IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-13, May 2019 E-ISSN: 2347-2693

A Framework for Lung Cancer Survivability Prediction Using Optimized-Deep Neural Network Classification and Regression technique

Pradeep K.R [1*], Naveen N.C [2]

[1]Dept. of CSE, K.S. Institute of Technology, VTU, Bengaluru, India
[2]Dept. of CSE, JSS Academy of Technical Education, VTU, Bengaluru, India

[*]*Corresponding Author: pradeepkr22@gmail.com., Tel.: +91-9886171714*

DOI: https://doi.org/10.26438/ijcse/v7si13.5765 | Available online at: www.ijcseonline.org

***Abstract*—** Lung cancer disease is the most widely recognized deadly disease in the world for loss of life. Throughout this research, Electronic Health Records (EHRs) textual data are investigated and survivability rates for lung cancer affected patients are predicted. If the patients are survivable more than one year, chemotherapy treatment can be started for those patients. This research paper examines an effective Batch Size-Optimizer based Deep Neural Network (Op-DNN) classifier framework model, which is developed to predict the patient's survivability based on status dead or alive. Considering only the patients who are alive, prediction is done to know how many months the patients will survive by Op-DNN regression technique. Here the textual data set is classified and processed in batches for each iteration. The errors generated from the original classification of the first batch size is fed back to the Op-DNN algorithm for further iterations with the reduced error loss that are free from underfitting and overfitting. The proposed method is compared with various parameters for Machine learning classifier algorithms demonstrating that the Op-DNN model has achieved better accuracy.

***Keywords*—** Lung Cancer, Diabetes, Survivable Rate, Artificial Neural network, DNN, Op-DNN, Classifier, SVM, NBs, C4.5, Optimizer, Adam, Relu, Epoch, Batchsize,Op-DNN Regression.,

I. INTRODUCTION

Lung Cancer is the leading cause of death each year worldwide. As of today, there has been limited research focused upon the lung cancer survivability prediction. However, the improvement in medical Artificial Intelligence (AI) has been connected to the development of AI programs planned to support in the design of a prediction model that can be used to make decisions. The DNN concept is based on Neural Networks and deep learning approaches such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) systems that are widely used for expert prediction behaviour modelling. The key objective of this research work is to implement a novel approach for lung cancer survivability prediction model based on Optimized - Deep Neural Network (OP-DNN). This model with the combination of batch size, epoch and optimizer provides survivability prediction by addressing the problem of over fitting and under fitting capabilities as compared to using a single methodology.

The classification based on DNN is primarily a Back Propagation (BP) algorithm, whose prototype usually implements BP Neural Network (BPNN) [1]. Relative to other kinds of classification approaches, the classification approach centered on neural network has a high learning capability. However, the degree of convergence and the generalization capability is not strong and it is easy for running into local minimum point which increases data grouping accuracy. However, it increases classification period efficiency by computing the gradient of a multi-variable function.

Rest of the paper is organized as follows; Section 2 contains related work. The proposed Methodology is presented in Section 3 The experimental details and results obtained are presented in Section 4. Section 5 contains the conclusion part.

II. RELATED WORK

Artificial Intelligence (AI) find its applications in several healthcare applications and for structured data ML can be used. For the unstructured data the neural network, classical support vector machine, natural language processing and modern deep learning may be applied. The AI is applied in identifying the diseases related to neurology, cancer and cardiology [2].

This research work presents a novel way which classifies the clinical text automatically at the sentence level. The complex features are represented by the application of deep conventional neural network. The network is trained based on the dataset which provides a wide classification of heath data. The new method achieves an increased 15% performance in comparison to other natural language

processing tasks. The convolutional deep networks based on multi-layer are capable of generating many optimal features in the course of training stage to denote semantics related to the sentence under analysis [3].

In the research work carried out for prediction of the Parkinson disease, relates with a new type of systems helping the diagnosis and the personalized evaluation of Parkinson disease. The directed systems are basically end to end deep neural architectures. Several works are completed by combining the neural networks along with knowledge illustration, it consists of deep neural networks and based on extracting rules after networks which are trained [4].

The cancer stage is mainly based upon its size with stage 1 and stage 2 referring to cancer related to the lungs. The present diagnostic techniques available are biopsy, imaging and CT scans. The survivability of the patient is based on the early detection of cancer which is the key point for the treatment [5]. AI finds its application in the diagnosis of medical related diseases such as diagnosis of the lung cancer, hepatitis etc. The development of several AI algorithms increases the decision capability of the network with the increase in hidden layers [6].

Researchers have combined ANN weights after optimization with pruning as well as using Genetic Algorithm (GA). The ANN along with GA shows higher convergence, greater success rate and lesser execution time during the test stage. The method minimizes the resources required for computation and also proves to be an alternative for neuro-genetic design of neural classifiers [7].

