A NOVEL APPROACH ON LUNG CANCER PREDICTION USING ENHANCED PCAWITH SMOTE

[1]S.Premkumar, [2]Dr.N.Revathy, [3]Dr.T.Guhan
[2]Professor, [1]Research Scholar, [3]Associate Professor
[1,2]Department of Computer Science,
[3]Department of Information Technology,
[1,2]Hindusthan College of Arts & Science, Coimbatore.
[3]Karpagam College of Engineering, Coimbatore.
[1] revathy.n@hicas.ac.in

Abstract

The performance of the classification models are highly degraded on large dataset with high dimension. The high dimensional dataset has both relevant as well as irrelevant features results in performance degradation of the classification model. Moreover, more number of datasets is imbalanced in nature. The imbalanced data also poses hindrance for the classification models. The imbalanced data set leads classification performance bias towards the majority class. In this work, PCA with smote is applied to derive an effective subset of lung dataset. The PCA reduces the dimensionality of the data set into lower dimension. The SMOTE is the technique that creates synthetic samples on the dataset. PCA eliminates the irrelevant features from the dataset and SMOTE creates new synthetic samples to increase the number of representative samples in minority class. Finally, SVM classifier is applied on the pre-processed dataset as well as performanceof the model is compared using evaluation metrics. The experimental results proved the effectives of the proposed methodology in terms of accuracy, precision, recall, false positive rate.

Keywords: PCA, SMOTE , SVM, Synthetic Sampling

1. INTRODUCTION

Data miming is the task of exploring or identifying useful and hidden patterns from vast amount of data using data mining algorithms. Machine learning is the process on which machines are trained to predict new samples or data. Classification is the foremost task in data mining and machine learning. Classification belongs to supervised machine learning algorithm that classifies samples into the pre-defined classes or groups. Data plays an important role in machine learning. In classification, the dataset is divided into training and testing as well as training data is used to build the classification model. Finally, the test data is used to test the performance of the classification model. Data poses various challenges for the machine learning scenario for effective learning. There are various issues related on data like missing data, redundant data, high dimensional data and lack of representative samples classes. These issues pose challenges for the model to perform better. The research work concentrates on two challenges based on data. The pre-processed dataset with the help of PCA and SMOTE will improve the classification performance. The research work considers the lung dataset that has issues like high dimensionality and imbalance. Cancer is the leading cause for death globally. Early detection and prevention of carcinoma or cancer can save human life [1]. There are various types of cancer exists like liver, stomach, breast, and lung. Among various categories, lung cancer is the foremost cause for death. The two categories of lung cancer are small cell as well as non-small cell. Tobacco is the foremost cause for lung cancer. Even, non-smokers also impacted due to the air pollution, chemical exposures or working atmosphere and more. There are different stages of lung cancer like stage 1, stage 2, stage 3, stage 3A, stage 3B, stage 4 etc. Detection using traditional methods like computed tomography and other techniques like bronchoscopic are not yielding better prediction results [2]. Machine learning plays an important role in predicting the lung cancer early and increase the survival rate of human.

The dataset contains both relevant as well as irrelevant features. Moreover, the number of representative samples for the minority class is low compare with the majority class. The proposed methodology will make use of the PCA with SMOTE to create an effective dataset to improve the classification performance.

2. PROBLEM OBJECTIVES

To protect the life of human being from one of the deadly disease carcinoma is vital for the society. Accurate and effective prediction is highly required for the medical industry and society. The work makes use of machine learning algorithms to build the model with the use of UCI dataset.

3. LITERATURE SURVEY

Features define the characteristics of data. Features play an important role on the performance of machine learning models. The performance of the machine learning algorithm may be degraded due to having more number of features in the data set. Similarly, when the dataset does not have adequate number of features, the model will likely to under-fit to predict the data. In general, too much of data leads to over-fit or under-fit as well as lack of sufficient data leads to under-fit. This situation is called as curse of dimensionality.

