**Abstract**

Heart Disease is a leading cause of death and morbidity globally, early detection is a good start to preventing this. Diagnostic technology has been increasingly used, with Healthcare Artificial Intelligence models being a tool.

The trustworthiness of these tools is important, as users must understand how reliable, reproducible, and fair the diagnosis are.

Using a heart failure prediction dataset publicly available on Kaggle, this project trains 5 models and evaluates them on fairness, performance and explainability. Adaboost, Random Forest, Support Vector Matrix, K-Nearest Neighbour, and a linear Neural Network are models with varying complexities and implementation making the comparison fair. This dataset included sensitive demographic features, age and sex, which shall be stratified.

The best performance by traditional metrics

**Introduction**

**Related Works**

**AI in Healthcare**

**The Heart Failure Prediction Dataset**

This dataset is accessible through a free account on Kaggle, the open-source machine learning platform, and can be used with appropriate citation. It combined 5 independent popular heart disease datasets across 11 common features [1], and a single label io the heart disease or normal, loosing the type of heart disease data in the process. A study from 1989 [2] was the source of 4 out of 5 of the datasets, with 1 dataset from 303 patients at Cleveland Clinic being the reference group. Other datasets were the test groups 200 patients from Long Beach California, 425 Budapest, Hungary and 143 from 2 Switzerland Hospitals. The last dataset curated was the Statlog heart disease database [3] with

All datasets had 13 features, but the author removed the number of major vessels an thal due to missing values within he first paper [4], **table 1** shows the final features. There was a total of 918 observations after the author removed duplicate data.

These were gotten from:

Table 1: Dataset Features and Descriptions

|  |  |  |
| --- | --- | --- |
| Features | Description | Feature Type |
| Age | age of the patient [years] | Numerical |
| Sex | sex of the patient [M: Male, F: Female] | Categorical |
| ChestPainType | chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic] | Categorical |
| RestingBP | resting blood pressure [mm Hg] | Numerical |
| Cholesterol | serum cholesterol [mm/dl] | Numerical |
| FastingBS | fasting blood sugar  [1: if FastingBS > 120 mg/dl, 0: otherwise] | Categorical |
| RestingECG | resting electrocardiogram results [Norma, ST: having ST-T wave abnormality, LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria] | Categorical |
| MaxHR | maximum heart rate achieved  [Numeric value between 60 and 202] | Numerical |
| ExerciseAngina | exercise-induced angina [Y: Yes, N: No] | Categorical |
| Oldpeak | oldpeak = ST [measured in depression] | Numerical |
| ST\_Slope | the slope of the peak exercise ST segment [Up: up, Flat: flat, Down: down] | Categorical |
| HeartDisease | output class [1: heart disease, 0: Normal] | Categorical |

Although this dataset was created less than 3 years ago, the data collation and observations are over 2 decades old. The reliability of this with a change of time would need to be addressed. While there is some variation in location of the data collected, more demographic data is desired to avoid perpertuating historical biases.

**Responsible AI**

Responsibility refers to the considerations taken to increase the trustworthiness of the tool in healthcare when developing the model. This project follows the fairness, universality, traceability, usability, robustness, and explainability (FUTURE-AI) guideline [5] created to increase the trustworthiness and ethics of healthcare AI tools. Fairness as it relates to equity and robustness in terms of reliable outputs independent of subgroups. This means the performance of the model should be sensitive feature agnostic, and that variability in data like outliers should not affect performance. Explainability is also an important factor for diagnosis tools.

This de-identified dataset came with some sensitive demographic information, age, and sex, and the experimental design was developed to mitigate the exacerbation of existing biases [6], [7]. Although the dataset only considers certain demographics in the United States of America, Switzerland, and Hungary. Heatmaps will be used to display feature importance, and the model limitations will be outlined.

