A REVIEW ON LUNG CANCER PREDICTION USING MACHINE LEARNING TECHNIQUES IN PYTHON

M. RAVIKUMAR
PG Student
Department of Computer Science
Bharathiar University
Coimbatore-641046
Ravikumarcs345@gmail.com

Dr.R. PORKODI
Professor
Department of Computer Science
Bharathiar University
Coimbatore-641046
Porkodi_r76@buc.edu.in

Abstract--**In this study, Lung cancer is the growth of cancer cells in the lungs. Because of the rising incidence of cancer, both men's and women's cancer rates have increased. Lung cancer is a disease that causes uncontrollable cell growthin the lungs. Lung cancer cannot be avoided, but it can be reduced. As a result, early detection of lung cancer is critical for patient survival. The number of people affected by lung cancer is dependent on the number of chain smokers. This paper presents the study of machine learning algorithms such as SVM, KNN, Naive Bayes, Decision Tree, Logistic Regression and Random Forest predicting lung cancer from the experimental dataset which is unbalanced, which is balanced using the SMOTE algorithm. The above algorithms are implemented successfully and the performance of these algorithms is validated using the performance metrics and proves that the SVM algorithm outperforms other algorithms.**

Keywords—*Lung cancer Prediction, SVM, KNN, Naïve Bayes, Decision Tree, Logistic Regression, Random Forest.*

I. INTRODUCTION

Lung cancer is the principal cause for cancer-related death. Lung cancer is one of the most common cancers that kills both men and women. The windpipe, main airway, or lungs can all be the site of lung cancer onset. It results from the unchecked growth and spread of some lung cell types. People with lung conditions like emphysema and a history of chest pain have a higher risk of developing lung cancer. The main risk factor for developing lung cancer in Indian men is excessive tobacco use, which includes smoking cigarettes and beedis. However, Indian women do not smoke as frequently, suggesting that there may be additional risk factors. Exposure to radon gas, air pollution, and chemicals at work are additional risk factors. Primary lung cancers begin in the lung, whereas secondary lung cancers begin in the lung and spread to other body parts. The stage of cancer is determined by the size of the tumour and the extent of its spread. A lung cancer that has not spread to nearby tissue or to another part of the body is considered to bein the early stages of the disease. A better understanding of risk factors can aid in the disease's prevention. Early detection using machine learning techniques is the key to increasing survival rates, so if we can use these techniques to make radiologists' diagnosis processes more effective and efficient, that will be a significant step in the right direction. They frequently begin in the bronchi, which are located towards the middle of the chest.

The fatality rate from lung cancer is proportional to the total number of cigarettes smoked. Primary preventative efforts include quitting smoking, changing one's diet, and chemoprevention. Secondary prevention is achieved by screening. Our approach to identifying potential lung cancer patients is based on a thorough examination of symptoms and risk factors. Some of the general indications of cancer disorders are non-clinical signs and risk factors. Human cancer is influenced by environmental variables. The air we breathe, the food we consume, and the water we drink all contain carcinogens [1]. The examination of cancer causation in humans is complicated by frequent and sometimes unavoidable exposure to environmental toxins. The intricacy of human cancer causes is particularly difficult to unravel for malignancies with a lengthy incubation period, which are often linked to exposure to common environmental toxins. Pre-diagnosis techniques. Pre-diagnosis aids in identifying or narrowing down the likelihood of lung cancer screening.

In the pre-diagnosis stage, symptoms and risk factors (smoking, alcohol intake, obesity) had a statistically significant influence. Lung cancer diagnostic and prognostic issues fall mostly within the umbrella of the much-discussed categorization issues. Many academics in the domains of artificial intelligence, data mining, and statistics have been drawn to these issues. The datasets on lung cancer used in this investigation were provided by the Kaggle Repository. Then using the classification algorithms such as SVM, KNN, Naïve Bayes, Decision Tree, Logistic Regression, Random Forest, respective classification models are implemented using the given training data and the related models are tested using testdata to check the model accuracy [2].

This paper is organized as follows: The section 2 presents related works. section 3 describing methodology for lung cancer prediction using smote (Oversampling) and machine

learning algorithms. section 4 discusses the results and discussion and the paper is concluded in section 5.

