Efficient multi-level lung cancer prediction model using support vector machine classifier

Manju B R1*, Athira V2, Athul Rajendran3
1Department of Mathematics, Amrita Vishwa Vidyapeetham, Amritapuri, India
2Centre for Wireless Networking and Applications, Amrita Vishwa Vidyapeetham, Amritapuri, India
3Department of Electrical and Electronics Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
E-mail: manjubr@am.amrita.edu

Abstract- This paper aims at the requirement for an interactive learning framework which empowers the successful checking of disorder in a patient. Principal component analysis stands out as an outstanding algorithm to significantly classify the target classes. PCA blends associated characteristics and makes a dissipated showcase of its components well. Scree plot examination gives solidarity of how many principal components are to be retained. Support Vector Machines (SVM) is a fast and dependable classification algorithm that outperforms other techniques with a limited amount of data. The obtained components will be served to Support Vector Machine for further classification. The pre-dangerous stage will remind the clinical experts to give additional consideration to those patients. The expectation ability is estimated in terms of the confusion matrix. The model developed gives a high and uncompromising accuracy in early detection of different levels of malignancy

Keywords: Premalignant, PCA, SVM, confusion matrix

1. Introduction

In this time, the primary reason of death around the globe is malignant growth or carcinoma. Among the different types of existing cancers like breast, liver, stomach and colorectal, majority of deaths are caused by lung carcinoma (1.6 million) [1]. According to the overall information in 2012, lung cancer contributes 13% all cancers considered. Lung Carcinoma seems to be a propagation of morphological changes [8]. The two varieties of lung cancer are Small Cell and Non-Small Cell Cancers (SCLC and NSCLC) [2]. The reasons for lung disease incorporate hereditary and non-hereditary reasons [3]. Tobacco smoking is the far leading cause of lung cancer. It can severely affect the health of chain smokers as well as people in contact with smokers [4]. The lungs of the non-smokers can be affected by this malignancy due to air pollution, working atmosphere with radon, asbestos or any other chemical exposure [4]. Hereditary causes include the gene mutations in lungs. Individuals who have had taken radiation therapy for other diseases can also be the victims of this malignancy [4].

Mostly this condition is recognized at the advanced stage. This can be limit the cure of therapeutic intervention [5]. The different stages of lung cancer [27] are tabulated below with explanation in Table.1:
Table 1. Lung Cancer stages explained

| Stages | Explanation |
|--------|-------------|
| Stage 1 | The cancer is confined only in the lungs. |
| Stage 2 | The cancer is in the lungs and can have metastasized to the lymph nodes adjacent to the lungs. |
| Stage 3 | Cancer is localized in the lungs and has spread to the lymph nodes in the center of the chest. |
| Stage 3A | The cancer has become extensive, however, is localized to one side of the lung. |
| Stage 3B | The cancer has metastasized to the lymph nodes on both the sides of the lungs. |
| Stage 4 | The cancer has metastasized to both the lungs or has spread to other parts of the body. Stage 4 is the most advanced stage of lung cancer. |

The primary care of medical practitioners is needed for the fast survival. The identification of lesion type will give the signals for the further treatment. The malignant stage is red signal, benign stage counts for green stage and an intermediate stage can be introduced as pre-malignant stage which is indicated as yellow signal. If a yellow signal is appeared, the doctors should continuously monitor those patients and it will be helpful for the fast cure. The popular examples declared by World Health Organization (WHO) are typical adenomatous hyperplasia; diffuse idiopathic pulmonary neuroendocrine cell hyperplasia, squamous metaplasia with dysplasia and carcinoma in-situ [6]. These conditions can be advance to harmful stages like carcinoid tumors, squamous cell carcinoma and adenocarcinoma [7]. Manual evaluation of lung features for the qualitative result is mentioned in [10]. Ground Glass Opacity [GGO] is considered to be pre-malignant, but this shadow is not differentiable by radiologist [9]. People reached a stage to be aware of the pre-malignant stage.

