A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR BREAST CANCER PREDICTION AND DETECTION

BY

Mohamad Showrov Hassan Shajib
ID: 191-15-2754

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised By

Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Md Assaduzzaman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2024

APPROVAL

This Project titled **"A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR BREAST CANCER PREDICTION AND DETECTION"**, submitted by **Mohamad Showrov Hassan Shajib** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13 July 2024

BOARD OF EXAMINERS

Hossain 13-07-2024

Dr. Md. Fokhray Hossain (MFH) **Chairman**
Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Shah Md. Tanvir Siddiquee (SMTS) **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

13-07-2024

Mr. Md Umaid Hasan (MUH) **Internal Examiner**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

13/7/24

Dr. Shamim H Ripon (DSR) **External Examiner**
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mr. Saiful Islam, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

30/06/24

Mr. Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

30.6.24

Mr. Md Assaduzzaman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

01.7.24

(Mohamad Showrov Hassan Shajib)
ID: 191-15-2754
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Mr. Saiful Islam**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining, Image Processing, Algorithm, Optimization*" to carry out this project. His endless patience, scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor & Head of the Department of CSE**, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Many efforts have been devoted to developing efficient methods for early detection and prediction in order to address this urgent problem. My suggested strategy makes use of risk prediction algorithms and encourages recent breast cancer awareness. My strategy offers a simple way to forecast breast cancer and stands out for its practical usefulness in real-world circumstances. Using dataset provided on the UCI platform, I included a variety of classifiers, such as “K-Nearest Classifier (KNN), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression (LR)”. The findings were encouraging. Moreover, Boosting Decision Tree showed accuracy that met this exacting requirement. I applied a variety of techniques, including as Bagging, Boosting, and Voting, to further improve performance. In addition to adding to the corpus of information on breast cancer detection and prediction, my experimental analysis revealed the most accurate model to be the Logistic Regression model, which achieved an exceptional accuracy rate of 97.37% for breast cancer. The main aim of this thesis patient outcomes enhances by presenting a viable early intervention approach and important new insights into the management.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Introduction

Cancer is simply the uncontrolled proliferation of damaged tissues, threat to individual at large. Detecting this type of cancer, which arises from the unregulated growth of breast tissue, is a pressing concern, as its incidence is rapidly on the rise. The primary challenge in addressing this issue is the timely and accurate identification of affected areas during diagnosis. Artificial intelligence (AI) plays a pivotal role in this context, as it leverages private medical information to evaluate various features and patient diagnostic data to detect the presence of cancer in women. In my research, I meticulously examined data and identified key components that prove instrumental in disease determination. The dataset I utilized contained comprehensive information about an individual's body size, composition, and breast cancer status. Many researchers have collaborated on the quest to identify malignant cells in the body using machine learning techniques. However, their methods often lacked practicality and simplicity when it came to predicting breast cancer. My action for improvement of accurate cancer identification for women. In supervised learning, models are created based on labeled data, where the algorithm learns from in and out pairs. This multi-faceted approach aims to improve the diagnosis of breast cancer, ensuring that individuals can receive timely and accurate medical attention.

To improve patient outcomes and survival rates, breast cancer must be detected early and accurately predicted. An early diagnosis enables quick action and specialized treatment strategies, which can greatly boost the likelihood of a full recovery. As a result, creating accurate analysis and prediction techniques for breast cancer has grown to be a substantial area of study and therapeutic application. The complexity of breast cancer has recently been better understood thanks to improvements in medical technology, including molecular profiling and enhanced imaging methods. These developments have made it easier to combine different data sources, including as demographic information about patients, clinical records, genetic markers, and imaging data, to create complex analytical models. Applying machine learning, data mining, and statistical tools to huge datasets in order to find patterns and trends for breast cancer investigation and prediction. Researchers and healthcare professionals can construct solid predictive models by utilizing these strategies to pull relevant data from a variety of sources. In order to improve patient care, these models can help with risk assessment, early detection, and tailored treatment planning. This paper will also cover potential future paths and opportunities for additional study, as well as the difficulties and restrictions related to breast cancer analysis and prediction.

Symptoms

Breast cancer can cause a range of symptoms, especially in the later stages. Over time, malignant cells can migrate different organs, including the lungs, liver, brain, and bones. Here are some of the common symptoms.

- The formation of an additional mass or enlargement of the breast or armpit area. It may have a different sensation than the surrounding breast tissue.
- Any noticeable change in breast size, shape, or appearance, such as swelling, distortion, or asymmetry between breasts, could be a potential sign.
- This might include redness, dimpling, puckering, or scaling of the breast skin. The skin might be compared to an orange's exterior.
- Changes in the nipple, like inversion, retraction, or a sudden change in position, as well as discharge, can be indicative of an underlying issue.
- Enlarged lymph nodes in these areas may indicate that breast cancer has expanded to surrounding lymph nodes as well.
- Any alteration in the texture of the breast tissue, such as the development of an area that feels thicker or firmer than the surrounding tissue, can be a potential symptom.
- Unusual changes in the appearance or prominence of veins on the breast surface, particularly when not related to breastfeeding or pregnancy, might be a symptom.

Risk Reduction

Lifestyle changes are critical in lowering the risk of breast cancer. One important factor is keeping a healthy weight. Research indicates that postmenopausal women with excess body weight are at a higher risk of developing breast cancer. This increased risk is attributed to the elevated levels of estrogen produced by adipose tissue. Adopting a healthy diet rich in fruits, vegetables, and nutritious meats while limiting junk foods can improve general well-being and perhaps lower the risk of breast cancer. Regular physical activity is also important since moderate exercise has been shown to minimize risk. Limiting or eliminating alcohol use can significantly lower the risk of breast cancer. Because even occasional drinking of alcohol has been related to an elevated vulnerability, restricting alcohol consumption can be a viable risk-reduction approach. Breastfeeding over an extended period of time may have a preventative effect on breast cancer in women who have the option and choose to do so. Additionally, some medical disorders increase the chance of developing breast cancer. Hormone replacement treatment and various types of birth control might affect risk levels. It is recommended for women who are concerned about breast cancer to discuss alternatives or reduce the duration of HRT and investigate non-hormonal birth control techniques. Individuals with a known abnormalities in their genes can benefit greatly from expert consultation and screening. These sessions give insights into individual risk levels as well as risk-reduction techniques such as increased screenings, monitoring, or preventative operations such as mastectomy. Environmental factors might also contribute to breast cancer risk. Minimizing exposure to environmental pollutants, toxins, and radiation, whenever possible, are among risk reduction strategies. Engaging in regular screenings and maintaining open communication with healthcare professionals for personalized risk assessments.

1.2 Motivation

Breast cancer is an increasingly common health concern that affects a significant portion of the female population, and its prevalence continues to rise. Various factors, including dietary habits and the use of cosmetic creams, contribute to the occurrence of breast cancer. As per the international study, in 2020, a total of 2,261,419 reported cases of individuals affected by breast cancer, and it resulted in 684,996 recorded deaths [1].

The study also identified several risk factors associated with breast cancer, such as alcohol consumption, higher birth weights, and women who reached menarche at an early age. Numerous research efforts have been made to predict and understand breast cancer better. However, many of these studies have struggled to achieve better accuracy in their predictions. In response to this challenge, I embarked on rigorous research and developed a method that demonstrates the best accurate in forecasting the presence of breast cancer in both patients under regular medical observation and those under suspicion of having the disease. My innovative approach aims to provide a more reliable and precise means of early detection and diagnosis for breast cancer, thereby contributing to improved healthcare outcomes for individuals at risk of this disease.

1.3 Rationale of the Study

Develop a comprehensive methodology for predicting breast cancer, considering the increasing impact of this disease on society and identifying the essential treatment resources needed to combat it. In resource-constrained environments, the cost associated with cancer diagnosis and the analysis of symptoms can be prohibitively high, presenting a significant barrier to timely and effective treatment.