**Materials and Methodology**

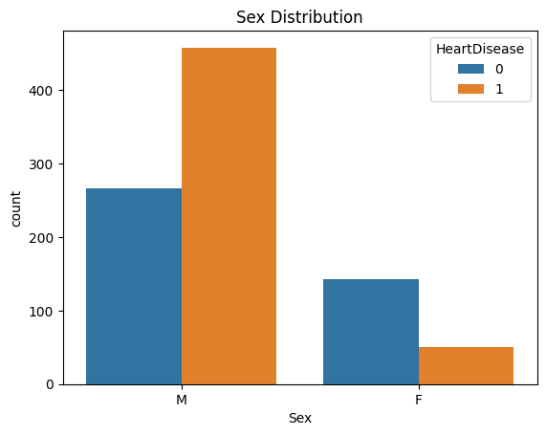
This section highlights the findings from the EDA, the reasoning behind the experimental setup, and the evaluation criteria used to determine the performance and fairness of the models. Python was used to develop this detection algorithm particularly the pandas and sci-kit learn packages for development and metrics. To evaluate the fairness of the model, Fairlearn [8] will be used, this is built with sci-kit learn as a base ensuring environment compatibility.

**Exploratory Data Analysis and Data Prep**

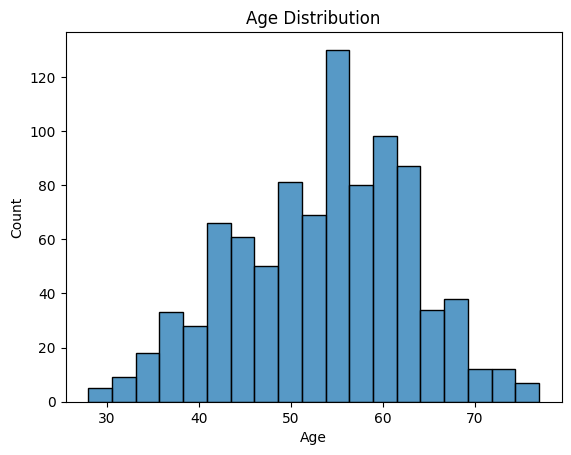
EDA was performed to understand the feature characteristics and develop appropriate preprocessing methods. No missing data was found in the 918 observations. Due to the limited number of features and the strong correlation between the features and labels all 11 features will be used for development.

The numerical features have a data type that is either integer or float values. For categorical features, some related to binary integer values, and others were object data types with qualitative features. Sex, ChestPainType, RestingECG, ExerciseAngina, and ST\_Slope are qualitative features with an object data type. Label encoding shall be applied to update the qualitative categories to numerical. Categorical qualitative features were used to stratify the label data to better understand their characteristics.

More observations with asymptomatic (ASY) chest pain type are likely to have heart disease, compared to the opposite behaviour for atypical (ATA), similar to observations having exercise-induced angina positive predictions. The dataset was imbalanced, across the sensitive features of age and sex with more observations from Males and patients over 40 as seen in Figures 4 and 5, with over 70% of the observations being of men. Stratification was the method chosen to combat this which is used in literature for medical datasets of this modality [5], [6].



**Fig. 4.** Stratified histogram of the label by Sex category.



**Fig. 5.** Histogram shows the distribution of Age feature values.

For Numerical feature types, a detailed statistical analysis was performed for features with numerical values, Table 1 shows some of the results.

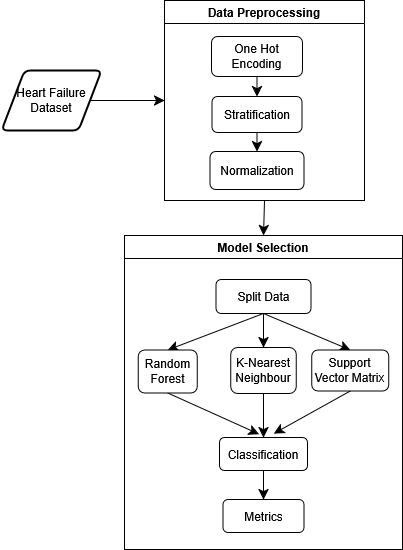
**Table 1**: Statistical Analysis Results for Numerical Features

