MACHINE LEARNING APPROACHES FOR EARLY DETECTION OF MULTIPLE DISEASES USING HEALTHCARE DATA

BY

MOMINUR RAHMAN
ID: 203-35-3131

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Software Engineering

Supervised By

A.H.M SHAHARIAR PARVEZ
Associate Professor
Department of Software Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

August 2024

APPROVAL

This thesis titled on **"Machine Learning Approaches for early Detection of Multiple Diseases Using Healthcare Data"**, submitted by **Mominur Rahman (ID: 203-35-3131)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

______________________________ **Chairman**

Dr. Imran Mahmud
Associate Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

______________________________ **Internal Examiner 1**

Md. Maruf Hasan
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

______________________________ **Internal Examiner 2**

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

______________________________ **External Examiner**

Dr. Md. Sazzadur Rahman
Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

I hereby declare that this project has been done by us under the supervision of **A.H.M Shahariar Parvez, Associate Professor, Department of Software Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

A.H.M Shahariar Parvez
Associate Professor
Department of SWE
Daffodil International University

Submitted by:

Mominur Rahman
ID: -203-35-3131
Department of SWE
Daffodil International University

ACKNOWLEDGEMENT

First and foremost, we would like to say the words of sincere thanks and appreciation to Almighty for the divine intervention gracing it possible for us to complete the final year project / internship to the successful end it has attained today.

I would like to express our immense thanks to **A. H. M Shahariar Parvez Department of SWE** Daffodil International University, Dhaka. Knowledge and the strong interest that my supervisor has in the field of “Machine Learning” to undertake this project. Much appreciation to him for his tireless patience, scholarly direction, encouragement, consistent, and energetic monitoring, feedback, counsel, reading numerous substandard exemplars, and correcting at every juncture that has made the accomplishment of this project possible.

It is my pleasure to acknowledge our warmest thank to the **Head of the Department of SWE** for his kind cooperation in completing our project and others members of the faculty as well as of the staff of the Department of SWE, Daffodil International University.

I would like also to thank all our course mates in Daffodil International University for active participation in this discussion while carrying out the course work.

Last but not least, I have to pay my respects to the much-watched support and wonderful patience of both my parents.

ABSTRACT

This paper aims at exploring the application of machine learning algorithms for early disease diagnosis through health information. A comprehensive dataset from Kaggle was used, involving 500 patient records with 23 variables, including symptoms of common diseases like as diabetes, malaria, dengue, typhoid, and hepatitis B. For improved data integrity and relevance, the dataset underwent significant preprocessing, which included missing value removal and feature selection. Five machine learning techniques, including GBC, GNB, RF, LR, and SVC, were tested for their ability to predict disease start. The results showed great accuracy rates across many methods, with RF having the best accuracy of 99.00%. Feature significance analysis offered information on the most selective symptoms for illness prediction. Furthermore, the study showed the social benefits of early disease identification, stressing better patient outcomes and lower healthcare expenses. Ethical considerations, sustainability strategies, and future research opportunities were discussed to help guide the appropriate deployment and continual improvement of predictive models in healthcare settings. This study improves the field of early detection of illness and highlights the transformational potential of machine learning in healthcare delivery. Future research topics include hybrid techniques, algorithmic bias protection, and adding predictive models into clinical decision support systems to improve diagnostic accuracy and patient care.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

1.1 Introduction

Early detection of diseases is critical in modern healthcare for successful treatment and better patient outcomes. Machine learning techniques, particularly those applied to healthcare data, have emerged as an attractive option for early disease identification. In this study, we focus on creating predictive models for the early identification of numerous diseases using a large dataset collected by verified sources such as Kaggle [1]

The dataset has 500 patient records, each with 23 relevant variables being a variety of symptoms and health markers. These features include common symptoms like skin rash, joint ache, weariness, and more, offering an in-depth picture of the patient's health. Furthermore, the information includes five target variables describing the frequency or absence of specific diseases, including diabetes, malaria, dengue, typhoid. Typhoid, and Hepatitis B. [2]

Prior to model creation, a careful data selection process was carried out to ensure the inclusion of key elements while removing redundant or irrelevant ones. Following that, a variety of statistical analyses were carried out to gain insight into the dataset's properties, including descriptive statistics, frequency distribution, correlation analysis, and visualization methods. [3]

The goal of this work is to create strong prediction models for early illness identification using machine learning methods such as GBC, GaussianNB, RF, LR, and SVC. Using these models, healthcare practitioners may be able to recognize and intervene in illness progression at an early stage, improving treatment efficacy and patient care. This study contributes to ongoing attempts to use technology to change healthcare delivery and reduce the disease burden on society.

Machine Learning Model

Logistic Regression

Logistic regression is a fundamental statistical model for binary classification that is widely utilized. In contrast to linear regression, it represents the relationship between the input variables and the likelihood of belonging to a specific class. The logistic function (sigmoid) is used in logistic regression to transfer the linear combination of input features to a probability score. This probabilistic technique enables intuitive interpretation as well as the estimation of class probabilities. It is noise-resistant and computationally efficient, and it works well with both continuous and categorical predictors.

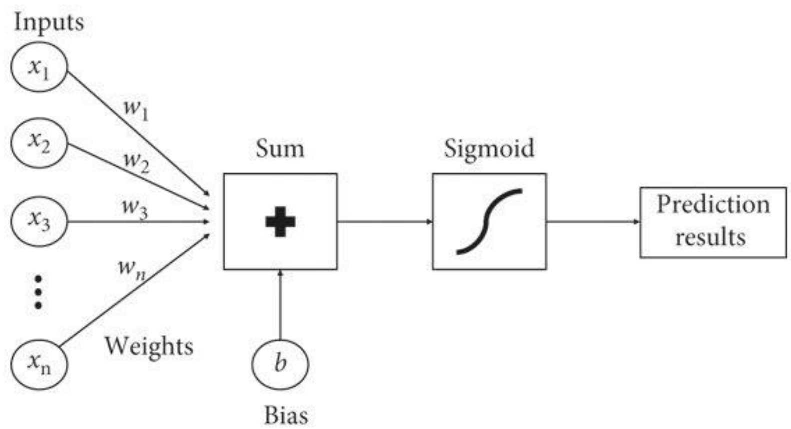

Figure 1.1: Proposed LR Architecture M. Bari Antor et al 2021[4]

Gradient Boosting Classifier

Gradient Boosting is an ensemble learning technique that combines the predictions of multiple weak learners, most commonly decision trees, to produce a strong predictive model. Gradient Boosting, unlike Random Forests, builds trees sequentially, with each tree correcting the errors of its predecessor. On the basis of the original data, a simple model, often a shallow decision tree, is built. Following trees are built to correct the errors introduced by the combined predictions of the existing trees. The main idea is to fit each new tree to the ensemble's residuals (the differences between the actual and predicted values). Each tree is trained to capture patterns that the previous trees were unable to adequately model. Gradient Boosting term "Gradient" refers to the use of gradient descent optimization to minimize the loss function, The learning rate parameter governs each tree's contribution to the ensemble. Gradient Boosting is more sensitive to overfitting than

Random Forests, but it often achieves higher predictive accuracy when properly tuned.

