ISSN 2959-6157

Enhancing Heart Disease Prediction through Machine Learning: A Comparative Analysis of Algorithm Generalization across Datasets with Various Distributions

Siming Lyu[1, *]

[1]Dundee International Institute, Central South University, Changsha, China
*Corresponding author: 2543514@dundee.ac.uk

Abstract:

Heart disease, due to its high prevalence and mortality, remains a key area of global research. Although traditional diagnostic methods are effective, they are often invasive and time-consuming, highlighting the need for non-invasive, AI-based approaches. A significant challenge in real-world applications is ensuring model generalization across different datasets, particularly when the datasets are small. In this study, the performance of machine learning models, including Decision Tree, Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP), was evaluated on two distinct heart disease datasets with different distributions and relatively small sizes. Two datasets with varying distributions were used for training and testing, with the primary focus on assessing model generalization in cross-dataset applications. It is shown in the results that, while the Decision Tree model performed best after hyperparameter tuning, the improvements in Random Forest and MLP were limited, and SVM exhibited a decline in performance after tuning in the cross-dataset task. It was found that grid search tuning has limitations in cross-dataset scenarios, especially with small datasets, where complex models are prone to overfitting. The study demonstrates that, with smaller datasets, simpler models like Decision Trees often adapt better to different datasets. Furthermore, transfer learning and domain adaptation techniques are suggested as crucial for improving model generalization. Future research should focus on employing these techniques to enhance the robustness and accuracy of heart disease prediction models across diverse datasets.

Keywords: Machine learning; generalization; grid search.

1. Introduction

Heart disease continues to be one of the leading causes of morbidity and mortality globally, underscoring its significance in medical research. The high prevalence and serious outcomes associated with heart disease underscore the need for early detection and accurate diagnosis. However, traditional diagnostic methods are often not recognized at the earliest possible time, leading to treatment delays. Therefore, there is a pressing need to explore innovative, non-invasive approaches for early and accurate diagnosis.

Artificial Intelligence (AI) has become a transformative tool in the medical field, providing innovative approaches for diagnosing and predicting heart disease. AI algorithms can process large datasets with exceptional accuracy, revealing patterns that traditional methods might miss. Machine learning and deep learning techniques, in particular, have shown promise in identifying subtle signs of heart disease from medical imaging, Electrocardiograms (ECGs), and Electronic Health Records (EHRs). By leveraging AI, researchers and clinicians can improve diagnostic accuracy and timeliness, predict patient outcomes and ultimately enhance patient care and saving lives.

The development of AI algorithms has seen remarkable progress over the past few decades. Advances in machine learning and deep learning have enabled AI systems to achieve high performance in various tasks. For instance, in medicine, AI has made significant strides, particularly in areas such as radiology, pathology, and cardiology. For instance, Convolutional Neural Networks (CNNs) have been successfully used to detect cardiovascular abnormalities from imaging data, while Recurrent Neural Networks (RNNs) have shown efficacy in analyzing time-series data from ECGs.

Numerous studies have emphasized the role of AI in predicting and classifying heart disease. One example is the use of machine learning models like random forests and SVMs, which have outperformed traditional statistical approaches. These models excel at managing complex interactions and non-linear patterns in data, leading to significant improvements in prediction accuracy [1-3]. In another study, a combination of six algorithms—including random forest, K-nearest neighbor, logistic regression, Naïve Bayes, gradient boosting, and AdaBoost—was

employed to boost prediction accuracy, reaching up to 95% through ensemble methods [4]. Additionally, deep learning models have been applied to ECG signal analysis, achieving high accuracy in detecting various heart conditions. Methods such as deep Convolutional Neural Networks (DCNNs) and Generative Adversarial Networks (GANs) have played a key role in enhancing the diagnostic process, enabling earlier detection and treatment [5, 6].

In real-world applications, a common challenge is the discrepancy between training and testing data, which can lead to reduced performance of AI models. Medical data can vary significantly across different populations, regions, and healthcare systems, making it difficult for AI models trained on one dataset to generalize to others.