Detection using screening procedures includes Computed Tomography (CT) and other bronchoscopic techniques [6]. CT visualizes the spongy nodules of the windpipe within single breath threshold [12]. These techniques can capture 2-3 mm of lesions and it can create confusions on determining whether it is malignant, benign or the intermediate stage. Until this point in time, screening preliminaries have had no critical effect on survival [7]. Machine learning and feature selection has a golden chair in Lung cancer Detection, nowadays. A numerical measurement fed to a most accurate machine learning system is more preferable than the distorted image obtained by radiologists. In [9], GGO is extracted for the study using Deep Learning, which is as intelligent as human brain. Automated system to view histopathological part of respiratory tract is explained in [10]. The trending state of art to do lung cancer diagnosis with the aid of computer knowledge is explained in [11], touching the technical concerns and validation. CAD procedure has a golden standard in terms of speed and accuracy, when compared to other methods [12].

In the world, main part of datasets are available via web based networking and from this tremendous piece of information, features useful for the tasks like classification and prediction can be fetched [13]. These types of feature vectors are classified into target classes using Random Forest Classifier in [14]. Along with the existing blocks of classifiers, weighted models of SVM, ANN and NB created sequential diagnostic model for lung cancer patients in [15].

Some of the reported studies regarding machine learning applications on lung cancer is tabulated in TABLE.2. This paper introduces an automated platform for the lung cancer victims belongs to premalignant stage. The application is overlying on a PCA – SVM combined model and all the statistical parameters are well drawn for the further analysis.
Table 2. Reported studies on Lung Cancer Detection

| SL.No. | Name                                                                 | Year | Overview                                                                                                                                 |
|--------|----------------------------------------------------------------------|------|-----------------------------------------------------------------------------------------------------------------------------------------|
| 1.     | Comparison of Lung Cancer Detection Algorithms [19].                 | 2019 | Applied Principal Component Analysis, K-Nearest Neighbors, Support Vector Machines, Naïve Bayes, Decision Trees and Artificial Neural Networks machine learning methods before and after preprocessing to detect anomaly. |
| 2.     | A Deep Learning Based Approach to Lung Cancer Identification [16].   | 2018 | A 3D neural network based diagnosis with tensorflow library which effectively works even with less number of samples.                     |
| 3.     | Lung Cancer Detection Using Image Processing and Machine Learning HCare [17]. | 2018 | SVM classifier and image processing is collaborated on CT scan lung images to enable early detection.                                    |
| 4.     | Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm [18]. | 2017 | A novel neural-network based algorithm with entropy degradation method (EDM), to detect small cell lung cancer (SCLC) from CT images.     |
| 5.     | A Survey on Lung Cancer Detection using Image Data Analysis and Machine Learning [25]. | 2017 | This study aims to highlight the significance of data analytics and machine learning.                                                 |
| 6.     | Lung cancer classification using deep learned features on low population dataset [20]. | 2017 | A deep Autoencoder classification mechanism which works on low population unstructured data.                                           |
| 7.     | Deep learning application trial to lung cancer diagnosis for medical sensor systems [22]. | 2016 | Personal diagnosis is enabled using sensing data and deep learning techniques for the feature selection.                              |
| 8.     | Convolutional neural networks for lung cancer screening in computed tomography (CT) scans [28]. | 2016 | Special properties of CNN networks like spatial invariance which extract features easily is applied for lung cancer detection.       |
| 9.     | Computer aided diagnosis system based on machine learning techniques for lung cancer [21]. | 2012 | Feature extraction using wavelet transform is applied on lung cancer dataset to achieve high accuracy.                              |
| 10.    | Predictive Analysis of Lung Cancer Recurrence [24].                  | 2011 | The predictive analysis of lung cancer recurrence based on non-small cell lung cancer cancerous gene expression data using data mining and machine learning techniques. |

2. Methodology

The strategy followed in this work is drawn in Fig.1

Figure 1. Block Diagram

2.1 Data Collection

Database used in this research work is obtained from Lung Cancer Dataset UCI repository [26]. This multivariate data consists of 600 instances with three target classes like Malignant (M, count-215), Pre-malignant (P, count-194) and Benign (B, count-190) indicates three conditions of lung cancer tumors. The features listed in this csv dataset are Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Obesity, Smoking, Chest Pain, Coughing of blood, Fatigue, Weight loss, Shortness, Wheezing, Swallowing difficulty, Clubbing of fingers, Frequent cold, Dry cough and Snoring.
2.2 Data Visualization

Visualizing the data is the method of converting the data into abstract images which follows certain patterns or trends. It encourages the analysts to think of huge choices in application level. Python offers various incredible diagramming libraries like pandas, Matplotlib etc., which creates various highlighted histogram, scatter plot and density curve [28]. The features that are skewed in TABLE.2 are stretched as combination of histograms and density curves in Fig. 2. The followed pattern of each feature corresponding the target class through histograms. The normal or exponential distribution of features helps in parametric analysis in terms of maximum value, standard deviation, minimum value, mean etc.

