A Project Report On

Analysis of Breast Cancer Detection using Ensemble Method

Submitted in partial fulfillment of the requirement for the 8^{th} semester **Bachelor of Engineering**

in

Computer Science and Engineering

DAYANANDA SAGAR COLLEGE OF ENGINEERING

(An Autonomous Institute affiliated to VTU, Belagavi, Approved by AICTE & ISO 9001:2008 Certified)

Accredited by National Assessment & Accreditation Council (NAAC) with 'A' grade

Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560078

Submitted By

Shalini Singh 1DS19CS146

Kota V Vishnu 1DS19CS723

Reena Jasmine Edwin 1DS19CS738

S Sai Brinda 1DS19CS741

Under the guidance of

Mrs Annapoorna B R

Asst Professor, CSE , DSCE

and

Ms Pooja

Co-Guide, Industry

2022 - 2023

Department of Computer Science and Engineering

DAYANANDA SAGAR COLLEGE OF ENGINEERING

Bangalore - 560078

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
Dayananda Sagar College of Engineering

(An Autonomous Institute affiliated to VTU, Belagavi, Approved by AICTE & ISO 9001:2008 Certified)

Accredited by National Assessment & Accreditation Council (NAAC) with 'A' grade

Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560078

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the project entitled **Analysis of Breast Cancer Detection using Ensemble Method** is a bonafide work carried out by **Shalini Singh [1DS19CS146]**, **Kota V Vishnu [1DS19CS723]**, **Reena Jasmine Edwin [1DS19CS738]** and **S Sai Brinda [1DS19CS741]** in partial fulfillment of 8th semester, Bachelor of Engineering in Computer Science and Engineering under Visvesvaraya Technological University, Belgaum during the year 2020-21.

Annapoorna BR	**Dr. Ramesh Babu D R**	**Dr. B G Prasad**
(Internal Guide)	Vice Principal & HOD	Principal
Asst Prof. CSE, DSCE	CSE, DSCE	DSCE
Signature:.....	Signature:.....	Signature:.....

Name of the Examiners:	Signature with date:
1.....	.....
2.....	.....

Acknowledgement

We are pleased to have successfully completed the project **Analysis of Breast Cancer Detection using Ensemble Method**. We thoroughly enjoyed the process of working on this project and gained a lot of knowledge doing so.

We would like to take this opportunity to express our gratitude to **Dr. B G Prasad**, Principal of DSCE, for permitting us to utilize all the necessary facilities of the institution.

We also thank our respected Vice Principal, HOD of Computer Science & Engineering, DSCE, Bangalore, **Dr. Ramesh Babu D R**, for his support and encouragement throughout the process.

We are immensely grateful to our respected and learned guide, **Mrs Annapoorna B R**, Professor CSE, DSCE and our co-guide **Pooja**, for their valuable help and guidance. We are indebted to them for their invaluable guidance throughout the process and their useful inputs at all stages of the process.

We also thank all the faculty and support staff of Department of Computer Science, DSCE. Without their support over the years, this work would not have been possible.

Lastly, we would like to express our deep appreciation towards our classmates and our family for providing us with constant moral support and encouragement. They have stood by us in the most difficult of times.

Shalini Singh 1DS19CS146

Kota V Vishnu 1DS19CS723

Reena Jasmine Edwin 1DS19CS738

S Sai Brinda 1DS19CS741

Analysis of Breast Cancer Detection using Ensemble Method

Kota V Vishnu, Reena Jasmine Edwin,Sai Brinda, Shalini Singh

June 10, 2023

Abstract

This study delves into the intricate analysis of breast cancer, employing four powerful machine learning algorithms: k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes, and Random Forest. To further enhance the predictive performance, an ensemble method harnessing XGBoost is utilized. The dataset comprises an array of clinical and histological features extracted from breast cancer patients. The team applies cutting-edge preprocessing techniques to address missing values, normalize features, and tackle class imbalance issues. The results reveal the sheer efficacy of KNN, SVM, Naive Bayes, and Random Forest algorithms in breast cancer analysis. The ensemble method, with its ability to amalgamate the predictions of multiple models, brings forth an outcome that is not only precise but also resilient. A feature importance analysis is conducted using the ensemble method, revealing the most significant features that play a vital role in breast cancer prediction. The findings are a testament to the rapid progress in machine learning research for breast cancer analysis and open up new avenues for further advancement in this crucial field.

Table of Contents

Abstract . . . . . **I**

Table of Contents . . . . . **II**

List of Figure . . . . . **IV**

List of Tables . . . . . **V**

I Introduction **1**

1 **1**

1.1 Current Situation 1

1.2 How ML Algorithms help us 3

1.3 Real World Solutions 5

II Problem Statement **8**

2 **8**

2.1 Questions which are addressed 9

2.2 Why Xgboost is used 11

III Literature Survey **13**

3 **13**

IV Architectural Design **17**

4 **17**

V Implementation **19**

5 **19**

5.1 Dataset 20

5.2 Data Preprocessing 20

5.2.1 Deletion of Unwanted Columns and Columns with Null Values 20

5.2.2 Encoding Categorical Data 20

5.2.3 Dataset Splitting 20

5.2.4 Feature Scaling 21
5.2.5 Hyperparameter Optimization: 21
5.3 Algorithms 22
5.3.1 KNN 22
5.3.2 SVM 23
5.3.3 Naive Bayes 24
5.3.4 Random Forest 24
5.3.5 XGboost Ensemble model 25
5.4 UML Diagram 26

VI Performance Metrics 29

6 29

6.1 Confusion Matrix : 29
6.1.1 Accuracy : 30
6.1.2 Precision : 30
6.1.3 Recall : 31

VII Result Analysis 32

7 32

7.1 Confusion Matrixes of Models: 33
7.2 Accuracy Comparision: 34

VIII Conclusion and Future Enhancement 35

8 35

List of Figures

1.1 Current Situation 1
1.2 Machine Learning 4
1.3 Usecase Diagram 5
4.1 Architectural Design 17
5.1 Flow Diagram 19
5.2 K-Nearest Neighbors (KNN) 22
5.3 Support Vector Machine 23
5.4 Naive Bayes 24
5.5 Random Forest 25
5.6 XGBoost 26
5.7 UML Diagram 27
7.1 Confusion Matrix of KNN 33
7.2 Confusion Matrix of SVM 33
7.3 Confusion Matrix of Naive Bayes 33
7.4 Confusion Matrix of Random Forest 33
7.5 Confusion Matrix of XGBoost Model 33
7.6 Bar Plot of Accuracy comparisions 34

List of Tables

6.1 Confusion Matrix Table 29
7.1 Performance Metrics Indices 32

List of Abbreviations

Abbreviation	Full Form
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
RF	Random Forest
NB	Naive Bayes
XGBoost	Extreme Gradient Boosting
ML	Machine Learning
WBCD	Wisconsin Breast Cancer Dataset
UML	Unified Modeling Language
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
HCRF	Hierarchical Clustering Random Forest
VIM	Variable Importance Measure
BCCD	Blood Cell Count and Detection
DT	Decision Tree
FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive

Chapter 1

Introduction

1.1 Current Situation

Breast cancer is a monstrous and life-threatening disease that plagues millions of women globally. Timely detection and precise diagnosis hold the key to successful treatment and improved patient outcomes. With technological advancements and massive datasets at our disposal, machine learning algorithms have emerged as formidable tools for breast cancer analysis.

This analysis endeavors to unravel the potential of diverse machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes, and Random Forest, for breast cancer analysis. Additionally, we will use the enigmatic XGBoost ensemble method to bolster the predictive performance and robustness of the models.

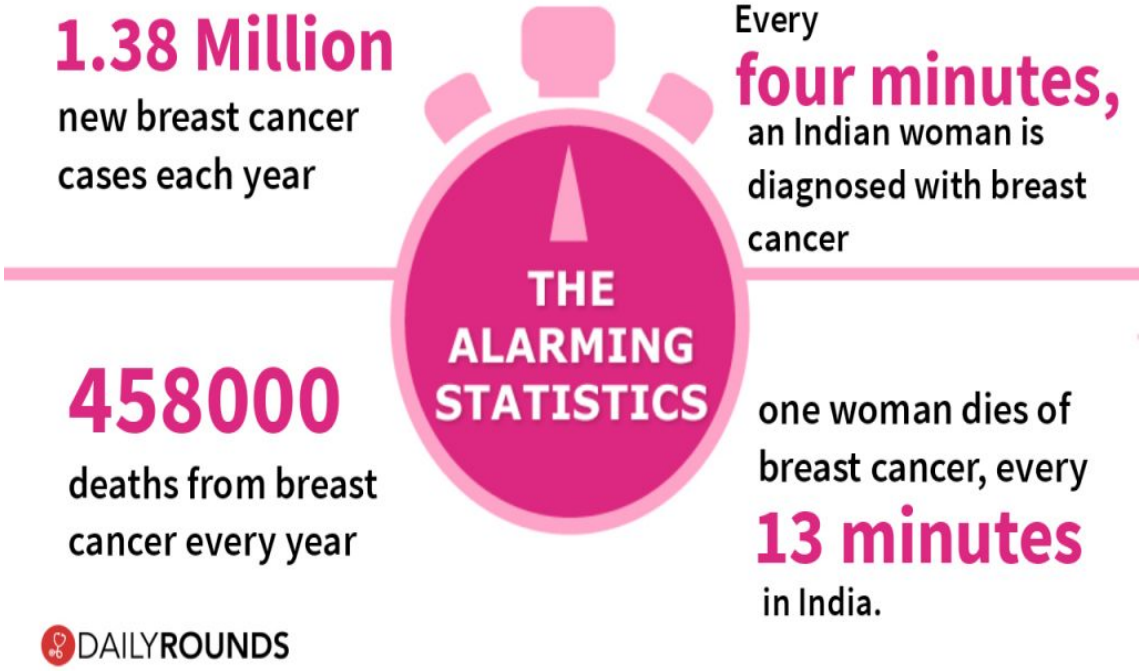

Figure 1.1. Current Situation

1. Breast Cancer and its Conundrums:

Breast cancer is a labyrinthine disease marked by the unbridled growth of malignant cells in breast tissue. Its complexity lies in its heterogeneity, varied clinical manifestations, and divergent treatment responses. Conventional diagnostic methods like mammography and biopsy have limitations in terms of accuracy, cost, and invasiveness. Machine learning algorithms have the potential to overcome these conundrums by harnessing the power of data-driven analysis.

2. Machine Learning in Breast Cancer Analysis:

Machine learning algorithms offer a promising avenue for breast cancer analysis by leveraging computational techniques to extract meaningful patterns from massive amounts of data. These algorithms learn from historical patient data, including clinical features, imaging results, genetic markers, and pathology reports, to build predictive models for early detection, risk assessment, and prognosis evaluation.

