# Interpretable Ensemble Learning for Cardiovascular Disease Prediction: A Large-Scale SHAP-Based Analysis

*Yifeng* Wang[1*]

[1]School of Computer Science and Technology, Tongji University, Shanghai 200092, China

**Abstract.** Cardiovascular disease is now the leading cause of death globally; however, there remains a significant barrier to the use of machine learning models to predict cardiovascular disease due to their lack of generalizability from smaller databases. The aim of this study was to develop an interpretable ensemble learning model based on a large database of over 70,000 patients. Specifically, this research created three advanced Gradient Boosting Models and enhanced these by Synthetic Minority Over-sampling Technique (SMOTE) for improving the data balance as well as Bayesian Hyperparameter Tuning for optimizing the models. The top performing Voting Ensemble had a test accuracy of 73.61% and an Area Under the Curve of 0.8022. SHAP results demonstrated that systolic blood pressure was the most important feature and that meaningful clinical thresholds existed at 120 mmHg for blood pressure and 48 years for age that aligned with established medical standards. Although the model's accuracy was somewhat lower than the benchmark of 78.42%, the increased interpretability of the model makes it a promising tool for clinical decision support. Therefore, this research demonstrates that by combining powerful ensemble methods with strong explanations for decisions, more trustworthy AI systems in the field of cardiovascular medicine can be created.

## 1 Introduction

The main challenge to be addressed in the field of CVD diagnosis is the high number of deaths due to CVDs and the need to develop accurate and reliable early diagnosis systems. It is well known that ML can be useful to discover hidden patterns in clinical data related to CVDs; however, in practice, the "small data" and "black box" problems represent significant obstacles. While some studies have developed several solutions to these challenges using ML approaches, such as feature analysis and prediction, there is still little knowledge about the ability of these approaches to provide satisfactory results in large and heterogeneous populations.

There are many recent studies on CVD prediction that have demonstrated the efficiency of various ML approaches and the improvement of results obtained when applying feature selection and gradient boosting to CVD prediction [1]. For example, Abdollahi and Nouri-

* Corresponding author's email: 2151929@tongji.edu.cn

Moghaddam combined a genetic algorithm-based feature selection with stacking ensembles to reach a 97.57% accuracy, while Khan et al. applied a modified Artificial Bee Colony algorithm with KNN to reach 98% accuracy [2, 3]. Gradient boosting models have also improved predictive power, with Zhang et al. demonstrating that XGBoost achieves an AUC of 0.988 for coronary artery disease prediction, and Rezk et al. reporting 96.5% accuracy with a hybrid XGBoost-LightGBM model [4, 5].

However, while strong performance is essential for clinical applications, it is not sufficient. Explainable AI has played a key role in enabling the application of ML in healthcare, particularly through the use of SHAP. Tarabanis et al. used SHAP to explain the XGBoost predictions for myocardial infarction mortality (AUC 0.922), and Luo et al. developed SHAP-based interpretable models to identify heart failure readmission risk factors (AUC 0.836) [6, 7]. Thus, SHAP enables black-box algorithms to be transformed into transparent, trustworthy, and clinically applicable diagnostic aids that enable clinicians to understand both the outcomes and the reasoning behind them.

Sreekumari et al. conducted a large-scale study with the same dataset of 70,000 subjects as the present study, and established a benchmark of 78.42% accuracy with basic ensemble classifiers [8]. Although this study did not investigate the potential of advanced gradient boosting models or a detailed interpretability analysis, this study aims to close this gap by developing an interpretable ensemble framework that combines XGBoost, LightGBM and CatBoost, and uses SMOTE and Bayesian hyperparameter tuning. Furthermore, SHAP was used to explain the model's predictions at both the population and individual patient levels to transform the ensemble into a transparent, clinically reliable diagnostic aid. By combining strong predictive accuracy with clear interpretability, this work contributes to the development of AI as a practical and trustworthy tool in cardiovascular medicine.

