Predictive Modelling Coronary Artery Diease and Hypertension Using Machine Learning

MSc Research Project
MSc DAD_A_JAN24I

Maitreya Govind Lavalekar
Student ID: X23209976

School of Computing
National College of Ireland

Supervisor: Prof. Abdul Qayum

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Maitreya Govind Lavalekar
Student ID:	X23209976
Programme:	MSc DAD_A_JAN24I
Year:	2024
Module:	MSc Research Project
Supervisor:	Prof. Abdul Qayum
Submission Due Date:	12/12/2024
Project Title:	Predictive Modelling Coronary Artery Diease and Hypertension Using Machine Learning
Word Count:	9005
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Maitreya Govind Lavalekar
Date:	25th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Modelling Coronary Artery Diease and Hypertension Using Machine Learning

Maitreya Govind Lavalekar
X23209976

Abstract

A Diseases related to the heart such as Hypertension (HT), and coronary artery disease (CAD) are major global health hazards. The timely prediction can help in performing preventive measures, thus providing a better patient outcome.Present models. Current techniques, despite their progress in predictive modeling, are often inapplicable for generalizing across complex and heterogeneous patient populations. This limitation has reduced their accuracy and reliability in real-world clinical settings, indicating that there is a need for more robust models that could address these issues and provide better prediction performance. In this study, by overcoming the data imbalance issue and utilizing ensemble methods with the data balancing through CTGAN, developed machine learning models that can predict the HT and CAD correctly with high accuracy. HT model generalized well over datasets attaining a test accuracy of 97% with balanced precision and recall. Test accuracy for the CAD model: 92% with a recall of 0.91 of CAD-positive cases, meaning it is able to reliably classify patients at risk of CAD. The CAD model shows minimal overfitting, with the training accuracy at 94%. The findings suggest that balancing the data can improve the accuracy to levels that can be clinically useful, and the ensemble model provides a reliable tool for accurate risk assessment for healthcare providers in the early stages of patient care. Larger datasets and advanced efforts over the model development on how to be more sturdy can be seen in future works.

Keywords : Coronary Artery Disease (CAD), Hypertension (HT), Synthetic data generation , Predictive Modelling, Machine Learning , Ensemble Method.

1 Introduction

In Cardiovascular diseases have become global risk factors for human health. If the healthcare providers can identify these conditions earlier, they may be able to take preventive measures that will lead to an improved healthcare outcome or decreased complications Jaiswal et al. (2017) . Although predictive models have been trained well, current models still struggle, particularly when working with imbalanced datasets, in which the rising(high) risk cases are rare Singh et al. (2024),Perkonoja (2020). We intend to solve these problems by developing a better predictive model that has both lower bias and variance compared with the current models and investigates the use of advanced machine learning techniques and data balancing with realistic synthetic data. A reliable prediction model for HT and CAD could have a great public health impact. Present models often have low generalizability to more heterogeneous patient populations and have suboptimal predictive accuracies in the clinical settings García-Vicente et al. (2022),Wang et al. (2024). This study aims at ameliorating model prediction by addressing data imbalance and model ensemble techniques. If such models are implemented successfully, it can serve as a decision-support tool for healthcare providers to identify high-risk individuals so that appropriate interventions can be implemented at a timely stage Acherjya et al. (2023)

Research Question: *How information regarding cardiovascular health can be beneficial for building integrated prediction models for Hypertension and coronary Artery Disease Verification?*

Sub-Research Questions: *How the application of these prediction models in a clinical setting can improve early identification and preventative treatment of hypertension and CAD?*

What important health indicators have the greatest impact on risk forecast for hypertension and CAD?[Blood Pressure ,Cholestrol Level,Systolic Blood pressure ,Diastolic Blood pressure ,Age, Family History , Diabetes]

Research Objective: To address the above questions, this study first evaluates the previous research on hyper-tension (HT) and coronary artery disease (CAD) prediction models and then provides a comprehensive review of existing models, including the current approaches and their limitations. A machine learning model will then be built on the basis of this review, with ensemble methods and data balancing techniques, to improve prediction performance. After the model will be applied focusing on proper data challenges in an environment where high risk cases are imbalanced. Third, it will assess the performance of the model using the three key performance metrics (accuracy, precision, and recall) to determine the reliability of the model across different clinical scenarios.

A novel model that integrates ensemble learning techniques with some optimization needed, for instance, CTGAN to create a predictive model, is the primary contribution of this research work. The model applies an integrated approach, meaning it can enhance predictions by potentially reducing biases through higher versatility, which is advantageous for reliable evaluation in various datasets, making the model a useful clinical tool to identify and monitor patients at risk for HT and CAD Alqulaity and Yang (2023),Mroz et al. (2024). There are several limitations in the present study. The dataset one uses may not cover all population groups adequately, while the model presets the data quality to be stable all the time. Testing of the model with larger, diverse datasets in various healthcare contexts is recommended for future work, making this research a preliminary result.

Structure of the Report: The report has been organised in a way which best addresses the question and objectives laid out in the introduction. Introduction describes the purpose of the study, research problem and significance. Abstract: The Literature Review illustrates existing models of prediction for Hypertension (HT) and coronary artery disease (CAD), as well as emphasizes the demand for the enhanced accuracy and adjustment data balancing. The Methodology includes detailed information on the applied research design, processing of the data, and feature engineering (feature extraction/selection) and machine learning (ML) methods (SMOTE, Regression, and KNN). The effectiveness of the achieved model will be evident with metrics like accuracy and precision through the results. The Discussion mainly interprets, brings forth key health indicators and limitations of the study.

Finally, the Conclusion and Future Work captures the main findings and possible ways in which the study can be extended to improve the model and broaden the analysis. In the next section, we discuss some of the related work and existing methods in predicting modelling for coronary artery disease and hypertension.

2 Related Work

In this section, we will critically review the main literature on predicting modelling for CVD focused on HT and CAD. This review discusses the evolution of predictive models since the early days of statistical modelling to contemporary machine learning-based approaches, showcasing a combination of foundational and recent studies. We synthesize approaches as part of the Knowledge Discovery in Databases (KDD) process to tackle ongoing issues of data imbalance and lack of generalizability of models in real-world clinical practice. In this regard, this analysis main objectives is to identify gaps in extant research and provide justification for the method used in this study, which combines synthetic data generation with ensemble methods to improve prediction accuracy and reliability in predicting HT and CAD risk..

2.1 An Overview of Cardiovascular Health Predictive Models

Choi et al. (2021)'s model had proven its ability for large-scale CVD prediction using data from the UCI repository And achived 0.67% of accuracy in Logistic regression model . In a similar fashion, Jeanson et al. observed that the accuracy of the model would be improved by 20% achieving 0.87% accuracy with the same dataset and help of ensemble learning methods such as bagging and boosting while applied in clinical environments. However, these methods have somehow to be continuously adapted and adjusted for different real datasets; however, researchers did not perform feature engineering and hyperparameter tuning in there research if they did, the result might be more generalize Choi et al. (2021); Jeanson et al. (2024).

2.2 Advanced Feature Selection Methods

In a different study, the authors implemented advanced feature selection techniques in their work for prediction of HT and CAD Abdullahi et al. (2024). They obtained data from live patient presenting to hospital in clinical settings. By reducing data dimensionality via certain algorithms, they made progress toward achieving an accuracy level as high as 75%. Using their feature selection process, they improved interpretability and achieved a gain of 12% in relation to baseline models. Petmezas et al. (2022) also applied Kaggle electrocardiogram data and showed noise reduction so that feature selection ensured 18% accuracy gain. Although these models are very promising, they require a lot of data preprocessing to perform well on different datasets but after hyperparameter tunning there accuracy got decreased from 0.95% to 0.75% Abdullahi et al. (2024); Petmezas et al. (2022).

2.3 Conventional Risk Analysis Models

Chowdhury et al. (2022) performed the conduct of regression and ML analysis on a set of national clinical data using samples based on traditional. Using samples from National Health Insurance Service, 75% accuracy for prediction of hypertension as outcome was achieved with regression models. But they discovered ML models (especially the deep learning included ones) performed 20% more accurately than regression, climbing to about 90%. Chang et al. (2019)also supported these results using UK Biobank data, stating that the ML models had superior performance when predicting HT among large cohorts of patients Chowdhury et al. (2022),Chang et al. (2019).But both authors didn't applied Statistical test ,l2 regularization for generalized model and co-relation-based feature selection .

