Prediction of Heart Failure Using Machine Learning Models

Nigel D'Costa
nlnu3@hawk.iit.edu
A20536696
Illinois Institue of Technology
Chicago, Illinois, USA

Sandeep Preetham
spreetham@hawk.iit.edu
A20538521
Illinois Institue of Technology
Chicago, Illinois, USA

Mehak Mathur
mmathur@hawk.iit.edu
A20554928
Illinois Institue of Technology
Chicago, Illinois, USA

ABSTRACT

Cardiovascular diseases (CVDs) are the leading cause of mortality globally, with heart failure representing a significant health threat within this spectrum. In light of these challenges, our research presents a novel machine learning-based approach aimed at early detection and prediction of heart failure. We have developed a series of predictive models using a dataset enriched with key attributes known to influence heart health, such as demographic details, clinical parameters, and physiological indicators.

By implementing algorithms such as Support Vector Classifier, Random Forest, Decision Tree, K-Nearest Neighbors, Logistic Regression, and Convolutional Neural Networks, we have established a comparative framework to assess each model's performance. This paper details our data preprocessing methods, including normalization and feature selection, our exploratory data analysis which unearthed critical insights into feature distributions and their correlation with heart failure, and our robust testing strategy to evaluate model efficacy.

Our findings demonstrate the potential of machine learning models in transforming cardiovascular healthcare by providing accurate, early predictions of heart failure, thereby paving the way for timely intervention and improved patient outcomes. With the goal of reducing the global burden of CVDs, our study contributes a scalable solution that promises to enhance preventive care and inform clinical decision-making processes.

KEYWORDS

Heart Failure, Machine Learning, Cardiovascular, Predictive, Data Preprocessing, Exploratory, Classification, Features, Comparison, Metrics, Classifier, Forest, Decision, Neighbors, Regression, Neural Networks, Clinical, Informatics, Management, Risk, Preventative, Public Health, Outcomes, Angina, Electrocardiogram, ST_Slope, Blood Sugar, Cholesterol, Heart Rate, Gender, Disparities, CVD.

1 INTRODUCTION

Amidst an abundance of medical information and the surge of Data Science, numerous of startups are rising to do the challenge of developing indicators for potential diseases. Cardiovascular diseases (CVDs) stand as the primary global cause of mortality, claiming an approximately of 17.9 million lives annually, representing 31% of all global deaths.

Heart failure it is a prevalent consequence of CVDs, necessitates early detection and management for individuals with cardiovascular disease or those at elevated cardiovascular risk due to factors such as hypertension, diabetes, hyperlipidemia, or existing illnesses. In this context, machine learning models[3] offer significant potential for assistance which will help in detecting the heart failure easily.

2 PROBLEM STATEMENT

Cardiovascular diseases (CVDs)[4] emerge as the primary culprit of mortality on a global scale, claiming an estimated 17.9 million[2] lives annually. This staggering figure accounts for roughly 31% of deaths worldwide, highlighting a critical health crisis that spans across continents and communities. The complexity and severity of CVDs demand a comprehensive approach to healthcare, one that harnesses the potential of the ever-expanding reservoir of medical data and the innovative solutions offered by Data Science.

As we delve into the era where technology meets healthcare, a multitude of startups and health-tech companies are stepping up to the challenge. They are dedicated to creating sophisticated predictive models that not only pinpoint potential health risks but also offer a proactive framework for disease management. Heart failure, a prevalent outcome of various cardiovascular conditions, stands as a stark emblem of the health risks associated with CVDs. It represents a substantial threat to public health, necessitating early detection and timely intervention to mitigate its impact.

Individuals diagnosed with or predisposed to cardiovascular conditions—owing to risk factors such as hypertension, diabetes, hyperlipidemia, or pre-existing health issues—find themselves on the front lines of this battle. It is for these individuals that advancements in machine learning (ML) hold great promise. By leveraging intricate algorithms and deep learning models, ML has the capability to sift through complex datasets to identify early signs of heart failure.

Our project aims to harness the power of ML to aid with the detection, diagnosis, and management of heart failure. With a rich dataset detailing numerous attributes associated with heart conditions, we employ ML techniques to build models that can accurately classify and predict the likelihood of heart failure among patients.

This endeavor not only serves as a testament to the transformative power of machine learning in healthcare but also sets the stage for addressing broader challenges within the domain of public health. By automating key processes in disease detection and management, we are paving the way towards a future where the burden of cardiovascular diseases is significantly reduced. Our ultimate goal with this project is to provide a reliable, scalable solution for predicting heart failure, offering a beacon of hope for millions at risk and crafting a new paradigm in the preventative care landscape.

3 RELATED WORK

There is abundant work related to the field of cardiovascular disease for Heart failure prediction) that has garnered significant attention recently due to its potential to enhance patient care and improve healthcare resource allocation. Numerous studies have explored the application of machine learning and deep learning techniques in predicting the onset and progression of heart failure. Various techniques and models have been implemented for the Heart Failure Prediction such as Support Vector Classifier, Decision tree classifer Random Forest classifier as well as other algorithms such as Gradient Boosting. We plan to utilise a few of these algorithms as well as a few more Deep Learning algorithms to further enhance our analysis.

4 PROPOSED SOLUTION

In order to achieve our goal of Heart Disease Prediction, we will be using the following models:

(1) Support Vector Classifier (SVC): Known for its effectiveness in high-dimensional spaces, the SVC will be used to find the hyperplane that best separates cases of heart failure from non-failure cases.
(2) Random Forest Classifier: This ensemble method, which operates by building multiple decision trees and merging their outputs, offers robustness and accuracy, making it suitable for handling the complex interdependencies within our data.
(3) Decision Tree Classifier: Provides a clear indication of feature importance and is easy to interpret, which is crucial for clinical decision-making processes.
(4) K-Nearest Neighbors (KNN): This algorithm will be utilized for its simplicity and effectiveness in classification by analyzing the proximity of data points in feature space.
(5) Logistic Regression Classifier: A fundamental technique for binary classification problems, offering a probabilistic framework that is easy to implement and interpret.
(6) Convolutional Neural Network (CNN): Although typically used for image data, CNNs can be adapted for sequence data in healthcare, capturing spatial hierarchies in structured data like electronic health records.