As researchers, our primary objective is to tackle this issue by leveraging advanced technologies and innovative methodologies. By doing so, we aim to make the prediction of breast cancer more accessible and economically feasible, particularly in underserved regions. Through our research, we strive to enhance healthcare services and facilitate the early detection of breast cancer, ultimately contributing to improved patient outcomes and better management of healthcare resources in areas with limited access to medical infrastructure.

Our approach incorporates advanced machine learning algorithms and data analysis techniques to analyze vast amounts of medical data with high accuracy. This will enable healthcare providers to identify potential breast cancer cases at an earlier stage, thereby improving the chances of successful treatment. Our goal is not only to advance scientific understanding but also to deliver tangible benefits to communities that are most in need of effective and affordable healthcare solutions.

1.4 Expected Output

It is increasingly spreading the general population, and its presence often remains uncertain. I propose an effective method for predicting and detecting this condition by examining diagnostic reports. my approach is capable of identifying breast cancer patients, improving decision-making, and providing precise assessments of its impact. Furthermore, it can measure the quality of life and address associated issues, contributing to increased awareness of breast cancer. The suggested model enables quick and accurate assessment of the disease.

1.5 Report Layout

First of all, introduce the thesis topic then next chapter start with review previous papers. After then, chapter 3 thesis methodology and proposed system given then result performance analysis.

Chapter 1: INTRODUCTION

Chapter 2: BACKGROUND

Chapter 3: RESEARCH METHODOLOGY

Chapter 4: RESULT & DISCUSSION

Chapter 5: CONCLUSION

CHAPTER 2

Background

2.1 Introduction

To pinpoint specific patterns associated with breast cancer, a variety of advanced techniques were employed. This research delves into studies related to the analysis of diagnostic reports, utilizing a broad spectrum of algorithms. A range of machine learning models were implemented during the exploratory phase of this research, drawing from a wealth of prior studies where diverse models had been utilized by various researchers.

The methodology includes a thorough examination of historical diagnosis reports to identify common indicators and anomalies that could signify the presence of breast cancer. The integration of these models into the diagnostic process is expected to significantly reduce the time and cost associated with traditional diagnostic techniques, making breast cancer screening more accessible to populations in resource-limited settings.

Furthermore, this research emphasizes the importance of continuous learning and adaptation. As new data becomes available, the models can be retrained and refined, ensuring that they remain current with the latest medical knowledge and technological advancements. The ultimate aim is to create a robust, scalable, and cost-effective solution that can be deployed widely, improving early detection rates and patient outcomes globally.

2.2 Related Work

The literature reveals a growing interest in leveraging advanced algorithms for accurate and early detection. Various studies have explored different ML models to enhance breast cancer diagnosis, emphasizing the importance of improving prognosis through early identification. Amrane et al. focused on “Naive Bayes (NB) and k-nearest neighbor (KNN)” classifiers, comparing their actions in breast cancer classification.

Their findings demonstrated that KNN outperformed NB, achieving a higher accuracy of 97.51%. This study underscores the effectiveness of ML algorithms for systematic and objective prognostic purposes [1].

Abdulla et al. delved into the utilization of “Support Vector Machine (SVM), K-Nearest Neighbour (K-NN), and Random Forest (RF)” in classifying breast tumors. They highlighted SVM's remarkable accuracy of 97%, showcasing its potential for rapid and precise diagnosis. The incorporation of AI, particularly ML algorithms, has proven beneficial in distinguishing between malignant and benign cases, contributing to improved detection methodologies [2].

Michael et al. show an optimized framework emphasizing the importance of early detection for improved survival rates. Their CAD system, utilizing machine learning classifiers, showcased the superiority of the LightGBM classifier, achieving an impressive 99.86% accuracy. This research contributes to addressing challenges associated with misclassification and false-positive rates in breast lesion analysis [3].

Murtaza et al. give a comprehensive review on cancer classification by medical imaging. Focusing on CNN and other deep neural network approaches, the review outlined the current landscape, emphasizing the significance of mammograms and histopathologic images for breast cancer classification [4].

Osareh and Shadgar explored the combination of SVM, K-NN, and probabilistic neural networks for breast cancer diagnosis. Their approach, incorporating feature ranking and selection techniques and this study reinforces the potential of ML, showcasing its efficacy in breast cancer diagnostics [5].

Yue reviewed ML techniques' applications, emphasizing the pivotal role of early detection. The overview included popular algorithms such as “artificial neural networks (ANNs), SVMs, decision trees, and k-NNs”. The study drew insights from the Wisconsin Breast Cancer Database (WBCD), showcasing ML's versatility in pattern classification and forecast modeling [6].

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks [7].

Pathan, S., Savarimuthu, T. R., & Al-Zahrani, A. (2019). Application of ML, datasets, and performance evaluation methods used in existing studies, offering a comprehensive understanding of the field [8].

Focusing on ensemble-based machine learning approaches, this paper explores their efficacy in breast cancer diagnosis and prognosis. The study highlights the advantages of combining multiple classifiers to enhance prediction accuracy and offers insights into feature selection and model optimization techniques [9].

The study demonstrates improved prediction accuracy compared to individual algorithms, emphasizing the potential of hybrid models in healthcare applications [10].

Examining the use of hybrid intelligent techniques, this paper explores the combination of genetic algorithms and support vector machines for breast cancer diagnosis. The study presents promising results and suggests the importance of feature selection and optimization algorithms in improving prediction accuracy [11].

This study focuses on the utilization of polygenic risk scores derived from genetic data for breast cancer prediction. While not directly related to machine learning, the paper highlights the importance of incorporating genetic factors into prediction models and presents a potential avenue for enhancing machine learning-based breast cancer prediction [12].

Although centered on lung cancer prognosis prediction, this paper showcases the significance of utilizing image analysis and pathology features in machine learning models. The research provides insights into the integration of image-based features for breast cancer prediction using similar computational techniques [13].

2.3 Scope of the Problem

Consisting of 569 instances with 32 features, requires careful analysis to extract meaningful patterns. Understanding the complex relationships between these features, such as cell characteristics and patient data, is crucial.

1. Dataset Complexity: The dataset comprises 569 instances with 32 features, demanding a detailed examination to uncover underlying patterns.

2. Feature Importance: Identifying the most influential features, such as cell characteristics, is essential for accurate breast cancer classification.

3. Model Evaluation: The developed predictive model needs thorough evaluation, comparing its accuracy against existing models to ensure superiority.

4. Impact on Diagnosis: Successful implementation of an accurate predictive model can significantly impact breast cancer diagnosis, aiding in early detection.

5. Ethical Considerations: Given the sensitive nature of healthcare data, ethical implications of the model's predictions must be carefully examined.

The primary objective is to enhance breast cancer classification accuracy, contributing to more effective diagnostic tools and potentially improving patient outcomes. The analysis involves utilizing machine learning algorithms, conducting exploratory data analysis, and leveraging advanced techniques.

2.4 Challenges

Embarking on this breast cancer classification project presented multiple challenges. Firstly, delving into the intricate world of medical data demanded a comprehensive understanding of breast cancer, its types, and the underlying factors contributing to its diagnosis. The sheer volume of the dataset, consisting of 569 rows and 32 columns, posed a significant challenge in terms of data processing and analysis. Understanding the medical intricacies of breast anatomy and cancer progression was a formidable task, requiring research and comprehension beyond standard machine learning practices. Additionally, the ethical responsibility associated with healthcare data added an extra layer of complexity, emphasizing the need for precision and reliability in model predictions.

1. Navigating the complex medical terminology and concepts related to breast cancer.
2. Processing and analyzing a large dataset with 569 rows and 32 columns.
3. Ensuring the accuracy and reliability of the predictive model in a critical healthcare context.
4. Adapting machine learning techniques to the unique challenges posed by medical data.
5. Addressing the ethical considerations associated with healthcare data and ensuring patient privacy.