2.4 Models in Coronary Artery Disease and Hypertension Prediction

Park et al. (2023) developed Powerful algorithms a gradient-boosting machine that was trained on real-time patient monitoring data features to predict hypoxemia with 78% accuracy. where as researcher Chang et al. (2019) Random Forest was used to identify the most important risk factors for rectal toxicity, with correct predictions of 88% on the kaggle's heart dieases dataset, again showing that these machine learning algorithms can be translated into applicable clinical data. Yet all models demanded a lot of computing resources and tuning Park et al. (2023),Chang et al. (2019)

2.5 HT and CAD Prediction by Ensemble Techniques

Singh et al. (2024) Alqulaity and Yang (2023) used ensemble learning methods like Random Forest, SVM and Gradient Boosting to predict CAD and HT, obtaining an increase of 16% in accuracy over single model predictions. They obtained a maximum accuracy of 88% using their models trained of Kaggle 's heart disease dataset. Even though the ensemble models were found to be more robust, transferring their results across different patient demographics continues to be a problem also most important gap in there study was model validation if they worked on validation it might more clear about how there model perform on training as well as in testing. Singh et al. (2024) Alqulaity and Yang (2023)

2.6 Synthetic Data Generation for Solving Data Imbalance

One of those improvements is synthetic data generation used to mitigate data imbalance. García-Vicente et al. Clinical Synthetic Data Generation García-Vicente et al. (2022)leveraged KDD's framework to utilized CTGAN method in cardiovascular datasets, resulting in better representation of minority classes. They improved model robustness by creating synthetic samples for underrepresented high-risk cases, without losing data quality . In the same way used CTGAN to create synthetic data on cardiovascular data from the VLDB workshop and increased model accuracy for minority cases by 20% and reached an overall accuracy of 84%. Perkonoja (2020) exercised the Priv Bayes on health data from UCI with more balance in datasets and semi-limited privacy for cardiovascular-specific predictions; however, exercise was constrained. Although these synthetic data methods are encouraging, they may need further testing to establish clinical importance García-Vicente et al. (2022)Perkonoja (2020)

2.7 Improvement in Data Processing Techniques

Wang et al. (2024)implemented KDD Methodology in there research and on a UK primary care dataset they applied machine learning algorithms like traditional models and under sampling methods like smote ; by following an orderly approach they achieved 82% accuracy in predicting patient outcomes. Kumar et al. (2023). A 10% improvement in model reliability on Big Data healthcare analytics was achieved by incorporating KDD with ensemble approaches on different datasets, reinforcing that if data is handled systematically, consistent model performance will be obtained but both researcher did not perform hyperparameter tunning as well as model validation Kumar et al. (2023),Wang et al. (2024)

2.8 Development and Optimization for Model Framework

The prediction of cardiovascular disease was implemented using KDD framework for model optimisation Tompra et al. (2024)to achieve 88% accuracy over the Z-Alizadeh Sani dataset. The structured optimization-based learning and Advanced machine learning model like ANN conveyed them a high accuracy, yet they were not unable to do hyperparameter tunning .Similarly, Fang et al. (2023) designed a deep learning-based ensemble strategy together with traditional machine learning methods to surpass state-of-the-art accuracy of UCI health datasets and they achieved 86% of highest accuracy from ensemble model that combined knn and lightgbm .These frameworks are effective, but they involve a trade-off between and the computational efficiency and prediction accuracy and also some how they did not consider features like family history blood pressure types. Tompra et al. (2024),Fang et al. (2023).

2.9 Model generalizability and clinical applicability problems

In a study of hypertension models from different clinical environments, García-Vicente et al. (2022)reported that performance decreased by 15% and got 0.75% accuracy when applied to patients in a new setting. More worryingly Cuevas-Chavez et al. (2023) and Mroz et al. (2024) showed a reduced prediction accuracy of approximate 10% on datasets coming from four regional healthcare centres in Europe, highlighting the difficulty to get final models that perform well everywhere.García-Vicente et al. (2022) , Cuevas-Chavez et al. (2023),Mroz et al. (2024)

2.10 Research Niche and Novelty

The novelty of this study is its integration of ensemble learning, synthetic data generation using the CTGAN algorithm, and KDD framework in order to improve predictive accuracy for HT and CAD.

Earlier, synthetic data techniques or ensemble learning were applied separately, but very few studies incorporated these approaches within a more structured KDD process. This integration is expected to make a combination of a more robust and flexible model to cope with imbalanced real clinical data. Utilization of KDD facilitates a more systematic handling, balancing and evaluation of data, addressing the current limitations in predictive performance and generalizability of the models. Fang et al. (2023),García-Vicente et al. (2022),Kumar et al. (2023),Wang et al. (2024). Below, Table 1 shows the summary of related work.

Author	Dataset Name	Technique	Accuracy	Research Gap
Choi et al. (2021)	National Health Insurance Service Database from Republic of Korea	Logistic Regression and Long Short-Term Memory (LSTM)	LR: 0.679% LSTM:-0.75%	Feature Engineering, hyperparameter tuning.
Jeanson et al. (2024)	National Health Insurance Service Database from Republic of Korea	Ensemble Method	Ensemble method: 0.894%	Feature Engineering, hyperparameter tuning.
Abdullahi et al. (2024)	CVD Prediction UCI Repository	Dimensionality Reduction + Ensemble Technique	0.75%	Did not perform Hyperparameter tuning.
Petmezas et al. (2022)	Samples from National Health Insurance Service	Perform regression and deep learning	Regression: 0.75%, Deep Learning: 0.95%	Feature engineering for Hypertension model.
Chowdhury et al. (2022)	Z-Alizadeh Sani dataset	Gradient Boosting and Random Forest	Gradient Boosting: 0.78%, Random Forest: 0.83%	Did not perform Hyperparameter tuning.
Chang et al. (2019)	Kaggle's heart disease dataset	Random Forest, SVM, and Gradient Boosting combined with Ensemble Technique	0.88%	Feature selection-based model training.
Alqulaity and Yang (2023)	Kaggle's heart disease dataset	GANs and Ensemble Technique	0.88%	Model Validation (Cross Validation).
Park et al. (2023)	Z-Alizadeh Sani dataset	Random Forest	0.78%	High computational resources, hyperparameter tuning.
Singh et al. (2024)	Kaggle's heart disease dataset	Ensemble Technique	0.88%	Model validation (Cross Validation).
García-Vicente et al. (2022)	Cardiovascular data from the VLDB workshop	CTGAN for synthetic data generation and deep learning	0.84%	Hyperparameter tuning, Statistical Analysis, generated fewer synthetic data.
Perkonoja (2020)	UK Bio Bank Data	Deep Learning models	0.90%	Did not perform Statistical test on Categorical features.
Wang et al. (2024)	UK primary care dataset	GANs, Traditional ML Models	0.82%	Generated fewer synthetic data.
Kumar et al. (2023)	Big Data healthcare analytics	GANs and Ensemble approach	0.92%	Model Validation, Hyperparameter Tuning.
Tompra et al. (2024)	Z-Alizadeh Sani dataset	Advanced Sampling method (SMOTE), ANN	0.88%	Hyperparameter Tuning.
Fang et al. (2023)	UCI health datasets	Ensemble method (LightGBM + KNN)	0.86%	Didn't consider features like family history and Blood pressure type.
Cuevas-Chavez et al. (2023)	Dataset from 4 regions in Europe	Deep learning	0.90%	Difficult to make real-time prediction.
Mroz et al. (2024)	Dataset from 4 regions in Europe	XGBoost	0.80%	Difficult to make real-time prediction.

Table 1: Literature Summary

3 Methodology

The present study systematically develops and evaluates predictive models that include two major cardiovascular diseases namely, Hypertention (HT) and coronary artery disease (CAD), through the methodology of knowledge discovery in databases. KDD process, from data selection to preprocessing, transformation, mining and evaluation together ensure the methodical rigor and should allow a straight-forward re-production of results (Figure 1). Breaking them down as three distinct steps leads to granular operation with great generalization performance on heterogeneous and imbalanced datasets while facilitating model interpretability. Consistent with KDD framework, this study offers a reproducible roadmap for researchers to validate its findings and use it as stepping stone for future works that ultimately lead to reliable and clinically useful models for HT and CAD predictionAbdullahi et al. (2024)Tompra et al. (2024). In next section we will discuss phases of KDD Methodology.

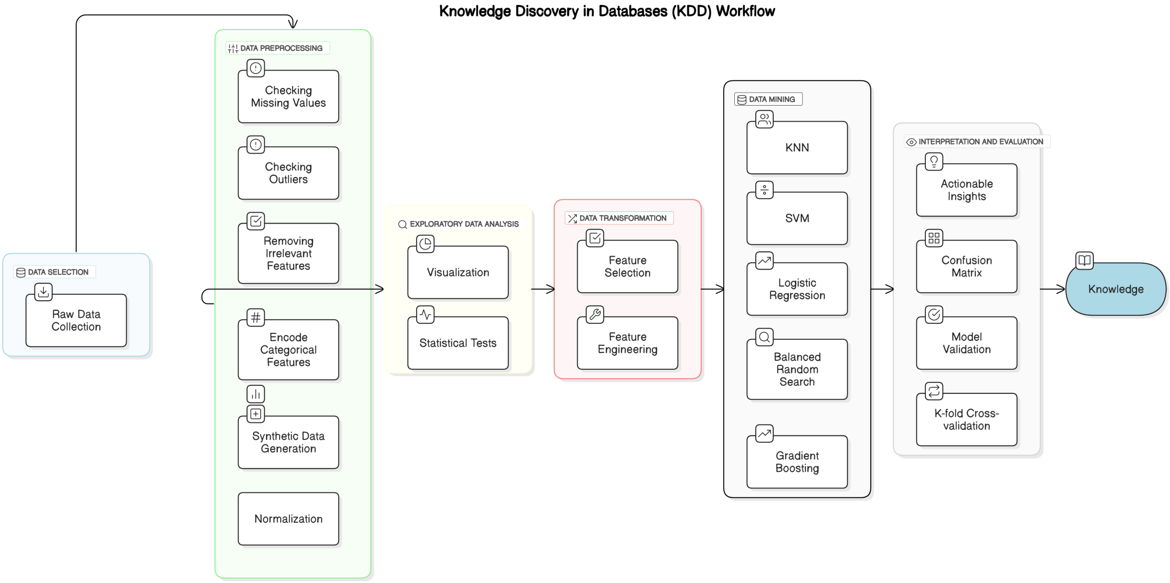


Figure 1: KDD Methodology

3.1 Data Selection

In this section we did data selection from kaggles .In next sub-section we discuss how data was colleted and what features are in datasets.