3. K-Nearest Neighbors (KNN):

KNN is a rudimentary yet potent algorithm that classifies instances based on their similarity to neighboring data points. It calculates the distance between the new instance and its k nearest neighbors, and assigns the class label based on majority voting. KNN is especially useful for breast cancer analysis due to its ability to handle multi-class classification and its ease of implementation.

4. Support Vector Machines (SVM): SVM is a robust supervised learning algorithm that separates data points into distinct classes by building an optimal hyperplane. Its goal is to maximize the margin between different classes, thereby boosting generalization and robustness. SVM has shown promising results in breast cancer diagnosis and prognosis prediction by efficiently handling high-dimensional data and capturing intricate relationships.

5. Naive Bayes: Naive Bayes is a probabilistic classifier that uses Bayes' theorem with an assumption of independence between features. Despite its simplistic assumption, Naive Bayes has displayed remarkable performance in various domains, including breast cancer analysis. It is computationally efficient, handles high-dimensional data, and can deal with missing values.

6. Random Forest: Random Forest is an enigmatic method that combines multiple decision trees to make predictions. It leverages the concept of bagging, where each tree is trained on a random subset of the data, and the final prediction is made by aggregating the outputs of individual trees. Random Forest is robust against overfitting, handles noisy data, and provides feature importance rankings, making it ideal for breast cancer analysis.

7. XGBoost : XGBoost is an optimized gradient boosting framework that amalgamates the strengths of boosting algorithms with scalable tree learning. It has gained popularity in machine learning competitions and has shown impressive performance in various domains. XGBoost's ensemble method combines multiple weak models iteratively to create a strong learner, capturing complex interactions and improving overall prediction accuracy.

8. Objectives of the Analysis: This analysis aims to: - Evaluate the performance of KNN, SVM, Naive Bayes, and Random Forest algorithms in breast cancer analysis.

- Harness the enigmatic XGBoost ensemble method to enhance the predictive accuracy and robustness of the models. - Compare the performance of individual algorithms with the ensemble approach.

- Provide insights into the strengths and limitations of each algorithm for breast cancer analysis.

The use of machine learning algorithms for breast cancer analysis holds immense potential in improving early detection, risk assessment, and treatment decisions. By exploring the potential of algorithms such as KNN, SVM, Naive Bayes, and Random Forest, in conjunction with the enigmatic XGBoost ensemble method, we aim to contribute to the ongoing efforts to advance breast cancer diagnosis and management.

1.2 How ML Algorithms help us

ML algorithms help combat breast cancer by analyzing medical images like mammograms to detect early signs of the disease, enabling prompt diagnosis and treatment. They assist in predicting treatment outcomes and prognosis based on patient-specific data, facilitating personalized treatment plans and improved patient care. Additionally, ML algorithms aid in identifying high-risk individuals, allowing for targeted interventions and preventive measures to reduce the incidence of breast cancer. This analysis aims to explore the effectiveness of several popular machine learning algorithms, namely K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes, and Random Forest, in diagnosing breast cancer. Additionally, an ensemble method using XGBoost will be utilized to harness the collective power of these algorithms, further enhancing the accuracy and reliability of the predictions.

Figure 1.2. Machine Learning

1.3 Real World Solutions

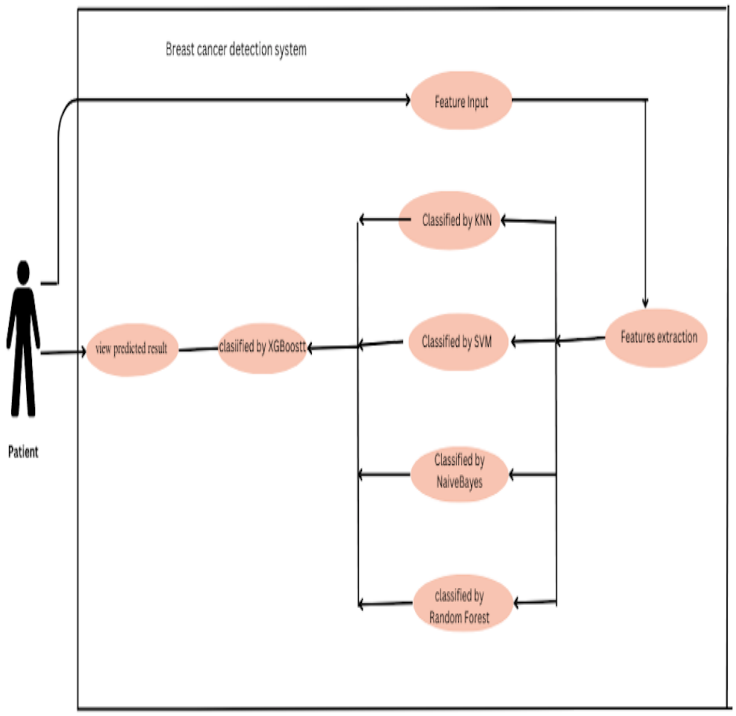

Figure 1.3. Usecase Diagram

Usecase Diagram : A use case diagram manifests a visual representation in the Unified Modeling Language (UML) that portrays the interactions among sundry actors (users, external systems, or entities) and the system being crafted. It presents a bird's-eye view of the functionality and behavior of a project or system, painting a picture of the different use cases or actions carried out by the actors.

In a use case diagram, actors are personified as stick figures, and the interactions between actors and the system are depicted using ovals entitled use cases. Arrows symbolize the communication or flow of actions between actors and use cases.

The fundamental purpose of a use case diagram is to capture and communicate the requirements and functionalities of a system in an informative and visual manner. It assists stakeholders, including project managers, developers, and users, to comprehend the broad scope of the project and the intended behavior of the system.

Here are the key components and their roles within a use case diagram:

1. Actors: Actors epitomize the diverse users or external systems that interact with the system being crafted. They can be individuals, roles, or other systems that trigger or partake in the use cases.

2. Use Cases: Use cases epitomize the specific functionalities or actions performed by the system. Each use case exemplifies a distinct interaction between an actor and the system, capturing a specific goal or task.

3. Relationships: Relationships or associations between actors and use cases are portrayed with arrows. They indicate the involvement or participation of actors in specific use cases. For example, an actor may initiate a use case or be involved in multiple use cases.

4. System Boundary: The system boundary is a rectangle or box that envelops the use cases, representing the scope or boundary of the system being crafted. It defines what is included within the system and what is external to it.

Use case diagrams facilitate a lucid understanding of the system's functionality and how different actors interact with it. They serve as a foundation for requirements analysis, system design, and development planning. Use cases assist in identifying and prioritizing key functionalities, defining user roles, and guiding the creation of detailed use case specifications.

Overall, use case diagrams are an efficacious communication tool, enabling stakeholders to grasp the essence of a project or system's requirements, user interactions, and desired functionality. They foster collaboration, enhance clarity, and provide a visual representation of the project's structure and objectives.

KNN, SVM, Naive Bayes, and Random Forest are well-established algorithms that have been widely applied in diverse fields, including healthcare. KNN is a non-parametric classification algorithm that classifies a data point based on the majority of its nearest neighbors. SVM, on the other hand, constructs a hyperplane that maximally separates the data points into different classes. Naive Bayes is a probabilistic classifier that applies Bayes' theorem with the assumption of independence between features. Lastly, Random Forest employs an ensemble of decision trees to make predictions based on the aggregation of individual tree results.

While these algorithms have shown promise in various medical applications, individually they may have limitations in terms of prediction accuracy or generalizability. To overcome these limitations, ensemble methods can be employed to combine multiple models and leverage their strengths. In this analysis, XGBoost, a popular ensemble technique, will be used to integrate the predictions of KNN, SVM, Naive Bayes, and Random Forest, resulting in a more robust and accurate breast cancer diagnosis model.

The Wisconsin dataset used in this analysis will consist of a collection of patient records, including

clinical features and diagnostic outcomes. By training the machine learning algorithms on this dataset, we can develop models capable of accurately predicting whether a patient is diagnosed with breast cancer or not.

The significance of this analysis lies in its potential to provide healthcare professionals with an efficient and reliable tool for diagnosing breast cancer. By harnessing the power of machine learning algorithms and the ensemble method, we can improve the accuracy and efficiency of breast cancer detection, leading to earlier interventions, more personalized treatments, and ultimately, better patient outcomes.

In the following sections, we will delve into the methodologies and results of the analysis, discussing the performance of each algorithm individually and the collective power of the ensemble method using XGBoost.

Chapter 2

Problem Statement

Breast cancer continues being a pressing health issue, affecting many women worldwide. To maximize treatment efficacy and generate positive patient outcomes, detecting breast cancer early on remains critical. Recently machine learning algorithms have offered promising results when detecting breast cancer effectively towards accurate diagnostic assistance for medical professionals. To achieve this goal efficiently, this study aims at exploring popular classification algorithms like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Random Forest, and Naive Bayes(NB) while leveraging XGBoost as an ensemble model for breast cancer identification employing various information like family history timelines, hormonal levels diagnostics such mammograms or biopsies among others The diagnostic appeals shown by traditional assessment methods that highly depend on expert opinions can be subjective and time-consuming hence employing modern AI technology offers increased accuracy through automated detection and early diagnosis towards upgraded treatment success.

Detecting breast cancer at early stages significantly improves chances of survival among diagnosed individuals. In our study, we investigated two prominent machine learning algorithms designed for the accurate characterization of tumor malignancy: Support Vector Machines (SVM) and k Nearest Neighbors (k NN). SVM is known for its ability to handle complex decision boundaries while coping with large datasets and high dimensions.

Utilizing mammogram features together with extensive patient information to create an augmented dataset was one way we tested this feature extraction technique. On the other hand. KNN is a non-parametric algorithm that involves finding similarities between existing data points from previous training sets and new occurrences using distance measures. By testing these algorithms' ability on real-world datasets containing multiple elements we intend to draw conclusions about their suitability as screening tools. Our team remains dedicated to producing cutting-edge tools critical in delivering precision medicine treatments focused on enhanced outcomes for all patients worldwide.

The study shall wield an all-encompassing dataset of breast cancer cases, encompassing pertinent patient intel and diagnostic traits extracted from medical imaging data. The SVM, k-NN, Random Forest, and NB models, both singularly and collectively through XGBoost, shall undergo evaluation by means of standard metrics such as precision, accuracy, recall, and F1 score. A comparative analysis shall impart insights into the strengths and weaknesses of each algorithm, empowering medical professionals to elect the most fitting approach for breast cancer detection.

