International Journal of Applied Mathematics, Sciences, and Technology for National Defense
**Volume 4, Issue 1, 71-82**
e_ISSN: 2985-9352, p_ISSN: 2986-0776
https://www.journal.foundae.com/index.php/JAS-ND
DOI : https://doi.org/10.58524/app.sci.def.v4i1.1148

# Hybrid random forest–catboost ensemble for heart disease prediction on imbalanced datasets: Toward applications in military health systems

**Mahyus Ihsan**
Universitas Syiah Kuala
INDONESIA

**Zahnur**
Universitas Syiah Kuala
INDONESIA

**Iftahul Fadlan**
Universitas Syiah Kuala
INDONESIA

**Ikhsan Maulidi***
Universitas Syiah Kuala
INDONESIA

**Article Info**

***Article history:***

Received: February 3, 2026
Revised: April 2, 2026
Accepted: April 21, 2026

***Keywords:***

CatBoost
Ensemble Learning
Heart Disease Prediction
Machine Learning
Random Forest

**Abstract**

**Background:** Heart disease is one of the main causes of death worldwide, with cases increasing every year. This situation highlights the urgent need for early detection systems that are not only fast but also accurate and reliable. In recent years, machine learning has emerged as a promising alternative approach for analyzing medical data, particularly for disease classification and risk prediction tasks.
**Aims:** This study aims to develop a heart disease prediction model by integrating Random Forest and CatBoost in a hybrid ensemble framework and evaluating its performance on an imbalanced medical dataset.
**Method:** This study employs a quantitative approach based on supervised learning using the Behavioral Risk Factor Surveillance System (BRFSS) 2021 dataset, which consists of more than 300,000 observations. Data preprocessing includes duplicate removal, BMI categorization, encoding of categorical variables, and exploratory analysis. To address class imbalance, the Borderline-SMOTE technique was applied before splitting the dataset using an 80:20 train-test split. Random Forest and CatBoost models were trained and combined using a soft voting ensemble.
**Result:** The evaluation results indicate that Random Forest achieved the highest accuracy of 0.94, with well-balanced precision and recall across all classes. CatBoost demonstrated relatively stable performance with accuracy around 0.84. The ensemble approach achieved an accuracy of 0.91 with strong metric stability and good sensitivity to positive cases.
**Conclusion:** The results indicate that Random Forest performs best for the dataset used in this study, while the ensemble model provides a balanced compromise between predictive performance and robustness. The analysis also shows that Age Category, General Health, and BMI are the most influential predictors of heart disease risk. This model can support early cardiovascular risk detection in military personnel, contributing to maintaining operational readiness in defense systems. Furthermore, the proposed approach provides a reliable decision-support tool for large-scale medical screening in resource-constrained healthcare environments.

**To cite this article:** Ihsan, M., Zahnur, Fadlan, I., & Maulidi, I. (2026). Hybrid random forest–catboost ensemble for heart disease prediction on imbalanced datasets: Toward applications in military health systems. *International Journal of Applied Mathematics, Sciences, and Technology for National Defense,* 4(1), 71-82

***Corresponding Author:**
Ikhsan Maulidi, Universitas Syiah Kuala, Indonesia, Email: ikhsanmaulidi@usk.ac.id

## INTRODUCTION

Cardiovascular disease continues to rank among the foremost causes of death globally, with its overall impact steadily rising year by year. This situation highlights the urgent need for early detection systems that are not only fast but also accurate and reliable. In recent years, machine learning has gained recognition as a compelling alternative for the analysis of medical data, particularly for disease classification and risk prediction tasks (Ashri et al., 2021; Murphy, 2012). Various algorithms have been applied in this context, demonstrating that data-driven approaches can improve sensitivity and specificity compared to conventional statistical methods.

Recent developments in medical data analytics have demonstrated that machine learning techniques can significantly outperform traditional statistical models in cardiovascular risk prediction when applied to large-scale clinical datasets. Studies utilizing electronic health records and population-scale data, such as those conducted by Weng (2017) and Alaa et al. (2019) , show that non-linear models and ensemble approaches can capture complex interactions among risk factors more effectively than conventional regression-based methods. Furthermore, systematic reviews indicate that machine learning models consistently achieve higher predictive accuracy in cardiovascular disease detection across diverse populations (Johnson, 2016; Karna, 2025; Krittanawong, 2020). These findings highlight the growing importance of advanced predictive modeling in modern healthcare systems

Current advances in heart disease prediction are largely driven by decision tree–oriented methods and ensemble-based approaches. Random Forest is known for its robustness to data variability and its effectiveness in handling heterogeneous feature spaces, while boosting algorithms such as CatBoost offer advantages in processing categorical features and reducing prediction bias (Latha & Jeeva, 2019; Zhou, 2012). Several studies have reported improved classification performance using ensemble techniques compared to single-model approaches (Shorewala, 2021). However, most existing studies focus on evaluating algorithms individually or employ homogeneous ensembles without exploring combinations of algorithms with different learning paradigms.