## 2 Methods

### 2.1 Dataset and Preprocessing

This study used the Kaggle Cardiovascular Disease Dataset, which comprises 70,000 patient records and 13 clinical features, thus providing a solid basis for building models that can generalize well in real-world scenarios [1]. Each record contains demographic information (age, gender), body measurements (height, weight), physiological data (systolic and diastolic blood pressure), and lifestyle factors like smoking, alcohol use, and physical activity. The target variable indicates whether a person has cardiovascular disease or not.

To ensure that the data is cleaned and realistic, values that do not correspond to physiological conditions were excluded. Systolic blood pressure was therefore limited to 80–220 mmHg, and diastolic to 50–150 mmHg, and all records in which the diastolic blood pressure was higher than the systolic blood pressure were deleted. Anthropometric outliers were also deleted by retaining only those data points between the 1st and 99th percentiles. Thus, after filtering out the unrealistic values, 66,256 samples (94.7%) remained, finding a good balance between data quality and sample size [8].

Four new derived features were created from the existing ones: age converted from days to years, body mass index (BMI), pulse pressure (systolic minus diastolic blood pressure), and blood pressure category (Normal, Elevated, Hypertension Stage 1/2 according to the medical guidelines). All categorical variables were one-hot encoded, resulting in a total of 18 features (from the original 14).

The data was almost perfectly balanced (50.5% disease vs. 49.5% healthy). To further promote the learning of models on minority cases, SMOTE was applied to the training set, producing a perfectly balanced distribution (50% vs. 50%), without duplicating samples [9].

### 2.2 Feature Selection

Good feature selection is an essential component of building accurate and stable models. As previous studies have shown, metaheuristic optimization techniques like genetic algorithms and swarm intelligence can significantly improve accuracy while reducing computational costs [2, 3]. In the current study, the Boruta algorithm—a wrapper-based feature selection technique—was used to evaluate the importance of each feature by comparing its values to randomly shuffled "shadow" values using a Random Forest classifier [10].

After 57 iterations, the Boruta algorithm confirmed 14 features as important: age_years, height, weight, ap_hi, ap_lo, active, BMI, pulse_pressure, cholesterol_2, cholesterol_3, gluc_3, and three blood pressure category indicators. The four features—smoking, alcohol consumption, gender, and glucose level 2—were rejected as noise. This result was in line with previous findings indicating that physiological factors tend to have stronger predictive values than lifestyle-related factors. All subsequent modeling used the 14 selected features.

### 2.3 Model Construction and Optimization

First, four baseline models were tested: Logistic Regression (as an interpretable benchmark), Random Forest (a bagging ensemble), XGBoost (a widely used gradient boosting framework) [11] and LightGBM (a fast, histogram-based version of gradient boosting) [4]. Given the superior performance of gradient boosting models in CVD prediction tasks, CatBoost, a new boosting algorithm capable of handling categorical data efficiently, was also tested [4].

Bayesian optimization with the Tree-structured Parzen Estimator (TPE) was used instead of the grid search method, which is slow and ineffective for large parameter spaces, to optimize the hyperparameters. For each of the three boosting models (CatBoost, XGBoost, LightGBM), 50 optimization trials were run on the SMOTE-balanced training data, using 5-fold cross-validation accuracy as the evaluation metric.

After having optimized each of the single models separately, a soft voting ensemble was constructed that combined the three best performing boosting models. The ensemble computes the average probabilities of the output of each model, weighted by the learned weights, thus exploiting model diversity to increase robustness [5].

### 2.4 Model Validation and Evaluation

The preprocessed dataset was divided into training (80%, 53,004 samples) and testing (20%, 13,252 samples) subsets using stratified sampling (randomstate=42). Only the training data were subjected to SMOTE, whereas the test data were left unchanged to avoid biased evaluation [9].

All of the final models were evaluated using 10-fold stratified cross-validation on the SMOTE-balanced training data, to obtain performance estimates with confidence intervals across folds.

