Lung Carcinoma Detection Using Deep Learning

Shivani Rao K, Swathi S Rao and Sathvik M N

[1,2,3]Information Science & Engineering, NMAMIT,Nitte

ABSTRACT

Nowadays, the changes in food habits and working culture among the individuals are causing many health issues, in which cancer is among the most life threatening disease amongst people. Lung cancer is most common among the cancer patients all over the world. Early detection and diagnosis of cancer helps in long survival for patients, whereas failure in early detection may cause fatal end. Data mining techniques like classification, regression and clustering are more powerful in disease detection. In this study, the deep learning algorithm, Convolutional Neural Networks (CNN) has been used on lung cancer CT scan image data set. From the experimental results, it is observed that the model achieves good accuracy.).

Keywords: CT scan, Convolutional Neural Network, Prediction, Training, Decision Tree.

1. INTRODUCTION

From many surveys, it has been proven that lung cancer is the topmost reason of most cancer deaths in human beings. The dying tempo can be decreased if humans go for early analysis in order to find suitable remedy. In short, the cancer is an uncontrolled abnormal boom of unusual cells and invades the encircling tissues. Lung cancer can be generalized in two ways, the first one is non-small cell lung cancer (NSCLC) and the second is small cell lung cancer (SCLC). Here, the focus is primarily on NSCLC patients on account that it is more complicated and difficult to treat. There are many contrasts regarding the detection and remedy of SCLC and NSCLC. There are diverse methods to locate the lung cancers, one among them is to apply its datasets SVM and LR algorithms which increases the category and prediction version.

Artificial intelligence, Machine Learning and Deep Learning are the emerging technologies, which is widely used in medical diagnosis. Nowadays living standards among people in urban cities are developing, due to which the health issues and life threatening disease such as cancer are also increasing. Artificial intelligence uses a set of machine learning and deep learning algorithms, in which regression model is vastly used and brings more efficiency on learning and detection. Lung cancer, are of two types, one is Small cell and non-small cell. The use of automated regression analysis on lung disease as affected or normal (a binary class problem) is very important in computer aided diagnosis and for accurate and early detection.

In this proposal the lung cancer detection is done through various regression techniques, predicting the presence or absence, the algorithm efficiency is compared through evaluation metrics. The further chapters describe the existing systems, proposed solution, algorithm details, results and discussion.

Many existing techniques have been studied by the researchers on lung cancer, few of them are discussed below. K. Narmada [1], the author studied through imaging techniques. They trained the dataset from kaggle bowel 2017 using convolutional neural network (CNN) algorithm. The author also used SVM classification for categorizing as Model A or Model B. Through the CT image scan, they segmented the portion of lung cancer and extracted the 15 features as text value and used it for SVM algorithm. They found this method achieves good results on classification. Hossam M. Zawbaa et.al in this paper [2] defined the work on early stage detection through an unique strategy called Ant lion optimizer plate detection from images. The author used optimization technique for the input feature before sending it to the machine learning part. The author compared the work with Genetic algorithm, wolf optimization and Ant lion optimization, from which they concluded that ALO method outperforms in terms of accuracy of detection. V.Krishnaiah author in [3] proposed lung disease prediction through rule based and classification techniques. For data preprocessing, the author applied One Dependency Augmented Na¨ıve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2), the aim was to exploit the use of hidden patterns for classification. The author experimented ODANB and proved better accuracy more than 70%, also proved that it determines early stage of cancer through symptoms and patient details. Animesh Hazra in [4], the author discussed logistic regression and SVM for lung cancer prediction. Through the experimental studied various evaluation metrics including true positive, true negative and false positives and false negatives. Through experimental results it is proved that logistic regression is achieving more accuracy than SVM algorithm. This work motivates us to do regression research than classification and clustering. Fei Xie in [5], the author discussed about license plate detection for Chinese plates, they used Back propagation Neural network (BPNN) for character recognition. The author specifically designed the Chinese plate detection with specific length and width parameter. For identifying characters they used BPNN, which is trained with 50 epoch. However their model is good, but applicable only to Chinese number plates. Numan et.al [6] discussed about lung cancer prediction through Artificial neural networks (ANN) algorithm, in which the exploited the used of macro neural networks, the simple neural network connected with no hidden layers to avoid over fitting problem. Their model shows significance performance on lung cancer prediction accuracy. J James A. Bartholomai in [7], the author studied the lung disease detection through SEER data. The author used five ensemble models of classification to identify ROC curve. The prediction accuracy achieved is as high as 90% for the SEER data. The dataset used in this study was lung image dataset,were pre-processed. Faezeh Hosseinzadeh in paper [8] discussed the prediction through protein attributes as Small cell Lung cancer and NonsmallcellLungcancerasbinaryclassproblemon1497protein attributes. They used feature selection technique and filtered most important feature of only 12 attributes. They used SVM and ANN for prediction of disease. SVM outperforms in terms of accuracy of lung cancer detection. A.Goyal in [9], the author discussed lung cancer prediction through classification techniques. Use of Naive bayes and J48 algorithm for prediction using Weka tool and classified as three classes namely A, B, and C. In their experimental results J48 model outperforms the other models. However, their technique is not much effective as they used all are classification models. Shraddha G. Kulkarni in [10], the author discussed lung cancer detection through image processing technique. The author applied Gabor filter and Gaussian rules for preprocessing. The proposed technique was efficient through segmentation of tumor area. However, to use image processing for tumor detection, the image quality should be high.