In addition to classical ensemble methods, recent research has increasingly explored hybrid and deep learning-based approaches with the aim of enhancing predictive accuracy, including approaches based on gradient boosting frameworks (Chen & Guestrin, 2016; Ke, 2017), deep neural networks (Miotto, 2016), and hybrid ensemble architectures (Rajkomar, 2018) have shown promising results in modeling complex, high-dimensional healthcare data. Moreover, studies have emphasized that combining different learning paradigms, such as bagging and boosting, can enhance model generalization and robustness, particularly in heterogeneous and noisy datasets (Dong, 2020; Sagi & Rokach, 2018). Despite these advances, the integration of heterogeneous ensemble models specifically tailored for imbalanced cardiovascular datasets remains underexplored.

Recent advances in heart disease prediction have also explored the use of more sophisticated machine learning and hybrid models. For instance, Detrano et al. and subsequent studies have shown that combining clinical attributes with machine learning models significantly improves diagnostic accuracy in cardiovascular risk assessment. More recent work by Alaa et al. (2019) introduced personalized risk prediction using machine learning, demonstrating improved predictive performance over traditional risk scoring systems. In addition, studies such as (Weng, 2017) have compared multiple machine learning algorithms against conventional clinical models and found that ensemble-based and non-linear approaches outperform traditional statistical methods in large-scale electronic health records. Furthermore, recent research has highlighted the growing role of hybrid and ensemble frameworks in handling complex and imbalanced medical data, suggesting that integrating different learning paradigms can enhance both predictive performance and model robustness.

A further challenge frequently encountered in medical datasets is class imbalance, in which the number of individuals without heart disease greatly outnumbers those diagnosed with the condition. This imbalance often leads predictive models to be biased toward the majority class, thereby reducing their ability to accurately detect positive cases (He & Garcia, 2009). Numerous studies have sought to mitigate this problem by employing oversampling methods, including SMOTE and its variants (Chawla et al., 2002; Han et al., 2005).

Extending these initial efforts, more recent research has broadened strategies for managing class imbalance by introducing advancements at both the data and algorithmic levels. Variants such as ADASYN adaptively generate synthetic samples in regions that are harder to learn, while more advanced techniques incorporate density-aware or cluster-based oversampling to better preserve the underlying data distribution (Douzas & Bacao, 2018; He et al., 2008). In parallel, hybrid approaches combining resampling with ensemble learning, such as SMOTEBoost and Balanced Random Forest—have demonstrated improved performance in imbalanced classification tasks by simultaneously addressing data distribution and model bias (Chawla et al., 2003). More recently, deep learning–based and generative methods, including GAN-based oversampling, have been explored to produce more realistic synthetic samples, particularly in complex medical datasets (Douzas & Bacao, 2018; Frid-Adar et al., 2018). Even with these developments, significant issues persist, including preserving data integrity, preventing overfitting, and achieving robust generalization on large-scale tabular healthcare datasets.

These limitations indicate that improvements at the data level alone may not be sufficient, thereby motivating the exploration of more integrated modeling strategies. However, most existing approaches still focus on applying these techniques within single-model frameworks or homogeneous ensembles, such as multiple tree-based or boosting-based models. Studies that integrate heterogeneous ensemble strategies, particularly combining algorithms with fundamentally different learning paradigms, such as bagging (e.g., Random Forest) and boosting (e.g., CatBoost) in conjunction with imbalance handling techniques remain relatively limited. Furthermore, the joint evaluation of such hybrid ensembles in large-scale tabular medical datasets is still insufficiently explored. This gap highlights the need for a more comprehensive framework that simultaneously addresses class imbalance and leverages complementary model characteristics to improve both predictive performance and generalization.

Motivated by these research gaps, this study proposes a hybrid ensemble approach that integrates Random Forest and CatBoost within a unified prediction framework. While ensemble methods and hybrid models have been widely explored in previous studies (Nissa et al., 2024; Shorewala, 2021), most approaches focus on homogeneous ensembles or apply imbalance handling techniques separately. The novelty of this study lies in the integrated use of algorithms with different learning paradigms, bagging and boosting (Zhou, 2012) combined with Borderline-SMOTE (Chawla et al., 2002; Han et al., 2005) in a single framework to simultaneously address class imbalance and improve model generalization. Furthermore, unlike prior works that evaluate models independently, this study systematically analyzes the complementary behavior of Random Forest and CatBoost on a large-scale imbalanced medical dataset, providing deeper insight into how heterogeneous ensembles can enhance both predictive performance and robustness in cardiovascular disease prediction.

In the context of national defense, maintaining the health and operational readiness of military personnel is of critical importance. Cardiovascular diseases can significantly impair physical endurance, cognitive performance, and mission effectiveness, thereby posing risks to both individual personnel and overall operational capability. Early detection of cardiovascular risk factors is therefore essential to support preventive interventions and ensure long-term readiness. In this context, machine learning–driven predictive models such as the approach introduced in this study, offer the potential to support military healthcare systems in detecting high-risk individuals and facilitating timely clinical decisions.

## METHOD

### Machine Learning for Heart Disease Prediction

Machine learning is a field within artificial intelligence that allows systems to learn from data and carry out prediction or decision-making tasks without requiring explicit programming (Murphy, 2012). In the healthcare domain, this approach has been widely utilized to detect and predict various diseases, including cardiovascular diseases, through the analysis of risk factor patterns derived from clinical and demographic data.