3.1.1 Raw Data Collection

Primary dataset was collected from Kaggle's Heart Attack Risk Prediction Dataset. It includes 8764 rows, and it includes 26 features like Patient ID, Age, Sex, Cholesterol, Blood Pressure, Heart Rate, Diabetes, Family History, Smoking, Obesity, Alcohol Consumption, Exercise Hours Per Week, Diet, Previous Heart Problems, Medication Use, Stress Level, Sedentary Hours Per Day, Income, BMI, Triglycerides, Physical Activity Days Per Week, Sleep Hours Per Day, Country, Continent, Hemisphere, Heart Attack Risk. Synthetic and real-world datasets were used in this study to reflect as completely a picture of HT and CAD indicators as possible. To deal with the problem of data imbalance, synthetic data was generated using state-of-the-art Generative Adversarial Networks (GANs) specifically CTGAN Arora and Arora (2023). Furthermore, synthetic data enhances privacy without loss of utility.García-Vicente et al. (2022)In next section we will discuss about Data Preprocessing .
[1]

3.2 Data Preprocessing

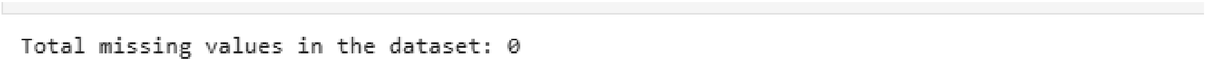
The following steps illustrates important processes in the data preprocessing stage such as null value treatment, normalization, encoding categorical features and outlier removal Each step is required, for preparation of data for further analysis, and it will lead to im-

[1] https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

proving model performance and also most important getting meaningful results effective for machine learning Petmezas et al. (2022).

3.2.1 Handling Missing Value

After analysing the dataset where used. isnull (). sum () methods for finding a missing values. Method. isnull () identifies Nan or None values in the dataset and returns a same shape data frames or list of series and. sum () method written a true value. In taken dataset we did not came across missing values. In (Figure 2) we can see that there were no missing values found in the dataset. Mroz et al. (2024)

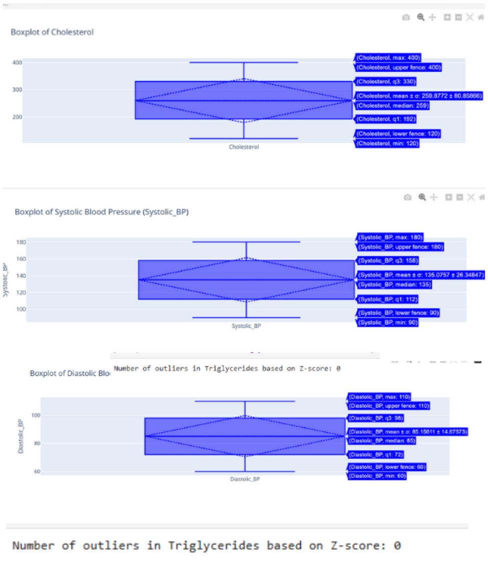


```
Total missing values in the dataset: 0
```

Figure 2: Missing Value

3.2.2 Checking Outlier

Finding an outlies is important step to identify unusual datapoints that will effect on machine learning models accuracy, and it may indicate data quality issue. Handling these outliers in data will improve overall accuracy as well as robustness of machine learning model and give us meaningful insights. For this study we applied boxplot as well as z-score method for finding outliers from data, and we did not find outliers from data it indicates that data is consistent and free from extreme values (Figure 3). Researcher Mahmood et al. (2024) state that Using Z-score method for finding outliers from data is important because z-score method indicates the number of standard deviations a data point is from the mean of a dataset. Analyzing z-scores is often used to detect outliers, standardize data, or compare scores across different distributions.Mahmood et al. (2024)



```
Number of outliers in Triglycerides based on Z-score: 0
```

```
Number of outliers in Triglycerides based on Z-score: 0
```

Figure 3: Checking Outliers

3.2.3 Removing Irrelevant Features

In this study we removed irrelevant columns patientid, Country, Continent, Hemisphere based on co-relation matrix method. (Figure 4)Because we could not find any relation with other features that contribute to study.Wang et al. (2024)

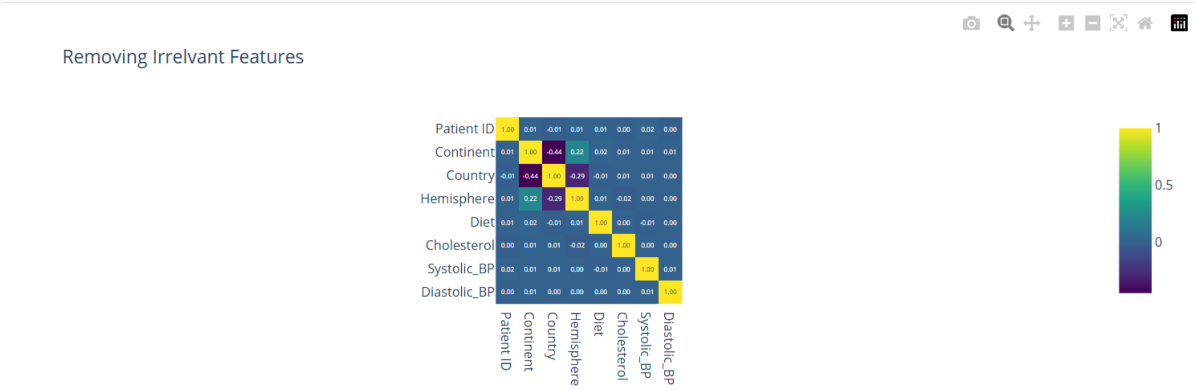


Figure 4: Removing Irrelevant Features

3.2.4 Encoding Categorical Features

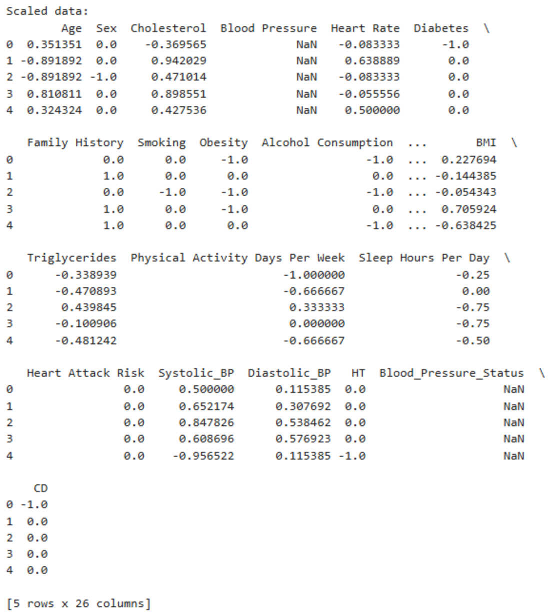
In this study we did encode Sex columns in binary if sex is male then we allocated value 1 and 0 for Female. it also uses one-hot encoding to transform Country into binary columns but drops one category to avoid duplication and we can say to avoid multicollinearity.Choi et al. (2021)

3.2.5 Synthetic Data Generation

In this Study we generated synthetic data using CTGAN library .Researcher Alqulaity and Yang 2023 suggested that using Generative Adversal Network for handling data imbalance and testing robustness of model will be useful in Particularly in healthcare domain Alqulaity and Yang (2023).