For model performance evaluation, five standard metrics were used: Accuracy (global prediction correctness), Precision (reducing false positives), Recall (sensitivity to true disease cases), F1-Score (balancing precision and recall), and AUC (distinguishing between classes). An AUC of 0.80 or better is generally considered satisfactory for clinical use [6]. All metrics were expressed as mean ± SD over the 10-folds, and subsequently validated on the independent test set.

### 2.5 Interpretability Analysis

While the high accuracy of ensemble models is beneficial for clinical applications, the difficulty of interpreting them often restricts their applicability. Therefore, SHAP was applied to explain the predictions of the Voting Ensemble, as SHAP is a methodology that provides a fair allocation of the contribution of each feature to each prediction through the use of Shapley values, a concept from game theory [12]. Recent studies in CVD prediction have already demonstrated the feasibility of SHAP for making model behavior more understandable [4, 6]. Specifically, a KernelExplainer with 100 background training samples was employed to approximate the decision process of the Voting Ensemble, and subsequently computed SHAP values for 500 test cases. This produced a multi-layered interpretation, which included global feature importance rankings, beeswarm plots illustrating how feature values influence predictions, dependence plots to capture non-linear trends, and individual force plots that break down the prediction for each patient. Consequently, this detailed analysis transformed the ensemble from a black-box predictor into a transparent, clinician-friendly diagnostic tool [7].

## 3 Results

### 3.1 Model Performance Overview

Classification performance is summarized in Table 1 for all models examined. The individual model with the highest accuracy on the holdout set was LightGBM with 73.73%, followed closely by XGBoost (73.61%), CatBoost (73.56%), and Random Forest (73.39%). Logistic Regression, serving as the simpler linear baseline, was moderately accurate at 72.70%. The Voting Ensemble, using CatBoost, XGBoost, and LightGBM in a soft-voting ensemble, showed 73.61% accuracy and an AUC of 0.8022, the best in the analysis and the best discrimination capability for clinical decision making.

**Table 1.** Performance Comparison of Models on Test Set

| Model | Accuracy (%) | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 72.70 | 0.7554 | 0.6617 | 0.7054 | 0.7936 |
| Random Forest | 73.39 | 0.7618 | 0.6713 | 0.7137 | 0.7988 |
| CatBoost | 73.56 | 0.7480 | 0.7009 | 0.7237 | 0.8018 |
| XGBoost | 73.61 | 0.7521 | 0.6950 | 0.7224 | 0.8015 |
| LightGBM | 73.73 | 0.7527 | 0.6974 | 0.7240 | 0.8017 |
| Voting Ensemble | 73.61 | 0.7501 | 0.6986 | 0.7234 | 0.8022 |
| Sreekumari et al. [8] | 78.42 | - | - | - | - |

The narrow range of accuracy (72.70%–73.73%) among the models (only 1% or so) suggests that prediction is based to a large degree on a small quantity of significant features, which is further substantiated by the findings in Section 3.2, the SHAP values. All models demonstrate sufficiently adequate predictive strength, but their accuracy is slightly below the 78.42%, the benchmark presented by Sreekumari et al. [8]. This difference in performance could be due to differing data cleaning methods or, more possibly here, the intentions of the present study are directed toward interpretability rather than the maximization of performance alone.

LightGBM has a slightly better accuracy (0.12% increase, as it is just more accurate than the Voting Ensemble), however, it does not account for all the interpretability evaluation, rather the ensemble is used for all aspects of interpretability due to two main reasons. The Voting Ensemble first, obtained the best AUC (0.8022), which is vital in the medical background as it dictates models which can be discriminated across thresholds reliably [6]. More importantly, the Voting Ensemble combines in its prediction some average of the outputs of the three boosting models, obtaining all three differing methods of prediction, which improves the model structure diversity and results in better performance and validity over a field [5]. As this 0.12% increase in accuracy is not insubstantial to differentiate between the predictions, but rather falls within ±0.69% of the cross-validation tests, the two are statistically equivalent models. The small AUC advantage as well as the greater robustness justify the use of the Voting Ensemble as the main interpretability model.

