A Mini Project with Seminar On

DETECTING EARLY PHASE BREAST CANCER USING HYBRID MACHINE LEARNING ALGORITHMS

Submitted in partial fulfillment of the requirements for the award of the

Bachelor of Technology

in

Department of Computer Science and Engineering
(Artificial Intelligence and Machine Learning)

by

P. Mamatha	**20241A6643**
M. Krishnamai	**20241A6637**
Ch. Sruthi	**20241A6655**
M. Tejaswini Reddy	**21248A6602**

Under the Esteemed guidance of

Dr. E. Poornima

Associate Professor

Department of Computer Science and Engineering
(Artificial Intelligence and Machine Learning)

GOKARAJU RANGARAJU INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Approved by AICTE, Autonomous under JNTUH, Hyderabad)

Bachupally, Kukatpally, Hyderabad-500090

GOKARAJU RANGARAJU INSTITUTE OF ENGINEERING AND TECHNOLOGY
(Autonomous)

Hyderabad-500090

CERTIFICATE

This is to certify that the mini project entitled "**DETECTING EARLY PHASE BREAST CANCER USING HYBRID MACHINE LEARNING ALGORITHMS"** is submitted by **P. Mamatha (20241A6643), M. Krishnamai (20241A6637), Ch. Sruthi (20241A6655) and M. Tejaswini Reddy (21248A6602)** in partial fulfillment of the award of degree in BACHELOR OF TECHNOLOGY in Computer Science and Engineering (Artificial Intelligence and Machine Learning) during Academic year 2023-2024.

Internal Guide	Head of the Department
Dr. E. Poornima	**Dr. G. Karuna**

External Examiner

ACKNOWLEDGEMENT

There are many people who helped us directly and indirectly to complete our project successfully. We would like to take this opportunity to thank one and all. First, we would like to express our deep gratitude towards our internal guide **Dr. E. Poornima, Associate Professor,** Department of Computer Science and Engineering(Artificial Intelligence and Machine Learning), for her support in the completion of our dissertation. We wish to express our sincere thanks to **Dr. G. Karuna,** Head ofthe Department**,** and to our principal **Dr. J. PRAVEEN,** for providing the facilitiesto complete the dissertation. We would like to thank all our faculty and friends for their help and constructive criticism during the project period. Finally, we are very much indebted to our parents for their moral support and encouragement to achieve goals.

P. Mamatha(20241A6643)
M. Krishnamai(20241A6637)
Ch. Sruthi(20241A6655)
M. Tejaswini Reddy(21248A6602)

DECLARATION

We hereby declare that the mini project titled **" Detecting Early Phase Breast Cancer using Hybrid Machine Learning Algorithms "** is the work done during the period from 17th January 2023 to 12th June 2023 and is submitted in the partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning) from Gokaraju Rangaraju Institute of Engineering and Technology (Autonomous under Jawaharlal Nehru Technology University, Hyderabad). The results embodied in this project have not been submitted to any other University or Institution for the award of any degree or diploma.

P. Mamatha(20241A6643)
M. Krishnamai(20241A6637)
Ch. Sruthi(20241A6655)
M. Tejaswini Reddy(21248A6602)

ABSTRACT

Breast cancer is the most common cancer among women. It occurs when few breast cells begin to grow abnormally. The national average for 2022 is 100.4 cases per 1,00,000 people, with a large number of women being diagnosed with breast cancer. Our main motive is to design a prediction system that can predict breast cancer at early stages using a set of attributes that have been selected from a critical dataset. The Wisconsin Kaggle dataset is used for this experiment. This study is aimed to predict breast cancer using hybrid machine learning approaches, applying ml models like SVM and PCA.ML algorithms that could help to predict cancer, as the early detection of this disease would help to slow down the progression of other diseases. In our project, we are implementing Hybrid algorithms like PCA and SVM and optimizing SVM with k-fold cross-validation for predicting Breast cancer at early stages with high accuracy. The goal is to increase the proportion of breast cancer detection at early stages and to reduce error rate with maximum accuracy. The Hybrid model ,a combination of PCA with SVM got an accuracy of 96% and its performance is better when compared with traditional SVM.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

PCA	Principal Component Analysis
SVM	Support Vector Machine
WBCD	Wisconsin Breast Cancer Dataset
ML	Machine Learning
FPR	False Positive Rate
TPR	True Positive Rate

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Worldwide, breast cancer affects many women. The World Health Organization estimates that in 2020, millions of women worldwide received a breast cancer diagnosis, and 6,85,000of them passed away from the disease, accounting for about 15% of all cancer-related fatalities in females. When cancer is detected early on, amazing changes are observed. Therefore, there is always a chance that the patients will receive a diagnosis fairly soon. The distinct stages of cancer are brought on by cancer cells spreading throughout the tissue. Cancer is formed by abnormal growth of fatty tissues. Numerous techniques for making an accurate diagnosis of breast cancer have been suggested in light of this. Here, a dataset has been used to develop a fullyautomatic classification and prediction of breast cancer using a hybrid machine learning technique.

Here, we will use the Wisconsin Breast Cancer Dataset (WBCD) dataset from the machine learning repository to identify and categorize the two kinds of cells: benign (non-cancerous) and malignant (cancerous). If the cancer is discovered when it is in the benign stage, successful outcomes will be attained. If the tumor does not spread to other body tissues or organs, it is regarded as benign. If a tumor's cells are able to invade neighboring tissues, it is regarded as malignant. Therefore, in our project, we combined PCA and hybrid PCA to find usefulcomponents and reduce the dimensions of the data. Support Vector Machine (SVM), a model for early cancer diagnosis that may detect the disease at an early stage, was employed and provided accuracy of 96%.

The technology revealed a number of risk variables, including cell shape, normal nuclei, lump thickness, mitosis, bland chromatin, and bare nuclei. Therefore, it is crucial to use pre-processing techniques to clean up the data before identifying breast cancer. In order to improve accuracy, the data is typically pre-processed to remove the noisy, redundant data, and inaccuracies. A common machine learning strategy that can help to increase the accuracy of the classification model by removing redundant or pointless features is the use of PCA to find valuable sections of

the data and minimize its dimensions. SVM is a potent algorithm that is frequently employed for classification tasks and has demonstrated promising outcomes in the detection of breast cancer.

Pre-processing techniques are essential for cleaning up the data and removing noisy, redundant data and errors, to increase the classification model's accuracy. These pre-processing methods include outlier elimination, feature scaling, and data normalization. Other factors, such as age, family history, and exposure to environmental factors, may also affect the development of breast cancer in addition to the risk factors you stated. These elements must be taken into account while creating a breast cancer categorization model. Overall, the application of machine learning methods to the prediction of breast cancer is a promising area of research that has the potential to enhance the precision and effectiveness of diagnosis, resulting in better outcomes of the patient.

1.2 OBJECTIVE

The goal of applying a hybrid machine learning algorithm to identify breast cancer at an early phase is to increase the accuracy and efficiency of breast cancer diagnosis, resulting in earlier cancer identification and treatment with appropriate assessment. The Support Vector Machine technique was applied to both linear and non-linear classification in the method that was proposed. To make the issue less difficult, principal component analysis was used. Correctly predicting whether the cancer is benign or malignant is the objective here. The experiment averaged 10 times of k-fold cross validation to confirm the suggested approach. To easily apply new alternative ways with various datasets and pre-processing processes, new alternative methods must be discovered. Here, a model combined with hybrid machine learning algorithms like SVM is proposed to give the best accuracy. A potent method to accomplish this is to employ the Support Vector Machine algorithm for classification along with Principal Component Analysis to reduce complexity. A common way to check a model's performance and make sure it is reliable and not overfitting the training data is k-fold cross-validation. Multiple instances of k-fold cross-validation can give a more precise assessment of the model's performance.

For simple implementation with various datasets and pre-processing methods, it is crucial to find new alternative ways. This is because the caliber of the data and the pre-processing methods employed can have a significant impact on how well a machine learning system performs. Other machine learning algorithms besides SVM, like decision trees, random forests, and neural

networks have also demonstrated promising outcomes in the diagnosis of breast cancer. To find the best method for a certain dataset, it is crucial to investigate various algorithms and evaluate their performance. Overall, the proposed method of merging hybrid machine learning algorithms for breast cancer diagnosis is a promising field of research that could result in increased precision and effectiveness in the identification of breast cancer. To increase the precision andeffectiveness of breast cancer diagnosis, it is critical to continue researching various machine learning algorithms and methodologies.

These algorithm's effectiveness is strongly influenced by the caliber of the input data and the pre-processing methods employed. Therefore, before using any machine learning method, the data must be properly examined and prepped. Furthermore, the suggested strategy of fusing hybrid machine learning algorithms like SVM and PCA is a potent method that can result in a precise detection of breast cancer. To guarantee these algorithms are reliable and appropriate for various applications, it is crucial to assess their performance on various datasets. In general, using machine learning algorithms for early breast cancer diagnosis is a significant field of research that has the potential to save thousands of lives by identifying the disease early and delivering prompt care.

1.3 METHODOLOGY

The methodology of this project involves using PCA to reduce the dimensionality of the dataset, SVM as a classification algorithm and k-folds cross-validation to estimate the performance of the model.

A. **Support Vector Machine**: It is an essential tool for classification and regression in machine learning.

Malignant or benign tumors are classified using the SVM algorithm. SVM produces results that are accurate and precise. It is an effective classification and regression technique. The percentage of accurate predictions made by an SVM model is referred to as accuracy. The percentage of correctly predicted benign or malignant tumors in the context of tumor classification. On the other side, precision assesses how precise or accurate the model's optimistic forecasts are. Precision in the context of tumor categorization would mean how accurate the model is at identifying malignant tumors. SVM is a useful technique for data classification even when the data points cannot be separated linearly. The objective of SVM is to maximize the margin between the

hyperplane and the data points of each class, which is accomplished by creating a hyperplane that divides the data points into various classes. The ability of SVM to handle high-dimensional data makes it helpful for a variety of applications, including text classification, image classification, and many more.

B. Principal Component Analysis: This analysis is required to reduce the dimensionality of complex dimensions.

With the aid of PCA's dimensionality reduction algorithms, the data is pre-processed and cleaned. Large multidimensional datasets are condensed into two to three dimensions. The strongest features are found using a supervised linear dimensionality reduction approach based on a covariance matrix. This covariance matrix is then subjected to an eigen decomposition in order to identify the primary components. The data is translated onto this new set of axes using the principal components that have the highest eigen values.

Principal components that come from this analysis are the strongest patterns in the initial data and are orthogonal to one another. We can decrease the dimensionality of the data while keeping the majority of the information by choosing only the top k principal components, where k is typically much fewer than the initial number of features. In order to minimize the dimensionality of data while preserving the most significant patterns in the data, PCA is a linear transformation technique. This is accomplished by utilizing the covariance matrix of the initial features to identify the major components that account for the most volatility in the data.

C. K-Fold Cross-Validation: This technique is used in validating and evaluating the data of the model.

A model is developed to forecast and test for the cancerous and non-cancerous tumor after more data analysis. A small number of suitable algorithms are employed to test the accuracy. The test, which comprises of the next several phases, uses the 10-fold cross-validation. The model's accuracy is examined using 10-fold cross-validation. The data are divided into 10 equal folds using this procedure. The remaining fold is utilized for testing once the model has been trained on nine of the folds. Each fold serves as the test set once during the course of this 10 times procedure. For each iteration, the evaluation score—such as accuracy or precision—is calculated and averaged across all 10 folds. The steps are as follows:

- Divide the data into 10 equal sections.
- For each fold, use it as the test set and the remaining 9 folds as the training set.
- Train the model using the training set and evaluate it.
- Record the evaluation score for that iteration.
- Discard the model and repeat steps for each fold.
- Compute the average evaluation score across all 10 folds.

1.4 ARCHITECTURE DIAGRAM

The system's architecture offers a thorough diagnosis of breast cancer. The system as a whole is made up of 7 main building elements, which are: data collection, data pre-processing, feature extraction, model evaluation, model training, hybrid model, and the result.

Data collection: This block entails gathering information from a variety of sources, including medical records, cell size, cell shape, lumps, and more. The dataset under consideration was the Wisconsin Breast Cancer dataset, which had 32 attributes and 569 cases overall. Machine learning models are trained using these features to determine whether a tumor is malignant or not. The dataset is frequently used as a benchmark to assess how well breast cancer diagnosis systems perform.

Data pre-processing: This block involves cleaning and transforming the raw data into a format that can be used for analysis. This includes removing missing values, handling outliers, and standardizing the data. Additionally, Principal Component Analysis (PCA) is a commonly used technique in data pre-processing for dimensionality reduction. PCA Is an unsupervised linear dimensionality reduction technique that aims to find the strongest features in the data based on the covariance matrix. The PCA technique is used to reduce the dimensionality of the data by projecting it onto a lower-dimensional space while retaining the maximum amount of information. When working with high-dimensional datasets, this can be very helpful as it can lower the computational cost of the analysis and increase the precision of the machine learning models. With fewer dimensions and no information loss, the raw data is converted through various pre-processing procedures into a format that can be used for analysis and machine learning modelling.

Feature extraction: Finding the most crucial characteristics or variables from the pre-processed data that are crucial for determining whether a tumor is malignant or not is the goal of this block. This calls for the application of statistical methods like PCA.

Hybrid model: Support vector machine (SVM) and principal component analysis (PCA) approaches were combined with k-fold cross-validation with a k value of 10 to create a hybrid model for the breast cancer diagnosis system. Popular machine learning methods for classification and regression analysis include the SVM algorithm, especially when dealing with non-linear or highly dimensional data. To simplify the data and improve the analysis's accuracy, dimensionality reduction using the PCA method is applied. The hybrid model can increase the precision of cancer detection by combining these methods and employing k-fold cross-validation, which entails dividing the data into k subgroups, training the model on k-1 subsets, then validating it on the remaining subset. The k-fold cross-validation procedure is especially helpful in assessing the model's performance since it lowers the danger of overfitting and gives a more accurate estimate of the model's performance on brand-new, untested data. It entails combining the strengths of the top-performing machine learning models into a hybrid model to improve performance.

Model training: On the pre-processed data, this block trains the machine learning models to determine if a tumor is cancerous or not. The training of the model employs SVM. The hybrid machine learning model, which combined SVM and PCA approaches, was trained using the cleaned and previously processed data. The data were classified into benign or malignant tumors using the SVM algorithm, and the model was trained and tested using training splits of 0.25 and 0.75. The model learned from the training data during the training process and applied what it had learned to generate predictions on the testing data. 25% of the data was utilized to train the model, while the remaining 75% was used for testing and performance evaluation, according to a training split of 0.25 and testing split of 0.75. This enables the model to be trained on a smaller fraction of the data while maintaining its ability to correctly predict results on fresh, untainted data. It's also vital to remember that hyperparameter adjustment was done to enhance the model's performance throughout the model training phase. To achieve the optimum performance on the training data, this entails modifying the SVM algorithm's parameters.

Model evaluation: In this block, the effectiveness of the machine learning models that will be used to determine whether a tumor is cancerous or not will be evaluated. Metrics including accuracy, precision, recall, and F1-score are used in this.

Result: This phase entails giving the doctor or patient the final diagnosis on the malignant or non-cancerous status of a tumor. This entails utilizing a variety of visualization strategies to convey the findings in a straightforward and understandable manner.

1.5 ORGANIZATION OF THE PROJECT

The organization of this project consists of the structure, surveys conducted and the required architecture. The references along with the code and future enhancements are present here.

CHAPTER 1- Introduction: This chapter gave a succinct overview of the project before going on to discuss its correct purpose and methods. There was also provided an architecture diagram that illustrated the diagnosis process.

CHAPTER 2- Literature Survey: The summaries of 15 manuscripts (research articles) include a proper description, results, plus the benefits and downsides of each strategy, which are then listed.

CHAPTER 3- Proposed Method: The problem statement and project goal are presented in this chapter, which is then followed by information on the architecture diagram, the modules and their descriptions, and the software and hardware requirements (functional and non-functional). Additionally offered were analysis and design using UML (class diagrams, sequence diagrams, use case diagrams, and activity diagrams).

CHAPTER 4- Results and Discussion: This chapter includes a description of the dataset that was used, followed by a thorough explanation of the experimental findings and the importance of the suggested approach.