3.2.6 Normalization

For a normalization we are using robust scalar method this method measures the median and IQR i.e. Inter Quartile Range to normalize the data, therefore it is robust to very extreme outliers. This is useful when your data contains outliers. This step we did after feature engineering that's why two columns that shows follwoing figure Blood pressure and Blood pressure status showing missing values. Researcher Morgan et al. (2023) state that if the new features provide the necessary and valid information for analysis or for machine learning model this approach is perfectly valid and improves data quality by avoiding redundancy or outdated features in datasets. Morgan et al. (2023) Figure 3.2.6



```
Scaled data:
        Age  Sex  Cholesterol  Blood Pressure  Heart Rate  Diabetes  \
0  0.351351  0.0    -0.369565             NaN   -0.083333      -1.0   
1 -0.891892  0.0     0.942029             NaN    0.638889       0.0   
2 -0.891892 -1.0     0.471014             NaN   -0.083333       0.0   
3  0.810811  0.0     0.898551             NaN   -0.055556       0.0   
4  0.324324  0.0     0.427536             NaN    0.500000       0.0   

   Family History  Smoking  Obesity  Alcohol Consumption  ...       BMI  \
0             0.0      0.0     -1.0                 -1.0  ...  0.227694   
1             1.0      0.0      0.0                  0.0  ... -0.144385   
2             0.0     -1.0     -1.0                 -1.0  ... -0.054343   
3             1.0      0.0     -1.0                  0.0  ...  0.705924   
4             1.0      0.0      0.0                 -1.0  ... -0.638425   

   Triglycerides  Physical Activity Days Per Week  Sleep Hours Per Day  \
0      -0.338939                        -1.000000                -0.25   
1      -0.470893                        -0.666667                 0.00   
2       0.439845                         0.333333                -0.75   
3      -0.100906                         0.000000                -0.75   
4      -0.481242                        -0.666667                -0.50   

   Heart Attack Risk  Systolic_BP  Diastolic_BP   HT  Blood_Pressure_Status  \
0                0.0     0.500000      0.115385  0.0                    NaN   
1                0.0     0.652174      0.307692  0.0                    NaN   
2                0.0     0.847826      0.538462  0.0                    NaN   
3                0.0     0.608696      0.576923  0.0                    NaN   
4                0.0    -0.956522      0.115385 -1.0                    NaN   

    CD  
0 -1.0  
1  0.0  
2  0.0  
3  0.0  
4  0.0  

[5 rows x 26 columns]
```

Figure 5: Normalization

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the key first step in this work, enabling a full understanding of the dataset as well as informing later modelling directions. The EDA step consists of understanding how variables are distributed, locating missing values, dealing with outliers and analysing correlations. EDA ensures that noise is minimized, and predictive insights are more reliable in the later stages by preparing and optimizing for model training with underlying data characteristics revealed Gupta et al. (2022),Mroz et al. (2024). Below are some Visualizations that we are going to look at.

3.3.1 Visualizations

According to researcher pramanik Visualization becomes an integral part of Exploratory Data Analysis (EDA) as it helps to reveal the patterns, trends or anomalies that are not obvious by reading numercal summaries alone.Pramanik et al. (2023) Below are Some visualizations that are help to understand data.

Average Cholesterol By Age The below line graph (Figure 6) depicts the changes in cholesterol levels of individuals with hypertension and without hypertension. The red color shows the individual with hypertension and blue color shows individuals without hypertension. This graph also describes the relationship between two features, i.e. high cholesterol and high blood pressure, among individuals with different ages.

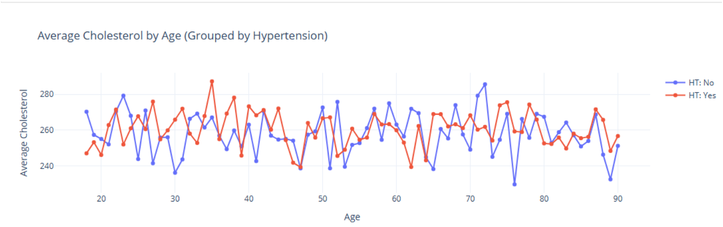


Figure 6: Average Cholestrol By Age

Cholesterol vs Systolic Blood Pressure Below (Figure 7) shows the Scatter plot between cholesterol and systolic blood pressure We have plotted cholesterol and systolic BP readings for each person (each dot) and fitted a linear trend line since we just want to see how things behave you can see that there is a slight upward direction – where also please note the very low correlation coefficient between higher cholesterol and systolic BP.

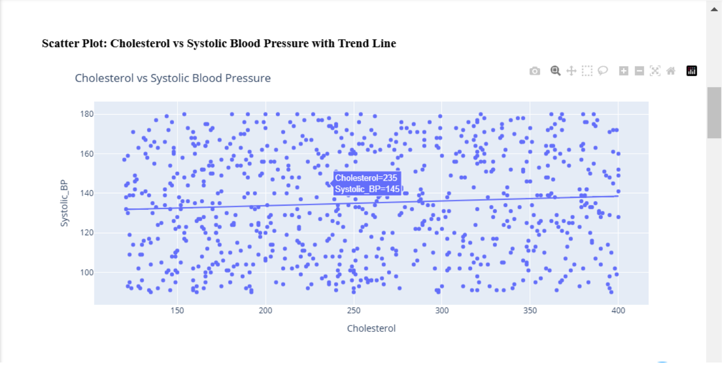


Figure 7: Cholestrol vs Systolic Blood Pressure

Original Data Vs Synthetic Data The following (Figure 8)highlights that the patterns are very similar between most variables in terms of the key features, where both datasets have a density close to 1 and shows how this would look with respect to original dataset (green) and synthetic (blue). Visualising both the datasets side-by-side gives one an idea of how well the synthetic data mimics statistical properties of real dataset, allowing a complete summary of synthetic data quality. For continuous features, like 'Age' and 'BMI', as well as categorical ones, such as 'Smoking' and 'Diabetes', the synthetic data has distributions that closely follow those of the original data. This shows that the generator learns to approximate the trends and properties of the real data. Non-negligible differences are noticed in some features like 'Income' and 'Stress Level' which are either overwhelmed or understated within specific ranges of the original data by the point cloud of synthetic ones. Even these variations provide a more accurate fit but perhaps not entirely to the tails or details of the data distribution. The fidelity of the synthetic data to the original dataset indicates that it is suitable for training predictive models in this study. This guarantees that the synthetic dataset can be a valid replacement for real data when we have issues with privacy or small data availability.

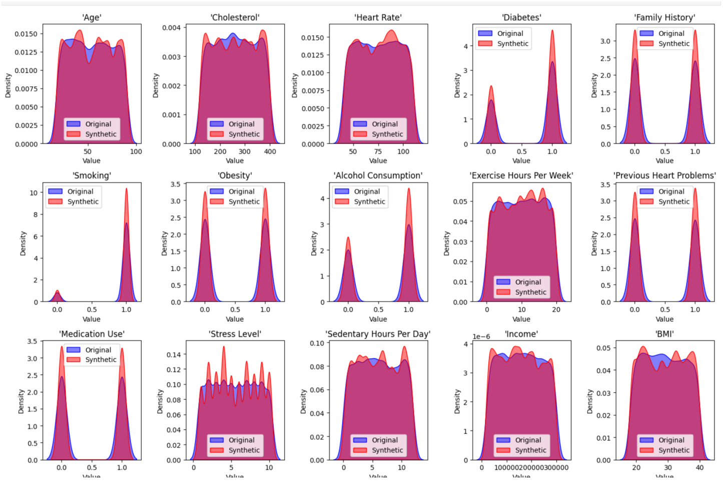
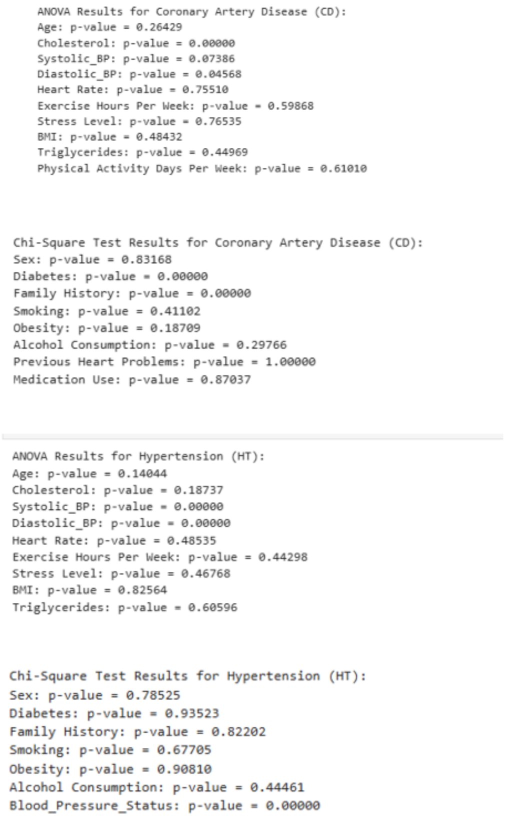


