

Symptom-Based Disease Detection by Applying Bayesian Probability and Machine Learning

*Note: Sub-titles are not captured in Xplore and should not be used

Student Name – Kuber Dhama
Student Number - 220996071
Supervisor Name – Dr. Paulo Rauber
Programme of study – FT Msc.
Computer Science

Abstract— Machine Learning has revolutionized the healthcare domain by enabling the detection of diseases with remarkable accuracy. Leveraging the power of statistical models, machine learning techniques are widely employed by organizations to automate disease detection using historical patient records. In this research, the focus has been on detecting heart disease using patient records from the esteemed UCI dataset, which includes 303 records from Cleveland, Hungary, and Switzerland. This dataset is highly regarded and has been extensively utilized by researchers. To prepare the data for heart disease detection, the datasets were merged and preprocessed, and important features were selected using a combination of correlation and the Bayesian model. Five machine-learning models, inspired by literature reviews, were then applied to the data. Ultimately, the decision tree model emerged as the most accurate, detecting heart disease with an impressive 95.21% accuracy, surpassing existing approaches in the field.

Keywords—heart disease detection, Bayesian method, machine learning, classification

I. INTRODUCTION (HEADING 1)

A. Overview of Project

Heart disease is a major health issue worldwide, and early detection is crucial for effective treatment. Recognizing important symptoms can help identify heart disease early on (Aggarwal & Kumar, 2022). Some warning signs include chest pain or discomfort, shortness of breath, nausea, and extreme fatigue. Additionally, high blood pressure, high cholesterol, and diabetes are risk factors that increase the likelihood of developing heart disease. Regular check-ups with a healthcare provider can help detect and monitor these risk factors (Aljaaf & Jumeily, 2015). With early detection and proper management, heart disease can be prevented or treated effectively. It is important to prioritize heart health and take steps to reduce the risk of heart disease.

B. Research Scope

In recent years, the incorporation of machine learning (ML) techniques in healthcare has revolutionized the process of disease detection and management. In particular, heart disease, as the leading cause of death worldwide, has become a primary target for the application of ML algorithms. The application of machine learning in heart disease detection has shown immense potential to improve diagnostic accuracy and reduce healthcare costs. Various algorithms are capable of analyzing electrocardiogram (ECG) data, blood tests, and other patient information to deliver fast and accurate results (Andreu-Perez, et al., 2015). This made it easier for

healthcare providers to identify risk factors and recommend early interventions, which could potentially lower mortality rates. Moreover, ML-based systems also show the potential to provide clinicians with valuable insights into the development and progression of heart diseases (Gupta, et al., 2020).

Additionally, the integration of wearable and mobile health technologies can facilitate the continuous monitoring of patients with heart diseases (Gupta et al., 2020). Data from these devices can be integrated with ML algorithms to identify early warning signs, such as abnormal heart rates or blood pressure levels, and automatically notify healthcare providers and patients to take pre-emptive measures.

C. Challenges and Potential Issues

Despite the promising potential of ML in heart disease detection, certain challenges need to be addressed to optimize the technology's impact. One major challenge is the quality and reliability of patient data. For an ML-based diagnostic tool to be accurate and effective, large amounts of high-quality and standardized data are required (Mortazavi, et al., 2016). There could be potential errors or inconsistencies in the data collected, ranging from measurement inaccuracies to missing and incomplete information, posing difficulties when developing and validating ML models.

Another challenge is the generalizability of ML models, which might not work effectively across various geographic and demographic populations (Gupta, et al., 2020). The development of ML algorithms using biased datasets might lead to skewed results and inaccurate predictions when applied to diverse populations, leading to potential misdiagnosis and inappropriate medical interventions.

D. Project Particulars

1) Research Questions

The research question for the project is discussed below:

Research Question-1. How the necessary and indicative symptoms can be identified for heart disease at the early stage?

Research Question-2. What are the suitable processes that can be applied to prepare the data through which the best accuracy for heart disease detection can be achieved?

Research Question-3. Is there any improvement in the approach for heart disease detection compared to the existing methodologies?

2) Project Aim

The aim of the project is to detect disease by applying the Bayesian probability factor in choosing necessary symptoms based upon which the disease will be detected with the application of Machine Learning.

3) Objectives of Project

The objectives of the project to satisfy the aim and to address the research questions are discussed below

- To study the disease and the underlying symptoms for which the disease can be seen in patients.
- To review the existing research papers to gather the necessary knowledge regarding the application of machine learning for disease detection by emphasizing symptoms.
- To select the disease dataset wherefrom the historical records of the patients regarding that disease can be obtained and study the features from there.
- To apply the Bayesian model to determine the probability of the essential features of the disease and identify the influence of the symptoms on the disease.
- To apply the chosen algorithms to classify and detect the disease along with the determination of classification metrics such as accuracy, precision etc. so that the comparison of performances can be done and the best-performing model can be chosen.

II. LITERATURE REVIEW

A. Feature Engineering And Applications

1) Feature Selection Using Correlation

(Razak, et al., 2022) employed the Pearson correlation method for the selection of necessary features from the data. The researchers set forth a procedure that utilizes the Pearson relationship coefficient to pick relevant qualities online from a dataset progressively as new information opens up. The algorithm determines each feature's Pearson correlation with the target variable, and features with higher correlation values are chosen as relevant for prediction. Execution measures including accuracy, precision, recall, and F1-score were utilized to assess the methodology on a dataset and survey how precise the picked highlights were. On the test set, for example, the proposed procedure accomplished an accuracy of 87%, precision of 88%, recall of 85%, and F1-score of 87%.

(Zhang, et al., 2018) suggested a way for choosing pertinent features from meteorological data in order to create accurate prediction models. This method makes use of information value and maximum correlation measurements. Each feature's information value and maximum correlation with the goal variable (in this case, a forecast of the weather) are calculated by the algorithm, and characteristics with greater information value and highest correlation values are chosen as important features for the prediction model. Utilising performance criteria like accuracy, precision, recall, and F1-score, the method's efficacy was assessed using a meteorological dataset and the selected features. For instance, on the test set, the suggested technique obtained accuracy, precision, recall, and F1-score of 88%, 86%, and 89%, respectively, demonstrating its efficacy in picking pertinent features for meteorological prediction.

2) Feature Selection Using Feature Elimination Process

(Adorada, et al., 2021) proposed selecting microRNA expression characteristics in breast cancer using the Support Vector Machine - Recursive Feature Elimination (SVM-RFE) approach. The most pertinent microRNA expression features for breast cancer prediction were chosen by the authors using SVM-RFE, an iterative feature selection approach that combines support vector machine classification with recursive feature elimination. Until the ideal set of features is obtained, the algorithm eliminates the least significant features in each iteration based on their ranking by SVM weights. Performance criteria like accuracy, precision, recall, and F1-score were used to assess the suggested method's performance on a dataset of breast cancer cases. For instance, the SVM-RFE approach produced an F1-score of 92% and test set accuracy, precision, recall, and recall of 91%.

(Zhao, et al., 2022) outlined a study that proposes a Recursive Feature Elimination (RFE) based feature selection approach to improve the performance of classifying multiple-cause deaths in colorectal cancer. To categorise multiple-cause fatalities in colorectal cancer, the authors used RFE, a prominent feature selection technique, to iteratively remove less pertinent information and determine the most useful features. The programme classified the fatalities with numerous causes using RFE in conjunction with a classification technique that might be stated in the publication, such as logistic regression or support vector machine (SVM). The suggested method was assessed using a colorectal cancer-related dataset, and the classification accuracy was assessed using suitable performance indicators. For instance, the RFE-based feature selection method classified multiple-cause fatalities from colorectal cancer with an accuracy of 87%, demonstrating its efficacy in enhancing the performance of the classification model.

3) Feature Selection Using Ensemble Model

(Al-Mashagbeh & Ababneh, 2021) suggested the application of Correlation-based Feature Selection (CFS) with the Best First search and Random Forest algorithm in combination to detect Tor. To find the most pertinent properties for Tor identification, the authors used CFS, a well-liked feature selection method. To efficiently find the best feature subset, the Best First search method was combined with CFS. The classifier for Tor detection used an ensemble learning system called Random Forest. Using a relevant dataset linked to Tor detection, the proposed approach's accuracy was assessed. Appropriate performance criteria were employed to gauge the classification model's correctness. For instance, the suggested method identified Tor traffic with a 95% accuracy rate, proving its effectiveness in identifying Tor traffic using a combination of feature selection and machine learning techniques.

B. Heart Disease Detection

1) Using Machine Learning

(Sahoo, et al., 2022) framed a review that presents an AI-based coronary illness expectation model for home customized care. The authors used decision trees, random forests, support vector machines, and other machine learning techniques to create a predictive model for the prediction of heart disease. A dataset of patient well-being records was utilized in the review to prepare and test the model. The model's efficiency was evaluated using a number of performance metrics, including accuracy, precision, recall, and the F1 score. In order to produce a precise and

dependable model for the prediction of heart disease, the study's algorithm included steps for feature selection, model training, and model evaluation. For instance, the study correctly predicted heart disease with an accuracy of 92% and a precision of 95%, demonstrating the effectiveness of the suggested machine learning-based approach for individualized home care.

(Atallah, 2019) proposed a machine learning majority voting ensemble strategy for a heart disease diagnosis model is described in this article. For the purpose of diagnosing cardiac illness, the investigators employed an ensemble model using various machine-learning techniques. A large collection of patient well-being information was used to create as well as assess the model. Performance criteria including precision, recollection, precision, as well as the F1 score were used to assess the algorithm's effectiveness. The heart disease detection model's accuracy and robustness can be improved by integrating the predictions of multiple base classifiers using the proposed ensemble method's majority voting approach. For instance, the study's effectiveness is demonstrated by its 94% accuracy in identifying cardiac disease.

