Ensemble ***learning, Predicting CAD***

Farinaz Tanhaei

23044867

# Abstract

Early detection is important in heart disease to ease intervention. In this study, we explore the Comprehensive Heart Disease Dataset (Siddhartha, 2020) , which combines data from five distinct datasets. Our objective is binary prediction, where we aim to classify instances as either heart disease-positive or negative.

We employed several models, such as K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Random Forest. We used 10-fold cross-validation for Hyperparameter optimization for our model selection. In addition, we create a custom ensemble by combining Random Forest and KNN. Our voting classifier (soft) achieves the best accuracy among the classifiers considered.

# Introduction

Coronary artery disease (CAD), also known as coronary heart disease (CHD), is a condition in which blood vessels supplying oxygen to the heart muscles are narrowing down or blocked. A shortage of blood supply to the heart and brain is one of the leading causes of heart attacks and strokes. CAD is affecting millions of people and is a major cause of mortality in the world and the UK, causing around 64,000 deaths in the UK alone in 2021. Please see Figure 1 for more information.

Early detection allows prompt intervention and prevents complications of treatments. Specialists employ standard procedures such as electrocardiogram (ECG), magnetic resonance imaging (MRI), Cardiac computed tomography Scan (CT scan), and stress testing to predict heart disease. Employing machine learning models often needs a reliable dataset and a larger dataset to allow more complex models. A comprehensive heart disease dataset (Siddhartha, 2020) has been released, combining five different datasets. In this paper, we investigate different models and their performance on this dataset along with different data preparation.

Figure 1- Age-standardised death rates per 100,000 from CAD, among all ages in United Kingdom between 2000 and 2021

# Background

KNN calculate the distance between a test sample and surrounding samples in the train set. It borrows the label from the majority of labels in the K neighbours. When p is 1, the distance measurement is Manhattan distance; this measurement is useful when the dimension is high. When p is 2, the Euclidean distance is used; this distance is mostly used; however, it tends to be more sensitive to outliers compared to others. In the case when P is other, distance is Minkowski. The Minkowski distance is a generalization of both the Euclidean distance and the Manhattan distance.

SVM find an optimal hyperplane that separates data points into different classes (usually binary. However, there are some strategies that can extend it to multi-class classification) or continuous values in case of regression. C is a regularization which controls between overfitting and underfitting, small C increases regularisation, which makes it simpler; larger C allows the model to fit the training data more closely. Gamma determines the impact of a training sample. Therefore, a smaller gamma leads to wider decision boundaries, and a larger gamma leads to a more complex model.

# Related word

CAD is the most common type of heart disease in the world. The main root causes of CAD are plaques, when fat deposits and other substances build within the coronary arteries’ wall. These plaques become hard, narrow the arteries, and reduce blood flow (prevention, 2024). Many varied factors have been named development of CAD, such as Age, Gender, Hypertension, Cholesterol, Smoking, and diabetes (WHF, 2023).

CAD can usually be diagnosed using different approaches, such as medical history and physical examination (e.g., stress testing) and Image testing. Machine learning models have been proposed to enhance the prediction’s accuracy in image testing and automatic diagnosis of CAD (Rahendra et al. (Rajendra , et al., 2017) Used CNN to gain high accuracy in distinguishing between patients with and without CAD using ECG data. Using CNN and LSTM has been proposed by Liang et al., which improved the accuracy (Liang, et al., 2020). Using Angiography Imaging for detecting CAD has been investigated in Rangraz et al. (Jeddi, et al., 2023). A different version of U-Net++ (Zhu, et al., 2021) the model has shown that it can achieve high performance.

Different datasets exist to identify CAD, such as Cleveland (Janosi, 1988) , UCI Heart Disease dataset in the UCI Machine Learning Repository (Ahsan & Siddique, 2022). Usually, due to the size of the dataset, complex models cannot be applied. In contrast, models that have been used on these can be divided into two parts: 1) Traditional ML models and 2) ensemble learning. Traditional models such as SVM (Rose, et al., 2023), K-NN (Shah, et al., 2020), Random forest (A, et al., 2020), Decision Tree (S & D, 2021) has been shown a good accuracy but not compared to the Ensemble learning. In ensemble learning, XGB (Ozhan & Z. Kuçukakcali, 2022) and voting ensemble learning (Das & Sinha, 2023) (Doppala, et al., 2022) and Stacked learning (Tiwari, et al., 2022) has been used often.

