IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.5, May 2020 E-ISSN: 2347-2693

Multi-Class Cancer Classification Using Dimensionally-Reduced Breast Cancer Data

Jency Gracy Bai A.[1*], Lathikaa Sri M.[2], Jayalakshmi M.[3], Harinii M.[4], K. Amshakala[5]

[1,2,3,4,5]Dept of Computer Science and Engineering, Coimbatore Institute Of Technology, Coimbatore, Tamil Nadu, INDIA

DOI: https://doi.org/10.26438/ijcse/v8i5.6169 | Available online at: www.ijcseonline.org

Received: 06/May/2020, Accepted: 18/May/2020, Published: 31/May/2020

Abstract – **Breast cancer is an uncontrolled growth of breast cells and the most common invasive cancer in women, the second leading cause of cancer death in women next to lung cancer.** Cancer starts from breast and spreads to other parts of the body. People are unable to identify the disease before it becomes dangerous. It can be cured if the disease is identified at an earlier stage. Awareness of breast cancer, public attentiveness, and advancement in breast imaging has made a positive impact on the identification and screening of breast cancer. The interpretation of a tumor image is taken from patients and stored in datasets. This study suggests a feature extraction method such as PCA (Principal Component Analysis) which is used for pre-processing the data and extracting the most relevant features. Several classifiers like K-Nearest Neighbour (KNN), Naïve Bayes (NB), Linear Support Vector Machine(L-SVM), Gaussian Kernel Support Vector Machine(K-SVM), Logistic Regression(LR) are used to build machine learning model, among these classifiers Linear kernel Support Vector Machine (L-SVM) gives better accuracy. The proposed system uses a Linear kernel Support vector machine(L-SVM) to perform staging. The objective of the project is to carry out dimensionality reduction on cancer datasets and to build a predictive model for multi-class cancer stage classification using a linear kernel SVM classifier.

Keywords- Classification Techniques, Feature extraction, Principal Component Analysis(PCA) k-Nearest Neighbor (KNN), Linear Support Vector Machine (L-SVM), Gaussian Kernel Support Vector Machine(K-SVM) , Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR).

I. INTRODUCTION

Breast cancer is the most common cancer among women, impacting 2.1 million women each year and also causes the greatest number of cancer deaths among women. In 2019, it is estimated that 6,27,000 women died from breast cancer. To improve breast cancer diagnosis outcomes and survival, early detection is critical. Early diagnosis and screening are the two detection strategies for breast cancer. The early diagnosis of cancer was not possible due to limited medical resources. The early detection of cancer type has become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The strategy of classifying cancer patients into the high or low-risk category has led many research teams, from the biomedical and bioinformatics field, to study the application of machine learning (ML) methods. Machine Learning models are getting better than pathologists at accurately predicting the growth of cancer. Therefore, these techniques have been utilized to represent the progression and treatment of cancerous conditions. Also, the ability of ML tools detect key features from complex datasets reveals their importance Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs), and Decision Trees (DTs) are the techniques, widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. The objective of the project is to carry out dimensionality reduction on cancer datasets and build a predictive model for multi-class cancer stage classification using an SVM classifier.

This paper has been organized in eight different sections. In section I, the introduction part is detailed. Section II explains the work done by reputed authors in this domain. In section III, the ML techniques have been discussed for the classification of the data. Section IV deals with the proposed work. Results are presented in Section V. The last part of the paper is followed by the conclusion, future work, references as the VI, VII and VIII section respectively.

Machine Learning

Machine learning is a subfield of artificial intelligence that allows systems to get access data themselves and the ability to automatically learn and improve from experience without being explicitly programmed. This is done through ML algorithms and statistical models. The types of machine learning methods are,

- Unsupervised Learning
- Supervised Learning

Unsupervised Learning

In Unsupervised Learning, input data(X) only present and there is no corresponding output variable. The goal of the unsupervised learning is to model the underlying structure or distribution in the data to learn more about the data. Clustering and Association are the main categories of unsupervised learning methodologies.

- **Clustering**: Inherent groupings of data will be found in

clustering problems such as grouping customers by purchasing behavior.
- **Association**: Associations and relationships will be found in large portions of data such as people who buy X also tend to buy Y.

Supervised Learning

Supervised learning is the task of learning a function that maps an input to an output based on example input-output pairs.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that the output variable(Y) can be predicted from new input data(X). Regression and classification are the main categories of supervised learning methodologies.