Heart disease prediction is generally formulated as a classification problem within the supervised learning framework, where models are trained using labeled data to distinguish between individuals at risk and those not at risk. Tree-based classification algorithms and ensemble

techniques have become particularly popular due to their ability to handle tabular data with mixed feature types and capture non-linear relationships among variables (Chen & Guestrin, 2016; James et al., 2013; Ke, 2017; Sagi & Rokach, 2018). Recent studies have also demonstrated that ensemble and non-linear models consistently outperform traditional statistical approaches in cardiovascular risk prediction tasks (Krittanawong, 2020; Weng, 2017).

Body Mass Index (BMI) is a commonly used anthropometric measure for evaluating an individual's weight status, calculated as the ratio of body weight in kilograms to the square of height in meters. The BMI value is then used to classify individuals into several weight status categories, referred to as BMI Category, namely underweight, normal, overweight, and obesity. This classification is commonly applied in epidemiological studies because it represents different levels of health risk associated with obesity and cardiovascular diseases (GBD 2015 Obesity Collaborators, 2017; Hruby & Hu, 2015; World Health Organization, 2000). Numerous large-scale studies have confirmed that elevated BMI is strongly associated with increased risk of cardiovascular morbidity and mortality.

**Ensemble Learning: Bagging, Boosting, and Voting**

Ensemble learning is a method that integrates multiple models to achieve greater predictive accuracy and robustness than any individual model alone (Breiman, 2001; Zhou, 2012). The fundamental principle is to leverage the collective strength of several models in order to simultaneously reduce bias and variance, as discussed in theoretical studies on ensemble methods in machine learning (Dietterich, 2000).

Bagging methods, such as Random Forest, construct multiple decision trees using different bootstrap samples to reduce variance and mitigate the risk of overfitting (Breiman, 2001). In contrast, boosting methods build models sequentially by assigning greater weights to instances that were misclassified by previous models, as implemented in various modern gradient boosting algorithms (Latha & Jeeva, 2019).

A voting classifier is an ensemble technique that combines outputs from several independent models. In soft voting, the final prediction is obtained by averaging the probability estimates produced by each model (Ashri et al., 2021). This strategy enables the integration of bagging-based and boosting-based models within a unified and more stable prediction framework.

**Random Forest and CatBoost**

Random Forest is a bagging-based ensemble method that combines multiple decision trees with random feature selection at each node (Breiman, 2001). The main advantages of this algorithm lie in its ability to reduce variance, handle high-dimensional feature spaces, and provide relatively interpretable measures of feature importance (Belgiu & Drăgut, 2016; Biau & Scornet, 2016; Müller & Guido, 2016). These characteristics make it one of the most widely used algorithms for tabular data, particularly in medical and healthcare applications due to its robustness and stability (Cutler, 2007; Qi, 2012).

CatBoost is a gradient boosting method developed to natively manage categorical variables by leveraging an ordered boosting technique (Prokhorenkova, 2018). This algorithm helps reduce the risk of target leakage and overfitting while demonstrating competitive performance on complex tabular datasets. Recent studies have shown that CatBoost achieves superior performance compared to other boosting algorithms in various real-world applications, particularly when dealing with categorical and heterogeneous data (Dorogush et al., 2018; Hancock & Khoshgoftaar, 2020).

The differences in learning mechanisms between Random Forest and CatBoost create an opportunity to integrate both models within an ensemble framework to achieve more robust performance. Combining bagging and boosting paradigms has been shown to improve generalization performance by leveraging complementary strengths and reducing both bias and variance (Dong, 2020; Sagi & Rokach, 2018; Zhou, 2012).

**Class Imbalance and SMOTE**

Imbalanced class distribution is a frequent challenge in medical datasets, where positive instances are typically far fewer than negative ones. This condition may cause predictive models to become biased toward the majority class and reduce their ability to detect disease cases.

The Synthetic Minority Over-sampling Technique (SMOTE) was developed to address this problem by generating synthetic samples for the minority class using a *k-nearest neighbors* approach (Chawla et al., 2002). Variants such as Borderline-SMOTE focus on samples located near the class boundary, which are more prone to misclassification (Han et al., 2005). This approach is particularly relevant in this study as it improves the model's sensitivity to heart disease cases without simply duplicating minority class samples.

**Related Work and Research Position**

Several previous studies have evaluated the performance of ensemble algorithms for heart disease prediction. The study by (Latha & Jeeva, 2019) showed that ensemble approaches combining multiple classifiers can improve prediction accuracy compared to single models by reducing variance and classification error. Another study by (Shorewala, 2021) reported that ensemble techniques improve sensitivity and performance stability in early heart disease detection compared to individual algorithms. In addition, comparative studies employing both boosting and bagging approaches within an ensemble framework have demonstrated better capability in capturing complex patterns in cardiovascular datasets, as reported by (Nissa et al., 2024).

Unlike previous studies that generally compare algorithms independently, this study integrates Random Forest and CatBoost through a *soft voting* approach while combining it with the Borderline-SMOTE technique to address class imbalance. This approach aims to produce a model that is not only accurate but also stable and sensitive to minority classes, making it more suitable for implementation in clinical decision support systems.

This study employs a quantitative supervised learning approach to develop a heart disease prediction model. The overall research workflow is illustrated in Figure 1. Experiments were conducted using Python with libraries including pandas, NumPy, scikit-learn, CatBoost, and imbalanced-learn. The dataset used is the BRFSS 2021 "Cardiovascular Diseases Risk Prediction Dataset" obtained from Kaggle, consisting of 308,854 observations with 19 variables, where Heart_Disease is defined as the target variable.