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature** | **mean** | **std** | **min** | **50%** | **max** |
| **Age** | 53.51 | 9.433 | 28 | 54 | 77 |
| **RestingBP** | 132.4 | 18.51 | 0 | 130 | 200 |
| **Cholesterol** | 198.8 | 109.4 | 0 | 223 | 603 |
| **FastingBS** | 0.2331 | 0.423 | 0 | 0 | 1 |
| **MaxHR** | 136.8 | 25.46 | 60 | 138 | 202 |
| **Oldpeak** | 0.8874 | 1.067 | -2.6 | 0.6 | 6.2 |
| **HeartDisease** | 0.5534 | 0.497 | 0 | 1 | 1 |

An important point to consider is outliers in the dataset, which was calculated using the z-score, 1 observation was removed due to the impossible RestingBP reading of 0. Other outliers were kept because the difference between anomalies and cases the model should not misclassify was determined. Data was scaled using the RobustScaler, which uses the interquartile range as a scaling factor to reduce the influence of existing outliers.

**Experimental Setup**

A traditional machine learning approach is taken, with an 80% development and 20% testing dataset split. This avoids data leakage by keeping the development and test set separate; the split is done with no randomness to ensure reproducibility. The development set was split into training and validation using a stratified k-fold cross-validation method to ensure a balanced dataset within each fold, and similar results across folds [9].



**Fig. 2.** Proposed Framework for Heart Failure Prediction.

A total of 8 models were developed and will be compared in this project. These models were chosen based on data characteristics, and performance in testing. Figure 2 is the flowchart of the experimental design.

Explainable methods will be applied to the best performing models using SHAP values

**Evaluation**

The evaluation criteria chosen are based on the nature of the task, and the metrics recommended to measure performance that was used in similar diagnostic performance classification models from the literature [10], [11], [5]. The number of heart disease predictions that are right (True Positive), the number of wrong positive predictions (False Positive), the number of correct normal predictions (True Negative), and the number of false normal predictions (False Negative) will be a basis for evaluation.

Recall or sensitivity value is the ratio of the positive values correctly predicted and the total positive outcomes. This is also known as the True Positive Rate and can be stratified across sensitive features to find the equal opportunity [6] fairness metric:

|  |  |
| --- | --- |
|  | **(1)** |

Precision describes the reliability of the model, also known as the positive parity.

|  |  |
| --- | --- |
|  | **(2)** |

Specificity, or Negative Predictive Value is the ratio of the True negative values and the total negative outcomes, it describes the reliability of the model.

|  |  |
| --- | --- |
|  | **(3)** |

The F1 score is a weighted average of precision and recall values, and is a popular metric for classification problems [5]:

|  |  |
| --- | --- |
|  | **(4)** |

The confusion matrix for the True Positives and False Negatives of the predicted and target values will be created to capture overall errors.

Demographic parity requires equal decision rates between subgroups, and Equal odds will be applied to the sensitive features to evaluate fairness [6].

|  |  |
| --- | --- |
|  | **(5)** |

**3.2. Reproducibility**

The dataset used is open source and is easily accessible on the Kaggle Platform with an account. Code will be hosted on a GitHub repository that is publicly accessible so interested users can run it on platforms like Google Colab or clone the repository to make improvements. A guide on how to use the repository shall be created within the wiki [12].