In addition to the size of the dataset, the imbalanced dataset was another challenge that should be solved before training the models on the dataset. This problem has been dealt with different techniques, such as the Synthetic minority oversampling technique (SMOTE) (P, et al., 2014) (Saba, et al., 2016) and under-sampling (Nan, et al., 2012) and oversampling (Shivam & Rahul, 2021)

# Methodology

## Dataset:

In this paper, we use a Heart disease dataset (Comprehensive) Dataset that is a combination of Cleveland, Hungarian, Switzerland, Long Beach VA (Janosi, 1988), and Statlog (Heart) (UCI Repository, 2024) Data Sets. The table shows some characteristics of the dataset.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Count** | **Description** |
| Number of features | 11 | 6 Categorical + 5 Continues |
| Target | Binary | 561 Negative, 629 Positive |
| Null values | No |  |
| Repeated rows | Yes | 151 Negative, 121 Positive |
| Feature reduction (e.g., PCA and ..) | No | The number of features was low |

Table 1- Dataset properties and data preparation.

Preparation techniques can generally improve the model’s performance. We use Standard Scaler, which ensures that each feature has a mean of 0 and a standard deviation of 1. This is typically important for KNN, as it uses distance.

The dataset is mostly clean and does not have any null values; however, there are duplicate rows. After dropping duplicate rows, 410 negative samples and 508 positive samples remained. This means the results from (Akkur, 2023) is not reliable or comparable with our results as they did not identify the existence of duplicate rows. Finally, we divide the dataset into two random sections: 80 train set and %20 test set. We did not use the test set until producing the last results.

## Method:

The objective is binary prediction, and many machine learning models can assist in this regard. We wanted to use ensemble learning as the traditional models tend to overfit in this dataset. Ensemble learning is a combination of multiple models to improve overall performance. To do so, we selected different models, such as KNN and SVM, and compared them with Random Forest, Ensemble learning (soft voting) and another Stacked Ensemble learning that we made of random forests and KNN. We use the Support Vector Classifier as a Meta-Classifier. However, the performance of the ensemble learning algorithms depends on the individual models. Therefore, we need to optimise each classifier at its best. Machine learning models can perform better when they employ an efficient hyperparameter optimization approach ( (Monica and Agrawal, 2024)).

In Table 2, we show the range of hyper parameters that we have searched for optimising the models. K-fold cross-validation is useful when we have limited samples. We use 10-fold cross-validation to find the best hyperparameters for each model in this experiment. This setup avoids using a test set when finding the best models and their parameters to avoid bias. In the end, after we find the best hyperparameters, we train models on the whole train set and then reuse them in our ensemble learning.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyperparameters** | **Range** | **Selected** |
| KNN | P (power parameter) | [1:2:0.1] | 1.1 |
| Number of neighbours | [1:10:1] | 7 |
| Random Forest | Number of estimators | [50:250:10] | 200 |
| Max depth | [1:10:1] | 9 |
| Min samples leaves | [1:10:1] | 1 |
| SVM | Kernel | [linear, rbf] | rbf |
| Gamma | [0.01,0.1,0.2,0.3,0.4] | 0. 01 |
| C | [1,51,101,151,201,251] | 1 |

Table 2- hyperparameters in selected models, [a: b: c] refers a range from a to b with step c.

# Results

In the result section, we measure model performance with different metrics. Using only accuracy metrics cannot show which model is the best. Table 3 describes each metric.

|  |  |
| --- | --- |
| **Metric** | **Description** |
| Accuracy | The ratio between correct prediction and all samples |
| Precision | Ratio between true positive and all positive predictions |
| Recall | The ratio between true positive and all positive |
| F1 Score | A balance measure of precision and Recall |

Table 3- List of metrics that have been used and its description.

After finding the best hyperparameters, we have train models on the whole train set. We illustrate the model’s performance for all models in table 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1\_score** | **Precision** | **Recall** |
| KNN | 85.86 | 86.73 | 86.73 | 86.73 |
| SVM | 84.24 | 85.43 | 84.16 | 86.73 |
| Random Forest | 87.5 | 88.44 | 87.13 | 89.79 |
| EL- Soft voting | 86.95 | 87.88 | 87.00 | 88.78 |
| EL- Stacked | **88.04** | **89.00** | **87.25** | **90.82** |

Table 4- Models performance on the Heart datasets. Our model worked slightly better than others.

KNN performs better in accuracy and other metrics than the SVM. The KNN model also has a good F1 score. In general, having more true positives seems much more desirable than missing true positives in critical health systems.

Random forest is an ensemble learning method that combines many decision trees. It is more accurate than a single decision tree and can better handle outliers. As we can see, Random Forest performed better than SVM and KNN. In general, the random forest was slightly biased toward the true positive.