II.RELATED WORKS

J. D'Cruz, A. et al. [3] Experimented to evaluate the referral course's impact and side effects on delays in a quick outpatient indicative programme for patients suspected of having lung cancer, as well as to ascertain whether delays were related to the severity of the illness and its outcome. The traits of tumours, their structure, and the numerous deferrals that have taken place have all been thoroughly examined. For this study, a total of 565 patient restoration schematics were gathered. A total of 111 participants (19.6%) had radiological anomalies that were not considered potentially life-threatening, and 51% of the participants had lung growths. The other half (8.5%) of the participants had a variety of injuries. First-line wait times for haemoptysis were significantly lower than in others.

B. R. Manju et al. [4] They looked into many methods of measurement, including lung expansion There were numerous them, such as the application. includes self-organizing maps, linear dependency analysis (LDA), artificial neural networks, and image processing (SOM). Support vector machines should be utilised as a characterisation approach, in summary. Support vector machines are a tool for machine learning that may be used to analyse data and spot patterns.

A. Binson et al. [5] Their initial research involved developing a method to identify lung development. The Authors achieved the desired outcome by classifying information images as either harmful or benign using back-propagation neural networks (BPNN). Used ant colony optimization with ANN and SVM to predict the accuracy of -98% and 93.2% respectively on 250 lung cancer CT images.

J. Kuruvilla et al. [6] showed how to use computed tomography images and a previously described computer-aided diagnostic (CAD) order technique to detect the proliferation of neural systems. The CT scan highlights were combined, then the lung was reconstructed using those pieces. The data were classified as malignant using the mean, standard deviation, skewness, kurtosis, and the fifth and sixth central moments.

Y. Zhang et al. [7] Hybrid feature extraction claims that as a result enhancing the reliability of ECG authentication as a Data were suggested. Creation of further development of a parallel ECG pattern recognition system, which increased the recognition's effectiveness across different ECGs feature rooms. Extracted features from chest x ray images and used concept of back propagation neural network method to improve the accuracy. Outlined a review of various machine learning approaches on several cancer data and concluded that application of integration of feature selection and classifier will provide a promising result in analysis of cancer data.

S. Xiao et al. [8] A back-propagation neural network was used by them to build a model for predicting the price of commodities. The author then suggests a self-evolving trading strategy in accordance with the market regulations for futures trading and the results of the testing. To show how their strategy has changed through time, the new strategies are then contrasted with the traditional techniques. Their approach outperforms those of the other researchers for the suggested assessment indicator, according to experiments.

V. A. Binson et al. [9] They came up with a method to identify lung growth early on in their research. In this way, data preprocessing is done to begin the process of photograph enhancement. The datasets can be tested once they are prepared for information mining and neural systems, which are both essential for differentiating between rehabilitative methods. The researchers achieved the desired outcome by using back-propagation neural networks to categorise information images as either dangerous or benign (BPNN).

III. LUNG CANCER PREDICTION WITH SMOTE AND MACHINE LEARNING ALGORITHMS

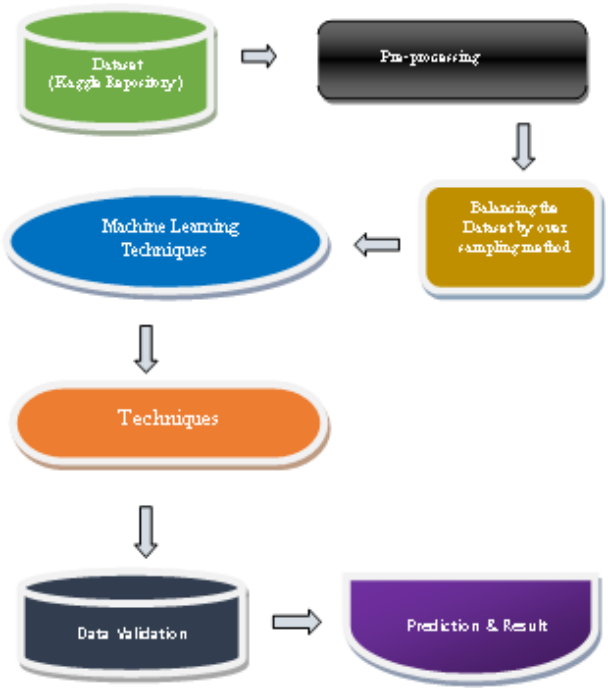


Fig.1. Framework

Fig.1 depicts the framework of this study. This framework contains Four categories namely Dataset, Pre-processing, balancing the dataset using smote and applying machine

learning algorithm and finally predict result and validate.