Each model will be trained using an 80:20 train-test split to ensure a balanced approach between learning and validation. Model performance will be assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve. This multi-metric evaluation will help identify the most effective model in terms of both predictive power and reliability.

Upon training, the models will undergo a rigorous comparison to determine which algorithm performs best under the conditions specific to our dataset. This step is critical as it will inform the selection of the most appropriate model(s) for deployment in real-world scenarios.

5 DATA DESCRIPTION

The dataset which we have used in this project have various data attributes related the project. To the heart failure cases, for building the machine learning models we are going to predict the likelihood of heart failure in patients. These attributes typically include:

- **Age** : Age of the patient [years]
- **Sex** : Sex of the patient [M: Male, F: Female]
- **ChestPainType** : Chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- **RestingBP** : Resting Blood Pressure [mm Hg]
- **Cholesterol** : Serum cholesterol [mm/dl]
- **FastingBS** : Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- **RestingECG** : Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV),LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- **MaxHR** : maximum heart rate achieved [Numeric value between 60 and 202]
- **ExerciseAngina** : exercise-induced angina [Y: Yes, N: No]
- **Oldpeak** : oldpeak = ST [Numeric value measured in depression]
- **ST_Slope** : the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- **HeartDisease** : output class [1: heart disease, 0: Normal]

These attributes provide valuable information about the patient's demographics, medical history, and physiological characteristics, which can be used to train machine learning models to predict the likelihood of heart failure.

5.1 Data Pre-Processing

Before applying machine learning models, our dataset undergoes a preprocessing phase to ensure optimal model performance. This includes handling missing values through imputation strategies tailored to the nature of the data, encoding categorical variables to transform them into machine-readable formats, and normalizing or standardizing numerical features to reduce bias due to scale differences. Additionally, feature selection techniques are employed to identify and retain the most relevant variables, thereby enhancing model efficiency and reducing overfitting. This preparatory step is critical as it directly influences the accuracy and reliability of our predictive models, setting a strong foundation for effective heart failure prediction.

5.2 Feature Engineering

Feature engineering plays a pivotal role in the success of machine learning models, especially in complex domains like healthcare. In our project on heart failure prediction, we performed tests like Chi-Squared Test and Annova test in order to see which features have the most influence and correlation with the target variable and accordingly, drop features that have little or no correlation

6 EXPLORATORY DATA ANALYSIS

In the initial phase of our project, we conduct a comprehensive exploratory data analysis (EDA) to gain insights into the underlying patterns and distributions within the dataset. This process involves visualizing data through graphs such as histograms, box plots, and scatter plots to understand the distribution of variables and identify any outliers or anomalies. Correlation matrices are also utilized to examine the relationships between different features and their impact on heart failure outcomes. This critical analysis helps in pinpointing significant predictors and assessing the data's structure, which guides the subsequent steps of feature engineering and model selection, ensuring that our predictive models are both accurate and interpretable.

Figure 1 shows the visualisation of the target variable. We see that 44.7% (410 Records) of the records are classified as No Heart Diesease whereas 55.3% (508 records) are classified as having heart disease, showing a relatively balanced dataset.

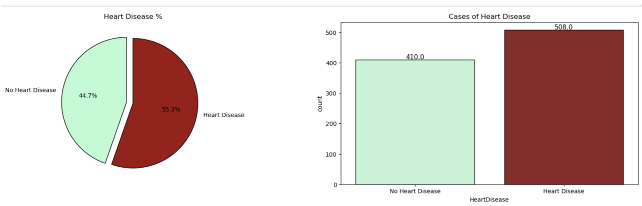

Figure 1: Targert Variable Visualisation

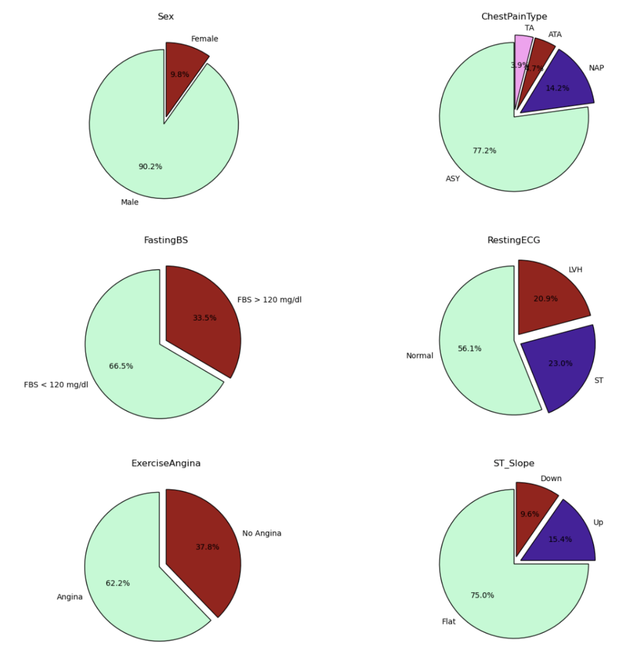

Figure 2: Categorical Features vs Positive Heart Disease Cases

Figure 2 shows the distribution of Categorical Features vs Positive Heart Disease Cases. In our analysis of heart disease patients, it is found that an overwhelming 90% are male. As for the chest pain types observed, the ASY variant is predominant, accounting for 77% of the cases leading to heart diseases. Observations show that a fasting blood sugar level below 120 mg/dl is frequently associated with a higher risk of heart conditions. Regarding RestingECG results, a normal reading is more commonly associated with heart diseases, at a rate of 56%, compared to LVH and ST levels. The presence of exercise-induced angina is also a significant indicator of heart diseases. Lastly, in terms of ST Slope readings, a flat slope is observed in a substantial 75% of cases, which could be key in detecting heart problems.

Figure 3 shows the distribution of Categorical Features vs the Target Variable. The incidence of heart disease is higher among males compared to females, with males making up a larger proportion of heart disease patients. The ASY chest pain type is strongly associated with a higher risk of heart disease. The relationship between fasting blood sugar and heart disease is complex: both those with and without elevated fasting blood sugar levels are represented among heart disease patients. The RestingECG readings do not delineate a specific category that could be linked to heart disease, as all three recorded values show a high prevalence of heart disease patients. The occurrence of exercise-induced angina markedly increases the likelihood of a heart disease diagnosis. As for the ST Slope readings, a flat slope is strongly correlated with a high probability of heart disease diagnosis, while a downward slope also indicates heart disease, albeit represented in a smaller portion of the data.