The apex goal of this research is to forge a dependable and precise breast cancer detection system that can expedite healthcare providers in formulating prompt diagnoses and tailored treatment regimes. By harnessing the potency of machine learning algorithms, we strive to augment the ongoing endeavors in combating breast cancer, thus enhancing patient outcomes and salvaging lives.

2.1 Questions which are addressed

Expansive and intricate, detecting breast cancer is a problem that demands a thorough and meticulous analysis of diverse machine learning algorithms. Our objective is to delve into the intricacies of breast cancer detection and explore the capability of k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Random Forest, Naive Bayes (NB), and the ensemble model XGBoost. By scrutinizing these algorithms and their performance, we aim to make a significant contribution to the development of precise and efficient breast cancer detection techniques.

How do k-NN, SVM, Random Forest, NB, and XGBoost fare in terms of their classification accuracy for breast cancer detection?

Our primary goal is to evaluate the classification accuracy of k-NN, SVM, Random Forest, NB, and XGBoost in detecting breast cancer. We will train and assess these algorithms on a comprehensive dataset of breast cancer cases, using standard metrics such as accuracy, precision, recall, and F1 score to compare their performances. This analysis will provide a comprehensive insight into the strengths and weaknesses of each algorithm and help us identify the most efficient approach for breast cancer classification.

Which algorithm exhibits the best performance in terms of sensitivity and specificity for breast cancer detection?

The sensitivity and specificity of an algorithm are crucial metrics for evaluating its performance in detecting breast cancer. Sensitivity measures the ability of an algorithm to identify malignant cases accurately, while specificity measures the ability to classify benign cases correctly. By calculating these metrics

for k-NN, SVM, Random Forest, NB, and XGBoost, we can pinpoint the algorithm with the highest sensitivity and specificity, indicating its effectiveness in detecting breast cancer and minimizing false positives and false negatives.

How do k-NN, SVM, Random Forest, NB, and XGBoost handle high-dimensional and complex breast cancer datasets?

Breast cancer datasets are often multifaceted, with a high number of features, such as patient information and diagnostic data. Handling such intricate and high-dimensional datasets is a challenge for many machine learning algorithms. By examining the performance of k-NN, SVM, Random Forest, NB, and XGBoost on these datasets, we can assess their ability to handle the complexity and dimensionality of breast cancer data. This analysis will provide insights into the algorithms' scalability and suitability for real-world applications in breast cancer detection.

Can the amalgamation of individual algorithms using XGBoost as an ensemble model increase the accuracy of breast cancer detection?

Ensemble learning methods, such as XGBoost, have shown great potential in enhancing classification accuracy by combining the predictions of multiple algorithms. In this study, we aim to explore the effectiveness of using XGBoost as an ensemble model to merge the outputs of k-NN, SVM, Random Forest, and NB for breast cancer detection. By analyzing the performance of the ensemble model and comparing it to the individual algorithms, we can determine if the combination leads to improved accuracy and better overall performance in breast cancer detection.

How does the choice of hyperparameters impact the performance of k-NN, SVM, Random Forest, NB, and XGBoost in breast cancer detection?

The performance of machine learning algorithms is heavily influenced by the selection of hyperparameters, such as the number of neighbors in k-NN, the kernel type in SVM, and the number of trees in Random Forest. In this study, we will scrutinize the impact of different hyperparameter settings on the performance of k-NN, SVM, Random Forest, NB, and XGBoost in breast cancer detection. By undertaking a comprehensive hyperparameter tuning analysis, we can determine the optimal parameter configurations for each algorithm, leading to improved accuracy and robustness in breast cancer classification.

By addressing these crucial questions, our study aims to provide valuable insights into the performance, capabilities, and suitability of k-NN, SVM, Random Forest, and NB, and the ensemble model XGBoost

for detecting breast cancer.

2.2 Why Xgboost is used

XGBoost, a popular ensemble learning algorithm, has garnered significant attention and success in diverse machine learning tasks, including classification. Its exceptional performance and consistent achievement of state-of-the-art results in various competitions and benchmarks are attributed to its gradient boosting framework, which combines multiple weak learners (base models) to form a robust ensemble. XGBoost captures complex relationships and patterns in the data, leading to improved classification accuracy.

Breast cancer datasets often suffer from class imbalance, where the number of benign cases outweighs the malignant cases or vice versa. Imbalanced datasets pose challenges for classification algorithms, as they may favor the majority class and produce suboptimal results. XGBoost provides specific techniques, such as weight adjustment and sampling strategies, to handle class imbalance effectively. These techniques ensure that the ensemble model is not biased towards the majority class, thus enhancing the detection of both benign and malignant cases.

XGBoost offers built-in mechanisms for assessing feature importance, allowing researchers and medical professionals to gain insights into which features contribute significantly to breast cancer detection. This analysis can aid in understanding the underlying factors and biomarkers associated with breast cancer. By identifying the most relevant features, medical professionals can refine diagnostic protocols and focus on the most informative factors during the detection process.

Overfitting is a common concern in machine learning, where the model becomes overly specialized to the training data and performs poorly on unseen data. XGBoost incorporates regularization techniques, such as L1 and L2 regularization, which help control model complexity and prevent overfitting. These techniques ensure that the ensemble model generalizes well to new and unseen breast cancer cases, enhancing its reliability and accuracy in real-world scenarios.

XGBoost is designed for efficiency and scalability, making it suitable for large-scale datasets encountered in breast cancer research. It is optimized to handle high-dimensional data and can efficiently process a vast number of features and samples. With optimized implementations and parallel processing capabilities, XGBoost can effectively handle the computational demands of breast cancer detection, making it a practical choice for real-time or high-throughput applications.

XGBoost is compatible with various programming languages and platforms, including Python and R,

making it easy to integrate with existing machine learning workflows and tools used in breast cancer research. Furthermore, XGBoost provides flexible options for customization, allowing researchers to fine-tune the ensemble model according to their specific requirements and experimental setups.

Breast cancer datasets may contain noisy or outlier data points, which can negatively impact the performance of classification algorithms. XGBoost's robustness to noise and outliers is attributed to its utilization of an ensemble of base models. By aggregating predictions from multiple models, XGBoost reduces the impact of individual noisy or outlier instances, leading to more robust and reliable predictions.

In summary, XGBoost's collective knowledge of diverse algorithms, including k-NN, SVM, Random Forest, and NB, makes it an attractive choice as an ensemble model for breast cancer detection. Its exceptional performance, ability to handle imbalanced datasets, feature importance analysis, regularization techniques, scalability, flexibility, and robustness to noise make it a valuable tool in breast cancer research.

Chapter 3

Literature Survey

The authors[5] use WEKA to analyze data from the UCL ML repository to predict the condition of Breast Cancer in a tumour scan. Several factors like cell size, shape , nucleoli, Clump Thickness, Marginal Adhesion is considered before coming to a conclusion. Naive Bayes uses Gaussian Distribution to cluster the data based on the results obtained. Using WEKA along with Naive Bayes yields an accuracy of 94.08 percent after segmenting the results into benign and malignant clusters.

[9]The research work provides in-depth analyses of the technical and usability aspects of histopathological image characteristics and performs breast cancer diagnosis using the Breakhis and breast histopathology image datasets.A well structured dataset is generated by repeatedly extracting 13 Haralick texture characteristics from each histopathology image. The dataset generated is subjected to dimension reduction techniques like PCA and LDA. The machine learning technique used to identify breast cancer is K-Nearest Neighbor Classifier. Accuracy score of KNN using LDA was 80.0 percent,which was higher than the accuracy score of KNN using PCA, which was 56.0 percent.Whenever a dataset has texture features, the approaches suggested by authors may be used to get insights into which factors contribute the most to the target features.

The authors[22] used Grid search to present a model for predicting breast cancer using Support Vector Machine. The Initial Support Vector Machine model is evaluated in the absence of grid search. The Support vector machine model is then evaluated using grid search. Ultimately, a comparison study was performed, and a new model was created based on the results. The new model uses a grid search of data prior to fitting it for classification, which optimizes the outcome and produces much improved outcome than a conventional SVM model. It can be observed that the correct parameter values for gamma and C are crucial for a certain quantity of data. This approach could also be employed to anticipate other ailments, acting as a decision-support system in the healthcare division.

The authors[20] have employed five primary algorithms: Random Forests, SVM, K-NN,Logistic Regres-

sion, Decision Tree to compute, contrast and assess various findings attained elicited from sensitivity, confusion matrix, AUC, accuracy, and precision to discover the superlative machine learning algorithm that is exact, dependable, and finds the highest accuracy. In the Anaconda environment, all algorithms were written in Python using the scikit-learn module. After a thorough evaluation of the models, it was discovered that the Support Vector Machine outperformed all other methods in terms of efficiency (97.2%), precision (97.5%), and AUC (96.6%). However, To achieve greater accuracy, new parameters are can be used for larger sets of data with more illness types.

The proposed method in the paper[25] is Hierarchical Clustering Random Forest (HCRF) and Variable Importance Measure(VIM), for classification and feature selection based on the Gini Index respectively. The parameters of our model are selected using the grid search algorithm. Datasets utilized for the study include WBC and Wisconsin Diagnosis Breast Cancer. From the specified training set, several different training subsets are created using the bootstrap sampling technique. The trees that share similarities are grouped, together. In the end, we choose the decision tree from each cluster that has the highest area under the curve and discard the others. The developed model performs better when tried to compare to other classifiers like Adaboost and decision tree. The Selected Tree for Random Forest are of low similarity.On the WDBC dataset, our suggested technique achieves an accuracy of 97.05 %, and on the WBCdataset, it achieves an accuracy of 97.76%.

The authors[2] proposed a breast cancer detection model using microarray breast cancer gene expression data. A hybrid of two choice of feature selection techniques: the filter method using Fisher-score and the C5.0 algorithm's inner feature selection capability are applied.This is employed because the most prevalent issue with data on gene expression is its high dimensionality.Support vector machines, C5.0 Decision Trees, Logistic Regression, and artificial neural networks are the classification methods that were employed to evaluate the predictive accuracy of this strategy. Prior to the application of feature selection, 24481 genes were chosen, with ANN showing a better accuracy score of 86.99 percent and C5.0 showing the lowest accuracy score of 79.01 percent. When feature selection is applied,the number of genes chosen was reduced to 5 and all shrinkage models provided classification accuracy greater than 80 percent. The authors intend to examine the effectiveness of the suggested strategy using new datasets from microarrays that have varied qualities that differ in the quantity of classes, genes, and samples.