CHAPTER 5- Conclusion and Future Enhancements: A detailed summary of the project was given, including its goals, significance, chosen strategy, final results, and recommendations for improvement.

CHAPTER 6- Appendices: The sample code that was used in the tests was provided in this chapter.

CHAPTER 7- References: There were 25 references included, including a review of the literature, applications, a description, and a block diagram

Chapter 2

Literature Survey

This chapter includes a description and summary of current strategies, their advantages, results and their shortcomings.

2.1 Summary of Existing Approaches

To improve clustering accuracy and processing time, Hao Quan Lin and Zhengzhou J et al. [1] coupled conventional SOM neural networks with K-means methods. Using a hybrid technique of K-Means and SOM neural network algorithms in the medical field that can precise data sets, the goal was to detect breast cancer. This hybrid algorithm increases runningspeed and accuracy since it is unable to shorten run time while maintaining greater accuracy. The ideal number of clusters was determined using an elbow chart. The Wisconsin Breast CancerDataset (WBCD), which includes 569 cases and 32 patient features, was used for this experiment.The findings indicate that K-Means and SOM neural networks perform better in terms of accuracy(95%) and running speed (4.62), making them more suitable for breast cancer forecast.

The following are the drawbacks in [1].

In order to predict breast cancer, the study relies on a small dataset that may not be typical of the whole population. The paper fails to indicate the sample size or the number of variables used in the study.

The efficiency of the proposed hybrid algorithm against various breast cancer prediction algorithms is not evaluated in the study. This makes it challenging to determine how efficient the suggested algorithm is.

The hybrid K-means and SOM algorithms stages and criteria for choosing the best number of clusters are not explained in depth.

Sam Khozama and Ali M. Mayya [2] created a new range-based breast cancer prediction model by combining ensemble learning with the Bayes theorem. The goal was to forecast BC with a range of 0% to 100%. The confusion matrix was utilized in this study to evaluate several parameters. The BCSC dataset, which contains 67632 records and 13 risk variables, was employed in this method. The weighting method is used with this model as well. The ensemble model with 30 students is then trained using the revised BCSC data set. AdaBoost provides the peculiar ability to differentiate the cancer outcome as a percentage rather than a predetermined binary outcome. It is necessary to compute the updated BCSC data set. AdaBoost provides the peculiar ability to differentiate the

cancer outcome as a percentage rather than a predetermined binary outcome. It is necessary to compute the updated BCSC dataset. The study came to the conclusion that the dataset was utilized to develop an ensemble learning model using the Bayesian hyperparameter optimization technique and that the model had a high accuracy of 91.33%.

Limitations of [2] are as follows.

The lack of an in-depth description of the feature selection procedure used in the study makes it challenging to evaluate the quality of the features selected.

It is uncertain if the model went through independent dataset validation or if the small dataset utilized in the study led it to overfit.

For early BC identification, MUAWIA A. ELSADIG [3] recommended using machine learning (ML). Utilizing seven different classifiers to identify breast cancer at an early stage was the goal. MLP, RF, DT, NB, SVM, LR, and KNN are the seven classifiers. With an appropriate feature selection strategy that takes into account just the characteristics with high influence and ignores the others, an efficient classifier approach based on SVM was presented. The relief feature algorithm methodology was applied to this strategy. The confusion matrix and ROC were two of the several criteria used to assess how well the suggested strategy performed. The WBCD datasets were utilized, and the dataset was divided into three sets: set 1 contains 70% of the training data, set 2 contains 80% of the training data, and set 3 contains 90% of the training data. For each experiment, the random sampling technique is used 20 times, yielding accurate and realistic findings. The software called Orange Data Mining was used for all of the trials. The study found that, when compared to other classifiers, the SVM had the highest accuracy (97.3%). It is unclear whether the results were validated using cross-validation or an external dataset using appropriate validation techniques. There are no specifics on how the dataset was gathered, how the features were chosen, or how the SVM was implemented through the use of a support vector machine (SVM) with a feature selection method.

In order to predict the outcomes of breast cancer using various datasets, Ramik Rawal [4] examined four algorithms: SVM, Logistic Regression, Random Forest, and KNN. The goal was to distinguish between benign and malignant patients, and it was planned to parametrize classification methods to attain high accuracy. SVM, Logistic Regression, Random Forest, and KNN were the four algorithms utilized by the author for risk assessment, cancer recurrence prediction, and cancer survival rate prediction. The Wisconsin using cross-validation or an external dataset using appropriate validation

techniques. There are no specifics on how the dataset was gathered, how the features were chosen, or how the SVM was implemented through the use of a support vector machine (SVM) with a feature selection method.

In order to predict the outcomes of breast cancer using various datasets, Ramik Rawal [4] examined four algorithms: SVM, Logistic Regression, Random Forest, and KNN. The goal was to distinguish between benign and malignant patients, and it was planned to parametrize classification methods to attain high accuracy. SVM, Logistic Regression, Random Forest, and KNN were the four algorithms utilized by the author for risk assessment, cancer recurrence prediction, and cancer survival rate prediction. The Wisconsin breast cancer dataset, the Haberman's Survival dataset, the SEER database, and the Wisconsin Diagnostic Breast Cancer dataset were the four different datasets used for the research. SVM beat all other models (by 97%), according to the results. Future improvements are required for a number of reasons. Although SVM provides. Considerable limitation of SVM is, it was more productive and efficient in terms of accuracy but disappointed in predict malignant class precisely even with lesser FP rate at the time of SVM classifiers usage.

An automated technique that uses patient health records to forecast the likelihood of developing breast cancer was introduced by Shawni Dutta and her team [5]. The goal was to assess the viability of using prior medical records and estimate the likelihood of contracting malignant breast cancer. In this study, gradient boosting algorithms were utilized and compared against a variety of other algorithms, including Naive Bayes, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and AdaBoost classifiers, as well as Decision Tree (DT) and Random Forest (RF) classifiers. Accuracy, the f1-score, the mean square error, the Cohen-Kappa score, and Matthew's correlation coefficient (MCC) are used to gauge overall performance. In order to produce predictions, this article attempted to use the gradient boosting method on the Breast Cancer Wisconsin (Diagnostic) Data Set. The gradient-boosting technique has the lowest error rate and is quite effective. This study came to the conclusion that the gradient boosting method, which had an accuracy of 97.3%, performed better than all other algorithms and aided in the early detection of breast cancer. Limitation of this approach is, when they did cancer prediction with Gradient Boosting algorithm, it gave lowest error rate and high efficiency but, the drawback is the mammographic data is not sufficient to detect the small lobules of cancer cells.

The efficiency of classification algorithms like KNN, SVM, LR, and NB was compared by Gaurav Singh [6]. The goal was to assess the accuracy of various ML classification methods, such as k-nearest neighbor, support vector machine, logistic regression, and Gaussian Nave Bayes, to predict

breast cancer in its early stages. The dataset will be divided into training and testing phases, each with an allocation of 80% and 20% of the dataset, respectively. The UCI Machine Learning Repository's WBCD data, which was used in the experiment, was used; however, fewer samples were used for the training and testing phases. A larger dataset should be used as guidance for the evaluation of details in relation to the clinical environment. Even with the aforementioned restrictions, the results demonstrate that KNN outperforms the other algorithms, providing an accuracy of 99%, which is an impressive performance across the board. Each with an allocation of 80% and 20% of the dataset, respectively. The UCI Machine Learning Repository's WBCD data, which was used in the experiment, was used; however, fewer samples were used for the training and testing phases. A larger dataset should be used as guidance for the evaluation of details in relation to the clinical environment. Even with the aforementioned restrictions, the results demonstrate that KNN outperforms the other algorithms, providing an accuracy of 99%, which is an impressive performance across the board.

They must focus more on following in [6].

As they trained their KNN, SVM, LR, NB models with small datasets it leads to model under-fitting, which means those model gives wrong prediction for train data also. So, I think the assessment of details with reference to clinical atmosphere should be guided with a larger dataset.

For the exact discovery of cancer cases, Muhammad Umer [7] suggested an ensemble literacy-grounded voting classifier that combines the logistic regression and stochastic grade descent classifiers. The thing was to give a better frame that directly and precisely distinguishes between nasty and benign excrescences. This suggests an ensemble literacy- grounded voting classifier that combines logistic regression with largely sophisticated features. The effectiveness of ML and the suggested ensemble model were examined using CNN features. In this trial, 32 attributes from the bone Cancer Wisconsin dataset were used. In this case, the dataset was only gathered from one source. As a result, it's insolvable to draw conclusions about the results of multicenter exploration. The findings for the other variables, still, show that using a voting classifier with complicated features leads to the stylish bracket delicacy of 100.

Issue in [7], Algorithms are overfitted as they collected dataset just from a single origin. Due to this it isn't possible to conclude the issues with respect to multicenter exploration.

Using ML and DL algorithms, Krishna Mridha [8] examined a few widely used evaluation techniques for the early identification of breast cancer. The goal was to identify the algorithms with

the best and lowest accuracy. The grades for malignant and benign readings are 0 and 1. ML and DL algorithms are used, and their precision is compared. It focused on the number of models that were asserted on the used dataset. The WBCD dataset, which contains 31 features assessed via breast mass fine needle aspiration (FNA), was used in the trial. There are more ANNs that need to be trained for the parameters. The dynamic architecture and parameter refresh procedure result in a higher estimated cost and longer training period requirements. Results indicate that the KNN model has the least accuracy (91.22%), while the ANN model has the best accuracy (99.73%). With this precision, the female death rate can be reduced, and breast cancer can be detected early using ANN.

The drawback detected in [8] is described below.

The parameters that must be trained are more in number of ANN. Because of the dynamic architecture and parameter renew process, the estimation cost needed is more and duration required for training period also raises.

To produce accurate forecasts, Shudipti Rani Mondal and her team [9] compared several data mining methods employing categorization. There are several approaches that can be used to get a precise diagnosis of breast cancer. The goal was to improve diagnosis through the use of effective techniques like LR, SVC, KNN, RF, and DT. Hidden features are extracted using deep learning methods and numerical dataset machine learning methods. On the basis of tenfold cross-validation, the effective method was identified. In general, numerical datasets are taken into account, machine learning techniques are applied, and accuracy is anticipated; however, in this case, image datasets are also taken into account, and convolutional neural networks are utilized forthis. The experiment made use of the WBCD dataset, which was downloaded from the UCI machine learning library. Results show that, with a prediction rate of 99.2% for breast cancer, logistic regression offers greater accuracy.

The weak points of [9] are noted below.

1. To forecast and analyze cancer databases and improve accuracy, which is unclear and does not explicitly state the objectives of the study.
2. The application of machine learning algorithms for breast cancer diagnosis makes no mentionof other potential factors, like genetic predisposition or lifestyle decisions, that can affect outcome or treatment.

Machine learning and deep learning ideas have been applied by Apoorva V, Yogish H K, and Chayadevi ML [10] to diagnose breast cancer early and treat it effectively. The goal was to forecast

and study cancer databases in order to increase accuracy by utilizing convolutional neuralnetworks for an image dataset whose features are extracted from digitized images of breast mass.The K-Nearest Neighbour (KNN), Decision Tree (CART), Support Vector Machine (SVM), and Naive Bayes methods are utilized for the numerical dataset. A standard scalar is also used to scaleand center the data to improve accuracy. Large datasets with numerous illness classifications are not used to test the parameters. SVM is taken into consideration for additional analysis and prediction because the results demonstrate that it provides high accuracy (96%). CNN, which was built using a sequential API, gets the results. The predicted image data was cancerous.

The flaws of [10] are outlined accordingly.

1. They contemplated only parameters that are in existence, no new parameters were established and cross-examined. Higher accuracy might be obtained if they had applied these algorithms using new parameters on bigger datasets with a greater number of disease classes.

2. Without elaborating on the variables that may have influenced the outcomes or their therapeutic importance, the abstract solely presents the performance metrics of each algorithm.

On the Breast Cancer Wisconsin Diagnostic dataset, Mohammed Amine Najia* [11] and his team have implemented five machine learning (ML) algorithms for early detection and prediction. The goal was to use machine learning algorithms to predict and diagnose breast cancer. Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (c4.5), and K-Nearest Neighbour (KNN) are the five ML techniques that the authors utilized. Accuracy, precision, and the confusion matrix were used to gauge the algorithms' performance. The Wisconsin Diagnostic dataset, which was employed in this study, comprised a total of 569 occurrences (benign: 357; malignant: 212). The results of the prediction were more sensitive, accurate, and specific thanks to genetic algorithms (GA). Results demonstrate that SVM, with an accuracy of 97.2%, is more effective than all other methods. This method relies on chosen features; hence, it might not deliver the desired results if the chosen attributes are not pertinent. Even though the Wisconsin dataset demonstrates high accuracy, the paper may need to be tested with other datasets to produce similarly impressive results. The only drawback is, they worked with KNN, RBF, DT,(ANN), FNN, PNN, and pattern recognition, the work was done with retrospective-single center dataset which was an imbalanced dataset. So, in the end they suggested to utilize a dataset which consists bigger sample proportions in multicenter. Like how others had drawbacks,[11] is not usedprimary real-time hospital data to enhance the precision of BC.

A prediction model for breast cancer was created by Mohammad R. Afrash and his team [12], utilizing genetic algorithms (GA) in conjunction with a number of machine learning (ML) techniques. Predicting breast cancer at an early stage was the goal. In order to forecast breast cancer based on demographics, history, and lifestyle factors, GA was integrated with a number of algorithms, including KNN, RBF, DT, ANN, FNN, PNN, and pattern recognition. The studies are carried out in a concurrent setting using the Anaconda Python platform and associated libraries.

The authors' patient case record data were gathered from the Breast Cancer Registry Database at the Ayatollah Taleghani Hospital. The suggested method reduces human error in the diagnostic process and lowers the cost of cancer diagnosis, but it does not leverage the primary real-time hospital data to increase BC accuracy. According to the experimental findings, the DT performs well (99% correct) compared to other ML models.

Identified limitations in [12] are summarized as follows.

1. The precise research question that is being explored isn't explained in full. The task at hand is not explicitly stated in the abstract, but it does emphasize the necessity for an improved technique.

2. The study's restrictions or its relevance to additional scenarios are not mentioned in the abstract. It is crucial to explore the application of the findings to various contexts and to understand any potential research flaws.

In her research proposal, Dr. S. Sridevi [13] combined a deep neural network model with customized hyper-parameters, validated to produce the required accuracy. The performance of the suggested hybrid prediction model was compared to that of various older techniques like SVM and regression. The data of the mammography images were compiled using the Kaggle mammography image repository, and PCA techniques were employed for feature selection. Here,the Breast Cancer dataset's 699 instances and 11 attributes are included in the UCI Machine Learning repository. In order to achieve high efficiency with high accuracy, these datasets are tested on classifiers. Results from appropriate models, including MCP, SVM, and KNN, were compared. As a result, this model, which uses the SVM-RBF kernel and has well-organized validations and algorithms, produces the maximum accuracy. In Wisconsin datasets, it has an accuracy rating of 96.8%.

The disadvantage of [13] are summed up as follows.

Although it is stated in the paper's abstract that the significance of classifying people with cancer

into high or low risk clusters has resulted in the introduction of machine learning techniques, it lacks further information on how this categorization is crucial or the way it is currently carried out without machine learning.

To improve breast cancer (BC) detection, Sarthak Vyas and Abhinav Chauhan [14] examined three different algorithms utilizing supervised learning approaches. In order to diagnose breast cancer effectively and accurately, the goal was to predict breast cancer using decision trees and KNN. Artificial neural networks and support vector machines were used as machine learning algorithms. The suggested model uses alternative methods that are simple to implement with various datasets, safe, trustworthy, and economical. 32 distinguishing factors from the Wisconsin breast cancer dataset helped reduce the muti-dimensionally big dataset. Results indicate that when compared to other methods, support vector machines offered the highest accuracy, at 92.7%. The accuracy is 92.7%; however, if the grouping techniques are improved, this accuracy could increase.

The problems with the implementation of [14] are outlined below.

1. About 86% of all diagnostics are accurate, yet the source of this number or the standards used to assess whether a diagnosis is accurate are not mentioned. It is so challenging to determine the significance of the suggested model for improving diagnosis accuracy.