Conclusively, this project not only tested my technical skills in machine learning but also required a deep dive into the medical domain, highlighting the interdisciplinary nature of the challenge at hand.

CHAPTER 3
Research Methodology

3.1 Working Model

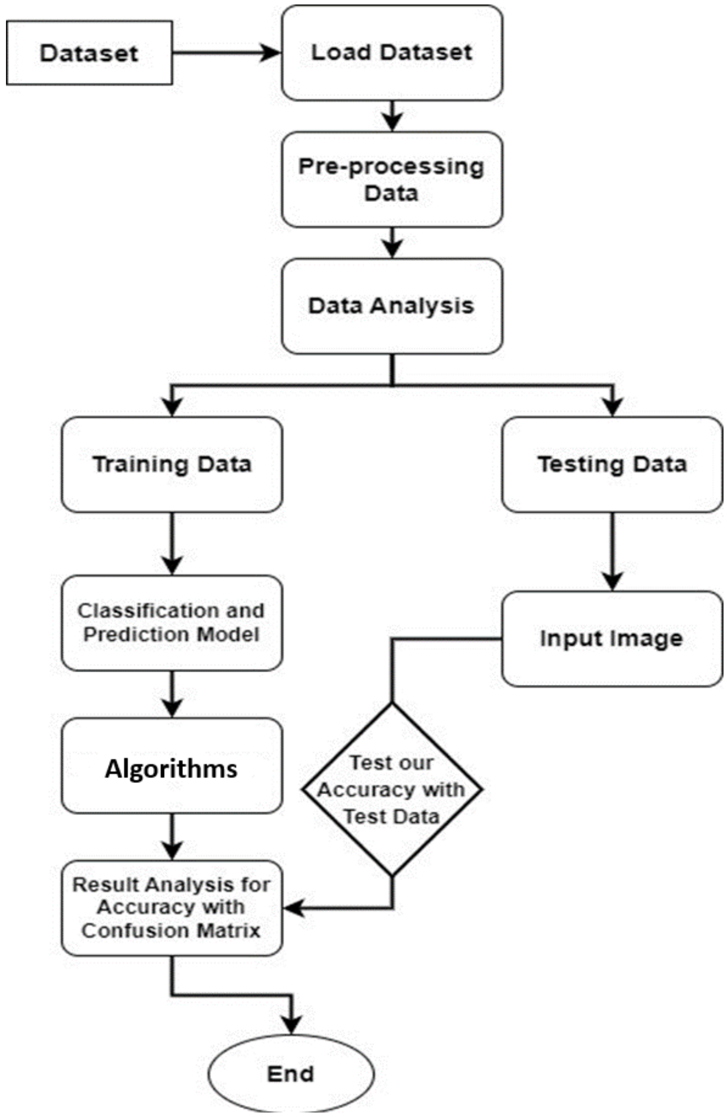

Fig 3.1.1: Flow Chart

The proposed methodology involves a systematic three-step approach. First, I will focus on Data Preprocessing, ensuring the dataset is cleaned and formatted for analysis. Following that, I'll delve into Model Development, where models like "Decision trees, logistic regression, Naive Bayes, KNN, SVM, and random forests" will be trained using the prepared data. Finally, in the Evaluation and Interpretation stage, I will assess the performance of these models, make comparisons, and interpret the findings.

3.2 Dataset Collection

Dataset from UCI platform, stands as a significant health concern for women globally. Originating from uncontrolled growth of cells in the breast, this condition manifests through detectable tumors, either visible on X-rays. The dataset at the core of this analysis is sourced from Kaggle, providing valuable insights into breast cancer diagnoses. As a collection of records, similar to a relational database table, the dataset encompasses critical information for each case. Beyond traditional strings and numbers, the dataset may contain diverse data structures such as lists, maps, and records, adding layers of complexity to the analysis. Comprehending its intricacies and ensuring necessary cleanup. Subsequently, I aim to construct classification models employing machine learning techniques. The exploration extends to fine-tuning hyperparameters and conducting a comparative evaluation of various classification algorithms. This undertaking not only sharpens my analytical skills but also contributes to the broader understanding of cancer detection [7].

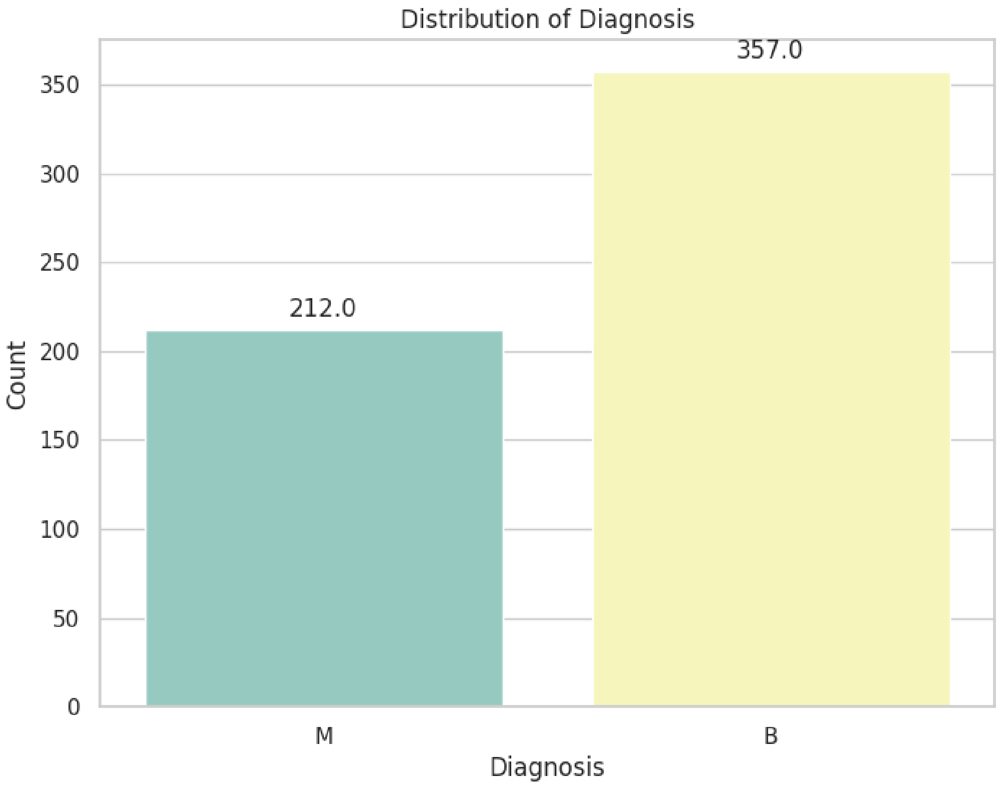

Fig 3.2.1: Diagnosis (M = malignant, B = benign)