For the diagnosis of lung cancer, the deep learning framework using computer aid is applied. The framework based on the multi stage identifies the nodules in 3D lung CAT scans. It determines whether the nodule is malignant and grounded on these results it assigns the probability of cancer. Deep convolution neural network performs better in the classification of the images, different visual tasks and object detection. Deep neural networks perform better in the 3D segmentation whereby a neural network contains several connections of processors known as neurons [8]. Each neuron produces an order of real-valued activations. The activation of the input neurons is by the sensors recognizing the setting. The different neurons are activated by the weighted connections through the earlier active neurons [9]

DNN is a subarea of machine learning. The increase in algorithms which are based on the human intelligence, innovations in the hardware devices to process and store bulk data sets enabled the DNN development. The different machine learning algorithms like fuzzy logic, Naïve Bayes, clustering, genetic, support vector machine, neural network, random tree and decision tree etc. finds extensive applications in the detection, diagnosis, classification and the risk assessment of the cancer [10].

Machine learning is effective in mining features where the extraction of specific feature depends on the algorithm. Deep learning provides a foundation for the generation of novel and enhanced algorithms for data generalization by the computer. Deep neural networks are unique in comparison to the conventional neural network due to the enhanced universal approximation properties. This is basically due to the application of many numbers of layers which gives flexibility of estimates of diverse function classes [11].

Deep learning technique works on the concept of Artificial Intelligence, in order to overcome the dimensionality reduction problem [12]. Deep Neural Networks (DNNs) includes the classification and the regression techniques which evaluates the parameters that are outside the scope of remaining algorithms. DNNs concepts are prominent in simplifying to the new data where it requires less time in pre-processing the data [13].

In order to forecast continuous values, regression techniques are extensively used. Regression methods are used in medical field to predict the detection of a facial landmark, estimation of head pose, estimation of age, registration of image or estimation of human pose [14]. The architectures of deep learning outpaced conventional vision tasks like object detection or image classification [15].

The convolutional neural networks are useful in solving the regression problems. Regression is also conveyed as classification [16]. Here researchers found out how Multiple linear regression analysis adopts a linear relationship amongst multiple independent variables (X1, X2, ..., Xn) and dependent variables (Y), which determines the effects of every independent variable (β) using the subsequent expressions [17].

The author compared the Normal ANN regression model with the novel OP-DNN linear regression model which provided a good prediction rate by increasing the number of hidden layers by removing the overfitting of result [18].

The literature review is carried out in the machine learning algorithms applied to detect the cancer at the early stage to predict the probability of survivability.

III. PROPOSED OP-DNN APPROACH

In the field of medical research to diagnose the cancer at the early stage is a challenging task. The patient who suffers the consequences of cancer may not know the symptoms. The early detection and proper medicine taken at the initial stage itself may give result to treat and cure the cancer. In this direction the ML algorithms are important in the present scenario to attempt to predict the survivability of lung cancer patients and its features but not focusing on diabetes and smoking as main features. The planned methodology adopted in the research work attempts to predict the survivability of lung cancer patients with Deep learning technique considers, gender, diabetes, smoking along with

other features is performed such that it will be help full for the doctors to recommend for chemotherapy based on the number of months that the patients will survival. The proposed methodology is based on the DNN for the classification, where the classifier is forecasting the status of lung cancer patient survivability status is as shown below.
The proposed system architecture of Op-DNN is shown in Figure 1, consisting of four phases, which undergoes the following steps.

- Lung cancer dataset acquisition.
- Exploratory data analysis includes missing value action and normalization along with the process of splitting data for training and testing.
- Op-DNN classifier approach to predict survival status(dead/alive)
- Comparison of algorithms with visualization.
- Op-DNN regression to predict survival month for alive patients.

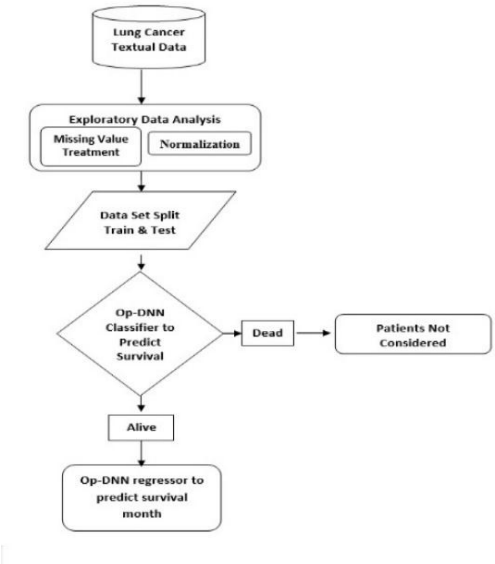

Figure 1: OP-DNN system architecture

A. Lung cancer dataset acquisition

The data collection is done from the patients after admitting to the hospital. the various data gathered are age, gender, tumor location, stage timing, n-stage, t-stage, diabetes, smoker. The survey is carried out along with the doctors as well as pathologists at different hospitals to select the features. The North Central Cancer Treatment Group (NCCTG) lung cancer data set [19] along with new patient data is used and data features are extracted for evaluating the performance. This forms the required features which are needed for the prediction with a probability of more information and non-redundant [20].
The feature set mined is added to OP-DNN and prediction is performed. The results of the proposed OP-DNN model is compared with other MLT classifier algorithms like DNN, SVM- Linear, SVM- polynomial, Naive Bayes (NBs), classification trees (C4.5), J48 where the proposed model provides the best predicted results.