Data preprocessing includes duplicate removal, verification of missing values, BMI categorization based on WHO standards, and encoding of categorical and ordinal features. Exploratory data analysis was conducted to assess the distribution of the data and to investigate relationships between variables. To address class imbalance, the Borderline-SMOTE technique was applied prior to dataset splitting using an 80:20 train–test ratio with a fixed random state.

Model development involves Random Forest and CatBoost as base learners, followed by their integration using a soft voting ensemble. Model performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Finally, permutation importance is used to analyze feature contributions to the predictive performance.

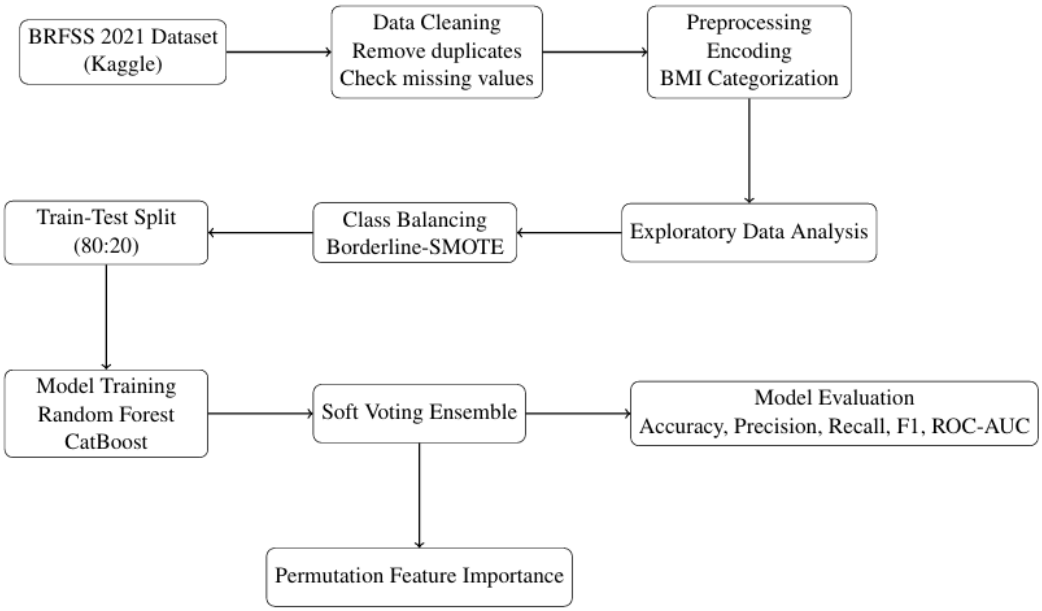



**Figure 1**. Research workflow for heart disease prediction using a hybrid ensemble approach.

## RESULTS AND DISCUSSION

### Data Exploration and Analysis of Variable Relationships

The final dataset after the cleaning process consists of 308,774 observations without missing values, as shown in Table 1.

**Table 1**. Duplicate data cleaning.

| Description | Value |
|---|---|
| Total duplicate rows observed | 80 |
| Shape before removing duplicates | 308,854 |
| Shape after removing duplicates | 308,774 |
| Total duplicates after removal | 0 |

Exploratory analysis indicates that BMI exhibits the highest variation among the numerical variables. Bivariate analysis shows that heart disease cases are concentrated in the overweight and obesity BMI categories. In addition, individuals with high alcohol consumption and low intake of fruits and vegetables tend to have a higher risk of heart disease. The visualization of categorical variables indicates that the history of smoking significantly increases the proportion of heart disease cases, while physical activity acts as a protective factor.

Pearson correlation analysis reveals a strong relationship between *Weight* and *BMI* ($r = 0.86$), as well as between *BMI* and *BMI_Category* ($r = 0.83$). With respect to the target variable, *Age_Category* ($r = 0.23$), *Diabetes* ($r = 0.18$), *Arthritis* ($r = 0.15$), and Smoking_*History* ($r = 0.11$) show positive correlations, while *General_Health* exhibits a negative correlation ($r = -0.23$). Although the initial feature selection applied a threshold of $|r| \geq 0.10$, all features were retained because tree-based models are capable of capturing non-linear relationships and complex feature interactions.

The class distribution before and after the balancing process is presented in Figure 2.