Figures 4-7 show the distribution of the numerical features vs the target variable Heart Disease.Analysis of the RestingBP data in figure 4 reveals that readings between 95 (19x5) and 170 (34x5) are most commonly associated with heart diseases. In figure 5, Cholesterol levels in the range of 160 (16x10) to 340 (34x10) are observed to be at a higher risk for heart diseases.In Figure 6 Heart diseases appear across the spectrum in MaxHR readings, but a concentration of cases is noted between 70 (14x5) and 180 (36x5) values. Similarly,in figure 7 Oldpeak values ranging from 0 (0x5/10) to 4 (8x5/10) are indicative of a heightened likelihood of heart diseases.

6.1 Numerical features vs Categorical features w.r.t Target variable (HeartDisease) :

Figure 8 shows the distribution of the categorical variable Sex vs the numerical features.The male demographic shows a prevalence of heart diseases across nearly all numerical feature values, particularly for individuals over 50 with positive Old Peak values and a maximum heart rate of less than 140; here, we see a clustering of heart disease cases. Conversely, data points for the female demographic are much fewer when compared to males, making it challenging to identify specific numerical ranges or values with a clear correlation to heart disease occurrences.

Figure 9 shows the distribution of the categorical variable ChestPainType vs the numerical features. The ASY variant of chest pain significantly outnumbers other types in relation to all numerical features.

Figure 10 shows the distribution of the categorical variable FastingBS vs the numerical features. Heart disease cases are prevalent across the dataset in individuals over the age of 50, regardless of whether they have been diagnosed with Fasting Blood Sugar. Elevated Resting BP readings above 100 in conjunction with Fasting Blood Sugar have shown a greater incidence of heart diseases compared to those without elevated fasting glucose levels. Interestingly, the presence of Fasting Blood Sugar does not appear to influence the impact of cholesterol levels in the context of heart diseases. However, patients without elevated Fasting Blood Sugar levels who exhibit a maximum heart rate of less than 130 are at a heightened risk for developing heart conditions.

Figure 11 shows the distribution of the categorical variable RestingECG vs the numerical features.Cases of heart disease are observed with RestingECG readings categorized as Normal, ST,

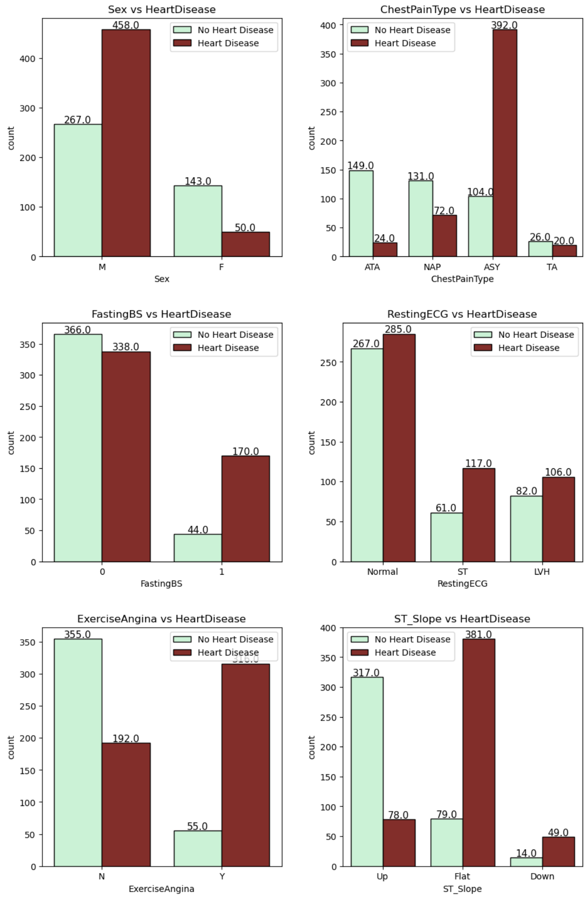

Figure 3: Categorical Features vs Target Variable (HeartDisease)

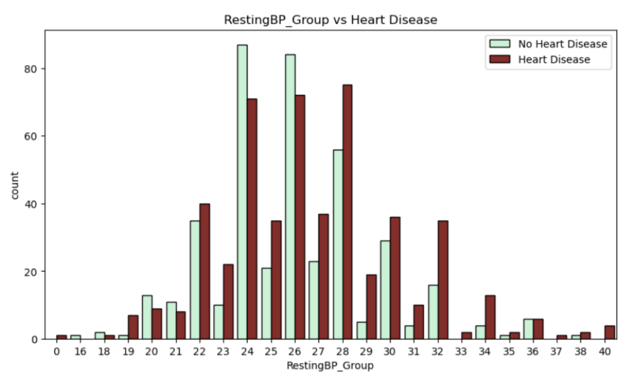

Figure 4: RestingBP vs Target Variable (HeartDisease)

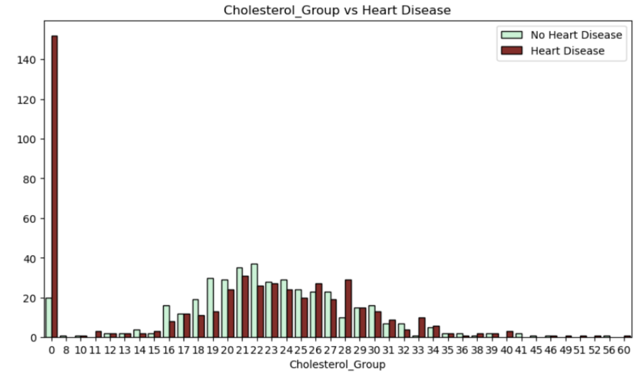

Figure 5: Cholestrol vs Target Variable (HeartDisease)

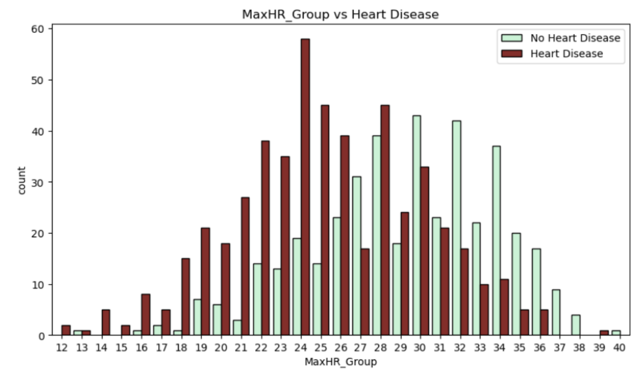

Figure 6: MaxHR vs Target Variable (HeartDisease)