- **Regression**: A regression problem is when the output variable is a real value, such as "dollars" or "weight".
- **Classification:** A classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease".

Types of some classification algorithms in machine learning are Logistic Regression, Nearest Neighbor, Support Vector Machine and Artificial Neural Network

Staging

Tumor size is an important factor in breast cancer staging, and it can affect a person's treatment options and outlook. However, the size of the tumor is only one of the factors that doctors consider when staging a person's breast cancer. T1a means that breast cancer is at a very early stage and has not yet spread. T4 is late-stage breast cancer, in which cancer has spread to other parts of the body.

Stage 1:

These are small tumors that either has not spread to the lymph nodes or are only affecting a small area of the sentinel lymph node. It has tumor size from 1mm to 10 mm.

Stage 2:

These are large tumors that have spread to some nearby lymph nodes. It has tumor size from 10mm to 20 mm.

Stage 3:

These tumors are large or growing into surrounding tissues, such as breast skin, muscle, and lymph nodes. It has tumor size from 20mm to 50 mm.

Stage 4:

These are tumors that started in the breast but have spread to other parts of the body. It has tumor size larger than 50mm.

II. LITERATURE SURVEY

Lot of works has been done in applying dimensionality reduction and ML based techniques for cancer diagnosis. This section explains about the study of such important research papers.

1) Breast Cancer Prediction system (2018)

A detailed survey on "Breast Cancer Prediction system" is published in the year 2018 by Madhu Kumari and Vijendra Singh in International Conference on Computational Intelligence and Data Science (IJCSES)[1]. They proposed a classification model with boosted accuracy to predict the breast cancer patient. They used filter method (Pearson's linear correlation coefficient measure) to select most relevant features from the dataset. . Performance of the system is evaluated by considering the actual and predicted classification. Accuracy of the system is calculated by using the confusion matrix obtained for the classifier used.

2) Dimensionality Reduction in Gene Expression Data sets (2019)

A detailed survey on "Dimensionality Reduction in Gene Expression Data sets" is published in the year 2019 by Jovani Taveira De Souza, Antonio Carlos De Francisco and Dayana Carla De Macedo in IEEE Access[2]. They proposed two approaches to reduce the dimensions in gene expression data using Attribute selection and principal Component analysis. The Attribute selection method is divided into two approaches filter and wrapper approach. Filter method is used to rank the features based on statistical calculations techniques used most commonly to evaluate subsets, such as correlation-based feature selection (CFS), consistency-based subset evaluation (CSE), minimum redundancy maximum relevance (mRMR), and fast correlation-based filter (FCBF). Wrapper approach performs the evaluation criteria of the data. The WEKA software is used to analyze the performance of seven different classifier such as NB, J48, SVM, 1-NN, 3-NN, 5-NN,and 7-NN. Finally, they measured the performance of the classifier.

3) Machine Learning Classification Techniques for Breast Cancer Diagnosis

A detailed survey on "Machine Learning Classification Techniques for Breast Cancer Diagnosis" is published in the year 2019 by David A. Omondiagbe, Shanmugam Veeramani and Amandeep S. Sidhu in IOP Conference series on Materials Science and Engineering[3].The main aim of this study is to integrate feature selection and feature extraction methods in machine learning classification techniques (Support Vector Machine, Artificial Neural Network and Naive Bayes) to compare their performance to identify the most suitable approach for breast cancer diagnosis and to solve classification efficiency. Simulation results showed that SVM-LDA and NN-LDA outperforms the other ML classifier models.

4) Comparative Study of Classification Techniques for Breast Cancer Diagnosis (2019)

A detailed survey on "Comparative Study of Classification Techniques for Breast Cancer Diagnosis" is published in the year 2019 by Ajay Kumar, R. Sushil and A. K. Tiwari in International Journal of Computer Sciences and Engineering [4]. They proposed a comparative study on major and popular classification techniques for performance based analysis. Two

Wisconsin dataset WBC (Original) and WDBC (Diagnostic) are used for classification purpose. The main and popular algorithm has been taken to analyse the data and predict the accuracy of being cancerous or not. The accuracy of the classification algorithm is determined on the basis of various parameters such as TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area etc. This study evaluates that the Bayesian Network gives the best accuracy with less featured dataset while Support Vector Machine gives best accuracy for more featured dataset.