The researcher (Khan, 2020) used two machine-learning techniques to create a cardiac disease diagnosis model: support vector machines (SVM) and k-nearest neighbours (KNN). For the purpose of training and evaluating the model, a clinical dataset comprised of patient demographics, medical histories, and diagnostic test results. The area under the curve (AUC), accuracy, sensitivity, and specificity were the model's performance metrics. The proposed model achieved an accuracy of 86%, demonstrating the potential of data-driven heart disease identification and the efficacy of machine learning in this context.

(Yadav, et al., 2020) have detected cardiac disease with the application of machine learning based on the patient database with symptoms. Using a variety of methods and calculations, the researchers developed expectation models for coronary disease proof using support vector machines, k-nearest neighbours, and choice trees. In order to train and test the models, the researchers made use of a dataset that contained significant variables like patient demographics, medical history, and diagnostic test results. Performance was assessed utilizing exactness, awareness, particularity, and other huge measurements. The proposed models' high levels of accuracy support the study's findings that machine learning can be used to diagnose cardiac disease. This study adds to the developing group of examination on AI-based coronary illness location by featuring its true capacity for exact and proficient recognition.

2) Using Hybrid Models

A heart disease prediction model based on a hybrid machine learning strategy employing decision trees and neural network algorithms is described by (Bakhshi, et al., 2022). The researchers proposed a strategy that joins the choice tree and brain network calculations to improve the precision of coronary illness forecast. The hybrid model was trained on a dataset of heart disease cases that was used in the study. The proposed model's accuracy was evaluated, and the findings pointed to promising outcomes. For example, the creators detailed a precision of 91% in accurately foreseeing coronary illness cases utilizing their cross-breed AI model, showing the capability of their methodology in further developing coronary disease expectations and findings.

(Esfahani, 2017) have employed machine learning models to detect cardiovascular disease based on symptom data of patients. The authors proposed a novel approach that makes use of an ensemble classifier, which combines multiple base classifiers, to improve disease detection accuracy. The study classified and predicted the presence or absence of cardiovascular disease based on a dataset of cases using the ensemble classifier. The accuracy of the proposed method was reported, and the outcomes were promising. Using their proposed ensemble classifier, the authors were able to correctly identify cardiovascular disease cases with an accuracy of 87%, demonstrating the potential of their strategy to improve disease detection and prediction.

(Geweid, 2019) used a dataset of important clinical information and elements and applied better SVM calculations to characterize cardiovascular disease cases. The SVM model's accuracy and efficacy were improved by including the duality optimization method. The study demonstrated the improved SVM with dual optimization's potential for accurately identifying heart failure cases by reporting the accuracy achieved by the proposed method. The authors' ability to accurately identify heart failure cases with their proposed method of 92% demonstrates the accuracy of their strategy in improving heart failure diagnosis and detection.

III. METHODOLOGY

A. Proposed Methodology

The methodology for the detection of heart disease is described below:

1. In the first step of the methodology for the research, the literature review has been conducted from where the methods and algorithms have been identified and chosen.
2. The data has been chosen from UCI and preprocessed using data cleaning, and feature encoding.
3. The Bayesian method has been applied to the feature to compute the probabilities of the feature to detect heart disease and the features have been finalized using probability filtering.
4. The machine learning models have been applied to the data to detect heart disease and the best-performing model will be chosen by comparing the accuracies

B. Data Selection

1.1.1 Selected Data

The data from UCI machine learning will be used to detect heart disease by gathering information from the draft literature review. The database contains heart disease patient records from Switzerland, Hungary and Cleveland with 303 records in each data (UCI, 1988). Additionally, the data contains 13 features (initial symptoms) and the target feature. This data has been used by all researchers as seen in Table 1. So, the selection of data is justified.

1.1.2 Justification of Data Choice

The data has been selected based on the review of existing research papers that have used the same data in their research. This choice will lead to performing a homogeneous comparison because similar data has been chosen. Hence, the researchers who have used the data (UCI Heart Disease data) earlier are as follows:

Atallah (2019), Iqbal et al. (2020), Khan (2020), Gaikwad et al. (2022), Yadav et al. (2020), Mahmood (2010), Esfahani (2017), Geweid (2019)

1.1.3 Feature Description

The details of the enlisted features in the UCI Heart Disease Data have been discussed below:

Table 1 Features of Heart Disease Data

Features	Usability	Type of Feature	Definition
trtbps	Predictor	Independent	Blood Pressure
thalachh	Predictor	Independent	Heart Rate
thal	Predictor	Independent	Defect types
target	Target	Dependent	Target Feature
slope	Predictor	Independent	Depression Level
sex	Predictor	Independent	Gender
restecg	Predictor	Independent	Value of Electrocardiography
oldpeak	Predictor	Independent	Depression Type
fbs	Predictor	Independent	Blood sugar
exng	Predictor	Independent	Agina value
cp	Predictor	Independent	Type of pain in the chest
chol	Predictor	Independent	Cholesterol in Blood
ca	Predictor	Independent	Major Vessels
age	Predictor	Independent	Age

C. Tool Selection

Machine learning has emerged as a significant and highly beneficial technology in the modern world. The use of programming languages to design algorithms capable of learning from and making predictions based on data has transformed how industries operate, solve problems, and make decisions. Python, a versatile and powerful programming language, has become a popular choice for machine learning development. In this project, Python has been chosen as the tool for programming and preparing the artefact.

D. Technologies to be Applied

1) Data Analysis and Visualization

Data analysis is the process of inspecting, cleaning, and transforming data to discover useful information, draw conclusions, and support decision-making. In heart disease detection, data analysis involves gathering and analyzing patient data such as age, blood pressure, and cholesterol levels.

Data visualization, on the other hand, involves representing data in a visual format such as charts and graphs. It helps in presenting data in an easily understandable format, making it easier to interpret and analyze. For example, a scatter plot can be used to show the relationship between blood pressure and age.

2) Bayesian Feature Selection

Feature selection is the process of selecting a subset of relevant features that will be used to predict heart disease. In heart disease detection, relevant features include age, blood pressure, and cholesterol levels. Machine learning algorithms can be used for feature selection, which involves identifying the most significant features that contribute to predicting heart disease. In this present research, the Bayesian feature selection process will be applied to determine the important features of the heart disease data. A brief description of the method has been given below:

a) Bayesian Feature Selection – An Overview

The Bayesian feature selection process is a probabilistic model that employs the Bayesian framework for evaluating the relevance of individual features based on the estimated joint probability of the features and the target variable. By considering the uncertainty and inherent stochasticity of the data, this method provides a more robust and accurate selection of features compared to traditional methods, such as stepwise regression and regularized linear regression (George, 2019).

Bayesian feature selection follows a three-step procedure:

- Defining the prior probability for each feature being relevant or irrelevant,
- Updating the probability based on the evidence provided by the data
- Choosing the features with the highest posterior probability.

As opposed to other methods, Bayesian feature selection naturally incorporates the uncertainties related to model parameters and data distribution, resulting in better feature selection and improved model performances (Masoudi-Nejad & Goliaei, 2012).

b) Importance of Bayesian Feature Selection

The adoption of Bayesian feature selection in various fields stems from its numerous advantages over traditional methods.

- **Robustness:** Since Bayesian feature selection inherently accounts for model uncertainty and parameter estimations, it is more robust in identifying the most informative features compared to other methods.
- **Interpretability:** By providing a probability score for each feature, Bayesian feature selection enables a better understanding of the contributions of individual features towards the target variable.
- **Scalability:** This method is highly scalable and can be easily applied to high-dimensional data while avoiding overfitting, a common issue faced by other algorithms.

c) Process of Application

Correlation and the Bayesian model represent two viable methods for feature selection. Employing correlation allows for the identification of the most pertinent features, determined by their relationship with the target variable. Conversely, the Bayesian model considers the uncertainty inherent in the data and the model itself. Each method has its own unique strengths and weaknesses, and the choice between the two often hinges on the specific problem at hand and the nature of the data involved. The Bayesian model stands out as one of the more advanced methods of feature selection, though it does come with its own set of challenges. For instance, it necessitates higher RAM for execution and feature selection from a set. If a large set of features is employed, this could potentially slow down the execution rate. Moreover, the use of the Bayesian model necessitates a GPU. To mitigate the complexity of execution involved in feature selection, a combination of correlation and the Bayesian model is often the recommended course of action. Hence, the process of applying the Bayesian feature selection method has been shown below:

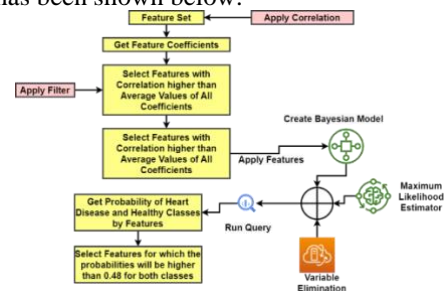


Figure 1 Process of Applying the Bayesian Method

E. Algorithm Selection

1) Logistic Regression

Logistic Regression is a machine learning algorithm used to predict the probability of an event by fitting data to a logistic function (Srivastava, et al., 2022). It is particularly useful in predicting binary outcomes, such as the presence or absence of heart disease. The algorithm utilizes input features (e.g., age, blood pressure, cholesterol levels) to generate a probability score that indicates the likelihood of a patient having heart disease. Logistic Regression provides a simple, yet effective method for predicting heart disease, allowing physicians to focus on patients with a high-risk score for early intervention (Zhao, et al., 2022).

2) K-Nearest Neighbours

The KNN algorithm is a non-parametric, instance-based method that classifies input data based on the majority of its 'k' closest neighbours' output labels in the feature space (Rahman, 2019). The algorithm calculates the distance between the input data point and its nearest neighbours using methods such as Euclidean, Manhattan, or Hamming distance (Khan, 2020). KNN's simplicity and effectiveness in heart disease prediction have prompted researchers to explore its usefulness in detection and diagnosis.