### 3.2 Global Feature Importance and Interpretability

In order to analyze which areas contribute most to the predictions, the SHAP values were calculated on 500 test examples. Figure 1 shows the global ranking of features based on their mean absolute SHAP values. The strongest predictor was ap_hi (systolic blood pressure), which had a mean SHAP value of 0.1285. This is more than twice that of age (0.0570). The top five features were ap_hi (0.1285), age_years (0.0570), weight (0.0555), cholesterol_3 (0.0427) and BMI (0.0274). These results agree very well with previously known medical risk factors [4], evidencing the clinical validity of the model.

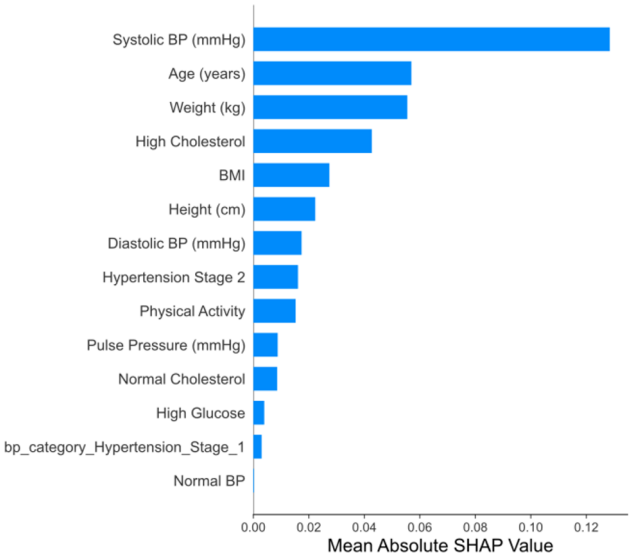


**Fig. 1.** Global feature importance ranking based on mean absolute SHAP values.

Figure 2 shows the SHAP beeswarm plot. This is a useful way to summarize both the magnitude of the effect of each feature, as well as the direction (positive or negative). The higher the values of ap_hi, age, and weight, the more positively the model tended to predict higher risk for CVD (positive SHAP values). This is in accord with medical knowledge and understanding [7, 13]. The lower the value of these features, the more the predictions were in the opposite direction, and lower risk was inferred. From the color-coded plot, it can be seen that high values of systolic pressure (red points) consistently increase the risk, whereas normal values (blue points) tend to be protectively effective. Thus, a clear and interpretable

pattern emerges making the model a more transparent tool which physicians sufficiently trust to render rationalisations of predictions on the basis of true clinical logic [6].