Dimensionality reduction techniques are essential to select optimal subset of data with reduced number of features to improve the accuracy of classification. Dimensionality reduction techniques can be of two forms such as feature selection and feature extraction. There are various issues can be raised while working with high dimensional dataset. Dimension represents the number of features of a dataset.

High dimensional data set may have more than hundreds of features. High dimensional dataset results in issues during analysis as well as visualization. It will be difficult for the analyst to derive patterns.

High dimensionality also leads to issues when training machine learning models. The aspects of curse of dimensionality are data sparsity and distance concentration. High dimensional dataset appears to be sparse and dissimilar as well as prevents from generalization. Supervised machine learning is trained for accurately predicting the outcome of the given data sample. Model generalization defines the ability of the machine learning models for predicting the unseen data. Distance concentration is the issue of convergence of all pair wise distance to the same value when the dimensionality increases.

Machine learning models like nearest neighbor or clustering make use of the distance based metrics for identifying the samples proximities. High dimensionality dataset results in lack of qualitative relevant proximity and similarity. High dimensionality results in infeasibility on optimization problems. The issues of curse of dimensionality can be tackled with the concepts such as feature selection and dimensionality reduction. Dimensionality reduction mainly falls into two major categories such as feature selection and extraction. Feature selection technique deals about selection or elimination of attributes based on its worthiness. Common feature selection methods are low variance filter, high correlation filter, multi-colinearity, feature ranking, and forward selection.

Low variance filter includes the following steps such as variance of the attributes in a dataset is compared, attributes with low variance will be discarded and attributes that do not possess much variance can be assumed as constant value and these attributes do not have any contribution to the predictability of the model. High correlation filter involves the following steps such as correlations between attributes are determined and one of the attribute in the analyzed pair will be eliminated when the correlation is high.

In the eliminated attribute, variability is gained using the retained attribute. Multi-collinearity deals about correlation among one or more number of independent attributes. Multi-collinearity means one of the independent variable can be determined by other independent variable. Multi-collinearity can be detected using the method called inflation factor.

The variables having high inflation factor mostly more than 10 usually discarded. Feature ranking is the task of ranking the features or attributes of the dataset based on their importance. One of the techniques like decision tree models such as CART – Classification and regression trees can be used to rank the attributes based on their features [3].

Forward selection is the technique of selecting or adding features one by one to build the machine learning model by analyzing its importance. While building, multi-linear regression model, features are added one by one based on its adjusted R2 value. The adjusted R2 value decides the importance of the selected attribute.

Feature extraction is the techniques are used to derive subset of attributes from high dimensional attributes. Different feature extraction techniques are independent component analysis, principal Component analysis, auto-encoder and partial least squares.

Principal component analysis

PCA is the dimensionality reduction technique. This technique performs transformation on high dimensional correlated data to convert into low dimensional uncorrelated data. Principal component analysis is the technique of transforming high correlated high dimensional data into uncorrelated lower dimensionality data.

In high dimensional data, lower dimensional principal component capture most of the information. In PCA, n principal components are derived by transforming n-dimensional data. PCA is an exploratory approach for reducing the dimensionality of the dataset from 3D to 2D. PCA is the linear transformation of dataset that specifies new coordinate rule as below the first axis shows the highest variance of the dataset and second axis demonstrates the next biggest variance and so on.

Purposes of principal component analysis are as follows visualizing high dimensionality data, for introducing improvements in classification problems, for obtaining compact descriptions, for capturing as much variance in the data as possible, for decreasing the number of dimensions of the dataset, for searching patterns in the high dimensionality dataset and discarding noise [4].

PCA is the task of transforming as well as reshaping high dimensional correlated variables into lower uncorrelated variables for capturing much of the variance in the dataset. PCA sorts the dimensions based on its importance, capturing the orthonormal basis for the data, eliminating the low significant dimension and focusing on uncorrelated as well as Gaussian components [5].

The main steps involved in PCA are PCA standardization, covariance matrix calculation, deriving eigen vectors and eigen values for the covariance matrix as well as vectors plotting on scaled data [8].