3) Support Vector Machine

SVM is a supervised machine learning algorithm that works by finding the best hyperplane that separates data points into different classes, maximizing the margin between classes (Rahman, 2019). It is robust in handling high-dimensional data and capable of modelling complex relationships between input features. SVM has been successfully applied to predict and detect heart disease using various types of data, such as electrocardiogram signals and clinical records, exhibiting high accuracy and promising results (Arslan, 2017).

4) Naive Bayes

Naive Bayes is a probabilistic, generative classification algorithm based on Bayes' theorem. It assumes that input features are conditionally independent, given the output label (Patra, 2019). Although this assumption is considered naive, the algorithm performs well in many real-world applications, including heart disease detection. Naive Bayes can provide probabilistic predictions that enable physicians to consider multiple factors when assessing a patient's risk of heart disease (Srivastava, et al., 2022).

5) Decision Tree

A Decision Tree is a recursive, flowchart-like structure that partitions the input data into subsets based on the most significant feature at each node (Tanuku, et al., 2022). It is an interpretable model that produces a set of decision rules that can be followed to predict the output class. Decision Tree models have been employed in heart disease detection with notable success, offering interpretable results that facilitate clinical decision-making (Sahoo, et al., 2022).

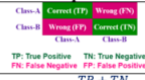
6) Random Forest

Random Forest is an ensemble method that combines the predictions of multiple Decision Tree classifiers, increasing the overall accuracy and reducing overfitting (Al-Mashagbeh & Ababneh, 2021). It has shown great potential in detecting heart diseases due to its high predictive power and ability to handle high-dimensional datasets with a large number of features.

F. Performance Evaluation Method

The below-discussed performance evaluation metrics will be used to compare the effectiveness of the employed models and to judge the performance outcomes:

Table 2 Performance Evaluation Metrics and Representations

Metric	Outcome	Equation
Confusion Matrix	It will show the result and outcomes of the classification	 <p>TP: True Positive TN: True Negative FP: False Positive FN: False Negative</p>
Model Accuracy	Accuracy of the model employed for the classification of heart diseases	$AC = \frac{TP + TN}{All\ Test\ Data}$
Model Precision	The preciseness of the model employed for the classification of heart diseases	$PR = \frac{TP}{TP + FP}$
Model Recall	The recall of the model employed for the classification of heart diseases	$RC = \frac{TP}{TP + FN}$
Model F1-Score	The f1-score of the model employed for the classification of heart diseases	$F1 = \frac{2 \times PR \times RC}{PR + RC}$

IV. ANALYSIS AND RESULT

The analytical outcomes of the data will be presented and interpreted in this chapter. Additionally, the detection of heart disease using the selected models will be presented and the results will be discussed accordingly.

A. Reading Heart Disease Data

The datasets for heart disease for three different locations contain the same number of instances which have been presented below:

```
Instances in cleveland is 303
Instances in hungarian is 303
Instances in switzerland is 303
Total Data Instances: 909
```

Figure 2 Instances in Datasets

These three datasets have been merged and one single data has been prepared which has been shown below:

	age	sex	cp	trestbps	chol	fbis	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0

Figure 3 Merged Heart Disease Dataset

B. Data and Feature Information

The feature information of the data has been investigated and it has been observed that the data contains 11 decimal features, 2 object features and one integer feature

C. Cleaning Data Noise

The noise or missing values have been checked initially in the data and it has been observed that there are no missing values available as shown below:

```
Missing Values
age      0
sex      0
cp       0
trestbps 0
chol     0
fbis     0
restecg  0
thalach  0
exang    0
oldpeak  0
slope    0
ca       0
thal     0
target   0
```

Figure 4 Initial Detection of Missing Values

To cross-check the result, the data has been investigated manually. In that investigation, it has been observed that some of the data points contain the special character "?" which is signifying the missing values for UCI data. As that special character cannot be detected directly, a manual process has been applied to remove those. As a result, those have been removed and the data has been cleaned.


```

Int64Index: 909 entries, 0 to 302
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype
---  -
 0   age         909 non-null    float64
 1   sex         909 non-null    float64
 2   cp          909 non-null    float64
 3   trestbps    909 non-null    float64
 4   chol        909 non-null    float64
 5   fbs         909 non-null    float64
 6   restecg     909 non-null    float64
 7   thalach     909 non-null    float64
 8   exang       909 non-null    float64
 9   oldpeak     909 non-null    float64
10   slope       909 non-null    float64
11   ca          909 non-null    float64
12   thal        909 non-null    float64
13   target      909 non-null    int64
dtypes: float64(13), int64(1)

```

Figure 5 Feature Information after Data Cleaning

D. Data Target Feature Preparation

The dataset contains five classes in the target feature 0-4 and those are in integer format as seen in Figure 9. So, those classes behave like continuous data rather than nominal. For classification purposes, the class should be nominal or categorical. So, the classes have been transformed to categorical as shown below:

```

0 492 -> No_Disease 492
1 165 -> HD_Severity_2 165
2 108 -> HD_Severity_1 108
3 105 -> HD_Severity_3 105
4 39 -> HD_Severity_4 39

```

Figure 6 Target Class Transformation

After transforming the target feature, the data type has been converted from integer to object or categorical as shown below:

```

#   Column      Non-Null Count  Dtype
---  -
0   age         909 non-null    float64
1   sex         909 non-null    float64
2   cp          909 non-null    float64
3   trestbps    909 non-null    float64
4   chol        909 non-null    float64
5   fbs         909 non-null    float64
6   restecg     909 non-null    float64
7   thalach     909 non-null    float64
8   exang       909 non-null    float64
9   oldpeak     909 non-null    float64
10  slope       909 non-null    float64
11  ca          909 non-null    float64
12  thal        909 non-null    float64
13  target      909 non-null    object

```

Figure 7 Feature Information after Class Transformation

E. Data Splitting

1) Data Balancing

In Figure 10, it has been seen that the data is imbalanced. To balance the data, the data classes have been equalised by upscaling and the final class distribution has been shown below:

```

No_Disease      492      HD_Severity_2      2460
HD_Severity_2   165      HD_Severity_1      2460
HD_Severity_1   108      HD_Severity_3      2460
HD_Severity_3   105      HD_Severity_4      2460
HD_Severity_4    39      No_Disease         492

```

Figure 8 Data Balancing

2) Splitting

The balanced data has been split into three segments namely training data (containing 60% of all instances), validation data (containing 20% of all instances) and test data (containing 20% of all instances). The class distribution of all three segments has been shown below:

```

Main Data
├── Train Data
│   ├── Class Distribution in Train Data
│   ├── HD_Severity_4 1511
│   ├── HD_Severity_3 407
│   ├── HD_Severity_2 405
│   ├── HD_Severity_1 405
│   ├── HD_Severity_4 405
│   ├── HD_Severity_3 405
│   ├── HD_Severity_2 405
│   └── No_Disease 285
├── Validation Data
│   ├── Class Distribution in Validation Data
│   ├── HD_Severity_4 407
│   ├── HD_Severity_3 405
│   ├── HD_Severity_2 405
│   ├── HD_Severity_1 405
│   ├── HD_Severity_4 405
│   ├── HD_Severity_3 405
│   ├── HD_Severity_2 405
│   └── No_Disease 94
└── Test Data
    ├── Class Distribution in Test Data
    ├── HD_Severity_4 515
    ├── HD_Severity_3 510
    ├── HD_Severity_2 473
    ├── HD_Severity_1 466
    ├── HD_Severity_4 466
    ├── HD_Severity_3 466
    ├── HD_Severity_2 466
    └── No_Disease 111

```

Figure 9 Class Distributions in Data Segments

F. Data Preprocessing

1) Outlier Treatment

1.1.3.1 Outlier Detection

The outliers have been detected in all three segments of the data and the boxplots have been visualized for each to observe those outliers. To detect outliers, the quantile method has been used with the flow of operations presented in Figure 3 (section 4.1). The boxplots below are showing that the data contains outliers:

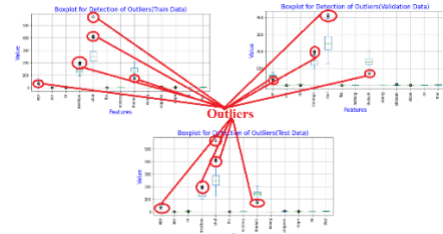


Figure 10 Outlier Detection

2) Data Normalization

To remove the outliers, data normalization has been done. In this context, MinMaxScaler has been used which scales the data points within the 0-1 range eliminating the possibility of having outliers. The normalized training data has been shown below:

```

age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
0.291867 1.0 1.000000 0.245283 0.116438 0.0 1.0 0.374048 1.0 0.403226 0.5 0.000000 1.000000 HD_Severity_3
0.087500 0.0 1.000000 0.432962 0.080750 0.0 1.0 0.564886 0.0 1.000000 1.0 1.000000 1.000000 HD_Severity_3
0.541667 0.0 0.333333 0.380792 0.283105 0.0 1.0 0.687023 0.0 0.225906 0.5 0.000000 0.428571 No_Disease
0.541667 1.0 1.000000 0.433962 0.207763 0.0 0.0 0.305344 1.0 0.903226 1.0 0.000000 1.000000 HD_Severity_3
0.562500 1.0 1.000000 0.336623 0.358447 1.0 1.0 0.244275 1.0 0.258065 1.0 0.000000 1.000000 HD_Severity_1
...
0.502500 1.0 1.000000 0.350491 0.130420 0.0 1.0 0.259542 1.0 0.338718 0.5 0.333333 0.857143 HD_Severity_2
0.645033 0.0 1.000000 0.528302 0.301370 0.0 0.0 0.656489 0.0 0.419355 0.5 0.000000 1.000000 HD_Severity_2
0.250000 1.0 1.000000 0.150943 0.105023 0.0 0.0 0.664122 0.0 0.000000 0.0 0.000000 1.000000 HD_Severity_2
0.770833 0.0 1.000000 0.762453 0.232877 1.0 0.0 0.717557 1.0 0.181290 0.5 0.000000 1.000000 HD_Severity_2
0.520833 1.0 1.000000 0.264151 0.365297 0.0 1.0 0.343511 1.0 0.516129 0.5 0.000000 0.428571 HD_Severity_3