III. MACHINE LEARNING (ML) TECHNIQUESFOR CLASSIFICATION

There are many algorithms in Machine Learning viz. Support Vector Machine (SVM), K-Nearest Neighbors(KNN), Bayesian Network (BN), and Naïve Bayes (NB) used for data classification. Following are the key features of these algorithms

A. *Support Vector Machine*

SVM is a Supervised Machine Learning classification algorithm. In the case of linearly separable 2D data al typical Machine Learning algorithm will try to find the boundary that divides the data in such a way that misclassification error is minimized. SVM works differently from other classification algorithms in the way that it chooses decision boundary that maximizes the distance from the nearest neighbour data points of classes. It not only finds the decision boundary; it finds the optimal decision boundary. The main idea behind SVM is to find the Maximum Marginal Hyperplane (MMH) that can best separate the dataset into classes. It iteratively performs to find the optimal hyperplane that can minimize the misclassification error.

Types of SVM

1. Linear SVM

Linear SVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time which scales linearly with the size of the training data set. It can be used as normal dot product between two given observations.

$$K(x, xi) = sum(x * xi)$$

2. Gaussian Kernel SVM

The SVM classifier with the Gaussian kernel is simply a weighted linear combination of the kernel function computed between a data point and each of the support vectors. The Gaussian kernel transforms the dot product in the infinite dimensional space into the Gaussian function of the distance between points in the data space.

$$K(x,xi) = exp(-gamma * sum((x – xi^2))$$

B. *K-Nearest Neighbors*

KNN is one of the foremost easy and simple data processing techniques. It is known as Memory-Based Classification as the coaching examples have to be in the memory at run-time. When KNN is employed for classification, the output is calculated because of the category with the very best frequency from the K-most similar instances. Every instance in essence votes for their class and therefore the class with the foremost votes are taken for the prediction.

C. *Naive Bayes Classifiers*

Naive Bayes classifiers are a group of classification algorithms supported Bayes Theorem. it's not one algorithmic rule however a family of algorithms wherever all of them share a typical principle, each try of options being classified is freelance of every different. Bayes theorem uses the contingent probability that successively uses previous information to calculate the probability that a future event can happen. Naive Bayes algorithmic program is employed for binary and multi-category classification and might even be trained small low information set that could be a huge advantage. It's additionally terribly quicker and climbable.

D. *Logistic Regression*

Logistic Regression is a linear algorithm (with a non-linear transform in output). Logistic regression assumes a linear relationship between the input variables with the output. In logistic regression, the model can overfit if there are multiple highly correlated inputs. Logistic regression intended binary classification problem, it will predict the probability of instances belonging to the target class which can be of 0 or 1 classification.

For Example,

- To predict whether an email is spam (1) or (0)
- To predict whether the tumor is malignant (1) or not (0)

Types of Logistic Regression are,

1.Binomial: target variables can be only 2 possible types: "0" or "1" which may be represented by "win" vs "loss", "pass"vs"fail".

2. Multinomial: The target variable can have 3 or more possible types that are not ordered (i.e types have no quantitative significance) like "disease A" vs "disease B" vs "diseaseC".

3. Ordinal: : It deals with target variables with ordered categories. For example, a test score can be categorized as "very poor", "poor", "very good", "good". Here each category can be given score like 0,1,2,3.

IV. PROPOSED WORK

System Architecture for multi class classification

Step 1: As the diagnostic dataset (WDBC) has high dimensions, it is preprocessed using PCA. The reduced Set will be given to the models such as KNN, NB, LR, K-SVM, L-SVM. The performance of the models is validated based on the accuracy and the SVM classifier gives higher accuracy than the other models. As this dataset doesn't contain tumor size, the cancer stages can't be predicted using this diagnostic dataset. Thus the prognostic dataset has come into the picture.

Step 2: Initially, the preprocessing steps are done on the prognostic dataset and the reduced set will be given to the models such as KNN, NB, LR, K-SVM, L-SVM.As the SVM classifier gives higher accuracy than the other models, it is selected in the validation for predicting the cancer stages and these cancer stages are predicted based on the tumor size.

Dimensionality reduction in breast cancer dataset predictive model consists of the following steps:

- To scale the data and to extract the features from the original dataset using PCA.
- Create training and testing datasets.
- Apply machine learning techniques to the training set.
- Generate the predictive model.
- Evaluate model using testing dataset.
- Compare performance among the machine learning techniques.

