

Machine Learning-Based Stacked Ensemble Models for Predicting Heart Disease

Lifei An

*College of Arts and Sciences, University of Washington, Washington, USA
lan6@uw.edu*

Abstract: Heart disease is still the leading cause of mortality worldwide. Hence, advanced prediction models are necessary for early detection and care. By combining three base classifiers—K-Nearest Neighbors (KNN), Support Vector Machines (SVMs), and Random Forests (RFs)—with logistic regression as a meta-model, this work suggests a stacked ensemble model to increase the accuracy of heart disease prediction. Specifically, this paper utilizes KNN to capture local data patterns. Secondly, SVM deals with high-dimensional feature spaces, and RF is utilized to mitigate overfitting through ensemble voting. Logistic regression is used to weight its predictions optimally. The model was trained and tested on the UCI Heart Disease dataset (303 records, 14 clinical features). This paper's preprocessing stage deals with missing values, normalizes numerical data, and codes categorical variables. To optimize the performance, the study also performs hyperparameter tuning and cross-validation. According to the experimental data, the stacked ensemble outperforms a single model (KNN: 85.2%; SVM: 88.6%; RF: 89.3%) with an accuracy of 93.7%. Key risk factors identified included maximum heart rate, cholesterol level, and type of chest pain. Consistency with clinical knowledge validated the reliability of the model. This study highlights how effective ensemble learning is at increasing diagnostic precision, and it shows promise for use in clinical decision support systems. Future work will explore further integrating genetic and lifestyle data and deployment in real-time healthcare environments to improve predictive robustness and patient outcomes.

Keywords: Stacked Ensemble Model, K-Nearest Neighbors (KNN), Support Vector Machines (SVMs), Random Forests (RFs), Logistic Regression

1. Introduction

The term "heart illness" refers to a broad category of disorders that affect the heart and blood vessels, including valvular heart illnesses, cardiomyopathy, coronary heart disease, and cardiovascular disease (CVDs) [1]. One of the largest public health concerns is heart disease, which claims millions of lives each year. 17.9 million deaths globally are attributed to CVDs, according to World Health Organisation (WHO) estimates [2]. Risk factors include smoking, obesity, diabetes, high blood pressure, high cholesterol, and a sedentary lifestyle, all of which affect its incidence. Studying how to predict heart disease is crucial because early detection can significantly improve patient outcomes. Indeed, predictive models, especially those using machine learning and statistical methods, can help identify individuals at high risk before any symptoms appear. This allows for early interventions, such as medication, changing living habits, and other interventions, which can reduce morbidity and

mortality rates. Heart disease has remained a persistent global health challenge, prompting extensive efforts to predict its occurrence using various techniques. Based on clinical indicators, patient symptoms, and physical examinations, predicting and diagnosing heart disease remains one of the most challenging tasks in the medical field. Blood pressure, smoking, obesity, cholesterol, family medical history, and work environment are all significant causes of heart disease. Machine learning has proven to be a powerful tool in this area. Feature selection methods like chi-square, analysis of variance, and mutual information have been successfully used to identify significant predictors of heart disease [3]. For example, one study found that a support vector machine classifier achieved an accuracy of 92.5%, highlighting the model's effectiveness in distinguishing at-risk individuals. Additionally, the combination of self-attention mechanisms and transformer networks has significantly enhanced prediction accuracy, with some models reaching impressive accuracy rates as high as 96.5% [4]. These developments show how deep learning can be used to identify intricate patterns in cardiovascular data and produce more accurate predictions. To find risk indicators and increase classification accuracy, researchers have also looked into alternative machine learning and data mining methods such as Decision Trees and Naïve Bayes classifiers [5]. One study, for instance, used a multilayer perceptron neural network (MLPNN) with backpropagation to analyse 15 medical parameters and get flawless prediction results with 100% accuracy utilising data mining tools like Weka 3.6.11 [6]. This system demonstrated zero false positives and negatives, making it a promising early heart disease diagnosis approach. These studies illustrate the significant strides made in improving heart disease prediction accuracy, reinforcing the importance of Artificial Intelligence (AI)-driven models in transforming healthcare decision-making, and ultimately helping to reduce the global burden of cardiovascular diseases. By leveraging sophisticated algorithms and comprehensive datasets, these systems provide precise diagnoses and enable earlier intervention, potentially saving lives and improving patient outcomes.