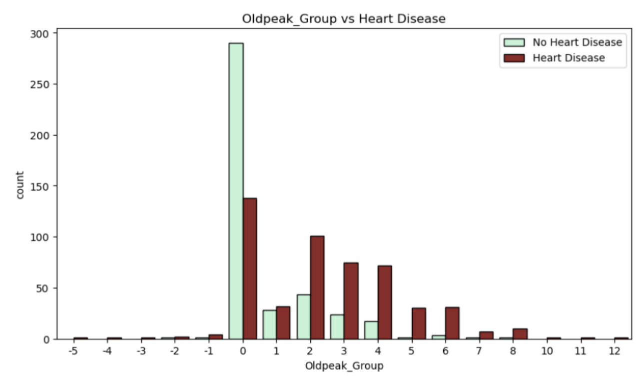

Figure 7: OldPeak vs Target Variable (HeartDisease)

and LVH, beginning at age thresholds of 30, 40, and 40, respectively. Individuals over the age of 50 show a higher susceptibility to heart diseases, and this trend holds true across all types of RestingECG values. The occurrence of heart diseases is distributed evenly across the entire range of RestingBP and RestingECG values. There is a notable concentration of heart disease cases among patients with cholesterol levels in the 200 - 300 range who also have an ST RestingECG value. Additionally, a higher density of heart disease is found in patients with a maximum heart rate below 140

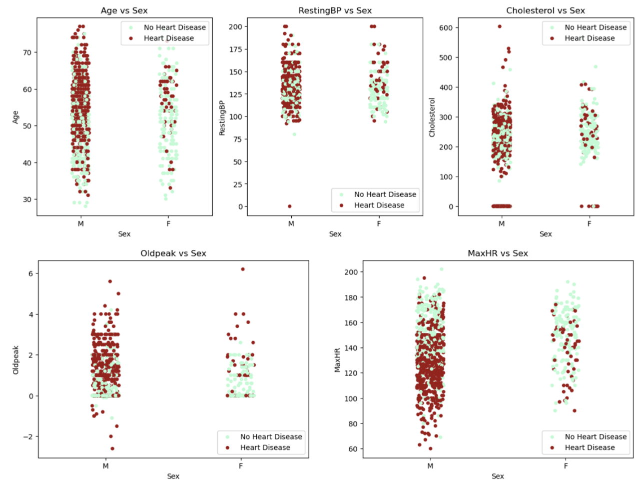

Figure 8: Sex vs Numerical Features.

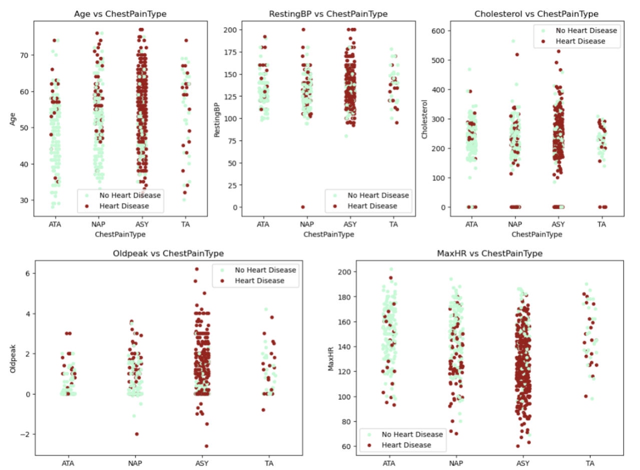

Figure 9: ChestPainType vs Numerical Features

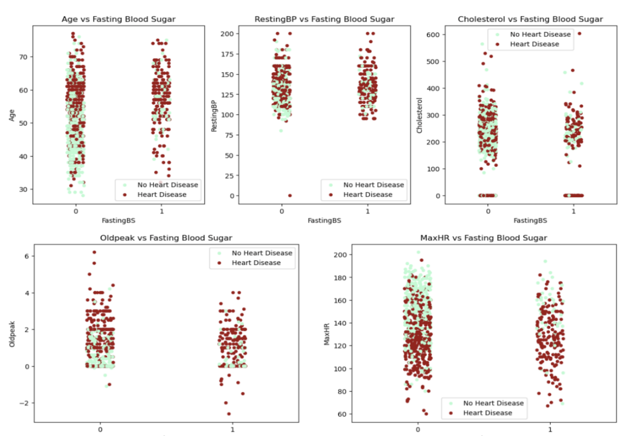

Figure 10: FastingBS vs Numerical Features

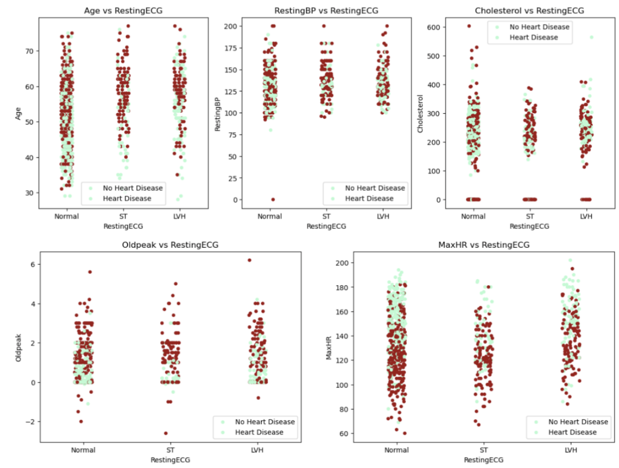

Figure 11: RestingECG vs Numerical Features

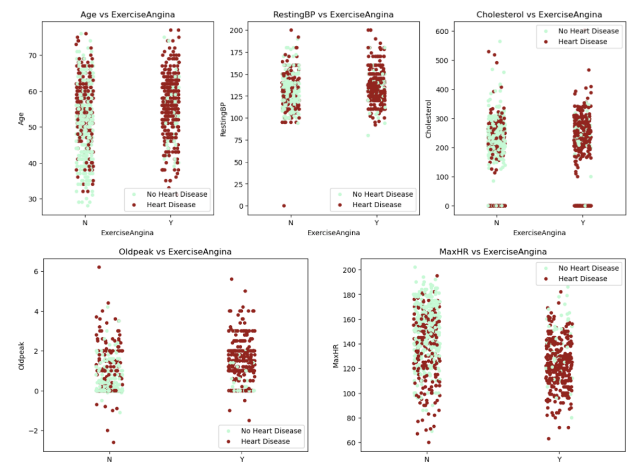

Figure 12: ExerciseAngina vs Numerical Features

and a Normal RestingECG reading. However, heart disease cases are also present across all values of maximum heart rate for those with ST and LVH RestingECG readings.