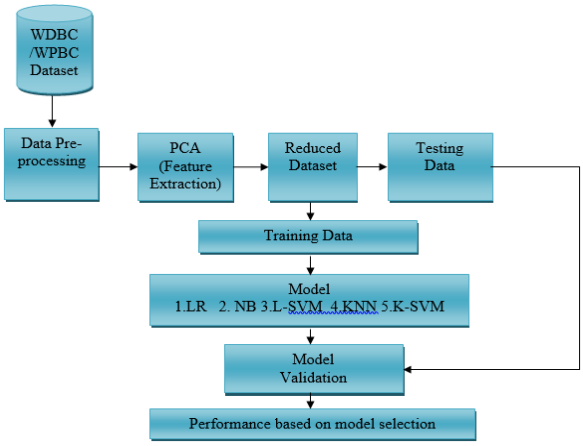

Fig. 1 Experiments are carried out with different datasets. Diagnosis is done on the WDBC dataset and prognosis is done on the WPBC dataset along with Multi-Class Classification.

Dataset Description

Dataset 1:

The Breast Cancer Wisconsin (Diagnostic) dataset was taken from the UCI Machine Learning Repository. The dataset has a dimension of 569 x 32, it has 30 real attributes and one numeric attribute (id field), and one categorical attribute, which is a class label. This dataset has two class values for diagnosis they are M (Malignant) and B (Benign). Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

They describe the characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus: The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

This dataset doesn't contain any missing attribute values. The class distribution is 357 benign and 212 malignant.

Dataset 2:

The Breast Cancer Wisconsin Prognostic Breast Cancer (WPBC) dataset was downloaded from the University of California, Irvine (UCI) Machine learning repository, which is available through open access. The data set consists of recordings collected from biopsies of real patients in different hospitals of Wisconsin. The dataset has a dimension of 198 x 35, it has 33 real attributes and one numeric attribute (id field), and one categorical attribute, which is a class label. This dataset has two class values for diagnosis they are R (Recurrence) and N (Non-Recurrence).

Ten real-valued features are computed for each cell nucleus.

This dataset doesn't contain any missing attribute values. The class distribution is 150 Non- Recurrence and 48 Recurrence.

Data pre-processing

- Data pre-processing is performed to improve the quality of a dataset to get clean data which can be useful for modeling. There are several processes involved in data pre-processing. These processes include data cleaning, feature extraction, etc.
- Data cleaning involves removing noise and the inconsistencies which are present in the data, thereby improving the quality of the data.
- Data selection is to reduce features by employing feature extraction methods on the training dataset. Feature extraction reduces the number of dimensions by transforming features in high dimensional space into fewer dimensions. Principal component analysis (PCA) is a feature extraction method used to transforms the original dataset into a reduced number of derived variables that do not correlate, and they are called principal components (PC). Performing PCA, the cumulative variance is used in reducing the feature dimension of the dataset. This chooses the number of principal components according to the Eigenvalue sizes or the proportion of variance each principal component.
- During the data pre-processing stage, the data is partitioned into the training dataset and validation dataset. The training dataset is used in training the machine learning model, while the validation dataset is used during the prediction stage.

ML Model Building

After pre-processing the data, the next stage is to apply ML classification techniques on the reduced data. During this stage, reduced data will be used to train and build the ML model. The ML classification algorithms considered in this work are LR, Linear SVM, Kernel SVM, NB and K-NN.

In LR, the regression co-efficient of the model are calculated by maximum likelihood and the features or independent variables. The probability of binary outcome is calculated and using the probabilities the breast cancer is classified into two categories malignant (1) or benign

(0). In liner SVM, the algorithm creates a line or hyperplane which separates the cancer diagnosis into two classes malignant (1) and benign (0). In Gaussian Kernel SVM, projecting the data to a higher dimensional space where the points are linearly separated. In NB, is used to predict the probability or each class such as the probability that given record belongs to particular class. The class with highest probability is considered as the most likely class. In KNN, the class probabilities are calculated within the set of most similar instances for a new data instance malignant (1) and benign (0).

The training data which have the entire feature attributes is used to train the models to classify the data into benign and malignant tumors.

Model Building and Performance Evaluation for WPBC Dataset

After pre-processing the data, the next stage is to apply ML classification techniques to the processed data. During this stage, the processed data will be used to train and build the ML model. The training data which have the entire feature attributes are used to train the models to classify the data into recurrence and non-recurrence tumors. After training the model the validation set is given to the model to evaluate the performance of the classification techniques used.