DATASET

The dataset used here for predicting lung cancer is taken from Kaggle Machine learning repository. Kaggle repository is a collection of databases that are used for implementing machine learning algorithms. The dataset used here is a real dataset which contains 310 Instances with 16 Attributes.

***Link**: https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction*

***Reference:** Sandra Grace Nelson*

PRE-PROCESSING

Data preprocessing is the process of preparing raw data for usewith a machine learning model. It is the first and most important step in developing a machine learning model. Pre-processing refers to the transformations applied to the data before giving to the algorithm.

BALANCING THE EXPERIMENTAL DATASETUSING

The SMOTE stands for Synthetic Minority Oversampling Technique. The component creates new instances from minority cases that you provide as input that already exist. The quantity of majority cases remains unchanged as a result of theSMOTE implementation. The new occurrences are distinct from minority cases that already exist. The algorithm instead selects samples from the feature space for each target class and its close neighbours. The algorithm then creates fresh examples that incorporate traits from both the target case and its neighbours. With this method, each class has access to more features, and the samples are more inclusive. SMOTE increases the percentage of only the minority cases after taking the entire dataset as input. Several categorization algorithms are used after the oversampling process and the data is reassemble.Fig.2. shows SMOTE technique operation [11].

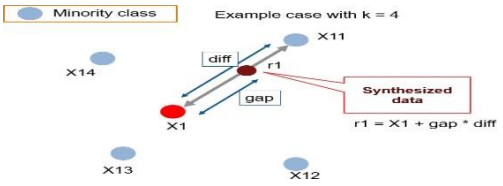


Fig.2. Smote Process Model

It is used for balancing the data before oversampling, counts of label '1': 192.

Before Oversampling, counts of label '0': 24.

After OverSampling, the shape of train_X: (384, 16)

After OverSampling, the shape of train_y: (384,) After

OverSampling, counts of label '1': 192After

OverSampling, counts of label '0': 192

IV. PREDICTION METHODS

A) Support Vector Machine (SVM)

A supervised learning technique called SVM analyses data used for classification analysis. SVM is more appropriate for non-linearly separable datasets because it lowers the misclassification rate. When using SVM, the goal is to minimise the distance between a point and the classes while attempting to maximise the distance. Fig.3 shows the SVM structure. The image has Two distinct classes are represented by the pink and blue, which are divided by a hyperplane. Additionally, the support vectors and margin are clearly marked below [12].

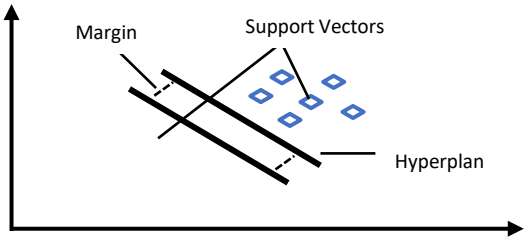


Fig.3. Support Vector Machine (SVM)

B) K-Nearest Neighbour (KNN)

One of the most straightforward approaches to classification inmachine learning techniques is the K-Nearest Neighbour (KNN) algorithm. It has been applied in numerous data mining applications. This algorithm uses the feature vector from the reference space to determine the label that the classifying object should assign to its k-nearest neighbour or neighbours (training data).

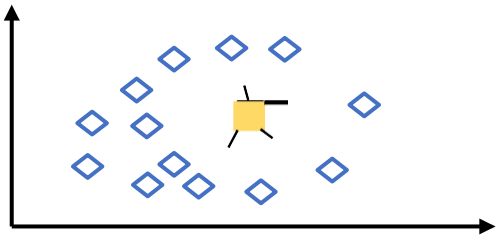


Fig.4. KNN (for k=4 neighbourhood)

This algorithm is known as the k-nearest neighbour algorithmbecause it is categorised based on the distance and the determined number of k [13]. If placed on the two-dimensional coordinate system in Fig.4 the five nearest

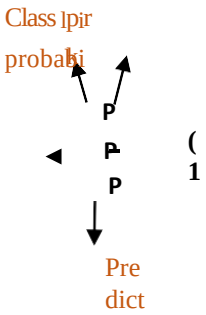
neighbour examples are used. If our k value is 5, we can categorise the object by taking and testing its 5 nearest neighbours. Circles make up 4 of the 5 samples; the fifth, square, will be placed in the circle class [14].