2. The abstract does not go into depth regarding how the suggested approach satisfies these requirements or how it differs from currently used techniques. This calls into doubt the proposed model's generalizability and practical relevance to different contexts or groups.

To improve the precision of earlier prediction models and address issues with breast cancer prediction, P Sivakumar [15] presented the Majority-Voting Based Hybrid Classifier (MBHC). There were five different machine learning techniques used: decision tree, random forest, SVM, and MBHC. The classification algorithms are well-trained using the historical data, making it possible to forecast new patterns from the recently obtained data. The experimental findings demonstrated that, when compared to other cancer prediction systems, the system using MBHC provided an accuracy of 79.2%. The University of Wisconsin Hospitals Madison Breast Cancer Database was used as the source of the medical data set. In the future, this model can be used with a real-world dataset. The Breast Cancer Prediction System experiment findings showed a respectable accuracy score of 79% and used a hybrid classifier based on a majority vote.

The imperfections of [15] are listed below.

1. The dataset implemented for breast cancer prognosis is described in the abstract as consisting

of just regarding the image attributes, which raises the possibility that the research may be restricted to a specific kind of data.

2. This highlight concerns regarding the suggested hybrid classifier's generalizability and practical applicability to various kinds of breast cancer data.

3. Although the suggested majority-voting-based hybrid classifier has a precision rating of 79%, the abstract fails to compare this result to the precision of existing prediction models or algorithms.

4. Due to this, it is challenging to comprehend the significance of the provided outcome and assess the success of the suggested solution in comparison to current practices.

Table 1. Summary of existing approaches

Ref. No.	TITLE	ALGORITHM	ADVANTAGES	ACCURACY
1.	Breast Cancer Prediction Based on K-Means and SOM Hybrid Algorithm	K-Means and SOM neural network	Adds productive help guidance and additional applications	95%
2.	A New Range-based Breast cancer Prediction Model Using the Bayes Theorem and Ensemble Learning	AdaBoost	AdaBoost has an eccentric characteristic of distinguishing the cancer result as percentage rather than a settled binary result	91.33%
3.	A machine learning approach for breast cancer early detection	MLP, RF, DT, NB, SVM, LR, and KNN	Random sampling, technique is repeated 20 times for each experiment and attained reliable and realistic results	SVM-97.4%

4.	Breast cancer prediction using Machine learning	SVM, LR, RF, and KNN	The experiments are conducted within a simulation environment and using JUPYTER platform, ensuring the results are reliable and replicable	SVM-97%
5.	Early Breast Cancer Prediction using Artificial Intelligence Methods	Gradient Boosting algorithm	Gradient Boosting gives least failure rate and great efficiency	97.3%
6.	Breast cancer prediction using machine learning	KNN, SVM, LR, NB	The breast cancer dataset is publicly available and partitioned into 80% training and 20% testing to ensure reliable and accurate results	KNN-99%
7.	Breast Cancer Detection Using Convoluted Features and Ensemble Machine Learning Algorithms	ensemble learning-based voting classifier	When voting Classifier together with convoluted features are used greater classification accuracy is accomplished	100%
8.	Early prediction of breast cancer by using Artificial neural network and ML techniques	ML and DL	Machine learning algorithms can reduce the Limitations of early-stage breast cancer diagnosis	ANN-99.73%

9.	Breast Cancer Prediction Using Machine Learning Techniques	KNN, CART, SVM, NB	Along with the numerical dataset, Image dataset is used by the convolution neural Network	SVM-96%
10.	Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis	SVM, KNN, LR, DT, RF	The task is done in Anaconda using python and Scikit-learn, making it easy to replicate and build upon	SVM-97.2%
11.	Developing breast cancer risk prediction system using hybrid machine learning algorithms	KNN, RBF, DT, (ANN), FNN, PNN, pattern recognition	Genetic Algorithm enhanced sensitivity along with specificity of prognosis results	DT-99%
12.	Breast cancer prediction with hybrid ML models	SVM-RBF	Reduces human errors in the diagnosis procedure	96.8%
13.	Prediction of breast cancer using machine learning	SVC, RF, LR, DT, KNN	Datasets can be tested on various classifiers to bring great efficiency	LR-99.2%
14.	Breast Cancer Detection Using Machine Learning Techniques	DT, ANN, KNN, SVM	The model achieved by integrating ANN, KNN, and DT, etc.is worthwhile and safer	SVM-92.7%

15.	Breast Cancer Prediction System: Novel approach to predict the accuracy using Majority-Voting Based Hybrid Classifier (MBHC)	LR, DT, RF, SVM, MBHC	This paper proposed a novel MBHC to enhance the accuracy of various prediction models for Breast Cancer	MBHC-79.2%

In conclusion, the literature survey on breast cancer early detection provides an essential overviewof the various methods and technologies used for early detection of breast cancer. It can offer valuable insights on the advantages and limitations of different approaches, and help inthe development of new and more effective methods of early detection.

CHAPTER 3

PROPOSED METHOD

3.1 PROBLEM STATEMENT

The following lines says about the problem which we are considered in our project. Nowadays the mortality rate due to breast cancer increasing rapidly due to inadequate knowledge of predicting the disease at its early stage. The cancerous cell forms a small duct in the initial stages at this stage it is called carcinoma in situ also called stage 0 in this stage cancer cells grow abnormally in one place and these cells become more cancerous and then spreads to nearby tissues and then spreads to the entire body. If cancerous cells in an in-situ stage are not recognized by people at their early stage then it may lead to hazardous cancer spreading throughout the entire body. The worldwide Organization for health (WHO) says that the number of women who received breast cancer diagnosis in 2020 was 2.3 million among them 685000 women died having do with breast cancer. And 7.8 million women were alive at the end of 2020 who were discovered to have breast cancer. This result says how rapidly the disease isgrowing and also suggests that there is a need for predicting the disease at its early stages. Breast cancer can occur at any age irrespective of gender.

The studies say that the occurrence of breast cancer in rural areas is more when compared to urban areas. Some studies say that women with low vitamin D also were targeted by breast cancer. So, there is a need of encouraging people to take care of their health. By understanding many realistic situations and observing the cases we came to know that the death rate or mortality rate is due to not identifying disease at its early stage. The problem that this paper is dealing with, is predicting, also reducing the mortality rate of many people.

OBJECTIVES

- This project's goal is to make breast cancer predictions. at its early stages using a combination of hybrid machine-learning methods.
- Project aims in selecting the best algorithms and then converting to hybrid algorithms for breast cancer forecasting at the initial stage.
- To achieve prediction of breast cancer at its early stages we have used two algorithms one from supervised machine learning technique the other from unsupervised machinelearning technique.

- The initial stage of the breast cancer in this paper is predicted with help of accuracy obtainedby classification algorithms.

3.2 EXPLANATION

A. ARCHITECTURE DIAGRAM

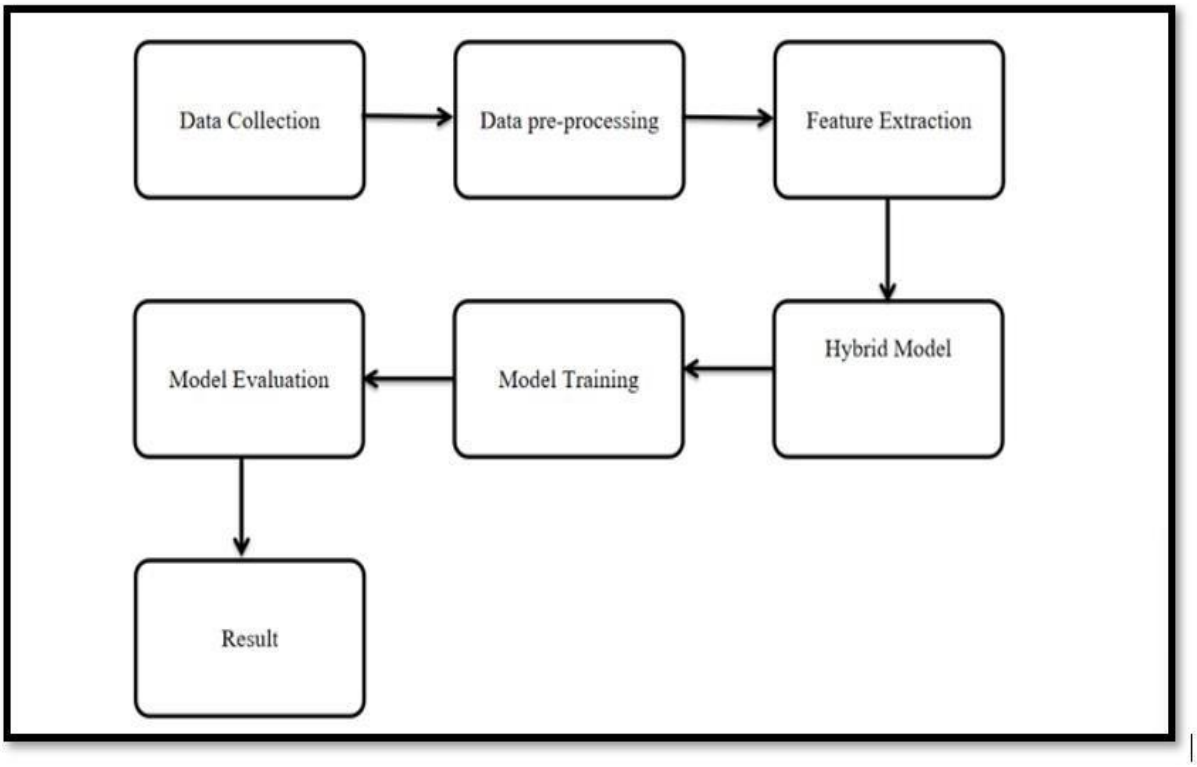

Figure 1: Architecture Diagram

MODULE 1: DATA COLLECTION

The dataset we have collected is from the Wisconsin Breast Cancer Dataset from Kaggle consists of a total of 569 instances and 32 attributes. The dataset shows the presence of benign and malignant diseases indicated with the letter 'B' and 'M' respectively. The dataset also contains some attributeslike texture, area, radius, perimeter, concave points, concavity, compactness and etc. All these qualities help in identifying the radius, area, texture, and perimeter of cancer cells and also help us in identifying which stage the disease is the dataset also provides a unique identification numberto identify the place where it is showing which cancer stage. The dataset collected shows the casesof benign and malignant. There are a total of 357 benign variants and a total of 212 malignant

variants. So even with this, there are also some extra attributes are also provided which are not very important in predicting the disease. Further, the process continues by following data preprocess which is one of key phase and one important step in this paper. So, this processgives a detailed explanation for data collection now let us see about data preprocessing and furthersteps given below with a detailed explanation.

MODULE 2: DATA PRE-PROCESSING

Data processing considered to be a part of data preparation. It is considered to be a crucial phase in the data mining process. Sometimes the data we have got can contain many unwanted or irrelevant data so it is better to remove them before applying other techniques to the data so that we can get efficient output. Data preparation alters the data's format so that is very easy is work on that data. In any project data preprocessing is important because it results in reliable, precise, and robust outputs.

The data is stored in Excel format and the entire experiment is done in a googlecolab environment the steps involved in preprocessing are also done in the googlecolab environment. In data preprocessing, we have removed unnamed and irrelevant data with the help of Python programing language. For this data processing we have used PCA (principal component analysis) is a technique for reducing dimensionality. Here in data preprocessing some of the unnamed data is dropped because of redundancy and it may lead to less usefulness in further processes.

MODULE 3: FEATURE EXTRACTION

Feature extraction is a process of reducing dimensions through which initial raw data can be converted to a different set of manageable groups through which further process continues. Feature Extraction is helpful in improving the performance of classification algorithms and helps in obtaining effective and accurate output. The Wisconsin dataset consists of a total of 32 attributes which also includes many unwanted features with the help of the googlecolab environment and using machine learning methods some of necessary features are selected using PCA (principal component analysis) algorithms and further process are continued on the selected relevant features. After feature extraction, the attribute size was reduced from 32 to 11 which showsthat the PCA was very much helpful in dimensionality reduction. After performing feature extraction, the data goes on to the next step.

MODULE 4: HYBRID MODEL

A hybrid model occurs when we integrate one algorithm with another and sometimes a hybrid model can be constructed by giving outputs of one algorithm as inputs to another and optimizing the resultant algorithms. We have constructed a hybrid algorithm taking help of PCA and SVM (Support Vector Machine) and optimized PCA-SVM with cross-validation. Cross-validation is used by using training models, machine learning models can be evaluated on a subset of raw data. We have applied cross-validation on PCA_SVM with the help of k-folds and here we have used the k valueas 10. We have noticed that with the help of cross-validation, the algorithm produced accurate andeffective outputs which were not provided initially and this hybrid model is measured with the help of metrices like recall, f1 score, accuracy, and precision. The results obtained are very relevant to express how our hybrid model works. So, this process says about the how we are goingto follow hybrid algorithms in our project.

MODULE 5: MODEL TRAINING

Model training is the step where the data is trained by machine learning models and the trained model results in giving better performance results. The training is crucial to keep in mind since the training of raw data can be done only when providing high-quality data. Model training should be applied once you have selected relevant features. If the data provided was not cleaned properly then the trained model may not have resulted in giving good outputs which are not up to our expectations. Here wehave removed all unnecessary data and then given the data is trained with a machine learning model like SVM (support vector machine) using a value of 0.25 for training and 0.75 for testing.

MODULE 6: MODEL EVALUATION

When evaluating a model, assessment metrics are helpful for understanding how well it performs and for identifying the advantages and disadvantages of the machine learning model that we have employed. We used a confusion matrix to test our hybrid model. let us have a brief intro to the confusion matrix

- True Positive: It says that the number of cases predicted to be correct and the actual output was also correct

- True Negative: It says that the number of cases was predicted to be wrong but the actual output was correct.
- False Positive: It says that the number of cases predicted to be correct but the actual output was wrong.
- False Negative: It says that the number of cases was predicted to be wrong but the actual output was correct.

We have also used cross-validation for model evaluation. let us see a brief note on how this cross-validation works:

- Step 1: Firstly, we divide data into equal-sized data points and they have grouped these points called folds.
- Step 2: Then we train our data excluding 1-fold.
- Step 3: we are going to test our data.
- Step 4: Repeat steps 2 and step 3 for all the folds.
- Step 5: To estimate the capability of the algorithm the performance is measured across all these folds.

So, we have evaluated our model by following these evaluation techniques for hybrid model evaluation and further processes.

MODULE 7: RESULT

The final stage of any project ends with knowing the results. To know how accurate our hybrid model is we have compared our hybrid model with traditional SVM which does not include PCA. The result of the accuracy provided by traditional SVM was 95% while the accuracy provided by our hybrid model made of PCA and SVM with k-folds provided an accuracy of 96%. Hence, it is clear that our proposed hybrid model works more accurately when compared to traditional SVM. This process says about complete and detail description about result of our project so that we can say these words are reflects the results of our project in a very well-defined way.

