JOURNAL OF
EMERGING INVESTIGATORS

Article

Cardiovascular Disease Prediction Using Supervised Ensemble Machine Learning and Shapley Values

Aahan Shah[1], Munib Mesinovic[2]
[1]Milpitas High School, Milpitas, California
[2]Department of Engineering Science, University of Oxford, Oxford, UK

SUMMARY

Cardiovascular disease (CVD) is the leading cause of death globally. The lack of awareness of coronary heart disease (CHD), a type of CVD, symptoms can potentially increase the vulnerability to experiencing a heart attack or cardiac arrest, making the early diagnosis and treatment of CHD imperative. The predictive modeling of clinical data has seen exponential growth over the past decade. Enhancing the traditional prognosis capacity with predictive modeling presents a lucrative and viable approach for doctors to predict the risk of CVD. This research is focused on evaluating multiple machine learning and deep learning algorithms to predict the onset of CVD. We hypothesized that supervised machine learning models with feature interpretability and ensemble learning could be deployed using clinical diagnosis data for reasonably accurate cardiovascular disease prediction. We observed that the smaller CVD dataset had a class imbalance problem, which was minimized by employing the adaptive synthetic (ADASYN) sampling technique to improve model performance. This study demonstrated that boosting algorithms can efficiently be deployed on small or large clinical datasets to predict diseases more accurately. The results indicated that while deep learning performs better on larger unstructured datasets, it is less efficient on tabular data, and ensemble boosting models outperformed other supervised machine learning and deep learning models, with 74% prediction accuracy. Shapley values were utilized to identify the risk factors that contributed most to the classification decision with XGBoost, demonstrating the high impact of systolic blood pressure and age on CVD, which aligned with findings in the field of clinical research.

INTRODUCTION

According to the World Health Organization, cardiovascular diseases (CVDs) cost an estimated 17.9 million lives yearly and are the leading cause of death globally (1). Coronary heart disease (CHD), a specific type of CVD, belongs to a family of heart problems, such as angina, a kind of chest pain or discomfort, heart attack caused by blockage of blood flow to the heart, and cardiac arrest, a condition when the heart stops pumping blood. CHD's underlying cause is coronary artery disease (CAD), a condition that develops when the coronary arteries cannot deliver oxygen-rich blood to the heart due to blockage created by the build-up of a fatty substance called plaque. The main risk factors of coronary artery disease are diabetes, high blood pressure, high cholesterol, obesity, smoking, age, and family history (2). According to the Centers for Disease Control and Prevention, about 1 in 20 adults aged 20 and older have CAD, making it the most common type of heart disease in the United States (3). Early prediction of the onset of CHD enables preventative measures such as a healthy diet, quitting smoking, managing stress, and appropriate medication to significantly prevent disease progression and reduce fatalities.

Machine learning (ML) algorithms have been proven to increase the proactive identification of patients at risk of cardiovascular risks (4). ML algorithms outperform conventional prognosis and standard statistical modeling techniques by analyzing complex and heterogeneous data from various sources, such as electronic health records, lab tests, medical images, and genomic sequences (5). ML algorithms use this data to capture the nonlinear relationships and interactions between the risk factors to improve cardiovascular risk prediction (5). Different approaches used by machine learning algorithms include supervised learning, defined as when the algorithm learns from labeled; well-defined structured data that has a known outcome; unsupervised learning, defined as when the algorithm learns from unlabeled data without the need for human interference; semi-supervised learning, defined as when the algorithm learns from a combination of labeled and unlabeled data; and reinforcement learning, defined as when the algorithm learns from its actions and feedback (6). Conventional supervised machine learning algorithms encounter challenges with medical datasets containing outliers and class imbalance. The class imbalance problem arises when the number of patients in each class is not equal or balanced, causing the model to be biased towards the majority class containing a higher number of patients, resulting in poor model learning performance. Research indicates a proportion of minority class of <20% is considered a moderate to high imbalance (7-8). This affects the model's learning ability and reduces its prediction accuracy. Ensemble learning, a technique that combines multiple machine learning algorithms, can achieve high performance and prediction accuracy by overcoming the limitations and errors of individual algorithms, such as high variance, low accuracy, or bias (9).

Previous research has highlighted the application of ML methods and their prediction accuracy comparison for patients with CHD. One such study compared multiple ML algorithms — decision tree, random forest, support vector machines, neural networks, and logistic regression — using R-Studio and RapidMiner software platforms for analyzing the model effectiveness. The authors concluded that ML

DOI: https://doi.org/10.59720/23-257

algorithms can enhance traditional techniques (10). Another study comparing the ML algorithms using data from a large cohort of 42,676 patients with hypertension concluded that the XGBoost ensemble method performed better than the logistic regression and k-nearest neighbor models in predicting a 3-year CHD (11). This inspired us to incorporate the XGBoost ensemble method in our research. Another paper focused on image-based data of CAD to construct the pooled area curve, a graphical method to summarize the accuracy of diagnostic tests, to account for variation in medical images to predict CAD (12). This approach demonstrates wide applications of machine learning in predicting cardiovascular disease and is another example that inspires us to pursue our study.