```

Figure 11 Normalized Training Data

The normalized validation data has been shown below:

```

age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
0.520833 1.0 1.0 0.264151 0.365297 0.0 1.0 0.343511 1.0 0.516129 0.5 0.000000 0.428571 HD_Severity_3
0.503333 1.0 1.0 0.547170 0.337900 0.0 0.0 0.129771 1.0 0.103548 0.5 0.333333 1.000000 HD_Severity_2
0.770833 1.0 1.0 0.306226 0.431507 0.0 0.0 0.412214 1.0 0.200323 0.5 0.000000 0.857143 HD_Severity_1
0.729167 1.0 0.0 0.150943 0.194064 0.0 1.0 0.557252 1.0 0.200323 0.5 0.000000 0.428571 No_Disease
0.000000 1.0 1.0 0.415004 0.091324 0.0 1.0 0.412214 1.0 0.080645 0.5 0.333333 0.428571 HD_Severity_4
...
0.503333 1.0 1.0 0.264151 0.212026 0.0 0.0 0.528716 0.0 0.103548 0.5 0.000000 1.000000 HD_Severity_3
0.087500 0.0 1.0 0.433962 0.324201 0.0 0.0 0.673089 0.0 0.000045 1.0 0.000000 0.428571 HD_Severity_3
0.750000 0.0 1.0 0.528302 0.226027 0.0 1.0 0.328244 0.0 0.181290 0.5 0.000000 1.000000 HD_Severity_4
0.604167 1.0 1.0 0.056604 0.246575 0.0 0.0 0.648855 0.0 0.016129 0.0 0.333333 1.000000 HD_Severity_1
0.708333 0.0 1.0 0.528302 0.641553 0.0 1.0 0.633088 0.0 0.645161 0.5 1.000000 1.000000 HD_Severity_4

```

Figure 12 Normalized Validation Data

The normalized test data has been shown below:

```

age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1.000000 1.0 1.0 0.292453 0.399538 0.0 1.0 0.694056 1.0 0.000000 0.0 1.000000 0.428571 HD_Severity_4
0.503333 1.0 1.0 0.150943 0.471132 0.0 0.0 0.549618 1.0 0.483871 0.5 0.333333 1.000000 HD_Severity_1
0.645033 1.0 1.0 0.336623 0.175210 0.0 0.0 0.488569 1.0 0.387087 0.0 0.000000 1.000000 HD_Severity_4
0.645033 1.0 1.0 0.433962 0.374134 0.0 1.0 0.757525 0.0 0.103548 0.5 0.000000 1.000000 HD_Severity_1
0.503333 0.0 1.0 0.245283 0.515012 0.0 1.0 0.702290 1.0 0.006774 0.0 0.000000 0.428571 No_Disease
...
0.541667 1.0 1.0 0.358491 0.512702 0.0 0.0 0.488569 1.0 0.103548 0.5 0.333333 1.000000 HD_Severity_3
0.503333 1.0 1.0 0.150943 0.471132 0.0 0.0 0.549618 1.0 0.483871 0.5 0.333333 1.000000 HD_Severity_1
0.208333 1.0 1.0 0.264151 0.203233 0.0 0.0 0.528716 0.0 0.103548 0.5 0.000000 1.000000 HD_Severity_3
1.000000 1.0 1.0 0.292453 0.399538 0.0 1.0 0.694056 1.0 0.000000 0.0 1.000000 0.428571 HD_Severity_4
1.000000 1.0 1.0 0.292453 0.399538 0.0 1.0 0.694056 1.0 0.000000 0.0 1.000000 0.428571 HD_Severity_4

```

Figure 13 Normalized Test Data

3) Outlier Elimination

The outliers have been eliminated by applying feature normalisation which has been reflected in the below-presented figure:

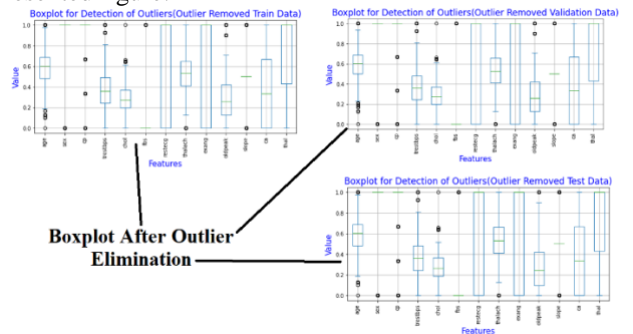


Figure 14 Outlier Elimination

4) Feature Selection using the Bayesian Model

a) Feature Correlation

Primarily, the feature correlation has been applied to check the feature relationship with the target feature. The correlation coefficients of the features have been shown below:

Features	Correlation
sex	0.242698
cp	0.368103
exang	0.370535
oldpeak	0.428425
slope	0.333847
ca	0.462930
thal	0.479050

Figure 15 Feature Correlation

b) Creating Feature and Target Combination

Those features have been combined with the target feature by creating a tuple of features. This process is required for the Bayesian model to determine the class probabilities concerning each of the features available or taken.

Feature and Target Combination for Graph:

```
('sex', 'target')
('cp', 'target')
('exang', 'target')
('oldpeak', 'target')
('slope', 'target')
('ca', 'target')
('thal', 'target')
```

Figure 16 Feature Combination

c) Applying Bayesian Model

Now, by applying the features in the Bayesian model, the class probabilities have been obtained. This defines the effectiveness of the features to detect the classes. The almost equal probability will ensure that the classes can be detected precisely resulting in higher accuracy in heart disease detection using those features.

1. Heart Disease Probability for: sex	2. Heart Disease Probability for: cp	3. Heart Disease Probability for: exang
target phi(target)	target phi(target)	target phi(target)
target(HD_Severity_1) 0.1842	target(HD_Severity_1) 0.1853	target(HD_Severity_1) 0.1857
target(HD_Severity_2) 0.2004	target(HD_Severity_2) 0.2072	target(HD_Severity_2) 0.2007
target(HD_Severity_3) 0.1823	target(HD_Severity_3) 0.1812	target(HD_Severity_3) 0.1947
target(HD_Severity_4) 0.1745	target(HD_Severity_4) 0.1707	target(HD_Severity_4) 0.1827
target(No_Disease) 0.2549	target(No_Disease) 0.2556	target(No_Disease) 0.2362
4. Heart Disease Probability for: oldpeak	5. Heart Disease Probability for: slope	6. Heart Disease Probability for: ca
target phi(target)	target phi(target)	target phi(target)
target(HD_Severity_1) 0.2000	target(HD_Severity_1) 0.1081	target(HD_Severity_1) 0.1945
target(HD_Severity_2) 0.2000	target(HD_Severity_2) 0.1078	target(HD_Severity_2) 0.1952
target(HD_Severity_3) 0.2001	target(HD_Severity_3) 0.1073	target(HD_Severity_3) 0.2002
target(HD_Severity_4) 0.2000	target(HD_Severity_4) 0.1074	target(HD_Severity_4) 0.2046
target(No_Disease) 0.2000	target(No_Disease) 0.2095	target(No_Disease) 0.2055

Figure 17 Class Probabilities using the Bayesian model

The Bayesian network for the features with the target feature has been shown below:



Figure 18 Bayesian Network Graph

d) Final Feature Selection by Probability Filtering

Finally, the features have been selected concerning the distribution of the equal or almost equal probabilities of the features to detect the data classes. To select the final feature, the filter has been applied to the features and those features have been selected which have the probability of detecting all classes are almost equal. The selected features and the probabilities have been listed below:

Table 3 Final Feature Selection

Features	Healthy(Probability)	HD_1(Probability)	HD_2(Probability)	HD_3(Probability)	HD_4(Probability)
sex	0.1842	0.2041	0.1823	0.1745	0.2549
cp	0.1853	0.2072	0.1812	0.1707	0.2556
oldpeak	0.2000	0.2000	0.2001	0.2000	0.2000
slope	0.1981	0.1978	0.1973	0.1974	0.2095
ca	0.1945	0.1952	0.2002	0.2046	0.2055
thal	0.1942	0.2174	0.1929	0.1824	0.2131

G. Heart Disease Detection Result

The result of the heart disease detection will be presented and interpreted in this section. Primarily, the training and validation data will be applied to the models to check the

effectiveness of the training and validation process. As the test data is totally separate from the training and validation, the final testing has been done using the test data to observe the effectiveness of the models on the test data. This will provide the final result of the detection of heart disease.