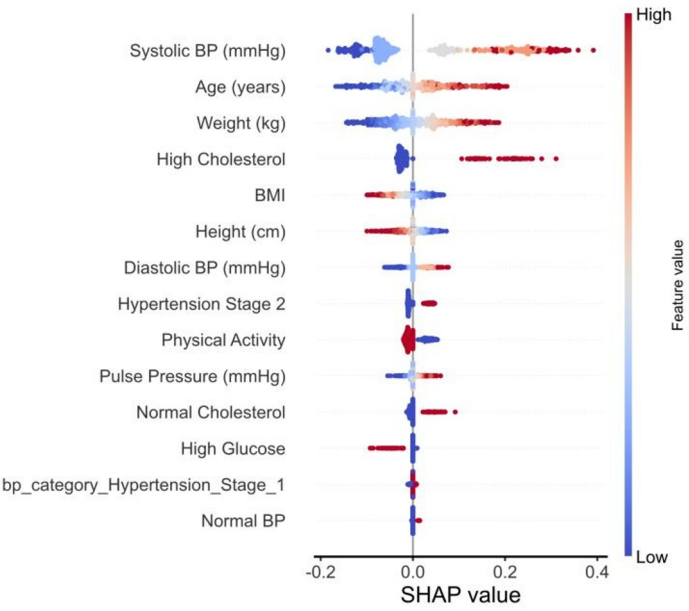


**Fig. 2.** SHAP summary plot showing feature value distributions and directional impacts on CVD risk predictions.

### 3.3 Feature Dependence Analysis

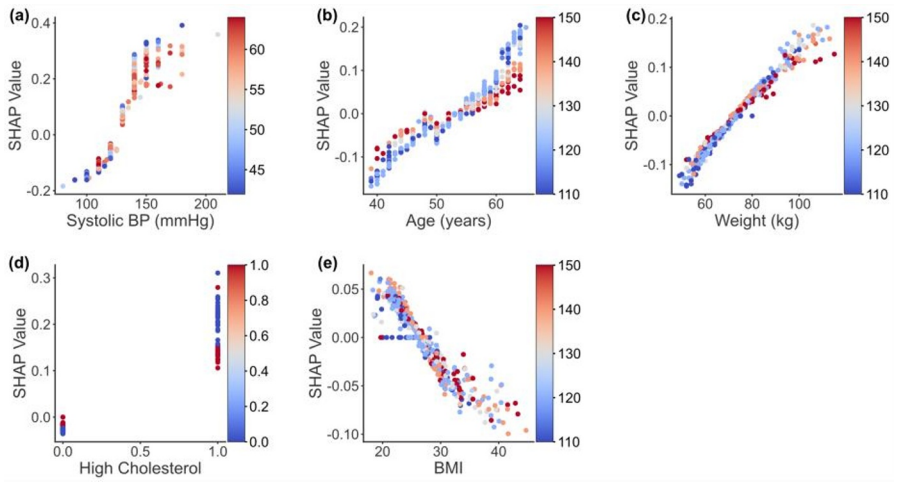


**Fig. 3.** SHAP dependence plots for the top five features: (a) Systolic BP, (b) Age, (c) Weight, (d) High Cholesterol, and (e) BMI.

The SHAP dependence plots of Figure 3 demonstrate a non-linear relationship between feature values and cardiovascular risk [6]. Important features include:

Systolic Blood Pressure (ap_hi): There is a clear "turning" point at about 120 mmHg, the clinical threshold of elevated blood pressure, where below it SHAP values are flat and low

indicating little additional risk. On exceeding this level, the contribution to risk rises steeply [4], being in perfect accord with the clinical definitions.

Age: There is a marked increase in risk after about 48 years reflecting the substantial increase in cardiovascular events in mid-life [7].

Weight and BMI: Risk shows a fairly uniform rise to well over about 74 kg in weight and 24.69 BMI, near the clinical "overweight" cut-off of 25 [4, 13].

High cholesterol: This feature demonstrates always positive SHAP values further demonstrating its established role as one of the large contributors to CVD risk.

It can be seen from these plots that it does not only make excellent predictions but also learns medically useful turning points, making it not only interpretable, but also clinically useful [6, 7].

### 3.4 Individual-Level Prediction Explanation

It can also be demonstrated how prediction for individual patients can be explained. SHAP force plots were constructed for typical patients to show illustrative cases (Figs. 4-6). These diagrams provide a breakdown of each prediction into additive feature contributions with respect to the global baseline risk of 45.99% [5].

High risk scenario (figure 4): this patient had a predicted CVD risk of 90.02%. The major contributory risk was high systolic pressure (160 mmHg, SHAP = +0.3104) which gave a great increase to this risk. High BMI (32.46) and elevated cholesterol also increased it, showing clearly how several risk factors combine to a higher risk score overall [6].

