Heart Disease Prediction

A Comparative Study of Machine Learning Techniques and Optimizations

07 April 2025

Emily Kiddle
Jamie MacDonald
Gabriel Rodriguez

Winter 2025 CPSC 544 - Machine Learning

UNIVERSITY OF CALGARY

Engaging activity Here (i was wondering if we have two samples form our dataset, and like we give people the symptom profile and we ask them if the patient may have heart disease or not? then we compare the actual ML outputs on that symptom profile.

Engaging activity time !!
(i was also thinking of not telling them the answer until the end of the presentation so attention remains throughout our presentation)

Introduction and Background

What is this about?

Introduction and Background

- We aim to build a machine learning model that **predicts heart disease from patient data.**
- Heart disease is difficult to diagnose early because symptoms often overlap with other conditions.
- By using a combination of datasets and several ML models, we can **test which approaches are most accurate.**
- The dataset includes 920 samples from four different hospitals, each with 14 clinical features.
- This project explores whether machine learning can improve early detection and support doctors in real decisions.

Motivation

- Heart disease is one of the leading causes of death.
- Early prediction can improve patient outcomes and lower costs.
- Many patients are misdiagnosed due to symptom overlap with other conditions.
- Machine learning can support faster and more accurate diagnoses.
- Our goal is to investigate which models are most effective and reliable for this task.

Related Works

- Previous studies found **Random Forest** and deep learning models **perform well** in disease prediction.
- Studies on LightGBM and XGBoost showed high accuracy but require more computation.
- Our work expands on these findings by using the full dataset and testing model improvements with PCA and tuning.
- We provide a detailed comparison across a wide range of models, not just one or two.

Methodology

What did we do?

Preprocessing the Dataset

- Our dataset had 920 samples and 14 features.
- We dropped three columns with over 10% missing values, filled missing values using median imputation, and applied z-score standardization to normalize the data.
- We also binarized the target variable into 0 (no disease) and 1 (disease). This helped simplify the classification task and addressed the imbalance across the original multi-class labels.
- We used PCA to reduce the number of features from 10 to 8 while keeping 85% of the variance, removing redundancy and noise in the data and improving model performance and generalization.

Preprocessing the Dataset (continued)

- Class Imbalance, visualized:

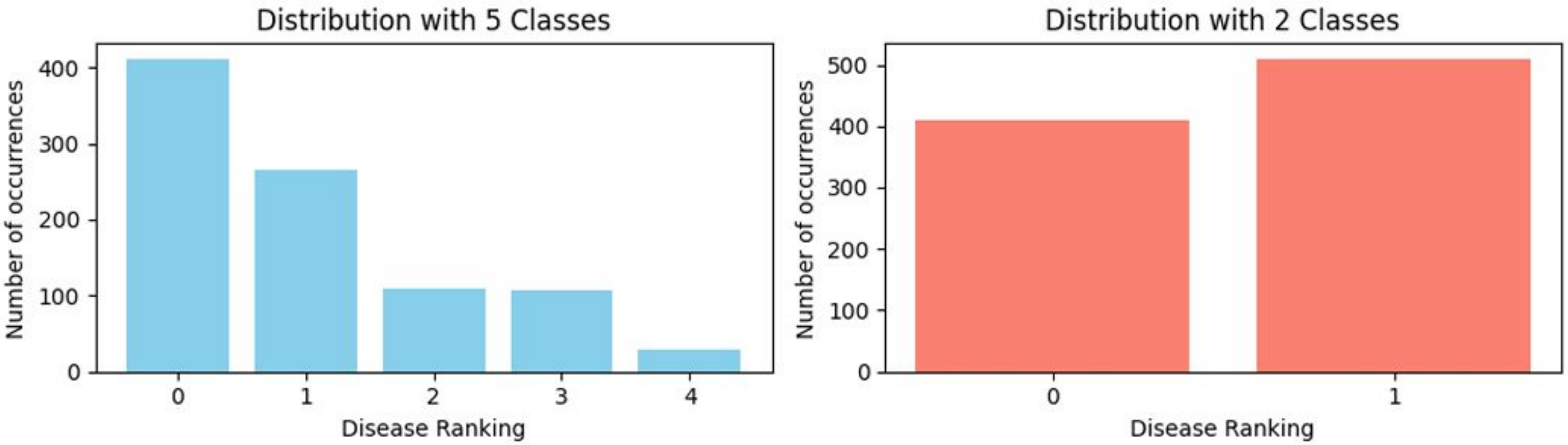

Fig. 1. Distribution of target class variables from 0 (no heart disease) to either a ranking of 1-4 with heart disease (left), or 1 (presence of heart disease, right)

Models and Tools

- We tested 11 models in total, including Random Forest, Extra Trees, Gradient Boosting, AdaBoost, Logistic Regression, SVC, and KNN.
- To explore the benefits of ensemble methods, we added Voting and Bagging Classifiers to the comparison.
- Hyperparameter tuning was done using GridSearchCV, where we adjusted settings such as tree depth, number of estimators, and learning rate.
- Model performance was assessed using accuracy, precision, recall, and F1-score, providing a well-rounded evaluation.
- All analysis and visualizations were carried out in Python using Scikit-learn, Matplotlib, and Seaborn.