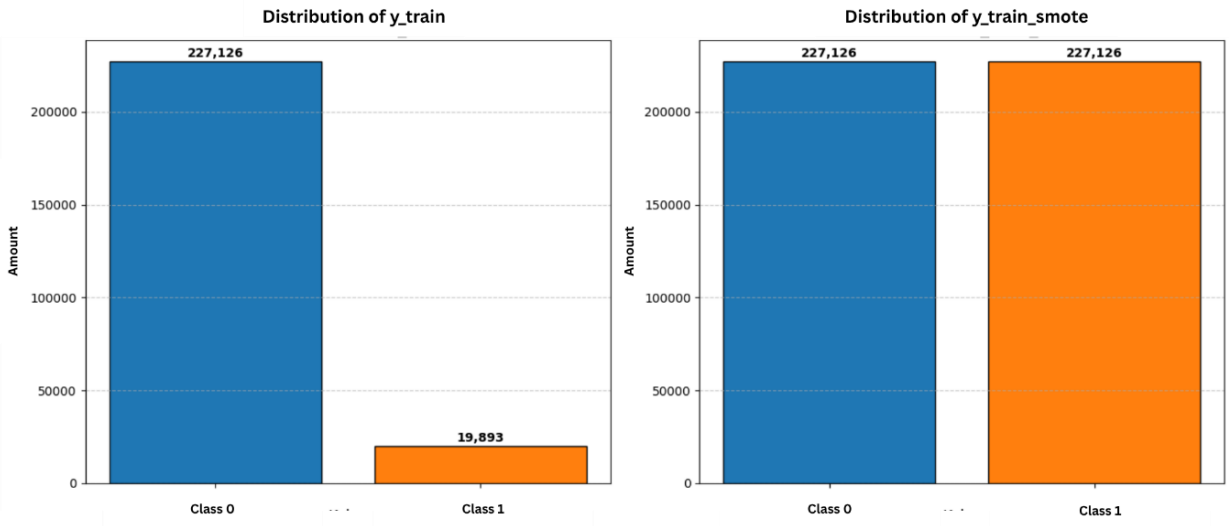



**Figure 2**. Data balancing.

The implementation of *Borderline-SMOTE* produced a more balanced distribution, thereby improving the model's ability to detect heart disease cases. Overall, the results of the exploratory analysis suggest that age, general health status, obesity, and smoking history are key determinants in predicting cardiovascular disease risk, which are further validated in the modeling stage.

### Model Evaluation and Performance Comparison

Based on the exploratory analysis and the identified relationships among variables in the previous subsection, all processed features were subsequently used in the modeling stage to evaluate the classification performance of each algorithm. This stage aims to assess how well the patterns

identified in the initial analysis can be translated into predictive models capable of accurately and consistently detecting the risk of heart disease.

The results of the evaluation indicate that Random Forest attains the top accuracy of 0.94, while maintaining a strong balance between precision and recall for both classes. In contrast, CatBoost achieves an accuracy of approximately 0.84, demonstrating relatively balanced performance, particularly in detecting the minority class.

The ensemble approach based on *soft voting* achieves an accuracy of 0.91 without indications of overfitting, as reflected by the very small difference between training and testing accuracy. Further analysis using the *confusion matrix* indicates that the model maintains good sensitivity for the positive class with a relatively low number of *false negatives*.

Overall, Random Forest achieves the highest accuracy, while the ensemble model provides more consistent metric stability. The integration of prediction probabilities from both algorithms offers a balance between performance and robustness, making it more suitable for implementation in clinical decision support systems.

**Model Comparison Analysis and Performance Implications**

The experimental results indicate that the three approaches—Random Forest, CatBoost, and the ensemble model—exhibit different performance characteristics. A visual comparison of model performance is presented in Figure 3.

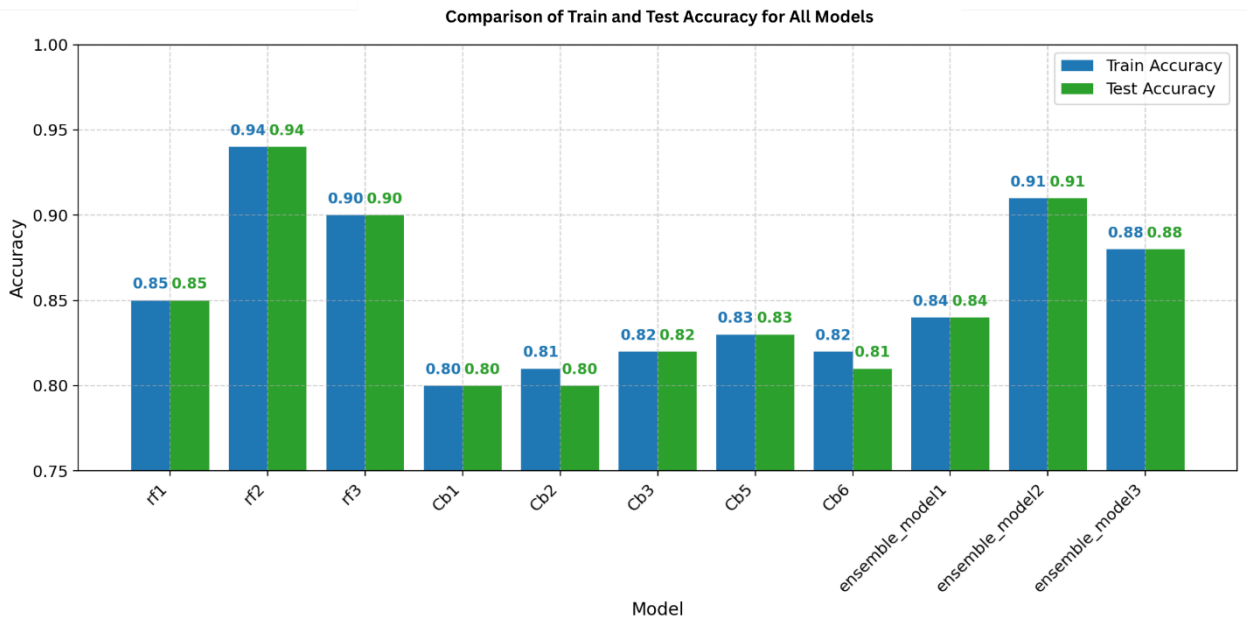



**Figure 3**. Comparison of accuracy across all models.

From a quantitative perspective, Table 2 presents a summary of the evaluation metrics for each model. Among them, Random Forest (rf2) obtained the highest test accuracy of 0.94, with precision and recall for the positive class reaching 0.93 and 0.96, respectively. This performance suggests that Random Forest is highly effective in handling tabular data with mixed feature structures and can optimally leverage the balanced dataset. The model's stability is also reflected by the minimal difference between training and testing accuracy, indicating no evidence of overfitting.