In the prediction phase the test the dataset is used to assess the performance of the models in classifying recurrence and non-recurrence tumors. The different performance metric is used to evaluate the performance of the ML model are confusion matrix, accuracy, precision, recall, and F1-score

Stages of Cancer

The stage of breast cancer is determined by cancer's characteristics, such as how large it is and whether or not it has hormone receptors. Staging helps to describe whether cancer is located or it has spread or if it is affecting other parts of the body.

The tumor size (T) category describes the original primary tumor:

- TX: means the tumor can't be assessed.
- T0: There is no evidence of the primary tumor in the breast.
- T1: The tumor in the breast is 10 millimeters (mm) or smaller in size at its widest area. The T1 stage is then broken into 2 sub-stages depending on the size of the tumor.
  - ❖ T1a is a type which has tumor size larger than 1 mm and less than 5 mm.
  - ❖ T1b is a type that has a tumor size larger than 5 mm and less than 10 mm.
- T2: The tumor size is between 10mm and 20mm.
- T3: The tumor size is between 20mm and 50mm.
- T4: The tumor size is larger than 50mm.

Multi-class Cancer Classification

A A multi-class classification system includes a supervised learning methodology. To learn a function using a labeled training set, consisting of several features and a class label in its simplest form. The classification problem is the optimization of a target function so that the input (X_i) transformed to an output (Y_i), where X_i €X, Y_i € Y, Y = {1,2,..,K} When K>2, where *K*is the number of classes.

V.RESULTS

Performance Evaluation

The next step after implementing machine learning models is to seek out how effective the models performed on the datasets. This is carried out by running the classification models on the validation dataset which was set earlier in pre-processing stage. Then the evaluation was done by comparing the model results with the real data value.

In order to determine and compare the performances of the different algorithms, several metrics have been used. The confusion matrix is built in comparing the predicted results with the actual values. The data in the matrix is used to compute the performance of the classifier.

Table 1 Confusion Matrix

		Predicated Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

- True Positive (TP): These are the occurrences where both the predictive and actual class is True (1).
- True Negative (TN): True negatives are the occurrences where both the predicted class and actual class is False (0).
- False Negative (FN): These are occurrences where the predicted class is False (0) but actual class is True (1).
- False Positive (FP): False positives are theoccurrences where the predicted class is True (1) while the actual class is False (0).

Different performances metric that can be used to evaluate the performance of the ML model are,

- Accuracy: Evaluation of classification models is done by one of the metrics called accuracy. Accuracy is the fraction of prediction. It determines the number of correct predictions over the total number of predictions made by the model. The formula of accuracy is:

  Accuracy = (TP+TN) / (TP+TN +FP+FN)
- Recall: It is a measure of the proportion of patients that were predicted to have the complications among

those patients that actually have the complications. Recall can be calculated as follows:

Recall = TP / (TP+TN)

- Precision: It is described as a measure of proportion of patients that actually have complications among those classified to have complications by the model. The formula for Precision is as follows:

Precision = TP / (TP+FP)

- F1-Score: Weighted average of precision and recall is known as F1score. It is calculated as follows:

F1Score=((Precision ∗Recall)/(Precision+Recall))*2

Result Analysis for WDBC dataset

The main objective is to improve the performance of classification and increasing diagnosis accuracy by reducing the dimensions of features in WDBC dataset by using the PCA.

Dimensionality Reduction in WDBC dataset

Feature extraction (PCA) methods were performed on the data to reduce the dimension of features, thereby producing reduced versions of the original dataset.

Feature extraction is done by transforming data from the original high dimensional one to new low dimensional one based on calculating the proportion of the variance explained for each feature and by picking a threshold and adding the feature until that threshold is reached (this is done until proportion of variance explained hits or exceeds 80%).

The main purposes of performing feature extraction are to improve the prediction performance and ensure faster prediction.