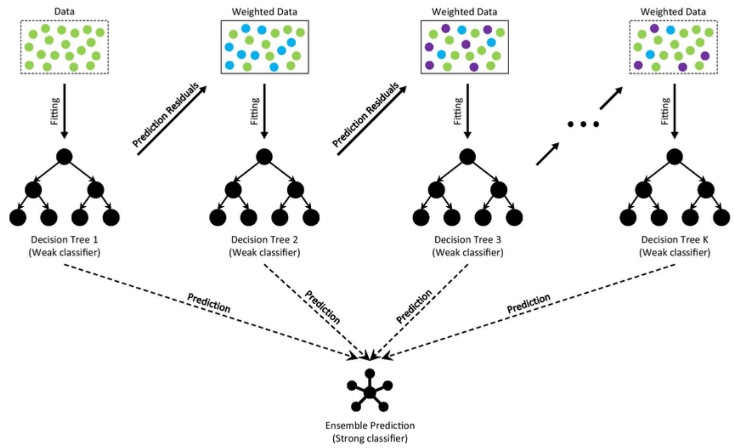

Figure 1.2: GB Architecture H. Deng, Y. Zhou et al.2021 [5]

Support Vector Machine

The Support Vector Machine (SVM) is a powerful supervised learning algorithm that we use in our research to detect early disease. SVM is particularly well-suited to binary classification tasks, such as identifying between disease and non-disease states using patient symptoms and health indicators. One of the primary benefits of SVM is its ability to effectively handle high-dimensional data, making it ideal for analyzing complex healthcare datasets with numerous features. Furthermore, SVM has a solid theoretical foundation and is well-known for its ability to identify optimal hyperplanes that maximize the margin between classes, thereby improving generalization performance.

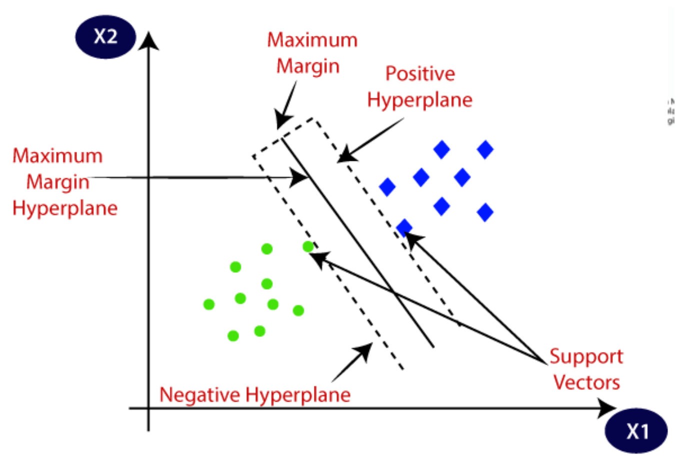

Figure 1.3: SVM Architecture [6]

Random Forest

The RF classifier is a method that creates a lot of decision trees at the same time using bootstrapping and aggregation, also called bagging. When the decision trees are trained in parallel on different subsets of the training data and the same, but only partially available features, it is call bootstrapping. By ensuring that every decision tree in the random forest is distinct, bootstrapping lowers the RF classifier's overall variance. Because the RF classifier aggregates the decisions made by individual trees to arrive at a final decision, it demonstrates strong generalization. With no overfitting problems, the RF classifier typically performs better in terms of accuracy than the majority of other classification techniques. RF classifier does not require feature scaling, similar to DT classifier. The RF classifier is more resilient to noise in the training dataset and the selection of training samples than the DT classifier. When compared to the DT classifier, the RF classifier is more difficult to understand but simpler to adjust the hyperparameter.

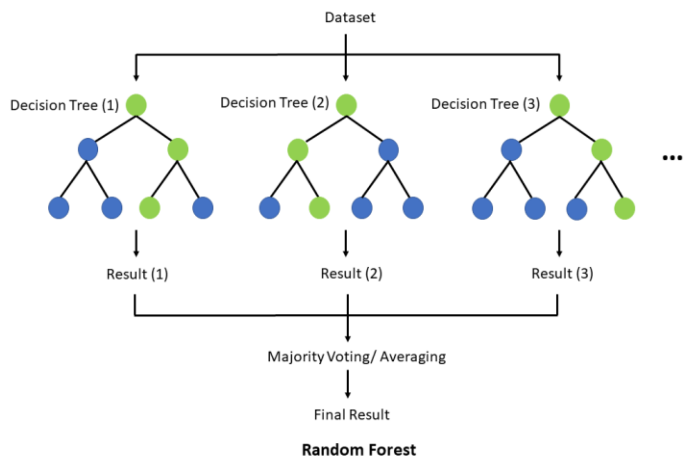

Figure 1.4: RF Architecture Alanazi, R et al. 2020[7]

Gaussian Naïve Bayes

Gaussian Naive Bayes (NB) is a logical classification algorithm based on Bayes' theorem that is especially well suited for situations where features are assumed to be normally distributed. A simple and fast algorithm that works well in high dimensions and is widely used for text categorization and spam detection. The term "Naive" refers to the assumption of feature independence, which means that the presence of one feature does not affect the presence of another, given the class label. Despite this simplifying assumption, Gaussian NB frequently performs admirably in practice.The algorithm computes the likelihood of a given sample belonging to a specific class by multiplying the individual probabilities of each feature within that class. It takes into account the prior probabilities of the classes and normalizes the results. The Gaussian NB model is based on the assumption that continuous features have a Gaussian (normal) distribution. When dealing with real-valued features, Gaussian Naive Bayes is especially useful because it is less sensitive to irrelevant features. While its feature independence assumption may not always hold true in real-world scenarios, Gaussian NB is a fast and effective classifier, especially when the underlying independence assumption is reasonable.

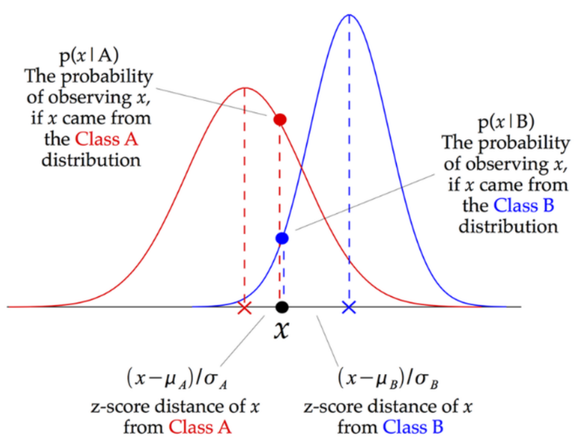

Figure 1.5: GNB Architecture Raizada, et al. 2013 [8]

1.2 Motivation

This research is motivated by an urgent need to enhance healthcare outcomes through early disease diagnosis. Timely diagnosis is critical in reducing the impact of diseases on individuals and society as a whole. However, traditional tests frequently rely on symptomatic presentation, which might postpone care until diseases have proceeded to advanced stages. Machine learning offers a harmful answer by analyzing huge amounts of healthcare data to detect subtle trends and signs of disease before symptoms show. We hope to develop algorithms that can detect early signs of diseases such as diabetes, malaria, dengue fever, typhoid, and hepatitis B by using the power of predictive modelling. The development has the potential to significantly affect healthcare delivery by enabling proactive interventions and individualized treatment programs. These models have the potential to decrease healthcare expenditures, reduce patient suffering, and even save lives by aiding early detection. This study fits in with the larger objective of using technology to address urgent healthcare issues and improve people's well-being around the world.

1.3 Rational of the Study

The rationale for this study is to address a key gap in current healthcare systems in terms of early disease diagnosis. We hope to construct predictive models capable of detecting subtle disease indications before symptoms appear obviously by using machine learning algorithms to comprehensive healthcare databases. Early detection allows healthcare practitioners to personalize treatment strategies and manage resources more efficiently, in addition to allowing for quick action. Furthermore, the ability to predict various diseases based on common symptoms is extremely valuable for improving diagnostic workflows and reducing diagnostic uncertainty. Through this research, we hope to contribute to the improvement of healthcare technology, ultimately improving patient outcomes, reducing healthcare costs, and reducing illness burden on individuals and healthcare systems alike.

1.4 Research Questions

1. How does machine learning improve early detection of common diseases using healthcare data?
2. What are the most effective predictive modeling techniques for identifying disease onset based on patient symptoms?
3. Can machine learning algorithms accurately differentiate between similar diseases like Malaria and Dengue?
4. How do different machine learning approaches compare in predicting the onset of Typhoid fever?
5. What ethical considerations arise when deploying machine learning for disease detection in healthcare?
6. How can predictive models be integrated into clinical decision support systems to aid healthcare professionals?
7. What implications do machine learning-based disease detection models have for public health and healthcare resource allocation?