1) Training and Validation

The confusion matrices have been produced for each model while training and validating those. The confusion matrix will imply the detection status by classes enlisted in the data. The confusion matrices of the models have been presented below:

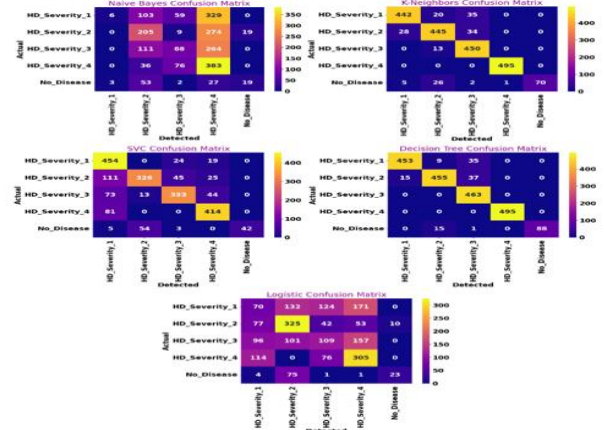


Figure 19 Confusion Matrices of Applied Models for Training and Validation

From the confusion matrices, it can be seen that the Decision tree, Support Vector Classifier and the K-neighbours model have performed well. To check the overall result through the classification metrics (section 4.6, Table 4), the metrics have been computed after applying the data to each of the models and are listed below:

Table 4 Training and Validation Result

Classifiers	Train Accuracy	Validation Accuracy	Precision	Recall	F1-Score	Overfit
Decision Tree	95.68	95.55	96	96	96	0.13
K-Neighbours	94.97	94.72	95	95	95	0.25
SVC	82.96	82.67	84	83	82	0.29
Logistic	37.89	36.45	33	36	34	1.44
Naive Bayes	40.47	38.63	54	39	33	1.84

2) Test Result

Finally, the test data has been applied to all five models to check the test status for detecting heart disease. The confusion matrices for testing have been shown below:

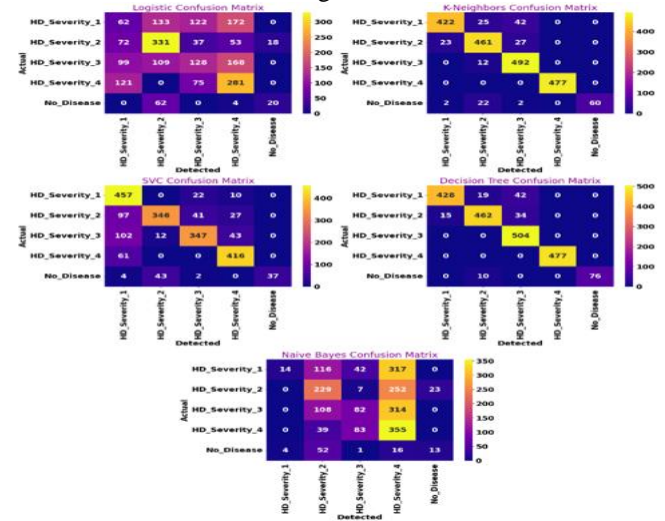


Figure 20 Confusion Matrices of Applied Models for Testing

From the test confusion matrices, it can be seen that the Decision tree, Support Vector Classifier and the K-

neighbours model have performed well same as the training results. The overall test result metrics are shown below:

Table 5 Test Result

Classifiers	Train_Accuracy	Test_Accuracy	Precision	Recall	F1-Score	Overfit
Decision Tree	95.68	95.21	95	95	95	0.47
K-Neighbours	94.97	94.24	94	94	94	0.73
SVC	82.96	83.16	84	83	83	0.2
Naive Bayes	40.47	39.72	57	40	35	0.75
Logistic	37.89	38.51	35	39	36	0.62

V. EVALUATION, DISCUSSION AND CONCLUSION

A. Performance Measure and Comparison

1) Performance Graphs

The results of heart disease detection have been presented through the classification metrics in Table 6 and Table 7. The metrics will be compared and the comparison graphs will be presented in this section. This will help decide the most effective model through which the detection of heart disease can be done with the highest accuracy. The performance graphs are presented below:

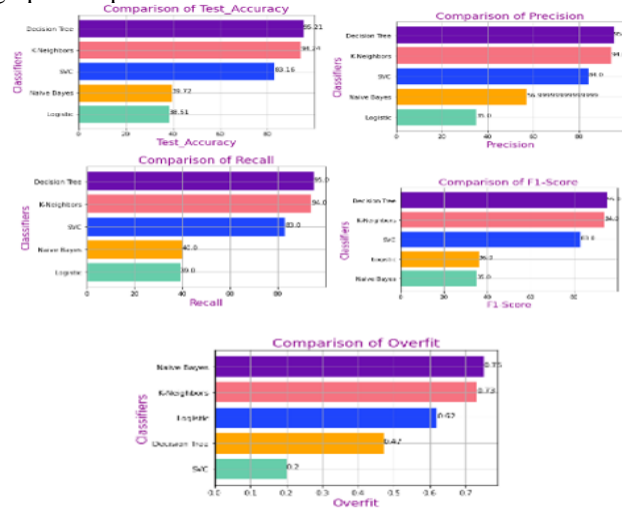


Figure 21 Performance Graphs

2) Selection of the Most Effective Model

From the overall comparisons of the test results (through classification metrics), it can be said that the Decision Tree is the most effective model that has detected heart disease with 95.21% accuracy. The test confusion matrix and classification report have been shown below:

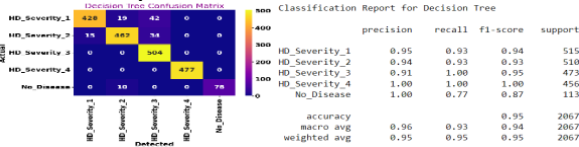


Figure 22 Test Result of Most-Effective Model

3) Research Improvement

The accuracy of the Decision Tree for heart disease detection has been compared with the existing research where the same dataset has been used. The comparison is shown below:

Table 6 Comparison of Present Research with Existing

Author	Detection	Data Source	Algorithm	Feature Selection	Accuracy
Atallah (2019)	Heart Disease	UCI	Majority voting approach	Correlation	94
Khan (2020)	Heart Disease	UCI	Support Vector Machines	Correlation	86
Yadav et al. (2020)	Heart Disease	UCI	Naive Bayes	Correlation	82.19
Esfahani (2017)	Heart Disease	UCI	Hybrid model	Not Applied	87
Geweid (2019)	Heart Disease	UCI	Support Vector Machines	Not Applied	92
Sahoo et al. (2022)	Heart Disease	UCI	Random Forest	Recursive Method	90.16
Present Research	Heart Disease	UCI	Decision Tree	Bayesian Feature Selection	95.21

From the above-presented comparison, it is clear that the present research with the application of the Decision tree has gained the highest accuracy for detecting heart disease. It has outperformed the existing models with a significant difference in the accuracy of detection. So, it can be stated that the present approach for detecting heart disease has a significant improvement. The comparison has been visualised below:

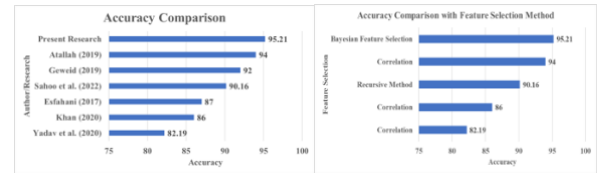


Figure 23 Accuracy Comparison with Existing Research

B. Discussion

1) Addressing Research Questions

a) RQ-1

The necessary symptoms that have a higher influence on the detection of heart disease have been detected with the application of the Bayesian model. In this context, the Bayesian model has been applied to determine the probability of detecting the data classes. With this application, the below-mentioned symptoms have been chosen:

Sex, cp, old peak, slope, ca, thal

Based on the selected features, the detection of heart disease has been performed. Finally, the detection of heart disease has been done with the highest accuracy 95.21% using the Decision Tree. This accuracy has been seen to be the highest compared to the existing research as well.

b) RQ-2

The selected dataset for heart disease has been prepared with the application of data noise removal (data cleaning), outlier detection & elimination along with the selection of features using the Bayesian model. While cleaning the data, a manual inspection has been performed to remove the special character "?". Next, the outliers have been detected using the quantile method and eliminated using data normalization. Finally, the features have been selected using the Bayesian method. Those features have been used to prepare the final data using which the detection of heart disease has been performed with 95.21% accuracy using the Decision Tree model.

c) RQ-3

The most effective model for the detection of heart disease has been achieved for the Decision Tree model. The model has acquired 95.21% accuracy and this performance has been compared with the previous approaches. The comparison has been presented in section 6.3. Concerning the comparison of heart disease detection, it can be stated that the present approach using the Decision Tree for detecting heart disease has a significant improvement over the previous approaches.

2) Challenges Faced and Resolution Taken

The challenges have been faced during the execution of the research to detect heart disease. The challenges faced and the respective resolutions which have been taken have been presented below:

1. One of the initial challenges encountered was determining the missing values in the data. Initially, the inspection of missing values did not provide any indication of their status. To address this issue, a manual inspection was conducted, which revealed that the data contained a "?" character instead of "NAN" to represent missing values. In order to remove these missing values, all instances of "?" were replaced with zero.
2. During the application of the Bayesian model, it became apparent that the execution time was significantly slower and the system eventually crashed. This was due to supplying all the available features of the data. To overcome this issue, the correlation method was implemented to select the initial features and reduce their number. By applying the Bayesian model to the reduced set of features, the final feature selection was successfully performed.

3) Research Gaps

The research gaps have been discussed below:

1. The existing research review suggests that the feature selection process has been explored using various methods such as correlation and Bayesian models separately. However, there is a research gap in terms of combining correlation and Bayesian models for feature selection, as this approach has not been previously applied in the reviewed studies. Therefore, this study aims to address this research gap by utilizing the combination of correlation and Bayesian models for feature selection.
2. The previous research for the detection of heart disease has applied two classes heart disease and healthy person. It means the previous research using the UCI heart disease data (which has been used in the present research also) has been done with the binary class. The new implication has been done in the present research by applying five classes.

C. Conclusion

Symptom-based disease detection plays a crucial role in the early identification and diagnosis of various medical conditions. Traditionally, healthcare professionals heavily rely on their expertise and clinical judgment to interpret symptoms and make accurate diagnoses. However, with the advancements in technology, particularly in the fields of Bayesian probability and machine learning, there is a promising opportunity to enhance disease detection by leveraging these techniques. The present research explored the potential of applying Bayesian probability and machine learning algorithms to improve symptom-based disease detection, ultimately leading to more accurate and precise identification of heart disease.

In light of existing research, the aim of this study is to effectively classify and detect heart disease. The datasets used were collected from Kaggle, while the inspiration was drawn from previous research papers. The algorithms employed were carefully selected based on the knowledge

gathered from previous researchers who had used them for similar purposes. To ensure accurate results, the data underwent various preprocessing techniques, including data cleaning, outlier detection and elimination (using the quantile method), and feature selection (using the Bayesian model). The selected model successfully detected heart disease, with the voting model achieving the highest accuracy rate of 95.21% and minimal model overfitting at 0.47%. This research has demonstrated significant improvements in heart disease detection accuracy compared to existing approaches.