CatBoost (cb4), although achieving a lower accuracy (approximately 0.83–0.84), demonstrates relatively balanced metrics, particularly for the positive class with an F1-score of 0.85. The performance trend of CatBoost across different experimental configurations can be observed in Figure 4.

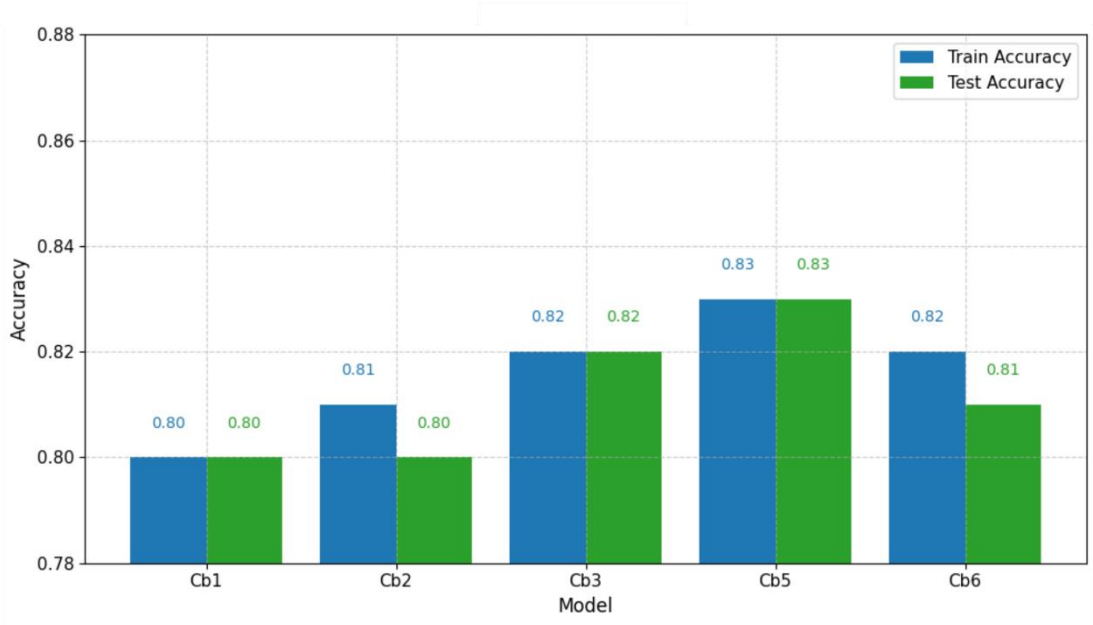



**Figure 4.** Accuracy comparison for CatBoost models.

As summarized in Table 2, this model shows relatively high recall for the positive class, suggesting that the boosting mechanism improves sensitivity to the minority class, although it does not maximize overall accuracy as effectively as Random Forest.

The best ensemble model (Ensemble 2), constructed using a *soft voting* approach, achieved an accuracy of 0.91 with balanced precision and recall across both classes, as shown in Table 2. The integration of prediction probabilities from Random Forest and CatBoost effectively maintains sensitivity for positive cases without significantly increasing classification errors. Although the ensemble accuracy is slightly lower than that of Random Forest, the stability of the evaluation metrics and the absence of significant differences between training and testing accuracy indicate that the ensemble model has strong generalization capability.

**Table 2**. Summary of the best model performance.

| Model | Test Accuracy | Precision(Positive) | Recall (Positive) |
|---|---|---|---|
| Random Forest (rf2) | 0.94 | 0.93 | 0.96 |
| CatBoost (cb4) | 0.83 | 0.81 | 0.89 |
| Ensemble (Model 2) | 0.91 | 0.88 | 0.94 |

From a practical perspective, these results indicate that model selection depends strongly on the intended application objective. If maximum predictive accuracy is the primary priority, Random Forest is the most optimal choice. However, if the balance between sensitivity and prediction stability is the main concern—as in clinical decision support systems—then the ensemble approach provides a more balanced and robust compromise.

**Feature Importance Analysis and Clinical Interpretation**

The *permutation importance* analysis across the three approaches reveals a consistent hierarchy of feature importance. The quantitative values of the three most influential features for each model are summarized in Table 3, Table 4, and Table 5.

For the Random Forest model, permuting the *Age_Category* feature leads to the largest decrease in accuracy (30.94%), followed by *BMI* and *General_Health*, as shown in Table 3. These results confirm that age, general health condition, and body mass index are the primary determinants in heart disease prediction when using bagging-based models.