Our stack ensemble learning classifier has gained the best accuracy among the other classifiers that we have selected. It is good to notice that the margin between our classifier and the random forest is so low (e.g., %0.5 in accuracy), and the performance has not increased very much. In general, ensemble learning models performed better than our individual models.

# Conclusion & Furtur discussions

Our findings highlight the importance of ensemble learning in heart disease prediction. While KNN outperforms SVM in accuracy and other metrics. Random Forest, as an ensemble, surpasses the other two models and two of our voting classifiers. Notably, our stack ensemble learning classifier achieves better accuracy, closely competing with Random Forest. Further research can explore additional features and fine-tuning to enhance model performance.

It is good to investigate possible solutions for the overfitting problem in this dataset. Techniques such as data augmentation are useful in image data processing; however, we could not find similar techniques on tabular data. We might, however, be able to use some techniques, such as domain adaptation (DA), to overcome overfitting (Khoee, et al., 2024).

# Bibliography

Ahsan, M. M. & Siddique, Z., 2022. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine,* p. 102289.

Akkur, E., 2023. Prediction of Cardiovascular Disease Based on Voting Ensemble Model and SHAP Analysis. *Sakarya University Journal of Computer and Information Sciences 6 ,* p. 3.

A, R. et al., 2020. Heart disease prediction using machine learning. *International Journal of Research and Technology 2020,* p. 659–662.

Das, T. & Sinha, B. B., 2023. A comprehensive study on machine learning methods for predicting heart disease: a comparative analysis,. *8th International Conference on Computing in Engineering and Technology (ICCET 2023),* pp. 205-210.

Doppala, B., D, D. B., Janarthanan, M. & Baik, N., 2022. A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques.. *J Healthc Eng.*

Janosi, A. S., 1988. *Heart Disease. UCI Machine Learning Repository..* [Online]   
Available at: 2024

Jeddi, F. R. et al., 2023. Machine Learning Approaches for Detecting Coronary Artery Disease Using Angiography Imaging: A Scoping Review. *Healthcare Transformation with Informatics and Artificial Intelligence,* pp. 2244-248.

Liang, Y. et al., 2020. Deep Learning Algorithm Classifies Heartbeat Events Based on Electrocardiogram Signals. *Computational Physiology and Medicine,* p. 11.

Nan, L. et al., 2012. An intelligent scoring system and its application to cardiac arrest prediction. *IEEE Trans Inf Technol Biomed 2012,* p. 1324–1331.

Ozhan, O. & Z. Kuçukakcali, 2022. Estimation of risk factors related to heart attack with XGBoost that machine learning model.. *Middle Black Sea Journal of Health Science,* pp. 582-591.

P, E. K. et al., 2014. A multiscale approach for modeling atherosclerosis progression. *IEEE J Biomed Health Informatics,* p. 709–719.

prevention, C. f. d. C. a., 2024. *Heart Disease Facts.* [Online]   
Available at: https://www.cdc.gov/heartdisease/facts.htm

Rajendra , U. . A. et al., 2017. Automated Detection of Coronary Artery Disease Using Different Durations of ECG Segments with Convolutional Neural Network. *Knowledge-Based Systems,* p. 132.

Rose, J. S. et al., 2023. Heart Attack Prediction using Machine Learning Techniques. *2023 9th International Conference on Advanced Computing and Communication Systems,* pp. 210-213.

Saba, B., Usman, Q. & Hassan, K. F., 2016. A multicriteria weighted votebased classifier ensemble for heart disease prediction. *Comput Intell,* pp. 32-34.

Shah, D., Patel, S. & Bharti, S., 2020. Heart disease prediction using machine learning techniques.. *SN COMPUT. SCI,* p. 345.

Shivam, D. & Rahul, K., 2021. Early detection of heart diseases using a low-cost compact ecg sensor. *Multimed Tools Appl,* p. 32615–37.

S, P. & D, H., 2021. Machine Learning Techniques for Heart Disease Prediction. *Journal of Cardiovascular Disease Research,* pp. 93-96.

Tiwari, A., Chugh, A. & Sharma, A., 2022. Ensemble framework for cardiovascular disease prediction.. *Computers in Biology and Medicine,* p. 146.

UCI Repository, 2024. *Statlog (Heart).* [Online]   
Available at: https://archive.ics.uci.edu/dataset/145/statlog+heart

WHF, 2023. *Cardiovascular disease infographic,* s.l.: World Heart Federation.

Zhu, H. et al., 2021. Segmentation of Coronary Arteries Images Using Spatio-temporal Feature Fusion Network with Combo Loss. *Cardiovascular Engineering and Technology,* pp. 407-418.