C) *Decision Tree*

Decision tree derives from the simple divide-and conquer algorithm. Leaves represent classes, and branches represent feature combinations that lead to those classes in these tree structures. At each node of the tree, the attribute that most effectively splits samples into different classes is chosen. To predict the class label of an input, a path to a leaf from the rootis found depending on the value of the predicate at each node that is visited. It's a knowledge representation structure made up of nodes and branches arranged in the shape of a tree, witheach internal non-leaf node labelled with attribute values. Themost common algorithms of the decision trees are ID3 and C4.5. The values of the characteristics of an internal node are indicated on the branches that emerge from it. A class is assigned to each node (a value of the goal attribute). Induction modelling is commonly implemented using tree-based models, which include classification and regression trees. [15].

D) *Naïve Bayes*

Naive Bayes is commonly used in machine learning. SVM classification is carried out using statistical methods. From a Bayesian viewpoint, a classification problem can be written as the problem of finding the class with maximum probability given a set of observed attribute values. Such probability is seen as the posterior probability of the class given the data, and
is usually computed using the Bayes theorem. Estimating this probability distribution from a training dataset is a difficult problem, because it may require a very large dataset to significantly explore all the possible combinations. To calculate the probabilities, we used the following equation. Explain the bayes theorem below Eq.1.



A conditional probability is the likelihood of some conclusion, *C*, given some evidence/observation, *E*, where a dependence relationship exists between *C* and *E*. This probability is denoted as P*(C | E)* where this conditional relationship allows an investigator to gain probability information about either *C* or *E* with the known outcome of the other. This results in the need for a very large amount of data. Fig.5 shows a new incoming X with labels C1, C2, and C3 representing the classes. The incoming X belongs to class C1 based on the probability values shown in the model [16].

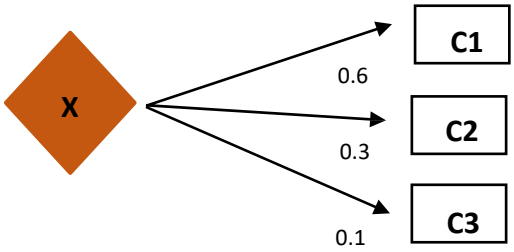


Fig.5. Naïve Bayes

E) ***Logistic*** *Regression*

Logistic Regression (LR) is a popular mathematical modelling procedure used in the analysis of epidemiologic datasets, particularly in the machine learning field.

The Logistic Regression method can be used in the following steps:

I. Use the logistic function to compute.
II. Discover the logistic regression model's coefficients.
III. Finally, use a logistic regression model to make predictions.The logistic function is shown in Eq.2.

Table.2 **L**

$$f(x) = \frac{}{1+e^{-k(x-x_0)}} \quad (2)$$

e = Euler's number.
x0 = Sigmoid function's middle x-value
L = Maximum value of curve
k = Curve abruptness.
The logistic regression equation is used to estimate an output value (y) based on input values (x). Eq.3 depicts the logistic regression model.

$$y = \frac{e^{b0+b1*x}}{1+e^{b0+b1x}} \quad (3)$$

Using training data, logistic regression parameters are estimated by maximising the logarithmic likelihood function. Fig.6 depicts an example of a logistic regression

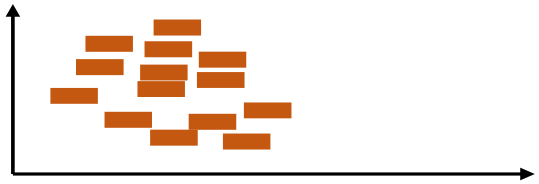


Fig.6 Naive Bayes

F) Random Forest

A random forest classifier is represented by a combination of classifier trees. One of the best methods for representing input variables in the form of trees that form a forest-like structure. Trees are used to represent input data, and each tree specifies a class label. The random forest is determined by its error rate.The error rate can be expressed in two ways. The

First is the relationship between trees, and the second is the tree's strength. The benefits of random forest.

- Appropriate method for representing noisy and unbalanced data.
- The best approach in machine learning platform for accuracy improvement.
- one of the most dependable algorithms [18].

V. RESULTS AND DISCUSSION

The experimental dataset is divided into training and testing. The training dataset is used to construct the classifier model, and the test dataset is used to validate it. In this study, 75% of the training data and 25% of the test data are used, respectively. Because our dependent variable has two categories, begin by thinking about classification problems with only two classes. Formally, each instance 1 is assigned to the positive and 0 is negative class labels. A classifier model is a mapping between instances and predicted classes. Table2. shows the confusion matrix of machine learning algorithms.