Figure 12 shows the distribution of the categorical variable ExerciseAngina vs the numerical features.It is distinctly observable that there is a strong positive correlation between the occurrence of heart disease and the presence of exercise-induced angina. This relationship is consistently evident across the entire spectrum of numerical features.

Figure 13 shows the distribution of the categorical variable ST Slope vs the numerical features.It is evident that there is a direct positive relationship between the ST Slope value and the incidence of heart disease. In descending order of likelihood for a heart disease diagnosis, Flat, Down, and Up ST Slope categories correspond to high, moderate, and low probabilities, respectively.

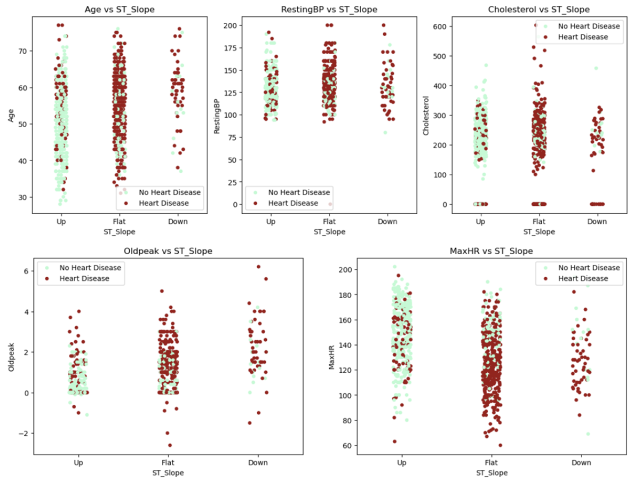

Figure 13: ST Slope vs Numerical Features

	precision	recall	f1-score	support
0	0.88	0.85	0.87	89
1	0.87	0.89	0.88	95
accuracy			0.88	184
macro avg	0.88	0.87	0.87	184
weighted avg	0.88	0.88	0.87	184

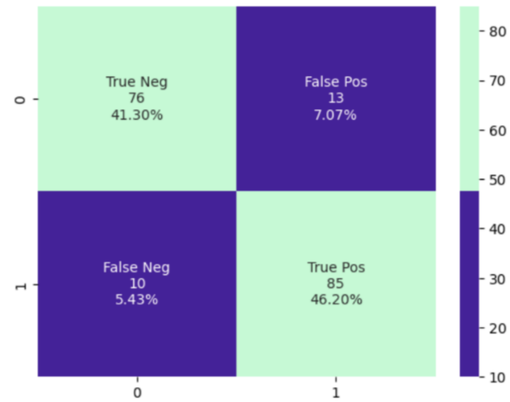

Figure 14: Confusion Matrix and Statistics for Logistic Regression

```
Accuracy :  87.50%
Cross Validation Score :  91.12%
ROC_AUC Score :  87.43%
```

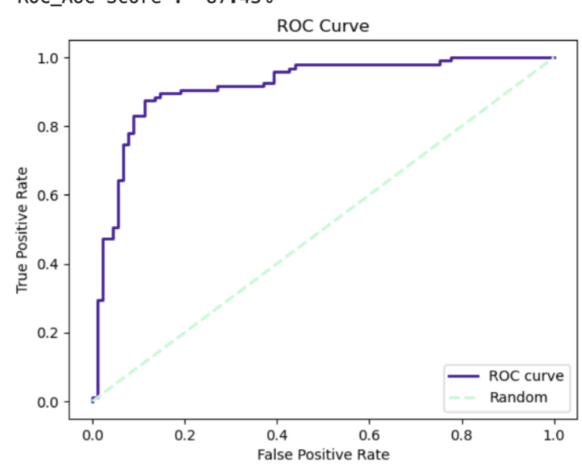

Figure 15: ROC Curve for Logistic Regression

7 MODELLING AND RESULTS

For the detection of heart failure among cardiovascular conditions, we tried to implement five state-of-the-art machine learning models for this problem, respectively chosen for their specific strong points in dealing with the complexity of the characteristics of the dataset. We used logistic regression[1], support vector classifier, decision tree classifier, random forest classifier, and k-nearest neighbors classifier models. The choice of these models is motivated by different abilities in handling non-linear relationships, feature interactions, and approaches to the classification boundary, finally, making them adaptable on the multidimensional and nuanced dataset in question.

7.1 Logistic Regression

First, of course, it is Logistic Regression. This is most often a default selection for a binary classification problem and uses a logistic function to model a binary dependent variable. In our project of predicting heart failure, it assesses the probability of correspondence of a given set of inputs to a particular class (presence or absence of heart disease). This model is even popular for its outputs, which give not only the classifications but also the probabilities linked to them, therefore giving clinicians an idea of the risk levels. Linear kernel works well on this dataset since it has the linearity property and gives a clear explanation of individual features, say age or cholesterol levels, in relation to the likelihood of heart disease.

This further means that the coefficients from the logistic model can be interpretable to ascertain both the direction and strength that they have an impact on predictor variables, hence invaluable for medical insights and interventions.

It's efficient in cases where the relations between the independent features and the outcome are linearly expected. In our project, it has an accuracy of 87.50% and the cross-validation scores of 91.12%, the former not the worst and the latter strong in generalizing over the heart failure dataset.