1) Limitations

The current research has certain limitations that are worth discussing.

- The data used in this study is derived from a single source, although the idea behind the data collection was inspired by previous research.
- Deep learning models have not been utilized for the detection of heart disease in this research.
- The quantile method, which has been employed solely for detecting outliers, may have limitations in accurately identifying heart disease.
- The training, validation, and test data have been prepared using a fixed ratio of 60:20:20 for which the justification has not been applied.
- The tuning of the hyperparameters of the algorithms has not been performed, which could potentially impact the accuracy and performance of the detection models.
- The detection of heart disease has been carried out using only a limited number of features obtained through a combination of correlation and the Bayesian model.

2) Future Recommendations

In order to further enhance the research in the future, there are several techniques that can be employed:

- The implementation of a Neural Network model using deep learning technology can greatly improve the detection of heart disease.
- Models from Transfer learning can also be applied to enhance the accuracy of heart disease detection.
- Utilizing various feature selection methods can help identify and prioritize important features that can aid in the detection of heart disease.
- An application interface can be designed where the symptoms can be taken from users and the detection results can be shown.

REFERENCES

- [1]. Adorada, A. et al., 2021. Support vector machine-recursive feature elimination (svm-rfe) for selection of microRNA expression features of breast cancer.. In *2018 2nd international conference on informatics and computational sciences (ICICoS) IEEE.*, pp. (pp. 1-4).
- [2]. Aggarwal, R. & Kumar, S., 2022. MLPPCA: Heart Disease Detection using Machine learning. *Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pp. 1-5.
- [3]. Ali, L., Khan, S. U. & Anwar, M., 2019. Early Detection of Heart Failure by Reducing the Time Complexity of the Machine Learning based Predictive Model. *International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pp. 1-5.
- [4]. Aljaaf, A. J. & Jumeily, D. A.-., 2015. Predicting the likelihood of heart failure with a multi level risk assessment using decision tree. *Third International Conference on Technological Advances*

- in *Electrical, Electronics and Computer Engineering (TAECE)*, pp. 101-106.
- [5]. Al-Mashagbeh, M. H. & Ababneh, M., 2021. Tor Detection using a Machine Learning Approach Using Correlation based Feature Selection with Best First and Random Forest.. In *2021 International Conference on Information Technology (ICIT) IEEE*, pp. (pp. 893-898)..
 - [6]. Andreu-Perez, et al., 2015. Big data for health. *IEEE Journal of Biomedical and Health Informatics*, 19(4), pp. 1193-1208.
 - [7]. Atallah, R. & A.-M. A., 2019. Heart disease detection using machine learning majority voting ensemble method.. In *2019 2nd international conference on new trends in computing sciences (ictcs) IEEE*, pp. (pp. 1-6)..
 - [8]. Bakhshi, M., Mirtaheri, S. L. & Greco, S., 2022. Heart Disease Prediction Using Hybrid Machine Learning Model Based on Decision Tree and Neural Network.. In *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCM) IEEE*, pp. (pp. 36-41)..
 - [9]. Elhaj, F. A. et al., 2017. Hybrid classification of Bayesian and extreme learning machine for heartbeat classification of arrhythmia detection.. In *2017 6th ICT International Student Project Conference (ICT-ISPC) IEEE*, pp. 1-4.
 - [10]. Esfahani, H. A. & G. M., 2017. Cardiovascular disease detection using a new ensemble classifier.. In *2017 IEEE 4th international conference on knowledge-based engineering and innovation (KBEI) . IEEE*, pp. (pp. 1011-1014).
 - [11]. Gaikwad, M. J., Asole, P. S. & Bitla, L. S., 2022. Effective Study of Machine Learning Algorithms for Heart Disease Prediction.. In *2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) IEEE*, pp. (pp. 1-6)..
 - [12]. George, 2019. An Introduction to Bayesian Feature Selection. Introduction to Bayesian Econometrics. *Cambridge University Press*, pp. 1-10.
 - [13]. Geweid, G. G. & A. M. A., 2019. A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique.. *IEEE Access*, pp. 149595-149611.
 - [14]. Gulati, S., Guleria, K. & Goyal, N., 2022. Classification and Detection of Coronary Heart Disease using Machine Learning. *2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 1-5.
 - [15]. Gupta, Wang & Ganesh, 2020. An overview on deep learning-based methods for ECG-related human health anomaly detections: Scope and challenges.. *International Journal of Information Technology*, Volume 17, pp. 27-40.
 - [16]. Han, S., Eom, H., Kim, J. & Park, C., 2020. Optimal DNN architecture search using Bayesian Optimization Hyperband for arrhythmia detection.. In *2020 IEEE Wireless Power Transfer Conference (WPTC) IEEE*, pp. (pp. 357-360)..
 - [17]. Hua, Q. et al., 2021. An interpretable model for ECG data based on Bayesian neural networks.. *IEEE Access*, pp. 57001-57009.
 - [18]. Iqbal, J., Iqbal, M. M., Khadam, U. & Nawaz, A., 2020. Ordinary Learning Method for Heart Disease Detection using Clinical Data.. In *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) IEEE*, pp. (pp. 1-6)..
 - [19]. Khan, M. I. H. & M. M. R. H., 2020. Effectiveness of Data Driven Diagnosis of Heart Disease.. In *2020 11th International Conference on Electrical and Computer Engineering (ICECE) IEEE*, pp. (pp. 419-422)..
 - [20]. Mahmood, A. M. & K. M. R., 2010. Early detection of clinical parameters in heart disease by improved decision tree algorithm.. In *2010 Second Vaagdevi International Conference on Information Technology for Real World Problems*, pp. 24-29.
 - [21]. Manikandan, S., 2017. Heart attack prediction system. *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 817-820.
 - [22]. Masoudi-Nejad & Goliaei, A. &, 2012. A pairwise nested logit model for gene and gene feature selection. *Computer Methods and Programs in Biomedicine*, 107(3), pp. 479-491.
 - [23]. Mohan, N., Jain, V. & Agrawal, G., 2021. Heart Disease Prediction Using Supervised Machine Learning Algorithms. *5th International Conference on Information Systems and Computer Networks (ISCON)*, pp. 1-3.
 - [24]. Mortazavi, et al., 2016. Analysis of machine learning techniques for heart failure readmissions.. *Circulation: Cardiovascular Quality and Outcomes*, 9(6), pp. 629-640.
 - [25]. Panigrahi, S. S. & K. N., 2022. Hybrid Classification Method for the Heart Disease Prediction.. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) IEEE*, pp. 494-499.
 - [26]. Patra, R. & K. B., 2019. Predictive analysis of rapid spread of heart disease with data mining.. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) IEEE*, pp. 1-4.
 - [27]. Razak, A. et al., 2022. Online feature Selection using Pearson Correlation Technique.. In *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE) IEEE*, pp. (Vol. 7, pp. 172-177)..
 - [28]. Sahoo, G. K., Kanike, K., Das, S. K. & Singh, P., 2022. Machine Learning-Based Heart Disease Prediction: A Study for Home Personalized Care.. In *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP) . IEEE*, pp. (pp. 01-06).
 - [29]. Shishah, W., 2022. An Efficient Early Stage Heart Disease Risk Detection Using Machine Learning Techniques.. In *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)IEEE*, pp. 1-6.
 - [30]. Srinivas, K. & Rao, G. R., 2015. Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. *5th International Conference on Computer Science & Educatio*, pp. 1344-1349.
 - [31]. UCI, 1988. *Heart Disease Data Set*. [Online] Available at: <https://archive.ics.uci.edu/ml/datasets/heart+disease> [Accessed 2022].
 - [32]. Yadav, S. S., Jadhav, S. M. & Nagrale, S. & P., 2020. Application of Machine Learning for the Detection of Heart Disease. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) IEEE*, pp. 165-172.
 - [33]. Zhang, D. et al., 2018. Meteorological Feature Selection Method Based on Information Value and Maximum Correlation.. In *2018 Chinese Automation Congress (CAC) . IEEE*, pp. (pp. 3159-3164).
 - [34]. Zhao, L., Deng, F., Zhang, X. & Yu, N., 2022. RFE Based Feature Selection Improves Performance of Classifying Multiple-causes Deaths in Colorectal Cancer.. In *2022 7th International Conference on Intelligent Informatics and Biomedical Science (ICIIBMS) IEEE*, pp. (Vol. 7, pp. 188-194)..
 - [35]. Zhou, Z. & Wang, Y. & L. M., 2020. Feature selection method based on hybrid SA-GA and random forests.. In *2020 International Conference on Computing and Data Science (CDS) . IEEE*, pp. (pp. 139-142).

MSc Project - Reflective Essay

Project Title:	Symptom-Based Disease Detection by Applying Bayesian Probability and Machine Learning
Student Name:	Kuber Dhami
Student Number:	220996071
Supervisor Name:	Dr. Paulo Rauber
Programme of Study:	FT Msc. Computer Science

CHALLENGES AND POTENTIAL ISSUES

Despite the promising potential of ML in heart disease detection, certain challenges need to be addressed to optimize the technology's impact. One major challenge is the quality and reliability of patient data. For an ML-based diagnostic tool to be accurate and effective, large amounts of high-quality and standardized data are required (Mortazavi, et al., 2016). There could be potential errors or inconsistencies in the data collected, ranging from measurement inaccuracies to missing and incomplete information, posing difficulties when developing and validating ML models.

Another challenge is the generalizability of ML models, which might not work effectively across various geographic and demographic populations (Gupta, et al., 2020). The development of ML algorithms using biased datasets might lead to skewed results and inaccurate predictions when applied to diverse populations, leading to potential misdiagnosis and inappropriate medical interventions.