2. METHODOLOGY

This chapter explains the implementation stage of the project. The implementation phase may be seen as 3 separate modules which are:

- Dataset assortment
- Data pre-processing
- Training and prediction victimization Regression Models

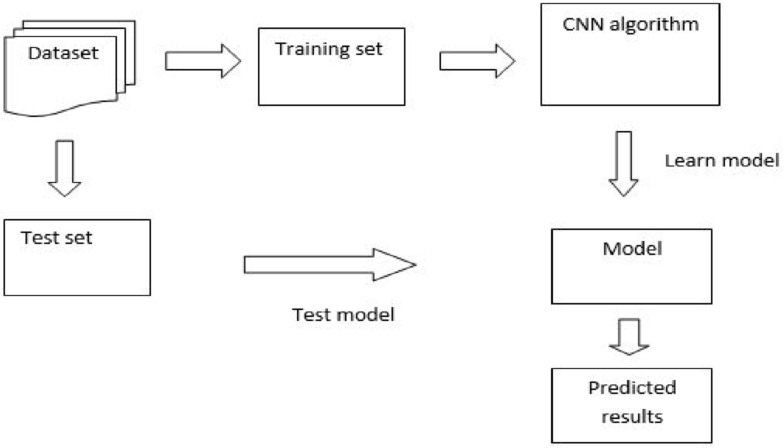


Figure 1 Overall System Design

Dataset Collection: The collection of feature datasets, derived from chest Computed Tomography (CT images) is what describes the dataset, which can be used in the treatment of Chronic Obstructive Pulmonary Disease (COPD).The images given in this database are weakly labelled, i. e. per image, a diagnosis (COPD or no COPD) is stated, but it is not known to us the parts of lungs that are affected. Furthermore, the images were obtained from different sites and scanners.

Data Preprocessing: The dataset obtained is split into train and test data set. The images which are to be given as input are pre-processed with colour conversion. After pre-processing the image, features are given as input to CNN algorithm. Training using Convolutional 2D Neural Network: For training and testing our model we used convolutional 2D neural network which is available in Keras.

A. Sequential Model Keras supports two types of models – they are Sequential and via the Functional API .The Sequential model is probably going for most deep learning networks .It permits us to easily
stack sequential layers (and even recurrent layers) of the network so as from input to output.

B. Adding 2D Convolutional Layer The addition of 2D convolutional layer is vital.The first argument passed to the Convo2D() layer function is the number of output channels-in this case we have thirty two output channels. The next input is that the kernel size ,which we have taken to be a 5x5 moving window, which would be followed by the strides in the x and y directions(1,1).Following we have a activation function which is a rectified linear unit and then finally we have to supply the model with the size of the input to the layer .Declaring the input shape is only necessary for the first layer – Keras is good enough to work out the size of the tensors flowing through the model.

C. Adding 2D max pooling Layer In this step we specify the size of the pooling in the x and y directions – (2,2) in this case, and the strides .Max pooling is used to reduce the size of the input image. It also helps in the reduction of spatial dimension of the output image.

D. Adding another convolutional max pooling Layer Next step is to add another convolutional max pooling layer ,which consists of 64 output channels .The default strides argument in Conv2D() function is (1,10) in Keras, so we can leave it out . The reason behind the default argument in Keras is to make it equal to the pool size. The input tensor for this layer is (batch size , 28,28, 32) the 28 X 28 is the size of the image, 32 is the number of output channels from the previous layer.