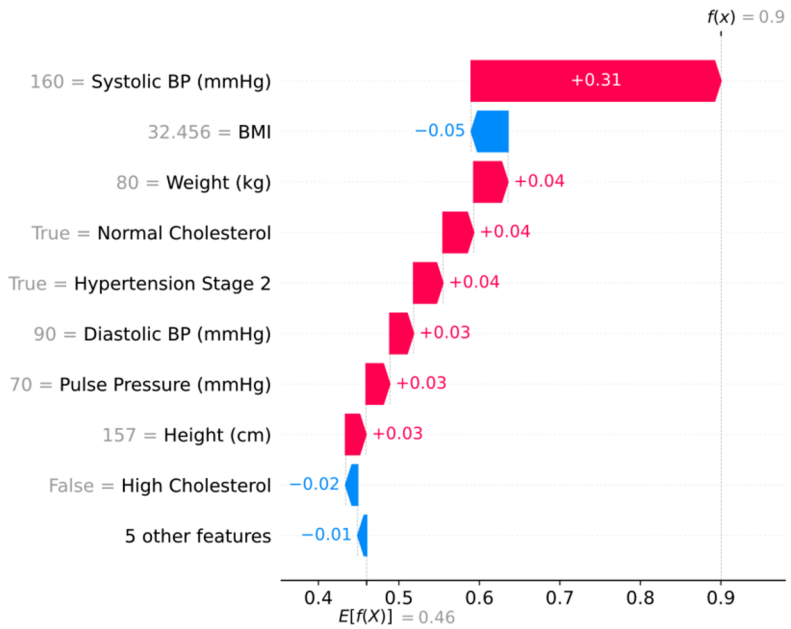


**Fig. 4.** SHAP force plot for a high-risk case (predicted probability: 90%).

Low risk scenario (figure 5): this patient had a predicted CVD risk of 18.94%. Normal systolic pressure (100 mmHg, SHAP = -0.1612) and lower weight (64 kg) acted as protective factors, cancelling out some of the mildly raised risk due to age (58 years). The model performs well in capturing this balance [7].

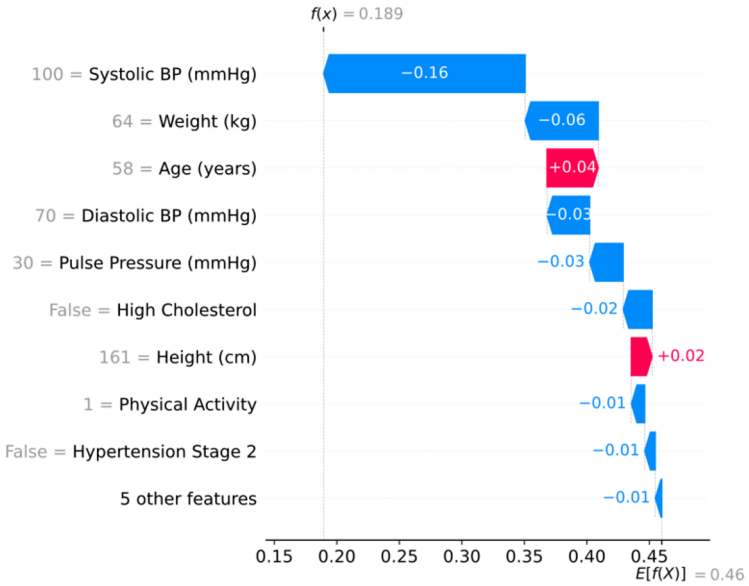


**Fig. 5.** SHAP force plot for a low-risk case (predicted probability: 19%).

Borderline case (figure 6): for a patient around the decision limit (49.90% risk), older age (62 years, SHAP = +0.1379) increased risk, whilst moderate systolic blood pressure (120 mmHg) and moderate BMI decreased risk slightly. This kind of mixed profile case illustrates how various determining factors combine in modifying ways, and this is usually what happens in real clinical assessment [5, 13].