Building on existing research, this study aims to systematically compare the training and prediction performance of different machine learning and deep learning algorithms on different data sets. Specifically, the performance of these algorithms on multiple heart disease data sets will be evaluated, including data sets from different healthcare systems and patient populations. By analyzing how different algorithms perform on different datasets, this study hopes to reveal the possible reasons behind these differences, thereby providing valuable insights for future AI model development.

2. Method

2.1 Dataset Preparation

In this study, two distinct datasets with various distributions were utilized to develop and evaluate machine learning models for heart disease prediction. The training dataset was sourced from Kaggle [7], containing 1,190 instances with 11 features, and the target variable indicating the presence or absence of heart disease. The testing dataset was also sourced from Kaggle [8], comprising 303 instances with 13 features and a similar binary target variable.

During the preprocessing phase, several steps were taken to ensure the quality and compatibility of the datasets. First, any records containing missing or duplicate data were removed. Following this, feature standardization was performed to normalize the data, ensuring that each feature had a mean of 0 and a standard deviation of 1.

A key part of the preprocessing involved comparing the distributions of features between the two datasets shown in Fig. 1 to identify and remove features that were not similarly distributed. This step aimed to enhance the model's ability to generalize across different datasets by focusing on features that behaved consistently between the training and testing sets.

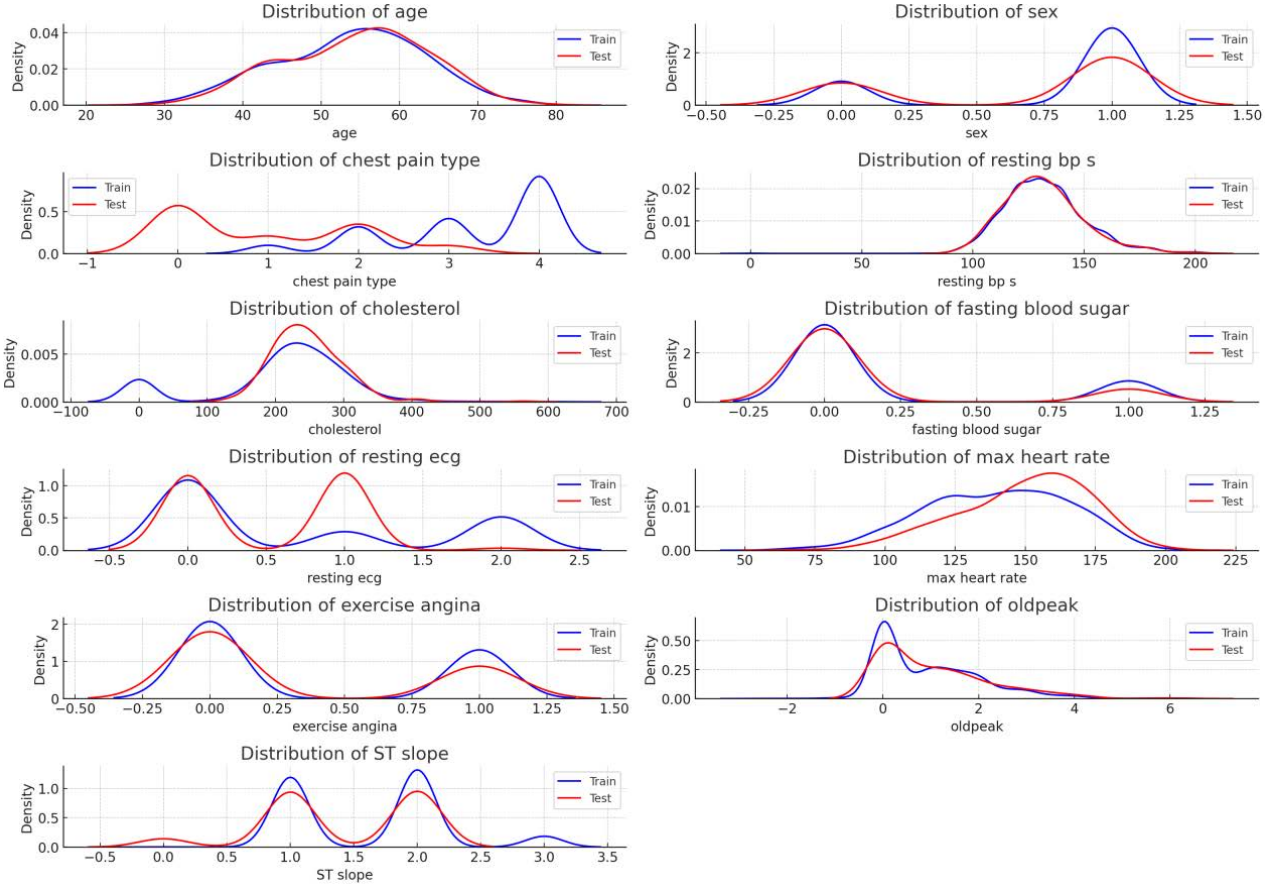

Fig. 1 The distributions of properties used (Photo/Picture credit: Original)

Unlike typical approaches where both datasets are split into training and testing sets, in this study, the model was trained exclusively on the training dataset and tested directly on the testing dataset. This approach was intentionally designed to assess the model's performance in cross-dataset generalization, simulating real-world scenarios where models are often deployed on data that differs from the data they were originally trained on.

2.2 Machine Learning-Based Prediction

Several machine learning algorithms were employed to predict heart disease based on the prepared datasets. The focus was on Decision Tree, Random Forest, SVM, and MLP classifiers [9, 10]. The performance of these models was evaluated using standard metrics, including accuracy and confusion matrix, to provide a comprehensive assessment of their predictive capabilities.

Each machine learning algorithm is optimized on the training set using five-fold grid search, with accuracy as the evaluation criterion.

2.2.1 Decision Tree