B. MODULES-CONNECTIVITY DIAGRAM

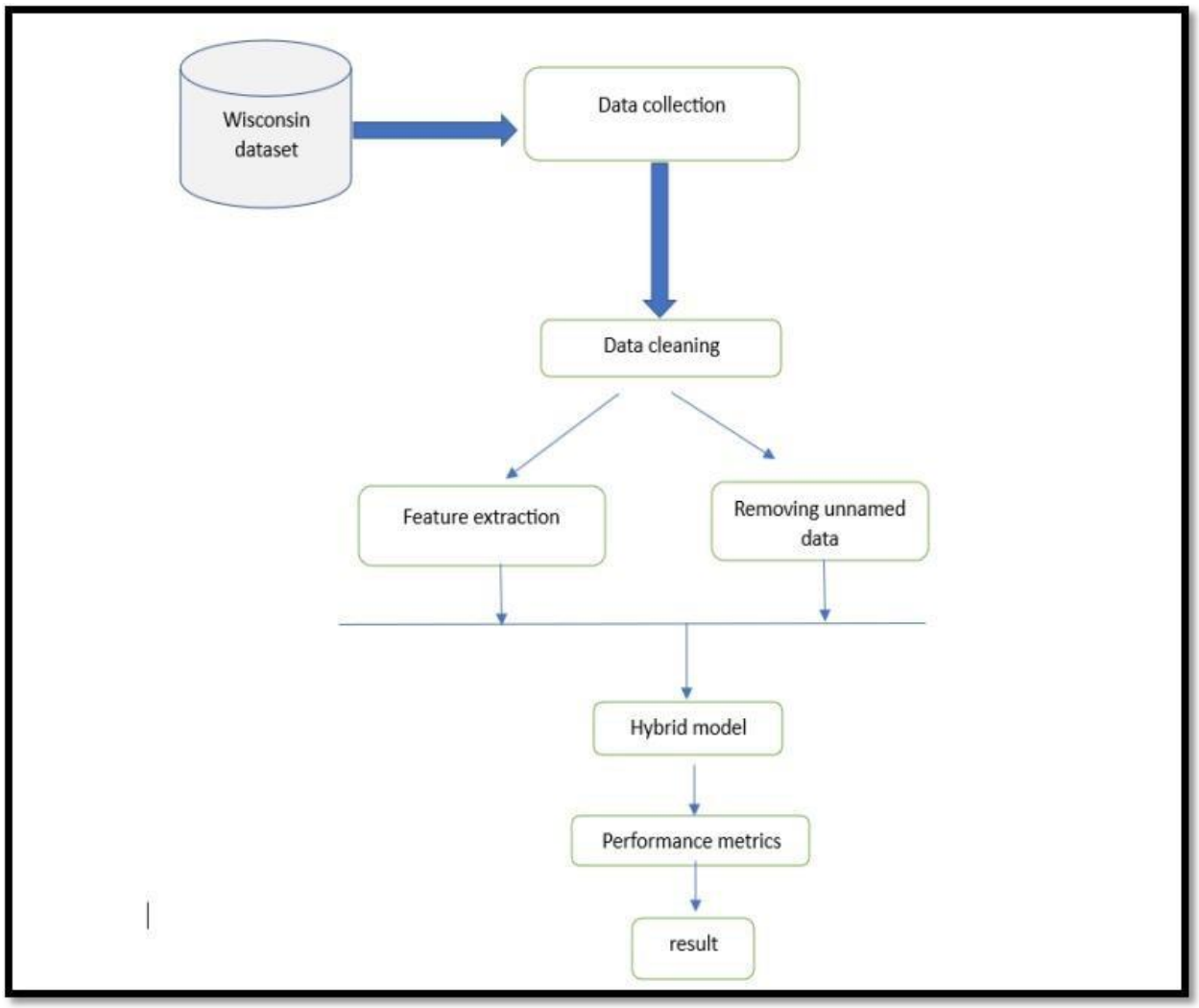

Figure 2: Module Connectivity Diagram

This figure shows how different modules are connected to one another it also says that any project needed to incorporate a proper plan before going to implement the project. So, to implement our project successfully we have followed the modules shown in the figure. The key start in this phase is to collect data this data relates to breast cancer so we have to gather the data related to breast cancer in this phase here we have gathered the Wisconsin breast cancer dataset. Then coming to the second step which is data collection even though the data is already stored in Wisconsin dataset, we have to collect the data in our own system to know the breast cancer cases and to detect breast cancer at an initial stage.

Then comes the third step which is data cleaning, this phase or step says that the collected data gets cleaned here and it is a very crucial step in any project to go further in applying their further applications because when the data is not cleaned properly then is a chance of not getting expected results. So, it is important and necessary to clean data. In our project, we have used PCA (principal component analysis) for data cleaning purposes because PCA is used for removing unnecessary data and also helps in reducing dimensions which makes our work simple for going to further processes.

After applying PCA then the process goes to feature extraction here we are going to extract only the necessary feature to which our project is more concerned and at the same time we will also check whether is there any unnamed data or redundant data. If we found any unnamed or redundant data then we have to drop them. After cleaning redundant and unnamed data, the data is provided to the hybrid model. In our project, we supplied PCA processed data to SVM (support vector machine) then we optimized SVM by using cross-validation and used the k value as 10.

The next step in this process is the performance matrix we have used different measures for example, recall, precision, f1-score, and accuracy to estimate the performance of optimized SVM. All performance measures expressed above provided a very satisfactory result that all have crossed 90%.

The next step in our module connectivity is results we have expressed the results with the help of ROC (receiver operating characteristic curve). The ROC curve presents graphical representation between optimized SVM and SVM without PCA and it clearly says that the optimized SVM is quite better when compared to normal SVM.

The above explanation says a detailed description of how our modules are connected and how each module passes the data after doing the process in their phase and then proceeds to pass the data to the next module without proper connectivity between modules it is expected to be difficult for getting the accurate results and with the help of proper planning of modules we have successfully reached to the end of each module without any trouble.

C. SOFTWARE AND HARDWARE REQUIREMENTS

C.1 SOFTWARE REQUIREMENTS

- Ms. Excel: MS Excel is one of the software requirements for storing the dataset we have downloaded from Kaggle and it is helpful in storing the data in a table format so that we can get a clear count of instances and attributes available with us in our dataset.
- We required software for implementing our whole code so we used googlecolab environment for our project implementation it is also an open-source software for various applications the below images show the software we have used.
- Here we have used colab because it can be accessed anywhere and is helpful when there are no laptops with us and we can show our results only by entering our mail id so it acts like a cloud.

Figure 3: Google Colab

C.2 HARDWARE REQUIREMENTS

The entire project processed only when we have the necessary hardware is available for running software, we need hardware and hardware is used for storing the data which also helps in executing our project and this all can be done when we have the required hardware facility.

The hardware requirements are:

- Operating system: windows

- Preprocessor: Intel core i5
- RAM (Random access memory): 4GB

3.3 MODULES AND ITS DESCRIPTION

This breast cancer diagnosis system mainly consists of 6 modules, which describe the functioning and procedure throughout the process of the model.

MODULE 1: DATA COLLECTION

Data is collected from various sources such as medical records, cell size, cell shape, and lumps. The dataset used in this particular case is the Wisconsin Breast Cancer dataset, which is available on Kaggle repository. This dataset consists of 569 instances and 32 attributes. The attributes include features such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, among others. These attributes are used to train machine learning models to predict whether a tumor is cancerous or non-cancerous. The dataset is a commonly used benchmark for enhancing the performance of breast cancer diagnosis system.

The WBCD dataset is a well-known dataset used in machine learning and data analysis for classifying breast cancer cells as either benign (B) or malignant (M). The dataset consists of features extracted from digitalized images of Fine of breast mass, including cell shape, cell size, and other attributes. The dataset contains 569 instances, with 357 instances labelled as benign and 212 instances labelled as malignant. To collect data using the WBCD dataset, you would need to first obtain the dataset, which is publicly available online. Onceyou have obtained the dataset, you can load it into a data analysis tool or programming language such as Python using a library like Pandas. To analyze the data, looking at summary statistics for the features, such as mean, standard deviation, and range, to gain an understanding of the distribution and variability of the data is essential. Visualizations can also create, such as histograms or box plots, to explore the data further.

To build a classification model using the WBCD dataset, you can use a machine learning algorithm such as logistic regression or decision tree classification. Before building the model, splitting the data into training and testing sets to evaluate the performance of the model is a necessity. Metrics such as accuracy, precision, recall, and F1-score to evaluate the model's performance can also be used. In conclusion, data collection using the WBCD dataset involves obtaining the dataset,

analyzing the features using statistical methods and visualizations, and building a classification model using a machine learning algorithm. The goal of this process is to accurately classify breast cancer cells as benign or malignant based on the features extracted from the FNA images.

MODULE 2: DATA PRE-PROCESSING

Data pre-processing for the breast cancer diagnosis system involves cleaning and transforming the raw data, handling missing values and outliers, scaling the features, selecting relevant features, and addressing class imbalance. These steps are critical in preparing the data for analysis and machine learning modelling, and can significantly impact the performance of the system. PCA is a widely used unsupervised dimensionality reduction technique that is commonly used in data pre-processing to transform high-dimensional data into a lower-dimensional space while retaining most of the relevant information.

PCA works by identifying the underlying structure in the data through linear combinations of the original features that capture the maximum amount of variance in the data. The principal components generated by PCA represent new dimensions that are a linear combination of the original features. The first principal component captures the maximum amount of variance in the data, and each subsequent component captures the maximum variance orthogonal to the previous components. By retaining the top principal components that capture the most variance, PCA reduces the dimensionality of the data, which can help to simplify the analysis and reduce computational complexity. In addition to reducing the dimensionality of the data, PCA can also be used for feature extraction, data visualization, and data compression. PCA is widely used in machine learning applications such as image recognition, natural language processing, and gene expression analysis, where the dimensionality of the data can be very high.

MODULE 3: FEATURE EXTRACTION

Feature extraction is an important step in machine learning, especially in cases where the dataset contains a large number of features or variables. PCA (Principal Component Analysis) is one of the most commonly used techniques for feature extraction.

In the case of the Wisconsin dataset, there are 32 attributes, which can make it challenging to identify the most relevant features or variables for predicting whether a tumor is cancerous or non-cancerous. Feature extraction techniques such as PCA can help to address this challenge. PCA can be applied to the 32 attributes of the Wisconsin dataset to identify the principal components that

capture the most significant variation in the data. By retaining the most important principal components, dimensionality of the dataset can be reduced while retaining most of the information needed for accurate tumor classification.

Once the most relevant features have been identified and can be used to build a predictive model such as logistic regression, decision tree, or support vector machine (SVM). The model can then be trained on the pre-processed and feature-selected dataset to predict whether a tumor is cancerous or non-cancerous based on the selected features.

MODULE 4: HYBRID MODEL

In the breast cancer diagnosis system, a hybrid model was built using a combination of support vector machine (SVM) and principal component analysis (PCA) techniques, along with k-fold cross-validation with a k value of 10. The SVM algorithm is a well-known machine learning techniqueused for classification and regression analysis, particularly in cases where the data is non-linear orhigh-dimensional.

The PCA technique is used for dimensionality reduction to reduce the complexity of the data and enhance the accuracy of the analysis. By combining these techniques and using k-fold cross validation, which involves splitting the data into k subsets, training the model on k-1 subsets, and validating it on the remaining subset, the hybrid model is able to improve the accuracy of the diagnosis of breast cancer. The k-fold cross-validation technique is particularly useful in evaluating the performance of the model, as it helps to reduce the risk of over fitting and provides a more reliable estimate of the model's performance on new, unseen data. It involves integrating the best-performing machine learning models into a hybrid model that combines their strengths to achieve better performance.

MODULE 5: MODEL TRAINING

Model training is a crucial step in the machine learning pipeline, where the model is trained on pre-processed data to predict whether a tumor is cancerous or non-cancerous. In the breast cancer diagnosis system, SVM was used in the model training, and a hybrid machine learning model was built by combining SVM with PCA techniques. The cleaned and pre-processed data was used to train the model, and a training split of 0.25 and testing split of 0.75 were used to train and evaluate the model. During the training process, the model learned from the training data and used that knowledge to make predictions on the testing data.

By using a training split of 0.25 and testing split of 0.75, we can train the model on a smaller subset of the data while still being able to accurately predict the outcomes on new, unseen data. Overall, the model training block is essential in the machine learning pipeline, as it allows the model to learn from the data and make accurate predictions on unseen data. The use of a hybrid machine learning model further improved the accuracy of the predictions.

MODULE 6: RESULT

The final stage involves presenting the diagnosis to the clinician or patient using various visualization techniques to make the results clear and understandable. The model has used support vector machines (SVM) and k-folds cross-validation with a k value of 10 to achieve an accuracy of 96%. It is a technique used to reduce the dimensionality of the data and can often improve the performance of machine learning models, but it's not always necessary or beneficial in all cases. Earlier, this model was compared with a traditional SVM based model which did not include PCA. It was clearly observed that after the involvement of PCA in combination with SVM, the new hybrid model gave better results.

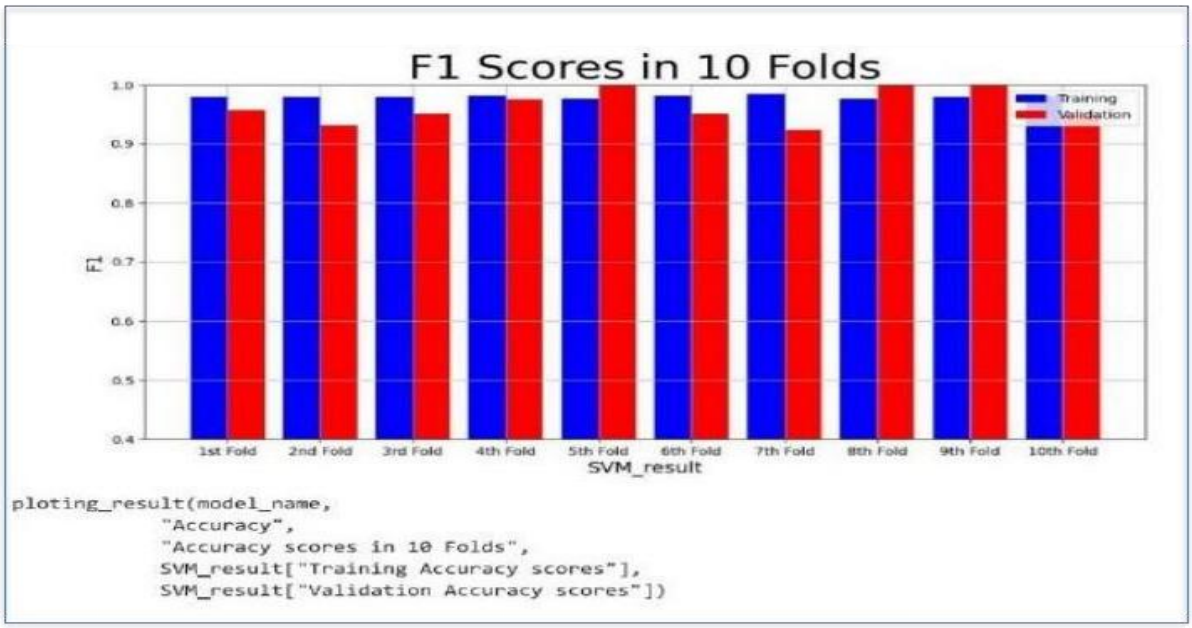

Figure 4: Result

While 96% accuracy is high, it's important to remember that machine learning models should be used as a tool to guide medical professionals in making diagnoses and treatment decisions, rather than a replacement for clinical judgment. Additionally, it's important to consider

other factors such as sensitivity, specificity, and potential biases when evaluating the performance of a diagnostic model.

3.4 REQUIREMENTS ENGINEERING

3.4.1 FUNCTIONAL REQUIREMENTS

1. **DATA COLLECTION:** In order to gather relevant details about breast cancer patients and train and comprehend hidden patterns in our model, we have to search data in various sites.

2. **DATA PREPROCESSING**: When collecting data, we obtain it in its raw form, which is unsuitable for training our model because it contains redundant data, different features are scaled differently, and many features have empty values. Redundancy and missing values must be handled via preprocessing techniques in order to prepare our data for model training.

3. **FEATURE SELECTION**: Preprocessing by itself cannot produce accurate results because adding extra features to our model during training increases computing complexity and decreases model performance. We require feature selection in order to simply retrieve pertinent features.

4. **MODEL TRAINING:** In order to predict the patient's health state, the model needs to be trained to produce as accurate results as feasible.

5. **REAL-TIME APPLICATION:** When patients get tested and come up with their reports, they have to manually enter readings from the report into the model. Then, our pretrained model must detect breast cancer.

NON-FUNCTIONAL REQUIREMENTS

- **Performance:** The system should be able to maintain tremendous quantities of data and provide predictions quickly.
- **Scalability:** The system must be able to handle rising user and data volumes.
- **Accuracy:** The system's predictions should be exact and definite.
- **Security:** To safeguard private victim data, the system must be designed with assurance.
- **Portability:** The system must be able to function on a variety of hardware platforms and operating systems.

3.5 ANALYSIS AND DESIGN THROUGH UML

Making a diagram or representation of any complex system is the best approach to understand it. These examples more strongly influence our understanding. Diagrams are not a novel idea, but they are utilized frequently in a variety of ways across a wide range of industries, as can be seen if we take a closer look.

We generate UML diagrams to better and more thoroughly understand the system. Every aspect of the system cannot be represented in a single diagram. UML offers different diagram kinds to account for the majority of a system's components. To satisfy your needs you can also develop your own set ofdiagrams.