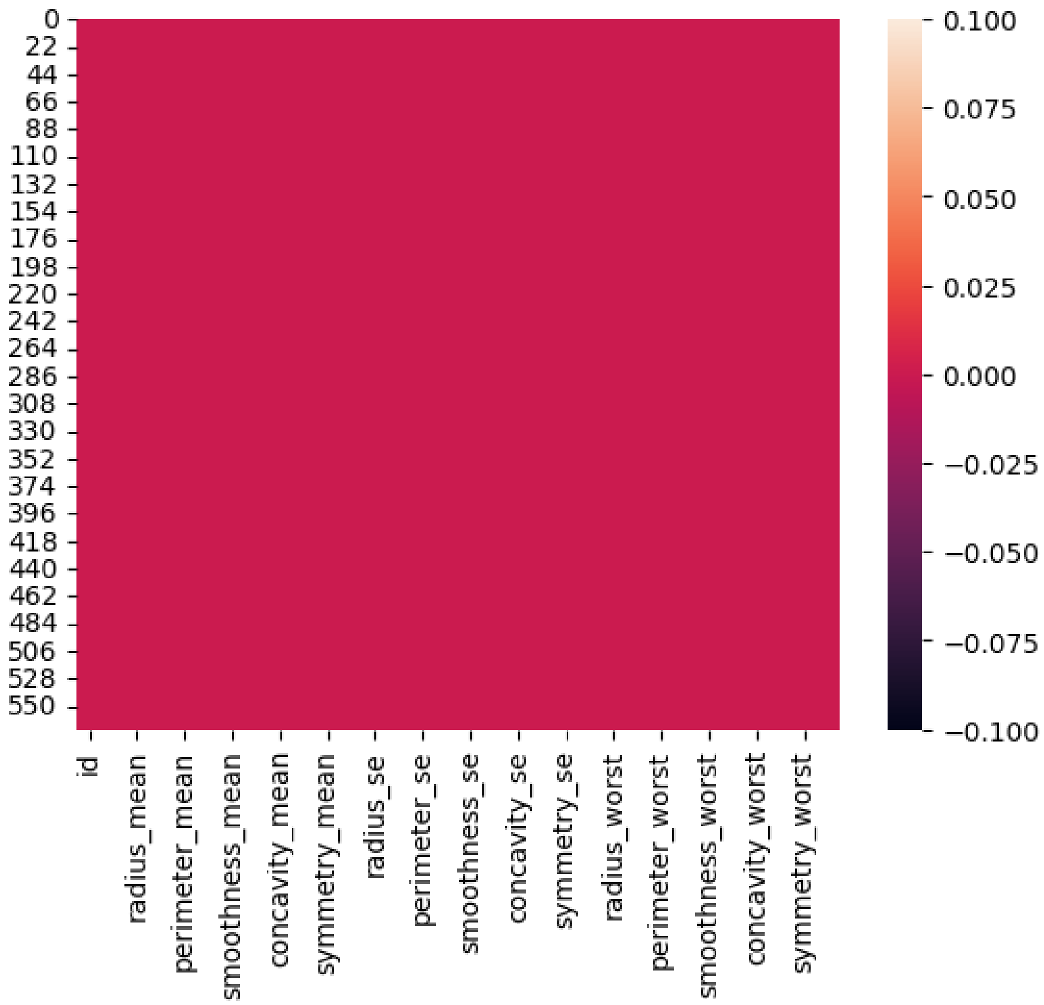

Fig 3.2.2: Dataset Value

3.3 Data Pre-processing

In the initial phase of my thesis work, I concentrated on data preprocessing, a fundamental step in preparing the dataset for comprehensive analysis. The primary objective was to transform the raw data into a clean and well-organized dataset, setting a solid foundation for accurate and meaningful outcomes in the subsequent stages of my research. This process involved meticulous checks for missing values, noisy data, and other inconsistencies that could adversely affect the accuracy of the algorithms.

I began by thoroughly examining the dataset to ensure it was free from any irregularities that could potentially skew the results. Addressing missing values was a top priority. I employed various techniques to handle these gaps, such as imputation methods to estimate missing values based on available data, or, in certain contexts, removing incomplete entries that could not be reliably filled.

In addition to managing missing values, I identified and corrected any noisy data points that could introduce errors into the analysis. This involved filtering out anomalies and ensuring that the data accurately represented the real-world scenarios it was meant to model. By standardizing the data and normalizing values, I aimed to eliminate any biases that might distort the results.

Through this meticulous data preprocessing phase, I ensured that the dataset was robust and reliable, ready for the application of advanced ML.

3.3.1 Feature Scaling

I focused on feature extraction, a crucial step in transforming raw data into numerical features for effective while retaining the essential information. Feature extraction is vital for enhancing the performance of machine learning models compared to directly applying them to raw data. Instead of using complex language, I opted for a simpler explanation to make the concept more accessible. Feature extraction involves converting the original data into numerical features, facilitating better analysis and prediction in breast cancer classification. This process is particularly important in the context of breast cancer diagnosis, where accurate classification is crucial for effective healthcare decisions. Overall, feature extraction serves as a bridge between raw data and machine learning algorithms, ensuring that the model can leverage meaningful information for improved classification outcomes in the study of breast cancer.

3.4 Label Encoding

The target variable in my dataset was the diagnosis, which was categorized into two groups: (i) benign = 0 and (ii) malignant = 1.

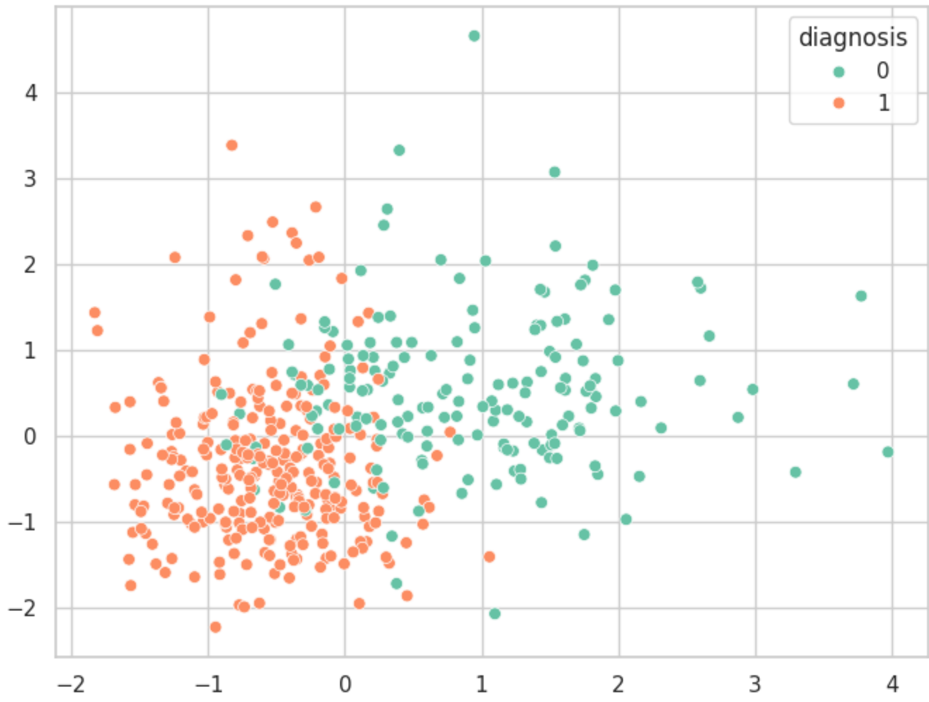

Fig 3.4.1: Diagnosis (1 = malignant, 0 = benign)