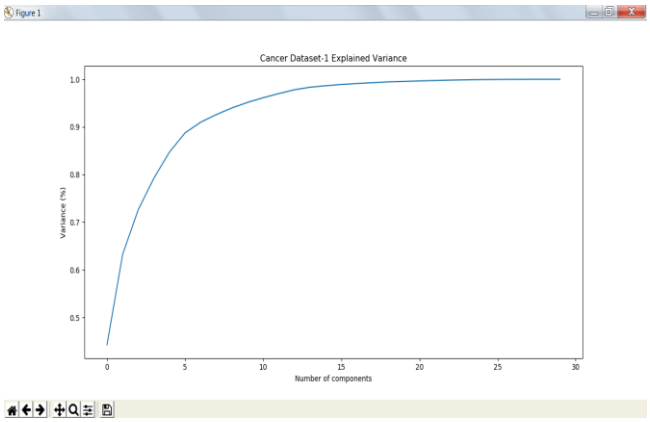

Fig. 2 principal component vs. explained variance ratio curve

The graph illustrates the cumulative variance over features in WDBC dataset. From the plot, the number of principal component can be easily selected based on the threshold and the remaining features are eliminated based on the cumulative explained variance ratio.

Training and testing set in WDBC dataset

Then the model is trained using ML algorithms such as SVM, LR, NB and KNN. After the learning and training phase, the next step is to test the intelligence of the model, for this purpose the test data is used. The test set has 114 data points with 2 independent features and one target label. To test the trained model, the test set with the exclusion of the target label is fed to the model for the model to make some predictions. The predictions (predicted outcome) from the model will be used to match the actual outcomes of the test set.

```
======= RESTART: C:\Users\sony\Favorites\Desktop\project work\WDBC.py =======

('Train set and Test set before PCA', (455, 30), (114, 30))

('Train set and Test set after PCA', (455, 2), (114, 2))
```

Fig. 3

Training and Testing set before and after PCA

For classification, the pre-processed data is fed to the ML classifier models. The data is split into two parts; the training set (80% of data) and test set (20% of data). To train the model using training set and to evaluate the performance of the model using test set. Before applying the PCA the training set has 455 data points and the test set is 144 data points with 30 features. After applying the PCA the training set has 455 data points and the test set is 144 data points with 2 features.

Performance of ML classification Model in WDBC dataset

Dimensionality reduction affects the performance of a ML algorithm. The simulation results revealed that reducing the high dimensionality of a dataset by PCA method can improve the performance of the classification models. Five different methods were used in the classification stage to train the dataset: LR, Linear SVM, Kernel SVM, NB, K-NN. The results of the different classifier model are viewed and compared with each other. Four performance metrics namely confusion matrix, precision, recall, and f1-score are used to evaluate the performance of the trained models.

The table 4 below demonstrates the results of different metrics for the algorithms to predict Breast Cancer with Principal Component Analysis:

Machine Learning Model	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-Score
LR	70	2	1	41	0.9736	0.9761	0.9534	0.9647
KNN	69	3	2	40	0.9561	0.9523	0.9302	0.9411
Linear SVM	70	2	1	41	0.9736	0.9761	0.9534	0.9647
NB	70	5	0	38	0.9561	1.0	0.8872	0.9382
Kernel SVM	70	2	1	41	0.9736	0.9761	0.9534	0.9647

Fig.4 Performance comparison of machine learning models in WDBC dataset

The results of the different classifier models are viewed

and compared with each other. From the results displayed in table 4, we can observe that LR + PCA, L-SVM + PCA and K-SVM + PCA combinations obtained the best performance of 97.39% in terms of accuracy. However, NB + PCA, combination and KNN + PCA combination obtained the best performance of 95.61% in terms of accuracy.

Result Analysis for WPBC dataset

The main objective is to improve the performance of classification and increasing diagnosis accuracy in WPBC dataset. The ML classification algorithm with high performance accuracy is used to classify the stages of cancer.

Training and Testing set in WPBC dataset

Then the model is trained using ML algorithms such as SVM, LR, NB and KNN. After the learning and training phase, the next step is to test the intelligence of the model, for this purpose the test data is used. The test set has 40 data points with 34 independent features and one target label. To test the trained model, the test set with the exclusion of the target label is fed to the model for the model to make some predictions. The predictions from the model will be used to match the actual outcomes of the test set. After prediction, the performance metric of classifier model are compared with ML algorithms SVM, NB, KNN and LR.

```
Python 2.7.12 (v2.7.12:d33e0cf91556, Jun 27 2016, 15:19:22) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
====== RESTART: C:\Users\sony\Favorites\Desktop\project work\WPBC.py ======


('Training and Testing set in WPBC dataset', (158, 34), (40, 34))
```

Fig. 5 Training and Testing set in WPBC dataset

Initially the data is split into two parts; the training set (80% of data) and test set (20% of data). To train the model using training set and to evaluate the performance of the model using test set. The training set has 158 data points and the test set is 40 data points with 34 features.