Figure 8: Original Vs. Synthetic Data

3.3.2 Statistical Test

Statistical tests are essential to exclude non-informative predictors of hypertension and CAD from the model, which may invalidate the final risk predictions as significant association is not proven between variables. Such variable had the clear statistically significant p-value good for making prediction tactics for instances like cholesterol, diastolic BP, diabetes and family history. This technique improves prediction accuracy and stability, according to research by Abdullahi (2024) as they emphasized that good feature select passing the predictive power of an excellent predictor.Abdullahi et al. (2024) Below (Figure 9) shows the Statistical tests for coronary artery disease which include two tests first one is ANOVA which we applied for Continues features and other is chi-squared test that applied for categorical features. In our statistical tests, CAD continuous variables of cholesterol and diastolic blood pressure are significant, and among categorical variables therefore from the category of datasets diabetes and family history also show significant associations. It may imply that these factors have, in fact, some importance regarding CAD risk or its presence. where we can see a relationship for continuous features like Cholesterol $p = 0.00000$ and Diastolic Blood Pressure (BP) $p = 0.04568$ have a meaning full relationship with CAD also Categorical features like Diabetes $p = 0.00000$ Family History $p = 0.00000$ have meaningful relationship with CAD which show how important this feature is. Below (Figure 9)Shows the statistical tests for hypertension which include two tests Anova and Chi -square tests. ANOVA applied for continues feature and Chi-square test applied for categorical features. Statistical tests for Hypertension (HT) show that continuous variables systolic and diastolic blood pressure are both significant, corroborating the close link between HT and the continuous versions of these parameters. The variable ”Blood Pressure Status”is highly regular $P < 0.001$, major it to be strongly associated in categorical analysis of HT. However this test since other parameters like cholesterol and diabetes are not linked to hypertension. In next section we will analyse the how data preprocessing helped for further analysis for coronary artery disease and Hypertension.



```
ANOVA Results for Coronary Artery Disease (CD):
Age: p-value = 0.26429
Cholesterol: p-value = 0.00000
Systolic_BP: p-value = 0.07386
Diastolic_BP: p-value = 0.04568
Heart Rate: p-value = 0.75510
Exercise Hours Per Week: p-value = 0.59868
Stress Level: p-value = 0.76535
BMI: p-value = 0.48432
Triglycerides: p-value = 0.44969
Physical Activity Days Per Week: p-value = 0.61010
```

```
Chi-Square Test Results for Coronary Artery Disease (CD):
Sex: p-value = 0.83168
Diabetes: p-value = 0.00000
Family History: p-value = 0.00000
Smoking: p-value = 0.41102
Obesity: p-value = 0.18709
Alcohol Consumption: p-value = 0.29766
Previous Heart Problems: p-value = 1.00000
Medication Use: p-value = 0.87037
```

```
ANOVA Results for Hypertension (HT):
Age: p-value = 0.14044
Cholesterol: p-value = 0.18737
Systolic_BP: p-value = 0.00000
Diastolic_BP: p-value = 0.00000
Heart Rate: p-value = 0.48535
Exercise Hours Per Week: p-value = 0.44298
Stress Level: p-value = 0.46768
BMI: p-value = 0.82564
Triglycerides: p-value = 0.60596
```

```
Chi-Square Test Results for Hypertension (HT):
Sex: p-value = 0.78525
Diabetes: p-value = 0.93523
Family History: p-value = 0.82202
Smoking: p-value = 0.67705
Obesity: p-value = 0.90810
Alcohol Consumption: p-value = 0.44461
Blood_Pressure_Status: p-value = 0.00000
```

Figure 9: Statistical Test Coronary Artery Diease And Hypertension

3.4 Data Transformation

In this sections we will see data transformation according to researcher Jeanson et al. (2024) Data Transformation is nothing but transformation of original data into format that well suitable for analysis or modelling .

3.4.1 Feature Selection

In this study we used method(Figure 10)co- relation-based feature selection for improving stability of model and we can avoid multicollinearity by performing corelation based feature selection. Based on co relation matrix we selected features like, Systolic Blood Pressure, Diastolic Blood Pressure, Cholesterol, Family history, Diabetes, Blood pressure status and so forth.

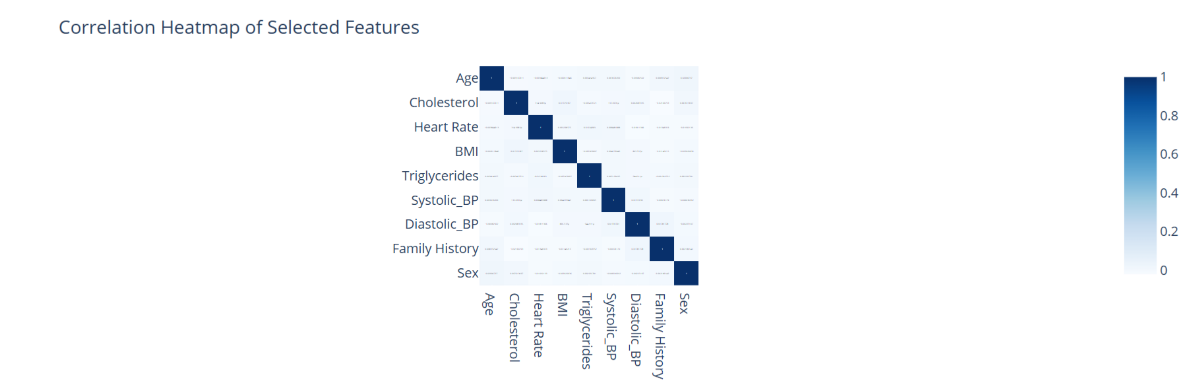


Figure 10: Statistical Test Coronary Artery Diease And Hypertension

3.4.2 Feature Engineering

In this study we splitted columns “Blood pressure” into systolic BP and Diastolic BP The systolic blood pressure measures the pressure in arteries when heart is pumping blood out to the body according to doctors normal systolic blood pressure is about less than 120$mmHg(millimetresofmercury)$ and higher is 140$mmHg(millimetresofmercury)$ or higher values than 140.However according to medical experts measures Diastolic Blood pressure measures the pressure in arteries when the heart is relaxed between the beats the normal range is for diastolic blood pressure is less than 80$millimetremercury(mmHg)$ and 90$mmHg$ or higher values is higher range for diastolic blood pressure according to new feature we add new feature “HT” which will be our Target variable for predicting Hypertension this feature takes both i.e. systolic blood pressure and diastolic blood pressure if both values is higher we allocate 1 and if the blood pressure is normal then we allocated 0 . Then we add another features “CD” for predicting coronary artery disease based on Family history and cholesterol level for “CD” features we checked in feature cholesterol level if it is equal to 240 $milligramsperdecimetre(mg/dL)$ or greater than 240mg/dL if its greater we allocate binary value 1 and if it is less than 240 then we allocate 0 values.Stevens et al. (2024)

3.5 Data Mining

According to researcher Polikar (2006)Data mining is the analysis of massive data sets in order to extract patterns, trends and useful information from them. This project is Exploration of the data using different Machine learning, and Predictive modelling. Some of the algorithms selected are K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Logistic Regression, Balanced Random Forest and Gradient Boosting Kumar et al. (2024) .We chose these methods because they have high accuracy with different types of data and enable us to improve our predictions and to capture nonlinear and linear relationship and complex patterns in data.Yadav et al. (2024) Furthermore, we are combining strength of each model, and it reduced each other individual model weakness. Researcher Polikar (2006) states that logistic regression struggled with nonlinear relationship which can limit its performance when we data complexity increases whereas Balanced Random search and gradient Boost can easily handle nonlinear relationship for identifying highly effective patterns in data we are using KNN and SVM and we also utilized ensemble techniques to join these models and increase their accuracy. Polikar (2006), Yadav et al. (2024),Kumar et al. (2024)In next section we will discuss algorithms or technique in detailed.

4 Design Specification

he following figure describe the model architecture (Figure 4) It trains an ensemble model on synthetic data for prediction complements the proposed architecture. The process starts with the creation of synthetic data, which acts a proxy for real-world data where it is not practical to obtain or use real-world data often because it is too sensitiveGarcía-Vicente et al. (2022). For base classifiers, we use 4 machine learning models: K-nearest Neighbours (KNN), Logistic Regression, Gradient Boosting and Support Vector Machines (SVM). And we used Balanced Random Search on top of each model so that the best hyperparameters are achieved on performance. A soft voting classifier combines these base models, where each model's output (predicting the class) is weighted based on its probabilities and the class with the highest overall probability is chosen Kumar et al. (2023). As a result, it composes an ensemble model that combines the powers of each single classifiers and produces more precise and robust predictions. The Ensemble architecture uses different models to learn separate patterns from the data and utilizes the strength of each model to minimize errors and enhance generalization and robustnessAbdullahi et al. (2024). This design ensures that even synthetic data-based training will yield a scalable and effective predictive model.

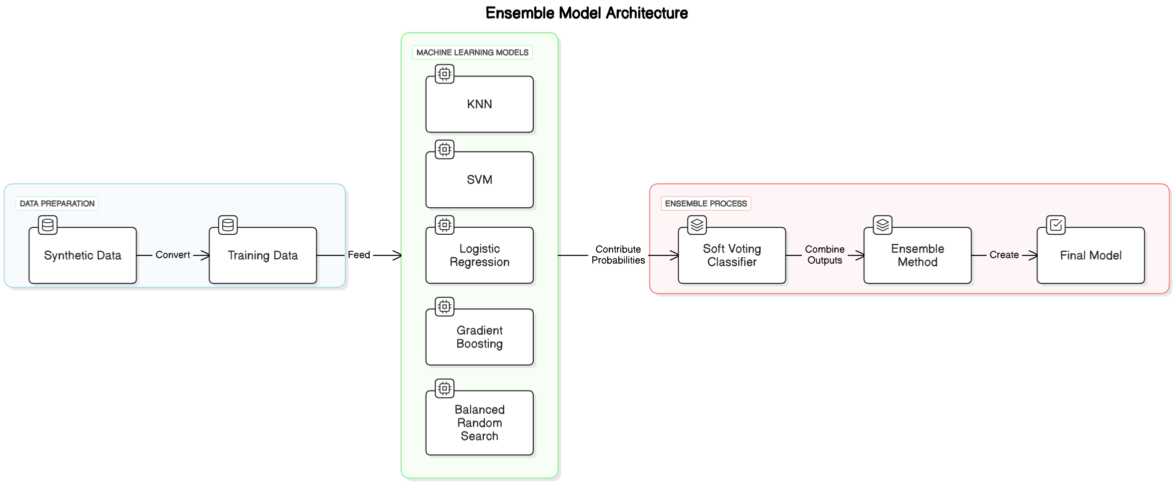


Figure 11: Ensemble Model Architecture

We will discuss the implementation details of our approach in the following section. It will cover a detailed description of the approaches, models and frameworks used in constructing the system providing an outline of each step performed to train/evaluate/optimize the ensemble model. The emphasis is going to be on showing how the theoretical concepts previously discussed were made practical.