2.3 Data Skewness.

The skewness of the data represents the asymmetry in statistics. The curve will seems to be skewed to either left or right [29]. The result can be quantified to analyse the difference between the obtained distribution and normal distribution.

- If mean > median, then “Positively skewed”.
- If mean < median, then “Negatively skewed”.

| Features                       | Skew Values |
|--------------------------------|-------------|
| Air Pollution                  | 0.195541    |
| Alcohol use                    | 0.023802    |
| Dust Allergy                   | -0.563203   |
| Occupational Hazards           | -0.170883   |
| Genetic Risk                   | -0.085225   |
| Chronic Lung Disease           | -0.109587   |
| Balanced Diet                  | -0.053694   |
| Obesity                        | 0.061515    |
| Smoking                        | 0.407958    |
| Passive Smoker                 | 0.372199    |
| Chest Pain                     | 0.189506    |
| Coughing of Blood              | 0.162642    |
| Fatigue                        | 0.857201    |
| Weight Loss                    | 0.370156    |
| Shortness of Breath            | 0.380384    |
| Wheezing                       | 0.222926    |
| Swallowing Difficulty          | 0.480487    |
| Clubbing of Finger Nails       | 0.787004    |
| Frequent Cold                  | 0.456050    |
| Dry Cough                      | 0.231364    |
| Snoring                        | 0.546970    |
Figure 2. Visualizing the feature
2.4 Data Preprocessing
Preprocessing is an important pre-requisite for any data examination. It is generally an excellent plan to set up the information in such a manner to uncover the structure of the data to the machine learning calculations that needs to use. Data preprocessing techniques are well known in enhancing the capability power of classification systems [31]. This includes various exercises [30] like:
- Allotting numerical qualities to target.
- Dealing with missing numerical.
- Normalizing the highlights
Some unimportant highlights like patient-id, Age and Gender are cleared out to expand the effectiveness of the model we consider. As of now the information contains 3 targets, 21 characteristics and 599 cases. The cleaned csv dataset completes the preprocessing, after reducing the dimensionality.

Dimensionality Reduction – Principal Component Analysis (PCA). PCA handles the feature selection strategies through the view of reduced dimensionality. The unfocussed features can bring tremendous decline in the performance rate of prediction models. Diminishing the dimensionality of a data [38] by picking the significant features present in the underlying dataset are named as feature selection. These informational indexes are simpler to investigate and break down information in simpler way for calculations without superfluous factors to process. PCA is figured as a powerful data representation tool in [37]. PCA works on the mathematical basis of linear algebra which analyses the correlation of features [39]. In Fig. 3, PCA feature space transformation is done with two target classes namely Malignant (M) and Benign (B) and samples are spread along the axes. This type of transformation is also applicable for the Pre-malignant class, which is displayed along with the conventional classes in Figure 5. Some of the relevant markings of PCA in medical domain are:

For distinguishing normal and abnormal prostate cells, PCA is used along with signal processing in [32]. The irrelevant points are cleared out from the original feature space using PCA in breast cancer detection [33]. PCA pre-forms the data and concentrates on relevant features for training the model [34]. The PCA based algorithm achieved improved accuracy with the absence of overfitting and outliers in [35]. In [36], a PCA-FNN system is proposed to tune parameters for categorizing liver cancer.

![Figure 3. PCA feature space](image)

**Scree Plot:** The numbers of PCA components displayed in the Fig. 3. are PCA_1 and PCA_2, which indicates the two components, used to transform the feature space. The significance of choosing principal component number in PCA can be vividly seen in the scree plot in Fig. 4. A scree plot is an analytic device to check whether PCA functions are good indicators or not [40]. The components are arranged by the measure of variety they spread.