1.5 Expected Output

This project is expected to produce powerful machine learning models capable of accurately predicting the start of numerous diseases, such as diabetes, malaria, dengue fever, typhoid, and hepatitis B, based on early symptoms and health markers. These models are expected to have good predictive performance, as shown by careful review using relevant measures such as accuracy, precision, recall, F1-score, and ROC-AUC. The insights generated by feature importance analysis will shed light on the key symptoms and health indicators that drive disease prediction, allowing for the discovery of early warning signs and customized treatments. Furthermore, comparative analyses of various machine learning algorithms will help guide algorithm selection for specific disease prediction tasks, allowing for more informed decision-making in healthcare settings. Finally, the successful implementation of predictive models is expected to improve early disease detection, patient outcomes, and advance personalized medicine and public health initiatives.

1.6 Project Management and Finance

This research project needs careful planning, organization, and task execution to ensure timely completion and adherence to project milestones. Data collection, preprocessing, model construction, evaluation, and distribution of results are all important processes. Effective communication and collaboration among team members, which include data scientists, healthcare professionals, and domain specialists, is critical to project success. Finance needs budget allocation for collecting data, software tools, computing resources, and people expenses. Research grants, institutional sponsorship, and collaborations with industry partners are all potential sources of funding. Transparent budget management and accountability are critical for improving resource usage and increasing research impact.

TABLE 1.6: ESTIMATED COST

SL. N	Components	Estimated Cost(BDT)
1	Hardware's	500-1600
2	Software and Tools	500-1500
3	Data Collection & Processing	5000-6000
4	Report Writing	500-1500
5	Miscellaneous	500-1500
6	Contingency	600-1200
Total Cost		7600-13,300

CHAPTER 2

2.1 Preliminaries

Before getting into the specifics of early disease detection using machine learning, it is critical to have a solid understanding of the fundamental concepts and principles that underpin this research subject. This section, titled "Preliminaries," serves as a primer, giving readers important background information. This section offers an introduction of machine learning techniques that are often used in healthcare applications, including GBC, GNB, LR, RF and SVC. Fundamental ideas in healthcare data analysis are also presented, including data pretreatment methods, feature selection strategies, and model evaluation metrics. The Preliminaries section aims to improve comprehension of subsequent chapters by alerting readers to these fundamental concepts, allowing them to engage more thoroughly with the study findings and techniques described in the report. This basic understanding serves as a platform for later debates about early disease diagnosis and its impact on society.

2.2 Related Works

This review provides a basic understanding of the field and informs the design of our research method in given below:

Anthony et al. [1] review explores the landscape of healthcare analytics, focusing on disease classification methodologies. It highlights the limitations of traditional approaches and proposes the use of multiclass supervised learning for enhanced accuracy. Previous studies are reviewed to inform the development of a robust model, aiming to improve patient outcomes and healthcare decision-making.

Mohammed Badawy et al [2] provides a discussion on how the fields of machine learning (ML) and deep learning (DL) have affected or contributed to healthcare predictive analytics. It highlights the importance of accurate disease prediction for patient outcomes and identifies ML and DL as crucial tools in analyzing complex healthcare data. Through a review of 41 papers from 2019 to 2022, the study explores methodologies and challenges in applying ML and DL in healthcare predictive analytics, emphasizing their potential in enhancing disease diagnosis and contributing to the field's advancement.

Explores the application of machine learning (ML) in heart disease detection within e-healthcare proposed by Joshua Cena, et al [3]. It emphasizes the need for accurate diagnostic tools in cardiovascular health and investigates various ML techniques using comprehensive datasets. The study highlights the performance of ML algorithms such as support vector machines (SVM), random forests, and artificial neural networks (ANN), showcasing their promising accuracy and efficacy. Additionally, it discusses the importance of preprocessing, feature selection, interpretability methods, and ensemble learning in enhancing model performance. Challenges such as data privacy concerns and model generalizability are acknowledged, with recommendations provided for implementing ML in clinical practice and future research directions. Overall, the review underscores the potential of ML in improving heart disease detection and patient outcomes in e-healthcare settings.

Mohammed Badawy et al. [4] presented the state-of-art studies and prospects of machine learning (ML) and deep learning (DL) schemes in healthcare predictive analytics. It emphasizes on the importance of the predictive analytics in the health sector especially in the use of the artificial intelligence in diagnosing diseases, disease diagnosis, and results in better outcomes. The review examines 40 papers from reputable journals, showcasing the impact of ML and DL algorithms in healthcare prediction. Despite acknowledging limitations, such as potential oversight of relevant approaches, the review underscores the transformative potential of AI in healthcare services. It concludes by advocating for further research to enhance the integration of AI with data quality management, aiming to optimize patient care and outcomes.

Senerath Mudalige et al. [5] review explores the vital role of machine learning (ML) in healthcare decision-making, emphasizing the need for early disease detection and accurate analysis of medical data. ML tools enable the identification of hidden abnormalities and relationships, aiding computational decision-making processes. Historical progressions highlight ML's evolution from processing medical data to advancing precision medicine. Applications include disease diagnosis, medical imaging, patient care, and resource allocation. Neural network-based deep learning methods exhibit high accuracy, driving computational biology advancements. ML's indispensable role in healthcare, particularly evident during the COVID-19 pandemic, underscores its significance in patient care, treatment research, and resource management. Overall, ML, as a subset of artificial intelligence, significantly enhances healthcare decision-making processes, ensuring efficient, accurate, and cost-effective outcomes across various healthcare domains.

Rayan Alanazi et al [7] focuses on the identification and prediction of chronic diseases using machine learning (ML) approaches. Acknowledging the importance of early disease detection, the paper aims to develop a reliable categorization method for identifying individuals with common chronic illnesses. ML techniques, particularly convolutional neural networks (CNN) for feature extraction and disease prediction, and K-nearest neighbor (KNN) for distance calculation, are employed for accurate disease prognosis based on patient symptoms and habits. The study highlights the advantage of using both structured and unstructured data for dataset preparation. Comparative analysis with Naïve Bayes, decision tree, and logistic regression algorithms demonstrates the proposed system's superior accuracy of 95%, potentially reducing the risk and cost of chronic diseases through early diagnosis and intervention.

Disease prediction using machine learning over big data in healthcare communities emphasizing the importance of accurate data analysis for early disease detection proposed by Min Chen, et al [8] . Challenges like incomplete data and regional variations in disease characteristics are highlighted. The study proposes a novel CNN-MDRP algorithm, integrating structured and unstructured data for multimodal disease risk prediction. Using real-life hospital data from central China, the algorithm achieves a high prediction accuracy of 94.8%, outperforming typical prediction methods. This approach demonstrates faster convergence speed and superior accuracy, showcasing the potential of machine learning in enhancing disease prediction and community healthcare services.

Ching-Hsien Hsu et al. [9] presents a computer-aided diagnosis system using machine learning models to improve cancer detection accuracy. Utilizing datasets from the University of California, Irvine repository for breast, cervical, and lung cancer, supervised learning algorithms are employed to train and validate optimal features. The proposed algorithm, GA-CFS, integrates genetic algorithm-based feature selection with correlation-based feature selection to identify informative features. MLP-NN shows the best performance across all datasets, followed by SVM, NB, and LDA. The GA-CFS model outperforms existing feature selection methods, providing promising diagnostic accuracy. The system serves as a diagnostic tool for medical practitioners, aiding in accurate cancer diagnosis and potentially expanding to handle complex, real-time datasets in the future.

Data mining for healthcare, stemming from database statistics, aids in evaluating medical interventions by K. Arumugam et al. [10]. Diabetes-related heart disease, a complication of diabetes, lacks sufficient data for prediction. Cardiovascular disease, a conjunction of illnesses affecting heart and blood vessels is difficult to foretell in diabetes. Despite various classification algorithms, the decision tree consistently outperforms naive Bayes and support vector machine models. Hence, fine-tuning the decision tree model optimizes forecasting heart disease likelihood in diabetic individuals. This interdisciplinary approach enhances medical therapy effectiveness assessment and aids in better predicting diabetes-related heart disease.