The analysis is based on observing various features of the dataset. Each record in the dataset has data in the form of vectors with the length of k. The number of features of a dataset is represented by n and number of records in the dataset is represented by k. The data set is in the form of k*n matrix. Each of the students belongs to the k-dimensional space. During PCA, some of the features of PCA are ignored and rests of the features are considered. PCA can ignore the collinear and linearly dependent features, constant features and noisy features [7]. PCA can consider the non-collinear features and features that are variable as well as high variance. Eigen vales and eigen vectors play crucial role in PCA. The source of the PCA is defined by the eigen vectors as well as eigen values of the covariance matrix. The direction of the new attribute is determined by the eigen vectors and the magnitude is determined by the eigen values [6].

3.1 Enhanced principal component analysis with SMOTE

PCA is an effective technique to extract data from set of features. It forms subset of features for reducing the dimensionality of the dataset. It transforms the high correlated and high dimensional dataset into low – correlate as well as low dimensional space [9]. It finds the best vector space for representing the data. The feature space derived by the eigen vector reduces the dimensionality of the original space results in reduced computation time to predict lung cancer patients [10]. The main objective of PCA is to reduce the high dimensional space into low dimensional data. It is considered as multivariate analysis method according to eigen vector. It is implemented by two different techniques. The first technique is attained by decomposing Eigen value of the covariance matrix. Another technique is attained by decomposing singular value of the data matrix [11]. The results of PCA are indicated as a factor or component scores as well as standardized component score weight [13]. Finally, the results can be expressed as set of eigen values [12]. After reducing the dimensionality the result is considered as Eigen faces or eigenvectors.

The enhanced principal component analysis makes use of the mean of set of values

Roy et. al [29] used the combination of image processing methods as well as information discovery for improving the accuracy and early prediction of lung cancer [14].

In this research work, authors make use the lung image acquired from CT scans. The images are pre-processed and SURF algorithm was used to extract features. SVM is the classifier to classify the images as malignant or benign [15].

Faisal et.al build models using various classifiers such as Naïve Bayes, Multilayer perceptron and decision tree, Neural network, SVM an gradient boosting. Gradient boosting algorithm outperforms than other algorithms [16].

[29] Boban et al uses machine learning algorithms including KNN, SVM and MLP. Among different algorithms MLP performs better. The research work makes use of lung disease videos as input and Gray level co-occurrence matrix is used for picking the relevant features.

[30] Sreekumar et al used deep learning to detect whether the image is malignant or benign.

[31] Banerjee et. al suggested a framework to detect cancer. The authors have used various

machine learning algorithms like SVM, random forest and ANN. ANN is the suitable algorithm for accurately predicting area as well as textual oriented features. Maleki etc al has used KNN and Genetic algorithm. GNN is for feature engineering and for reducing the dimensionality of the features. KNN outperforms due to using effective algorithms such as GNN [17].

Reddy et al [32] proposed a hybrid model with the algorithms such as decision tree, KNN, neural network and bagging for improving the accuracy. Gunaydin et al adopted the techniques such as PCA, KNN, SVM, Naïve Bayes, decision trees and ANN for detecting anomalies

4. METHODOLOGY

Data collection

The source of the data for this research is UCI repository. From UCI repository lung dataset is downloaded for this research work. The dataset has 600 instances and 3 target classes such as Malignant, Pre-Malignant, and Benign. The features can be classified into two categories such as independent features and dependent features. Independent features are used to derive the dependent feature.

In this scenario, the dependent feature is the target class. The remaining features are considered as independent features. The set of independent features are air pollution, alcohol use, allergy, dust, occupational, genetic risk, hazards, chronic lung disease, smoking, chest pain, coughing of blood, weight loss, shortness, fatigue, swallowing difficulty, wheezing, frequent cold, clubbing of fingers, snoring and dry cough [18].