Table 3. Top three most important features in random forest (rf2).

| Feature | Accuracy Decrease | New Accuracy |
|---|---|---|

| | | |
|---|---|---|
| Age_Category | 0.3094 | 0.6302 |
| BMI | 0.2822 | 0.6574 |
| General_Health | 0.2526 | 0.6871 |

In the CatBoost model, a similar pattern is observed, although with a smaller magnitude of contribution, as summarized in Table 4. The features *Age_Category* and *General_Health* remain dominant predictors, while the feature contribution distribution tends to be more evenly spread. This indicates that boosting-based methods distribute predictive contributions more proportionally across features.

Table 4. Top three most important features in CatBoost (cb4).

| Feature | Accuracy Decrease | New Accuracy |
|---|---|---|
| Age_Category | 0.1856 | 0.6524 |
| General_Health | 0.1389 | 0.6990 |
| BMI | 0.0774 | 0.7605 |

In the ensemble model, the consistency of the feature hierarchy becomes even more apparent. The feature contribution values for the ensemble model are presented in Table 4. Permuting the *Age_Category* feature reduces accuracy by 26.71%, followed by *General_Health* and *BMI*. In addition, all features exhibit relatively meaningful contributions, including variables with smaller effects such as *Exercise*. This indicates that the ensemble model is capable of capturing complex interactions among variables that may not be optimally detected by single models.

Table 5. Top Three Most Important Features in Ensemble Model 2.

| Feature | Accuracy Decrease | New Accuracy |
|---|---|---|
| Age_Category | 0.2671 | 0.6456 |
| General_Health | 0.2165 | 0.6962 |
| BMI | 0.2094 | 0.7034 |

From an application perspective, the identified key predictors have important implications in military health contexts. Higher BMI, particularly in the overweight and obesity categories, has been associated with reduced physical endurance, mobility limitations, and increased fatigue, all of which can negatively affect mission performance (GBD 2015 Obesity Collaborators, 2017; Hruby & Hu, 2015). Similarly, increasing age is closely related to declining cardiovascular capacity, reduced aerobic fitness, and slower recovery rates, which may impair operational readiness and physical performance in demanding environments (Fleg et al., 2005; Kodama et al., 2009). In addition, poor general health status may reflect underlying chronic conditions that can limit an individual's ability to perform physically intensive tasks and increase vulnerability to cardiovascular events (Mensah et al., 2019).

Therefore, the ability of the proposed model to accurately identify these risk factors provides practical value in supporting early intervention strategies. In military settings, predictive models have been increasingly recognized as valuable tools for enhancing personnel health monitoring, optimizing fitness management, and supporting preventive medical decision-making (Krittanawong, 2020; Rajkomar, 2018). Such approaches can assist in prioritizing medical evaluations, reducing the risk of sudden cardiovascular events during high-intensity operations, and improving long-term force readiness (Harion et al., 2018; Nindl et al., 2016). These findings demonstrate that the model is not only predictive but also actionable for real-world decision support, particularly in high-stakes environments such as military operations.

## CONCLUSION

This study evaluated the performance of Random Forest, CatBoost, and a soft voting-based ensemble approach for predicting heart disease risk on an imbalanced dataset. The findings demonstrate that Random Forest outperformed the other models, attaining an accuracy of 0.94 and balanced precision and recall, while CatBoost demonstrated stable but lower performance. The

ensemble model provided competitive and consistent results, although it did not outperform Random Forest. These results suggest that Random Forest demonstrates strong effectiveness. The permutation importance analysis consistently identified Age_Category, General_Health, and BMI as the most influential predictors. These results highlight the critical role of physiological condition, aging, and obesity in cardiovascular risk assessment, providing insights that are both predictive and clinically interpretable.

From an application perspective, the proposed model has strong potential to be implemented as an early warning system for cardiovascular risk detection, particularly in structured populations such as military personnel. The model can support military decision-making by enabling timely identification of high-risk individuals, thereby assisting in fitness evaluation, deployment readiness assessment, and preventive health interventions. Furthermore, it can contribute to risk-based personnel management by helping prioritize medical screening, optimize resource allocation, and reduce the likelihood of health-related performance degradation during operations.

Despite these findings, this study is subject to several limitations. First, the dataset relies on survey-based data, which may contain self-reported bias and lacks detailed clinical measurements such as laboratory test results. Second, the model is evaluated on a single dataset, which may limit its generalizability to other populations or healthcare settings. Third, although class imbalance is addressed using Borderline-SMOTE, other advanced imbalance handling techniques and cost-sensitive learning approaches were not explored.

For future research, integrating more comprehensive clinical features, such as cholesterol levels, blood pressure, family history, and electrocardiogram (ECG) data, is suggested to enhance the model's predictive performance. Additionally, exploring advanced ensemble techniques such as stacking and integrating the model into real-time decision support systems may further enhance its applicability in operational and healthcare environments.

## AUTHOR CONTRIBUTIONS

M.I.: Conceptualization, formal analysis, methodology, supervision. Z.Z.: Conceptualization, resources, supervision. I.F.: Data curation, visualization, formal analysis, methodology, writing - original draft. I.M.: Investigation, validation, writing - review & editing.

## CONFLICT OF INTEREST

The authors declare that have no conflict of interest.