R. Venkatesh et al. [11] explores the critical role of predictive analytics in healthcare, particularly in predicting heart disease using machine learning techniques. Focusing on the Naive Bayes approach within a Big Data Predictive Analytics Model (BPA-NB), the study emphasizes the significance of accurate health predictions for improved patient outcomes. By leveraging probabilistic classification and clustering techniques, the BPA-NB scheme achieves high prediction accuracy (97.12%) and early disease detection. Employing Hadoop-Spark for big data processing enhances computational efficiency, while simulation results demonstrate notable improvements in accuracy, CPU utilization, and processing time. The study underscores the potential of machine learning-driven predictive analytics in healthcare and suggests avenues for future research in heterogeneous environments.

the dire need for improved medical services in Bangladesh due to doctor shortages and low salaries. It proposes MD Samiul Islam et al. [12] by leveraging modern technologies, particularly machine learning, facilitated by electrical device interaction, to bridge the patient-specialist gap. By harnessing big data, a novel platform integrating databases, mobile, and web applications enables patient-doctor interaction and predictive analysis. Achieving over 95% accuracy in disease detection at significantly lower costs than local hospitals, this platform promises to enhance healthcare accessibility in rural Bangladesh, despite challenges in data collection and digitalization.

2.3 Comparative Analysis and Summary

The literature review discusses various methodologies for using machine learning (ML) in healthcare to detect diseases and perform predictive analytics. Joshua Cena et al. [3] concentrate particularly on heart disease detection, while Anthony et al. [1] and Mohammed Badawy et al. [2] highlight the importance of ML in disease classification and predictive analytics. While Ching-Hsien Hsu et al. [9] and K. Arumugam et al. [10] focus on ML for cancer and heart disease diagnosis, respectively, Rayan Alanazi et al. [7] and Min Chen et al. [8] suggest ML-based approaches for chronic disease prediction. R. Venkatesh et al. [11] and MD Samiul Islam et al. [12] draw attention to the ways that machine learning-driven predictive analytics can enhance patient outcomes and accessibility to healthcare. All studies highlight machine learning's transformative potential in improving disease diagnosis, prediction, and patient care, despite differences in applications and methodologies.

2.4 Scope of the Problem

This work aims to construct and evaluate machine learning models for early disease diagnosis using healthcare data. The study focuses on planning the start of common diseases such as diabetes, malaria, dengue fever, typhoid, and hepatitis B using early symptoms and health markers. The study's goal is for studying the prediction powers of various machine learning algorithms, including GBC, GNB, LR, RF and SVC, in properly knowing disease patterns and trends. While the study focuses primarily on disease detection, it also analyses broader implications for healthcare delivery and overall well-being. However, it does not include clinical diagnosis or treatment decisions, which require expert medical knowledge and a thorough patient examination. In addition, the study does not address ethical, legal, or regulatory issues with healthcare data use and privacy.

2.5 Challenges

Developing and implementing machine learning models for early disease diagnosis using healthcare data is expected to offer a number of obstacles. To begin, getting high-quality, well-curated healthcare datasets with enough samples and correct marking is difficult due to data privacy concerns and data silos inside healthcare systems. Second, the interpretability of machine learning models in healthcare is still a concern, since complicated models may lack transparency, limiting their popularity and use by physicians. Furthermore, solving class imbalance in illness datasets, where some diseases may be less common than others, necessitates specific methods to avoid model bias and maintain robust performance across all disease categories. In addition, creating the generalizability of predictive models across multiple patient populations and healthcare settings is difficult, as models trained on a single dataset may not perform well when applied to different situations. Finally, ethical considerations such as patient approval, data security, and algorithmic fairness must be properly addressed in order to protect patient rights and trust in healthcare AI systems.

CHAPTER 3

3.1 Research Subject and Instrumentation

3.1.1 Research subject:

The study focuses on early detection of illnesses via machine learning techniques applied to healthcare data. The study's goal is to create predictive models that can detect the development of common diseases such as diabetes, malaria, dengue fever, typhoid, and hepatitis B using early symptoms and health markers.

3.1.2 Instrumentation:

The instrumentation for this study includes a variety of techniques, technologies, and methodologies for collecting, preprocessing, and translating healthcare data. This comprises programming languages like Python for data manipulation and build of models, machine learning libraries like scikit-learn and TensorFlow for algorithm implementation, and data analysis software. Furthermore, tools include data collecting technologies such as electronic health records, medical imaging data, and wearable symptom monitors. Data pretreatment techniques such as data cleaning, feature engineering, and dimensionality reduction help prepare data for machine learning models. Evaluation metrics and validation methodologies are also important instrumentation components to calculate predictive model performance and generalizability.

3.2 Data Collection Procedure

The data choosing process involves collecting a large healthcare dataset from Kaggle, a dependable site that hosts a wide range of datasets across multiple disciplines. A thorough search on Kaggle turned up a dataset containing patient records with symptoms and disease labels relevant to the research aims.

The dataset was chosen based on its relevance to early disease identification, sufficient sample size, and the availability of variables capturing symptoms connected to the diseases under inquiry (Diabetes, Malaria, Dengue, Typhoid, and Hepatitis B).

After selecting a suitable dataset, a preliminary examination was carried out to ensure data quality, integrity, and completeness. This included reviewing data documentation, looking for missing information, and ensuring the accuracy of disease classifications. After creating the dataset's status, it was downloaded from Kaggle and subjected to additional preprocessing steps in preparation for subsequent analysis and modelling. This data selection method resulted in the creation of a strong and reliable dataset for studying early disease detection with machine learning techniques. In the graph below, there is a graph for the number of datasets in Figure 3.1:

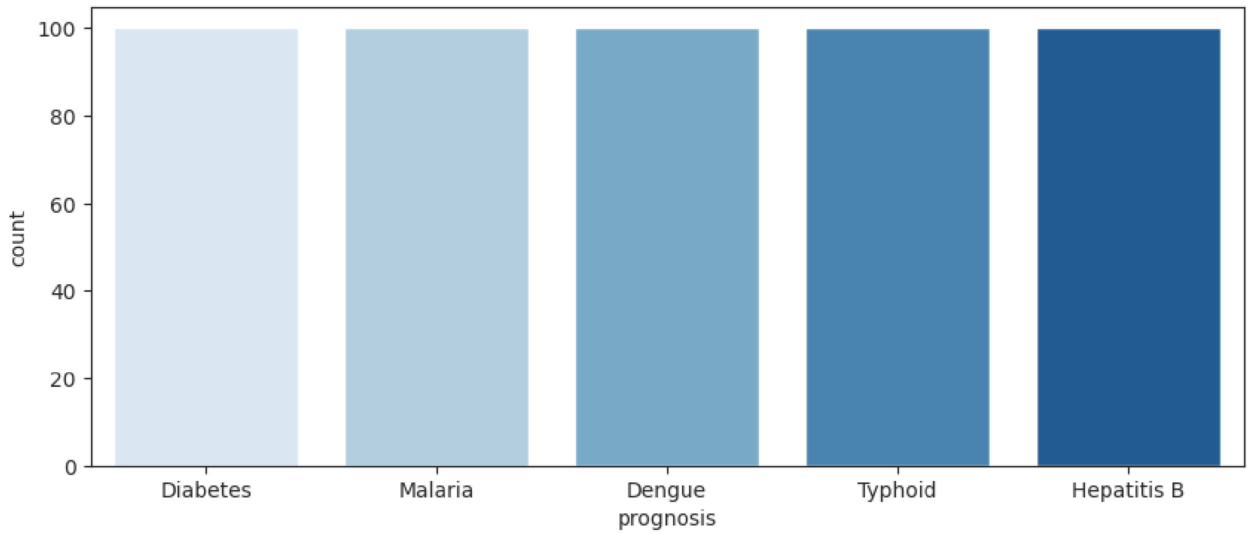

Figure 3.1: Number of datasets

Figure 3.1 shows the dataset includes 500 patient health records with 27 attributes that describe symptoms and indicators. It contains four target attributes that indicate the presence or absence of specific diseases: diabetes, malaria, dengue, typhoid, and hepatitis B. Each patient has 100 cases of diabetes, malaria, dengue, typhoid, and Hepatitis B.