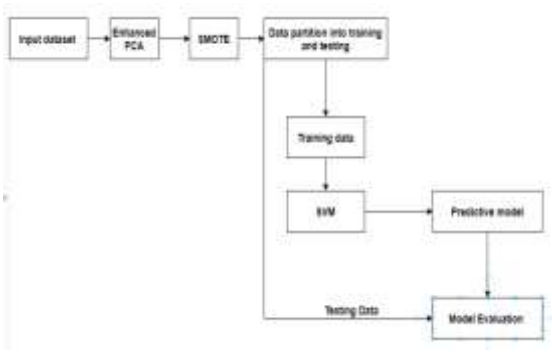


Figure 1. Proposed Framework

Data exploration

The exploratory data analysis is helpful to explore the data. It includes both descriptive and diagnostic analysis. Descriptive analysis uses statistical method to derive insights. Diagnostic analysis derives the causes for the descriptive results. Exploratory analysis enables the researchers to find insights from the data and define hypothesis from the dataset. It gives unknown insights from the dataset. It gives domain knowledge on the research topic. In this work, statistical tools like mean, median, mode, standard deviation are applied to perform uni-variate and multivariate analysis.

Data pre-processing

The exploratory data analysis provides various insights like issues in the dataset such as missing data, outliers, redundant data and more [19]. The identified issues of the dataset can be resolved in data pre-processing stage. It is an important phase for enhancing the capability of the classification system. During this pre-processing stage the activities such as allocating numerical qualities for the target, handling missing values and normalizing the data are performed. Some unimportant features are age, patient-id, as well as gender are eliminated from the dataset. After completing data pre-processing, the dataset contains three target variables, 21 features as well as 599 instances. Once completing the basic pre-

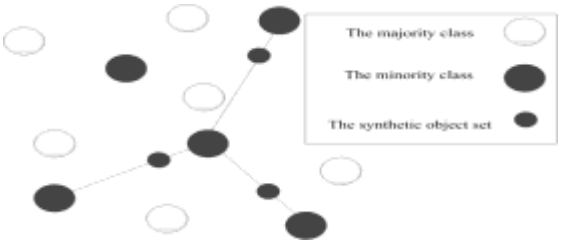
processing, it will be incurred for dimensionality reduction.

Dimensionality reduction (PCA)

Dimensionality reduction with PCA is also called as feature selection. In this phase, only the essential features will be selected from the dataset. PCA analyses the correlation of features. It transforms the high dimensional correlated features into low dimensional un-correlated features. PCA transformation is performed with two target classes such as benign and malignant. PCA focuses on only relevant features of the data as well as eliminates the irrelevant features from the dataset.

Dimensionality reduction (PCA)

Machine learning models also suffer due to the imbalances in the classes. The number of representative samples may be relatively low in one class compare with other class. This kind of situation is called as imbalance [20]. It poses difficulties for the machine learning models to learn on minority class. This research work focuses on improving model performance by resolving issues such as high dimensionality and imbalancing. Smote algorithm creates synthetic samples along the path that joins the two existing samples. The working principle of SMOTE is illustrated below:



The algorithm oversamples the minority class and equalize the number of samples between two classes. It resolves the performance issues imbalancing [21].

Support Vector Machine

SVM is one of the popular algorithms that separate non-linear data. Hyper plane is an important element that separates the data points in such a way that the margins between the classes are wide as well as the points are far as possible [22]. Hyper plane creates a decision boundary with support vectors closer to the left as well as right of hyper plane. The research work makes use of linear SVM to predict lung cancer.

5 EXPERIMENTAL STUDY

The experiments are conducted using python tool. It contains more packages and libraries to perform data analysis and machine learning [23]. The initial dataset is divided into training and testing. The training dataset along with machine learning algorithm is used to build the model. The performance of the model is evaluated using testing dataset. Evaluation measures are used to evaluate the performance of the model [24].

5. 1 Confusion matrix

Confusion matrix is the basic for various evaluation measures such as accuracy, precision, recall, F1 measure and so on. The confusion matrix represents true positive (TP), true negative (TN), false positive (FP) and false negative (FN) [25].