This research aims to explore the ability of various supervised machine learning models, including linear regression, decision tree, random forest, LightGBM, XGBoost, and deep learning model TabNet, to predict CVD. We built upon previous work by creating a robust supervised ensemble algorithm to achieve higher accuracy in disease prediction. Finally, Shapley values analyzed the most important features that impact cardiovascular risk prediction. We hypothesized that supervised machine learning models with feature interpretability and ensemble learning could be deployed on clinical diagnosis data for reasonably accurate cardiovascular disease prediction. Our results show that the ensemble boosting models outperformed other supervised machine learning and deep learning models in predicting CVD. Shapley values revealed that systolic blood pressure and age had the highest impact on CVD prediction.

RESULTS

This research aimed to predict cardiovascular disease based on the patients' risk data. We started with the first dataset of 3,390 patients' cardiovascular risk data to predict patients having a CHD in the next 10 years (13). Since the first dataset was relatively small and included the class imbalance problem resulting in poor model performance, we expanded the research to a second, more extensive cardiovascular risk data of 70,000 patients to predict CVD (14). On both datasets, we conducted exploratory data analysis (EDA), a method to analyze and investigate datasets and summarize their main characteristics. We deployed multiple machine learning and deep learning models and compared their accuracy in predicting cardiovascular disease. For tuning and improving model performance, we selected BayesSearchCV, a technique that uses Bayesian optimization to search for the optimal hyperparameters of the model.

10-year CHD prediction

The CHD data shows that only 15.1% (511 out of 3390) patients were classified as positive for the 10-year CHD, meaning the dataset is highly imbalanced. The proportion plots of the CHD dataset revealed that older patients have a higher risk of CHD, with the proportion of patients with CHD linearly increasing from 7% at age 40 to 41% at age 65 (**Figure 1**). Patients taking blood pressure (BP) medication have twice the risk of CHD than patients not taking BP medication (**Figure 1**). Patients with diabetes have two and a half times the risk of CHD than patients without diabetes (**Figure 1**).

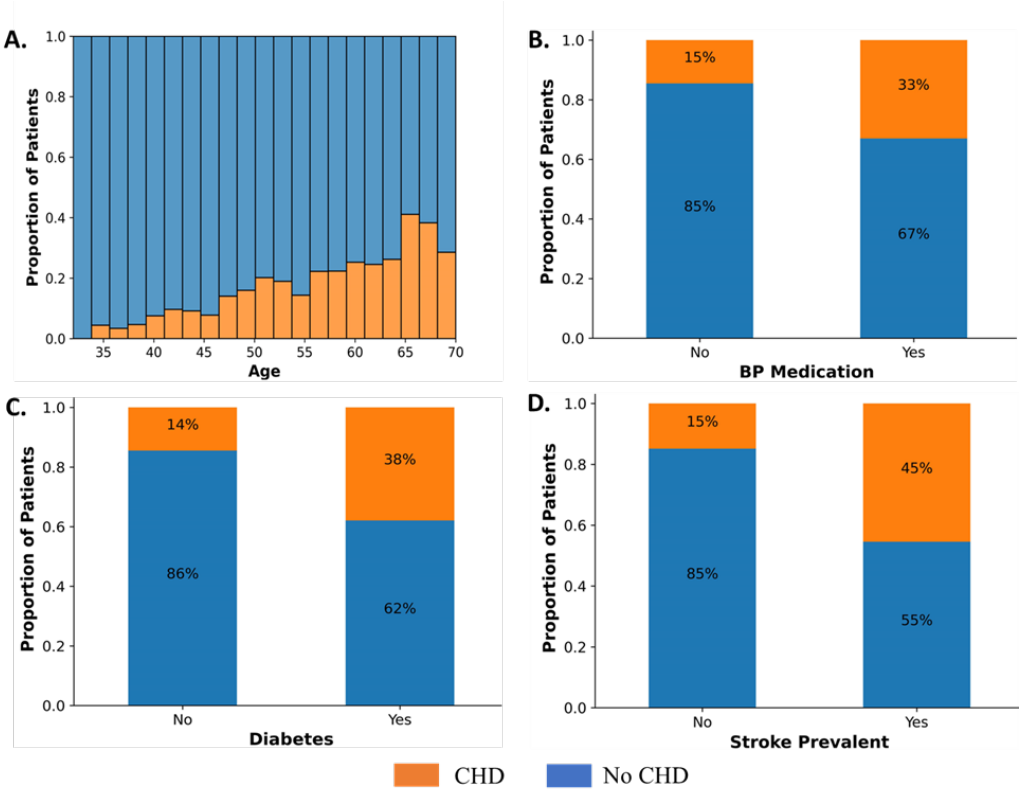

Figure 1: The proportion of patients with a CHD in the next ten years by risk factors. A) The proportion of patients with CHD increases linearly with age. B) Patients taking BP medication have twice the risk of CHD than patients not taking BP medication. C) Patients with diabetes have two and a half times the risk of CHD than patients without diabetes. D) Patients prone to stroke have three times the risk of CHD than patients not prone to stroke. Patient sample size, n=3390.