E. Flatten and Adding Dense Layer From the above step,the output is flattened so that it launches into the connected layers. The next two lines determine the fully connected layers –by using the Dense() layer in Keras, we can specify the size – in line with the architecture , we also specify 1000 nodes, each activated by a ReLu function .Next is the size of the number of classes, or output layer, which is our soft – max classification.

F. Training neural network A training model specifies the loss function, or determines the framework indicating the optimiser type to be used (i.e. gradient descent, Adam optimiser etc). Lass function of standard cross entropy for categorical class classification (keras.losses.categorical cross entropy) and Adam optimiser (keras.optimiser.Adam) are used. Eventually, we can then specify the metric that will be calculated when we run evaluate() on the model. We first pass all the composed training data – in this case x train and y train. The batch size is the next argument. In this case we are using a batch size of 32.As a next step the number of training epochs is passed (which is 2). The verbose flag is set to 1 here, specifies if you want detailed information being printed in the console about the progress of the training happening.

RECOGNITION: Finally, we pass the validation or test data
to the fit function so Keras recognises the data to test and the metric against which evaluate() function is made to run on the model.

ALGORITHM DESIGN

Image Data Input Parameters : number of images, image height, image width, number of channels, number of levels per pixel.

Image Preprocessing : Aspect Ratio : Square Cropping, Significance to the middle of the image.

Image Scaling : Need to use 60X60. Use of image size to the power of 2.

Dimensionality reduction: Conversion of RGB into Gray image (reduces 3 channels to

1) Architecture Inputs:
2) Number of Layers
3) Number of Neurons in each Layer
4) Regularization Parameters
5) Learning Rate
6) Dropout Rate
7) Weight Sharing
8) Activation Function (linear, sigmoid, tanh, relu)
9) Loss/Divergence Function (MSE, Cross Entropy, Binary Cross Entropy)
10) Algorithm for Weight Updates (SGD, Adam, RMSProp.)

G. Sequential Model

Sequential Model API : We can create a model layer by layer. We cannot create a model that can share layers or have multiple inputs and outputs.