$$P(y=c) = w_1 P_{LR}(y=c) + w_2 P_{KNN}(y=c) + w_3 P_{SVM}(y=c) + w_4 P_{GB}(y=c) + w_5 P_{BRS}(y=c) \quad (1)$$

Where: $PLR(y = c)$: Probability predicted by Logistic Regression. $PKNN(y = c)$: Probability predicted by KNN. $PSVM(y = c)$: Probability predicted by SVM $PGB(y = c)$: Probability predicted by Gradient Boosting. $PBRS(y = c)PRFC(y = c)$:Probability predicted by Balanced Random Search W1,W2,W3,W4,W5 weights for each model (default is equal, i.e., w=1N w= N1). For better understanding, we took weight, i.e., w, which go along with each model with w1,w2,w3,w4 and so on .According to Dietterich (2000)When we set the weights to 1 (meaning they all contribute equally to the ensemble prediction), i.e default weights it prevents any one model from overpowering the others. This is especially helpful as a baseline in the situations we have no idea which model performs better or contributes more to the prediction. Also,Researcher Zhou (2012) states that equal weights would not bias the ensemble towards a particular model, which would be a major concern for healthcare studies because these studies try to aggregate the perspectives of different algorithms like logistic regression, KNN, SVM, etc. Moreover study Conducted by Polikar (2006),states that it allows for simplified ensemble design furthermore researcher Kumar et al. (2023) says weighted ensemble models can outperform baseline methods of testing based on various weights, but starting with equal weights allows us to evaluate the win ratio of the combined model overall without complicating matters of weight tuning. For example, research has shown that mixing algorithms in equal weights does well compared to each individual one since each algorithm has its strengthes and weaknesses Zhou (2012).

Let's take an example Assume Logistic Regression says 60% chance, KNN says 70%, SVM says 50% and Gradient Boosting says 80% for "Hypertension" — the mean will be 65%. Because 65% is greater than that probability for "No Hypertension",the "No Hypetension "Probability is based on 1-Probability of hypertension so the Logistic regression says 40% chance ,knn says 30% chance and svm says 50% chance and gredient boosting says 20% chance after calculating the probababilty for "No Hypertension" is 35% .So the 65% is greter than "No-Hypertesnion" the model state the case is "Hypertension". Same goes for" coronary artery disease "prediction model.

[2]

5 Implementation

In this section, in order to achieve a good result we used Python Programming language with version 3.11.7, and for creating a web framework, we used Flask, a lightweight WSGI web framework with Hypertext Markup Language (HTML), CSS and JavaScript. In the following image (Figure 5) shows the implementation architecture that describe how each section works practically. Starting off with Input Data, we make use of real-world or synthetic datasets as a basis for model training and evaluation. The data will pass through Data Preprocessing in which it identifies cleaning, normalization, categorical encoding and missing value handling, to produce a standardized form of the data that is compatible with ML models. After this, Synthetic Data Generation using advanced methodologies (e.g. GANs) increases the number of samples in the dataset, balances classes

[2] https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.VotingClassifier.html

distributions and reduces restrictions from low data volume or protection from patient privacy by generating only sensitive data synthesizing new ones. All this is provided to a package of Machine Learning Algorithms comprising Logistic Regression, K-Nearest Neighbour (KNN), Support Vector Machines (SVM) and Gradient Boosting. Each of these base models operates independently, from the training to hyperparameter optimization using Balanced Random Search to aim for best performance. Trained model results are aggregated through an Ensemble Method. This ensemble uses a Soft Voting Classifier that takes the result of each base model and averages the probability to reach the final prediction. We also applied methods including but not limited to L2 Regularization to reduce the chance of an overfitting producing and 5-Fold Cross-Validation is conducted to confirm that the ensemble model generalizes well on unseen data. Through Train vs. Test Accuracy, we evaluate the performance of a model, which can help find overfitting or underfitting as issues. Final Model is serialized (e.g., joblib file) to deploy the optimized model. These model gets hosted onto a Web Framework built on Flask to make real time predictions and interactions with the user. Login Page help the web app requires users to log in before being able to use it. The Welcome page allows users to navigate through the application. The Result Page Displays predictions. Backend MySQL Database It supports the backend of the application, saving user inputted data, results from predictions and other data across the application. This framework encompasses the entire workflow, from preprocessing data to deploying a model and is a comprehensive solution for creating simple but scalable predictions in one place.

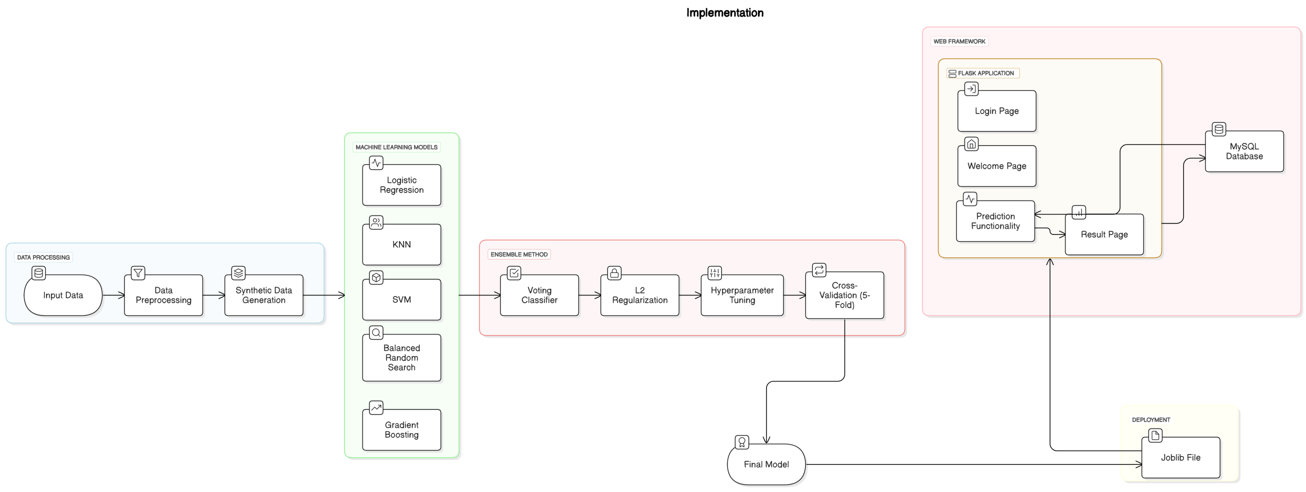
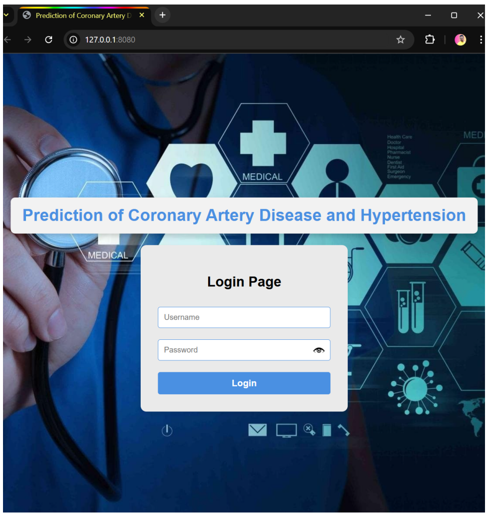


Figure 12: Implementation

5.1 Practical Implementation and User Interface

In Medical or in an clinical environment, web application can solve many problems and its worthy to used in daily life according to the researcher Menon et al. (2022) web application was built to prove the usefulness of the ensemble models developed. The interface ensures seamless interaction where users can input patient data, retrieve predictions, and visualize the results. Bridging the gap between theoretical modeling and real-world clinical utility, the application is designed to be both accessible and scalable. Here are the main parts of the interface:

Login Page A login page is important as it is a gateway page, where every user would enter to access your application securely. Authentication – among the most essential security measures used in systems, users need to sign in with valid credentials. This page was created in a way to ensure data privacy and that only an authorized entity can view the data per the norms of data security in the healthcare industry..(Figure 5.1)



Prediction of Coronary Artery Disease and Hypertension

Login Page

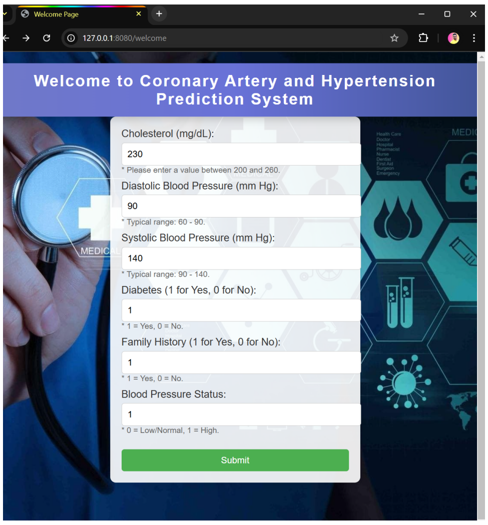
Username

Password

Figure 13: Login Page

5.1.1 Welcome Page

The welcome page shows you what the application is all about, how in medical environment use the application. It is a navigation panel that directs the users to different site sections, like input forms for predictions and the dashboard for results (Figure 5.1.1)



Welcome to Coronary Artery and Hypertension Prediction System

Cholesterol (mg/dL): 230
* Please enter a value between 200 and 260.