1. One of the initial challenges encountered was determining the missing values in the data. Initially, the inspection of missing values did not provide any indication of their status. To address this issue, a manual inspection was conducted, which revealed that the data contained a "?" character instead of "NAN" to represent missing values. In order to remove these missing values, all instances of "?" were replaced with zero.
2. During the application of the Bayesian model, it became apparent that the execution time was significantly slower and the system eventually crashed. This was due to supplying all the available features of the data. To overcome this issue, the correlation method was implemented to select the initial features and reduce their number. By applying the Bayesian model to the reduced set of features, the final feature selection was successfully performed.

PROJECT MANAGEMENT AND ISSUES

RESEARCH ISSUES

Heart disease is one of the leading causes of death worldwide, with an estimated 17.9 million deaths per year. In recent years, machine learning algorithms have been increasingly used to detect and diagnose heart disease in patients. While machine learning algorithms hold a lot of promise for the detection and diagnosis of heart disease, they also raise significant social, ethical, professional, legal, security and medical issues that need to be addressed (Boddington, 2009). To ensure the successful adoption and implementation of these algorithms, it is essential to develop safe, secure, and ethical data processing and

storage systems, establish clear standards and guidelines, and provide appropriate training and education to health professionals. These measures will help to ensure that machine learning-based heart disease detection improves healthcare outcomes while preserving patient privacy, security, and trust(Wilmot, 2001). While this technology holds a lot of potential to improve healthcare outcomes, it also raises social, ethical, professional, legal, security, and medical issues that need to be addressed.

Social Issues

One of the social issues that machine learning-based heart disease detection raises is the potential for discrimination. If the algorithm is trained on data from a biased sample, it may produce biased results that discriminate against certain groups of patients(Follath, 2009).

For instance, if the algorithm is trained on data predominantly from men, its accuracy in detecting heart disease in women may be lower, leading to misdiagnosis and undertreatment. To address this issue, researchers and practitioners should strive to use a diverse and representative sample of patients when training and validating machine learning algorithms.

Ethical Issues

Another ethical issue that machine learning-based heart disease detection raises is the issue of patient privacy. Patient data is highly sensitive and confidential, and the use of machine learning algorithms to process it raises concerns about data security and breaches(Ali, et al., 2021). To mitigate this risk, it is important to develop secure and robust data storage and processing systems that protect patient privacy and ensure data security. Also, strict ethical guidelines should be established for the collection, use, and storage of patient data.

Professional Issues

There are also professional issues that machine learning-based heart disease detection raises, particularly around the skills and competencies required to design, develop, and deploy these algorithms. Health professionals need to be trained not only in the technical aspects of machine learning but also in the ethical, legal, and social implications of this technology(Rhahla, et al., 2021). Also, there is a need to develop clear and consistent standards in the design, development, and validation of machine learning algorithms for heart disease detection.

Legal Issues

Legal issues also arise in the context of machine learning-based heart disease detection. Health professionals and institutions that use these algorithms must ensure that they comply with relevant laws and regulations governing the use and protection of patient data. Also, they must be transparent about the algorithms' limitations and potential biases when communicating with patients and other stakeholders (Ross, et al., 2023).

Security Issues

Finally, security and medical issues must not be overlooked when using machine learning for heart disease detection. Machine learning algorithms used in healthcare must be secure, reliable, and accurate. Furthermore, they should be designed and validated using appropriate evidence-based methods to ensure that they are safe and effective for use in clinical settings.

CHAPTER ARRANGEMENT

The chapters of the dissertations will be arranged as follows:

Introduction	<ul style="list-style-type: none">• Discuision of the Problem statement concerning the research• Presenting the Research Questions, Aim, Objectives
Project Management and Issues	<ul style="list-style-type: none">• Presenting of research Issues• Presentation of Planning and Gantt chart
Literature Reviews	<ul style="list-style-type: none">• Review of Previous Research Papers• Knowledge gathering and gaining ideas of approaches and methods for data analytics and classification regarding heart disease with feature selection processes
Research Methodology	<ul style="list-style-type: none">• Presenting Research Methodology and Justifying Components• Description of Data Collection• Description of Tool, Terchnology Selection• Description and Justification of Algorithm and Execution Methods selection
Analysis and Result	<ul style="list-style-type: none">• Presenting the Results of Analyses doen in the artefact and interpret the results• Present the result of heart disease detection
Evaluation of Performance	<ul style="list-style-type: none">• Evaluation of Performances of Models employed for Heart Disease Detection• Selecting the present best model with the highest accuracy of detection• Determination of research improvements
Discussion and Conclusion	<ul style="list-style-type: none">• Discussion on research questions, research gaps and challenges• Concluding research along with emphasizing the limitations and futrure scopes

Limitations

The current research has certain limitations that are worth discussing.

- The data used in this study is derived from a single source, although the idea behind the data collection was inspired by previous research.
- Deep learning models have not been utilized for the detection of heart disease in this research.
- The quantile method, which has been employed solely for detecting outliers, may have limitations in accurately identifying heart disease.
- The training, validation, and test data have been prepared using a fixed ratio of 60:20:20 for which the justification has not been applied.
- The tuning of the hyperparameters of the algorithms has not been performed, which could potentially impact the accuracy and performance of the detection models.
- The detection of heart disease has been carried out using only a limited number of features obtained through a combination of correlation and the Bayesian model.

Future Recommendations

In order to further enhance research in the future, there are several techniques that can be employed. Firstly, the implementation of a Neural Network model using deep learning technology can greatly improve the detection of heart disease. By leveraging the power of artificial intelligence and advanced algorithms, this approach can analyze a large amount of data and identify patterns that are not easily discernible to humans. Additionally, models from Transfer learning can be applied to enhance the accuracy of heart disease detection. This technique allows researchers to leverage knowledge from pre-trained models in related domains, effectively utilizing existing knowledge to improve the detection capabilities. Furthermore, utilizing various feature selection methods can help identify and prioritize important features that can aid in the detection of heart disease. By focusing on the most relevant variables, researchers can streamline the analysis process and potentially improve accuracy. Lastly, an application interface can be designed where users can input their symptoms, and the detection results can be shown. This user-friendly interface can facilitate early detection and timely intervention, potentially saving lives. By implementing these techniques, future research can significantly advance the field of heart disease detection and contribute to improving overall healthcare outcomes.

CONCLUSION

Symptom-based disease detection plays a crucial role in the early identification and diagnosis of various medical conditions. Traditionally, healthcare professionals heavily rely on their expertise and clinical judgment to interpret symptoms and make accurate diagnoses. However, with the advancements in technology, particularly in the fields of Bayesian probability and machine learning, there is a promising opportunity to enhance disease detection by leveraging these techniques. The present research explored the potential of applying Bayesian probability and machine learning algorithms to improve symptom-based disease detection, ultimately leading to more accurate and precise identification of heart disease.

Heart disease is a global epidemic, claiming countless lives each year. The key to effectively combating this deadly condition lies in early detection and diagnosis. Fortunately, the field of machine learning has shown great promise in aiding healthcare professionals in identifying and diagnosing heart disease. Machine learning is a branch of artificial intelligence that empowers computers to learn and enhance their performance through experience, without explicit programming. By analyzing vast datasets, machine learning algorithms can classify and predict outcomes accurately. When it comes to heart disease detection, these algorithms can scrutinize patient data, including medical history, symptoms, and test results, allowing for the identification of individuals at risk of developing heart disease. In addition to machine learning, another valuable tool in the fight against heart disease is Bayesian feature selection. This statistical method can determine the most pertinent features within a dataset for a given task. In the context of heart disease detection, Bayesian feature selection can pinpoint crucial risk factors, such as age, gender, blood pressure, cholesterol levels etc. By focusing on these factors, healthcare professionals can gain valuable insights into a patient's likelihood of developing heart disease. One of the primary advantages of utilizing machine learning classifications and Bayesian feature selection in heart disease detection is the significant improvement in diagnostic accuracy. Traditional diagnostic methods rely on manual analysis of patient data, which is time-consuming and prone to errors. Machine learning algorithms, on the other hand, are capable of analyzing extensive datasets in real-time, resulting in precise and timely diagnoses.

Moreover, machine learning classifications and Bayesian feature selection offer the ability to personalize treatment plans for heart disease patients. By thoroughly analyzing patient data, these algorithms can identify the most effective treatment options based on individual risk factors and medical history. This personalized approach ensures that patients receive tailored and optimized care, leading to better outcomes and improved quality of life.

In light of existing research, the aim of this study is to effectively classify and detect heart disease. The datasets used were collected from Kaggle, while the inspiration was drawn from previous research papers. The algorithms employed were carefully selected based on the knowledge gathered from previous researchers who had used them for similar purposes. To ensure accurate results, the data underwent various preprocessing techniques, including data cleaning, outlier detection and elimination (using the quantile method), and feature selection (using the Bayesian model). The selected model successfully detected heart disease, with the voting model achieving the highest accuracy rate of 95.21% and minimal model overfitting at 0.47%. This research has demonstrated significant improvements in heart disease detection accuracy compared to existing approaches.