Yixuan Li et. al.[18] employed the LR, DT, RF,SVM and NN models to prognosticate the kind of breast cancer with other features. The prediction findings will aid in lowering the rate of false - positive results and developing appropriate therapeutic plans for recovery. In this investigation, 2 datasets are

employed. This analysis initially gathers source data from the BCCD dataset, that has 116 participants along with nine characteristics, and source data from the WBCD dataset, which comprises 699 participants containing 11 features. The source data from the WBCD dataset was then preprocessed, yielding data including 683 participants with nine characteristics and an index signifying whether the volunteer had a malignant tumour. Off the back of collating the accuracy, The ROC curve and F-measure metric of five different classifiers were used to determine which model should be used as the principal classifier in this investigation. It performs well on huge datasets. They are, however, significantly more difficult and time-consuming to build. This experiment only analyzes the data on 10 features. The lack of source data has an impact on the correctness of the outcomes. Furthermore, the RF may be used in conjunction with other approaches to data mining to provide more precise diagnostic conclusions.

In the proposed model author[28] suggests a method for locating Micro calcifications, tiny calcium apatite crystals that, despite their tiny size and low contrast, are the first indication of breast cancer. A coded contour is available with an image containing microcalcification denoting the area of their presence. An automated method employing discrete wavelet transform for segmenting and RF for classifying breast microcalcifications in mammograms respectively. The Digital Database for Screening Mammography has 966 mammography images divided into three classes: benign, malignant, and normal. To enhance, mammography images were processed through a two-dimensional discrete wavelet transform. The tissue surrounding the microcalcification is removed using the maximum entropy approach. The sequential forward features selection procedure is used to minimize the set of features after the features are chosen using GLCM. Following that, Random Forest is used and a grid search was used to determine the parameters. It was trained using 10-fold cross-validation. In comparison to previous models, this one has a 95% accuracy rate.

The authors[23] used Logistic Regression for Breast Cancer Detection. It was observed that the logistic regression method had an accuracy of more than 94% in detecting whether the cancer was malignant or benign.The findings indicate that integrating multidimensional data with various categorization, feature selection, and dimension reduction strategies might give beneficial tools for analysis in this domain. More research is necessary to enhance the efficiency of classification systems so they are capable of predict additional variables.

A myriad of cutting-edge approaches for forecasting and detecting breast cancer have been unearthed through an extensive literary expedition. The analyses scrutinized a plethora of machine learning algorithms, including Naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random

Forest, Logistic Regression, and Decision Trees. Each algorithm demonstrated its own unique strengths and weaknesses in terms of precision, accuracy, and efficiency. The proposed system endeavors to capitalize on the strengths of these individual models by engineering an ensemble model employing XGBoost as the ensembler. By fusing the predictive prowess of KNN, SVM, Naive Bayes, and Random Forest, the proposed system aspires to concoct an innovative model that surpasses the accuracy of the conventional standalone models. This ensemble model harbors the potential to fortify breast cancer prediction and diagnosis, delivering more dependable outcomes and contributing significantly to better decision-making in the healthcare sector.

Chapter 4

Architectural Design

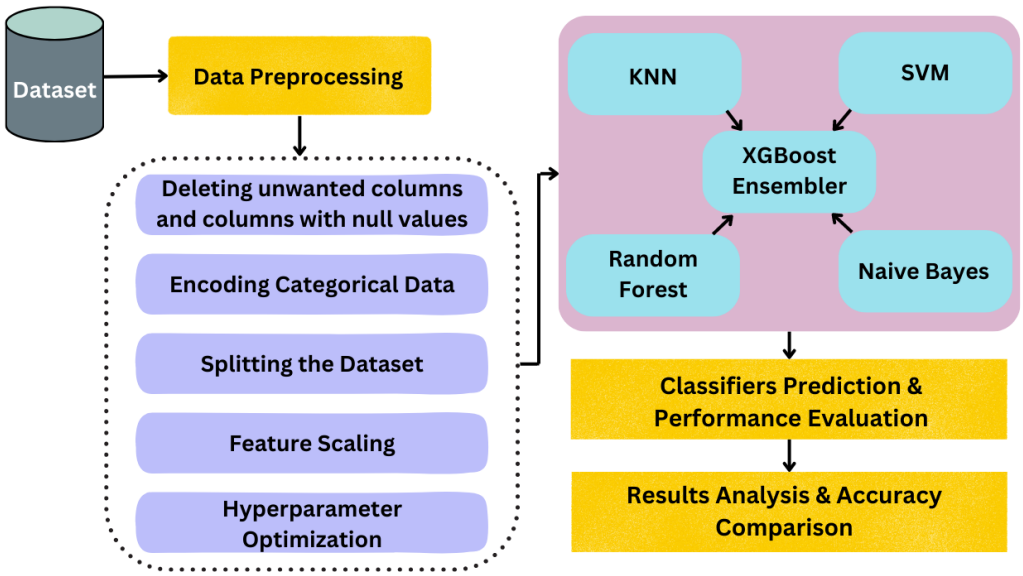

Figure 4.1. Architectural Design

Our solution's architecture entails exploiting the Wisconsin Breast Cancer Dataset as our primary data source. We employ various data preprocessing techniques to boost the quality of the dataset and prime it for analysis. These techniques involve excising undesirable and missing-value-laden columns, as well as encoding categorical data.

To assess the models' efficacy, we split the dataset into training and testing sets in an 80:20 proportion. Additionally, we enact feature scaling to standardize the self-reliant variables within a specified span. This exercise guarantees that no variable overwhelms the others by equalizing them to the same scale.

Afterward, we conduct hyperparameter optimization to fine-tune the machine learning models' parameters. Our goal is to attain the most exceptional accuracy possible. We consider individual models such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and Naive Bayes. These models are constructed and fitted with the preprocessed data, and we evaluate their respective

performance.

Additionally, we construct an ensemble model using these four individual models, employing XGBoost as the ensemble method. The ensemble model combines the predictions from each individual model to generate a final prediction. The performance of this ensemble model is then evaluated and compared with the performance of the individual traditional models.

By following this architecture, we aim to thoroughly analyze the performance and effectiveness of the individual models as well as the ensemble model in predicting breast cancer using the Wisconsin Breast Cancer Dataset.

Chapter 5

Implementation

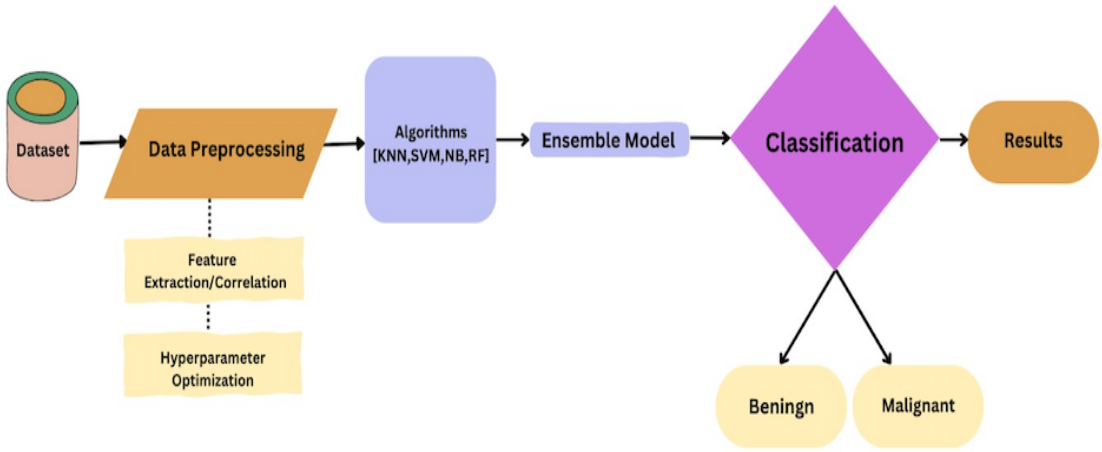

Figure 5.1. Flow Diagram

Our solution utilizes the Wisconsin Breast Cancer Dataset as the primary data source, as depicted in the data flow diagram. Several data preprocessing techniques are employed to enhance the dataset's quality and prepare it for analysis. These techniques involve the removal of undesirable and missing-value-laden columns and the encoding of categorical data. To evaluate the effectiveness of the models, we divide the dataset into training and testing sets in an 80:20 ratio. Furthermore, feature scaling is implemented to normalize the independent variables within a specified range. This ensures that no variable dominates the others by standardizing them to the same scale. Next, we conduct hyperparameter optimization to fine-tune the parameters of the machine learning models. This is done to achieve the highest possible accuracy. We consider several models, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and Naive Bayes. These models are constructed and fitted with the preprocessed data, and their respective performances are evaluated. Moreover, an ensemble model is developed using these four individual models and XGBoost as the ensemble method. The ensemble model combines the

predictions from each individual model to generate a final prediction. The performance of this ensemble model is then assessed and compared to that of the individual traditional models.

5.1 Dataset

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset has been subjected to the cutting-edge machine learning algorithms. It comprises the patient's unique ID, cell nuclei traits, and medical diagnosis. The ID serves as the patient's distinct identification number, while the cell nuclei features emanate from a digital image of a fine needle aspirate (FNA) of a breast mass. These features encapsulate ten distinct characteristics of each cell nucleus, each containing three attributes, namely: (1) mean, (2) standard error (3) worst. Astoundingly, a whopping 30 features of 569 patients were meticulously evaluated. Two additional columns, namely, id and unnamed:32, were also included. Out of all the cases, a staggering 357 benign cases and 212 malignant ones were identified.

5.2 Data Preprocessing

In the realm of data analysis and machine learning, the art of data preprocessing is a pivotal step. It's a process that involves metamorphosing raw data into a spick-and-span, well-organized, and fitting format for further analysis. By tackling knotty issues such as missing values, outliers, irrelevant features, and inconsistent data formats, data preprocessing elevates the quality and credibility of a dataset. Our solution employs a gamut of Data Preprocessing techniques, including:

5.2.1 Deletion of Unwanted Columns and Columns with Null Values

In the Wisconsin dataset, two columns, namely "id" and "unnamed:32", were spotted as superfluous and beset with null values. Consequently, these columns were expunged from the dataset to ensure data hygiene and minimize any potential prejudice.

5.2.2 Encoding Categorical Data

As the machine learning models thrive on numerical data, it's mandatory to transmute the categorical variables into numeric renditions. In the given dataset, the 'diagnosis' column flaunted categorical values of 'M' and 'B', which were transposed into 1 and 0, correspondingly, to expedite the subsequent scrutiny.

5.2.3 Dataset Splitting

To invigorate the potency and scrutiny of our machine learning models, we bifurcated the dataset into two subsets: a training set and a test set. This partitioning enabled us to educate the models on a fraction of the data and scrutinize their efficiency on unobserved data, facilitating precise extrapolation and

models' assessment.