The Decision Tree algorithm is a simple and interpretable model that makes decisions by recursively splitting the dataset based on the value of its features. The model continues splitting until it reaches a decision, which is represented as a leaf node in the tree. For a decision tree, the grid search will explore the following parameter ranges: 1) max_depth: Varies from None (no limit) to specific values (3, 5, 7, 10) to control the depth of the tree. 2) min_samples_split: Tests different thresholds (2, 10, 20, 50) for the minimum number of samples required to split an internal node. 3) min_samples_leaf: Evaluates different minimum numbers of samples required to be at a leaf node (1, 5, 10, 20). 4) max_features: Considers various methods for selecting the maximum number of features to consider when looking for the best split, including None (all features), 'sqrt' (square root of the number of features), 'log2' (log base 2 of the number of features), and fractions (0.5, 0.8) of the total number of features.

2.2.2 Random Forest

Random Forest is an ensemble learning technique that combines the predictions of multiple decision trees to improve model performance. It works by creating several decision trees during training, each constructed from different random subsets of the data. The final prediction is made by averaging the outputs of all individual trees, which helps reduce overfitting and improves generalization.

For a random forest, the grid search will explore the following parameter ranges: 1) n_estimators: Tests different numbers of trees in the forest (100, 200). 2) max_depth: Varies the maximum depth of the trees from None (no limit) to specific values (10, 20).

2.2.3 Support Vector Machine

Support Vector Machines (SVM) is a powerful classification algorithm that operates by identifying the optimal hyperplane to separate data points from different classes. It excels in high-dimensional spaces, making it ideal for scenarios where the number of features exceeds the number of samples. By maximizing the margin between classes, SVM promotes robust generalization, even with complex datasets.

For an SVM, the grid search will explore the following parameter ranges: 1) C: Tests different regularization parameter values (0.1, 1, 10), which controls the trade-off between achieving a low training error and a low testing error. 2) gamma: Evaluates different kernel coefficient values for the kernel (0.001, 0.01, 0.1), which determine the influence of individual training examples. 3) kernel: Considers two kernel types for mapping data to a higher-dimensional space: 'linear' and 'rbf'.

2.2.4 Multi-Layer Perceptron

A Multi-Layer Perceptron (MLP) is a type of artificial neural network that consists of multiple layers of interconnected neurons, where each neuron in one layer is connected to all neurons in the next layer. MLP is particularly effective at capturing complex patterns in data through its multiple layers of non-linear transformations.