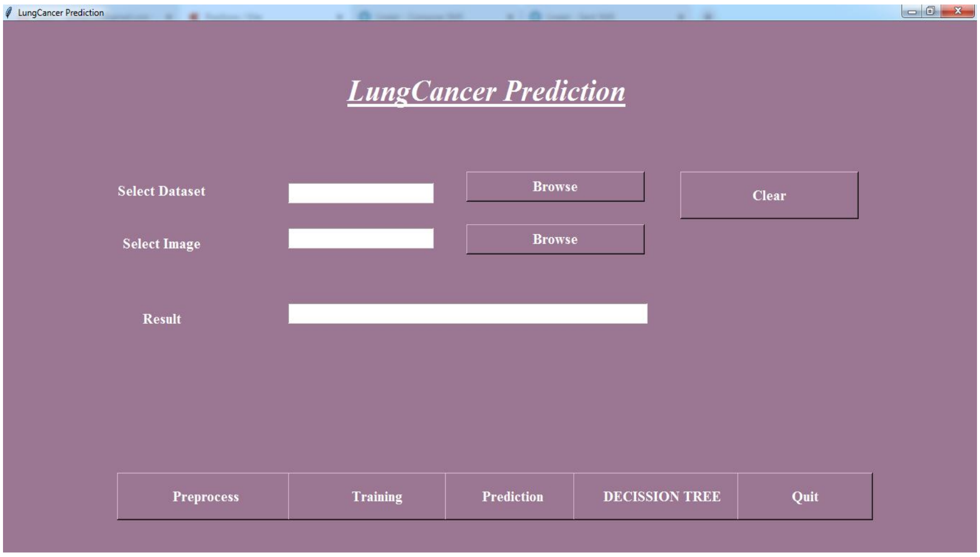


Figure 2 User interface of application

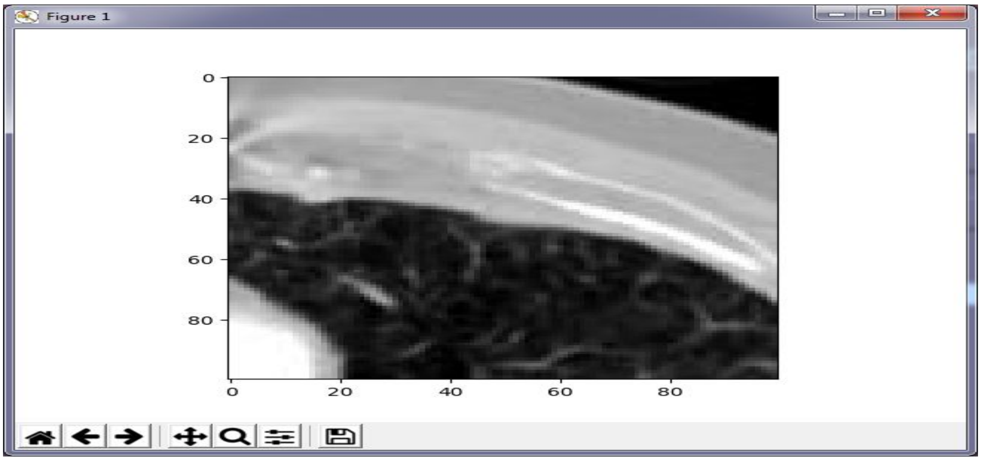
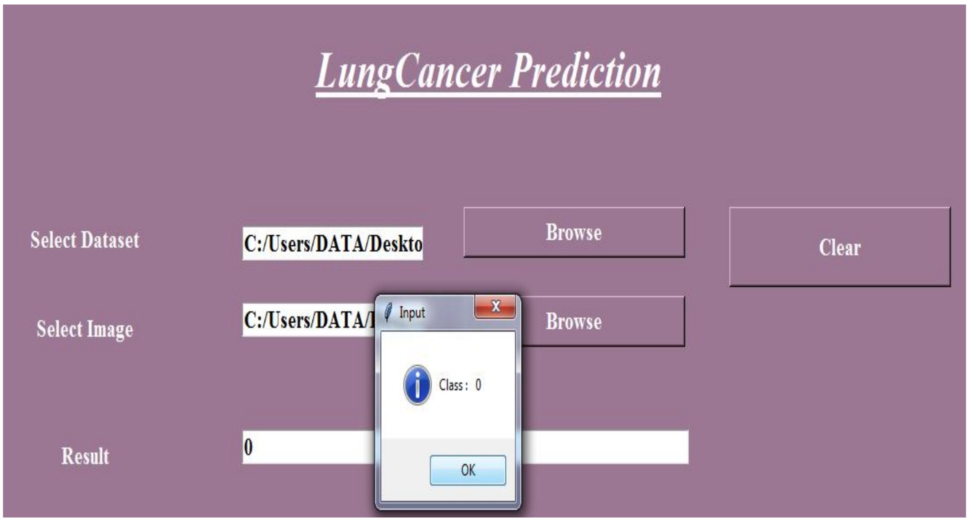