DOI: https://doi.org/10.59720/23-257

Model	Sampling Technique	Accuracy	Precision	Recall	F1 Score
Logistic Regression	Non-Sampled	86%	75%	6%	11%
Logistic Regression	SMOTE	85%	61%	39%	29%
Random Forest	Non-Sampled	85%	43%	6%	10%
LightGBM	Non-Sampled	84%	29%	6%	10%
XGBoost	Non-Sampled	83%	27%	9%	13%
Ensemble	Non-Sampled	83%	28%	7%	11%
Decision Tree	Non-Sampled	76%	23%	24%	23%

Table 1: 10-year CHD prediction model performance summary. The logistic regression model achieved higher Recall and F1 scores with the SMOTE sampling technique. All other models used non-sampled datasets. Patient sample size, n=3390.

Patients prone to stroke have three times the risk of CHD than patients not prone to stroke (**Figure 1**). The pre-trained models without any parameter tuning show a high prediction accuracy but a low F1 score, recall, and precision, indicating a class imbalance issue in the data (**Table 1**). A low F1 score indicates an unbalanced and poor model performance. The medical dataset contains multiple highly unbalanced categories. We applied the Synthetic Minority Oversampling Technique (SMOTE) sampling technique to reduce the class imbalance, slightly improving the logistic regression model performance with the F1 score of 29% and recall of 39% (**Table 1**). We applied an adaptive synthetic (ADASYN) sampling technique to improve the model performance. A precision-recall curve (PRC), which shows the precision values on the y-axis and recall values on the x-axis, is used to evaluate the model's performance, particularly with imbalanced datasets. An ADASYN optimal threshold of 0.9 is selected based on the highest Average Precision (AP), represented by the area under the PRC, and the exact sampling is performed before splitting the train and test datasets. The sampled data yielded higher precision, recall, and accuracy, indicating a higher ability of the model to predict the minority category correctly (**Table 2**). The deep learning TabNet model had the lowest prediction accuracy of 61% (**Table 2**). The XGBoost, LightGBM, and ensemble boosting models performed slightly better in prediction accuracy based on all measures than other models. The ensemble model performed the best with 88% accuracy, 92% precision, 85% recall, and 88% F1 score (**Table 2**).

CVD prediction

The CVD data is balanced with about 50% (34,979 out of 70,000) of patients classified as having a CVD. The proportion plots show older patients have a higher risk of CVD, with the proportion of patients with CVD linearly increasing from 21% at age 40 to 70% at age 65 (**Figure 2**). Patients with high cholesterol, glucose levels, and weight are susceptible to CVD (**Figure 2**). Models trained with optimally tuned hyperparameters did not significantly improve performance over most pre-trained models, except the decision tree model, with a 7% improvement in prediction accuracy (**Table 3**). We used TabNet, a deep learning model for tabular data, which is structured data that can be organized in rows and columns for CVD prediction. The optimum learning rate (LR) found by autoLR, which uses a learning rate finder implemented in the PyTorch Lightning library, was 0.063 for the TabNet model. TabNet model had 73% accuracy and performed slightly better than logistic regression, random forest, and decision tree models (**Table 3**). The XGBoost, LightGBM, and ensemble boosting models performed marginally better than other models in prediction accuracy based on all measures. The XGBoost and LightGBM combined ensemble boosting models performed best with 74% accuracy, 77% precision, 69% recall, 73% F1 score, and 0.74 AUC score, surpassing all other models in CVD prediction (**Figure 3**). We used SHAP (SHapley Additive exPlanations) values on the XGBoost model for feature interpretability. The beeswarm plot shows the impact of the features on the prediction (**Figure 4**). Systolic blood pressure, age, weight, cholesterol, and diastolic blood pressure had the highest impact, while physical activity, alcohol intake, and gender had a lower impact on the prediction (**Figure 4**).

DISCUSSION

Supervised ensemble machine learning and deep learning models were efficiently deployed on patients' health risk data to

Model	Sampling Technique	Accuracy	Precision	Recall	F1 Score
Ensemble	**ADASYN**	**88%**	**92%**	**85%**	**88%**
XGBoost	ADASYN	88%	91%	84%	87%
LightGBM	ADASYN	88%	91%	85%	88%
Random Forest	ADASYN	87%	90%	85%	88%
Decision Tree	ADASYN	78%	79%	79%	79%
Logistic Regression	ADASYN	70%	71%	70%	71%
TabNet	ADASYN	61%	61%	61%	61%

Table 2: 10-year CHD prediction model performance summary. All models have the ADASYN sampling technique applied with threshold = 0.9. The Ensemble model outperformed all other models with the best Accuracy, Precision, Recall, and F1 Score results. Patient sample size, n=3390.