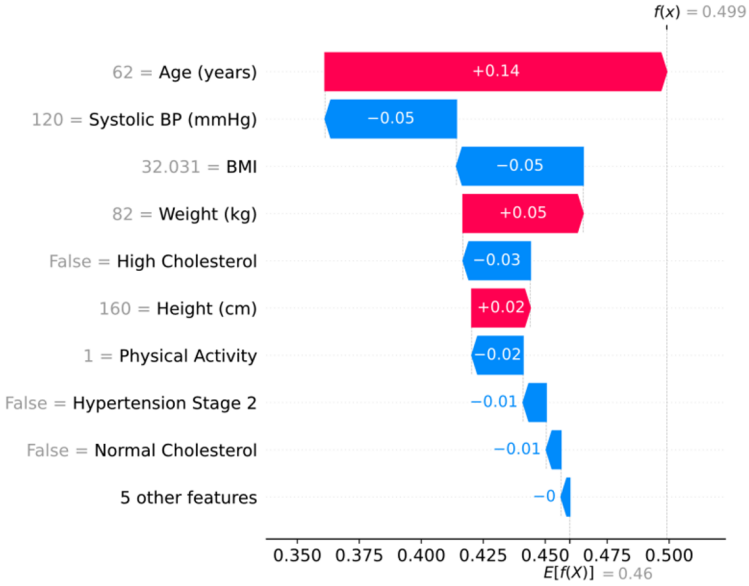


**Fig. 6.** SHAP force plot for a borderline case (predicted probability: 50%).

Overall the SHAP-based explanations confirm the known risk relationships, and provide detailed patient specific insights which are meaningful as far as clinicians are concerned, and are as readily interpretable [7]. Systolic blood pressure emerged as the most important risk component- more than twice as important as the next most important, and indicating its key

role in cardiovascular prevention. The accuracy of this method is slightly less than the above-mentioned methods, there is, however, a much greater degree of interpretability, which renders it far more suitable for real life clinical assessment, in which knowing the reasons for the predictions are equally as important as knowing the predictions themselves [5, 13].

## 4 Discussion

Although the study did not surpass the 78.42 % benchmark for Cardiovascular Disease prediction by Sreekumari et al. [8], the creation of an interpretable ensemble learning framework that optimizes Gradient Boosting Models through SMOTE and Bayesian Hyperparameter Tuning has created significant value for clinicians.

One major benefit of this study was the integration of SHAP, which helped in understanding how each feature contributes to predictions at both the group and individual level, thus addressing the "Black-Box" issue that limits clinician trust in many AI systems [6, 12].

The importance of systolic blood pressure (ap_hi) as the top predictor, along with clinically relevant threshold values like 120mmHg for blood pressure and approximately 24.7 for BMI confirmed that the model was able to learn patterns that align with current medical knowledge [4, 7].

These findings support the potential for this framework to be used as a clinically relevant decision support tool.

There are some limitations to consider when discussing this study. First, this model was developed and tested on one large dataset. Therefore, the applicability of the model to other demographics or clinical characteristics may be reduced [8].

While the small difference in accuracy between models (72.70%-73.73%) indicates that the model may have reached its peak performance based on the features and data included, the Boruta Algorithm chose a very strong set of predictors. The majority of the predictive power is attributed to a few well-established risk factors, therefore the predictive power may be similar in even simpler models.

Second, the incorporation of additional data modalities (such as Medical Imaging, Genomic Data, Longitudinal Follow-Up Data) that are currently unavailable, may be necessary to reach full predictive capacity of this framework.

Finally, although SHAP is highly effective, the computational requirements for the method remain a practical limitation for real-time deployment in a clinical setting [5].

In the future, researchers should evaluate this framework on multiple external datasets representing different populations, incorporate multimodal medical data, and assess the framework's feasibility in real clinical workflows to determine whether the framework will support clinicians' decisions. Additionally, the development of more computationally efficient interpretability methods would aid in making these types of systems more applicable for daily use in a clinical environment.

Ultimately, this work demonstrates that by combining state-of-the-art ensemble learning with Explainable AI, it is possible to develop cardiovascular disease risk prediction frameworks that are both transparent and clinically relevant. The remaining challenge is achieving a balance between accuracy and interpretability in order to ensure safe and effective adoption of AI into medicine.

## 5 Conclusion

This paper demonstrated the utility of an Interpretable Ensemble Framework for Predicting Cardiovascular Disease using data from 70,000 patients. By combining CatBoost, XGBoost,

and LightGBM and optimizing them through SMOTE and Bayesian Tuning, the final Voting Ensemble had a test accuracy of 73.61% and an Area Under Curve (AUC) of 0.8022. While the performance of the framework was slightly less than the 78.42% benchmark reported by Sreekumari et al., the framework stood out due to its focus on Interpretable AI, which is a necessity for medical applications in the clinic.