Diastolic Blood Pressure (mm Hg): 90
* Typical range: 60 - 90.

Systolic Blood Pressure (mm Hg): 140
* Typical range: 90 - 140.

Diabetes (1 for Yes, 0 for No): 1
* 1 = Yes, 0 = No.

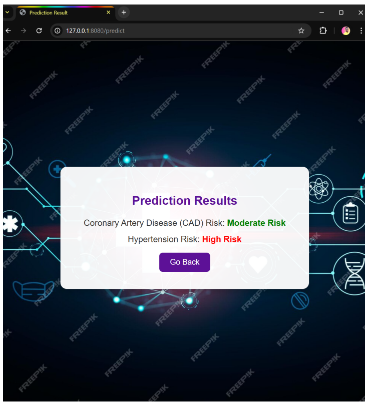
Family History (1 for Yes, 0 for No): 1
* 1 = Yes, 0 = No.

Blood Pressure Status: 1
* 0 = Low/Normal, 1 = High.

Figure 14: Welcome Page

5.1.2 Result Page

This has produced predictions that are presented on the result page from the ensemble models shows in (Figure 5.1.2) . The output of detailed information, including probability for coronary artery disease (CD) and Hypertension (HT) and visual insights such confidence score and feature importance can be viewed by the users.
[3]



Prediction Results

Coronary Artery Disease (CAD) Risk: Moderate Risk

Hypertension Risk: High Risk

Figure 15: Result Page

6 Evaluation and Result

The ensemble models built for predicting coronary artery disease (CD) and Hypertension (HT) performed excellently on both training and testing datasets, as visualized on the Confusion Matrices (Figure 6) and the ROC Curves (Figure 6). For the CD model, we received a training accuracy of 92% and a testing accuracy of 94%, indicating that the model generalizes well to unseen data. The confusion matrix (Figure 6). shows that the model was able to accurately identify most cases, capturing 320 true negatives, 265 true positives, 25 false positives, and 27 false negatives. The balanced precision (0.92), recall (0.92), and F1-scores suggest trustworthy performance for both classes (No CD and CD). The probability of goodness of fit of the model is supported by ROC curve (Figure 6)., where AUC score of 0.97 indicates the good separation ability of the model between classes Although it showed great performance, we also observed some false positives and false negatives where we could add discriminative features to improve it even more. The HT model performed even better, presenting an accuracy of 97% of the training and testing datasets. As depicted in the confusion matrix , almost all the samples were classified correctly, with 3850 true negatives, 3861 true positives, 104 false positives and 99 false negatives. All precision, recall and F1-scores were very high (0.97) for both classes i.e. No HT and HT, showing it is very reliable and accurate in predicting hypertensive cases. In conclusion, the model HT received an AUC score of 0.98 (Figure 6). on the ROC curve corroborating its strength and relative generalization. The results validate that the ensemble models proposed can effectively address the research problem by providing accurate predictions of coronary artery disease and Hypertension. The HT model performed marginally better than CD; this may be owing to more distinct class separability in the dataset, or less complicated feature interactions. ROC curves for both

[3] https://www.python.org/downloads/release/python-3117/

models show high AUC scores of 0.923 and 0.926 confirming strong discriminatory power. Both the models performed quite well, however the misclassifications observed suggest potential areas of improvement in terms of feature engineering and data preprocess. The implications of these findings are great academically and practically. Theoretically, they confirm the usefulness of ensemble methods for complex healthcare prediction problems. From a practical point of view, our models exhibit great promise for implementation as a part of clinical decision-support systems for sensitive conditions, as they are capable of reliably producing correct and timely predictions. In the following section, we will discuss the general implications of these findings, and possible avenues for future work.

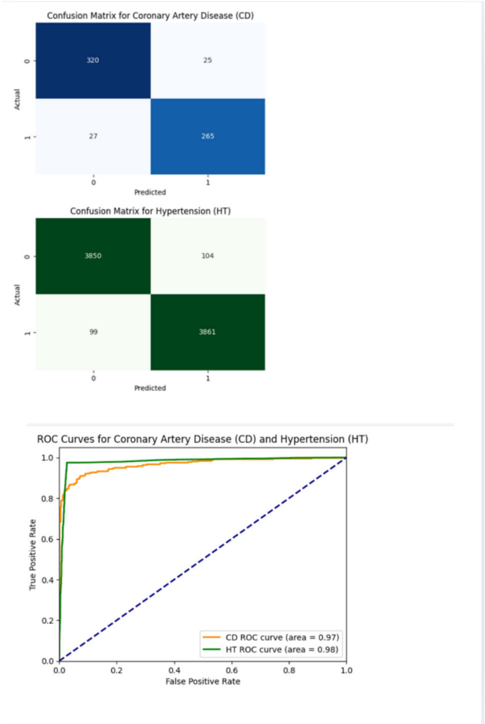


Figure 16: Confusion Matrix and ROC Curve

6.1 Experiment

In this section, we elaborate the experiments performed to answer the research questions from this study properly. These sets of experiments aim to compare different predictive models for coronary artery disease (CAD) and for Hypertension (HT). We explored the effect of feature selection, data preprocessing, and algorithm choice on model accuracy and generalizability through repeated experimentation. All experiments were designed to yield important information regarding the important features driving predictions and their clinical relevance. The following are the table will give a how this study improved in modelling stage.

6.1.1 Experiment 1

The following experiments were conducted on a data set with class imbalance to check the robustness and performance of the model over both classes (major and minor) to get pretentious precision, recall, and f1-score measurements.

Model	Class	Precision	Recall	F1-Score	Support
CD	Class 0	1.00	1.00	1.00	117
	Class 1	1.00	1.00	1.00	1636
HT	Class 0	1.00	1.00	1.00	554
	Class 1	1.00	1.00	1.00	1199

Table 2: Experiment 1

From the above Table 6.1.1, we can confirm that from the first experiment, this model was overfit by calculating ratio for both model(ratio for hypertension(HT) 1.2.2 and coronary artery diease (CD) 1.14). According to Mahmood et al. (2024)The overfitting occurred when the model memorized the data instead of identifying common patterns in the data. In next experiment, we did handled class imbalance using ctgan library, crated synthetic data and validate second model on synthetic data.

6.1.2 Experiment 2

The accuracy of this model (applying 5-fold cross-validation) significantly improves the previous one, as it reduces overfitting and shows better generalization we can show in Table 6.1.2. The CD model had a test accuracy of 0.910.910.91, with balanced performance per class. The HT model also had a test accuracy of 0.970.970.97 and a uniform metric for both classes. This model's realistic and solid performance across folds suggests its robustness and generalization power to out-of-sample data—in contrast to the previous overfit model, which created artificially perfect metrics.Fernández et al. (2018)

Model	Class	Precision	Recall	F1-Score	Support
CD	Class 0	0.90	0.93	0.92	345
	Class 1	0.91	0.88	0.90	292
HT	Class 0	0.97	0.97	0.97	3954
	Class 1	0.97	0.97	0.97	3960

Table 3: Experiment 2

6.1.3 Final Result

By using L2 Regulairization ,5-Fold Cross validation and Hyperparameter combined the models Performance was better than previous one which shown in Table 4.Fang et al. (2023)Gupta et al. (2022) The high accuracies for CD(0.92) and HT(0.97) directly translates to balanced metrics across classes, with very few amounts of errors. This facilitates a more solid and dependable experience for real-world applications.

Model	Class	Precision	Recall	F1-Score	Support
CD	Class 0	0.92	0.93	0.92	345
	Class 1	0.91	0.91	0.91	292
HT	Class 0	0.97	0.97	0.97	3954
	Class 1	0.97	0.97	0.97	3960

Table 4: Final Results

In the next section we will discuss how this study achived good results.

6.2 Discussion

This study confirmed the efficiency of the applied ensemble models such that we achieved testing accuracies of 94% for coronary artery disease (CD) and 97% for Hypertension (HT). Models being robust also being justified by balanced precision, recall and f1-scores which signifies that these models are not biased towards any class and give reliable predictions. This finding indicates that the HT model has higher capacity for generalization across heterogeneous patient profiles, while the CD model's marginally lower accuracy suggests capacity for further improvement around feature selection and engineering. One of the primary strengths of this study is an ensemble learning using Logistic Regression, KNN, SVM and Gradient Boosting with the soft votingAbdullahi et al. (2024). It combines the best aspects of each model, reduces their drawbacks, and improves the overall accuracy. The performance obtained from the models was highly attributed to hyperparameter tuning and the addition of synthetic data, which were considered to alleviate data scarcity and guarantee a balanced dataset. Though the models exhibited very high performance, small areas of misclassification indicate improvements can be made, specifically testing the models against larger, real-world datasets to further ensure generalizability. And while practical, the reliance on synthetic data may reduce applicability in cases with high data variability. These results have great significance. From an academic point of view, they showcase the promise of ensemble methods for predicting healthcare needs, and they provide a target for further studies. On a more practical note, the models show potential as diagnostic tools, providing accurate and timely predictions that may improve clinical decision-making and patient care. Further improvement of these models may motivate evolution of more sophisticated, scalable diagnostic systems.