Table.2 (a) SVM

N= 93	Actual NO	Actual YES	
Predict NO	TN=07	FP=08	15
Predic t YES	FN=02	TP=76	78
	09	84	

used to distinguish between two classes (green- orange images) [17].

Table.2 (b) Naïve bayes

N= 93	Actual NO	Actual YES	
Predict NO	TN=05	FP=10	15
Predict YES	FN=02	TP=76	78
	07	86	

Table.2 (c) Logistic Regression

N= 93	Actual NO	Actual YES	
Predict NO	TN=06	FP=09	15
Predict YES	FN=05	TP=73	78
	11	82	

Table.2 (d) Decision Tree

N=	Actual NO	Actual YES	
Predict NO	TN=03	FP=1	15
Predict YES	FN=0	TP=77	78
	0	8	

Table.2 (e) Random Forest

N= 93	Actual NO	Actual YES	
Predict NO	TN=05	FP=10	15
Predict YES	FN=14	TP=64	78
	19	74	

Table.2 (f) Random Forest

N= 93	Actual NO	Actual YES	
Predict NO	TN=06	FP=09	15
Predict YES	FN=02	TP=76	78
	08	85	

Positives consider the proportion of subjects who are "true positives (TP)," that is, correctly predicted as cases, among the instance true class. False Positive considers the proportion of subjects who are "false positives (FP)," that is, are falsely predicted as cases, among the instance estimated class. Classification performance measures are frequently used. Additionally, this dataset used statistical measurements such as precision, Recall, F-measure and Accuracy of performancematrix.

The performance of the considered machine learning algorithms are presented in Table.3 and the same is depicted in fig.7 with metrics Precision, Recall, F-measure and Accuracy. The accuracy of SVM, Random Forest, Decision Tree, Logistic regression, Naïve bayes and KNN are 89%, 88%, 87%, 86%, 84% and 74% respectively it is observed that SVM algorithm outperforms other algorithms.

Classifier	Precision	Recall	F-Measure	Accuracy
SVM	0.90	0.97	0.93	0.89
KNN	0.86	0.82	0.83	0.74
Naïve Bayes	0.87	0.93	0.88	0.84
Logistic Regression	0.92	0.98	0.94	0.86
Decision Tree	0.88	0.97	0.91	0.87
Random Forest	0.89	0.97	0.92	0.88

Table.3 Performance Matrix

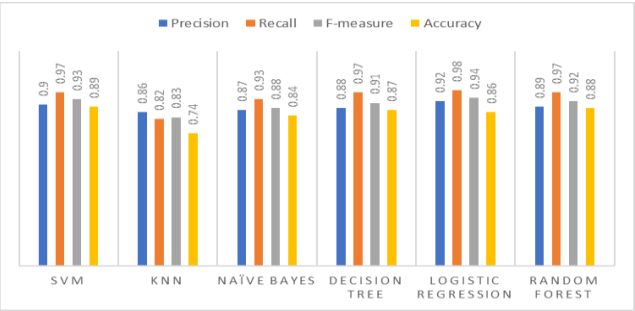


Fig.7. Performance Matrix Graph

AUC (AREA UNDER THE ROC CURVE)

The fitted model is given a receiver operating curve (ROC), and the area under the curve is computed as a measure of discriminatory performance. ROC has grown in popularity in recent years due to the availability of computer software that can easily generate such a curve and compute the area under the curve. The ROC curve is a plot of sensitivity by 1- specificity values derived from several classification tables corresponding to different cut-points

used to classify subjects into one of two or more groups, such as predicted cases and

non-cases of a disease. ROC graphs for all methods are shown in Fig.8.