5.2.4 Feature Scaling

Scaling the features is a pivotal step in preprocessing to guarantee that variables are on a level playing field. This entails homogenizing the independent variables within a restricted span between 0 and 1. By bringing the variables to the same level and magnitude, no solitary variable reigns supreme over the others, foiling any predisposed model outcomes. This technique is especially critical for models that hinge on distance-based computations, like the Euclidean distance..

5.2.5 Hyperparameter Optimization:

The art of machine learning is fraught with uncertainty and requires a deft hand to navigate the hyperparameters that affect performance. With the power to control the fate of a model, hyperparameters can twist and turn results in unexpected ways. The pursuit of accuracy requires a tireless search for optimization techniques that can unearth the optimal values. Techniques such as GridSearchCV and RandomizedSearchCV are the tools of the trade, allowing the intrepid explorer to boldly venture into the unknown and emerge with the highest model performance.

5.3 Algorithms

We have used four Algorithms KNN,SVM,Naive Bayes and Random forest for detection of breast cancer individually and calculated their accuracy. After that we created an ensemble model using above algorithms to improve accuracy of our model.In the beginning, we imported a number of libraries, such as Numpy for working with arrays, Pandas for working with datasets, Matplot and Seaborn for showing graphs, and train and test for dividing datasets.Other imported libraries were created specifically for the implementation of algorithms.Some are used to display performance metrics

5.3.1 KNN

K-Nearest Neighbors (KNN) is a machine learning technique that predicts the class or value of a new data point based on its similarity to already labeled data points. The classification process involves determining the distance between the new data point and its k nearest neighbors, and then assigning the majority class label to the new data point. As KNNs are non-parametric, the accuracy of the results depends on carefully choosing the value of k and managing feature scaling.

To build a KNN classifier, we specified the number of neighbors to be considered for classification using the ”neighbours” option. The distance metric used to measure the separation between data points was determined by the ”metric” parameter, with the Minkowski distance metric selected in this case. The ”p” parameter was also used when the Minkowski distance metric was selected, as it determines the power parameter needed to calculate the distance. After fitting the model with training data, it was tested and the accuracy was calculated. Finally, the best parameters were determined using randomized search CV, and the accuracy was recalculated using the best parameters.

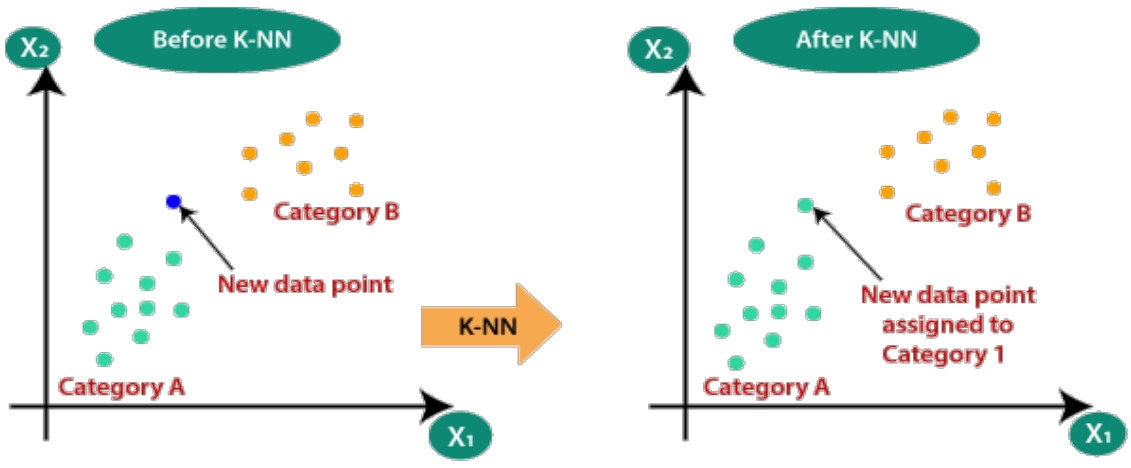

Figure 5.2. K-Nearest Neighbors (KNN)

5.3.2 SVM

The Support Vector Machines (SVM) technique is a powerful machine learning method used for both classification and regression applications. Essentially, the SVM algorithm seeks to identify the ideal hyperplane that maximizes the margin of separation between data points of different classes. This is achieved by mapping the input data into a higher-dimensional feature space and identifying the hyperplane that maximizes the distance between the support vectors (i.e. the data points closest to the decision boundary).

One of the strengths of SVM is its ability to handle both linearly and non-linearly separable data by utilizing various kernel functions. Additionally, SVM is known for its versatility in choosing kernels, its ability to handle high-dimensional data, and its resistance to outliers.

Before scaling up, the SVM classifier model must be developed, tested, and its accuracy assessed. The ideal hyperplane that separates different classes of data points is critical to the accuracy of the model, and the scales of the features can affect the position and orientation of the hyperplane. In cases where features have varying scales, SVM may prioritize features with larger scales, resulting in a biased or unfavorable decision boundary. To ensure that each feature contributes equally to the SVM model, the features can be scaled.

Once accuracy has been determined, the confusion matrix is printed. In this process, three hyperparameters are taken into account: SVM C, gamma, and kernel. The cost parameter C determines the penalty for misclassifying practice examples, while the gamma hyperparameter defines the influence of a single training example. Using the best parameters selected by GridSearchCV, accuracy is calculated again.

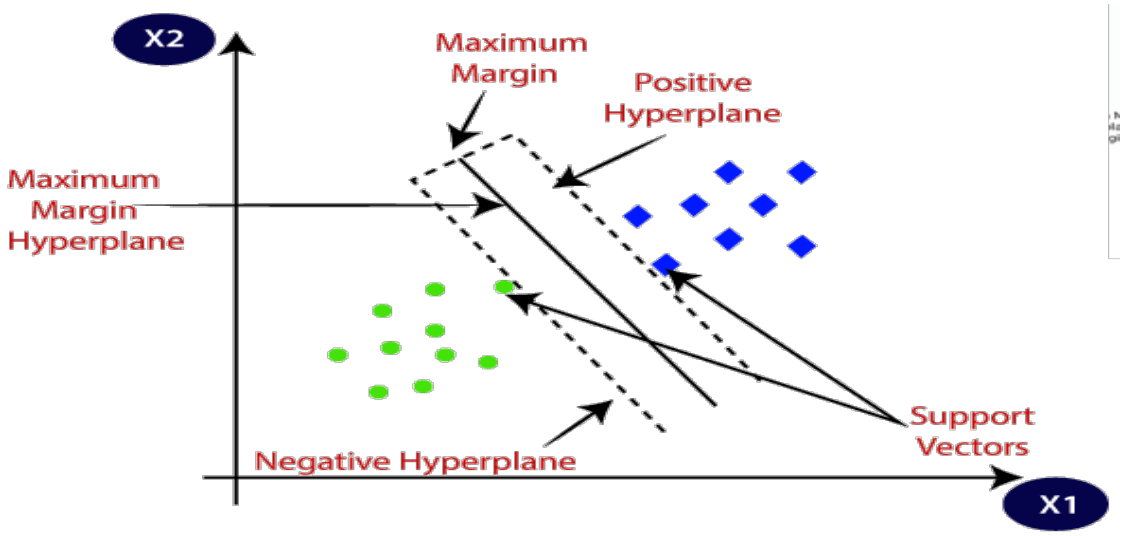

Figure 5.3. Support Vector Machine

5.3.3 Naive Bayes

Behold, a favored technique for classifying tasks in probabilistic machine learning is none other than the illustrious Naive Bayes. This wizardry is founded upon the Bayes theorem and, forsooth, assumes that features remain unconnected. Hence, it hath earned the moniker ”naive.” By reckoning the probabilities of each characteristic given the class, Naive Bayes ascertains the probability that a datum belongs to a particular class. Then, by multiplying these probabilities in concert, the overall likelihood doth emerge. Verily, the algorithm doth select the predicted class with the preeminent probability.

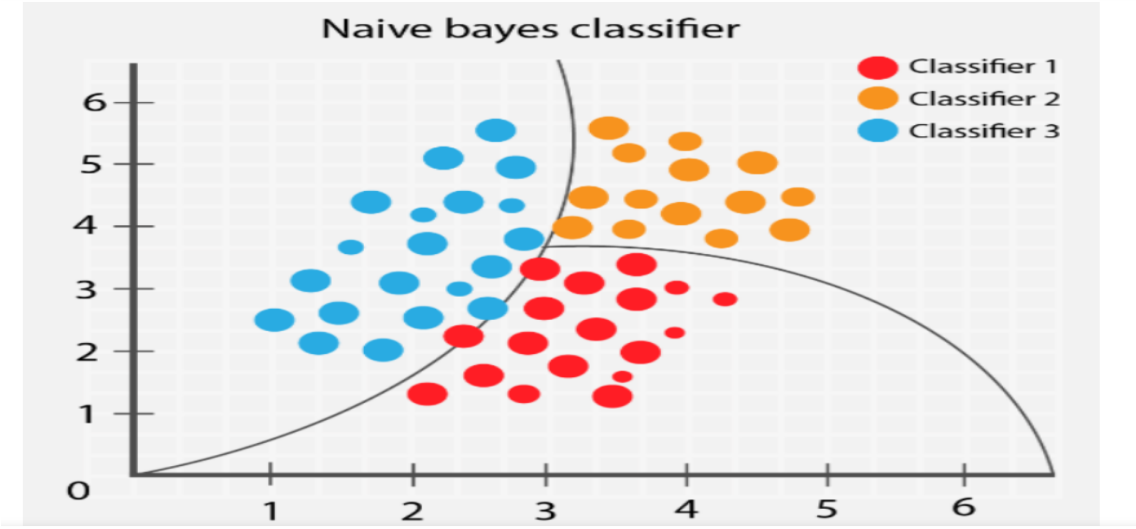

Figure 5.4. Naive Bayes

5.3.4 Random Forest

The Random Forest algorithm morphs into a chimerical force, capable of both classification and regression tasks. It amalgamates myriad decision trees, each one trained on a capricious subset of the data and features. During the prediction phase, the algorithm amalgamates the predictions of the individual trees to culminate a final decision. Random Forest is a warrior, slaying high-dimensional data, nonlinear correlations, and noisy data, while also providing feature importance measures, enabling the identification of the most influential features.

The Random Forest classifier metamorphoses into a golem, crafted with minmax scaling to bring all values to the same plane for better classification. We then performed a calculation of accuracy and drew a confusion matrix.