For an MLP, the grid search will explore the following parameter ranges: 1) hidden_layer_sizes: Tests different configurations for the number and size of hidden layers, including single layers with 50 or 100 neurons, and two-layer configurations (50, 50) and (100, 50). 2) activation: Considers different activation functions for the hidden layers, including 'relu' (Rectified Linear Unit) and 'tanh' (Hyperbolic Tangent). 3) solver: Evaluates different algorithms for weight optimization, specifically 'adam' (a stochastic gradient-based optimizer) and 'lbfgs' (a quasi-Newton optimizer). 4) learning_rate_init: Tests different initial learning rates for weight updates (0.001, 0.01).

3. Results and Discussion

Table 1. The performance of Decision Tree

Class	precision	recall	f1-score	support
0	0.51	0.59	0.55	138
1	0.6	0.52	0.56	164
accuracy	0.55	0.55	0.55	302
macro avg	0.56	0.56	0.55	302
weighted avg	0.56	0.55	0.55	302

Table 2. The performance of Random Forest

Class	precision	recall	f1-score	support
0	0.32	0.32	0.32	138
1	0.43	0.44	0.44	164
accuracy	0.38	0.38	0.38	302
macro avg	0.38	0.38	0.38	302
weighted avg	0.38	0.38	0.38	302

Table 3. The performance of MLP

Class	precision	recall	f1-score	support
0	0.32	0.32	0.32	138
1	0.43	0.43	0.43	164
accuracy	0.38	0.38	0.38	302
macro avg	0.38	0.38	0.38	302
weighted avg	0.38	0.38	0.38	302

Table 4. The performance of SVM

Class	precision	recall	f1-score	support
0	0.26	0.26	0.26	138
1	0.37	0.36	0.36	164
accuracy	0.31	0.31	0.31	302
macro avg	0.31	0.31	0.31	302
weighted avg	0.32	0.31	0.32	302

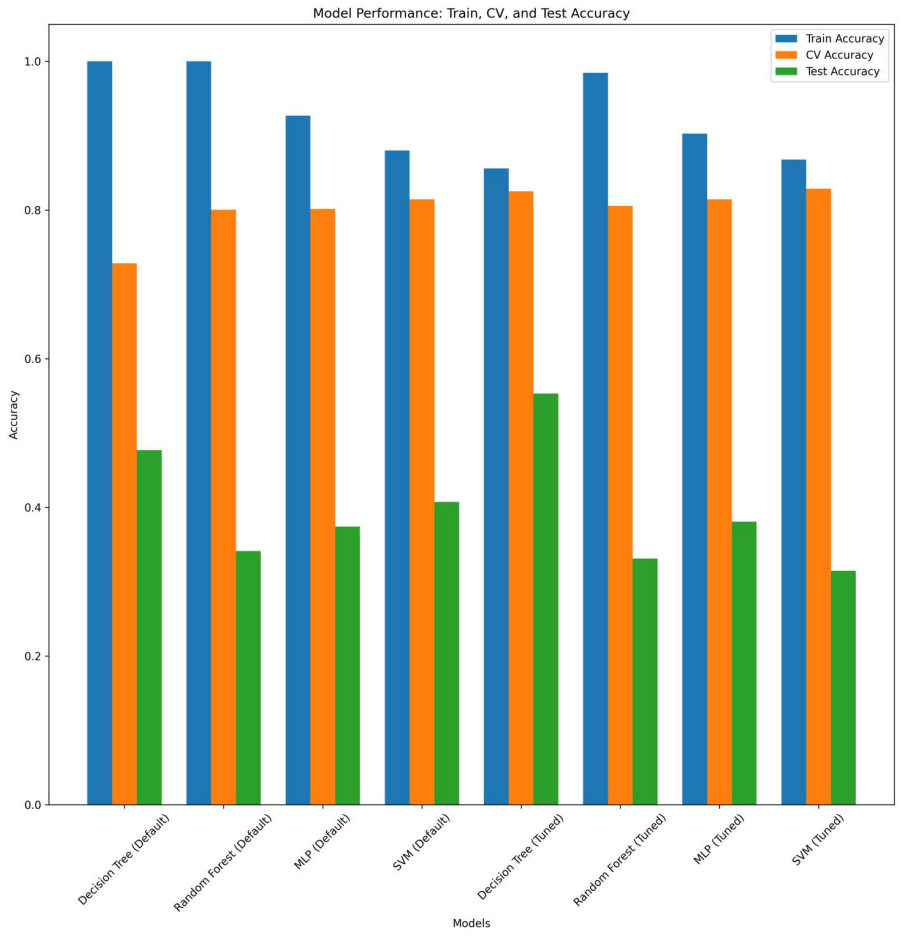

Fig. 2 Model Performance (Photo/Picture credit: Original)