**Results**

Performance for final fold value

* Before bias mitigation

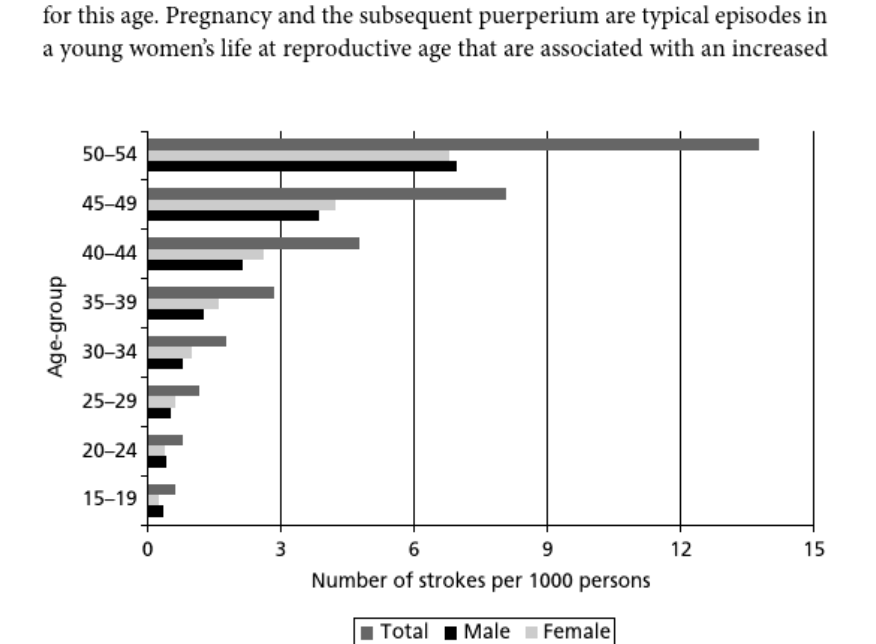
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | SVM | KNN | Random Forest | AdaBoost |
| Recall/ Sensitivity |  |  |  |  |
| Precision |  |  |  |  |
| F1-score |  |  |  |  |
| Demographic parity |  |  |  |  |
| Equal Odds/Opportunity |  |  |  |  |

* After bias mitigation

Images and Graphs

Stratified

* Fairness model plot
* Stratified accuracy, recall
* Combine sensitive features across the F1 Score
* Final statistics being used



**Analysis**

**Limitations**

**Conclusion and Future considerations**

**References**

[1] fedesoriano, “Heart Failure Prediction Dataset.” Kaggle, https://www.kaggle.com/fedesoriano/heart-failure-prediction, Sep. 2021. Accessed: Feb. 08, 2025. [Online]. Available: https://www.kaggle.com/fedesoriano/heart-failure-prediction

[2] R. Detrano *et al.*, “International application of a new probability algorithm for the diagnosis of coronary artery disease,” *The American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, Aug. 1989, doi: 10.1016/0002-9149(89)90524-9.

[3] Unknown, “Statlog (Heart).” UCI Machine Learning Repository. doi: 10.24432/C57303.

[4] W. S. Andras Janosi, “Heart Disease.” UCI Machine Learning Repository, 1989. doi: 10.24432/C52P4X.

[5] K. Lekadir *et al.*, “FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare,” *BMJ*, vol. 388, p. e081554, Feb. 2025, doi: 10.1136/bmj-2024-081554.

[6] Q. Feng, M. Du, N. Zou, and X. Hu, “Fair Machine Learning in Healthcare: A Review,” Feb. 01, 2024, *arXiv*: arXiv:2206.14397. doi: 10.48550/arXiv.2206.14397.

[7] J. E. Alderman *et al.*, “Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations,” *The Lancet Digital Health*, vol. 7, no. 1, pp. e64–e88, Jan. 2025, doi: 10.1016/S2589-7500(24)00224-3.

[8] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, “Fairlearn: Assessing and Improving Fairness of AI Systems,” *Journal of Machine Learning Research*, vol. 24, no. 257, pp. 1–8, 2023.

[9] M. T r, V. K. V, D. K. V, O. Geman, M. Margala, and M. Guduri, “The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification,” *Healthcare Analytics*, vol. 4, p. 100247, Dec. 2023, doi: 10.1016/j.health.2023.100247.

[10] D. K. Plati *et al.*, “Machine Learning Techniques for Predicting and Managing Heart Failure,” in *Predicting Heart Failure*, John Wiley & Sons, Ltd, 2022, pp. 189–226. doi: 10.1002/9781119813040.ch9.

[11] D. K. Lutfi and G. F. Shidik, “Improvement Heart Failure Prediction Using Binary Preprocessing,” in *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, Nov. 2023, pp. 236–241. doi: 10.1109/ICAMIMIA60881.2023.10427846.

[12] P. Vandewalle, J. Kovacevic, and M. Vetterli, “Reproducible research in signal processing,” *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 37–47, May 2009, doi: 10.1109/MSP.2009.932122.