Accuracy

It is the rate of correct prediction out of total number of samples. Accuracy is the basic measure to evaluate the model.

$$Accuracy = \frac{True\ positive + True negative}{True\ positive + True\ Negative + False\ positive + False\ negative}$$

Recall

Recall is also known as sensitivity. It is calculated by dividing true positive rate by sum of true

positives and false negatives. It is the measure that indicates the numbers of true positives were predicted as positives [26]. It is also known as completeness or sensitivity or True positive rate. It is represented as follows

$$Sensitivity \quad \frac{True\ positive}{True\ positive + False\ Negative}.$$

Precision

Precision is the measurement also known as positive predictive rate. It refers the rate of actual positive among the samples predicted as positive. It is the probability for predicted yes will become yes [27]. It is represented by dividing true positive by sum of true positive and false negative. It is a useful measure for medical data analysis.

$$Precision = \frac{True\ positive}{True\ positive + False\ positive}$$

5.1.3 F1-score

F-measure includes precision, recall, and precision's relative value. It is the harmonic mean of recall and precision. It is required while seeking balancing between recall and precision. Moreover, it is also useful for uneven class distribution.

$$F1score = \frac{2*(precision*Recall)}{Precision + Recall}$$

5.1.3 ROC curve

The region under the ROC curve (ROC) or AUC defines the relationship between a classifier's true as well as false positive rate for different judgment thresholds 18]. It is an effective to evaluate the classification model that involves imbalanced dataset.

The accuracy of the model is evaluated using various evaluation measures. The dataset is divided into 70% of training data and 30% of testing data. The performance of the model is evaluated using various evaluation measures.

The experimental results are illustrated in Table 1. for machine learning model with SVM

Class	TP rate	FP rate	Precision	Recall	F1 measurere	ROC
1	0.965	0.108	0.961	0.985	0.965	0.954
2	0.882	0.005	0.929	0.885	0.907	0.985
3	0.775	0.000	0.987	0.857	0.98	0.97
Average	0.874	0.037667	0.959	0.909	0.950667	0.969667