PCA_1 catches the most variety, PCA_2 the second most, etc. Fig.4 shows an elbow curve at the point 2, which points to the generation of PCA_1 and PCA_2 for the fed input.
Figure 4. Scree Plot

Figure 5. Three target classes with PCA
2.5 Selection of features through Feature correlation

Machine learning algorithms are always acknowledged with framework of features that are quantifiable. Data Scientists cleanse data, by eliminating the unwanted features or features having less contribution to the target class. Feature correlation is the inevitable step which indicates the contribution of each feature to the target class. Correlation can be positive or negative. Redundancy and relevancy of features are pestered in [42]. Positive correlation values indicate the direct proportionality of features with the class, while negative correlation tells about the inverse proportionality. Correlation value 1 means “perfectly correlated”. If two features $x$ and $y$ are directly proportional to malignant class, then one of the feature can be removed by considering their correlation value. To visualize the correlation values together on a plane, Heatmap is generally used by the analysts. The Heatmap for the lung cancer prediction model is displayed in Fig. 6.

![Figure 6. Heatmap representation](image)

2.6 Classification

Classification is the logical grouping of data. It has the capability to make decisions on real and unstructured data. It plays a major role in health care and data security. CVDs and diabetes are classified evidently using ANN in [43]. Neural network synced with fuzzy system to detect asthma in [44]. Diagnosis of epilepsy using signaling techniques and machine learning can be seen in [45]. The possibilities of Alzheimer’s disease are explored using classification task in [46]. In the field of data security, variable indexes are used to classify in [47] and mobility of big data is enhanced to improve data security in [48]. In [49], data mining is explored to eradicate malware and DoS attacks. Texts are pruned and classified to ensure security in an automated way [50]. Conventionally, the cancer dataset was classified as Malignant (M) and Benign (B), but this paper throws light into an intermediate condition called “Pre-malignant”, which provides an extra care for the patients. In data security, there will be a continuous monitoring of categories like copied, transmitted and retrieved data. Classification involves labeling information to make it effectively accessible and identifiable. It can eliminate duplications which reduces storage and backup costs. It can incredibly decrease the processing time.

**Support Vector Machine (SVM):** SVM is one of the most popular classification algorithms which have an elegant way of transforming nonlinear data. Classification strategy of SVM is well explained in [52]. Hyper plane is the important tool of SVM that separates the data points in such a manner that the margin between two classes will be wide and the data points will be as far as possible. In this way, hyper plane will be creating a decision boundary with support vector points nearer to the left and right hyperplane. Linear SVM model is used for this lung cancer prediction study. A sample of SVM classification is shown in Fig. 7. Class 1 belongs to Malignant (M) and class -1 belongs to Benign (B).
Advantages of SVM [51] are:
• Clear separation of data points into classes.
• Expertise in high dimensional analysis.
• Support vectors needs minimum memory storage.
• Parameters are tunable and user friendly.

SVM classifiers are prominent in health domain. 2-level DWT images using a GLCM (Gray level Co-occurrence Matrix) effected by means of an SVM (Support Vector Machine) in [53]. Chip based system with SVM flavor is designed in [54]. The various strategies of SVM to detect breast cancer in an accurate way is explained in [55] [56] [57].

3. Results – Performance measurement

Results are obtained through confusion matrix and classification report. Confusion matrix is generated from the binary classification outcomes [58]. The highlighted parameters like accuracy, error rate, truth positive rate (TPR), false positive rate (FPR), truth negative rate (TNR) and false negative rate (FNR) can be calculated on the basis of this matrix. Accurate model with high score is the uncompromising factor for detection systems. GPS based map matching systems [59], and processors with frequency clocks [60] question the score of accuracy and ask for improvement. So, data scientists will try to minimize the error rate to a great extent. Accuracy score and error rate of this model are equated below as eqn [1] and eqn [2]. Commonly, the cancer prediction is deployed by 2 x 2 matrices. Here, it is innovated as 3 x 3 matrix and drawn in Fig. 8 along with the classification report. Classification Report will gives the precision, recall, f1-score and support of this classification system. Parametric definitions are in Table 4.