3.3 Statistical Analysis

Statistical analysis is critical for getting relevant insights from healthcare data and detecting diseases early on. This involves employing a variety of statistical approaches to examine the dataset's features, detect trends, and evaluate correlations between variables.

Descriptive statistics, such as measures of central tendency and dispersion, offer an overview of data distribution and variability. Frequency distribution analysis may help you understand the prevalence of symptoms and diseases in your collection.

Correlation analysis analyses the degree and direction of connections between various symptoms and disease labels, which informs feature selection and model design. Visualization tools like histograms, box plots, and correlation matrices help show data patterns and trends, making interpretation and decision-making easy.

Statistical analysis gives researchers useful insights into the structure of the information, allowing them to make educated judgments during the subsequent preprocessing, modelling, and assessment stages of the research process.

3.4 Proposed Methodology

The proposed technique offers a systematic approach to creating machine learning models for early disease diagnosis using healthcare data. This section gives an organized framework for carrying out the research, such as crucial phases from data selection to model evaluation.

Data Collection:

Collect a comprehensive healthcare dataset from Kaggle, ensuring that it has essential attributes for early disease diagnosis and a large enough sample size to train and assess machine learning models.

Labeling:

Add disease labels to healthcare data to show which conditions are present or absent. For precise disease labeling, make use of medical standards, validated algorithms, or expert knowledge. Through quality assurance procedures and inter-rater reliability checks, make sure that labeling is consistent and reliable.

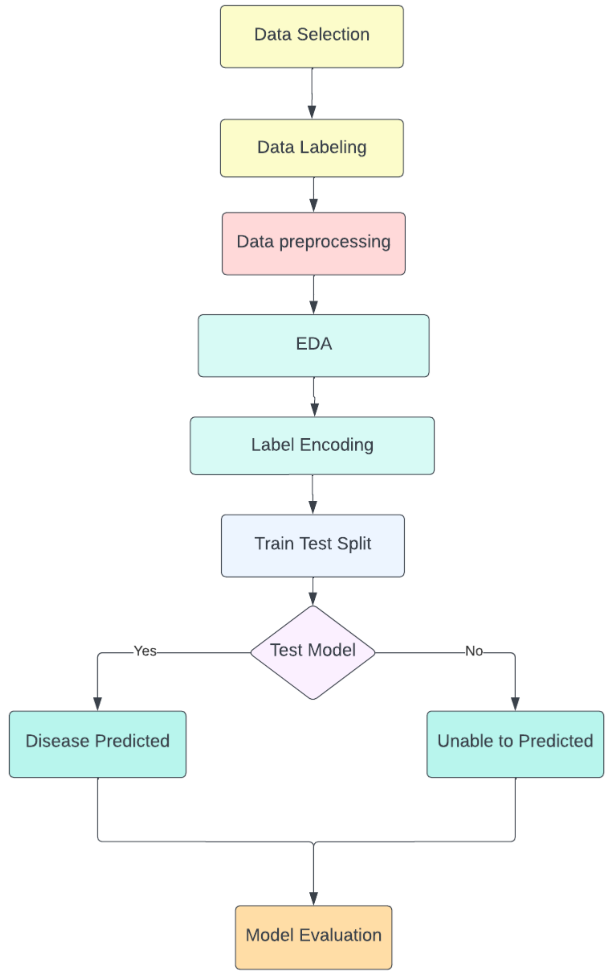

Figure 3.2: Working Flow

Data Pre- Processing:

To handle missing values, outliers, and noise, clean and preprocess raw healthcare data. Convert unstructured data into meaningful features by performing feature extraction and selection. To ensure consistent data representation, encode categorical variables and normalize numerical features.

Model Selection:

Look at a variety of machine learning algorithms, including support vector machines, random forests, logistic regression that are suitable for early disease detection. When selecting models for the task, take into account factors like interpretability, scalability, and algorithmic complexity.

Model Training:

Train machine learning models on preprocessed data using methods like GB Classifier, Gaussian NB, RF Classifier, LR, and SVC.

Model Evaluation:

Evaluate trained models' performance using relevant metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to ensure robustness and generalizability.

Test Model:

Apply the best-performing model to a different test dataset to evaluate its performance on before unseen data, verifying its suitability for real-world use in early illness detection scenarios.

3.5 Implementation Requirements

Machine learning models for early disease detection require a strong infrastructure involving hardware, software, data, knowledge, and validation methods. Adequate computer resources, such as processor power, memory, and storage capacity, are required to handle huge databases and complex methods well. Building models requires the use of computer languages such as Python or R, as well as machine learning libraries like scikit-learn and TensorFlow. High-quality healthcare datasets with relevant attributes must be collected from reliable sources while maintaining data privacy and standards of ethics. Successful implementation needs skilled professionals with skills in data science, machine learning, and healthcare analytics, as well as domain knowledge in healthcare. Collaboration between data scientists, doctors and nurses, and domain specialists makes model building easier. Predictive models' reliability and generalizability have been guaranteed through rigorous validation techniques such as cross-validation and testing on independent datasets.

CHAPTER 4

4.1 Experimental Setup

The experimental setup focuses on creating the environment for model training, evaluation, and validation. These covers breaking the dataset into training, validation, and test sets to effectively evaluate model performance. Model parameters are optimized using hyperparameter tuning methods such as grid search and random search. Cross-validation methods, such as k-fold cross-validation, are used to reduce overfitting and increase resilience. The selection of assessment factors, such as accuracy, precision, recall, F1-score, and ROC-AUC, is critical for evaluating model performance. Furthermore, tests may be undertaken to examine the effectiveness of several machine learning methods, such as GB Classifier, Gaussian NB, RF Classifier, LR, and SVC, in early disease diagnosis. The experimental setting was created to offer trustworthy information about the performance and generalizability of predictive models for real-world applications.

4.2 Experimental Results & Analysis

In this part experimental results show that machine learning models perform well in detecting early diseases using healthcare data. To assess predictive capabilities, models were evaluated using a variety of metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The results show the strengths and weaknesses of various algorithms for identifying disease patterns and trends. Furthermore, feature importance analysis reveals the most discriminative symptoms for each disease. Comparative evaluations show the efficacy of various machine learning techniques and help guide algorithm selection for specific disease prediction tasks. Furthermore, analyzing model predictions on test datasets validates their accuracy and generalizability. Overall, the experimental results and analysis help to gain a better understanding of the predictive modeling process for early disease detection, as well as valuable insights for future research and applications in healthcare.