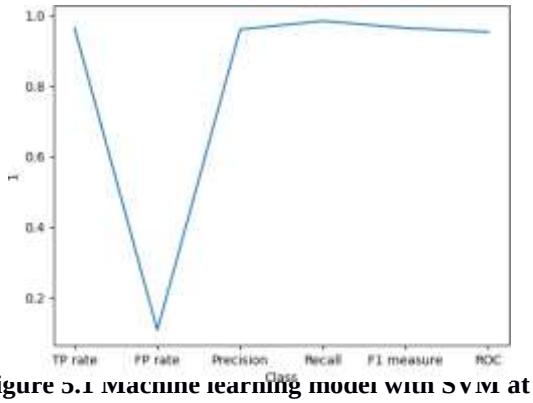


Figure 5.1 Machine learning model with SVM at Class I

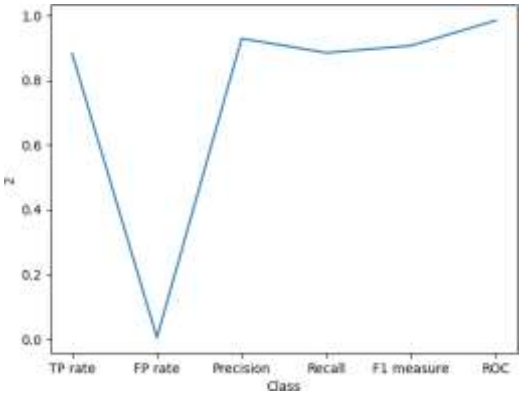


Figure 5.1 Machine learning model with SVM at Class II

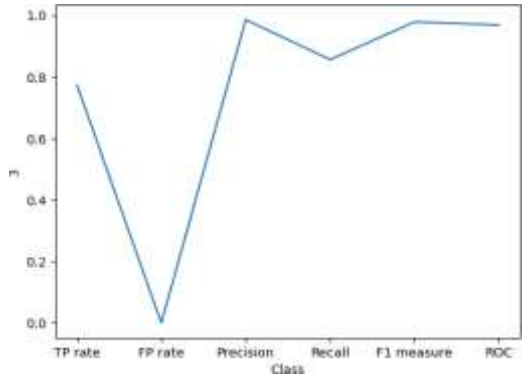


Figure 5.3 Figure 5.1 Machine learning model with SVM at Class III

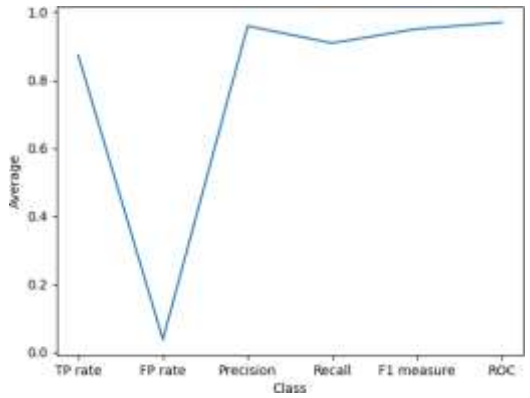


Figure 5.4 Figure 5.1 Machine learning model with SVM

The experimental results are illustrated in Table 2. for machine learning model with enhanced PCA and SVM

Class	TP rate	FP rate	Precision	Recall	F1 measure	ROC
1	0.985	0.108	0.961	0.985	0.965	0.954
2	0.882	0.005	0.929	0.885	0.907	0.985
3	0.775	0.000	0.987	0.857	0.98	0.97
Average	0.880667	0.037667	0.959	0.909	0.950667	0.969667