GridSearchCV transforms into a sorceress, utilizing a systematic exploration of a predefined hyperparameter grid to unearth the optimal combination for a Random Forest model. It performs an exhaustive search, scrutinizing each combination through cross-validation to determine the best hyperparameters.

This incantation helps to maximize the model's performance by finding the hyperparameters that yield the highest accuracy or other desired evaluation metric.

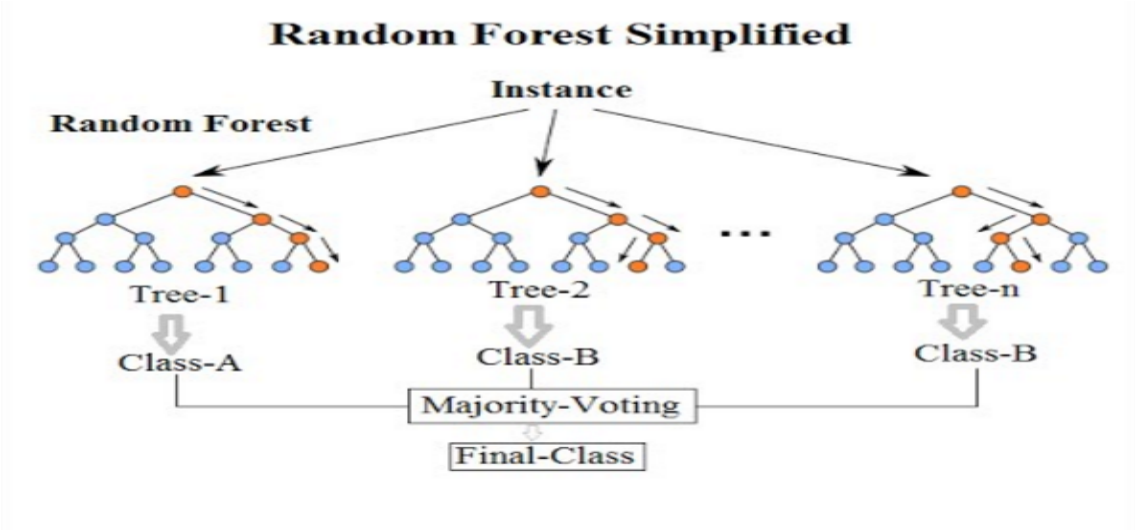

Figure 5.5. Random Forest

5.3.5 XGboost Ensemble model

Extreme Gradient Boosting (XGBoost) is a sophisticated ensemble learning technique that builds a strong predictive model by combining the predictions of several weak predictive models, often decision trees. Each succeeding model is trained to rectify the mistakes caused by the prior models using a gradient boosting framework. To enhance model generalisation and reduce overfitting, XGBoost uses regularisation techniques including shrinkage and feature subsampling. XGBoost builds an ensemble of models that together produce extremely accurate predictions by iteratively optimising a loss function.

SVM, RF, NB, and KNN were the four algorithms we gave xgboost as inputs. We also utilised certain standard parameters as inputs and calculated accuracy.Using GridSearchCV, we were able to calculate the input parameter with the highest level of accuracy.

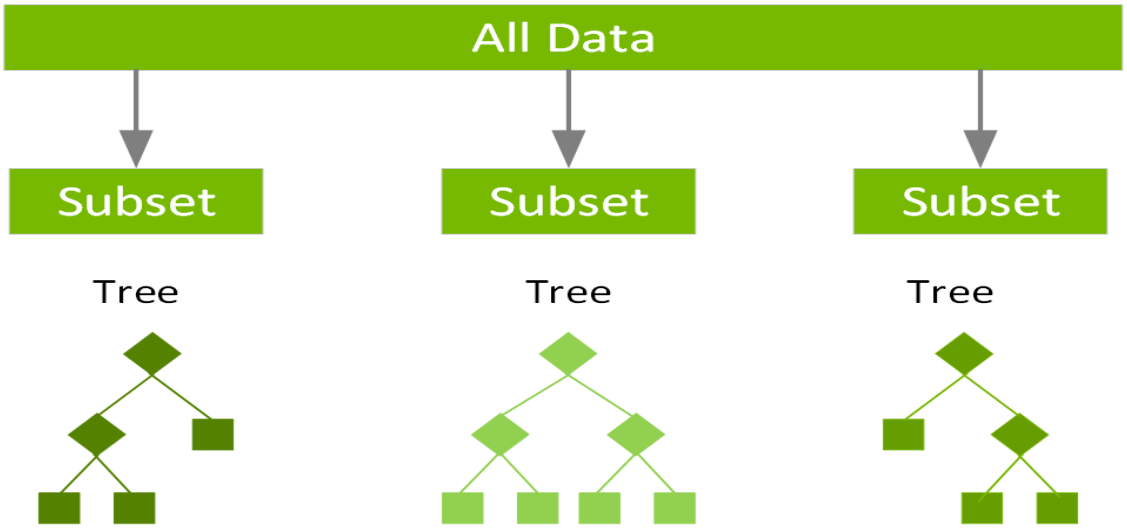

Figure 5.6. XGBoost

5.4 UML Diagram

The Unified Modeling Language (UML) is a dynamic and versatile modeling language that serves as a blueprint for visualizing complex system designs. It's like a conductor's score for a symphonic orchestra. When it comes to complex applications involving numerous teams, effective communication becomes vital.

UML is commonly associated with object-oriented design and analysis, utilizing various elements and associations to create meaningful diagrams. However, in machine learning projects, the use of UML diagrams is relatively minimal due to the nature of the workflow.

In machine learning projects, each component executes a specific task or operation on the data, and the outcomes from previous components are often passed along to subsequent steps without intricate relationships. The emphasis is on the sequential flow of data and the application of algorithms or models to produce the desired outcome. As a result, the use of UML diagrams, which rely heavily on associations between components, is less prevalent in the context of machine learning.

Instead, machine learning projects often employ other visualization techniques, such as flowcharts or pipeline diagrams, to illustrate the sequential steps involved in data preprocessing, model training, evaluation, and prediction. These visualization methods provide a clear overview of the data flow and the transformations performed at each stage, without the need for intricate associations between components.

UML diagram representing Breast Cancer Detection using Ensemble model is depicted below:

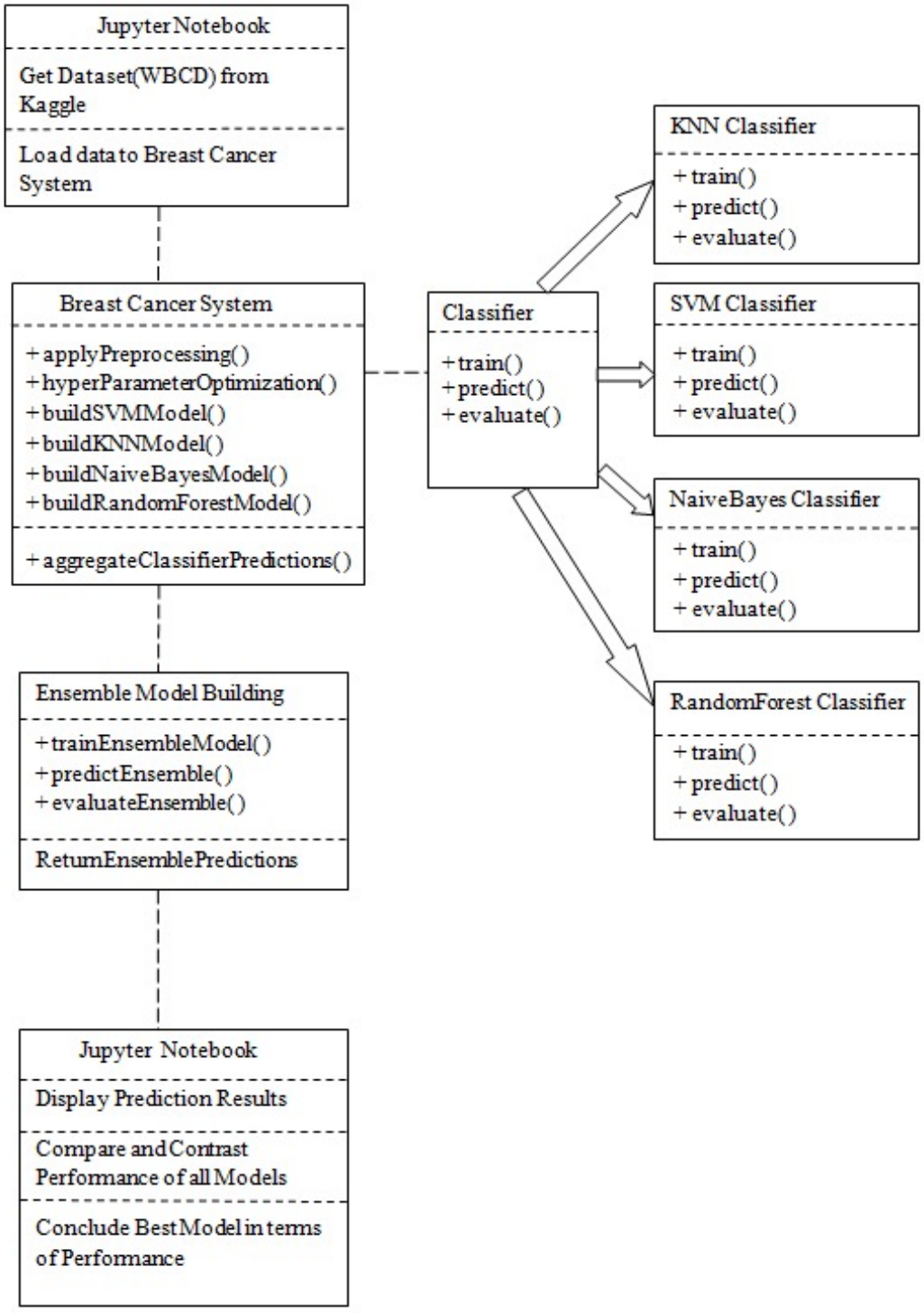

Figure 5.7. UML Diagram

Working:

Within Jupyter Notebook, we fetch the Wisconsin Breast Cancer dataset from Kaggle, which serves as

the foundation for our breast cancer system. To fortify the dataset's quality, we meticulously scrutinize it for any anomalies, striking them from existence to ensure immaculate data. Our data cleaning process is a cutthroat one, as we expunge null values and purge invalid columns from the dataset.

We then calculate an array of parameters, including mean, average, and correlation, for each feature in the dataset. This allows for a comprehensive overview of the data's characteristics and relationships among the different features, which we organize into a tabular form.

Our preprocessing steps and statistical measures not only enhance the data quality but also set a solid foundation for subsequent analysis and modeling in the breast cancer detection system. To construct individual classifier models, we employ KNN, SVM, Naïve Bayes, and Random Forest. Further, we magnify their performance by utilizing Hyperparameter Optimization Techniques.