Results

How did it turn out?

Results (Pre-PCA)

- Initial results showed that Extra Trees and the Voting Classifier achieved the highest test accuracies, both above 83%.
- Although Random Forest also performed well, it showed signs of overfitting with perfect training accuracy.
- Models like Logistic Regression and KNN demonstrated lower test scores but had better generalization due to simpler structures.
- These results revealed a clear trade-off between model complexity and real-world reliability.

Results (Post-PCA)

- After applying PCA, Random Forest showed improvement, with test accuracy increasing to 84.2%. SVC and KNN also benefited significantly, gaining over 8% and 7.6% in test accuracy respectively.
- These improvements suggest that PCA helped remove noise and redundant information that certain models are more sensitive to.
- In contrast, ensemble models like Extra Trees experienced a slight drop, likely because they already manage feature complexity well internally.
- All in all, PCA improved generalization and boosted performance in models that initially struggled.

Results (Pre-PCA)

Model	Accuracy	Recall	Precision	F1-Score	Training Accuracy
Random Forest	0.809783	0.809783	0.809783	0.809783	1.000000
Gradient Boosting	0.815217	0.815217	0.815217	0.815217	0.907609
AdaBoost	0.793478	0.793478	0.793478	0.793478	0.854620
Extra Trees	0.836957	0.836957	0.836957	0.836957	1.000000
Logistic Regression	0.820652	0.820652	0.820652	0.820652	0.805707
SVC	0.739130	0.739130	0.739130	0.739130	0.722826
Decision Tree	0.733696	0.733696	0.733696	0.733696	1.000000
K-Nearest Neighbors	0.739130	0.739130	0.739130	0.739130	0.789402
K-Nearest Neighbors (Manhattan)	0.744565	0.744565	0.744565	0.744565	0.793478
Voting Classifier	0.831522	0.831522	0.831522	0.831522	0.938859
Bagging Classifier	0.793478	0.793478	0.793478	0.793478	1.000000

Table 1. Initial models and performance on binary data.

Results (Post-PCA)

Model	Accuracy	Recall	Precision	F1-Score	Training Accuracy
Random Forest	0.842391	0.842391	0.842391	0.842391	1.000000
Gradient Boosting	0.820652	0.820652	0.820652	0.820652	0.918919
AdaBoost	0.804348	0.804348	0.804348	0.804348	0.875921
Extra Trees	0.798913	0.798913	0.798913	0.798913	1.000000
Logistic Regression	0.804348	0.804348	0.804348	0.804348	0.799754
SVC	0.820652	0.820652	0.820652	0.820652	0.842752
Decision Tree	0.733696	0.733696	0.733696	0.733696	1.000000
K-Nearest Neighbors	0.815217	0.815217	0.815217	0.815217	0.857494
K-Nearest Neighbors (Manhattan)	0.771739	0.771739	0.771739	0.771739	0.857494
Voting Classifier	0.826087	0.826087	0.826087	0.826087	0.934889
Bagging Classifier	0.809783	0.809783	0.809783	0.809783	1.000000

Table 2. Models and performance following principal component analysis (PCA) maintaining 85% variance.

Discussion of Results

What did we learn?

Discussion of Results

- **Random Forest and Voting Classifier emerged as the strongest models** due to their ability to combine multiple decision paths and reduce variance.
- We found that PCA made the biggest impact on models like SVC and KNN, which are more easily affected by noise in the data.
- Overfitting was common in deeper tree models, especially those left untuned, but this was addressed through parameter optimization. Despite the accuracy trade-off, we prioritized generalization, which is more important in real clinical use.
- From all evaluations, Random Forest stood out as the most balanced model, offering high accuracy, interpretability, and efficiency.

Future Work

- We want to explore better scaling techniques to address skewed feature distributions.
- Adding more patient data from different sources could help improve the model's robustness and fairness.
- It would also be valuable to test these models in clinical environments to understand how they perform in real-world decision-making. Furthermore, incorporating patient history or time-based features could support more personalized and dynamic predictions.
- We're also interested in exploring hybrid ensemble methods and deep learning models that may capture more complex patterns.

Engaging activity time !!
ANSWER!!!

References

[1] Rayan Alanazi. 2022. Identification and Prediction of Chronic Diseases Using Machine Learning Approach. Journal of healthcare engineering (Feb. 2022). doi:2826127

[2] Jonathan Asanjarani. 2025. Sex-Specific and Regional Analysis of Heart Disease Prediction Using Machine Learning Algorithms: Insights from the UCI Irvine Public Heart Disease Datasets (Cleveland and Long Beach). (2025). https://hdprediction.commons.gc.cuny.edu/

[3] Homin Lee Young-Jin Kim Yeongsic Kim Young Hoon Park Dong Jin Park, Min Woo Park. 2021. Development of machine learning model for diagnostic disease prediction based on laboratory tests. Scientific Reports 11 (April 2021), 7567. doi:10.1038/s41598-021-87171-5

[4] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. 2019. Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. BMC Medical Informatics and Decision Making 19 (2019). doi:10.1186/s12911-019-1004-8

Thank you for sitting in to our presentation!

07 April 2025

Emily Kiddle
Jamie MacDonald
Gabriel Rodriguez

UNIVERSITY OF CALGARY