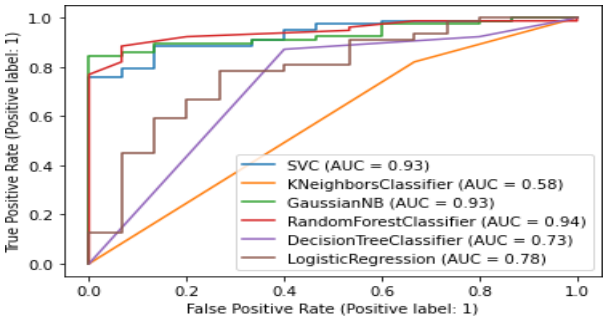


Fig.10. All Methods of AUC Curves

VI. CONCLUSION

In the past, a doctor had to perform a number of tests to decidewhether a patient had lung cancer or not. But this is a lengthy procedure. Machine learning algorithms are now used extensively in the prediction and classification of medical data.This technique can be used on medical datasets to assist physicians in making more accurate decisions about lung cancer detection. The machine learning algorithms used in this comparative study are SVM, KNN, Naive Bayes, decision tree, logistic regression and Random Forest. A comparison of the accuracy rates of each classifier is presented. Classifier prediction accuracy is statistically compared. The support vector machine (SVM) algorithm is giving the best results. The SVM algorithm used a high dimension to classify the finding, resulting in the best performance. This technique allows for more accurate lung cancer detection. Feature work may be collected with a real-time dataset and working with machine learning algorithms to predict a high accuracy rate.

References

[1] KwetisheJoroDanjuma, ” Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients” Department of Computer Science, ModibboAdama University of Technology, Yola, Adamawa State, Nigeria.

[2] Survey of Intelligent Methods for Brain Tumor Detection- IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 1, September 2014 .

[3] J. D’Cruz, A. Jadhav, A. Dighe, V. Chavan, and J. Chaudhari, “Detection of lung cancer using backpropagation neural networks and genetic algorithm,” Computing Technologies and Applications, vol. 6, pp. 823–827, 2016.

[4] B. R. Manju, V. Athira, and A. Rajendran, “Efficient multi- level lung cancer prediction model using support vector machine classifier,” in In IOP Conference Series: Materials Science and Engineering, vol. 1012, India, 2021no. 1, Article ID 012034 IOP Publishing.

[5] V. A. Binson, M. Subramoniam, Y. Sunny, and L. Mathew,“Prediction of pulmonary diseases with electronic nose using SVM and XGBoost,” IEEE Sensors Journal, vol. 21, no. 18, pp. 20886–20895, 2021.

[6] J. Kuruvilla and K. Gunavathi, “Lung cancer classification using neural networks for CT images,” Computer Methods and Programs in Biomedicine, vol. 113, no. 1, pp. 202–209, 2014.

[7] Y. Zhang, R. Gravina, H. Lu, M. Villari, and G. Fortino, “PEA: parallel electrocardiogram-based authentication for smart healthcare systems,” Journal of Network and ComputerApplications, vol. 117, pp. 10–16, 2018.

[8] S. Xiao, H. Yu, Y. Wu, Z. Peng, and Y. Zhang, “Self- evolving trading strategy integrating internet of things and bigdata,” IEEE Internet of Things Journal, vol. 5, no. 4, pp. 2518–2525, 2018.

[9] V. A. Binson, M. Subramoniam, Y. Sunny, and L. Mathew,“Prediction of pulmonary diseases with electronic nose using SVM and XGBoost,” IEEE Sensors Journal, vol. 21, no. 18, pp. 20886–20895, 2021.

[10] Metin N Gurcan, BerkmanSahiner, Nicholas Petrick, Heang-Ping Chan, Ella A Kazerooni, Philip N Cascade, and LubomirHadjiiski. Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system. Medical Physics, 29(11):2552{2558, 2012}.

[11] R. Blagus and L. Lusa, “SMOTE for high-dimensional class imbalanced data”.

[12] Lung Cancer detection and Classification by using Machine Learning & Multinomial Bayesian-IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 9, Issue 1, Ver.III (Jan. 2014), PP 69-75

[13] V. Atmaca, “Örme KumaĢlardaki Üretim Hatalarının Görüntü ĠĢleme Teknikleri Ġle Otomatik Tespiti ve Sınıflandırılması”, Ġstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Y.Lisans Tezi, Haziran 2005.

[14] P.-N. Tan, M. Steinbach, V. Kumar, “Introduction to Data Mining, International Edition”, Pearson Education Inc.,Boston, USA, 2006.

[15] R. Linder, T. Richards, and M. Wagner. Microarray data classified by artificial neural networks. METHODS IN MOLECULAR BIOLOGY CLIFTON THEN TOTOWA-, 382:345, 2007.

[16] Maria-Luiza Antonie, Osmar R. Za¨ıane, AlexandruComan Application of Data Mining Techniques for Medical Image Classification.Page 97.

[17] Isaac, J., Harikumar, S.,” Logistic regression within DBMS”, Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics,IC3I 20167918045, pp. 661-666,2016.

[18] Hussein, S., Kandel, P., Bolan, C.W., Wallace, M.B., Bagci, U.: Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches. IEEE Trans. Med. Imag. **38**(8), 1777–1787 (2019).