## REFERENCES

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., & der Schaar, M. (2019). Cardiovascular Disease Risk Prediction Using Automated Machine Learning: A Prospective Study of 423,604 UK Biobank Participants. *PLOS ONE*, *14*(5), e0213653. https://doi.org/10.1371/journal.pone.0213653

Ashri, S. E., El-Gayar, M. M., & El-Daydamony, E. M. (2021). HDPF: Heart disease prediction framework based on hybrid classifiers and genetic algorithm. *IEEE Access*, *9*, 146797–146809. https://doi.org/10.1109/ACCESS.2021.3122519

Belgiu, M., & Dr\uagu\ct, L. (2016). Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chawla, N. V, Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *European Conference on Principles of Data Mining and Knowledge Discovery*, 107–119. https://doi.org/10.1007/978-3-540-45167-9_13

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD*. https://doi.org/10.1145/2939672.2939785

Cutler, D. R. et al. (2007). Random Forests for Classification in Ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Dong, X. et al. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

Dorogush, A. V, Ershov, V., & Gulin, A. (2018). CatBoost: Gradient Boosting with Categorical Features Support. *ArXiv Preprint ArXiv:1810.11363*.

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, *91*, 464–471. https://doi.org/10.1016/j.eswa.2017.09.030

Fleg, J. L., Morrell, C. H., Bos, A. G., Brant, L. J., Talbot, L. A., Wright, J. G., & Lakatta, E. G. (2005). Accelerated longitudinal decline of aerobic capacity in healthy older adults. *Circulation*, *112*(5), 674–682. https://doi.org/10.1161/CIRCULATIONAHA.105.545459

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, *321*, 321–331. https://doi.org/10.1016/j.neurocomputing.2018.01.093

GBD 2015 Obesity Collaborators. (2017). Health Effects of Overweight and Obesity in 195 Countries Over 25 Years. *New England Journal of Medicine*, *377*(1), 13–27. https://doi.org/10.1056/NEJMoa1614362

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *International Conference on Intelligent Computing*, 878–887.

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for Big Data: An Interdisciplinary Review. *Journal of Big Data*, *7*(1), 94. https://doi.org/10.1186/s40537-020-00369-8

Harion, W. J. T., Friedl, K. E., Buller, M. J., Arango, N. H., & Hoyt, R. W. (2018). Evolution of Physiological Status Monitoring for Ambulatory Military Applications. In *Human Performance Optimization: The Science and Ethics of Enhancing Human Capabilities* (pp. 142–164). Elsevier. https://doi.org/10.1016/B978-0-12-813734-7.00007-0

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IJCNN)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hruby, A., & Hu, F. B. (2015). The epidemiology of obesity: A big picture. *Pharmacoeconomics*, *33*(7), 673–689. https://doi.org/10.1007/s40273-014-0243-x

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Johnson, A. E. et al. (2016). Machine learning and decision support in critical care. *Proceedings of the IEEE*, *104*(2), 444–466. https://doi.org/10.1109/JPROC.2015.2501978

Karna, V. V. R. et al. (2025). A Comprehensive Review on Heart Disease Risk Prediction Using Machine Learning and Deep Learning Algorithms. *Archives of Computational Methods in Engineering*, *32*(3), 1763–1795. https://doi.org/10.1007/s11831-024-10015-6

Ke, G. et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NeurIPS*.

Kodama, S., Saito, K., Tanaka, S., Maki, M., Yachi, Y., Asumi, M., Sugawara, A., Totsuka, K., Shimano, H., Ohashi, Y., Yamada, N., & Sone, H. (2009). Cardiorespiratory fitness as a quantitative predictor of all-cause mortality and cardiovascular events in healthy men and women: A meta-analysis. *JAMA*, *301*(19), 2024–2035. https://doi.org/10.1001/jama.2009.681

Krittanawong, C. et al. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports*, *10*, 16057. https://doi.org/10.1038/s41598-020-72685-1

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203. https://doi.org/10.1016/j.imu.2019.100203

Mensah, G. A., Roth, G. A., & Fuster, V. (2019). The global burden of cardiovascular diseases and risk factors: 2020 and beyond. *Journal of the American College of Cardiology*, *74*(20), 2529–2532. https://doi.org/10.1016/j.jacc.2019.10.009

Miotto, R. et al. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, *6*, 26094. https://doi.org/10.1038/srep26094

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python*. O'Reilly Media.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Nindl, B. C., Jones, B. H., Van Arsdale, S. J., Kelly, K., & Kraemer, W. J. (2016). Operational physical performance and fitness in military women: Physiological, musculoskeletal injury, and optimized physical training considerations for successfully integrating women into combat-centric military occupations. *Military Medicine*, *181*(suppl_1), 50–62. https://doi.org/10.7205/MILMED-D-15-00363

Nissa, N., Jamwal, S., & Neshat, M. (2024). A technical comparative heart disease prediction framework using boosting ensemble techniques. *Computation*, *12*(1), 15. https://doi.org/10.3390/computation12010015

Prokhorenkova, L. et al. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*, 6638–6648.

Qi, Y. (2012). Random Forest for Bioinformatics. In *Ensemble Machine Learning* (pp. 307–323). Springer. https://doi.org/10.1007/978-1-4419-9326-7_11

Rajkomar, A. et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*, 18. https://doi.org/10.1038/s41746-018-0029-1

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), e1249. https://doi.org/10.1002/widm.1249

Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked*, *26*, 100655. https://doi.org/10.1016/j.imu.2021.100655

Weng, S. F. et al. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, *12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

World Health Organization. (2000). *Obesity: Preventing and Managing the Global Epidemic* (Vol. 894).

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.