B. Exploratory data analysis

In this phase the total number of textual data present in a dataset is 3065 and the feature set available count is 18 that are facilitated for training plus testing of data essential for measuring decent accuracy. In a total count of 3065 lung cancer data set 2137 are male and 928 are female lung cancer patient's data in which 1726 lung cancer patients are Type 1 Diabetic (T1D), 916 are Type 2 Diabetic (T2D) and non-diabetic are of 424, as of to consider smoking feature set, 1762 are smoker and 1303 are non-smoker lung cancer patients. In the data set lung cancer patient's survival status includes the count of 2628 are alive and 437 are dead [21]

1) Missing value treatment and Normalization

To get rid of the missing values all the 18 features from the dataset samples are treated by normalization technique by mean, standard deviation with min 25%, moderate of 59%, 75% of max values.

2) Split data for training and testing

A large set of labelled data are trained by using deep learning models based on neural network architectures which learns features right from the data without the need for manual feature extraction. In this regard, out of 3065 lung cancer data sets, train data and test data are considered in the ratio of 80% and 20%, so that, train data count is 2452 and test data count is 613.

C. OP-DNN Classifier approach

The network based on the deep learning is separated with the single hidden layer with respect to the depth. The depth is the layers of nodes by which the data passes in the recognition of data. The basic type of neural network like the first perceptions is of only one input layer and one layer of output may contain only one hidden layer. Any network with greater that three layers called as deep learning. With respect to the unlabeled data training every node layer within the deep network acquire the features automatically. It is because of repetitive action of reconstruction of the input after getting the samples. It tries to lessen the network guessing and input data probability distribution. The OP-DNN architecture with optimizer as shown in Figure 2.

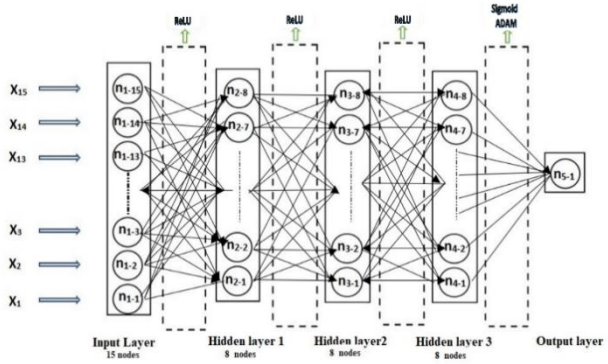

Figure 2. OP-DNN architecture with optimizer

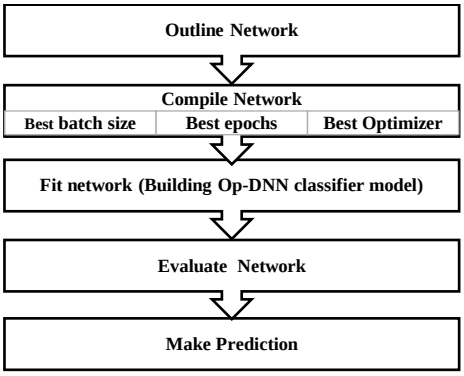

Figure 3. OP-DNN classifier

OP-DNN classifier follows six steps to predict the survivability of the lung cancer patients, flow chart of the OP- DNN classifier is shown in Figure 3.

1)Outline network: This network includes initializing the DNN classifier, here Keras Sequential model is applied for initializing the DNN classifier that involves linear stack of layers to identify specifying the input shape for the situation which would anticipate [22]. Therefore, the first layer is a Keras Sequential model and the rest all following layers will be able to do automatic outline inference which requests to accept knowledge about its input shape.