Performance of ML classification Model in WPBC dataset

The performance of the classifiers algorithm in breast cancer WPBC dataset can be evaluated from the analysis of confusion matrix, Accuracy, Precision, Recall, and F1 Score parameters are calculated.

The table 5 below demonstrates the results of different metrics for the algorithms to predict Breast Cancer.

Machine Learning Model	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-Score
LR	28	4	4	4	0.8	0.5	0.5	0.5
KNN	30	7	2	1	0.775	0.3333	0.125	0.1818
Linear SVM	29	3	3	5	0.85	0.625	0.625	0.625
NB	21	5	11	3	0.6	0.21	0.37	0.27
Kernel SVM	32	8	0	0	0.8	0.0	0.0	0.0

Fig. 6 Performance comparison of machine learning models in WPBC dataset

The results of the different classifier models are viewed and compared with each other. From the results displayed in table 5, we can observe that L-SVM algorithm obtained the best performance of 85% in terms of accuracy. However, KNN, LR, NB and Kernel SVM algorithms obtained the lowest performance in terms of precision, recall and f1-score.

Multiclass Cancer Classification using L-SVM

The Linear SVM classification model is selected based on the performance accuracy for multiclass classification. The testing data are evaluated by passing it to the optimized trained SVM with linear kernel. Using Linear SVM the stages of breast cancer T1a, T1b, T2, T3 and T4 are classified based on the size of the tumour in WPBS dataset. The performance of Multiclass Linear SVM is measured by precision, recall and F1-Score metrics.

Here to train the model using training set and to evaluate the performance of the model using test set. The training set has (189, 179, 169 and 159) data points and the test set is (9, 19, 29 and 39) data points with 34 features to classify the stages of cancer and the performance of Linear SVM is measured using precision, recall and F1 Score.

The table 6 below demonstrates the results of different performance metrics for the algorithm Linear SVM and outcomes of stages of cancer during different test scenarios.

Test set (%)	Test data randomly selected	Precision	Recall	F1-Score	T1a	T1b	T2	T3	T4
5	9	1.0	1.0	1.0	2	0	1	5	1
10	19	1.0	1.0	1.0	2	2	6	8	1
15	29	0.98	0.966	0.976	2	5	9	10	3
20	39	0.98	0.966	0.976	2	5	11	16	5

Fig. 7 Result of Multiclass Cancer Classification

Thus the results illustrates that linear support vector machine gives the optimal performance in classifying the stages of breast cancer.

VI.CONCLUSION

Cancer has become the leading cause of death worldwide. The most effective way to reduce cancer deaths is to detect it earlier. An important challenge in machine learning area is to build precise and computationally efficient classifiers for Medical applications. The Proposed system helps to remove the insignificant features from the large Wisconsin Diagnostic Breast Cancer (WDBC) dataset (569 instances) and the Wisconsin Prognostic Breast Cancer (WPBC) dataset (198 instances) which is a standard dataset utilized in the diagnosis of breast cancer. The experimental result shows that the number of features for classification of breast cancer from the original WBC dataset can be reduced by the proposed feature extraction method and still accurate results are obtained.

Step wise principal component analysis was used to explore the effects of the test parameters which are significantly affecting the patient's health condition. Feature extraction is done by transforming data from the original high dimensional one to new low dimensional one based on calculating the proportion of the variance explained for each feature and by picking a threshold and adding the feature until that threshold is reached (this is done until proportion of variance explained hits or exceeds 80%). The reduced dataset is given as an input to SVM which is used for training and testing using models (the training and testing dataset are randomly split into 80:20 ratio). LR, L-SVM, K-SVM, NB, KNN models are trained using the extracted feature in the original dataset and used to classify the test data. These models are used in cancer diagnosis where it says whether the tumor is malignant or benign and helps to make accurate breast cancer classification. Model accuracy is verified using confusion matrix. Thus the proposed system uses one of the best models to identify whether the patient is having breast cancer or not and if the result are positive it will classify the stages of breast cancer. As cancer stages are predicted using size this will tell the doctor about the growth and will help the doctor to decide on the appropriate treatment for the affected patients. Therefore the model makes more accurate prediction and also helps in early detection of cancer stages which enables to reduce deaths due to delayed diagnosis of cancer.