DOI: https://doi.org/10.59720/23-257

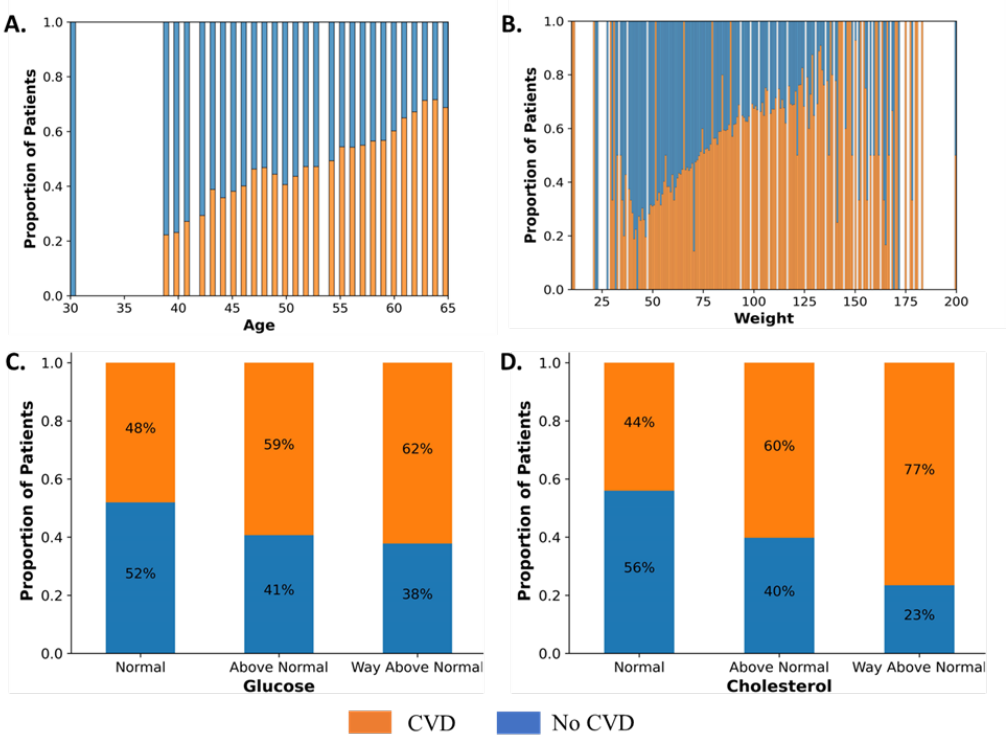

Figure 2: The proportion of patients with CVD by risk factors. A) The proportion of patients with CVD increases linearly with age. B) The proportion of patients with CVD increases linearly with weight. C) Patients with higher glucose levels are susceptible to CVD. D) Patients with higher cholesterol levels are susceptible to CVD. Patient sample size, n=70,000.

predict CVD and 10-year CHD, validating the initial hypothesis. Model predictions can enhance the detection of subtle and early signs of CVDs that may be overlooked by conventional clinical prognosis methods that are limited in accuracy and affordability, such as an electrocardiogram (ECG), which measures the heart's electrical activity. Different sampling techniques can be applied to small datasets with class imbalances to improve model performance and prediction accuracy. This study demonstrated that boosting algorithms can be deployed on small or large clinical datasets and are more efficient in predicting diseases with higher accuracy than conventional supervised machine learning and deep learning models. The ensemble gradient boosting performed better than the individual gradient boosting XGBoost and LightGBM models on the cardiovascular disease clinical data. Shapley values assessed the feature interpretability and its contribution to the prediction.

The cardiovascular risk data used for the 10-year CHD prediction was highly imbalanced, as only 15.1% of patients were classified as positive for the target class, 10-year CHD, compared to other classes present. This dataset class imbalance can be attributed to the relatively small sample of the medical diagnosis data collected from the study conducted on residents of just one town, Framingham, Massachusetts. The pre-trained model performance on this data showed high prediction accuracy but low F1 score, recall, and precision, as the models were highly effective in predicting only one class but ineffective in predicting other classes. SMOTE sampling technique was applied by under-sampling the majority class and over-sampling the minority class to reduce the class imbalance, but it did not improve the model performance much (15). The ADASYN sampling approach proved superior to SMOTE, as it applies the optimal weighted sampling distribution on minority classes according to the level of learning difficulty (16). ADASYN assisted in improving the model learning from the limited data available by generating synthetic samples, reducing the bias introduced by the class imbalance, and shifting the classification decision toward the difficult minority classes. Sampling techniques are essential