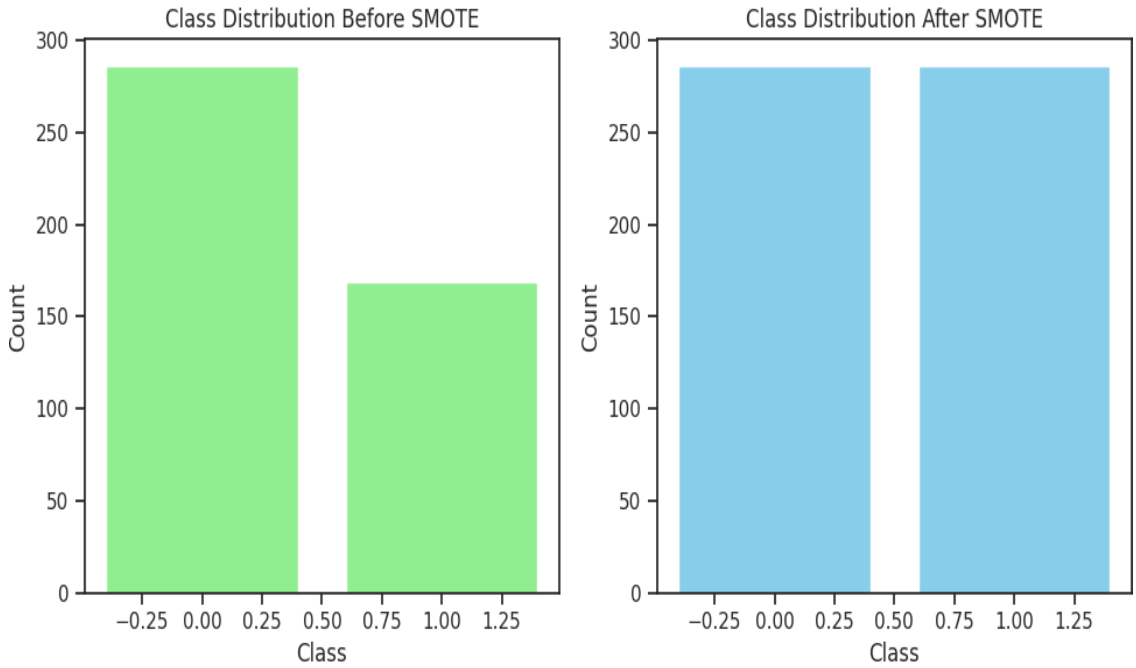

Fig 3.4.2: Diagnosis Class

3.5 Correlation

In my thesis on breast cancer classification using machine learning, I explored the relationships between various features in the dataset. The correlation matrix revealed the extent of connections among these features. Correlation coefficient indicated a strong positive relationship, 0 suggested a neutral relationship, and -1 pointed to a strong negative relationship. For instance, I observed a notable positive correlation between certain features, indicating they tend to increase or decrease together. On the other hand, some features showed a negative correlation, suggesting an inverse relationship. This matrix played a crucial role in understanding how different factors in the dataset are interrelated, providing valuable insights for building an effective predictive model [8][9].

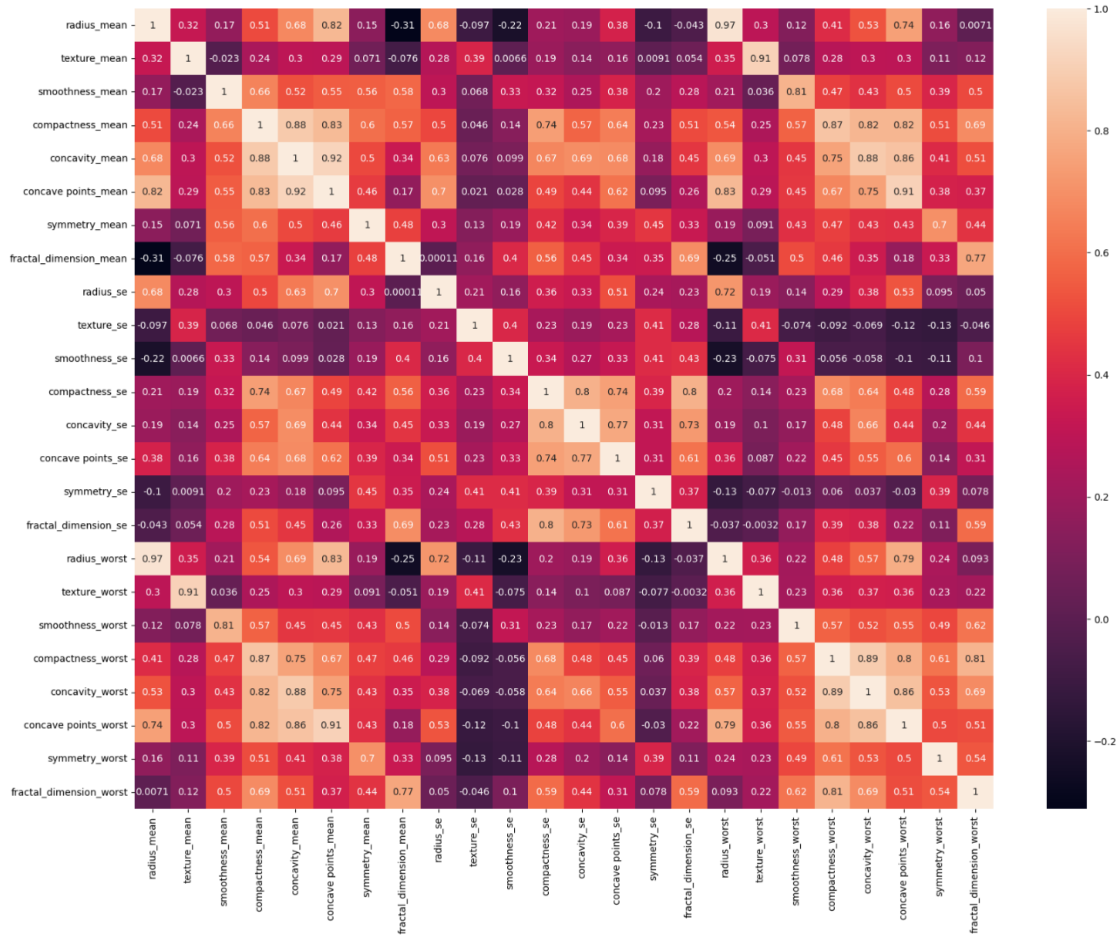

Fig 3.5.1: Correlation matrix

3.6 Training & Classification

Through supervised learning, the algorithm analyzes numerous instances, working to minimize loss and refine its predictive capabilities—a concept known as empirical risk minimization. Initially, the model is provided with a dataset containing features and corresponding labels, representing benign or malignant cases of breast cancer. The training phase involves adjusting the model's parameters iteratively to reduce the difference between predicted outcomes and actual labels. The goal is to achieve a well-tuned model that accurately classifies breast cancer types. By fine-tuning settings through methods, I strive to enhance the model's interpretability and reliability. This iterative process of learning from data forms the essence of model training in the context of breast cancer classification [11][12].

Essential Tools for Thesis

To facilitate the research and analysis in my thesis, a variety of essential tools and advanced libraries were utilized. These tools were chosen to ensure robust data processing, efficient algorithm implementation, and effective visualization.

Primary Tools:

Google Colab:

Description: Google Colab is an invaluable tool for data science and machine learning projects. It provides a cloud-based environment that allows for the execution of Python code, eliminating the need for local computing resources. Its support for GPU acceleration makes it particularly suitable for handling large datasets and complex computations.

Use in Thesis: In my thesis, Google Colab was used extensively for running machine learning algorithms, preprocessing data, and visualizing results. Its collaborative features also facilitated easy sharing and collaboration on the project.

Advanced Libraries:

Python:

Description: Python is a versatile, high-level programming language known for its simplicity, making it ideal for data analysis and machine learning.

Use in Thesis: Python was the primary language used for scripting, data manipulation, and implementing machine learning models throughout the thesis.

Pandas:

Description: Pandas is a powerful library for data manipulation, providing essential structures and functions to handle structured data.

Use in Thesis: Pandas was used for data cleaning, transformation, and analysis, ensuring the dataset was in optimal shape for machine learning.

Numpy:

Description: Numpy is a fundamental package for numerical computation in Python, supporting large, multi-dimensional arrays and matrices.

Use in Thesis: Numpy facilitated efficient numerical operations and mathematical functions, crucial for data preprocessing and algorithm implementation.

Seaborn:

Description: Seaborn is a statistical data visualization library built on top of Matplotlib, designed for creating informative and attractive graphics.

Use in Thesis: Seaborn was utilized to create detailed and aesthetically pleasing visualizations of the data, aiding in exploratory data analysis.

Matplotlib:

Description: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Use in Thesis: Matplotlib was used to generate various plots and charts to visualize the data's results and patterns.

Scikit-learn (Sklearn):

Description: Scikit-learn is a robust machine learning library offering efficient tools for data mining and analysis.

Use in Thesis: Sklearn was employed to implement various machine learning models, including linear regression, support vector machines, and decision trees.

Itertools:

Description: Itertools is a standard Python library providing functions for creating efficient looping iterators.

Use in Thesis: Itertools was used to facilitate iterative processes required in the analysis.

Plotly:

Description: Plotly is an interactive graphing library that creates interactive, publication-quality graphs.