VII. FUTURE WORK

The approach is applicable to even large dataset to perform similar analysis .It can be used for efficient selection of parameters for future projects. The idea of applying other feature selection on the currently used models is also under consideration, such as the Recursive Feature Elimination and the Gray Wolf Optimization (GWO). Future work can also consider comparing more ML algorithms used for breast cancer diagnosis. Various cancer types can also be considered in future works. This study helps in making more effective and reliable disease prediction and diagnostic system which will contribute towards better healthcare system into a potential practical method for aiding and assisting doctors with quick second opinion in diagnosing breast cancer.

REFERENCES

[1] MadhuKumari and Vijendra Singh**,** "Breast Cancer Prediction system ". In the proceedings of the 2018 International Conference on Computational Intelligence and Data Science (IJCSES), India, Vol.132, p.371-376, 2018.

[2] David A. Omondiagbe, Shanmugam Veeramani and Amandeep S. Sidhu ,"Machine Learning Classification Techniques for Breast Cancer Diagnosis". In the proceedings of the 2019 IOP Conference series on Materials Science and Engineering , Vol .495, 2019.

[3] J. Taveira De Souza, A. Carlos De Francisco and D. Carla De Macedo, "Dimensionality Reduction in Gene Expression Data Sets," in IEEE Access, vol. 7, pp. 61136-61144, 2019, doi: 10.1109/ACCESS.2019.2915519.

[4] Ajay Kumar, R. Sushil, A. K. Tiwari, "Comparative Study of Classification Techniques for Breast Cancer Diagnosis," *International Journal of Computer Sciences and Engineering*, Vol.7, Issue.1, pp.234-240, 2019.

[5] Pritom AI, Munshi MAR, Sabab SA, Shihab S. "Predicting breast cancer recurrence using effective classification and feature selection technique". In 19th international conference on computer and information technology (ICCIT). New York: IEEE; 2016. p. 310–4.

[6] Lu J, Keech M. "Emerging technologies for health data analytics research: a conceptual architecture". In 26th international workshop on database and expert systems applications (DEXA). IEEE; 2015. p. 225–9.

[7] Chaurasia V, Pal S. "A novel approach for breast cancer detection using data mining techniques". In International journal of innovative research in computer and communication engineering (an ISO 3297: 2007 certified organization), vol. 2; 2017.

[8] Kumar UK, Nikhil MS, Sumangali K." Prediction of breast cancer using voting classifier technique". In IEEE international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM).NewYork:IEEE;2017.p.108–14.

[9] Ajay Kumar, R. Sushil , A. K. Tiwari." Comparative Study of Classification Techniques for Breast Cancer Diagnosis". International Journal of Computer Science and Engineering(IJCSE), Vol.-7, Issue-1, p.234-240 Jan 2019.

[10] Vikas S, Thimmaraju S N. "Breast Cancer Diagnosis and Classification Using Support vector machines With Diverse Datasets". International Journal of Computer Science and Engineering(IJCSE), Vol.-7, Issue-4, p.442-446,April 2019.

AUTHORS PROFILE

Ms A.Jency Gracy Bai currently pursuing engineering in the stream of computer science at Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India and also completed diploma in computer science and engineering at Sri Ranganathar Institute of Polytechnic College(2014-2017),Coimbatore, Tamil Nadu, India.

Ms M.Lathikaa Sri currently pursuing engineering in the stream of computer science at Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India and also completed diploma in Computer Technology at Sri Krishna Polytechnic College (2014-2017), Coimbatore, Tamil Nadu, India.

Ms Jayalakshmi M currently pursuing B.E in Computer Science and Engineering at Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.

Ms Harinii M currently pursuing B.E in Computer Science and Engineering at Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.

Dr.K.Amshakala joined Coimbatore Institute of Technology in June 2002 and she is currently working as Associate Professor in the Department of Computer Science and Engineering.
She received Ph.D degree from Anna University, Chennai in the faculty of Information and Communication Technology in the year 2014. Her research thesis was on extracting semantics in the form of data dependencies from databases and applying it for data integration. Her research interests include data analytics, Artificial Intelligence and IoT. She has published articles in international and national journals and currently guiding UG,PG and Ph.D scholars in the area of data analytics and AI.