This section discusses in detail how each model performs with default and tuning parameters, and performs horizontal (between different models) and vertical (between the same model and different data sets) comparative analysis. Table 1, Table 2, Table 3, Table 4 and Fig. 2 provide the performance of models. In addition, theoretical expectations are compared with actual results to reveal differences in model performance and possible causes.

1) Horizontal Comparison between different models. Under the default parameter settings, the test set accuracy of each model is significantly different. Specifically, the decision tree model performed best, with an accuracy of 0.48 on the test set, higher than the other models. In contrast, the random forest and multi-layer Perceptron (MLP) models have test set accuracy of 0.34 and 0.37, which are close but lower than decision trees. The support vector machine (SVM) model performed in the middle, with an accuracy of 0.41 on the test set.

After tuning, the performance of each model on the test set is different, and the accuracy of the decision tree model is significantly improved to 0.55, which is excellent. Random forest model and MLP also have a certain degree of improvement, but the improvement is small. Surprisingly, after SVM model tuning, the accuracy on the test set dropped to 0.31, indicating that the tuning effect was not good, which may introduce more serious overfitting problems.

2) Vertical Comparison of the same model on different data sets. The accuracy of the training set in the 5-fold cross-validation of the decision tree model is 82.5%, and the accuracy of the test set is significantly different from that of the test set, which is only 0.55, indicating obvious overfitting. A similar situation occurs in the random forest model. The accuracy of the training set is 80.7% in the 50-fold cross-validation, and the accuracy of the test set is lower, 0.38.

The performance of MLP model is relatively stable after default parameters and tuning. Under the default parameters, the accuracy of MLP on the training set is high, but the accuracy of MLP on the test set is mediocre, which indicates that the model may have some underfitting problems. After tuning, the accuracy of the MLP model on the training set and the test set did not change much.

After the SVM model is tuned, the accuracy of the training

set is improved, but the accuracy of the test set is reduced from 0.41 under the default parameters to 0.31, which indicates that the optimized model overfits the training data, resulting in a decline in generalization ability and cannot be effectively applied on the previously unseen data.

3) Comparison with Theoretical Expectations. Based on theoretical expectations, SVM and MLP models generally perform well in terms of generalization across data sets. However, from the actual results, the performance of these two models on the test set is relatively average, even lower than the decision tree and random forest models. For SVM, the performance after tuning further declines, which is contrary to the robustness typically shown on structured data, possibly due to the introduction of too much complexity in the tuning process, resulting in overfitting of the model. In addition, the performance of MLP models failed to meet expectations, possibly due to the limited help of mesh search to improve the fitting ability of MLP models.

In contrast, the decision tree model is simple, but its performance on the test set is better than other models after tuning. This is related to its high ability to fit training data. After proper pruning and parameter tuning, the ability to predict new data can be improved while overfitting can be reduced. As an integrated method, random forest model should be stable in theory on different data sets, but from the actual results, its overfitting of training data is still a problem that needs attention, and the improvement effect after tuning is limited.

4. Discussion

Three sets of unexpected contradictions emerged in this experiment. First, decision trees are not only more accurate than random forests, but also perform best among all methods. This may be related to the small size of the data set, and a single decision tree may be sufficient to capture patterns in the data, resulting in a random forest that is less integrated than a single decision tree.

Second, adjusting the grid search in the training set does not necessarily improve the accuracy of the prediction set. The accuracy of training on the training set is not positively correlated with the accuracy of testing on the test set. This may be due to differences in the distribution of the two data sets. Therefore, when making predictions across data sets, training based on the original data set needs to be carefully handled, and mesh search tuning cannot be done using accuracy as the only criterion.

Finally, the accuracy of almost all methods in a binary classification problem is less than 50%, and the performance is similar to that of random guessing. This result highlights the potential value of Transfer Learning and Domain Adaptation techniques to address these challenges.