Use in Thesis: Plotly was utilized to create interactive visualizations, allowing deeper exploration of the data and results.

XGBoost:

Description: XGBoost is an optimized gradient boosting library designed for high efficiency, flexibility, and portability.

Use in Thesis: XGBoost was implemented to enhance the performance and accuracy of predictive models.

Required Tools and Resources

Windows 10:

Description: The Windows 10 operating system was used as the primary platform for running various tools and libraries.

Use in Thesis: The development environment, including all necessary software and libraries, was set up on a Windows 10 machine.

Dataset from UCI Platform:

Description: The dataset for the thesis was sourced from the UCI Machine Learning Repository, a renowned repository for machine learning research.

Use in Thesis: This dataset was critical for training and testing the machine learning models.

Algorithm

Artificial Neural Networks (ANN):

Description: ANN is a computational model inspired by the structure and function of biological neural networks.

Linear Regression (LR): Used for predicting continuous outcomes.

Support Vector Machine (SVM): Employed for classification tasks.

Naive Bayes (NB): Applied for probabilistic classification.

Decision Tree: Used for both classification and regression tasks.

CHAPTER 4

Experimental & Result

4.1 Result and Discussions

In assessing the model's effectiveness for breast cancer classification, I employed various evaluation metrics to gauge its performance. The primary focus was on “accuracy, precision, recall, and F1-score”, considering their relevance in healthcare applications. Improved accuracy to 97.3% after applying Grid Search to fine-tune parameters. This enhancement signifies a more accurate prediction of breast cancer cases. Precision, particularly crucial in avoiding false alarms, reached 100%, indicating better identification of actual cancer instances. Additionally, recall remained consistent, affirming the model's reliability in identifying true cases. Implementing k-Fold Cross-Validation ensured robust performance across diverse datasets, crucial for real-world applications.

The models demonstrated accuracy improvements, particularly the Support Vector Machine after Grid Search optimization. This optimization led to a more precise prediction, reducing false positives and increasing overall reliability. The k-Fold Cross-Validation approach further validated the models' robustness, ensuring consistent performance across different datasets. Random Forest, SVM, and SVC exhibited reliable accuracy, providing a basis for selecting the most suitable model for breast cancer classification. The provided background information on globally provides context for the significance of this work. With breast cancer being a prevalent and potentially fatal disease, accurate and interpretable classification models contribute to early detection. In conclusion, the study successfully addressed the goal of improving breast cancer classification using machine learning. The combination of optimized models, cross validation, and analysis enhances our understanding and trust in the models, offering valuable insights for future research and application in real-world healthcare scenarios.

4.2 Result Analysis

I applied the trained machine learning model to predict cancer types by new data. The model takes features like cell characteristics as input and provides predictions. After divided the data into train data 80% and test data 20%, and apply six-machine learning classifier. The testing phase involved making predictions on new data and evaluating the model's performance, showcasing its potential for practical application in healthcare.