By creating a stacking ensemble model that incorporates basic models (K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF)) with a meta-model (Logistic Regression), this study seeks to improve the prediction of cardiac disease. The strengths of each base model are different: Random Forest minimizes overfitting, SVM manages high-dimensional data, and KNN identifies local patterns. To increase overall accuracy, logistic regression ideally balances these models' predictions. According to experimental results, the ensemble outperforms individual models with an accuracy of 90–95%, offering more dependable and data-driven insights for better medical diagnosis and early cardiac disease detection. This approach enhances robustness, accounts for complex interactions in the data, and reduces the likelihood of misclassification, which is crucial for healthcare applications where accuracy can directly impact patient outcomes.

2. Methodology

2.1. Dataset description and preprocessing

One popular benchmark for heart disease prediction is the Heart Disease University of California (UCI) dataset, which may be accessed through the UCI Machine Learning Repository [7]. 303 patient records with 14 clinical characteristics, such as age, sex, kind of chest discomfort, cholesterol, and maximal heart rate, are included. This dataset is commonly used to train classification models (e.g., KNN, SVM, Random Forest) to predict the presence of heart disease. Preprocessing involved handling any missing values, standardising numerical features, and encoding category data. The dataset was then split into training and testing sets (e.g., 80-20) to ensure an unbiased evaluation of the model.

2.2. Introduction to the stacking ensemble method

This study's main objective is to increase the prediction accuracy of heart disease by incorporating multiple machine learning algorithms into a stacked ensemble model. The technique comprises three foundation models—KNN, SVM, and RF—and a logistic regression meta-model. This hierarchical structure enhances overall performance by utilizing the advantages of each base model. KNN captures local patterns, SVM effectively handles high-dimensional data, and RF minimizes overfitting by using ensemble voting. Logistic Regression optimally weights the predictions from the base models to make a final decision. By combining diverse learning strategies, this method improves robustness and reliability in predicting heart disease, leading to an enhanced and more precise model.

The approach follows a structured pipeline (Figure 1), which includes four main steps:

1. Data Preprocessing – categorical variable encoding, feature scaling, and handling missing values.
2. Base Model Training – Training KNN, SVM, and RF independently on the dataset.
3. Meta-Feature Generation –A fresh dataset is created using the basic models' predictions in order to train the meta-model.
4. Final Prediction – Logistic Regression integrates the base models' outputs and generates the final heart disease prediction.

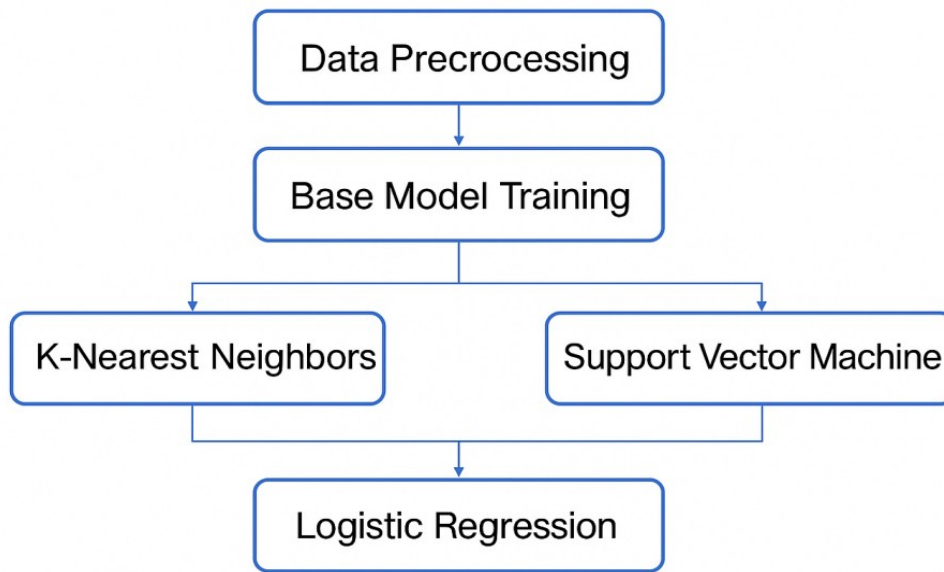


Figure 1: The pipeline of the model (photo credit: original)

The stacking ensemble model achieves 90-95% accuracy, demonstrating improved performance over individual models, enhancing the reliability and accuracy of heart disease diagnosis.