Model	Accuracy	Precision	Recall	F1 Score
Ensemble	**74%**	**77%**	**69%**	**73%**
Ensemble-Tuned	74%	76%	69%	73%
LightGBM	74%	76%	69%	73%
XGBoost-Tuned	74%	76%	69%	73%
LightGBM-Tuned	74%	76%	69%	72%
XGBoost	73%	76%	69%	72%
TabNet	73%	73%	73%	73%
Logistic Regression	72%	74%	68%	71%
Logistic Regression-Tuned	72%	74%	68%	71%
Decision Tree-Tuned	71%	71%	71%	71%
Random Forest	71%	72%	69%	70%
Random Forest-Tuned	71%	76%	63%	69%
Decision Tree	64%	64%	63%	64%

Table 3: CVD prediction model performance summary. Models not labeled as "Tuned" were pre-trained models with default parameters. "Tuned" labeled models were trained with different hyperparameters, including learning rates, number of epochs, batch size, regularization constant, etc., and optimal parameters selected using BayesSearchCV to improve the model performance. The Ensemble model outperformed all other models with the best Accuracy, Precision, Recall, and F1 Score results. Patient sample size, n=70,000.

DOI: https://doi.org/10.59720/23-257

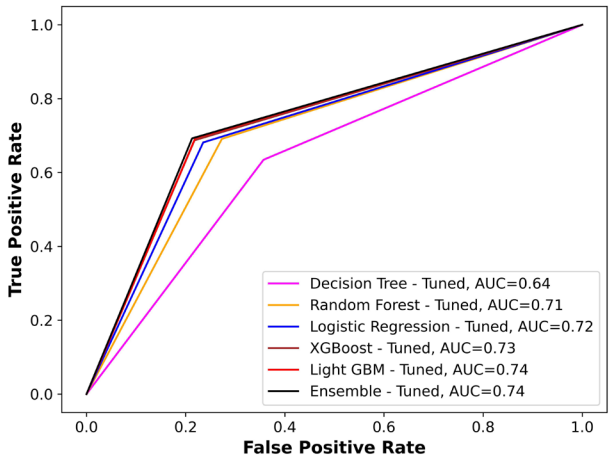

Figure 3: Model performance comparison using ROC plot. AUC is the area under the ROC curve, which quantifies the model's performance. The ROC curve closest to the top-left corner has the highest AUC value. A higher AUC represents superior model performance. The ensemble model performed best, surpassing all other models in CVD prediction.

in applications such as medical diagnosis, where generally small clinical data is available, and the risk tolerance for misdiagnosis is low.

The cardiovascular risk data used for the CVD prediction was balanced across all classes and did not require additional synthetic sampling to train the model. BayesSearchCV was selected for optimal hyperparameter estimation for tuned models for its higher efficiency and accuracy over other techniques, such as GridSearchCV, as it can handle complex and nonlinear relationships between hyperparameters. BayesSearchCV is also stated to outperform RandomSearchCV on various problems, including optimizing hyperparameters (17). Most tuned models did not significantly improve performance over the pre-trained models, which was unexpected. There are some limitations of this study. While the tuning was performed on all models, there is an opportunity to optimize the input range for hyperparameters to improve the grid search estimates and model performance. The other limitation is the quality of the cardiovascular risk data used for model training, which can significantly contribute to model prediction accuracy. The TabNet deep learning model performed better with the larger CVD dataset than with the smaller CHD dataset. TabNet model prediction accuracy didn't improve much by increasing the epoch runs as the model had achieved performance stability. TabNet has several hyperparameters that can be optimized further to achieve higher model performance. XGBoost and LightGBM, the highly efficient gradient-boosting models for tabular classification data, performed higher on all measures than logistic regression, decision tree, and random forest conventional supervised learning models. The superior performance of gradient-boosting models comes from their ability to automatically handle heterogeneous features, including nonlinearity and higher-order interactions (18). The combined XGBoost and LightGBM ensemble model achieved the best overall performance, surpassing the CVD prediction of all other models. The ensemble model was trained with "Soft" voting, which is more accurate than "Hard" voting, as it averages the probabilities and predicts the output class with the highest average probability. The ensemble model is more robust as it reduces the prediction error variance, improving the prediction accuracy and achieving better performance than any single model. The Shapley values feature ranking revealed systolic blood pressure as the top risk factor, followed by age contributing to cardiovascular disease. This aligns with the clinical research showing that cardiovascular disease risk increases with elevated systolic blood pressure levels and age (19, 20, 21, 22).

This study demonstrates the successful deployment of supervised ensemble machine learning and deep learning models to predict cardiovascular disease with significant accuracy that can augment the conventional clinical prognosis methods, such as ECG, for detecting the onset of cardiovascular disease. We addressed the class imbalance in the smaller dataset using the ADASYN sampling technique. Our findings showed that deep learning is not highly efficient on tabular data, and the ensemble gradient boosting models outperform other machine learning and deep learning models for cardiovascular disease prediction. The CHD feature importance using the Shapley values showed systolic blood pressure and age are the top risk factors for cardiovascular disease aligned with the clinical research. Future research may aim to explore improvement in prediction by further tuning the model hyperparameters and exploring other machine learning and deep learning techniques applied to diverse and more extensive datasets.