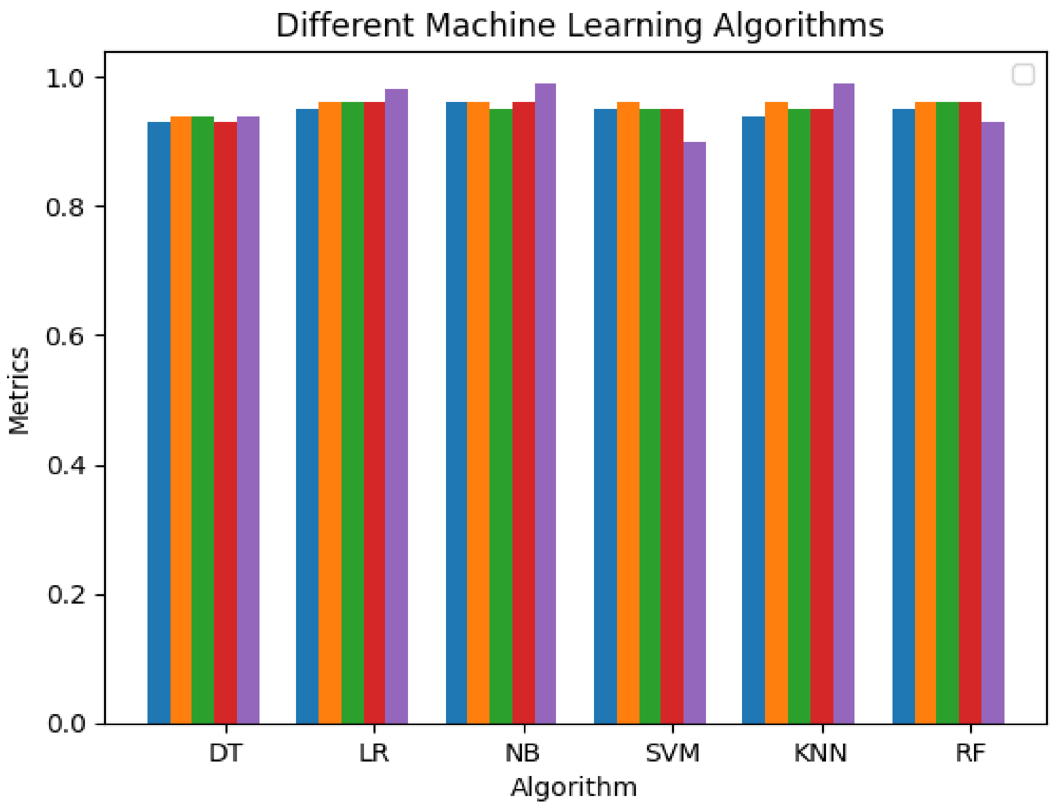

Fig 4.2.1: ML Algorithms

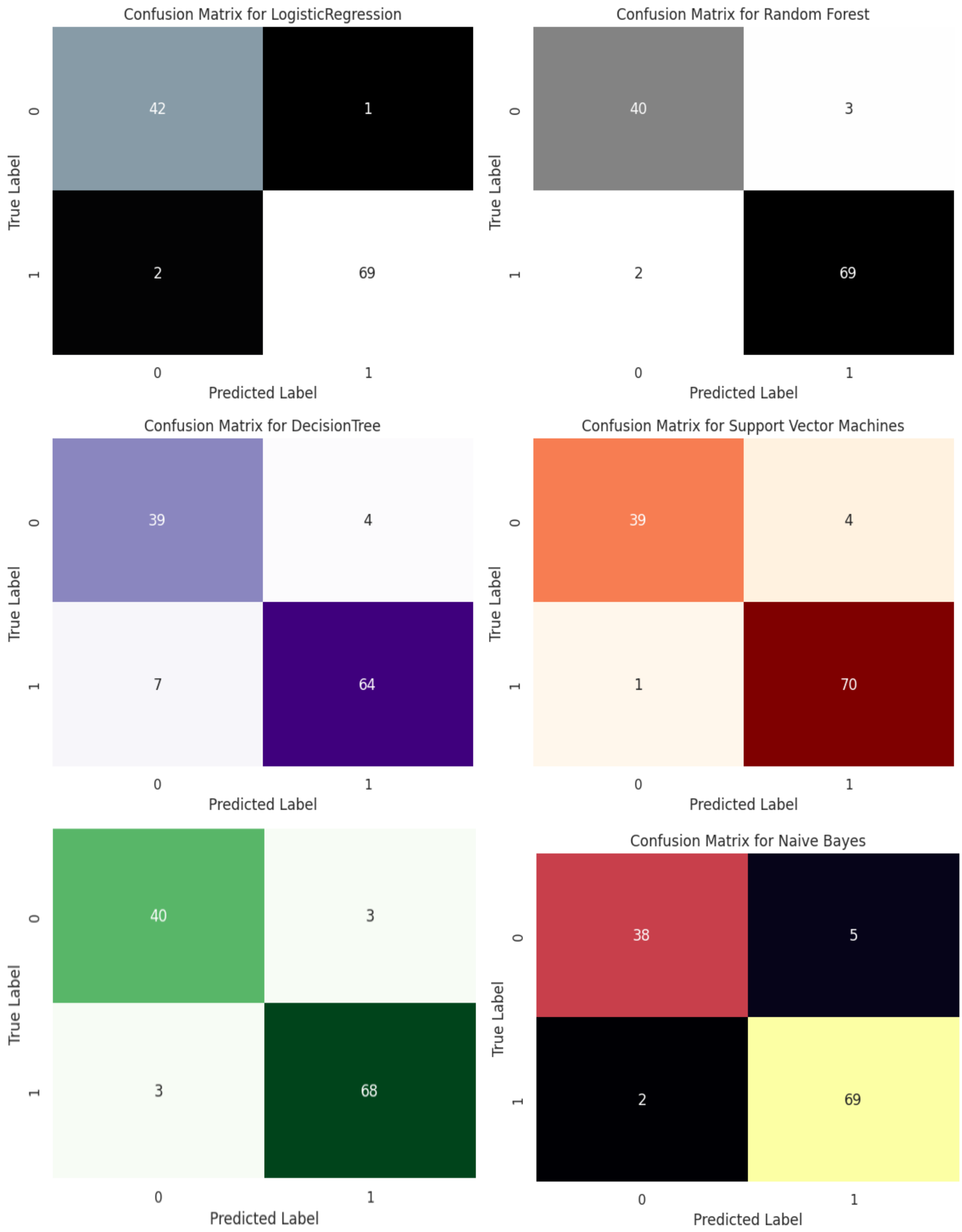

Fig 4.2.2: Algorithms Confusion Matrix

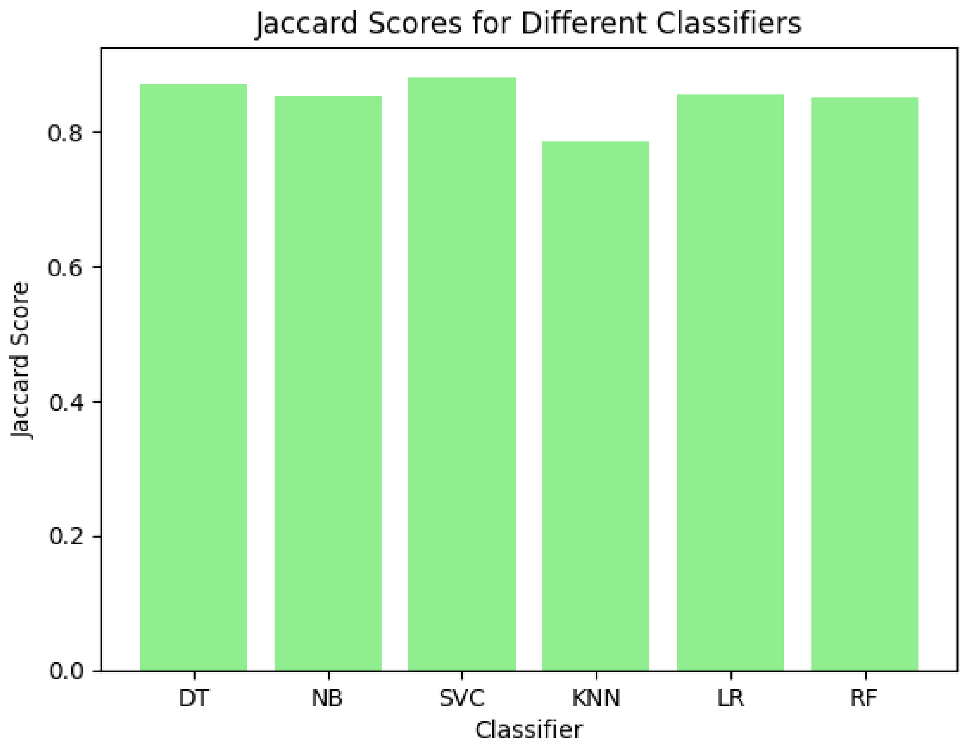

Fig 4.2.3: Jaccard Score

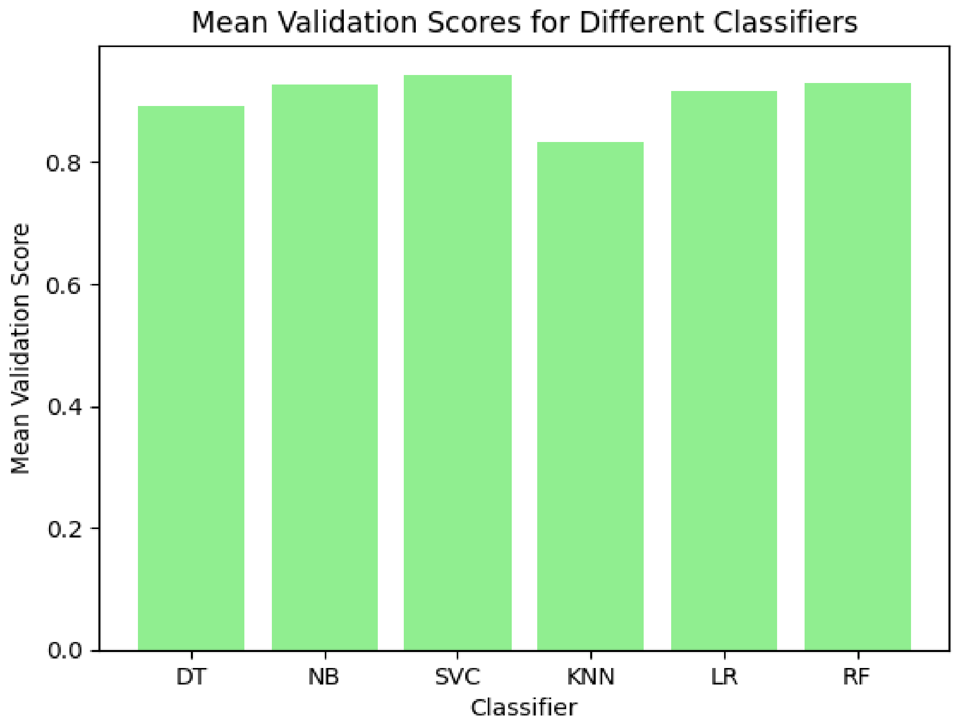

Fig 4.2.4: Mean Cross-Validation Score

4.3 Comparative Analysis

First, I looked at how well algorithmic classifiers performed. In Table 1, the visualization is displayed. My analysis of the algorithmic classifiers' performances revealed that LR had the highest accuracy, which was 97.37%. Using LR and DT, a precision score of 99.00% or better was reached. The four individuals with the highest recall scores were SVM, DT, KNN, and NB. The best F-1 score, which was achieved with LR and DT.

Table 1: Classifier Result with Dataset [16-19]

Algorithm	Accuracy %	Precision %	Recall %	F-1 Score %
LR	97.37	99.000	99.900	98.630
RF	95.54	98.019	99.900	97.290
DT	90.35	99.000	99.900	98.630
NB	93.86	97.029	99.900	96.000
SVM	95.61	97.029	99.990	96.000
KNN	94.74	97.029	99.900	96.000

CHAPTER 5

Conclusion

5.1 Summary of the Study

The Summary of Breast Cancer Prediction include determining who is impacted by the disease and assisting medical science in areas such as:

- Enhances the accuracy of breast cancer prediction, providing more reliable results. 2. Aids in early detection of breast cancer, leading to timely intervention and better treatment outcomes.
- Increases precision in diagnosing cancerous and benign tumors, reducing the likelihood of misdiagnosis.
- Empowers individuals and healthcare professionals with valuable insights for informed decision-making regarding treatment plans.
- Minimizes false alarms, ensuring a more efficient use of medical resources and preventing unnecessary stress for patients.
- Incorporates SHAP visualizations for a transparent understanding of how the model reaches its predictions, fostering trust and acceptance.