Figure 3 Visualization in pre-process stage



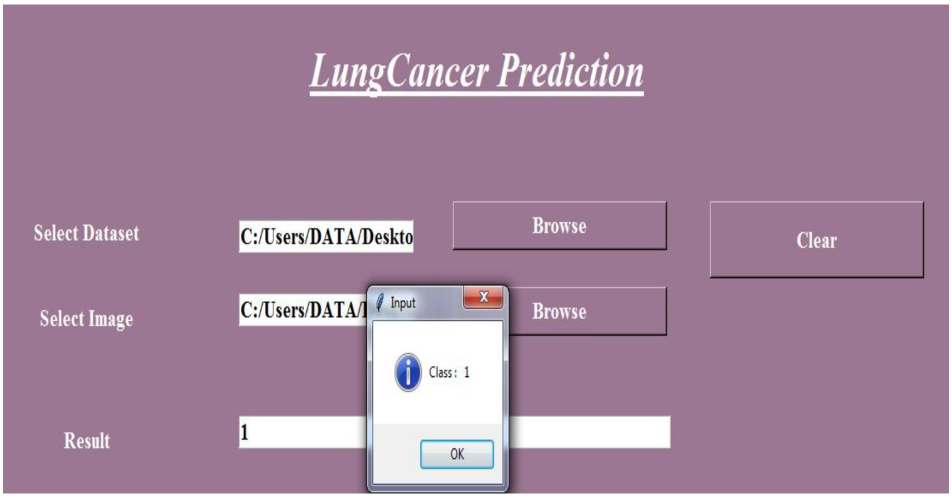
Select Dataset C:/Users/DATA/Deskto

Select Image C:/Users/DATA/I

Class : 0

Result 0

Figure 4 Detection of class 0



LungCancer Prediction

Select Dataset C:/Users/DATA/Deskto

Select Image C:/Users/DATA/I

Input

Class : 1

Result 1

Figure 5 Detection of class 1

3. RESULTS

The proposed work is implemented in Python 3.6.4 with libraries scikit-learn, pandas, matplotlib and other mandatory libraries. In general Neural Network consists of different hidden layer. In most of Conv2D will have two hidden layers with 16 or 32 neurons and more, Hidden layers are multiplied with different random weight of image pixel data which is between 0 to 1. But in Conv2D Neural Network was design with same two hidden layers and each hidden layers consists of large set of neurons i.e. we used 128 dense neurons are taken and this are multiplied with random weights. By using this deep network we got promising results. Here we displaying some labeling of characters in Table 1.

Table 1: Sample data representation of labeling

Layer Type	Layer operation	Feature map size	Total parameters
C1	Convo 2D	60x60	3032