3.5.1 CLASS DIAGRAM

Class diagrams are the most common UML diagrams. A class diagram includes all of the following: classes, interfaces, affiliations, and cooperation. Class diagrams effectively show the static object-oriented representation of a system.

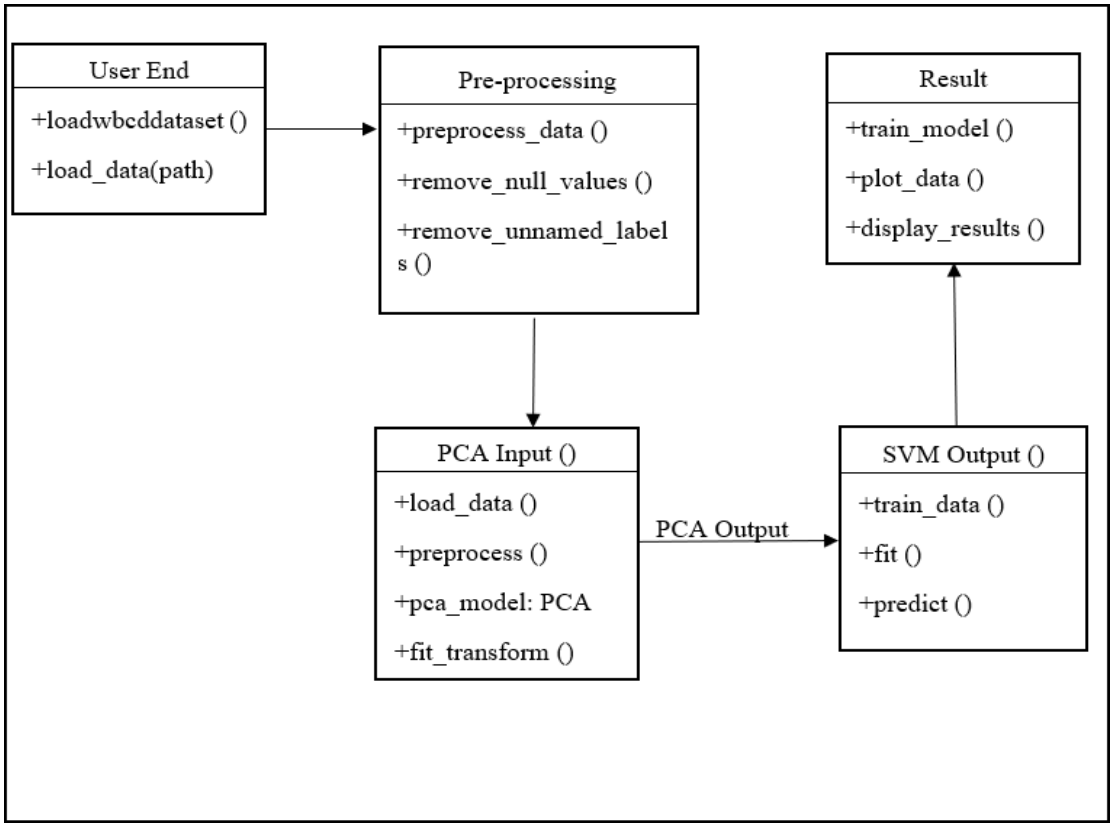

Figure 5: Class Diagram

In Figure 5, the class diagram for this project is displayed. Two classes—User End and Pre-processing make up the class diagram. There are two methods in the User End class, and they are the user can load the WBCD Dataset as an input to the system method, using the load wbcd dataset and load data(path) methods. In order to reduce noise, the pre-processing eliminates null values and unidentified labels and the dataset's dimensions are now reduced using PCA, and the output of the PCA model is fed into the SVM model to predict cancer.

3.5.2 SEQUENCE DIAGRAM

Sequence diagrams are interchangeable terms. The diagram's name makes it plain that it is concerned with some sequences, which are the transmission of messages from one thing to another.

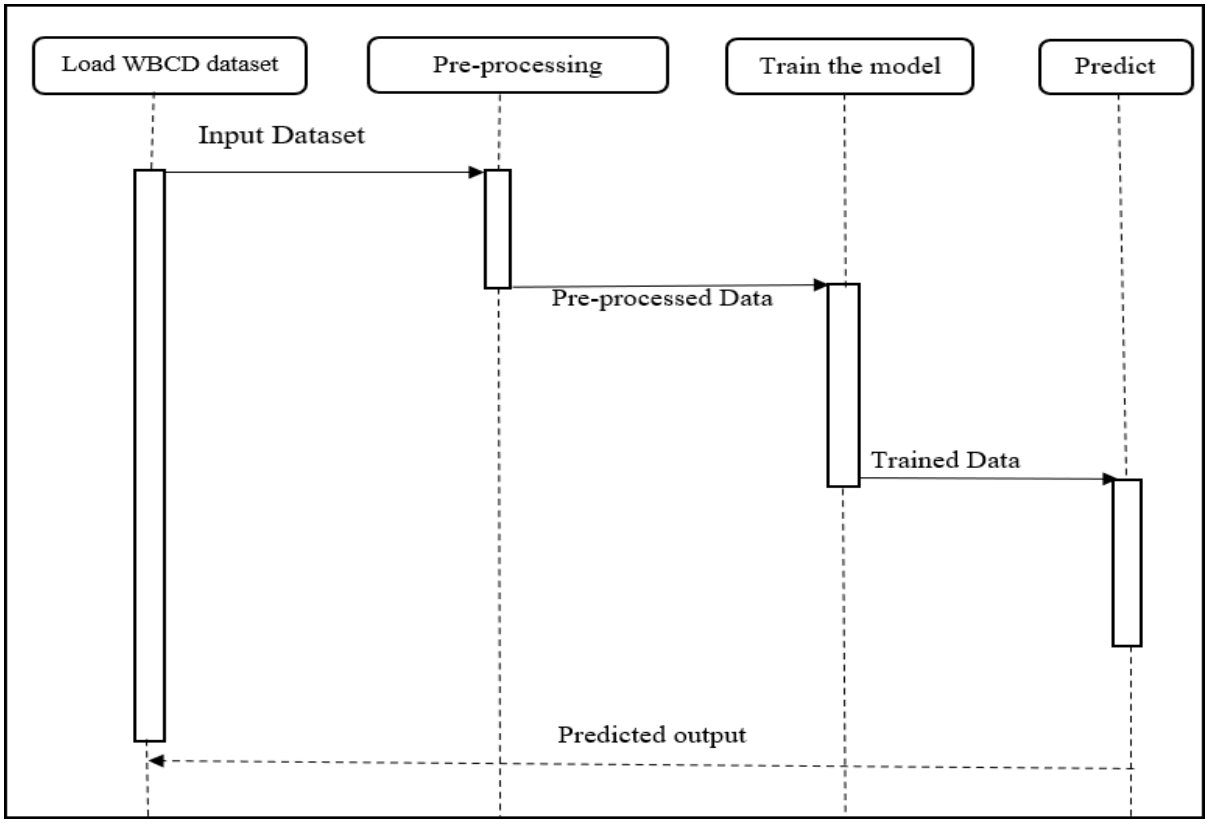

Figure 6: Sequence diagram

Figure 6 displays the sequence diagram for this project. The sequence diagram essentially depicts how messages move between objects in the system. There are 4 objects in Figure 2, including Loading the WBCD dataset, Pre-processing, Training the Model, and Predicting. The things are represented by rectangular boxes. Every object has a dotted line attached to it that symbolizes the

lifeline of the related object, and vertical bars on the lifeline, or dotted line, indicate the object's active condition. In the sequence diagram, the arrow marks signify the transmission of messages from one object to another.

3.5.3 USE CASE DIAGRAM

Actors, use cases, and their linkages are all included in use case diagrams. They reflect the use case perspective of a system. An example of a use case is a specific system capability.

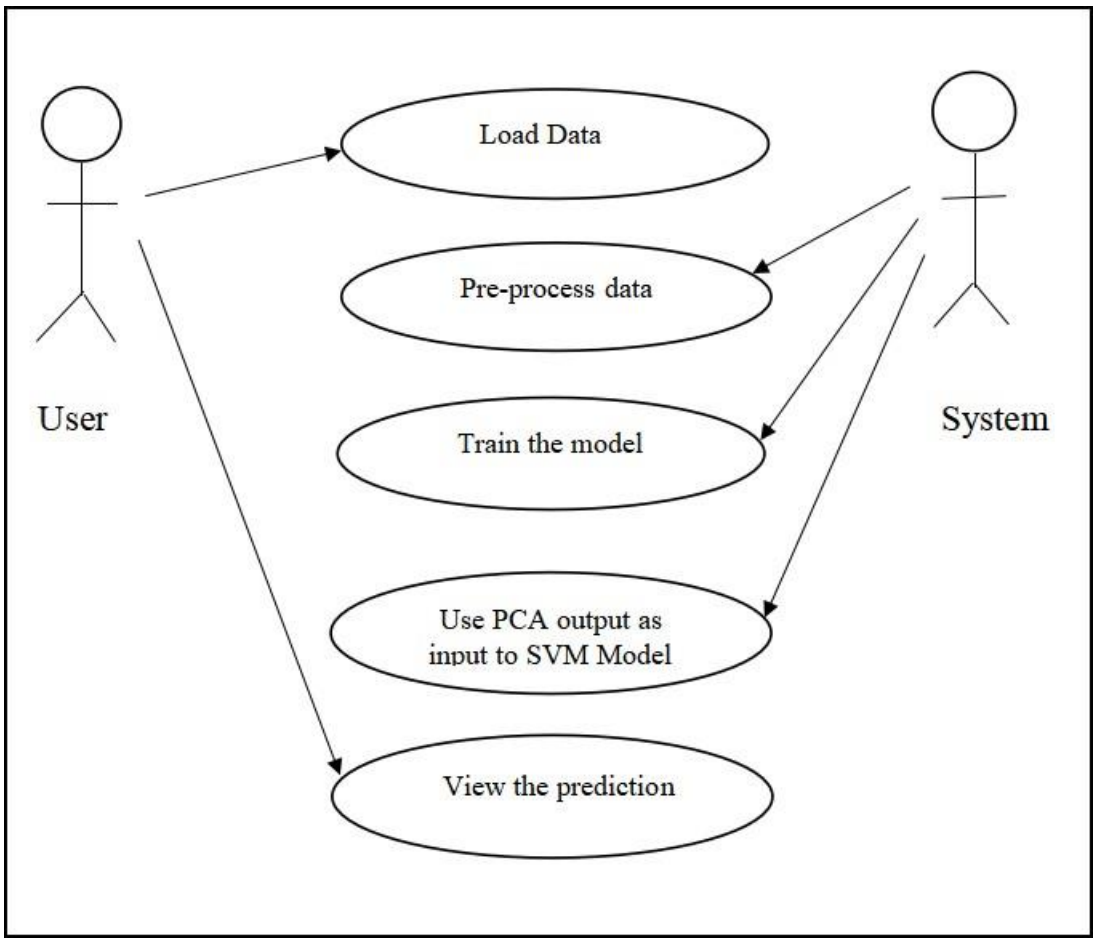

Figure 7: Use Case Diagram for Early Prediction of Breast Cancer.

3.5.4 ACTIVITY DIAGRAM

An activity diagram demonstrates how command moves throughout a system. It consists of connections and activities. A branched, concurrent, or sequential flow might be present.

Numerous activity diagrams are produced to show the entire system flow. Controls are moved across a system using activity diagrams. This is done on purpose so that you can see how the system works in action.

Figure 3 represents the Activity diagram of this project. There are mainly 4 activities that are represented in the diagram namely Loading the WBCD dataset, Pre-processing, Training, and Prediction. The process begins by loading the WBCD dataset into the system then the input dataset is pre-processed. The model is built by a training dataset. To forecast breast cancer, a trained model is employed. Once all operations or procedures have been completed, the outcome is shown.

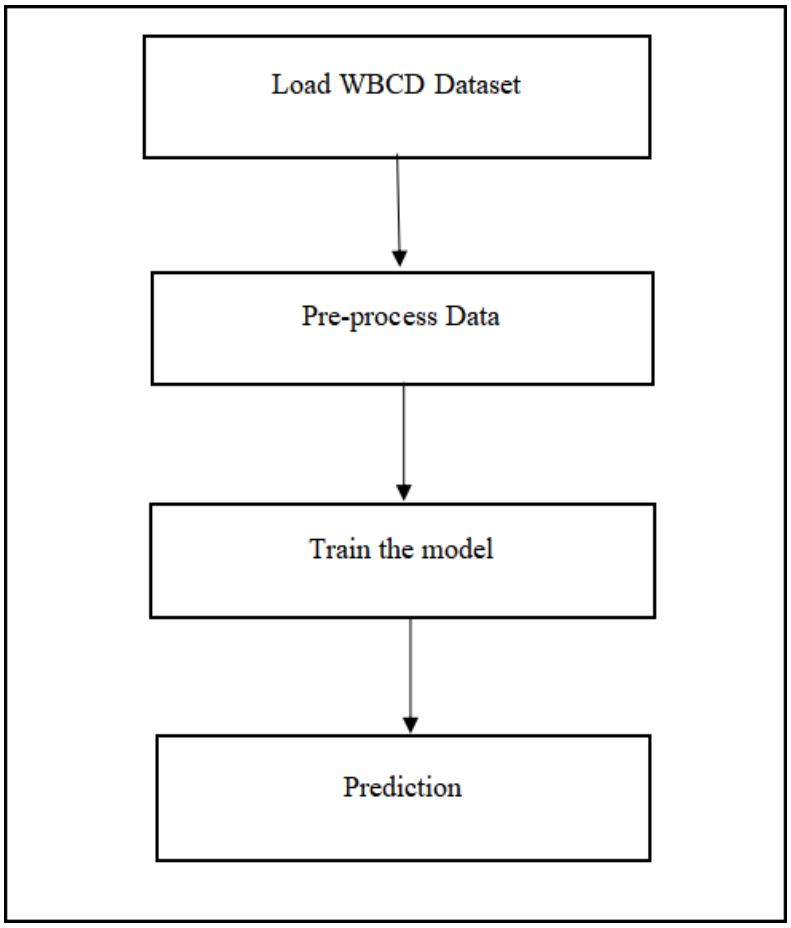

Figure 8: Activity Diagram for Breast Cancer Prediction

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 DESCRIPTION OF THE DATASET

Breast Cancer Wisconsin (Diagnostic) Dataset, which was retrieved via Kaggle, is the dataset gathered for the research. Dr. William H. produced the dataset. Ten real-valued characteristics are present in the dataset:

- **radius:** In a digital picture of a breast mass, this is the average distance between a cell nucleus's centre and its border points.
- **texture:** In the Wisconsin Breast Cancer Dataset (WBCD), texture is a characteristic that describes the visual properties of cells in a breast mass observed through a microscope. These features are derived from digital images and can reveal details about the uniformity, diversity, nucleus size, and shape of the cells. The texture is among the various features used in WBCD for the purpose of separating benign from cancerous breast tumours.
- **perimeter:** The dataset contains several measurements of breast masses that were made using fine needle aspirates (FNA) pictures that had been digitally captured. One of these measurements is the perimeter, which represents the total length of the outer edge or boundary of the mass.
- **smoothness**: It refers to a feature that describes the variation in the radius length of the cell nuclei in a microscopic image of a breast mass.
- **compactness**: compactness is an important measure of the quality of a classification model, as models that are able to achieve high levels of compactness for benign and malignant breast tumors are more likely to accurately predict the diagnosis of new cases.
- **concavity**: concavity refers to the presence of irregularities or depressions in the contour of a cell's surface. Concavity is one of several morphological features included in the WBCD dataset, which can be used to classify whether a given sample is malignant or benign. Specifically, the presence of concavity is often considered a characteristic of malignant cells and may be used as a predictor of malignancy in diagnostic settings.
- **concave points**: Concave points are a characteristic that defines how many concave areas of a cell nucleus contour are seen in an image obtained from a breast tumor that was aspirated with a tiny needle.