Accuracy:

Accuracy assesses the quantitative performance of the model in terms of the percentage of correct samples out of the total samples. However, in cases of class imbalance, it simply provides a general idea about the performance of the model, although it might not be a complete picture.

$$Accuracy=(TP+TN)/(TP+TN+FP+FN)$$

Precision:

In terms of accuracy of the outcomes, precision is concerned with the proportion of the actual positive predictions out of all the predicted positive situations by the model.

$$Precision=TP/(TP+FP)$$

Recall: This can also be called sensitivity or true positive rate and is the ratio of true positive predictions to the total quantity of true positive samples.

$$Recall=TP/(TP+FN)$$

F1 Score: The F1 score is the weighted average of recall and precision defined as F1 score. It offers a fair measure of evaluation that into account recall and precision into consideration. The F1 measure is particularly appropriate here since the number of classes is unbalanced and it takes into consideration both false positives and false negatives. It is articulated that a high F1 score signifies a good balance between precision and recall.

$$F\text{-}1\ Score=2*(Precision*Recall)/\ (Precision + Recall)$$

Specificity: The percentage of people who do not have the ailment who have a negative test result is known as specificity. A highly specific test is effective in excluding the majority of individuals without the disease. Because of this, a positive outcome on a very precise test can definitively rule out the condition for a specific person. The equation is given below:

$$Specificity =TN / ((TN+FP))$$

Machine Learning Architecture

Logistic Regression:

For Test accuracy of Logistic Regression is 98.75%. In below Figure 4.2.1 describing the confusion matrix of LR:

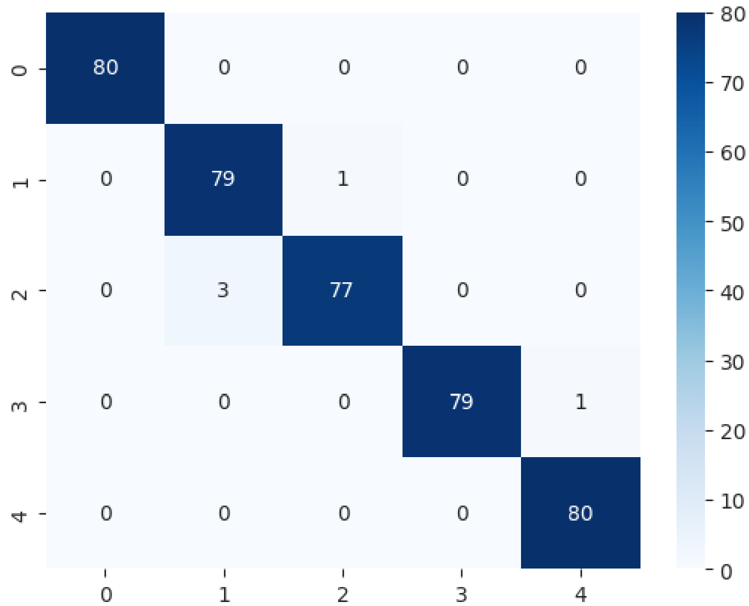

Figure 4.2.1: Confusion Matrix of LR

Figure 4.2.1 showed the confusion matrix of high accuracy across all classes, with minor variations. Class 0 shows perfect precision and recall, whereas class 1 has slightly lower precision. Class 2 shows a slight decrease in recall, while class 3 has perfect precision but a slightly lower recall. Class 4 shows high precision and recall. Overall, the model performs admirably, with a weighted average F1-score of 0.99, indicating robust classification across multiple classes.

GNB

Accomplishing Test Accuracy of GNB is 98.50%. In below Figure 4.2.2 about the confusion matrix of GNB.

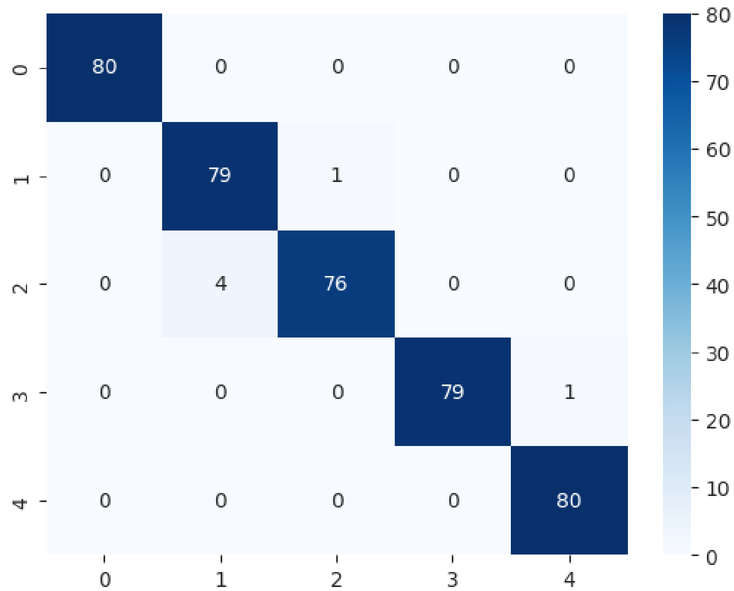

Figure 4.2.2: Confusion Matrix of GNB

Figure 4.2.2 the Gaussian Naive Bayes (GNB) model performs admirably with minor variations. Class 0 shows perfect precision and recall, whereas class 1 has slightly lower precision. Class 2 exhibits a minor decrease in recall while maintaining high precision. The model excels at classifying class 3, with perfect precision and high recall. Class 4 shows high precision and recall. Overall, the GNB model is robust, with a weighted average F1-score of 0.98, indicating accurate classification across multiple classes.

Support Vector Machine

In general, to reach the Test Accuracy of SVM namely 98.75%. In below Figure 4.2.3 discussing the confusion matrix of SVM.

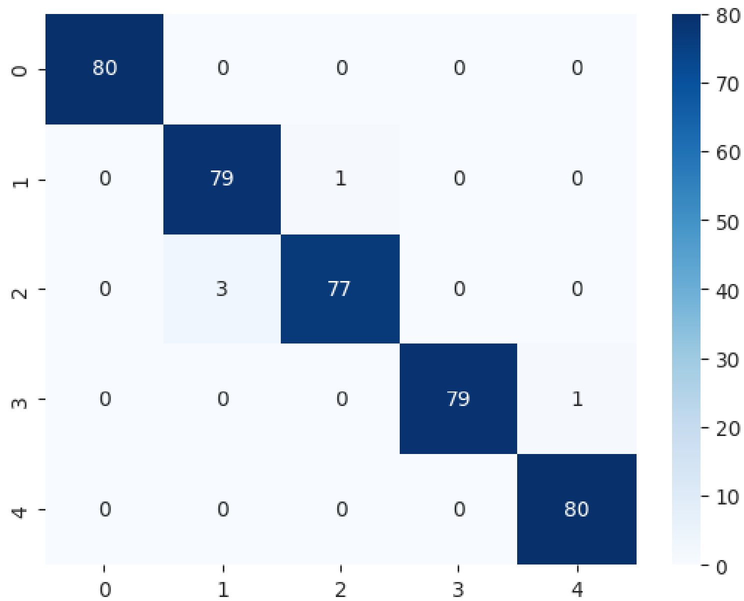

Figure 4.2.3: Confusion Matrix of DT

Figure 4.2.3 shows the SVM performs exceptionally well across all classes. Class 0 shows perfect precision and recall, implying precise classification. Class 1 and Class 2 show high precision and recall, with minor deviations. Class 3 shows perfect precision and high recall. Class 4 exhibits high precision and recall as well. Overall, the SVM model is robust, with a weighted average F1-score of 0.99, indicating accurate classification across a variety of classes.

GB Classifier

To achieve GBC Test Accuracy of 98.75%. In below Figure 4.2.4 that explains the confusion matrix of GBC.

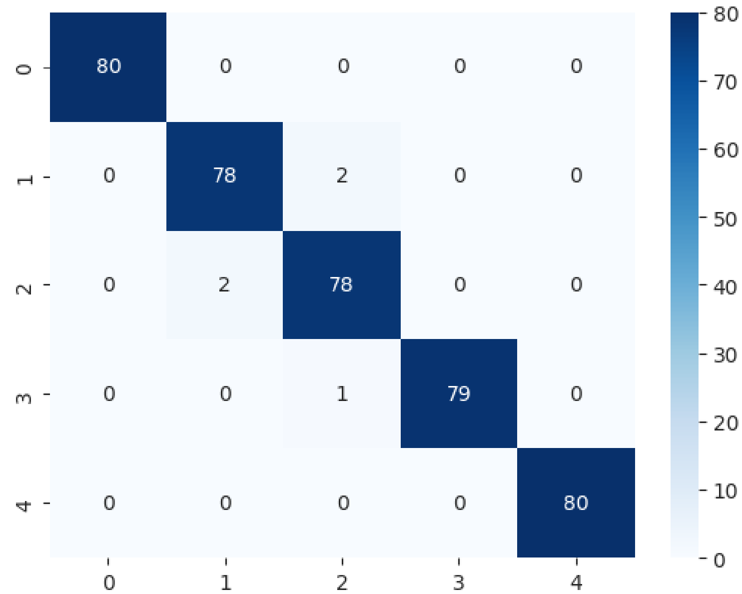

Figure 4.2.4: Confusion Matrix of GBC

Figure 4.2.4 shows the GBC model produces impressive results in all classes. Class 0 shows perfect precision and recall, indicating flawless classification. Classes 1 and 2 show high precision and recall, with minor deviations. Class 3 demonstrates perfect precision and high recall, whereas class 4 demonstrates both. Overall, the GBC model is robust, with a weighted average F1-score of 0.99, indicating accuracy in classifying various classes with precision and recall.