REFERENCES

- Adorada, A. et al., 2021. Support vector machine-recursive feature elimination (svm-rfe) for selection of microrna expression features of breast cancer.. *In 2018 2nd international conference on informatics and computational sciences (ICICoS) IEEE.*, pp. (pp. 1-4).
- Aggarwal, R. & Kumar, S., 2022. MLPPCA: Heart Disease Detection using Machine learning. *Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pp. 1-5.
- Ali, L., Khan, S. U. & Anwar, M., 2019. Early Detection of Heart Failure by Reducing the Time Complexity of the Machine Learning based Predictive Model. *International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pp. 1-5.
- Ali, M. M. et al., 2021. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, Volume 136, p. 104672.
- Aljaaf, A. J. & Jumeily, D. A.-., 2015. Predicting the likelihood of heart failure with a multi level risk assessment using decision tree. *Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAECE)*, pp. 101-106.
- Al-Mashagbeh, M. H. & Ababneh, M., 2021. Tor Detection using a Machine Learning Approach Using Correlation based Feature Selection with Best First and Random Forest.. *In 2021 International Conference on Information Technology (ICIT) IEEE.*, pp. (pp. 893-898)..
- Andreu-Perez, et al., 2015. Big data for health. *IEEE Journal of Biomedical and Health Informatics*, 19(4), pp. 1193-1208.
- Arslan, E. & B.-N. U. M., 2017. Bayesian top scoring pairs for feature selection.. *In 2017 51st Asilomar Conference on Signals, Systems, and ComputersIEEE.*, pp. (pp. 387-391)..
- Atallah, R. & A.-M. A., 2019. Heart disease detection using machine learning majority voting ensemble method.. *In 2019 2nd international conference on new trends in computing sciences (ictcs) IEEE.*, pp. (pp. 1-6)..
- Bakhshi, M., Mirtaheri, S. L. & Greco, S., 2022. Heart Disease Prediction Using Hybrid Machine Learning Model Based on Decision Tree and Neural Network.. *In 2022 9th International Conference on Soft Computing & Machine Intelligence (ISCM) IEEE.*, pp. (pp. 36-41)..

Boddington, P., 2009. HEART DISEASE AND SOCIAL INEQUALITY: ETHICAL ISSUES IN THE AETIOLOGY, PREVENTION AND TREATMENT OF HEART DISEASE. *Bioethics*, pp. 1-6.

Elhaj, F. A. et al., 2017. Hybrid classification of Bayesian and extreme learning machine for heartbeat classification of arrhythmia detection.. In *2017 6th ICT International Student Project Conference (ICT-ISPC) IEEE.*, pp. 1-4.

Esfahani, H. A. & G. M., 2017. Cardiovascular disease detection using a new ensemble classifier.. In *2017 IEEE 4th international conference on knowledge-based engineering and innovation (KBEI) . IEEE.*, pp. (pp. 1011-1014).

Follath, F., 2009. Ethical considerations in cardiovascular prevention. *Bioethics*, pp. 1-7.

Gaikwad, M. J., Asole, P. S. & Bitla, L. S., 2022. Effective Study of Machine Learning Algorithms for Heart Disease Prediction.. In *2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) IEEE.*, pp. (pp. 1-6)..

George, 2019. An Introduction to Bayesian Feature Selection. Introduction to Bayesian Econometrics. *Cambridge University Press*, pp. 1-10.

Geweid, G. G. & A. M. A., 2019. A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique.. *IEEE Access*, pp. 149595-149611.

Gulati, S., Guleria, K. & Goyal, N., 2022. Classification and Detection of Coronary Heart Disease using Machine Learning. *2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 1-5.

Gupta, Wang & Ganesh, 2020. An overview on deep learning-based methods for ECG-related human health anomaly detections: Scope and challenges.. *International Journal of Information Technology*, Volume 17, pp. 27-40.

Han, S., Eom, H., Kim, J. & Park, C., 2020. Optimal DNN architecture search using Bayesian Optimization Hyperband for arrhythmia detection.. In *2020 IEEE Wireless Power Transfer Conference (WPTC) IEEE.*, pp. (pp. 357-360)..

Han, S., Eom, H., Kim, J. & Park, C., 2020. Optimal DNN architecture search using Bayesian Optimization Hyperband for arrhythmia detection.. In *2020 IEEE Wireless Power Transfer Conference (WPTC) IEEE.*, pp. (pp. 357-360)..

Hostiadi, D. et al., 2022. A New Approach Feature Selection for Intrusion Detection System Using Correlation Analysis.. In *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS) IEEE.*, pp. (pp. 1-6)..

Hua, Q. et al., 2021. An interpretable model for ECG data based on Bayesian neural networks.. *IEEE Access*, pp. 57001-57009.

Iqbal, J., Iqbal, M. M., Khadam, U. & Nawaz, A., 2020. Ordinary Learning Method for Heart Disease Detection using Clinical Data.. In *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) IEEE.*, pp. (pp. 1-6)..

Kaushik, J. &, 2022. Comparison of Machine Learning Algorithms for Predicting Chronic Kidney Disease.. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE.*, pp. (pp. 1134-1139).

Kavitha, K. R., Harishankar, U. N. & Akhil, M. C., 2018. PSO based feature selection of gene for cancer classification using SVM-RFE.. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE.*, pp. (pp. 1012-1016).

- Khan, M. I. H. & M. M. R. H., 2020. Effectiveness of Data Driven Diagnosis of Heart Disease.. *In 2020 11th International Conference on Electrical and Computer Engineering (ICECE) IEEE.*, pp. (pp. 419-422)..
- Mahmood, A. M. & K. M. R., 2010. Early detection of clinical parameters in heart disease by improved decision tree algorithm.. *In 2010 Second Vaagdevi International Conference on Information Technology for Real World Problems*, pp. 24-29.
- Manikandan, S., 2017. Heart attack prediction system. *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 817-820.
- Masoudi-Nejad & Goliaei, A. &, 2012. A pairwise nested logit model for gene and gene feature selection. *Computer Methods and Programs in Biomedicine*, 107(3), pp. 479-491.
- Mohan, N., Jain, V. & Agrawal, G., 2021. Heart Disease Prediction Using Supervised Machine Learning Algorithms. *5th International Conference on Information Systems and Computer Networks (ISCON)*, pp. 1-3.
- Mortazavi, et al., 2016. Analysis of machine learning techniques for heart failure readmissions.. *Circulation: Cardiovascular Quality and Outcomes*, 9(6), pp. 629-640.
- Nugroho, A., Fanani, A. Z. & Shidik, G. F., 2021. Evaluation of feature selection using wrapper for numeric dataset with random forest algorithm.. *In 2021 International Seminar on Application for Technology of Information and Communication (iSemantic) IEEE.*, pp. (pp. 179-183)..
- Panigrahi, S. S. & K. N., 2022. Hybrid Classification Method for the Heart Disease Prediction.. *In 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) IEEE.*, pp. 494-499.
- Patra, R. & K. B., 2019. Predictive analysis of rapid spread of heart disease with data mining.. *In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) IEEE.*, pp. 1-4.
- Pour, A. F. & D. L. A., 2014. Optimal Bayesian feature selection on high dimensional gene expression data.. *In 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP) IEEE.*, pp. (pp. 1402-1405)..
- Pour, A. F. & D. L. A., 2017. Multiclass Bayesian feature selection.. *In 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP) IEEE.*, pp. (pp. 725-729)..
- Rahman, T. M. S. S. R. S. E. H. N. & I. M. H., 2019. Early detection of kidney disease using ECG signals through machine learning based modelling.. *In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) IEEE.*, pp. (pp. 319-323)..
- Razak, A. et al., 2022. Online feature Selection using Pearson Correlation Technique.. *In 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE) IEEE.*, pp. (Vol. 7, pp. 172-177)..
- Rhahla, M., Allegue, S. & Abdellatif, T., 2021. Guidelines for GDPR compliance in Big Data systems. *Journal of Information Security and Applications*, Volume 61, p. 102896.
- Ross, G. et al., 2023. Best practices and current implementation of emerging smartphone-based (bio)sensors – Part 1: Data handling and ethics. *TrAC Trends in Analytical Chemistry*, Volume 158, p. 116863.
- Sahoo, G. K., Kanike, K., Das, S. K. & Singh, P., 2022. Machine Learning-Based Heart Disease Prediction: A Study for Home Personalized Care.. *In 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP) . IEEE.*, pp. (pp. 01-06).

Shishah, W., 2022. An Efficient Early Stage Heart Disease Risk Detection Using Machine Learning Techniques.. *In 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)IEEE.*, pp. 1-6.

Srinivas, K. & Rao, G. R., 2015. Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. *5th International Conference on Computer Science & Educatio*, pp. 1344-1349.

Srivastava, A., Kumar, Mahesh, T. R. & V., V., 2022. Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms.. *In 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) IEEE.*, pp. (pp. 1-4)..

Suganthi, N., Jemin, V. M., Rama, P. & Chandralekha, E., 2022. Chronic Kidney Disease Detection using AdaBoosting Ensemble Method and K-Fold Cross Validation.. *In 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS). IEEE.*, pp. (pp. 979-983).

Tanuku, S. et al., 2022. Disease Prediction Using Ensemble Technique.. *In 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) IEEE.*, pp. (Vol. 1, pp. 1522-1525)..

UCI, 1988. *Heart Disease Data Set*. [Online] Available at: <https://archive.ics.uci.edu/ml/datasets/heart+disease> [Accessed 2022].

Wilmot, S., 2001. Nurses and whistleblowing: the ethical issues. *Bioethics*.

Yadav, S. S., Jadhav, S. M. & Nagrale, S. & P., 2020. Application of machine learning for the detection of heart disease.. *In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) IEEE.*, pp. 165-172.

Zhang, D. et al., 2018. Meteorological Feature Selection Method Based on Information Value and Maximum Correlation.. *In 2018 Chinese Automation Congress (CAC) . IEEE.*, pp. (pp. 3159-3164).

Zhang, J. et al., 2022. Research on Feature Selection Method Based on Bayesian Network and Importance Measures.. *In 2022 13th International Conference on Reliability, Maintainability, and Safety (ICRMS) IEEE.*, pp. (pp. 18-22)..

Zhao, L., Deng, F., Zhang, X. & Yu, N., 2022. RFE Based Feature Selection Improves Performance of Classifying Multiple-causes Deaths in Colorectal Cancer.. *In 2022 7th International Conference on Intelligent Informatics and Biomedical Science (ICIIBMS) IEEE.*, pp. (Vol. 7, pp. 188-194)..

Zhou, Z. & Wang, Y. & L. M., 2020. Feature selection method based on hybrid SA-GA and random forests.. *In 2020 International Conference on Computing and Data Science (CDS) . IEEE.*, pp. (pp. 139-142).