7.2 Support Vector Classifier (SVC)

Support Vector Classifier (SVC) is the most powerful and accurate algorithm that works on linearly separable and non-linearly separable datasets through the implementation of kernel functions. It, therefore, makes SVC highly effective for this application because of high-dimension spaces very well. For the dataset of heart failure,

	precision	recall	f1-score	support
0	0.88	0.85	0.87	89
1	0.87	0.89	0.88	95
accuracy			0.88	184
macro avg	0.88	0.87	0.87	184
weighted avg	0.88	0.88	0.87	184

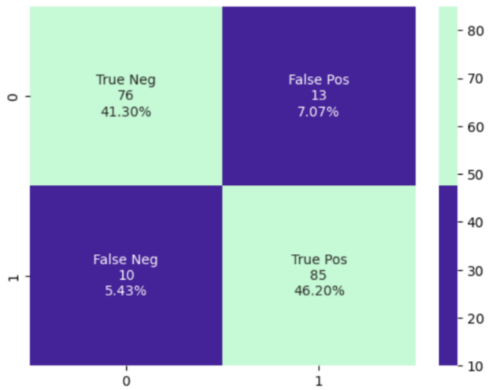

Figure 16: Confusion Matrix and statistics for SVC

the proper kernel choice would help the SVC be conformed to the distribution of points in that space.

The optimization process it uses deals with the points hardest to classify (the support vectors), so as to ensure the model is not too much affected by outlying. This precision makes it apt for medical datasets, whereby the boundaries between classes are not very distinct but are key for the right diagnosis. The kernel trick of support vector machine is pivotal in handling non-linear separations, which makes it a good choice for complex datasets.

With an accuracy score and ROC AUC score of 87.50% and 87.43% correspondingly, SVC guarantees high precision in class separation and, thereby, a highly reliable model in predicting heart disease.

7.3 Decision Tree Classifier

The Decision Tree Classifier is designed to create a model that predicts the value of a target variable from learning simple decision rules. This method is very visual and easy to understand. Each node of the tree corresponds to a feature of the dataset; each branch corresponds to a decision rule. This implies that every path in the tree will provide clear criteria from attributes such as blood pressure or the type of chest pain that will lead to either a diagnosis of heart disease or not. This model provides an advantage in the medical environment with the simplicity it offers in diagnostic paths and clinical validation ease. However, decision trees are overloaded, especially in complex datasets, so the careful tuning of parameters such as the depth of the tree and the size of the leaves is necessary. Hierarchical division of the dataset makes it easy to capture complex patterns. They are effective in handling mixed data types.

The model obtained an 84.78% accuracy score, and the tree structure presented clear insight for feature importance and decision paths.

```
Accuracy :  87.50%
Cross Validation Score :  90.53%
ROC_AUC Score :  87.43%
```

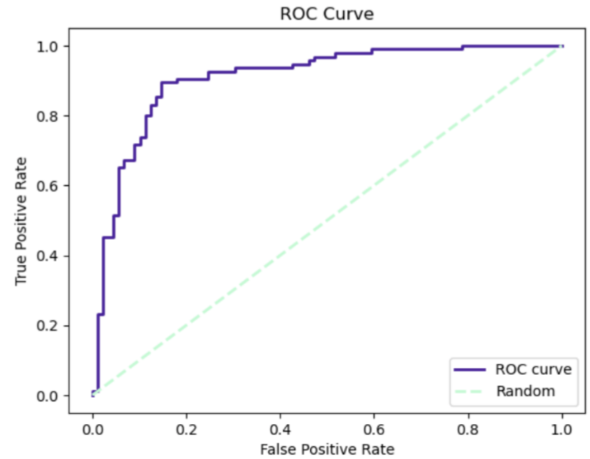

Figure 17: ROC Curve for SVC

	precision	recall	f1-score	support
0	0.88	0.80	0.84	89
1	0.83	0.89	0.86	95
accuracy			0.85	184
macro avg	0.85	0.85	0.85	184
weighted avg	0.85	0.85	0.85	184

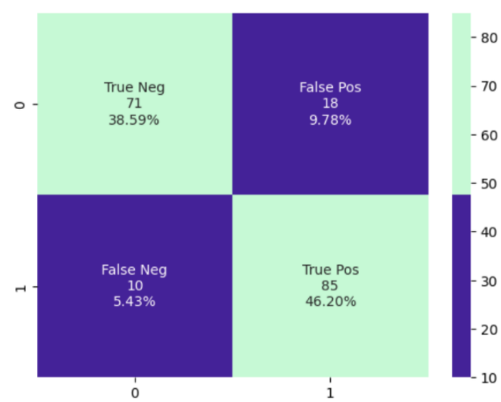

Figure 18: Confusion Matrix and Statistics for Decision Tree Classifier

7.4 Random Forest Classifier

Random Forest is an ensemble of the decision tree method, creating a "forest" of trees, of which every tree is a bit different from others. This model introduces randomness into the model building process. Each tree is built from a random sample of data points, and a random subset of the features is considered for each split. This randomness introduces a stronger model for better generalization to reduce overfitting, one of the key problems associated with decision trees.

```
Accuracy :  84.78%
Cross Validation Score :  89.09%
ROC_AUC Score :  84.62%
```

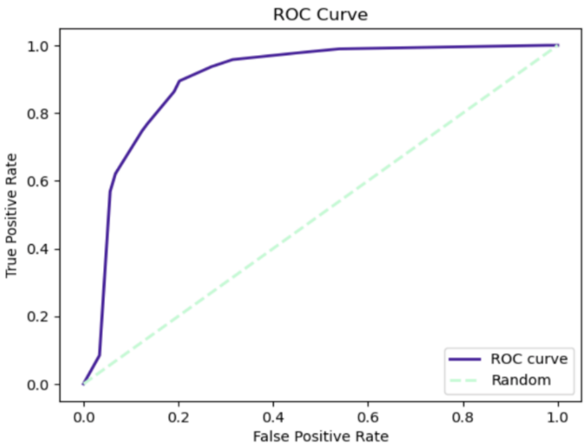

Figure 19: ROC Curve for Decision Tree Classifier

Prediction of heart failure using this ensemble approach means it is able to handle the various anomalies and non-linear relationships of the dataset without fitting excessively to noise, hence giving good reliability in prediction among different patient profiles. Random Forest improves the Decision Tree by building multiple trees and aggregating their predictions, which results in better performance and reduces overfitting of the model. This proves highly effective in high-dimensional and collinear datasets.

Of all the models, this has the highest cross-validation score at 92.91%. This points out its stability and predictive ability, making it highly reliable in predicting heart failure.