MATERIALS AND METHODS

Data Acquisition

This study used two cardiovascular risk datasets (**Table**

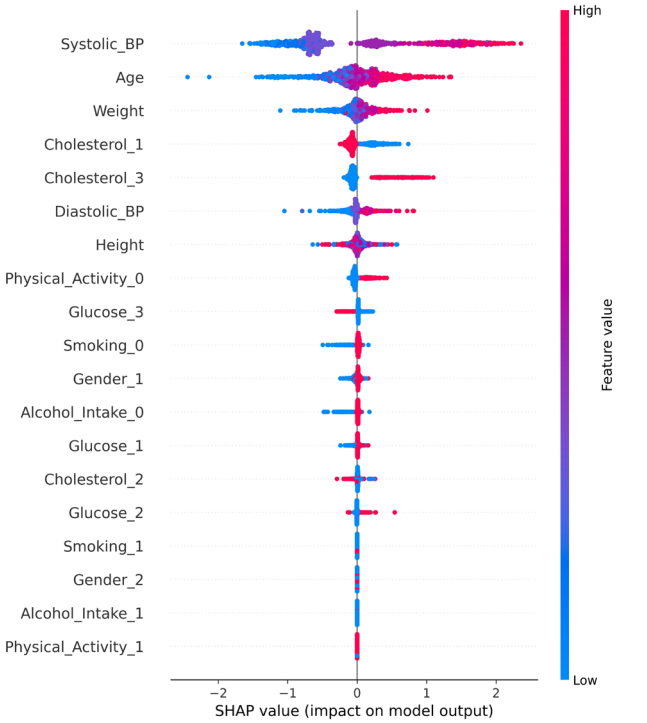

Figure 4: SHAP values for XGBoost model CVD prediction. Features in the CVD dataset are listed on the left in the order of feature importance from low importance at the bottom to high importance at the top.

DOI: https://doi.org/10.59720/23-257

Dataset	Category	Attribute
Dataset 1 (samples: 3,390)	Demographic	Sex, Age, Education
	Behavioral	Smoking, Cigarettes per day
	Medical (history)	BP Meds, Prevalent Stroke, Prevalent Hypertensive, Diabetes
	Medical (current)	Cholesterol, Systolic BP, Diastolic BP, Body Mass Index, Heart Rate, Glucose
	Prediction Target	10-year risk of coronary heart disease
Dataset 2 (samples: 70,000)	Demographic	Age, Gender, Height, Weight
	Behavioral	Smoking, Alcohol Intake
	Medical (current)	Cholesterol, Systolic BP, Diastolic BP, Glucose, Physical Activity
	Prediction Target	Cardiovascular disease

Table 4: Datasets for cardiovascular disease prediction. Dataset 1 includes 3,390 patient samples, including 15 features and 1 target. Dataset 2 includes 70,000 patient samples, including 11 features and 1 target.

4). The first dataset included 15 demographic, behavioral, and medical features representing a potential risk factor for CHD from 3,390 patients collected from an ongoing cardiovascular study on residents of Framingham, Massachusetts (13). The second dataset included 11 demographic, behavioral, and medical features representing a potential risk factor for cardiovascular disease from 70,000 patients collected during medical examination based on objective, tests, and subjective details (14).

Materials and Methods

The model development and coding were performed in Python version 3.10. Google Colab notebooks, which are Jupyter notebooks that run in the cloud and are integrated into Google Drive, were used for code development and execution. Models were implemented using pandas 1.5.3, numpy 1.25.2, scikit_learn1.2.2, and pytorch libraries and modules. The skopt library was used for optimization, shap 0.42.1 was used for Shapley values, and matplotlib was used for plotting and visualization. Data preprocessing, including filtering less relevant data such as identification numbers and scaling the data using the StandardScaler technique to standardize the features of the data, was performed to ensure a level playing field for different attributes that may have varying scales and distributions. EDA was performed to identify and keep the most relevant features, providing a strong correlation to the risk of cardiovascular disease, while the irrelevant features were abandoned. The cardiovascular risk patient datasets were split into 80% training and 20% test sets. A customized supervised learning approach that uses labeled datasets to train various algorithms that predict and classify outcomes effectively was applied to the pre-trained models. The CHD dataset was first resampled using SMOTE, followed by the ADASYN sampling techniques to address the class imbalance issue in the dataset. An optimal ADASYN threshold of 0.9 was selected, and the exact sampling was performed before splitting the train and test datasets. The identical training and testing datasets were utilized across the various models to ensure a fair comparison. Feature analysis was conducted to evaluate the relative importance of each feature to enable feature selection, achieving optimal prediction outcomes.