With each individual classifier providing predictions, we aggregate their results to build an ensemble model. This stage involves the XGBoost algorithm, whose predictions we compare with all other classifiers. To aid in our evaluation, we visualize the ensemble model's performance alongside the individual classifiers, allowing for a clearer understanding of their strengths and weaknesses.

Finally, based on overall performance, we identify the best model and conclude it within Jupyter Notebook.

Chapter 6

Performance Metrics

This chapter explains the parameters utilized to assess the performance of employed machine learning techniques. To evaluate performance, various metrics such as a Confusion Matrix, Accuracy, Precision, and Recall are derived.

6.1 Confusion Matrix :

The term ”confusion matrix” is often used interchangeably with an error matrix. It represents a table-like structure that presents the outcomes of an algorithm. The anticipated class is represented in each row, while the actual class is represented in each column (or vice versa) in this matrix.

Behold, the 2x2 matrix reigns supreme in visualizing the interplay between actual and predicted classes. The rows boast of the actual classes, while the columns stand tall as the predicted classes. From this matrix emerges a cornucopia of performance metrics, each more intriguing than the last. Behold, the metrics that may be derived from the confusion matrix include accuracy, precision, recall (sensitivity), specificity, and F1 score.

	Actual Positive(N)	Actual Negative (N)
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Table 6.1: Confusion Matrix Table

i) True Positive (TP) signifies the accurate identification of individuals with cancer as malignant.

ii) False Positive (FP) indicates the incorrect identification of non-cancerous individuals as malignant.

iii) True Negative (TN) represents the accurate identification of non-cancerous individuals as benign.

iv) False Negative (FN) indicates the incorrect identification of individuals with cancer as benign.

6.1.1 Accuracy :

Accuracy serves as a reliable indicator of the level of correctness achieved during model training and its overall performance. It quantifies the extent of correct predictions relative to incorrect ones. The provided equation can be utilized to calculate the accuracy value.

Let us delve into the enigma of accuracy, by delving into a binary classification scenario where we seek to prognosticate whether an email is mere spam or not. Our arsenal comprises a plethora of emails, complete with their respective labels, which we use to train our model. The model is then put to the test by making predictions on an entirely separate dataset. The accuracy is then deduced by dividing the number of emails that were correctly classified by the total number of emails in the test set.

Accuracy is an easily comprehensible metric, for it depicts the percentage of precise predictions. To illustrate, if the model attains an accuracy of 85%, it signifies that it accurately classified 85% of the instances in the test set.

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

6.1.2 Precision :

Precision measures the level of correctness in identifying positive outcomes. It is determined by calculating the ratio of true positives to the total number of positives. Precision primarily focuses on the system's ability to handle positive values but does not provide information about negative values. Let us delve into the depths of precision by scrutinizing a binary classification quandary. Our task is to prognosticate whether a patient is afflicted with a certain medical ailment or immune to it. Precision hones in on the positive class, which signifies the existence of the ailment. When we scrutinize the model's efficacy, precision computes the ratio of accurately predicted positive instances to the total instances forecasted as positive.

Precision is especially invaluable in scenarios where the ramifications of false positives are calamitous or where the cost of misdiagnosis is exorbitant. Consider medical diagnosis, where a false positive prophecy could engender unnecessary medical procedures and surgeries, subjecting patients to physical and emotional trauma. In such cases, a high precision classifier is indispensable to curtail false positives and ensure accurate prognostications.

Precision (P) = TP/TP+FP

6.1.3 Recall :

Recall, also known as sensitivity, quantifies the ratio of correctly identified positive instances to all observations. It serves as a measure of the system's effectiveness in predicting positives and determining associated costs. Recall provides insights into the system's ability to capture positive instances accurately. To fathom recall, let's contemplate a binary classification quandary where we prophesy whether an electronic missive is spam or not. Recall hones in on the affirmative class (spam missives) and appraises the model's knack to snag all the bona fide affirmative occurrences. It computes the ratio of veritable affirmative occurrences to the summation of veritable affirmatives and counterfeit negatives. A towering recall score evinces that the model possesses a meager counterfeit negative rate, insinuating it effectively detects the bulk of affirmative occurrences in the dataset. Conversely, a feeble recall score intimates that the model overlooks a momentous number of affirmative occurrences.

Recall = TP/TP+FN

Chapter 7

Result Analysis

Our groundbreaking work entails the identification of breast cancer utilizing sophisticated machine learning models like Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naïve Bayes, Random Forest, and XGBoost Ensemble Model. We executed these models with the aid of renowned open-source machine learning libraries in Python, specifically numpy, pandas, and Scikit-learn. The Jupyter Notebook, an open-source web application, enabled us to execute the program seamlessly. To conduct our experiment, we partitioned the dataset comprising of 569 observations into two sets: 80% for training and 20% for testing.

We scrutinized the classifiers' performance by gauging Accuracy, Precision, Recall, and scrutinizing the Confusion Matrix. The outcomes, showcased in Table Y, flaunt the might of each proposed model. Remarkably, the Ensemble model outshone other Conventional Machine Learning Algorithms, clinching a phenomenal accuracy of 97.7%.

	KNN	SVM	Naive Bayes	Random Forest	XGBoost Ensembler
Accuracy(%)	95.61	96.49	94.7	96.49	97.71
Precision(%)	1.0	1.0	94.11	95.89	95.89
Recall(%)	89.58	91.66	96.96	98.59	98.59