RF Classifier

Achieving outstanding Test Accuracy of RF is 99%. Below Figure 4.2.5 describing the confusion matrix of RF.

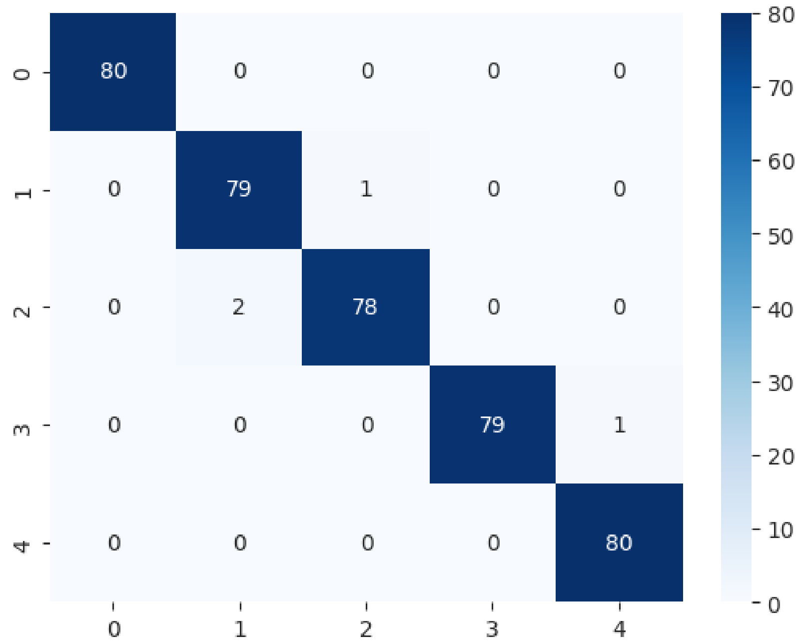

Figure 4.2.5: Confusion Matrix (RF)

Figure 4.2.5 show The Random Forest (RF) model performs exceptionally well across all classes. Class 0 demonstrates perfect precision and recall, indicating flawless classification. Classes 1, 2, and 4 show high precision and recall, with minor deviations. Class 3 shows perfect precision and high recall. Overall, the RF model is robust, with a weighted average F1-score of 0.99, indicating that it can accurately classify various classes with precision and recall.

The result of Machine learning model is compared on the basis of Accuracy, Precision, Recall, F1 Score in below table of 4.1:

Table 4.1. PERFPRMANCE EVA;UATION

Model Name	Accuracy	Precision	Recall	F1-Score
LR	98.75%	99%	99%	99%
RF	99%	99%	99%	99%
Gaussian NB	98.50 %	99%	98%	98%
SVM	98.75%	99%	99%	99%
GBC	98.75%	99%	99%	99%

In given below I am showing Comparative Model Accuracy Bar Plot:4.6:

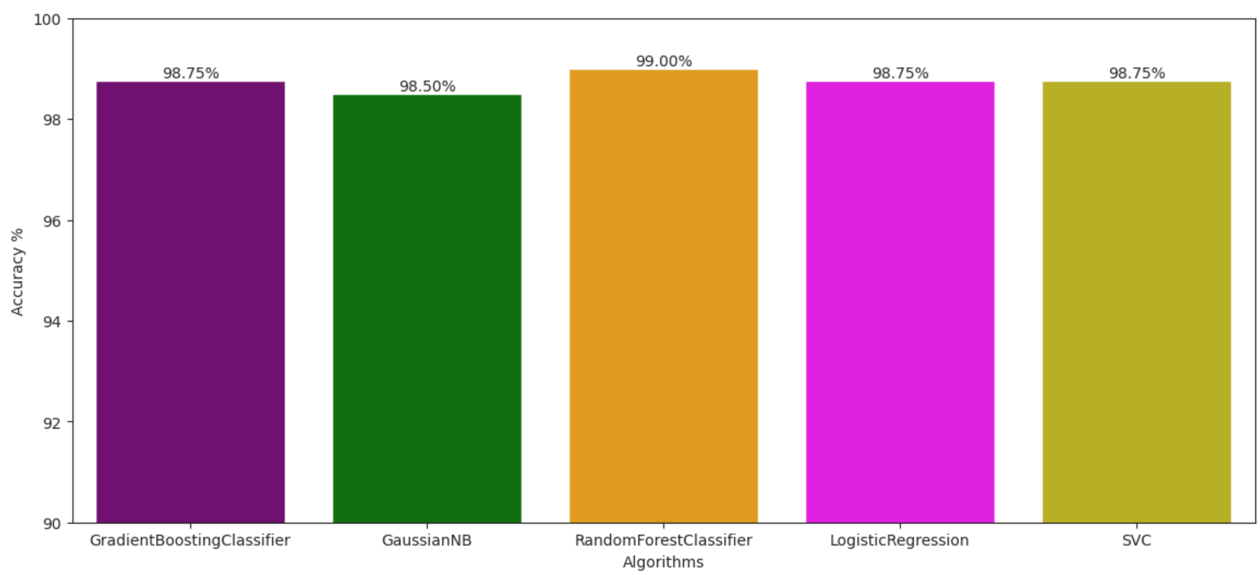

Figure 4.6: Comparative Model Accuracy Bar Plot

Figure 4.6 shows based on healthcare data, the performance of several machine learning algorithms, including GBC, GNB, RF, LR, and SVC, was assessed for early disease detection. Among these algorithms, RF had the highest accuracy of 99.00%, followed by GBC and LR, both with 98.75%. GaussianNB also performed well, with an accuracy of 98.50%, while SVC had an accuracy of 98.75%. The results show that ensemble methods like RF and GBC, as well as traditional classifiers like LR and GaussianNB, can accurately predict disease onset. Further analysis of feature importance and model predictions sheds light on the discriminatory power of various symptoms, as well as the robustness of predictive models for real-world use in healthcare.

4.3 Discussion

RF and GBC show great accuracy, showing the accuracy of ensemble learning techniques in early disease diagnosis. LR and GaussianNB also performed well, highlighting the value of using linear and probabilistic techniques in predictive modelling. While SVC showed comparable performance, its computational complexity needs further research into optimization methodologies. Overall, the findings illustrate the viability of applying machine learning algorithms to detect illness patterns in healthcare data. However, interpretability remains an issue, highlighting the importance of useful model explanations for improving clinical acceptance. Future research could look into hybrid approaches and ensemble methods to improve model interpretability and solve the difficulties of disease prediction in varied patient populations. Additionally, efforts to incorporate domain knowledge and expert views may improve model performance and facilitate real-world use in clinical settings.

CHAPTER 5

5.1 Impact on Society

The effect of machine learning for early disease identification extends beyond clinical settings, with major societal benefits. Diabetes, Malaria, Dengue, Typhoid, and Hepatitis B can be detected early, allowing for more effective treatment and lower healthcare expenses. By supporting early treatment and management methods, these prediction models help to reduce illness burden and improve public health outcomes. Furthermore, the use of such technologies promotes a move toward proactive healthcare practices that focus prevention over therapy alone. Enhanced disease surveillance and monitoring skills allow healthcare systems to respond more effectively to outbreaks and epidemics, thereby protecting community health on a larger scale. Finally, the societal benefit of machine learning in early disease identification extends outside individual healthcare encounters, resulting in a healthier and more resilient population.