Models Used

Logistic Regression

Logistic Regression is one of the most important statistical techniques for analyzing and classifying binary and proportional response datasets. The model can naturally provide probabilities and extend to multi-class classification problems (23). It is an equation where each predictor is multiplied by a coefficient and summed together. This sum becomes the argument for the logistic function to predict the class. Tuned model training was performed using optimal hyperparameter values for “penalty” and “C” obtained using BayesSearchCV.

Decision Tree

Decision trees are non-parametric supervised learning methods. Every root node in a tree signifies a single input variable (x) and a split point on that variable. The output variable (y) is in the tree’s leaf nodes and is used to make a prediction (24). Predictions are made by traversing through the tree’s splits before reaching a leaf node and then outputting the class value at that node. Tuned model training was performed using optimal hyperparameter values for max_depth, max_features, min_samples_leaf, and criterion, obtained using BayesSearchCV.

Random Forest

Random forest is a classification ensemble learning method consisting of many decision trees. The random forest algorithm delivers a consolidated prediction result by combining the outputs of these trees. Tuned model training was performed using optimal hyperparameter values for max_depth, max_features, max_samples, min_samples_split, n_estimators, and min_impurity_decrease, obtained using BayesSearchCV.

XGBoost

XGBoost is an efficient gradient boosting framework that creates strong learners by iteratively adding new decision trees. XGBoost supports parallel and distributed computing, enabling higher performance. It has a flexible and expressive interface that provides custom metrics, model analysis tools for feature importance, and tree visualization. XGBoost is highly

DOI: https://doi.org/10.59720/23-257

scalable, enabling it to solve many data science problems quickly and accurately using far fewer resources than existing algorithms (25). Tuned model training was performed using optimal hyperparameter values for n_estimators, subsample, learning_rate, colsample_bytree, and colsample_bylevel, obtained using BayesSearchCV.

LightGBM

A gradient boosting model that utilizes tree-based learning algorithms provides higher training speed and efficiency with support for GPU learning, lower memory usage, better accuracy, and the capability to handle large-scale data. LightGBM is preferred if higher predictive accuracy is required for multi-class classification (26). Tuned model training was performed using optimal hyperparameter values for learning_rate, max_depth, min_child_samples, min_child_weight, subsample, colsample_bytree, and n_estimators, obtained using BayesSearchCV.

TabNet

TabNet is a deep learning model for tabular data that utilizes a sequential attention mechanism that softly selects features to reason from at each decision step and then aggregates the processed information to make a final prediction decision. TabNet learns very efficiently as the most relevant features are evaluated at each decision point, which enables more interpretable decision-making (27). TabNet model was trained using batch_size=1600, epochs=50, and learning rate = 0.001.

Ensemble

The ensemble model was trained using the VotingClassifier technique, which combines predictions from multiple machine learning models. "Soft" voting was chosen as the VotingClassifier parameter. XGBoost and LightGBM models were selected for ensemble modeling.

Shapley Values

Shapley values reveal the contribution of each feature to an individual prediction. Shapley values are applied to the testing dataset for interpretability and to gain deeper insight into the model prediction by identifying the features that contribute most to the prediction (28, 29). To calculate the Shapley value of a specific feature *i*, sets of all possible unions are formed with all *n* features except for the feature *i*. The value of the *i*-th feature is obtained by calculating the difference between the results of the characteristic function *v* on *N* (the set of all features) and *S* (the subset of *N* without feature *i*). The Shapley value of a particular feature *i* is then calculated by taking the average of the marginal contributions of all possible combinations of the feature unions. The following equation calculates the Shapley value φ for feature *i* (28, 29):

$$\varphi_i(v) = \sum_{S \subset N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left(v\left(S \bigcup \{x_i\}\right) - v(S)\right)$$

Measurement Metrics

Since accuracy only can be a misleading metric, particularly for imbalanced datasets, we evaluated the model performance using additional metrics, recall, precision, F1 Score, and AUC. Accuracy is the sum of true positive and true negative divided by the sum of true positive, true negative, false positive, and false negative. Recall is true positive divided by the sum of true positive and false negative. Precision is true positive divided by the sum of true positive and false positive. F1 score is twice the precision times recall divided by the sum of precision and recall. AUC is the area under the receiver operating characteristic (ROC) curve, which quantifies the model's performance. A higher AUC represents superior model performance.