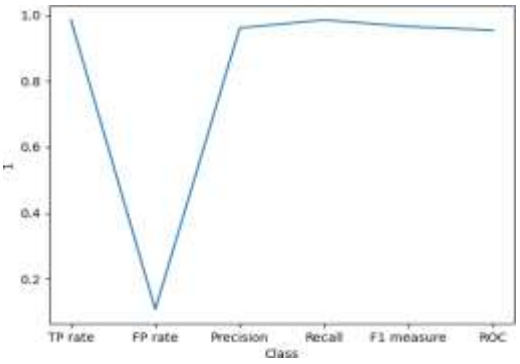


Figure 5.5 Enhanced PCA with SVM at class I

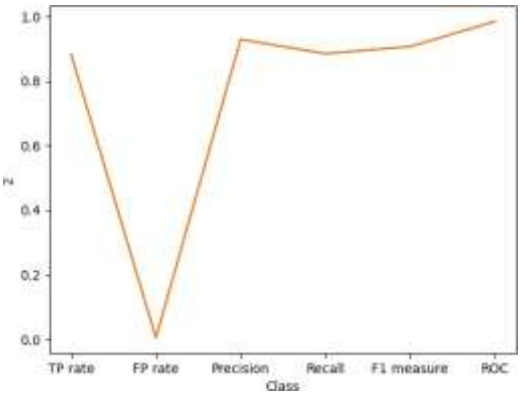


Figure 5.6 Figure 5.5 Enhanced PCA with SVM at class II

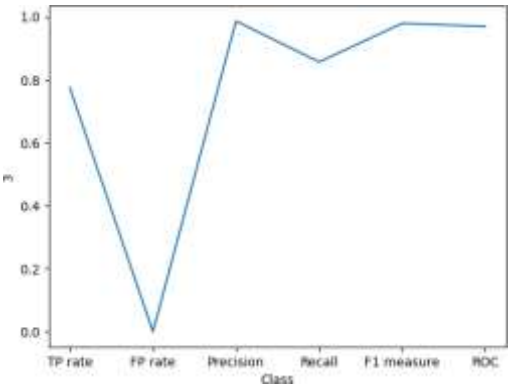


Figure 5.7 Figure 5.5 Enhanced PCA with SVM at class III

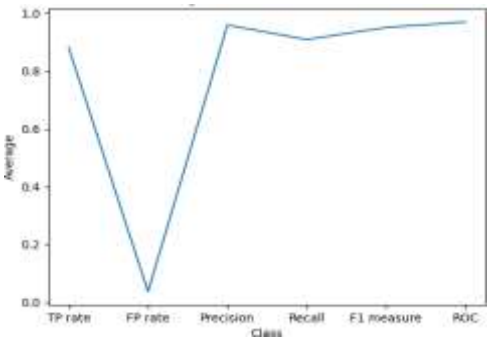


Figure 5.8 Figure 5.5 Enhanced PCA with SVM

5.1 Interpretation of results

The experimental results show that, on predicting the classes such as class 1, class 2 and class 3 using SVM and enhanced PCA with SVM, enhanced PCA with SVM perform better than SVM algorithm. The experimental results demonstrate that, the proposed methodology resolves the issues that exist in the dataset. Datasets have various issues such as missing data, redundant data, and outliers. The initial pre-processing stage, pre-processes the dataset and passed to the next phase such as principal component analysis. PCA transforms the high dimensional and into lower dimensional. SMOTE algorithm overcomes the issues of im-balancing. The proposed methodology resolves the issues of dataset thereby improve the predictive accuracy.

5.1 Summary and conclusion

Lung cancer or carcinoma is one of the leading causes for death. The existing methodologies to diagnose the disease are not sufficient to accurately diagnose the diseases. The existing techniques are not sufficient to predict the diseases at earlier. Machine learning algorithms play an important role on predicting the disease at earlier to take proactive decisions. The research work considers the lung dataset from UCI. The work makes use of PCA and SMOTE to address the issues of high dimensionality and im-balancing. The proposed methodology on PCA with SMOTE resolves the issues of the dataset that results in improvement in classification model performance. The proposed methodology outperforms than traditional SVM on various measures such as TP rate, FP rate, recall, precision, recall, F-score and ROC. The average TP rate is higher on PCA with SMOTE and SVM than SVM alone. It shows that, the proposed model classifies samples accurately and the FP rate is lower in the proposed frameworkrespectively.

6.REFERENCES

[1] P. Chaudhari, H. Agarwal, and V. Bhateja, "Data augmentation for cancer classification in oncogenomics: an improved KNN based approach," Evol. Intell., pp. 1–10, 2019.

[2] S. F. Khorshid and A. M. Abdulazeez, "BREAST CANCER DIAGNOSIS BASED ON K-NEAREST NEIGHBORS: A REVIEW," PalArch's J. Archaeol. Egypt/Egyptology, vol. 18, no. 4, pp. 1927–1951, 2021.

[3] F. Q. Kareem and A. M. Abdulazeez, "Ultrasound Medical Images Classification Based on Deep Learning Algorithms: A Review."

[4] D. Q. Zeebaree, A. M. Abdulazeez, D. A. Zebari, H. Haron, and H. N. A. Hamed, "Multi-Level Fusion in Ultrasound for Cancer Detection Based on Uniform LBP Features."

[5] J. R. F. Junior, M. Koenigkam-Santos, F. E. G. Cipriano, A. T. Fabro, and P. M. de Azevedo-Marques, "Radiomics-based features for pattern recognition of lung cancer histopathology and metastases," Comput. Methods Programs Biomed., vol. 159, pp. 23–30, 2018.

[6] I. Ibrahim and A. Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 10–19, 2021.

[7] P. Das, B. Das, and H. S. Dutta, "Prediction of Lungs Cancer Using Machine Learning," Easy Chair, 2020.

[8] G. A. P. Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," Neural Comput. Appl., vol. 31, no. 10, pp. 6863–6877, 2019.

[9] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 20–28, 2021.

[10] H. A. Hussein and A. M. Abdulazeez, "COVID-

19 Pandemic Datasets Based on Machine Learning Clustering Algorithms: A Review," PalArch's J. Archaeol. Egypt/Egyptology, vol. 18, no. 4, pp. 2672–2700, 2021.