- **symmetry**: Symmetry points refer to specific locations or pixels in an image where the left and right halves of the image are identical or nearly identical. In the context of the WBCD (Wisconsin Breast Cancer Diagnostic) dataset, which contains digitized images of breast mass lesions, symmetry points can be used to analyze the symmetry of the breast masses and potentially aid in the diagnosis of breast cancer.
- **fractal dimension**: fractal dimension is used to characterize the texture of the cell nuclei in the images.

Attribute Information:

1) ID number

2) Diagnoses (3–32) (M means malignant, B means benign)

This dataset contains details of 569 patient data samples with 33 columns. It contains Malignant and Benign tumors of breast cancer. The number of features in this dataset has been decreased by utilizing the PCA feature selection technique to improve it. Considering only relevant features and removing unnecessary factors like id, null values, and unnamed attributes to have a high influence in the dataset. A feature reduction method is PCA. The variables which do not affect the decision-making are eliminated by PCA. In this project, the original dataset contains id, unnamed and null values now we are removing them which are not required. In the below figure, we can see the cleaned dataset.

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimensi
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	

Figure 9: Cleaned Dataset.

4.2 EXPERIMENTAL RESULTS

A heat map is a graphical representation of where each cell in the matrix is color coded to represent value in the cell. These maps are often used for representing the patterns and their relationships in the data. In this project, the heat map is created using the Seaborn library in Python. To create a heat map, we need to remove unnecessary features. such that the heat map will display correlation values between each feature and target variable, with dark color indicating a stronger correlation. Now we will be assigning our dataset to the Seaborn library. It tells the most relevant features of the dataset. It shows the relationship between two variables and their significance on a plotted graph.

The figure below is heat map of the project it talks about the attributes and the relation between each of them. This heat map gives a clear understanding of the correlated features and this a way more understandable by looking at the correlation heat map and this says about different type of malignant and benign classes which are stored in dataset very clearly and the classes are represented using different colors such as yellow, blue, green for clear understanding about different disease classes present in the dataset. So, we can have a very clear understanding and also which allows us to explain others how the classes of disease are stored in dataset with much more efficient way and also understandable way.

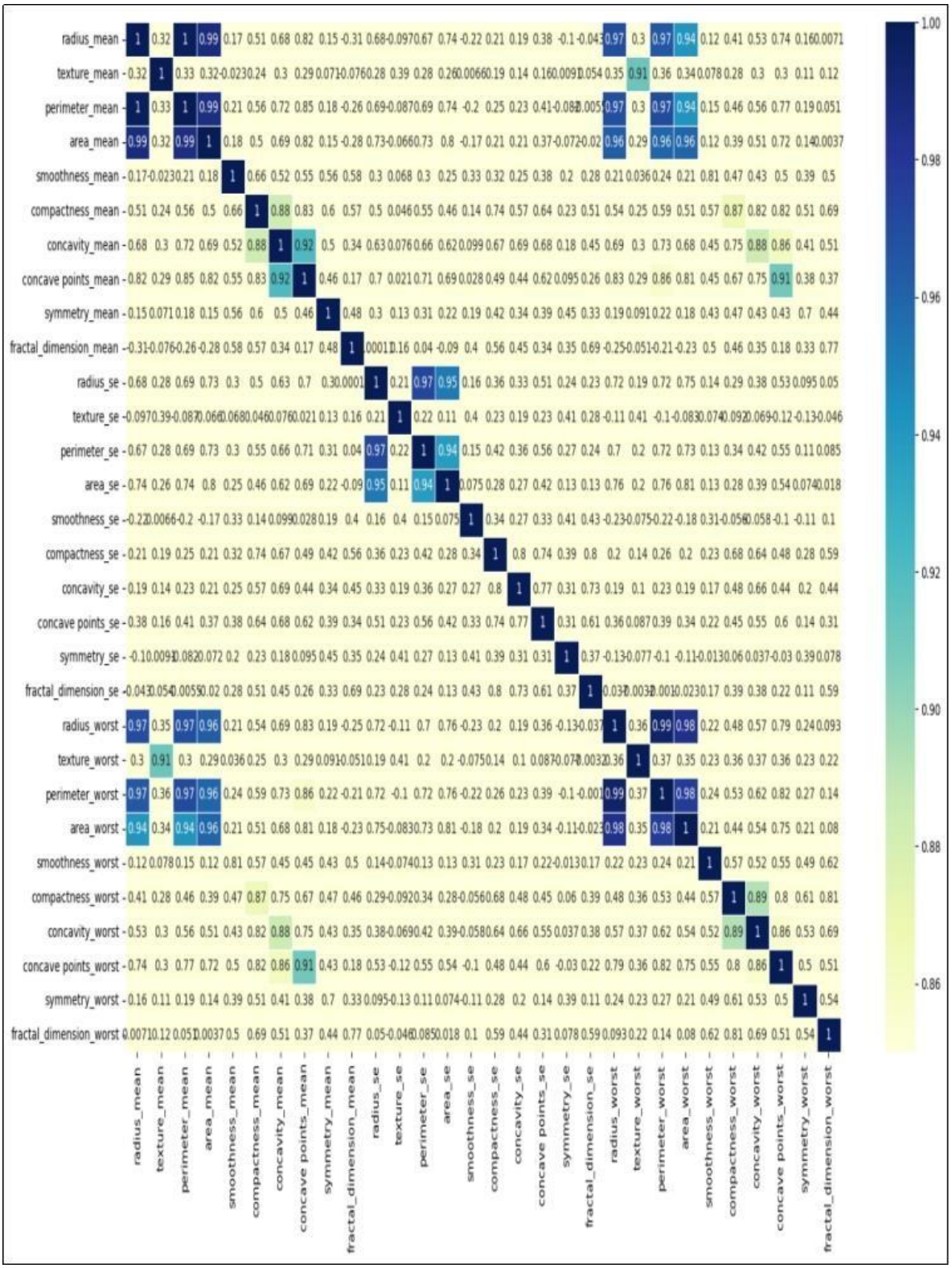

Figure 10: Heat Map of the WBCD Dataset

A table that displays various predictions' results is known as a confusion matrix. It plots a table of all predicted values with actual values. In the below table, we can see the accuratevalues in blue shading by dividing the training set into 80% and the testing set into 20%. To improve the model's accuracy, it is necessary to reduce false positives and false negatives.

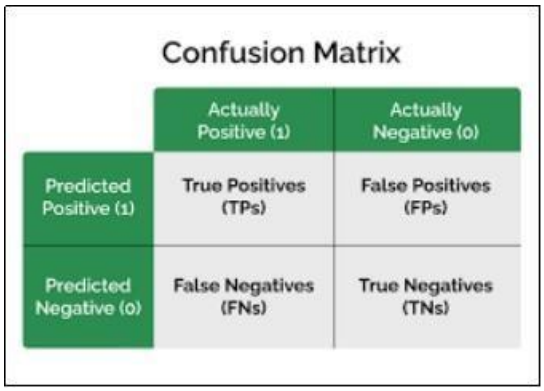

Figure 11: Confusion matrix

It is possible to quantify important performance indicators including recall, precision, accuracy, and the AUC-ROC curve using a confusion matrix, which is a useful tool in machine learning.

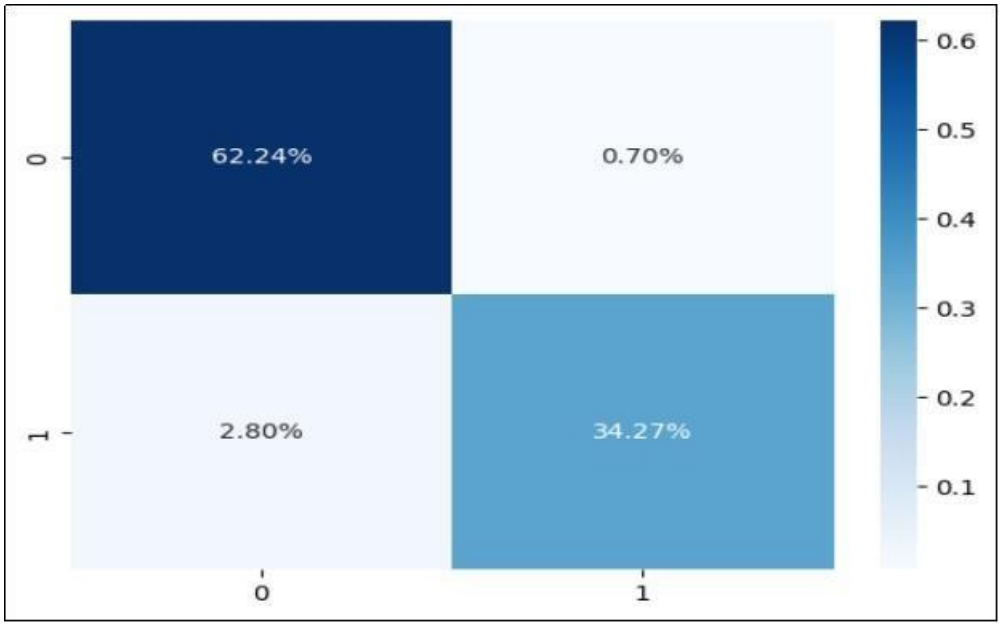

Figure 12: Confusion matrix of the hybrid model

The true positive of the model is 62.24% which represents the correctly predicted outcomes. A false negative of the model is 0.70% which represents incorrectly predicted outcomes. Type II errors include false negatives.

Let us analyze the table and perform some common performance measures from the confusion matrix.

Accuracy: Accuracy is a common performance metric used in machine learning to evaluate a classification model's efficiency. It calculates the percentage of examples in a dataset that themodel correctly categorizes. Accuracy is calculated as follows:

$$\textbf{Accuracy} = \frac{(TP+TN)}{TP+TN+FP+FN}$$

In other words, accuracy is the proportion of dataset occurrences the model correctly categorized. A model with high accuracy may accurately predict the class of most cases in the dataset, whereas a model with low accuracy tends to forecast the class of more examples incorrectly. Here we have given the PCA output as input to the SVM algorithm. This makes the hybrid model train the input and perform the calculations and gives us more accuracy rather than a single algorithm value.

Precision: When evaluating how effectively a classification model is doing, machine learning uses the precision statistic. It determines the percentage of "true positives"—predicted as positive and really positive—from all positive predictions the model generates.

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

Recall: Recall is a performance indicator used in machine learning that quantifies the proportion of true positive results that a model accurately identified out of all real positive cases.

$$\textbf{Recall} = \frac{TP}{TP+FN}$$

f1-score: The F1-score is a well-liked machine learning statistic for evaluating a classification model's efficiency. To assess a model's accuracy, it looks at its recall and precision.

$$\textbf{f1-score} = 2 * \frac{(precision * recall)}{(precision + recall)}$$

```
Accuracy: 0.965034965034965
Classification report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        90
           1       0.98      0.92      0.95        53

    accuracy                           0.97       143
   macro avg       0.97      0.96      0.96       143
weighted avg       0.97      0.97      0.96       143
```

Figure 13: Classification Report.

K-fold cross-validation: The K-fold cross-validation technique is frequently utilized for model evaluation and hyperparameter adjustment in machine learning. The original dataset is partitioned into k-folds of equal size to perform k-fold cross-validation. The dataset should be split into k equal-sized folds. For each of the remaining k folds, use the last k-1 folds as the test set and the final k folds as the training set. The model should be tested after being evaluated on the training set.