Received: October 1, 2023
Accepted: March 18, 2024
Published: August 6, 2024

REFERENCES

1. "Cardiovascular Diseases." *World Health Organization*. www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. Accessed 8 July 2023.
2. "Coronary Artery Disease - Coronary Heart Disease." *American Heart Association*. www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/coronary-artery-disease. Accessed 19 February 2024.
3. "Heart Disease Facts." *Centers for Disease Control and Prevention*, 15 May 2023, www.cdc.gov/heartdisease/facts.htm.
4. Weng, Stephen F., et al. "Can machine-learning improve cardiovascular risk prediction using routine clinical data?." *PloS one* 12.4 (2017): e0174944.https://doi.org/10.1371/journal.pone.0174944.
5. Ahmed, Zeeshan, et al. "Artificial Intelligence with Multi-Functional Machine Learning Platform Development for Better Healthcare and Precision Medicine." *Database*, vol. 2020, 1 Jan. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7078068/, https://doi.org/10.1093/database/baaa010
6. Sarker, Iqbal H. "Machine learning: Algorithms, real-world applications and research directions." *SN computer science* 2.3 (2021): 160. https://doi.org/10.1007/s42979-021-00592-x.
7. Krawczyk, Bartosz. "Learning from Imbalanced Data: Open Challenges and Future Directions." *Progress in Artificial Intelligence*, vol. 5, no. 4, 22 Apr. 2016, pp. 221–232, https://doi.org/10.1007/s13748-016-0094-0.
8. "Imbalanced Data." *Google Developers*, developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data.
9. Mienye, Ibomoiye Domor, and Yanxia Sun. "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects." *IEEE Access*, vol. 10, 2022, pp. 99129–99149, ieeexplore.ieee.org/abstract/document/9893798. https://doi.org/10.1109/ACCESS.2022.3207287.
10. Beunza, Juan-Jose, et al. "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)." *Journal of biomedical informatics* 97 (2019). https://doi.org/10.1016/j.jbi.2019.103257.
11. Du, Zhenzhen, et al. "Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: model development and performance evaluation." *JMIR medical informatics* 8.7 (2020). https://doi.org/10.2196/17257.

DOI: https://doi.org/10.59720/23-257

12. Zong Chen, Joy Iong, and Hengjinda P. "Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method - a Comparative Study." *March 2021*, vol. 3, no. 1, 16 Mar. 2021, pp. 17–33, https://doi.org/10.36548/jaicn.2021.1.002.
13. Sharma, Mamta. "Cardiovascular risk factor data." www.kaggle.com/datasets/mamta1999/cardiovascular-risk-data. Accessed 5 August 2023.
14. Ulianova, Svetlana. "Cardiovascular disease dataset." www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. Accessed 31 August 2023.
15. Blagus, Rok, and Lara Lusa. "SMOTE for high-dimensional class-imbalanced data." *BMC bioinformatics* 14 (2013): 1-16. https://doi.org/10.1186/1471-2105-14-106.
16. He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee, 2008. https://doi.org/10.1109/IJCNN.2008.4633969.
17. Turner, Ryan, et al. *Bayesian Optimization Is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020*. 20 Apr. 2021, pp. 3–26.
18. Zhang, Zhongheng, et al. "Predictive Analytics with Gradient Boosting in Clinical Medicine." *Annals of Translational Medicine*, vol. 7, no. 7, Apr. 2019, pp. 152–152, https://doi.org/10.21037/atm.2019.03.29.
19. Whelton, Seamus P., et al. "Association of normal systolic blood pressure level with cardiovascular disease in the absence of risk factors." *JAMA cardiology* 5.9 (2020): 1011-1018. https://doi.org/10.1001/jamacardio.2020.1731.
20. Kannel, William B. "Elevated systolic blood pressure as a cardiovascular risk factor." *The American journal of cardiology* 85.2 (2000): 251-255. https://doi.org/10.1016/S0002-9149(99)00635-9.
21. Benetos, Athanase, et al. "Influence of age, risk factors, and cardiovascular and renal disease on arterial stiffness: clinical applications." *American journal of hypertension* 15.12 (2002): 1101-1108. https://doi.org/10.1016/S0895-7061(02)03029-7.
22. Jousilahti, Pekka, et al. "Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14 786 middle-aged men and women in Finland." *Circulation* 99.9 (1999): 1165-1172. https://doi.org/10.1161/01.CIR.99.9.1165.
23. Maalouf, Maher. "Logistic Regression in Data Analysis: An Overview." *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, 2011, p. 281, https://doi.org/10.1504/IJDATS.2011.041335.
24. Motarwar, Pranav, et al. "Cognitive approach for heart disease prediction using machine learning." 2020 international conference on emerging trends in information technology and engineering (ic-ETITE). IEEE, 2020. https://doi.org/10.1109/ic-ETITE47903.2020.242.
25. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016. https://doi.org/10.1145/2939672.2939785.
26. Zhang, Dongyang, and Yicheng Gong. "The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure." *IEEE Access* 8 (2020): 220990-221003. https://doi.org/10.1109/ACCESS.2020.3042848.
27. Arik, Sercan Ö., and Tomas Pfister. "Tabnet: Attentive interpretable tabular learning." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. No. 8. 2021. https://doi.org/10.1609/aaai.v35i8.16826.
28. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).
29. Ibrahim, Lujain, et al. "Explainable prediction of acute myocardial infarction using machine learning and shapley values." *Ieee Access* 8 (2020): 210410-210417. https://doi.org/10.1109/ACCESS.2020.3040166.