Table 4. Parametric Definitions

| Sl.No. | Parameters          | Definitions                                                                 |
|--------|---------------------|-----------------------------------------------------------------------------|
| 1      | True Positive (TP)  | Correct positive predictions                                                 |
| 2      | False Positive (FP) | Incorrect positive predictions                                               |
| 3      | True Negative (TN)  | Correct negative predictions                                                |
| 4      | False Negative (FN) | Incorrect negative predictions                                               |
| 5      | Error Rate/Misclassification rate | Total number of incorrect predictions                                       |
| 6      | Accuracy            | Total number of correct predictions                                          |
| 7      | Sensitivity         | The number of correct positive predictions (TP) divided by the total number of positives (P). |
| 8      | Specificity         | The number of correct negative predictions divided by the total number of negatives. |
| 9      | Precision           | The number of correct positive predictions divided by the total number of positive predictions. |
| 10     | F1-score            | Harmonic mean of precision and recall.                                      |
Accuracy = (TPR + TN)/Total = 0.87

Error Rate = (FP + FN)/Total = 0.3

Confusion matrix:

\[
\begin{bmatrix}
160 & 4 & 6 \\
6 & 178 & 14 \\
6 & 36 & 126
\end{bmatrix}
\]

Confusion matrix plot of SVC

| Classification report | precision | recall | f1-score | support |
|-----------------------|-----------|--------|----------|---------|
| B                     | 0.96      | 0.94   | 0.95     | 170     |
| M                     | 0.82      | 0.83   | 0.87     | 192     |
| P                     | 0.87      | 0.76   | 0.81     | 178     |
| avg / total           | 0.88      | 0.88   | 0.88     | 540     |

Figure 8. Performance of SVM classification model

4. Conclusion and Future Work

This investigation draws thoughtfulness to the significance of pre-malignant stage in early detection of lung carcinoma. The real and inconsistent data is processed and cleansed. The precision of the SVM classifier could clearly signature with high accuracy of the prediction with high distinction. The confusion matrix obtained could effectively label and identify the information hidden in the dataset. Here the performance is coined in terms of classification report toasted on confusion matrix and its calculations. Accuracy score and error rate of this model is presented as 3 x 3 matrix in contrast to the conventional 2x2 matrix. Classification Report gives the precision, recall, f1-score and support of this classification system. The work gives a scope for extension to optimization level that can ensure leading accuracy in machine learning diagnostics.
References

[1] Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M. 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer, 136(5), 359–386.

[2] Gadgeel, S.M.; Ramalingam, S.S.; Kalekerian, G.P. 2012. Treatment of lung Cancer, adiol Clin North Am, 961–74.

[3] Sruthi Ignatious, Robin Joseph . 2015. Computer Aided Lung Cancer Detection System, Proceedings of 2015 Global Conference on Communication Technologies (GCCT 2015), 7342723.

[4] American Cancer Society, URL: https://www.cancer.org/cancer/non-small-cell-lung-cancer/causes-risks-prevention/what-causes.html.

[5] W. Wang and S. Wu. 2016. A Study on Lung Cancer Detection by Image Processing, International Conference on Communications, Circuits and Systems, Guilin, 371-374.

[6] A. Sarkar, A. Sadhu, S. B. Thakur, S. Sengupta, A. Mukherjee and J. Chatterjee. 2016. Multimodal characterization of radiologically detectable lung lesions, 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, 1422-1425.

[7] Greenberg, A. K., Yee, H., & Rom, W. N. 2002. Preneoplastic lesions of the lung. Respiratory Research, 3(1).

[8] K. Hirayama et al. 2017. Extraction of GGO regions from chest CT images using deep learning , 17th International Conference on Control, Automation and Systems (ICCAS), Jeju, 2017, 351-355.

[9] Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., & Snyder, M. 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature Communications, 7, 12474.

[10] Bhavanishankar K1 and Dr. M.V.Sudhamani2. 2015. TECHNIQUES FOR DETECTION OF SOLITARY PULMONARY NODULES IN HUMAN LUNG AND THEIR CLASSIFICATIONS -A SURVEY, International Journal on Cybernetics & Informatics (IJCI), 4.

[11] Hosseinzaeh, F., KayvanJoo, A., Ebrahimi, M., & Goliaei, B. 2013. Prediction of lung tumor types based on protein attributes by machine learning algorithms. SpringerPlus, 2(1), 238.

[12] CENGIL, E., & CINAR, A. 2018. A Deep Learning Based Approach to Lung Cancer Identification. International Conference on Artificial Intelligence and Data Processing (IDAP).