SHAP Analysis demonstrated that systolic blood pressure was the most important risk factor, contributing more than double the amount of any other risk factor. The model also produced clinically relevant thresholds for blood pressure (120 mmHg), Age (48), and Body Mass Index (BMI) (24.69).

The primary limitations of this study are the reliance on a single dataset and what appears to be a performance ceiling resulting from feature constraints. Future studies should focus on validating the model across multiple diverse populations, adding multimodal data to the model, and testing the model prospectively in clinical settings. However, this study demonstrates that by combining advanced ensemble Techniques with Strong Interpretability, it is possible to move AI-assisted cardiovascular medicine closer to being both trustworthy and useful; where the reason a prediction is made is as important as how accurate the prediction is.

## References

1. H. El-Sofany, B. Bouallegue, Y. M. A. El-Latif, A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Sci. Rep. 14, 23277 (2024). https://doi.org/10.1038/s41598-024-74656-2
2. J. Abdollahi, B. Nouri-Moghaddam, A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation. Iran J. Comput. Sci. 5, 229–246 (2022). https://doi.org/10.1007/s42044-022-00104-x
3. M. A. Khan, T. Mazhar, M. M. Yaqoob, M. B. Khan, A. K. J. Saudagar, Y. Y. Ghadi, U. F. Khattak, M. Shahid, Optimal feature selection for heart disease prediction using modified Artificial Bee colony (M-ABC) and K-nearest neighbors (KNN). Sci. Rep. 14, 26241 (2024). https://doi.org/10.1038/s41598-024-78021-1
4. Z. Zhang, B. Shao, H. Liu, B. Huang, X. Gao, J. Qiu, C. Wang, Construction and Validation of a Predictive Model for Coronary Artery Disease Using Extreme Gradient Boosting. J. Inflamm. Res. 17, 4163–4174 (2024). https://doi.org/10.2147/JIR.S464489
5. N. G. Rezk, S. Alshathri, A. Sayed, E. El-Din Hemdan, H. El-Behery, XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach. Bioengineering 11, 1016 (2024). https://doi.org/10.3390/bioengineering11101016
6. C. Tarabanis, E. Kalampokis, M. Khalil, C. L. Alviar, L. A. Chinitz, L. Jankelson, Explainable SHAP-XGBoost models for in-hospital mortality after myocardial infarction. Cardiovasc. Digit. Health J. 4, 126–132 (2023). https://doi.org/10.1016/j.cvdhj.2023.06.001
7. H. Luo, C. Xiang, L. Zeng, S. Li, X. Mei, L. Xiong, Y. Liu, C. Wen, Y. Cui, L. Du, Y. Zhou, K. Wang, L. Li, Z. Liu, Q. Wu, J. Pu, R. Yue, SHAP based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: Feature selection and model interpretation. Sci. Rep. 14, 17728 (2024). https://doi.org/10.1038/s41598-024-67844-7
8. S. Sreekumari, R. Bhalla, G. Singh, Feature selection and model evaluation for heart disease prediction using ensemble methods. Procedia Comput. Sci. 259, 1282–1295 (2025). https://doi.org/10.1016/j.procs.2025.04.083

9. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321-357 (2002). https://doi.org/10.1613/jair.953
10. M. B. Kursa, W. R. Rudnicki, Feature selection with the Boruta package. J. Stat. Softw. 36, 1-13 (2010). https://doi.org/10.18637/jss.v036.i11
11. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 785-794 (2016). https://doi.org/10.1145/2939672.2939785
12. S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in Advances in neural information processing systems 30, 4765-4774 (2017).
13. Q. Fu, Y. Wu, M. Zhu, Y. Xia, Q. Yu, Z. Liu, X. Ma, R. Yang, Identifying cardiovascular disease risk in the U.S. population using environmental volatile organic compounds exposure: A machine learning predictive model based on the SHAP methodology. Ecotoxicol. Environ. Saf. 286, 117210 (2024). https://doi.org/10.1016/j.ecoenv.2024.117210