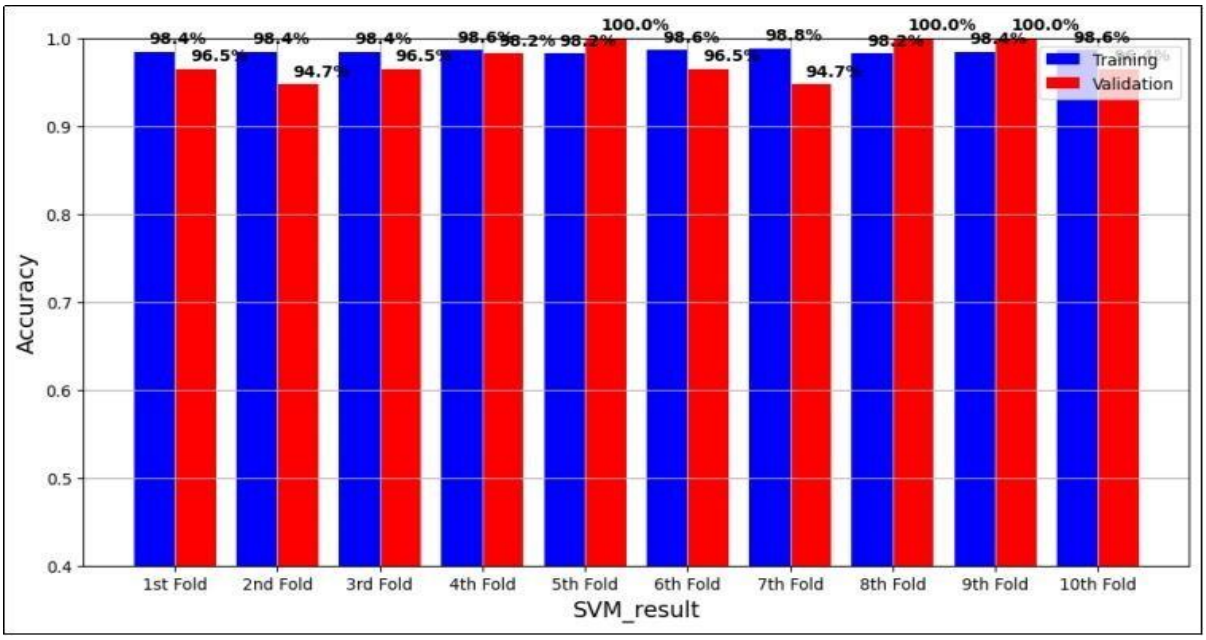

Figure 14: Accuracy of K-folds

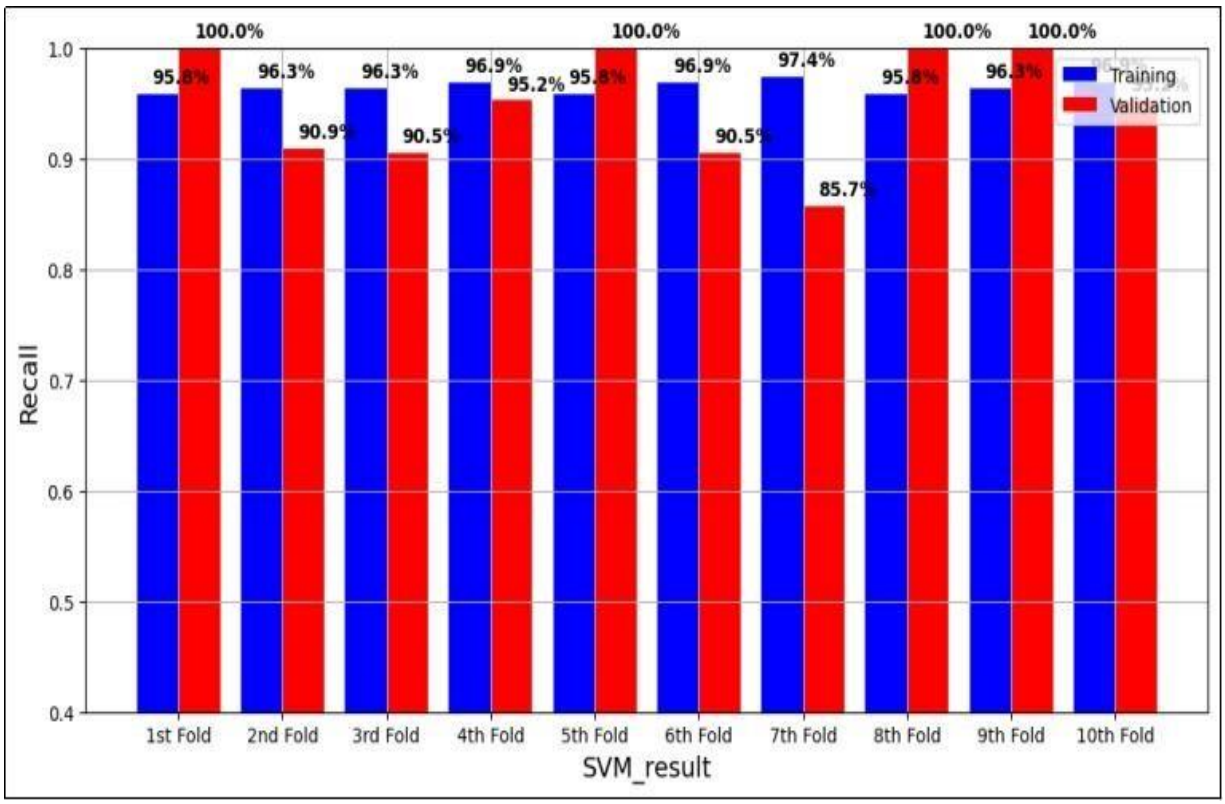

Figure 15: Recall of K-folds

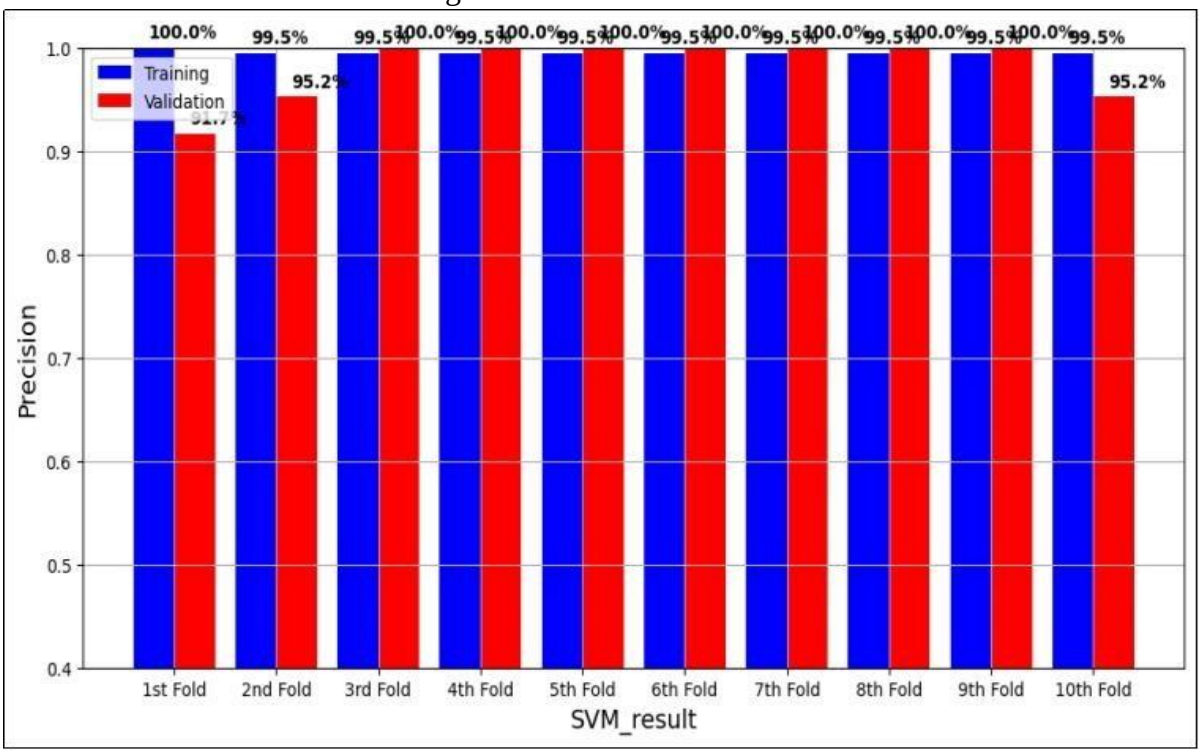

Figure 16: Precision scores of K-fold

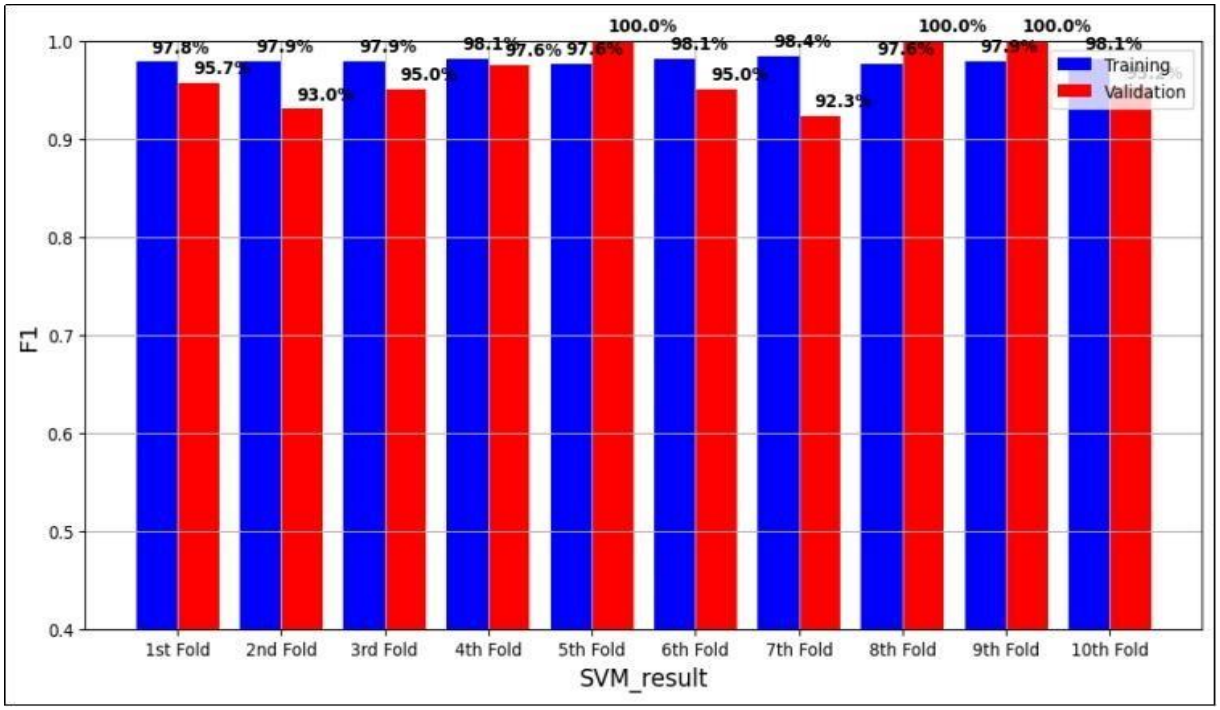

Figure 17: f1-score of K-folds

ROC curve: The performance of a binary classifier when its threshold for categorising data is altered is depicted graphically by the ROC curve. The true positive rate (TPR) and false positive rate (FPR) are plotted for various threshold levels. The proportion of positive samples that the classifier correctly identifies as positive is known as the TPR, also known as sensitivity or recall. The FPR measures how often negative samples are mistakenly classified as positive by the classifier.

The efficiency of binary classifiers is typically evaluated using the AUC, or area under the ROC curve. With a value of 1 denoting perfect classification and 0.5 denoting random categorization, AUC assesses the classifier's capacity to distinguish between positive and negative samples. In conclusion, the ROC curve is a useful tool for assessing the performance of binary classifiers, particularly when the cost of false positives and false negatives is not equal, and the AUC is an easy way to assess the overall efficacy of a classifier.

In our project, we have presented a graphical representation of the ROC curve for traditional SVM and SVM with K-folds. The graph indicates that SVM with K-folds exhibits higher positiveaccuracy compared to the traditional SVM.The blue line indicates traditional SVM which is having an accuracy of 98.8 whereas the orange line indicates svm with kfolds having

high accuracy of 99.7%. This shows that the hybrid model is better that the single model. A hybrid model increases the accuracy compared to a single algorithm.

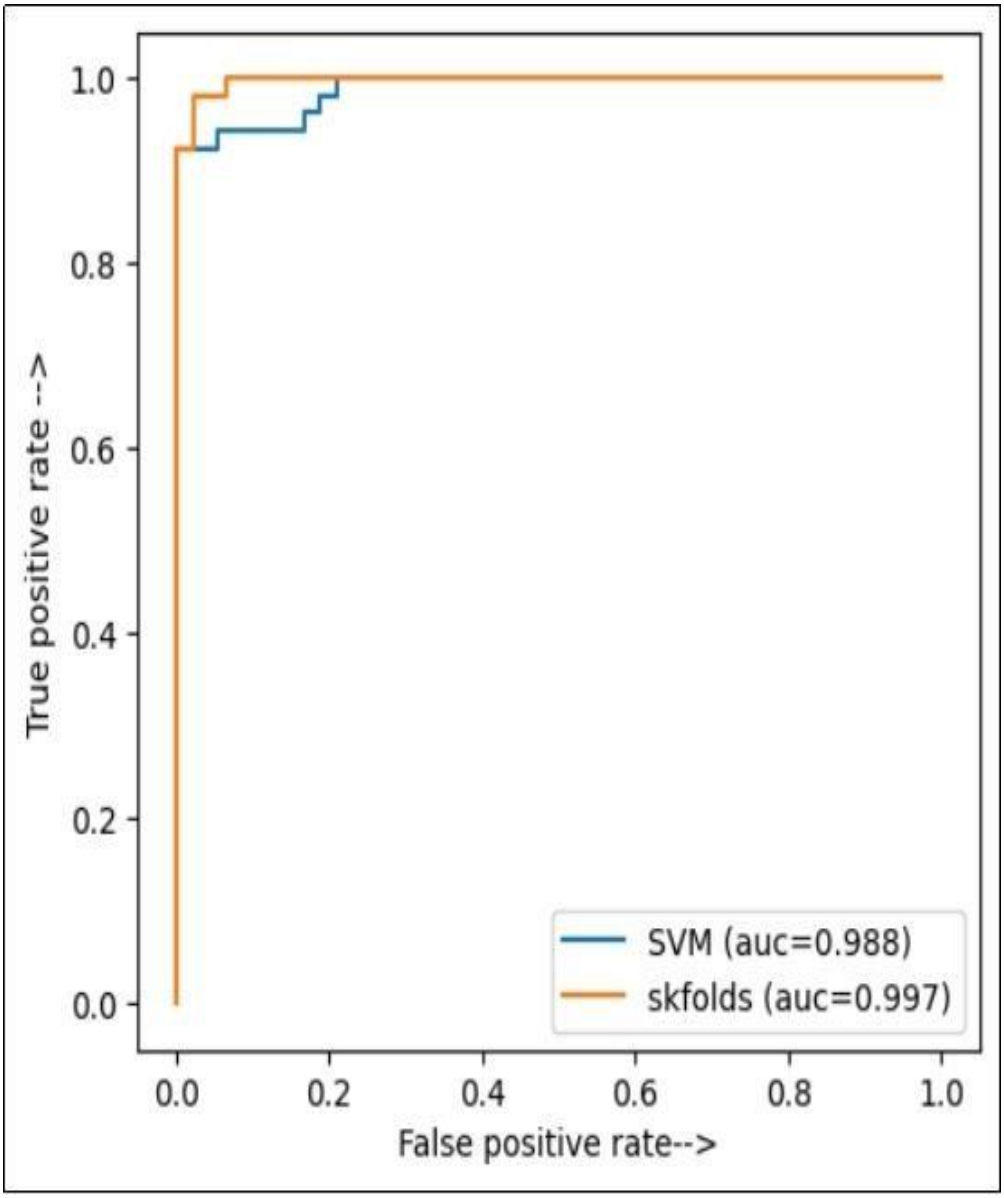

Figure 18: ROC curve.

4.3 SIGNIFICANCE OF PROPOSED METHOD WITH ITS ADVANTAGES:

Early identification is essential for effective treatment of breast cancer, a severe health concern. In recent years, Support Vector Machines (SVM) and Principal Component Analysis (PCA) have become prominent machine learning techniques for the early diagnosis of breast cancer. This study aims to evaluate the efficacy and advantages of PCA with SVM and k-fold cross-validation for the early breast cancer prediction. Early breast cancer identification and diagnosis are crucial for a successful course of therapy and increased survival rates. By identifying trends and predicting outcomes, machine learning approaches that leverage large datasets and complex algorithms have the potential to improve the detection of breast cancer.

PRINCIPAL COMPONENT ANALYSIS (PCA)

A common technique for extracting features and reducing dimensionality is PCA. PCA reduces the dimensions of a high-dimensional dataset while retaining the majority of its original variance. PCA can assist minimise overfitting, speed up computation, and enhance the effectiveness of machine learning algorithms by lowering the amount of features. PCA can assist in identifying the most important characteristics and reducing noise in the data when it comes to the identification of breast cancer.

SUPPORT VECTOR MACHINES (SVM)

SVM is a potent technique for classification and regression in machine learning. Finding the hyperplane that optimally divides the data points into distinct groups is how SVM operates. SVM is particularly well-suited for binary classification problems, such as breast cancer detection, where the goal is to classify samples as either malignant or benign. SVM has several advantages, including high accuracy, robustness to noise, and the ability to handle high-dimensional data.

BREAST CANCER DATASET

To demonstrate the effectiveness of PCA with SVM and k-fold cross-validation for predicting breast cancer at its early stages, we used the publicly available Breast Cancer Wisconsin (Diagnostic) Dataset. This dataset contains 569 samples with 33 features, including patient age, tumor size, and various measurements of cell nuclei. The dataset is labeled with either M (malignant) or B (benign) to indicate the diagnosis. On the Breast Cancer Wisconsin dataset, we used PCA with SVM and k-fold cross-validation to get a classification accuracy of 96.5%. To assess the effectiveness of the hybrid model, we employed k-fold cross-validation with k=10.

Because of its high accuracy, PCA and SVM are a potent combo for identifying breast cancer.

The outcomes show that PCA combined with SVM and k-fold cross-validation is a potent method for early breast cancer prediction. SVM produced a strong and reliable classification model, whereas PCA assisted in identifying the most useful features and reducing noise in the data. K-fold cross-validation aids in assessing the model's efficiency. The high degree of accuracy attained in this study indicates that this method may help in early identification and diagnosis of breast cancer. We have demonstrated the effectiveness of PCA with SVM and k-fold cross-validation as a method for early breast cancer prediction. The best features were found by combining PCA with SVM, which also increased the classification model's precision. the use of K- fold cross-validation helped assess the model's performance and avoid overfitting. The findings suggest that this strategy may enhance breast cancer detection and diagnosis, and more investigation into its practical applicability is necessary.

ADVANTAGES OF THIS MODEL

The following are the advantages of this hybrid model for predicting breast cancer at high speed:

- The hybrid model can leverage the strengths of different models and overcome their limitations.
- A hybrid model can reduce the risk of overfitting and improve generalization performance.
- A hybrid model can handle missing data and imbalanced datasets more effectively. By incorporating multiple models, the hybrid model can leverage each model's strengths to overcome others' weaknesses and provide more robust and reliable predictions.
- A hybrid model can offer predictions with more thorough and understandable justifications. The hybrid model, which combines various models, can offer a more detailed view of the underlying facts and more insightful insights for decision-making.
- PCA can reduce the effect of noise and outliers in the data by identifying the most important features and filtering out the noise. This can improve the robustness of the model and make it more reliable.
- By reducing the dimensionality of the dataset, SVM with PCA can reduce the number of features that need to be trained, thus reducing the training time and computational resources required.

- Using SVM with PCA for predicting breast cancer can provide several advantages, including improved accuracy, interpretability, robustness to noise, and reduced training time.
- ROC curve and k-fold cross-validation are both widely used techniques for assessing the effectiveness of machine learning models. By combining these techniques with SVM, we can obtain a more accurate evaluation of the model's performance and its adaptability to fresh data.
- K-fold cross-validation can assist in lowering the possibility of overfitting and by testing the model on various data subsets, the generalization performance of the model can be improved. This could increase the predictability of the model and reduce the likelihood of making false-positive or false-negative predictions.