[13] Q. Wu and W. Zhao. 2017. Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm, International Symposium on Computer Science and Intelligent Controls(ISCSIC),Budapest, 88-91.
[21] H. R. H. Al-Absi, B. B. Samir, K. B. Shaban and S. Sulaiman. 2012. Computer aided diagnosis system based on machine learning techniques for lung cancer, International Conference on Computer & Information Science (ICCIS), Kuala Lumpur, 295-300.

[22] R. Shimizu et al. 2016. Deep learning application trial to lung cancer diagnosis for medical sensor systems, International SoC Design Conference (ISOCC), Jeju, 191-192.

[23] P. Rao, N. A. Pereira and R. Srinivasan. 2016. Convolutional neural networks for lung cancer screening in computed tomography (CT) scans, 2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida, 489-493.

[24] Shweta Srivastava, Manisha Rathi, and J.P. Gupta. 2011. ACC 2011, Part I, CCIS 190, 260–269.

[25] Apoorva Mahale, Chetan Rawool et.al. A Survey on Lung Cancer Detection using Image Data Analysis and Machine Learning, International Journal of Innovative Research in Computer and Communication Engineering.

[26] UCI Lung Cancer Dataset, URL: https://archive.ics.uci.edu/ml/datasets/Lung+Cancer

[27] Stages and Survival Rate of Lung Cancer, URL: https://www.epainassist.com/chest-pain/lungs/stages-and-survival-rate-of-lung-cancer.

[28] Introduction to Data Visualization in Python, URL: https://towardsdatascience.com/introduction-to-data-visualization-in-python-89a54c97bed

[29] Machine Learning Basics: Data Skewness and handling, URL: https://medium.com/mlrecipies/machine-learning-basics-data-skewness-and-handling-778ec7ad18c.

[30] What Steps should one take while doing Data Preprocessing? URL: https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1ca.

[31] D. H. Deshmukh, T. Ghorpade and P. Padiya. 2015. Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset, International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, pp. 1-6.

[32] Ghosh, A., & Barman, S. 2013. Prediction of Prostate Cancer Cells based on Principal Component Analysis Technique. Procedia Technology, 10, 37–44.

[33] S. Ghosh, S. Mondal and B. Ghosh., 2014. A comparative study of breast cancer detection based on SVM and MLP BPN classifier," 2014 First International Conference on Automation, Control, Energy and Systems (ACES), Hooghy, 1-4.

[34] S. Jhajharia, H. K. Varshney, S. Verma and R. Kumar. 2016. A neural network based breast cancer prognosis model with PCA processed features, International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 1896-1901.

[35] Akshya Yadav, Imlikumla Jamir, Raj Rajeshwarai Jain, Mayank Sohani. Breast Cancer Prediction using SVM with PCA Feature Selection Method, International Journal of Scientific Research in Computer Science, Engineering and Information Technology.

[36] Ms. G.Ilakkiya and Mrs.B.Jayanthi. 2013. Liver Cancer Classification Using Principal Component Analysis and Fuzzy Neural Network, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 10.

[37] F. Song, Z. Guo and D. Mei. 2010. Feature Selection Using Principal Component Analysis, International Conference on System Science, Engineering Design and Manufacturing Informatization, Yichang, 27-30.

[38] M. Partridge and M. Jabri. 2000. Robust principal component analysis, Neural Networks for Signal Processing X. Proceedings of the IEEE Signal Processing Society Workshop (Cat. No.00TH8501), Sydney, NSW, Australia, 289-298.

[39] S. Sehgal, H. Singh, M. Agarwal, V. Bhasker and Shantanu. 2014. Data analysis using principal component analysis, International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), Greater Noida, 45-48.

[40] How to read PCA biplots and scree plots, URL: https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/.

[41] Why Feature Correlation Matters … A Lot! URL: https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4.
[42] G. Qu, S. Hariri and M. Yousif. 2005. A new dependency and correlation analysis for features, IEEE Transactions on Knowledge and Data Engineering, 17, 1199-1207.

[43] B. Alić, L. Gurbeta and A. Badnjević. 2017. Machine learning techniques for classification of diabetes and cardiovascular diseases, 6th Mediterranean Conference on Embedded Computing (MECO), Bar, 1-4.

[44] A Badnjevic, M Cifrek, D Koruga, D. Osmentovic. 2005. Neuro-fuzzy classification of asthma and chronic obstructive pulmonary disease, BMC Medical Informatics and Decision Making Journal.