The following figure shows accuracy of training and validation sets

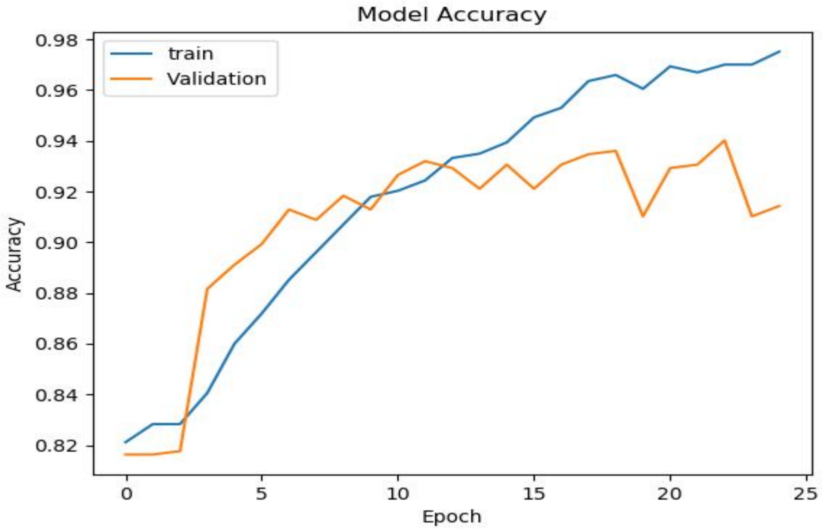


Figure 6 Training and Validation set Accuracy

The following figure shows loss of training and validation sets

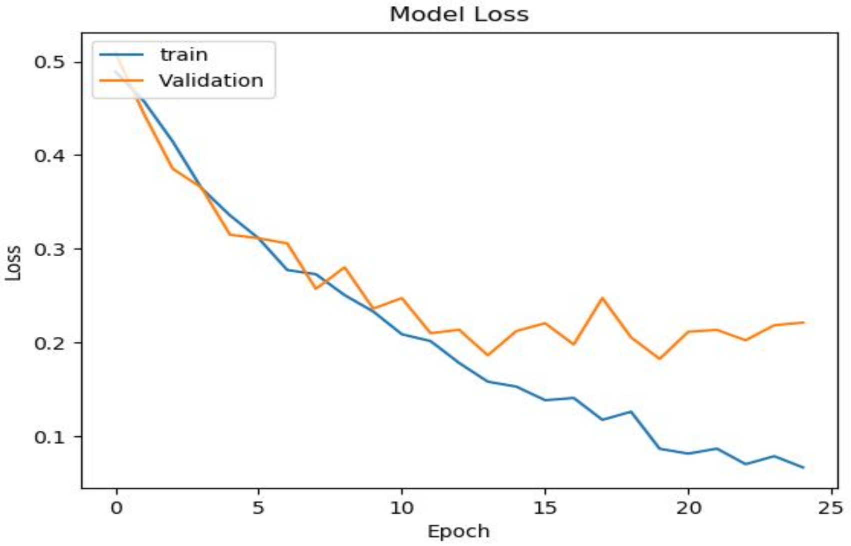


Figure 7 Training and Validation set Loss

The following figure shows the evaluation metrics of Decision Tree Algorithm

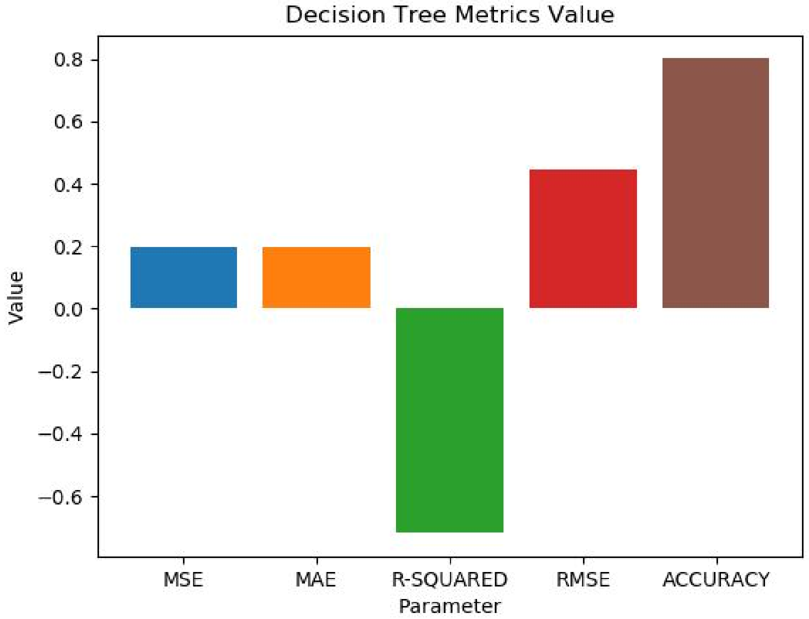


Figure 8 Evaluation Metrics of Decision Tree Algorithm

4. CONCLUSION

Lung cancer is a life threatening disease all over the world, changes in life style and people living in polluted area are the main basis of high rate deaths. The other root cause of death is detection leading to loss of life. The early detection would the save the patient’s life. There is a need for more effective system, for early detection lung cancer. In this proposed system, CNN has been applied for predicting lung cancer and evaluation metrics for training and validation sets are arrived at. Also, an experimental result of the Decision Tree algorithm performance for Lung cancer detection is drawn. The future scope of this work can include additional deep learning models such as Recurrent Neural Networks (RNN), LSTM models and achieve accurate detection rates.

References

[1] K. Narmada, G. Prabakaran, S. Mohan, ”Classification and Stage Prediction of Lung Cancer using Convolutional Neural Networks”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-10, August 2019. R. Caves, Multinational Enterprise and Economic Analysis, Cambridge University Press, Cambridge, 1982. (book style)

[2] Hossam M. Zawbaa,Eid Emary,Bazil Parv, ”Feature Selection Based on Antlion Optimization Algorithm”, IEEE Third World Conference on Complex Systems(WCCS), At Marrakech, Morocco.

[3] V.Krishnaiah , Dr.G.Narsimha*, Dr.N.Subhash Chandra, ”Diagnosis of LungCancer Prediction System Using Data Mining Classification Techniques”, V.Krishnaiah et al, / (IJCSIT) International Journal of Computer Science andInformation Technologies, Vol. 4 (1) , 2013, 39 - 45.

[4] Animesh Hazra,Nanigopal Bera,Avijit Mandal, ”International Journal of Com-puter Applications ”, (0975 – 8887) Volume 174 – No.2, September 2017 19 Predict-ing Lung Cancer Survivability using SVM and Logistic Regression Algorithms.

[5] Fei Xie , Ming Zhang ,Jing Zhao , Jiquan Yang,Yijian Liu, and Xinyue Yuan, ”ARobust License Plate Detection and Character Recognition Algorithm Based on aCombined Feature ExtractionModel and BPNN”.

[6] Numan Goceri, Evgin Goceri , ”Deep Learning In Medical Image Analysis: RecentAdvances And Future Trends”.

[7] J James A. Bartholomai and Hermann B. Frieboes, ”Lung Cancer Survival Pre-diction via Machine Learning Regression, Classification, and Statistical Techniques”, Proc IEEE Int Symp Signal Proc Inf Tech. 2018 Dec; 2018: 632–637.Published online 2019 Feb 18

[8] Faezeh Hosseinzadeh, Amir Hossein KayvanJoo, Mansuor Ebrahimi, and BahramGoliaei, ”Prediction of lung tumor types based on protein attributes by machinelearning algorithms”, Published on 2013 May 24.

[9] A.Goyal and R.Mehta, “Performance comparison of Na¨ıve Bayes and J48 classification algorithms”,published in 2012.

[10] Shraddha G. Kulkarni and Sahebrao B. Bagal , “Lung Cancer Tumor Detection Using Image Processing And Soft Computing Techniques”,Vol.No.5,IssueNo.05,May 2016.