[11] D. M. Abdullah and N. S. Ahmed, "A Review of most Recent Lung Cancer Detection Techniques using Machine Learning," Int. J. Sci. Bus., vol. 5, no. 3, pp. 159–173, 2021.

[12] M. I. Faisal, S. Bashir, Z. S. Khan, and F. H. Khan, "An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer," in 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology, 2018, pp. 1–4.

[13] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, "Gene selection and classification of microarray data using convolutional neural network," in 2018 International Conference on Advanced Science and Engineering (ICOASE), 2018, pp. 145– 150.

[14] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Trainable model based on new uniform LBP feature to identify the risk of the breast cancer," in 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019, pp. 106–111.

[15] H. Tang, J. Zhao, and X. Yang, "Explore machine learning for analysis and prediction of lung cancer related risk factors," in Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, 2018, pp. 41–45.

[16] P. R. Radhika, R. A. S. Nair, and G. Veena, "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," in 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1–4.

[17] A. I. Rahmani and M. Katouli, "Diagnosing Lung Cancer Using Grasshopper Optimization Algorithm and k-Nearest Neighbor Classification," J. homepage http//iieta. org/journals/rces, vol. 6, no. 4, pp. 69–75, 2019.

[18] Y. Nai et al., "Improving Lung Lesion Detection in Low Dose Positron Emission Tomography Images Using Machine Learning," in 2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC), 2018, pp. 1–3.

[19] S. Senthil and B. Ayshwarya, "Lung cancer prediction using feed forward back propagation neural networks with optimal features," Int. J. Appl. Eng. Res., vol. 13, no. 1, pp. 318–325, 2018.

[20]M. R. Mahmood, A. M. Abdulazeez, and Z. ORMAN, "A NEW HAND GESTURE RECOGNITION SYSTEM USING ARTIFICIAL NEURAL NETWORK."

[21] M. Somvanshi, P. Chavan, S. Tambade, and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," Proc. - 2nd Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2016, 2017, doi: 10.1109/ICCUBEA.2016.7860040.

[22] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine Learning Supervised Algorithms of Gene Selection: A Review," Mach. Learn., vol. 62, no. 03, 2020.

[23] O. Ahmed and A. Brifcani, "Gene Expression Classification Based on Deep Learning," in 2019 4th Scientific International Conference Najaf (SICN), 2019, pp. 145–149

[24] N. O. M. Salim and A. M. Abdulazeez, "Human Diseases Detection Based On Machine Learning Algorithms: A Review," Int. J. Sci. Bus., vol. 5, no. 2, pp. 102–113, 2021.

[25] N. M. Abdulkareem and A. M. Abdulazeez, "Machine Learning Classification Based on Radom Forest Algorithm: A Review," Int. J. Sci. Bus., vol. 5, no. 2, pp. 128–142, 2021.

[26] R. Sathishkumar, K. Kalaiarasan, A. Prabhakaran, and M. Aravind, "Detection of Lung

Cancer using SVM Classifier and KNN Algorithm," in 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019, pp. 1–7.

[27] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Med. Inform. Decis. Mak., vol. 19, no. 1, pp. 1–16, 2019.

[28] K. Roy et al., "A Comparative study of Lung Cancer detection using supervised neural network," in 2019 International Conference on Opto-Electronics and Applied Optics (Optronix), 2019, pp. 1–5.

[29] B. M. Boban and R. K. Megalingam, "Lung Diseases Classification based on Machine Learning Algorithms and Performance Evaluation," in 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 315– 320

[30] A. Sreekumar, K. R. Nair, S. Sudheer, H. G. Nayar, and J. J. Nair, "Malignant Lung Nodule Detection using Deep Learning," in 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 209–212.

[31] Banerjee and S. Das, "Prediction Lung Cancer–In Machine Learning Perspective," in 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1–5.

[32] D. Reddy, E. N. H. Kumar, D. Reddy, and P. Monika, "Integrated Machine Learning Model for Prediction of Lung Cancer Stages from Textual data using Ensemble Method," in 2019 1st International Conference on Advances in Information Technology (ICAIT), 2019, pp. 353–357.