[45] S. Avdakovic, I. Omerhodzic, A. Badnjevic, D. Boskovic. 2014. Diagnosis of Epilepsy from EEG Signals using Global Wavelet Power Spectrum, MBEC, Dubrovnik, Croatia.

[46] A Aljovic, A Badnjevic, L Gurbeta. 2016. Artificial Neural Networks in the Discrimination of Alzheimer's disease Using Biomarkers Data, IEEE MECO, Bar, Montenegro.

[47] F. Fatemi Moghaddam, M. Yezdanpanah, T. Khodadadi, M. Ahmadi and M. Esrami. 2014. VDCI: Variable data classification index to ensure data protection in cloud computing environments, IEEE Conference on Systems, Process and Control (ICSPC 2014), Kuala Lumpur, 53-57.

[48] I. Hababeh, A. Gharaibeh, S. Nofal and I. Khalil. 2019. An Integrated Methodology for Big Data Classification and Security for Improving Cloud Systems Data Mobility IEEE Access, 7, 9153-9163.

[49] M. Monshizadeh and Z. Yan. 2014. Security Related Data Mining, IEEE International Conference on Computer and Information Technology, Xi’an, 775-782.

[50] K. Alzhrani, E. M. Rudd, C. E. Chow and T. E. Boult. 2016. Automated big security text pruning and classification," IEEE International Conference on Big Data (Big Data), Washington, DC, 3629-3637.

[51] What do support vector machines trump other classification methods, URL: http://www.simafore.com/blog/bid/112816/When-do-support-vector-machines-trump-other-classification-methods.

[52] Chen Junli and Jiao Licheng. 2000. Classification mechanism of support vector machines, WCC 2000 - ICSP 2000. 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000, Beijing, China, 3, 1556-1559.

[53] D. P. Kaucha, P. W. C. Prasad, A. Alsadoon, A. Elchouemi and S. Sreedharan. 2017. Early detection of lung cancer using SVM classifier in biomedical image processing, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 3143-3148.

[54] S. Afifi, H. GholamHosseini and R. Sinha. 2017. SVM classifier on chip for melanoma detection, 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Seogwipo, 270-274.

[55] Z. Luo, X. Wu, S. Guo and B. Ye. 2008. Diagnosis of Breast Cancer Tumor Based on PCA and Fuzzy Support Vector Machine Classifier, Fourth International Conference on Natural Computation, Jinan, 363-367.

[56] M. Sewak, P. Vaidya, C. Chan and Zhong-Hui Duan. 2007. SVM Approach to Breast Cancer Classification, Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007), Iowa City, IA, 32-37.

[57] M. Hussain, S. K. Wajid, A. Elzaart and M. Berbar. 2011. A Comparison of SVM Kernel Functions for Breast Cancer Detection, Eighth International Conference Computer Graphics, Imaging and Visualization, Singapore, 145-150.

[58] Basic evaluation measures from the confusion matrix, URL: https://classeval.wordpress.com/introduction/basic-evaluation-measures/.

[59] M. Hashemi and H. A. Karimi. 2016. A Machine Learning Approach to Improve the Accuracy of GPS-Based Map-Matching Algorithms (Invited Paper),IEEE 17th International Conference on Information Reuse and Integration (IRI), Pittsburgh, PA, 77-86.

[60] Y. Tanaka, K. Oka, T. Ono and K. Inoue. 2016. Accuracy analysis of machine learning-based performance modeling for microprocessors, 2016 Fourth International Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC), Cairo, 83-86.
[61] Manju, B. R., and Anju R. Nair. 2019. Classification of Cardiac Arrhythmia of 12 Lead ECG Using Combination of SMOTEENN, XGBoost and Machine Learning Algorithms.” 9th International Symposium on Embedded Computing and System Design (ISED). IEEE.

[62] Manju, B. R., and M. R. Sneha. 2020. ECG Denoising Using Wiener Filter and Kalman Filter, Procedia Computer Science 171, 273-281.

[63] Manju, B. R., and Parvathy Rema. 2020. A Performance based comparative study on the Modified version of Empirical Mode Decomposition with traditional Empirical Mode Decomposition, Procedia Computer Science 171, 2469-2475.

[64] Manju, B. R., and B. Akshaya. 2020. Simulation Of Pathological ECG Signal Using Transform Method, Procedia Computer Science 171, 2121-2127.