7.5 K-Nearest Neighbors Classifier

The K-Nearest Neighbors (KNN) method follows very simple principles: to classify new data points, it searches 'K' number of close, labeled data points. It then attributes the majority class of its closest neighbors to this new data point. This model is purely nonparametric in the sense that it generally harbors no foundational assumptions about the data distribution. On this premise, it thus becomes more adaptive to real-world data, which rarely upholds such assumptions. In the case of heart failure detection, KNN might give good results if the dataset clearly clusters outcomes based on similar features from the patient.

Its major limitation, however, is that it is sensitive to the local data structure and can hence be very much affected by noisy or irrelevant features, needing careful selection of features and tuning of the parameter K (the number of neighbors). K-Nearest Neighbor (KNN) is an example of the proximity-based approach used in making predictions and assumes that similar cases with similar features lie near each other. This model is inherently nonparametric and captures the complex structure of the data space by way of considering the localized patterns of data points.

	precision	recall	f1-score	support
0	0.88	0.79	0.83	89
1	0.82	0.89	0.85	95
accuracy			0.84	184
macro avg	0.85	0.84	0.84	184
weighted avg	0.85	0.84	0.84	184

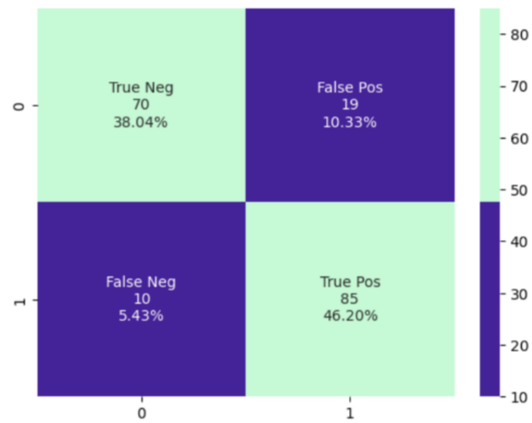

Figure 20: Confusion Matrix and Statistics for Random Forest Classifier

```
Accuracy :  84.24%
Cross Validation Score :  92.91%
ROC_AUC Score :  84.06%
```

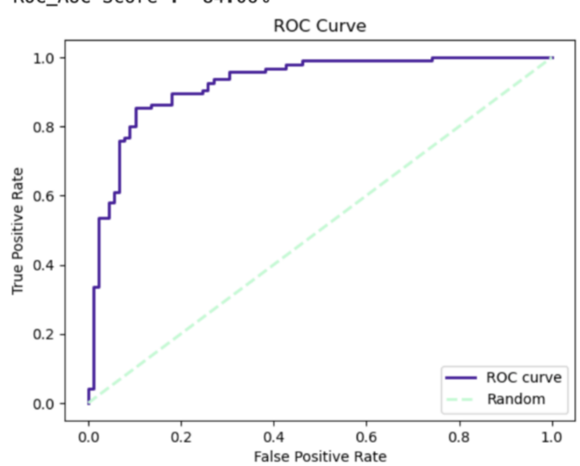

Figure 21: ROC Curve for Random Forest Classifier

Although it has the lowest accuracy of 81.52%, the simplicity and interpretability of the predictions based on the visible neighbors yield valuable clinical insight.

	precision	recall	f1-score	support
0	0.84	0.76	0.80	89
1	0.80	0.86	0.83	95
accuracy			0.82	184
macro avg	0.82	0.81	0.81	184
weighted avg	0.82	0.82	0.81	184

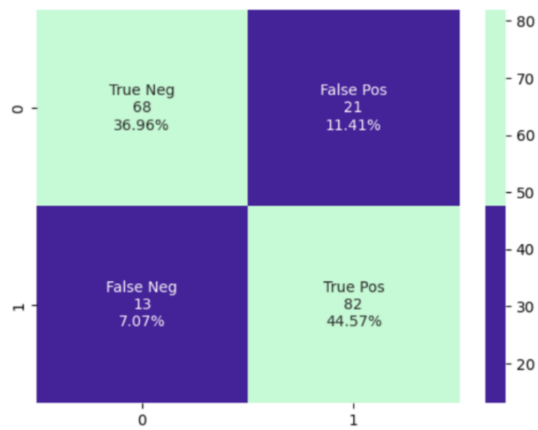

Figure 22: Confusion Matrix and Statistics for KNN Classifer

```
Accuracy :  81.52%
Cross Validation Score :  89.34%
ROC_AUC Score :  81.36%
```

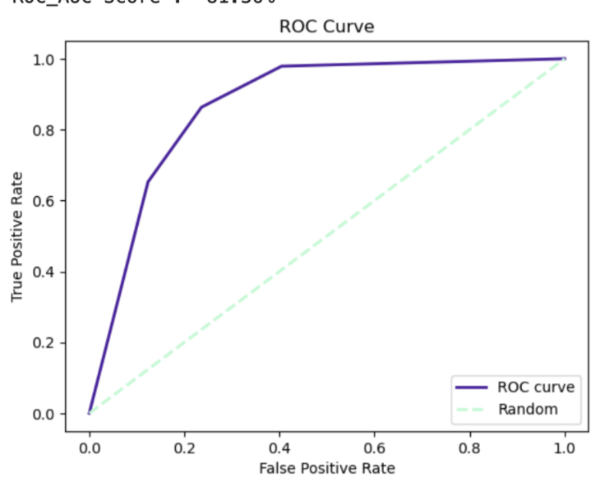

Figure 23: ROC Curve for KNN Classifier

7.6 Convolutional Neural Network

The Convolutional Neural Network (CNN) used in the heart failure prediction project represents a more complex and sophisticated approach than the traditional machine learning models previously discussed. Unlike simpler models that process input data in their raw form, CNNs introduce the concept of layers which can extract and learn features automatically through backpropagation. This ability to learn features directly from the data without needing manual feature extraction makes CNNs particularly well-suited this dataset.

In our project , the CNN model uses one-dimensional convolutions, as the data consists of structured tabular data. Convolutional Layers are the core building blocks of a CNN. Each convolutional layer applies a number of filters to the input. These filters are small matrices that move across the input data, perform element-wise multiplication, and sum up the results, producing a feature map. This process captures local dependencies and patterns in the data, such as the relationship between neighboring features like blood pressure and cholesterol levels.