Table 2. Comparison of proposed approach with existing approaches

REFERENCE NUMBER	EXISTING APPROACH	PROPOSED APPROACH
[1]	Accuracy of reference one existing approach is 95%	Accuracy of proposed approach is 96%
[14]	Accuracy of existing approach is 92.7% with SVM.	Accuracy of proposed approach is 96% i.e. (PCA-SVM)
[15]	Accuracy of existing approach is 79.2 %.	Accuracy ofproposed approach is 96%

CHAPTER 5

CONCLUSION WITH FUTURE ENHANCEMENTS

5.1 SUMMARY OF THE PROJECT

Breast cancers are one of the world's maximum risky diseases. This disease occurs due to the uncontrol growth of cells in body. According to the WHO, if breast cancer is not discovered in a timely manner, it is the leading cause of mortality. The percentage of occurrence of breast cancer in women is more when compared to men. This study used a variety of machine-learning algorithms to predict breast cancer in its early stages. One of the main subfields of artificial intelligence is machine learning. Machine mastering algorithms are the mathematical fashions it's far used for mapping strategies or to find styles withinside the data. Machine learning algorithms contain a large number of computational algorithms that are used in various situations such as computational, pattern recognition, and prediction after understanding the training data. The training data is inputted into selected algorithms for undergoing the further process.

These machine learning algorithms give reliable output to various healthcare applications. These machine-learning algorithms are fast and efficient in predicting disease at its early stages. The early prediction of disease helps in saving people. The purpose of this study is to use hybrid machine-learning algorithms to predict breast cancer in its early phases. It is coined after thoroughly following and understanding different research papers. The advantage of using a hybrid machine learning algorithm is, it offers high performance, and accuracy and is also convenient in applying to high dimensional data. Sometimes the output generated using a single algorithm is difficult and it is solved by some hybrid algorithms. Hybrid algorithms can be formed with the integration of various algorithms or by giving an output of one algorithm as input to other for further improvement. With hybrid machine learning techniques, this research hopes to predict breast cancer in its early phases.

CONCLUSION

This project says about combination of two machine learning algorithms such as PCA (Principal Component Analysis) and SVM (Support Vector Machine). There are many advantages of both algorithms let’s see some advantages of PCA. PCA comes an unsupervized machine learning

algorithm. PCA is employed for simple computing, for accelerating other machine learning methods, and for data cleansing. One of the supervised machine learning algorithms is the SVM. SVM also provides various advantages let us discuss some of them, SVM is more efficient in prediction when there is a dimension greater than samples and it is relatively well whenthe margins of classes are clear. With the help of the advantages of these two algorithms, we developed a hybrid system for breast cancer prediction.

On the Wisconsin Breast Cancer dataset (WBCD), we have utilised hybrid machine learning algorithms for the prediction of breast cancer. The process starts by cleaning the unnamed and unnecessary data. Then after cleaning the data by applying the PCA algorithm for dimensionality reduction and providing the reduced dimensional data as input to SVM. After getting results from the SVM algorithm optimization of SVM is done with the help of k-folds. This method results in an accuracy of 96.5%. SVM-k-folds make it more efficient in predicting the accurate output.

An effective validation and testing graph for precision, recall, and f1-score using SVM with k-folds was produced using 10-folds. This graph shows the improvement in recall, precision, accuracy, and f1-score. It is crucial to know about the accuracy of the terms means the state of correctness, precision says about quality, recall means the data samples are correctly identified bymachine learning algorithms, and the final one f1-score is used for measuring model accuracy. After analysis results show that recall, precision, accuracy, and f1-score all are above 90%. That says the approach worked accurately in predicting breast cancer. After proper implementation of the above algorithms there is little comparison with simple SVM and provided an accuracy of 95%it clearly says that the traditional SVM is not as accurate as the hybrid SVM algorithm. So, usinghybrid algorithms is very helpful for various prediction purposes. The utility of machine learning methods, both supervised and unsupervised, in reliably detecting breast cancer in its early stages is discussed in this project's conclusion. It also provides knowledge that hybrid machine algorithms are way better than traditional approaches. In this project, we have noticed that the hybrid algorithm we have used such as PCA-SVM(k-folds) outperformed when compared with SVM algorithms.

5.2 FUTURE ENHANCEMENTS

According to the analysis's findings, the Wisconsin Breast Cancer dataset is the only dataset where the recommended approaches produce reliable results. There is a need for further future enhancement should be carried out by applying various variables for obtaining better performance and there is a need of checking our proposed work with the different datasets to know whether our approach provides the same result on all other datasets or not. There is a lot of future work that can be enhanced from this as we have used only the Wisconsin dataset so even there are other methods of collecting data such as image data that is mammographic data which says the presence of disease in image format and it is even very more interesting to do such projects on image dataset because one can show whether the is a cancerous cell or not in a clear image format.

Another improvement we can suggest is to provide a computer-aided diagnosis with usage of computer-aided diagnosis it is more and more accurate to get the result whether patient is a Breast cancer patient, not and with the help of computer-aided diagnosis the results will be betterbecause as the computer systems can work more significantly and correctly in detecting disease and also the result provided by computer systems takes less time when compared to the previous situation and this is one of the enhancements one can follow and further on easy to detect diseaseand can give results accurately and in less time.

CHAPTER 6

APPENDICES

```
#importing required modules and dataset

#for loading data and for performing data analysis operations on it

import pandas as pd

import numpy as np

#for data visualization

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn import*

#for file operations

import os

print(\"All required libraries loaded! \")

#Load the dataset

data = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/data.csv', index_col=False)

data.shape

for PCA (feature engineering)

from sklearn.decomposition import PCA

#for data scaling

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_scaled = scaler.fit_transform(X)
```

```
X_scaled = pd.DataFrame(X_scaled)

X_scaled_drop = X_scaled.drop(X_scaled.columns[[2, 3, 12, 13, 22, 23]], axis=1)

pca = PCA(n_components=0.95)

x_pca = pca.fit_transform(X_scaled_drop)

x_pca = pd.DataFrame(x_pca)

print(\"Before PCA, X dataframe shape = \", X. shape,\"\\nAfter PCA, x_pca dataframe shape

=\",x_pca.shape)

 #for splitting dataset

from sklearn.model_selection import train_test_split

X=(Xy.iloc[:,0:11]).values

#75:25 train: test data splitting

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=0)

print(\"X_train shape \",X_train.shape)

print(\"y_train shape \",y_train.shape)

print(\"X_test shape \",X_test.shape)

print(\"y_test shape \",y_test.shape)

#for displaying evaluation metrics

from sklearn.metrics import classification_report

from sklearn.metrics import confusion_matrix

CM = confusion_matrix(y_test, y_pred_svc1)

print(\"Confusion matrix:\\n\", CM)

sns.heatmap(CM/np.sum(CM),annot=True,fmt='.2%', cmap='Blues')

plt.show()
```

```
from sklearn.metrics import accuracy_score

print(\"Accuracy:\",accuracy_score(y_test, y_pred_svc1))

creport = classification_report(y_test, y_pred_svc1)

print(\"Classification report:\\n\",creport)

from sklearn.model_selection import cross_validate

def Cross_validating(model, _X, _y, _cv=5):

_scoring = ['accuracy', 'precision', 'recall', 'f1']\n",

results = cross_validate(estimator=model,\n",

X=_X,\n",

y=_y,\n",

cv=_cv,\n",

scoring=_scoring,\n",

return_train_score=True)\n",

return {\"Training Accuracy scores\": results['train_accuracy'],

"Mean Training Accuracy\": results['train_accuracy'].mean()*100,

"Training Precision scores\": results['train_precision'],

"Mean Training Precision\": results['train_precision'].mean(),

"Training Recall scores\": results['train_recall'],

"Mean Training Recall\": results['train_recall'].mean(),

"Training F1 scores\": results['train_f1'],

"Mean Training F1 Score\": results['train_f1'].mean(),

"Validation Accuracy scores\": results['test_accuracy'],

"Mean Validation Accuracy\": results['test_accuracy'].mean()*100,
```

```
"Validation Precision scores\": results['test_precision'],
"Mean Validation Precision\": results['test_precision'].mean(),
"Validation Recall scores\": results['test_recall'],
"Mean Validation Recall\": results['test_recall'].mean(),
"Validation F1 scores\": results['test_f1'],
"Mean Validation F1 Score\": results['test_f1'].mean()
}
#from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
svc = SVC()
'"svc.fit(X_train, y_train)
y_pred_svc =svc.predict(X_test)
y_pred_svc.shape"
SVM_result = Cross_validating(svc, X, encoded_y, 10)
print(SVM_result)
def ploting_result(x_label, y_label, plot_title, train_data, val_data):
plt.figure(figsize=(12,6))
labels = [\"1st Fold\", \"2nd Fold\", \"3rd Fold\", \"4th Fold\", \"5th Fold\",\"6th Fold\",\"7th
Fold\",\"8th Fold\",\"9th Fold\",\"10th Fold]
X_axis = np.arange(len(labels))
ax = plt.gca()
plt.ylim(0.40000, 1)
plt.bar(X_axis-0.2, train_data, 0.4, color='blue', label='Training')
```

```
plt.bar(X_axis+0.2, val_data, 0.4, color='red', label='Validation')

plt.title(plot_title, fontsize=30)

plt.xticks(X_axis, labels)

plt.xlabel(x_label, fontsize=14)

plt.ylabel(y_label, fontsize=14)

plt.legend()

plt.grid(True)

plt.show()

#roc curve

from sklearn.metrics import roc_curve,auc

sk_fpr,sk_tpr,threshold=roc_curve(y_test,y_pred_svc2)

auc_sk=auc(sk_tpr,sk_fpr)

svm_fpr,svm_tpr,threshold=roc_curve(q_test,y_predict_svc3)

auc_svm=auc(svm_tpr,svm_fpr)

plt.figure(figsize=(5,5),dpi=100)

plt.plot(svm_tpr,svm_fpr,linestyle='-',label='SVM(auc=%.3f)'%auc_svm)

plt.plot(sk_tpr,sk_fpr,linestyle='.',label='skfolds(auc=%.3f)'%auc_sk)

plt.xlabel('False positive rate-->')

plt.ylabel('True positive rate -->')

plt.legend()

plt.show()
```

REFERENCES

[1] Lin, H., & Ji, Z. (2020). Breast Cancer Prediction Based on K-Means and SOM Hybrid Algorithm. In Proceedings of the 2nd International Conference on Computer Modeling, Simulation and Algorithm (pp. 042012). Journal of Physics: Conference Series, 1624(4). doi: 10.1088/1742-6596/1624/4/042012.

[2] Khozama, S., Mayya, A. M. (2022). A New Range-based Breast Cancer Prediction Model Using the Bayes' Theorem and Ensemble Learning. Information Technology and Control, 51(4), 757-770.

[3] Elsadig, M. A. (2021). A Machine Learning Approach for Breast Cancer Early Detection. Journal of Theoretical and Applied Information Technology, 99(5), 1044-1053. ISSN: 1992-8645. E-ISSN: 1817-3195.

[4] Rawal, R. (2020). Breast Cancer Prediction Using Machine Learning. Journal of Emerging Technologies and Innovative Research (JETIR), 7(5), 100-110. ISSN: 2349-5162.

[5] Dutta, Shawni & Bandyopadhyay, Samir. (2020). Early Breast Cancer Prediction using Artificial Intelligence Methods. Journal of Engineering Research and Reports. 13. 48-54. 10.9734/JERR/2020/v13i217105.

[6] Singh, Gaurav. "Breast Cancer Prediction Using Machine Learning." International Journal ofAdvanced Research in Computer Science 6, no. 4 (2020): 278-284.

[7] Umer, M.; Naveed, M.Alrowais, F. Ishaq, A. Hejaili, A.A.Alsubai, S. Eshmawi, A.A.Mohamed, A.Ashraf, I. Breast Cancer Detection Using Convoluted Features and Ensemble Machine Learning Algorithm. Cancers 2022, 14, 6015 https://doi.org/10.3390/cancers14236015

[8] Mridha, Krishna. "Early prediction of breast cancer by using artificial neural network and machine learning techniques." In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), pp. 582-587. IEEE, 2021.

[9] Shudipti,Aafreen,Rakesh (2021).Prediction of Breast Cancer using Machine Learning. IJIRAE::International Journal of Innovative Research in Advanced Engineering,Vol: VIII,28- 33.

[10] Apoorva V and Yogish H K, "Breast Cancer Prediction Using Machine Learning

Techniques," in Atlantis Highlights in Computer Sciences, volume 4, Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021).

[11] Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 191, 487-492.

[12] Afrash MR, Bayani A,Shanbehzadeh M, Bahadori M, Kazemi-Arpanahi H.Developing the breast cancer risk prediction system using hybrid machine learning algorithms. J Edu Health Promot 2022;11:272.

[13] YMER, S. Sridevi. "Breast Cancer Prediction with Hybrid ML Models." Journal Title, vol. 21, no. 5, May 2022, ISSN 0044-0477.

[14] Vyas, S., Chauhan, A., Rana, D., & Ansari, N. (2022). Breast Cancer Detection Using Machine Learning Techniques. International Journal of Recent Advances in Science, Engineering and Technology (IJRASET), 10(V), May 2022.

[15] Sivakumar, P., Lakshmi, T. U., Reddy, N. S., Pavani, R., & Chaitanya, V. (2020). Breast Cancer Prediction System: A novel approach to predict the accuracy using Majority-Voting Based Hybrid Classifier (MBHC). In 2020 2nd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 57-62). IEEE. DOI: 10.1109/INDISCON50162.2020.00024.

[16] Bhise, Sweta, Shrutika Gadekar, Aishwarya Singh Gaur, Simran Bepari, and D. S. A. Deepmala Kale. "Breast cancer detection using machine learning techniques." Int. J. Eng. Res. Technol 10, no. 7 (2021).

[17] Sharma, Shubham, Archit Aggarwal, and Tanupriya Choudhury. "Breast cancer detection using machine learning algorithms." In 2018 International conference on computational techniques, electronics and mechanical systems (CTEMS), pp. 114-118. IEEE, 2018.

[18] Lim, Tze Sheng, Kim Gaik Tay, Audrey Huong, and Xiang Yang Lim. "Breast cancer diagnosis system using hybrid support vector machine-artificial neural network." Int. J. Electr. Comput. Eng.(IJECE) 11, no. 4 (2021): 3059.

[19] Hamed, Ghada, Mohammed Abd El-Rahman Marey, Safaa El-Sayed Amin, and Mohamed Fahmy Tolba. "Deep learning in breast cancer detection and classification." In

Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), pp. 322-333. Springer International Publishing, 2020.

[20] Melekoodappattu, Jayesh George, and Perumal Sankar Subbian. "Automated breast cancer detection using hybrid extreme learning machine classifier." Journal of Ambient Intelligence and Humanized Computing (2020): 1-10.

[21] Wang, Xiaomei, Ijaz Ahmad, Danish Javeed, Syeda Armana Zaidi, Fahad M. Alotaibi, Mohamed E. Ghoneim, Yousef Ibrahim Daradkeh, Junaid Asghar, and Elsayed Tag Eldin. "Intelligent Hybrid Deep Learning Model for Breast Cancer Detection." Electronics 11, no. 17 (2022): 2767.

[22] Tahmooresi, Maryam, A. Afshar, B. Bashari Rad, K. B. Nowshath, and M. A. Bamiah. "Early detection of breast cancer using machine learning techniques." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 10, no. 3-2 (2018): 21-27.

[23] Mojrian, Sanaz, Gergo Pinter, Javad Hassannataj Joloudari, Imre Felde, Akos Szabo-Gali, Laszlo Nadai, and Amir Mosavi. "Hybrid machine learning model of extreme learning machine radial basis function for breast cancer detection and diagnosis; a multilayer fuzzy expert system." In 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1-7. IEEE, 2020.

[24] Sarkar, Suvobrata, and Kalyani Mali. "Breast Cancer Subtypes Classification with Hybrid Machine Learning Model." Methods of Information in Medicine 61, no. 03/04 (2022): 068- 083.

[25] Mangukiya, Manav, Anuj Vaghani, and Meet Savani. "Breast cancer detection with machine learning." International Journal for Research in Applied Science and Engineering Technology 10, no. 2 (2022): 141-145.

T17_Version2.pdf

ORIGINALITY REPORT

22%	19%	13%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.cse.griet.ac.in Internet Source	2%
2	www.mdpi.com Internet Source	1%
3	Submitted to The Robert Gordon University Student Paper	1%
4	www.researchgate.net Internet Source	1%
5	www.irjmets.com Internet Source	1%
6	Submitted to Sheffield Hallam University Student Paper	1%
7	Submitted to University of Northumbria at Newcastle Student Paper	<1%
8	Submitted to University of North Texas Student Paper	<1%
9	Submitted to SP Jain School of Global Management	<1%