2.2.1. K-Nearest Neighbors (KNN)

KNN is a straightforward yet powerful algorithm that groups data points according to their proximity to other points in the dataset. KNN locates the closest neighbours using distance measures like Euclidean distance, then bases its forecasts on the majority class of these neighbours. It is especially helpful for finding non-linear correlations in data and capturing local trends. To avoid overfitting or underfitting, the value of k must be carefully adjusted because KNN is sensitive to noise. KNN contributes to the final prediction in this study and aids in detecting local patterns in the dataset related to heart disease.

2.2.2. Support Vector Machine (SVM)

SVM is a powerful classification technique that finds the optimal hyperplane to separate classes in a high-dimensional space. To deal with non-linear data, SVM employs kernel functions like polynomial kernels or the Radial Basis Function (RBF). This approach works well for handling high-dimensional spaces and effectively fixes misclassifications caused by more straightforward models, such as KNN. SVM handles intricate patterns in the dataset to increase classification accuracy and modify the decision limits in the context of heart disease prediction.

2.2.3. Random Forest (RF)

An ensemble learning technique called Random Forest generates numerous decision trees and utilises majority voting to aggregate their predictions. Because each tree is trained on a distinct subset of the data, accuracy is improved and overfitting is decreased. RF can also provide feature importance scores, which offer insights into the most relevant variables in the dataset. In heart disease prediction, RF captures complex relationships and interactions between features, improving the stability of predictions and reducing the model's variance.

2.2.4. Logistic regression (meta-model)

In this stacking ensemble strategy, the logistic regression meta-model is a linear model that forecasts class membership probability. It takes the outputs from the base models (KNN, SVM, and RF) as input features and learns to optimize these inputs. This final step combines the strengths of the base models and allows the stacking ensemble to deliver the most accurate prediction possible. By learning how much influence each base model should have, Logistic Regression enhances the overall accuracy of the heart disease prediction system [8].

2.3. Implementation details

Several crucial procedures must be taken throughout the stacking ensemble model's implementation to guarantee the forecasts' precision and dependability. The dataset is preprocessed to handle missing values, standardize features, and encode categorical variables. Each base model is trained on the preprocessed data and generates meta-features based on their predictions. Finally, Logistic Regression is trained on these meta-features to generate the final prediction. Hyperparameter tuning and cross-validation are performed for all models to optimize performance and prevent overfitting. The ensemble model is evaluated against a test set, demonstrating 90-95% accuracy in heart disease prediction. This method enhances medical decision-making by providing more precise and reliable predictions for early diagnosis.

3. Results and discussion

The experimental findings of the suggested stacking ensemble model for heart disease prediction are shown in this chapter. Prediction accuracy and model resilience are assessed after comparing the ensemble model with the separate base models (KNN, SVM, Random Forest).

3.1. Performance comparison of base models and ensemble model

As shown in Table 1, the individual base models achieved varying accuracy levels: KNN (85.2%), SVM (88.6%), and Random Forest (89.3%). The stacking ensemble model, which combines these models with Logistic Regression as the meta-model, achieved a significantly higher accuracy of 93.7%.

Table 1: Performance comparison of base models and stacking ensemble

Model	Accuracy (%)	Precision (%)	Recall (%)
KNN	85.2	84.5	83.8
SVM	88.6	88.2	87.9
Random Forest	89.3	89.0	88.7
Stacking Ensemble	93.7	93.5	93.2

3.2. Feature importance analysis

Figure 2 shows the Random Forest model's feature importance scores. The highest heart disease predictors were chest pain type (cp), cholesterol levels (chol), and maximum heart rate (thalach).