7 Conclusion and Future Work

This work has shown a powerful and scalable approach to healthcare diagnostics by utilizing advanced algorithms and synthetic data. While predictive models have been well trained, existing models still face challenges, especially when dealing with imbalanced datasets where high risk cases are rare Singh et al. (2024),Perkonoja (2020). Ensemble models were developed to predict CD and HT with testing accuracies of 94% and 97%, respectively. The ensembling approaches of soft voting achieved strong and reliable performance by combining predictions for Logistic Regression, KNN, SVM, and Gradient Boosting with generated synthetic data, which could address the data scarcity and bias in the models. This study has important ramifications for clinical practice as it covers generalizable high-performance models that can be implemented into clinical workflows for better diagnostic and resource allocation. The study takes on limitations; the use of synthetic data as well as small misclassifications indicate design elements that could be improved upon, yet the research offers a groundwork for future research.Last but not least the models not only advance academic research but also have the potential for clinical usage, enabling tangible benefits to patient care.

7.1 Future Work

Involving and testing with more complex, real-world datasets, scaling to multi-class problems, and incorporating explainable AI can be improve interpretability. further more testing each individual model with weights and adding this weights to ensemble model either manually or using automatically via meta learner approach could be benifiicial in Preditcing Coronary Artery Diease and Hypertension.

References

Abdullahi, A., Ali Barre, M. and Hussein Elmi, A. (2024). A machine learning approach to cardiovascular disease prediction with advanced feature selection, *Indonesian Journal of Electrical Engineering and Computer Science* **33**(2): 1030.

Acherjya, G. K., Tarafder, K., Dutta, D., Mortuza, M. G., Sarkar, A. K., Das, N. L. and Ali, M. (2023). Frequency and risk factors stratification of hypertension among the rural population of bangladesh, *Journal of Family Medicine and Primary Care* **12**(10): 2488–2495.

Alqulaity, M. and Yang, P. (2023). Tabular generative adversarial networks with an enhanced sampling approach for high-quality cardiovascular disease dataset generation, *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, IEEE, pp. 2595–2600.

Arora, A. and Arora, A. (2023). Machine learning models trained on synthetic datasets of multiple sample sizes for the use of predicting blood pressure from clinical data in a national dataset, *Plos one* **18**(3): e0283094.

Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S. and Zhou, S. (2019). A machine-learning-based prediction method for hypertension outcomes based on medical data, *Diagnostics* **9**(4): 178.

Choi, Y. Y., Jeong, H., Lee, J. H., Sung, K. C., Shin, J.-H., Kim, H. C., Kim, J. Y. and Kang, D. R. (2021). Cardiovascular disease prediction model in patients with hypertension using deep learning: Analysis of the national health insurance service database from republic of korea, *CardioMetabolic Syndrome Journal* **1**(2): 145–154.

Chowdhury, M. Z. I., Naeem, I., Quan, H., Leung, A. A., Sikdar, K. C., O'Beirne, M. and Turin, T. C. (2022). Prediction of hypertension using traditional regression and machine learning models: A systematic review and meta-analysis, *PloS one* **17**(4): e0266334.

Cuevas-Chavez, A., Hernandez, Y., Ortiz-Hernandez, J., Sanchez-Jimenez, E., Ochoa-Ruiz, G., Perez, J. and Gonzalez-Serna, G. (2023). A systematic review of machine learning and iot applied to the prediction and monitoring of cardiovascular diseases, *Healthcare*, Vol. 11, MDPI, p. 2240.

Dietterich, T. G. (2000). Ensemble methods in machine learning, *International workshop on multiple classifier systems*, Springer, pp. 1–15.

Fang, M., Chen, Y., Xue, R., Wang, H., Chakraborty, N., Su, T. and Dai, Y. (2023). A hybrid machine learning approach for hypertension risk prediction, *Neural Computing and Applications* **35**(20): 14487–14497.

Fernández, A., del Río, S., Chawla, N. V. and Herrera, F. (2018). *Learning from Imbalanced Data Sets*, Springer.

García-Vicente, C., Chushig-Muzo, D., Mora-Jiménez, I., Fabelo, H., Gram, I. T., Løchen, M.-L., Granja, C. and Soguero-Ruiz, C. (2022). Clinical synthetic data generation to predict and identify risk factors for cardiovascular diseases, *VLDB workshop on data management and analytics for medicine and healthcare*, Springer, pp. 75–91.

Gupta, A., Kumar, R., Arora, H. S. and Raman, B. (2022). C-cadz: computational intelligence system for coronary artery disease detection using z-alizadeh sani dataset, *Applied Intelligence* **52**(3): 2436–2464.

Jaiswal, S., Natarajan, P., Silver, A. J., Gibson, C. J., Bick, A. G., Shvartz, E., McConkey, M., Gupta, N., Gabriel, S., Ardissino, D. et al. (2017). Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease, *New England Journal of Medicine* **377**(2): 111–121.

Jeanson, F., Farkouh, M. E., Godoy, L. C., Minha, S., Tzuman, O. and Marcus, G. (2024). Medical calculators derived synthetic cohorts: a novel method for generating synthetic patient data, *Scientific Reports* **14**(1): 11437.

Kumar, A., Singh, K. U. and Kumar, M. (2023). A clinical data analysis based diagnostic systems for heart disease prediction using ensemble method, *Big Data Mining and Analytics* **6**(4): 513–525.

Kumar, S., Sharma, A., Kaur, N., Pawar, L. and Bajaj, R. (2024). *Optimized Predictive Models in Health Care Using Machine Learning*, Wiley-Scrivener.

Mahmood, T., Rehman, A., Saba, T., Alahmadi, T. J., Tufail, M., Bahaj, S. A. O. and Ahmad, Z. (2024). Enhancing coronary artery disease prognosis: A novel dual-class boosted decision trees strategy for robust optimization, *IEEE Access* .

Menon, A. M., Jyothi, G. N., Bhavana, K. and Saara, K. (2022). Flask based web app on diabetes prediction using machine learning, *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*, Springer, pp. 785–793.

Morgan, B., Murali, A. R., Preston, G., Sima, Y. A., Marcelo Chamorro, L. A., Bourantas, C., Torii, R., Mathur, A., Baumbach, A., Jacob, M. C. et al. (2023). A physics-based machine learning technique rapidly reconstructs the wall-shear stress and pressure fields in coronary arteries, *Frontiers in Cardiovascular Medicine* **10**: 1221541.

Mroz, T., Griffin, M., Cartabuke, R., Laffin, L., Russo-Alvarez, G., Thomas, G., Smedira, N., Meese, T., Shost, M. and Habboub, G. (2024). Predicting hypertension control using machine learning, *Plos one* **19**(3): e0299932.

Park, J.-B., Lee, H.-J., Yang, H.-L., Kim, E.-H., Lee, H.-C., Jung, C.-W. and Kim, H.-S. (2023). Machine learning-based prediction of intraoperative hypoxemia for pediatric patients, *Plos one* **18**(3): e0282303.

Perkonoja, K. (2020). Generating synthetic longitudinal patient data with the privbayes method.

Petmezas, G., Stefanopoulos, L., Kilintzis, V., Tzavelis, A., Rogers, J. A., Katsaggelos, A. K. and Maglaveras, N. (2022). State-of-the-art deep learning methods on electrocardiogram data: systematic review, *JMIR medical informatics* **10**(8): e38454.

Polikar, R. (2006). Ensemble based systems in decision making, *IEEE Circuits and systems magazine* **6**(3): 21–45.

Pramanik, A., Rajput, P. and Aluvala, S. (2023). Applying healthcare analytics to diagnose and predict coronary artery disease using machine learning techniques, *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, IEEE, pp. 610–614.

Singh, M. S., Thongam, K., Choudhary, P. and Bhagat, P. (2024). An integrated machine learning approach for congestive heart failure prediction, *Diagnostics* **14**(7): 736.

Stevens, C. A., Vallejo-Vaz, A. J., Chora, J. R., Barkas, F., Brandts, J., Mahani, A., Abar, L., Sharabiani, M. T. and Ray, K. K. (2024). Improving the detection of potential cases of familial hypercholesterolemia: Could machine learning be part of the solution?, *Journal of the American Heart Association* p. e034434.

Tompra, K.-V., Papageorgiou, G. and Tjortjis, C. (2024). Strategic machine learning optimization for cardiovascular disease prediction and high-risk patient identification, *Algorithms* **17**(5): 178.

Wang, E., Mott, K., Zhang, H., Gazit, S., Chodick, G. and Burcu, M. (2024). Validation assessment of privacy-preserving synthetic electronic health record data: Comparison of original versus synthetic data on real-world covid-19 vaccine effectiveness, *Pharmacoepidemiology and Drug Safety* **33**(10): e70019.

Yadav, J., Nair, A. M., George, J. and Alapatt, B. P. (2024). Predictive modelling of heart disease: Exploring machine learning classification algorithms, *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, pp. 1–7.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*, CRC press.