Transfer Learning can be particularly useful when the datasets used for training and testing have different distributions, as seen in this experiment. Instead of training models from scratch on a new dataset, transfer learning allows the model to leverage knowledge from a pre-trained model on a similar but distinct dataset. This approach can significantly reduce the need for large amounts of labeled data and help the model generalize better to new, unseen data. For example, a deep learning model pre-trained on a large-scale cardiovascular dataset could be fine-tuned on a smaller, localized dataset, thus retaining the core features learned from the source domain while adapting to the specific nuances of the target domain.

Domain Adaptation, a subfield of transfer learning, focuses on adapting models to work well across different domains, especially when there is a shift in data distribution between the training and testing phases. In the context of heart disease prediction, domain adaptation techniques, such as domain adversarial training or domain-invariant feature learning, could help models generalize better across different patient populations or healthcare systems. These methods aim to reduce the discrepancy between the source and target domain data distributions, ensuring that the model learns features that are robust across different datasets.

The future potential of transfer learning and domain adaptation in heart disease prediction is substantial. By leveraging these techniques, researchers can develop more robust and adaptable AI models that perform well across diverse datasets, improving the accuracy and reliability of heart disease diagnosis and prognosis across various populations and healthcare environments.

5. Conclusion

Heart disease remains a major global health concern, underscoring the need for early and accurate diagnosis methods, as highlighted in the introduction. This study aimed to enhance the prediction of heart disease using machine learning and deep learning models by evaluating their performance on different datasets with varying distributions. By using Decision Tree, Random Forest, SVM, and MLP classifiers, this study compared the models' training and prediction accuracy, particularly focusing on cross-dataset generalization. The results revealed that while Decision Tree models achieved the highest accuracy after hyperparameter tuning, other models like Random Forest and MLP showed limited improvement. Surprisingly, the SVM model's performance decreased after tuning, indicating potential overfitting issues. When performing grid search across datasets, targeting basic metrics like accuracy or

F1 score often leads to suboptimal results. These findings suggest that model performance can vary significantly depending on dataset characteristics and hyperparameter optimization, highlighting the challenges of achieving generalization in real-world applications.

A limitation of this study is the use of only two datasets with relatively small sizes, which may not fully capture the complexities of real-world medical data. Furthermore, the differences in feature distributions between the datasets could have contributed to the models' inconsistent performance. Future work could explore the use of Transfer Learning and Domain Adaptation techniques to enhance model generalization across diverse datasets. Additionally, incorporating larger and more diverse datasets could provide a more comprehensive understanding of model behavior and improve the robustness of AI-based heart disease prediction systems.

References

[1] Joachims T. Training linear SVMs in linear time. InProceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining 2006 Aug 20 (pp. 217-226).

[2] Tang Y, Zhang YQ, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2008 Dec 9;39(1):281-8.

[3] Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. Advances in neural information processing systems. 2000;13.

[4] Subramani S, Varshney N, Anand MV, Soudagar ME, Al-Keridis LA, Upadhyay TK, Alshammari N, Saeed M, Subramanian K, Anbarasu K, Rohini K. Cardiovascular diseases prediction by machine learning incorporation with deep learning. Frontiers in medicine. 2023 Apr 17;10:1150933.

[5] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Medical image analysis. 2019 Dec 1;58:101552.

[6] Singh NK, Raza K. Medical image generation using generative adversarial networks: A review. Health informatics: A computational perspective in healthcare. 2021:77-96.

[7] Kaggle. Heart Disease Dataset. 2024. https://www.kaggle.com/datasets/mexwell/heart-disease-dataset

[8] Kaggle. Heart Disease Dataset. 2021. https://www.kaggle.com/datasets/yasserh/heart-disease-dataset

[9] Raczko E, Zagajewski B. Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. European Journal of Remote Sensing. 2017 Jan 1;50(1):144-54.

[10] Zhang C, Liu Y, Tie N. Forest Land Resource Information Acquisition with Sentinel-2 Image Utilizing Support Vector Machine, K-Nearest Neighbor, Random Forest, Decision Trees and Multi-Layer Perceptron. Forests. 2023 Jan 29;14(2):254.