Table 7.1: Performance Metrics Indices

7.1 Confusion Matrixes of Models:

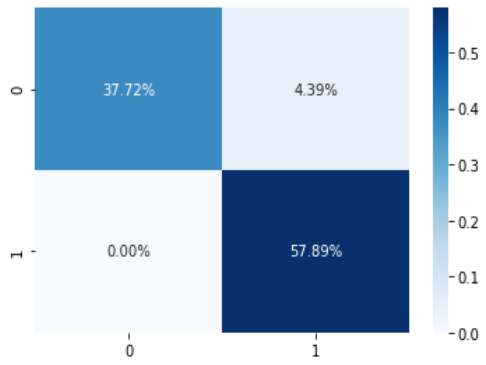

Figure 7.1. Confusion Matrix of KNN

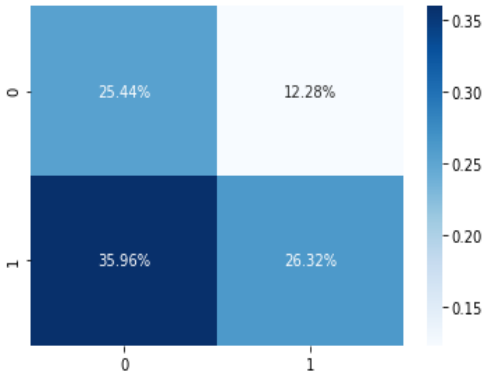

Figure 7.2. Confusion Matrix of SVM

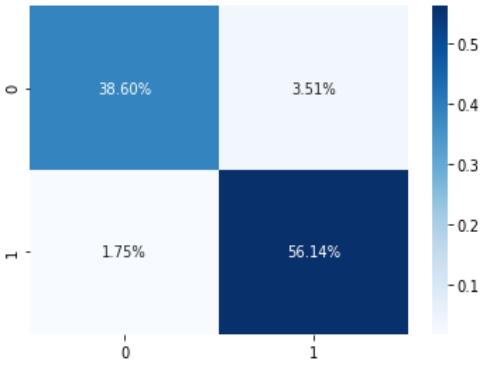

Figure 7.3. Confusion Matrix of Naive Bayes

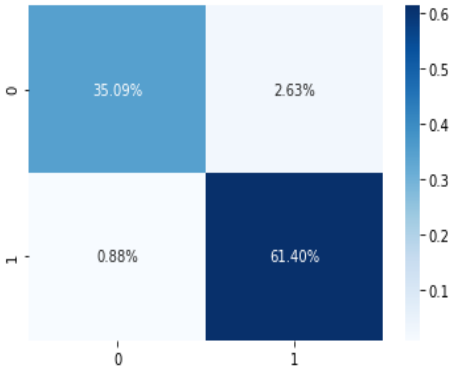

Figure 7.4. Confusion Matrix of Random Forest

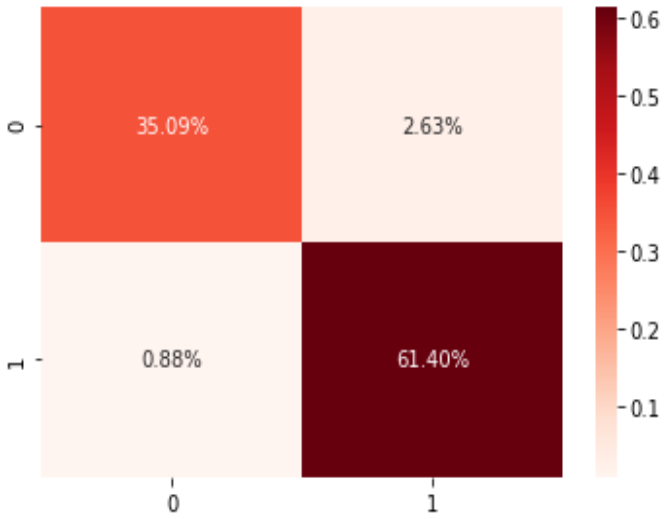

Figure 7.5. Confusion Matrix of XGBoost Model

7.2 Accuracy Comparision:

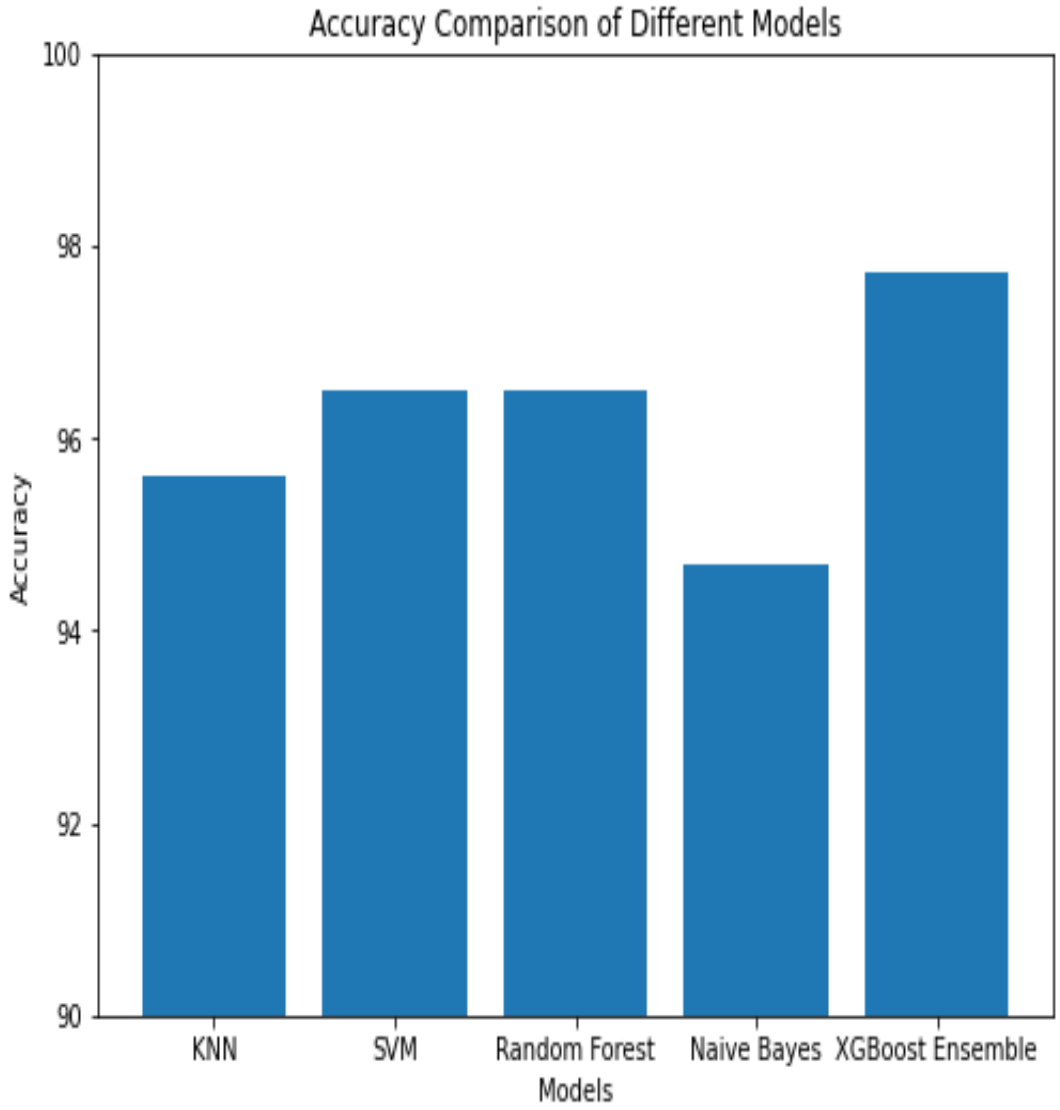

Figure 7.6. Bar Plot of Accuracy comparisions

Chapter 8

Conclusion and Future Enhancement

The aim of this investigation is to fabricate a formidable diagnostic scheme for breast cancer patients utilizing the WBCD benchmark database. The integration of data mining technologies in the medical arena is paramount as it contributes vastly to the decision-making process. We meticulously scrutinized the efficacy of various classifiers on patient data parameters and discovered that Support Vector Machine (SVM) and Random Forest (RF) achieved the utmost classification accuracy of 96.49% in prophesying breast cancer. Conversely, Naive Bayes displayed the lowest accuracy of 94.7% among the classifiers. Nevertheless, the Ensemble Model, constructed by interweaving all the classifiers, surpassed individual models with an accuracy of 97.71%. This accentuates the importance of machine learning in expediting premature prognosis of breast cancer, rendering it an indispensable instrument in healthcare research and medical centers.

Our forthcoming endeavors will encompass an intricate and meticulous examination of these datasets, amalgamating the prowess of machine learning with advanced deep learning models. We shall scrutinize the viability of deploying more sophisticated and intricate deep learning architectures to ameliorate the performance. Furthermore, we shall attempt to subject our deep learning approach to larger datasets, containing an extensive range of disease categories, to attain a superior level of accuracy in diagnosis. Moreover, we shall fabricate a user interface (UI) for the implemented system, simplifying the interpretation and evaluation of results through the medium of visual analysis. This UI shall offer a user-friendly platform, enabling one to interact with the system and glean valuable insights from the analysis.

References

[1] N. S. Ismail and C. Sovuthy, ”Breast Cancer Detection Based on Deep Learning Technique,” 2019 International UNIMAS STEM 12th Engineering Conference (EnCon), Kuching, Malaysia, 2019, pp. 89-92

[2] Hamim, M., El Moudden, I., Moutachaouik, H., Hain, M, ”Decision Tree Model Based Gene Selection and Classification for Breast Cancer Risk Prediction,” Communications in Computer and Information Science, vol 1207

[3] Kriti Jain, Megha Saxena and Shweta Sharma: “Breast Cancer Diagnosis Using Machine Learning Techniques”, IJISET - International Journal of Innovative Science, Engineering Technology, Vol. 5 Issue 5, May 2018.

[4] H. Kamel, D. Abdulah and J. M. Al-Tuwaijari, ”Cancer Classification Using Gaussian Naive Bayes Algorithm,” 2019 International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 165-170

[5] Megha Rathi, Arun Kumar Singh ”Breast Cancer Prediction using Na¨ıve Bayes Classifier” International Journal of Information Technology Systems, Vol. 1; No. 2: ISSN: 22779825 (July-Dec. 2012)

[6] T. A. Shaikh and R. Ali, ”A CAD Tool for Breast Cancer Prediction using Naive Bayes Classifier,” 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2020, pp. 351-356

[7] Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury ”Breast Cancer Detection Using Machine Learning Algorithms ” doi:10.1109/CTEMS.2018.8769187

[8] Than Than Htay,Su Su Maung, ”Early Stage Breast Cancer Detection System using GLCM Feature extraction and K-nearest Neighbor on Mammography image,” 2018 18th International Symposium on Communications and Information Technologies (ISCIT),IEEE

[9] Tevar Durgadevi Murugan,Mahendra G.Kanojia, ”Breast Cancer Detection Using Texture Features and KNN Algorithm,” In:HIS 2020-Part of the Advances in Intelligent Systems and Computing book series (AISC,volume 1375)

[10] Suhas Athani, Shreesha Joshi, B. Ashwath Rao, Shwetha Rai , N. Gopalakrishna Kini,”Parallel Implementation of kNN Algorithm for Breast Cancer Detection,”Advances in Intelligent Systems and Computing, vol 1176

[11] Youssef Aamer, Yahya Benkaouz, Mohammed Ouzzif, Khalid Bouragba, "A new approach for increasing K-nearest neighbors performance," 2020 8th International Conference on Wireless Networks and Mobile Communications (WINCOM),IEEE

[12] G. D. Rashmi, A. Lekha, Neelam Bawane, "Analysis of Efficiency of Classification and Prediction Algorithms (kNN) for Breast Cancer Dataset," Advances in Intelligent Systems and Computing, vol 434,

[13] Y. S. Deshmukh, P. Kumar, R. Karan and S. K. Singh, "Breast Cancer Detection-Based Feature Optimization Using Firefly Algorithm and Ensemble Classifier," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 1048-1054

[14] Nur Atiqah Hamzah, Sabariah Saharan , Khuneswari Gopal Pillay "Classification Tree of Breast Cancer Data with Mode Value for Missing Data Replacement," Proceedings of the 7th International Conference on the Applications of Science and Mathematics 2021,

[15] amim, M., El Moudden, I., Moutachaouik, H., Hain, M, "Decision Tree Model Based Gene Selection and Classification for Breast Cancer Risk Prediction," Communications in Computer and Information Science, vol 1207

[16] Kriti Jain, Megha Saxena and Shweta Sharma: "Breast Cancer Diagnosis Using Machine Learning Techniques", IJISET - International Journal of Innovative Science, Engineering Technology, Vol. 5 Issue 5, May 2018.

[17] Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta: "Diagnosis of Breast Cancer using Decision Tree Models and SVM". IJISET - International Journal of Innovative Science, Engineering Technology,Volume: 05 Issue: 03 Mar-2018

[18] Yixuan Li, Zixuan Chen, 2018: "Performance Evaluation of Machine learning methods for breast cancer prediction", Science publishing group 2018.

[19] Thomas Noel, Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, 2016, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Elsevier B.V. 2016.

[20] Mohammed Amine Naji , Sanaa El Filalib, Kawtar Aarikac, EL Habib Benlahmard, Rachida Ait Abdelouhahide, Olivier Debauchef "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis", Elsevier August 9-12, 2021

[21] S. Sathyavathi, S. Kavitha, R. Priyadharshini and A. Harini, “Breast Cancer Identification Using Logistic Regression” Biosci.Biotech.Res. Comm. Special Issue Vol 13 No 11 (2020)

[22] Vishal Deshwal, Mukta Sharma,” Breast Cancer Detection using SVM Classifier with Grid Search Technique”, International Journal of Computer Applications (0975 – 8887) Volume 178 – No. 31, July 2019

[23] Prof. Ajit N.Gedam, Kajol B. Deshmane, Nishigandha N.Jadhav, Ritul M.Adhav, Akanksha N.Ghodake,” Breast Cancer Detection using Logistic Regression Algorithm”, International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), Volume 11, Issue 5, May 2022.

[24] R. D. Ghongade and D. G. Wakde, ”Computer-aided diagnosis system for breast cancer using RF classifier,” 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 1068-1072

[25] Z. Huang and D. Chen, ”A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm,” in IEEE Access, vol. 10, pp. 3284-3293, 2022,

[26] S. Murugan, B. M. Kumar and S. Amudha, ”Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest,” 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), 2017, pp. 763-766

[27] S. Kabiraj et al., ”Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm,” 2020 11 th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-4,

[28] R. Fadil, A. Jackson, B. A. El Majd, H. El Ghazi and N. Kaabouch, ”Classification of Microcalcifications in Mammograms using 2D Discrete WaveletTransform and Random Forest,” ,

[29] L. Liu, ”Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning,” 2018 International Conference on Robots Intelligent Systems (ICRIS), 2018, pp. 157-160,

[30] B. Dai, R. -C. Chen, S. -Z. Zhu and W. - W. Zhang, ”Using Random Forest Algorithm for Breast Cancer Diagnosis,” 2018 International Symposium on Computer, Consumer and Control (IS3C), 2018, pp. 449-452,

[31] H. Rajaguru and S. K. Prabhakar, ”Expectation maximization based logistic regression for breast cancer classification,” 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 2017, pp. 603-606,

[32] T. A. Shaikh and R. Ali, ”Combating Breast Cancer by an Intelligent Ensemble Classifier Approach,” 2018 International Conference on Bioinformatics and Systems Biology (BSB), Allahabad, India, 2018, pp. 5-10