as the positive class without predicting many false positives.

B. Similar Results In Tree-Models And TabPFN

Our tree-based models and TabPFN have similar results because all models are trained and evaluated on the same preprocessed dataset, and the classification problem becomes high-signal. The features, particularly cardiac biomarkers, are strong indicators and correlate strongly with cardiovascular disease. For example, the dataset includes strong signals such as troponin, which is released into the bloodstream only when heart cells die. This will lead many models to achieve similar performance levels. The NN performs worse for reasons mentioned above.

C. Comparison of Models and Best Model

Comparison across the models is fair because all models are being trained and evaluated on the same preprocessed dataset and tested on the same proportions of the dataset. Each supervised-learning model was tuned with cross-validation, except TabPFN, as it doesn't require hyperparameter tuning. All models received the same data and evaluation criteria. The dataset has strong signals (bio-markers) like troponin, which is released into the bloodstream only when heart cells die, as seen in the correlation matrix between troponin and resulting heart attacks [7]. The tree-based models and TabPFN all achieve nearly identical results, which is a reflection of this. Furthermore, the bootstrap analysis of XGBoost and Random Forest shows lower variance in both models, making both of them more suitable for comparison to other models. In the individual model, TabPFN is marginally better than the tree-based models, with AUROC and AUPRC improvements of less than 1%. However, the ensemble soft-voting method remains the best one we have, with again, less than 1% improvement in AUROC and AUPRC compared to TabPFN. This level of improvement is likely not material in a clinical setting.

VI. Conclusion

This study explored various machine learning methods and models for predicting the presence of cardiovascular disease in patients. By reviewing existing research, certain limitations in previous approaches were identified, highlighting opportunities for improvement. The models evaluated in this study included Neural Networks, Random Forest, XGBoost, TabPFN, and an ensemble model combining all these approaches using a weighted average. Among these, XGBoost and TabPFN achieved the highest predictive performance, each attaining an accuracy of approximately 98%. But all in all, the ensemble model that combines all these models performs the best in terms of accuracy. The AUROC and AUPRC scores were less than 1% better than any other model employed to make predictions. These results suggest that, when applied

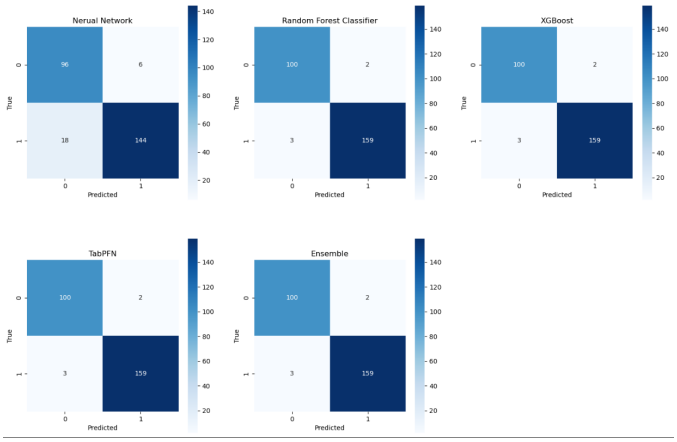

Fig. 3. Confusion Matrixes.

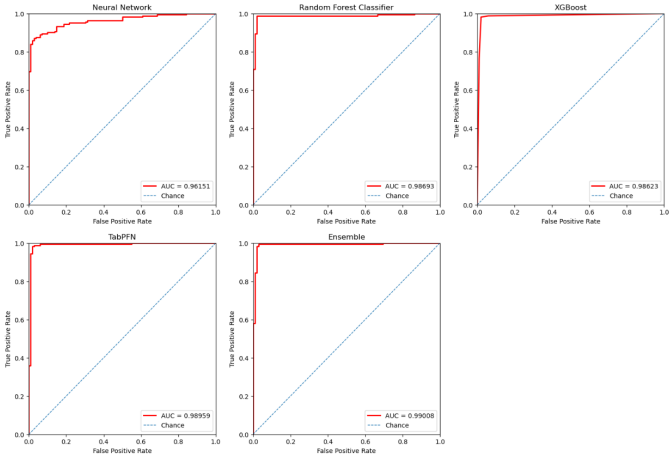

Fig. 4. ROC Curves

TABLE I
MODEL PERFORMANCE ON THE TEST SET.

Model	Accuracy	Precision	Recall	F1	AUROC	AUPRC
Neural Network	0.9091	0.9600	0.8889	0.9231	0.9615	0.9797
XGBoost	0.9811	0.9876	0.9815	0.9845	0.9862	0.9862
Random Forest	0.9811	0.9876	0.9815	0.9845	0.9869	0.9934
TabPFN	0.9811	0.9876	0.9815	0.9845	0.9895	0.9917
Ensemble	0.9811	0.9876	0.9815	0.9845	0.9899	0.9939

appropriately, these models can serve as early prevention and valuable decision-support tools for medical practitioners in diagnosing cardiovascular disease. A notable limitation is that the dataset originates from a specific region, which may lead to overfitting. Consequently, the models' performance may decrease slightly when applied to data from other populations. Future improvements could include training and validating the models on more diverse, multi-regional datasets to enhance their generalizability and robustness across different patient populations.

APPENDIX

Member 1: Oliver Olejar

- Contributed equally to the code with a focus on exploratory data analysis, random forest classifier, TabPFN, and ensemble implementation
- Contributed equally to the paper with a focus on the abstract, intro, related works, and system architecture.

Member 2: Jinil Desai

- Contributed equally to the code with a focus on XGBoost code implementation
- Contributed equally to the paper with focus on hyperparameter explanation, the abstract, conclusion, exploratory data analysis, and dataset selection

Member 3: Micky Huynh

- Contributed equally to the code with a focus on Neural Network, Stratified 5-fold CV, bootstrap analysis, and dropout regularization implementation
- Contributed equally to the paper with focus on system architecture, results, evaluation metrics, and neural network explanation.

REFERENCES

[1] World Heart Federation. (2023) World heart report 2023: Confronting the world's number one killer. [Online]. Available: https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf

[2] RBC Capital Markets. (2020) The healthcare data explosion. Healthcare Digitization series. [Online]. Available: https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion

[3] D. R. J, M. T. A, J. H, and J. C. J, "Enhancing cardiovascular disease diagnosis through bioinformatics and machine learning," Sep. 2024. [Online]. Available: http://dx.doi.org/10.21203/rs.3.rs-4716365/v1

[4] B. El-Din Waleed, E.-S. M. El-Kenawy, S. Ibrahim, H. El-Din Moustafa, and A. H. Rabie, "A new ensemble heart attack diagnosis (ehad) model using artificial intelligence techniques," *Scientific Reports*, vol. 15, no. 1, Sep. 2025. [Online]. Available: http://dx.doi.org/10.1038/s41598-025-18129-0

[5] B. N. Nava-Martinez, S. S. Hernandez-Hernandez, D. A. Rodriguez-Ramirez, J. L. Martinez-Rodriguez, A. B. Rios-Alvarado, A. Diaz-Manriquez, J. R. Martinez-Angulo, and T. Y. Guerrero-Melendez, "Heart attack risk prediction via stacked ensemble metamodeling: A machine learning framework for real-time clinical decision support," *Informatics*, vol. 12, no. 4, p. 110, Oct. 2025. [Online]. Available: http://dx.doi.org/10.3390/informatics12040110

[6] Bryar Hassan, "Heart attack dataset," 2022. [Online]. Available: https://data.mendeley.com/datasets/wmhctcrt5v/1

[7] M. Stark and S. Sharma. (2023) Troponin. Updated Apr. 23, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK507805/