5.2 Conclusion

In conclusion, breast cancer remains a significant health concern, impacting millions globally. The journey through understanding its complexities involved exploring data-driven models for classification. I contributed to the ongoing efforts in breast cancer prediction. The application of k-Fold Cross-Validation enhanced the models, making them more accurate, reliable, and robust. It's crucial to acknowledge the broader context of breast cancer, as it surpasses lung cancer in global diagnoses. The statistics highlight the substantial impact on women, emphasizing the importance of continued research, awareness, and early detection. Survival rates have seen improvement over the years, thanks to advancements in detection and treatment, but challenges persist, particularly for Black women. This thesis journey has not only deepened my understanding of machine learning applications but also underscored the broader societal impact of breast cancer. As I conclude this exploration, the significance of collaborative efforts in research, healthcare, and public awareness becomes evident. Moving forward, the focus should extend beyond algorithms to encompass holistic approaches, ensuring that advancements benefit all individuals affected by breast cancer.

5.3 Implementation of Future Study

Exploring the scope of this thesis, advancements in breast cancer classification using machine learning present promising opportunities. As a student, I envision potential applications in improving early detection methods, enhancing diagnostic accuracy, and personalizing treatment plans for breast cancer patients. The integration of sophisticated machine learning models could lead to the development. The successful application of grid search techniques to optimize support vector machines (SVM) demonstrates the potential for fine-tuning models, further enhancing their performance. Additionally, the k-fold cross-validation approach contributes to the robustness of the models, ensuring their effectiveness across diverse datasets. Looking ahead, the future may involve implementing these models in clinical settings to support healthcare professionals in breast cancer diagnosis and treatment planning. Furthermore, the insights derived from the SHAP analysis can contribute to ongoing research, fostering continuous improvement in breast cancer prediction models. As a student, I recognize the evolving landscape of machine learning in healthcare and the significant role it can play in advancing breast cancer research and patient care.

5.4 Limitations

The limitations of this thesis include the constraints inherent in the dataset and analysis methods. Firstly, the dataset, consisting of 569 rows and 32 columns, may have limitations in representing the diverse population affected by breast cancer. The information provided in the dataset may not encompass all potential factors influencing breast cancer, and there could be additional variables contributing to the complexity of the disease that are not included. Secondly, the model development and evaluation are based on machine learning algorithms applied to the available data. While these algorithms can provide valuable insights, they may not capture the full complexity of breast cancer and may have limitations in predicting individual cases accurately. Moreover, the predictive performance of the developed model should be interpreted with caution, as it might not generalize well to new and unseen data. Additionally, the thesis acknowledges the absence of a detailed exploration of the socio-economic, environmental, or genetic factors that could play a role in breast cancer. The focus on machine learning models may overlook the broader context in which breast cancer occurs, limiting the comprehensive understanding of the disease.

Reference

[1] Schneider, A. P., Zainer, C. M., Kubat, C. K., Mullen, N. K., & Windisch, A. K. (2014). The breast cancer epidemic: 10 facts. The Linacre Quarterly, 81(3), 244–277

[2] Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, *1*, 1- 14.

[3] Yassin, N. I., Omran, S., El Houby, E. M., & Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine, 156*, 25-45.

[4] Bayrak, E. A., Kırcı, P., & Ensari, T. (2019, April). Comparison of machine learning methods for breast cancer diagnosis. In 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT) (pp. 1-3). Ieee.

[5] Sharma, S., Aggarwal, A., & Choudhury, T. (2018, December). Breast cancer detection using machine learning algorithms. In *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS)* (pp. 114-118). IEEE.

[6] Obaid, O. I., Mohammed, M. A., Ghani, M. K. A., Mostafa, A., & Taha, F. (2018). Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology, 7(4.36), 160-166.

[7] Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th international conference on electronic devices, systems and applications (ICEDSA) (pp. 1-4). IEEE.

[8] Gupta, M., & Gupta, B. (2018, February). A comparative study of breast cancer diagnosis using supervised machine learning techniques. In 2018 second international conference on computing methodologies and communication (ICCMC) (pp. 997-1002). IEEE.

[9] Osmanović, A., Halilović, S., Ilah, L. A., Fojnica, A., & Gromilić, Z. (2019). Machine learning techniques for classification of breast cancer. In World Congress on Medical Physics and Biomedical Engineering 2018: June 3-8, 2018, Prague.

[10] Chaurasia, V., & Pal, S. (2020). Applications of machine learning techniques to predict diagnostic breast cancer. SN Computer Science, 1(5), 270.

[11] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, *8*, 150360-150376.

[12] Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J.,& Clarke, M. F. (2003). Prospective identification of tumorigenic breast cancer cells. Proceedings of the National Academy of ciences, 100(7), 3983–3988.

[13] Sinn, H. P., & Kreipe, H. (2013). A brief overview of the WHO classification of breast tumors. Breast Care, 8(2), 149–154

[14] B.M.Gayathri, C.P.Sumathi, and Santhanam. Breast Cancer Diagnosis Using Machine Learning Algorithms –A Survey, International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3,
May 2013

[15] G. Williams, "Descriptive and Predictive Analytics", Data Min. with Ratt. R Art Excav. Data Knowl. Discov. Use R, pp. 193-203, 2011.

[16] Kavitha R, Kannan E. An efcient framework for heart disease classifcation using feature extraction and feature selection technique in datamining. in: IEEE Int. Conf. on Emerging Trends in Engineering Technology and Science (ICETETS), 2016, pp 1–5.

[17] Uysal AK, Gunal S, Ergin S. The impact of feature extraction and selection on SMS spam fltering.
Electronics and Electrical Engineering. 2013;19(5):67–72

[18] Dua D, Graf C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2019.

[19] ames G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. 1st ed. New York: Springer; 2013.

[20] L. F. Carvalho, G. Fernandes, M. V. O. De Assis, J. J. P. C. Rodrigues,and M. Lemes Proenc¸a, "Digital signature of network segment forhealthcare environments support," Irbm, vol. 35, no. 6, pp. 299-309,2014

[21] "Google Colaboratory," [Online]. Available: colab.research.google.com. [Accessed: 12-May-2020].

Plagiarism Report

Showrov Hassan Shajib (191-15-2754)

ORIGINALITY REPORT

13%	11%	2%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	Submitted to Daffodil International University Student Paper	3%
3	iq.opengenus.org Internet Source	<1%
4	Submitted to Florida Gulf Coast University Student Paper	<1%
5	www.adequateinfosoft.com Internet Source	<1%
6	Submitted to University of Florida Student Paper	<1%
7	Submitted to Kingston University Student Paper	<1%
8	Submitted to Liverpool John Moores University Student Paper	<1%
9	www.coursehero.com Internet Source	<1%

10	tel.archives-ouvertes.fr Internet Source	<1 %
11	essay.utwente.nl Internet Source	<1 %
12	infosectoday.com Internet Source	<1 %
13	raw.githubusercontent.com Internet Source	<1 %
14	Houda AIT BRAHIM, Salah EL-HADAJ, Abdelmoutalib METRANE. "Machine Learning Analysis of Breast Cancer Treatment Protocols and Cycle Counts: A Case Study at Mohammed VI Hospital, Morocco", Systems and Soft Computing, 2024 Publication	<1 %
15	academic.oup.com Internet Source	<1 %
16	www.ijisrt.com Internet Source	<1 %
17	Hajar Saoud, Abderrahim Ghadi, Mohamed Ghailani. "Proposed approach for breast cancer diagnosis using machine learning", Proceedings of the 4th International Conference on Smart City Applications - SCA '19, 2019 Publication	<1 %

18	Lawrence, Leonardo C.. "Machine Learning Integration in Healthcare: Human Error Reduction in Breast Cancer Screening", Colorado Technical University, 2024 Publication	<1%
19	netserver.it.nuigalway.ie Internet Source	<1%

Exclude quotes	On	Exclude assignment template	On
Exclude bibliography	On	Exclude matches	Off