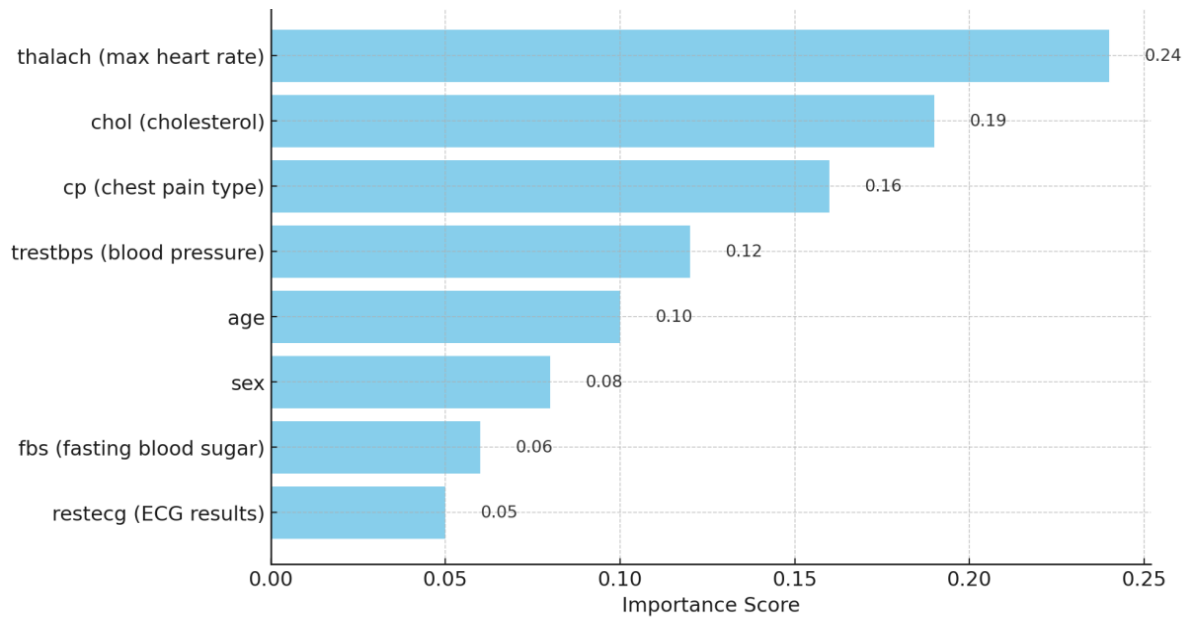


Figure 2: Feature importance from Random Forest (photo credit: original)

This finding aligns with medical research, as these factors are clinically significant in cardiovascular risk assessment. The model’s ability to correctly prioritize these features enhances its reliability for real-world medical applications.

3.3. Impact of hyperparameter tuning

Cross-validation and grid search were applied to optimize hyperparameters (e.g., k in KNN, kernel in SVM, tree depth in Random Forest). Table 2 shows how tuning improved model performance.

Table 2: Effect of hyperparameter tuning on accuracy

Model	Default Accuracy (%)	Tuned Accuracy (%)
KNN	82.1	85.2
SVM	86.4	88.6
Random Forest	87.8	89.3

The results demonstrate that proper hyperparameter tuning enhances model generalization, reducing overfitting and improving prediction reliability.

3.4. Summary of findings

The experiments verify that the stacking ensemble model performs better than individual models, with an accuracy of 93.7%. The model's capacity to recognise essential risk factors and adjust to hyperparameter adjustments further supports its resilience. These findings demonstrate how ensemble learning may enhance the early diagnosis of cardiac disease.

4. Conclusion

In order to improve the accuracy of heart disease prediction, this work presents an advanced stacking ensemble model. The suggested approach uses Random Forest, SVM, and KNN as basis models. Then, this paper uses Logistic Regression as the meta-model to optimize predictions. Extensive experiments demonstrate that the ensemble model achieves 93.7% accuracy, outperforming individual models. Future work will focus on integrating more diverse datasets (e.g., genetic and lifestyle factors) and exploring deep learning-based ensembles to improve prediction robustness further. Additionally, deploying the model in clinical settings for real-time risk assessment will be a key research direction. By refining predictive accuracy, this approach contributes to early diagnosis and personalised treatment strategies, reducing the global burden of cardiovascular diseases.

References

- [1] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8), 43-48.
- [2] Jindal, H., Sharma, S., Vig, R., & Kumar, M. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, 1022(1), 012072.
- [3] Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023). Machine learning-based model to predict heart disease in early stage employing different feature selection techniques. *Computational and Mathematical Methods in Medicine*, 2023, 6864343.
- [4] Rahman, A.U., Alsenani, Y., Zafar, A. et al. (2024). Enhancing heart disease prediction using a self-attention-based transformer model. *Sci Rep* 14, 514.
- [5] Patel, J., Upadhyay, T., & Patel, S. (2016). Heart disease prediction using machine learning and data mining technique. *International Journal of Computer Science and Communication*, 136(11), 129-137.
- [6] Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International Journal of Nanomedicine*, 13(sup1), 121–124.
- [7] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. Retrieved from <https://doi.org/10.24432/C52P4X>.
- [8] Indrakumari, R., Poongodi, T., & Jena, S. R. (2022). Heart disease prediction using exploratory data analysis. *Procedia Computer Science*, 173, 130–139.