We used a ReLU (Rectified Linear Unit) activation function following each convolutional layer. This function introduces nonlinearity into the model, allowing it to learn more complex patterns. To reduce the spatial size (i.e., dimensionality) of the feature maps we used Max pooling layers. The high-level reasoning in the neural network is done via fully connected layers. Where Neurons in a fully connected layer have connections to all activations in the previous layer. These layers act to classify the data based on the features extracted in the convolutional and pooling layers.

Finally,The final output layer uses a sigmoid activation function since we are dealing with binary classification and it learns through binary cross entropy. which provides the output probabilities of the presence of heart disease or not.

	precision	recall	f1-score	support
0	0.82	0.82	0.82	77
1	0.87	0.87	0.87	107
accuracy			0.85	184
macro avg	0.84	0.84	0.84	184
weighted avg	0.85	0.85	0.85	184

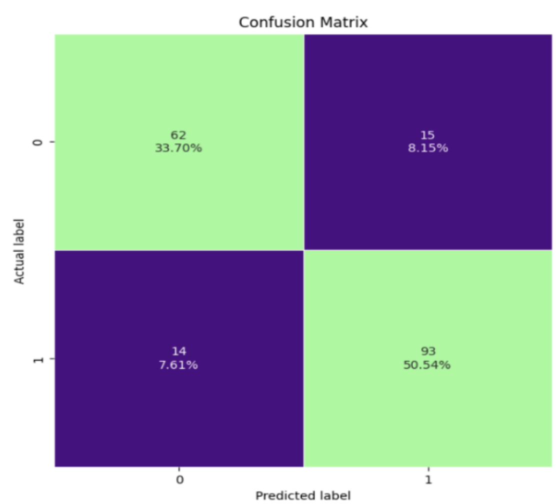

Figure 24: Confusion Matrix and Statistics for CNN

The CNN models accuracy reported at 84.24% with an ROC AUC of 89.16% indicates that the model was able to learn complex representations of the data, potentially capturing intricate interactions between features.

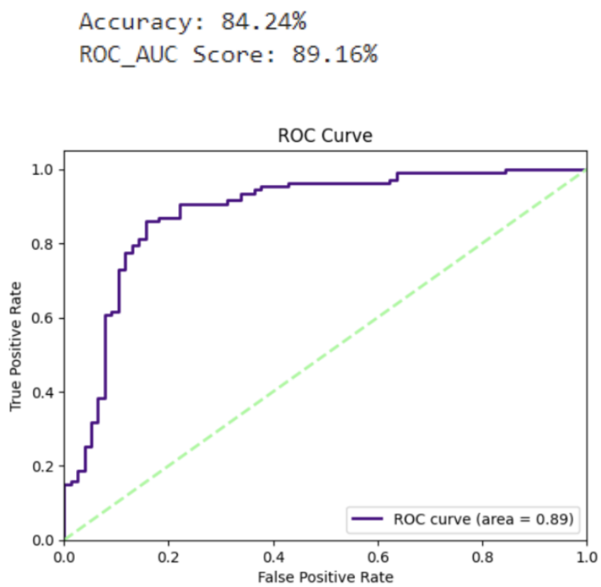

Figure 25: ROC Curve for CNN

8 CONCLUSION

This study underscores the transformative potential of machine learning (ML) in addressing the global challenge of cardiovascular diseases (CVDs), with a specific focus on predicting heart failure. By harnessing a diverse array of ML models—including Support Vector Classifier, Random Forest, Decision Tree, K-Nearest Neighbors, Logistic Regression, and Convolutional Neural Networks—we have developed a robust framework capable of early detection and timely intervention in heart failure cases.

Our comprehensive approach involved meticulous data preprocessing to optimize the quality and relevance of the dataset, which included numerous attributes associated with heart health. Through exploratory data analysis, we gained valuable insights into the relationships and patterns within the data, enabling more informed feature selection and model tuning. The implementation of a multi-model strategy allowed us to compare and evaluate the efficacy of different algorithms, ultimately identifying the most accurate and reliable models for heart failure prediction.

The findings from our study reveal that the integration of machine learning into healthcare workflows can significantly enhance the accuracy of heart failure predictions, offering a substantial improvement over traditional diagnostic methods. This not only aids in reducing the mortality rates associated with heart diseases but also facilitates a shift towards proactive and preventive healthcare paradigms.

9 FUTURE WORK

Future research would concentrate on combining more complex algorithms, as our current analysis has limitations. In order to enhance prediction methods, this research can also be enhanced by combining multiple machine learning algorithms.

One approach would be to Delve into powerful neural networks like LSTM/ Bi-LSTM to evaluate the models performance and explore the use of ensemble methods to potentially enhance the model performance. This optimization process will enable us to more effectively evaluate the presence of heart failure across various heart failure-related datasets. Additionally, we will focus on improving the accuracy of existing algorithms to enhance their performance in detecting heart disease. By leveraging these advancements, we aim to provide more robust and accurate methods for the diagnosis and evaluation of heart failure Prediction.

As we continue to refine our models and expand our datasets, our future work will focus on improving the scalability and adaptability of our solutions to accommodate a broader spectrum of demographic and geographical variations. Our ultimate goal is to deliver a universally accessible tool that empowers healthcare providers worldwide to predict and manage heart failure more effectively, contributing to the overarching aim of reducing the burden of cardiovascular diseases globally.

10 GITHUB REPOSITORY

The dataset and the code base for the project can be found at the following Github Repository: https://github.com/nigelthecosta/CS584.git

REFERENCES

[1] Hoogendoorn M. Aussems C. Bennis, F. C. and J. C. Korevaar. 2022. Prediction of heart failure 1 year before diagnosis in general practitioner patients using machine learning algorithms: a retrospective case–control study.. In *BMJ Open (12(8), e060458.).* https://doi.org/10.1136/bmjopen-2021-060458

[2] Sameer Pandey and Ravinder Kaur. 2022. Heart Failure Prediction Using Machine Learning Algorithm. In *IEEE Journals Magazine | IEEE Xplore.* .https://ieeexplore.ieee.org/document/9990484

[3] Chandrasegar Thirumalai Senthilkumar Mohan and Gautam Srivastava. 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. In *IEEE Journals Magazine | IEEE Xplore.* https://ieeexplore.ieee.org/document/8740989

[4] Abdelhalim H. DeGroat W. Zeeshan S. Venkat, V. and Z. Ahmed. 2023. Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine.. In *Genomics (115(2), 110584.).* https://doi.org/10.1016/j.ygeno.2023.110584