5.2 Impact on Environment

The effect of using machine learning for early disease diagnosis extends to the environment by many indirect channels. First, by facilitating early intervention and treatment, these predictive models help to reduce the burden on healthcare resources and infrastructure, which means lower energy consumption and waste generation connected with healthcare delivery. In addition, improved public health outcomes from early disease identification translate into lower environmental pollution, as fewer treatments and hospitalizations result in lower emissions from healthcare facilities. Furthermore, the transition to proactive healthcare techniques supports sustainability by emphasizing prevention over reactive treatment, which reduces healthcare systems' environmental footprint. In addition, using machine learning algorithms to optimize healthcare workflows and resource allocation helps streamline treatments and reduce inefficiencies, resulting in lower energy consumption and resource utilization throughout the healthcare sector. Overall, using machine learning into early disease detection helps to promote a more efficient and robust healthcare system, which adds to environmental sustainability.

5.3 Ethical Aspects

Moral issues with the use of machine learning for early disease identification are critical. To begin, protecting patient privacy and confidentiality is critical, which demands strong data protection procedures and following legal standards. In addition, honest communication and informed permission are critical to sustaining trust between patients and healthcare professionals, particularly in the gathering and use of personal health data. In addition, tackling algorithmic bias and fairness is critical to preventing disparities in disease identification and treatment among different demographic groups. Furthermore, healthcare personnel must keep to ethical principles when developing, understanding, and making decisions about models, with a focus on patient well-being and autonomy. Efforts to improve transparency, accountability, and inclusivity in predictive model development and deployment are critical for keeping ethical values and promoting equal access to early illness detection technology

5.4 Sustainability Plan

The sustainability plan for promoting machine learning in early illness detection includes several key techniques. To begin, continued investment in research and development is critical for continuously improving predictive models and adjusting to changing healthcare requirements and difficulties. Second, increasing collaboration across academia, industry, and healthcare institutions encourages information sharing and innovation, resulting in long-term advancements in predictive modelling methodologies. Third, investment in workforce training and education provides a competent worker pool that can effectively use machine learning technology in healthcare settings. In addition, prioritizing interoperability and standardization of healthcare data systems allows for the seamless integration of predictive models into existing healthcare workflows, which improves sustainability and capacity for growth. Furthermore, creating strong data governance frameworks and ethical principles promotes trust and accountability while limiting potential dangers and making sure the proper application of predictive analytics in healthcare. Finally, long-term relationships and financing methods allow ongoing support for research, development, and implementation activities, which promotes sustainability and resilience in early illness detection initiatives.

CHAPTER 6

6.1 Summary of the Study

In conclusion, our study focuses on the creation and testing of machine learning models for early detection of diseases using healthcare data. The study studied different algorithms, including, GNB, GBC, RF, LR and evaluated their ability for identifying the beginning of diseases such as Diabetes, Malaria, Dengue, Typhoid, and Hepatitis B. The results showed high accuracy across many methods, with RF Classifier having the best accuracy of 99.00%. Furthermore, feature importance analysis revealed the most selective symptoms for disease prediction. The study added the social benefits of early disease identification in terms of better patient outcomes and lower healthcare expenses. Ethical considerations, sustainability strategies, and future research implications were highlighted to help guide the ethical deployment and continual improvement of predictive models in healthcare environments. Overall, this study advances the field of early disease identification and highlights the power of machine learning in changing healthcare delivery.

6.2 Conclusions

Finally, this study shows that machine learning algorithms are useful in detecting early diseases using healthcare data. The examination of various methods indicated remarkable accuracy rates, with RF Classifier attaining a fantastic 99.00%. The feature importance analysis gave useful insights into the major symptoms which impact disease outlook. Early disease detection had a major societal benefit, with improved patient outcomes and lower healthcare costs. Ethical considerations were discussed, emphasizing the value of patient privacy, honesty, and justice in predictive modelling methods. In addition, a sustainability plan was offered to assure ongoing advancement and responsible deployment of predictive models in healthcare environments. Future research areas involve fixing algorithmic bias, improving model interpretability, and using domain knowledge to provide more robust illness detection solutions. Overall, this study advances the field of early disease identification and highlights the transformational power of machine learning in healthcare.

6.3 Implication for Further Study

Future research should look into hybrid approaches that mix machine learning with other predictive modeling techniques to improve accuracy and interpretability. In addition, exploring the integration of genetic and molecular data into prediction models could provide deeper insights into disease pathophysiology while also improving predictive performance. Addressing algorithmic bias and fairness concerns is an important focus of future research to ensure equitable healthcare results for varied patient populations. Furthermore, longitudinal studies are needed to determine the long-term efficacy and viability of prediction models in real-world healthcare settings. Partnerships between data scientists, medical experts, and domain specialists is critical for addressing complex healthcare concerns and driving innovation in early disease detection. Finally, adding predictive models into clinical decision support systems may speed up healthcare procedures and increase diagnostic accuracy, demanding additional exploration.

REFERENCES

[1] Paul, Anthony, and Kelvin Lawrence. "Improving Healthcare Analytics: A Multiclass Supervised Learning Method for Classifying Diseases." 2024.

[2] Badawy, M., Ramadan, N., & Hefny, H.A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. Journal of Electrical Systems and Information Technology, 10(1), 40.

[3] Cena, J., & Oluwaseyi, J. (2024). Investigating the Implementation of Machine Learning for Heart. Statistics, 14(1), 150-155.

[4] M. Bari Antor et al., "A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease," Journal of Healthcare Engineering, vol. 2021, p. e9917919, Jul. 2021

[5]]H. Deng, Y. Zhou, L. Wang, and C. Zhang, "Ensemble learning for the early prediction of neonatal jaundice with genetic features," BMC Medical Informatics and Decision Making, vol. 21, no. 1,

[6] JavaTPoint, "Support Vector Machine (SVM) Algorithm - Javatpoint,"

[7] Alanazi, R. (2022). Identification and prediction of chronic diseases using machine learning approach. Journal of Healthcare Engineering, 2022.

[8] Raizada, Rajeev DS, and Yune-Sang Lee. "Smoothness without smoothing: why Gaussian naive Bayes is not naive for multi-subject searchlight studies." PloS one 8.7 (2013): e69566.

[9] Hsu, C.H., Chen, X., Lin, W., Jiang, C., Zhang, Y., Hao, Z., & Chung, Y.C. (2021). Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning. Measurement, 175, 109145.

[10] Arumugam, K., Naved, M., Shinde, P.P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2023). Multiple disease prediction using machine learning algorithms. Materials Today: Proceedings, 80, 3682-3685.

[11] Venkatesh, R., Balasubramanian, C., & Kaliappan, M. (2019). Development of big data predictive analytics model for disease prediction using machine learning technique. Journal of Medical Systems, 43, 1-8.

[12] Islam, M.S., Liu, D., Wang, K., Zhou, P., Yu, L., & Wu, D. (2019, June). A case study of healthcare platform using big data analytics and machine learning. In Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference (pp. 139-146)

ORIGINALITY REPORT

18%	15%	8%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Submitted to Daffodil International University Student Paper	1%
4	Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, Shanay Rab. "Significance of machine learning in healthcare: Features, pillars and applications", International Journal of Intelligent Networks, 2022 Publication	1%
5	ijrpr.com Internet Source	1%
6	www.researchgate.net Internet Source	1%
7	en.zeugmakongresi.org Internet Source	1%
8	Submitted to University at Buffalo Student Paper	<1%
9	Submitted to University of Leeds Student Paper	<1%
10	www.ncbi.nlm.nih.gov Internet Source	<1%