The following are the sequence of actions that take place in defining the DNN for one 15 input feature set units of Lung Cancer patient's data, three hidden layers each consisting of eight units and one output layer having a total of 41 nodes and 264 edges.

a) Adding the input layer and the first hidden layer includes to input shape of a sequential classifier to the first layer consists of a tuple of integers where the batch dimension is not considered, for specification of the input shape 2D layer dense approach has been applied for the input dimension of 16 feature set of data with eight units in first hidden layer, neural network has neurons that work in correspondence of weight, bias and their respective activation function here the activation function used is Relu[23].

b) In each hidden layer back-propagation of neural network is done by Keras Dense and Keras Kernel Initializer that propagates the updating of weights and biases of the neurons on the source of the error at the output followed by a non-linear activation function.

c) Activation function decides, whether a neuron should be activated or not, by calculating weighted sum and further adding bias with it which updates the nodes passing of the information to next two hidden layers.

$$Y = activation\ function \sum(weights * inputs) + bias\ Z(X,1) = W(1)X + b(1) \quad (1)$$

Here,

Z (1) is the vectorized output of hidden layer 1

W (1) be the vectorized weights assigned to 8 units of neurons in hidden layer1 i.e. w_1, w_2, w_3.

X be the vectorized 15 input feature set of i.e. i_1, i_2 till i_{15}

b is the vectorized bias assigned to neurons in hidden layer i.e. b_1 and b2

b (1) is the vectorized form of any linear function.

d) Relu is a rectified linear unit used as a non-linear activation function in normal DNN which are applied to the hidden layers of neural network this helps to build an efficient and easy design that back-propagate the errors and activate the neurons multiple layers as in (2).

$$A(x) = max(0, x) \quad (2)$$

e) This provides an output x if x is positive and 0 otherwise with Value Range of $[0, inf]$

f) In this section we will present a minor analysis that provides insight towards the loss, as an estimate or a higher bound to per sample gradient norm. So, let x_i, y_i be the i^{th} input-output pair from the training set, $\Psi(\cdot; \theta)$ be a deep learning model parameterized by the vector θ, and $L(.;.)$ be the loss function to be minimized during training. To achieve it, let L (ψ, y): D → R be either the negative log likelihood through a sigmoid or the squared error loss function defined respectively as

$$L1(\psi, y) = -\log\left(\frac{exp(y\psi)}{1+exp(y\psi)}\right)\ y \in \{-1, 1\}\ \psi \in R \quad (3)$$

$$L2(\psi, y) = \|y - \psi\|_2^2\ y \in Rd\ \ \psi \in Rd \quad (4)$$

Given our upper bound to the gradient norm, from the (3) and (4), defines (5).

$$\|\nabla\theta t\, L(\Psi(x_i;\ \theta_t),\ y_i)\|\, 2 \leq L\rho\, \|\nabla\Psi L(\Psi(x_i;\ \theta_t),\ y_i)\|2 \quad (5)$$

g) With the estimation of the above due facts, hypothesis is confirmed from our trials, where the loss scuffles to attain a speedup in the initial phases of training where utmost trials still take moderately large loss values.

h) Adding the second, third hidden layer, each of 8 nodes develops the normal DNN and weights, bias regulation continues for back propagation adjusting the weights with the relu activation for each layer of nodes as stated in the step1 until it reaches the final output layer.

i) Adding the output layer includes Keras Dense with one output node, Keras Kernel Initializer and Sigmoid activation. The multilayer perceptron uses the sigmoid as the transfer function. For Sigmoid activation considers a two-class problem for lung cancer patient's data, with classes identified as stay alive for more than one year as S1 and Less representing that the patient might stay alive less than 1 year as S2 indicated to S1 to 1 and S2 as 0.

j) The sigmoid activation function produces a continuous value in the range 0 to 1, shown in the (6).

$$\text{output}_{i=}\frac{1}{1+e^{-x}} \quad (6)$$

Where $x = \text{gain} \cdot \text{activation}_i$

A variant of the sigmoid transfer function is the hyperbolic tangent function shown in the (7)

$$\text{output}_{i=}\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (7)$$

2) Compiling the Network: Involves training OP-DNN classifier used in best parameter combination of optimization method along with gradient descent [24]. This technique computes the gradient of a loss function with respect to all the weights in the network. The gradient is fed to the optimization method which in turn uses it to update the weights, in an attempt to minimize the loss function. Although, these gradient techniques are severely limited in their ability to find overall results. To overcome this problem, ADAM optimizer [25] is used along with the solution to minimize the loss, Binary Cross Entropy is used. Here the main objective of the function is having less value of mean square error function (loss/cost function). To find optimize values weights of normal DNN is to minimize the objective function. Along with this, calculation of metrics that includes, accuracy, f2 score, precision, recall is done based on Wrapper classes for turning Tensor Flow metrics into keras metrics. The equations below explain the working feature involved. The Adam optimization algorithm is a combination of gradient descent with momentum and RMSprop algorithms where it works well even with a slight fine-tuning of hyperparameters which is shown below.

a. First step, calculate an exponentially weighted average of previous gradients, store it in variables G_w and G_b earlier to bias correction and $G_{w^{corrected}}$ and $G_{b^{corrected}}$ done with bias correction.

b. The second step, at this instant calculate an exponentially weighted average of the squares of the past gradients, and store it in variables SG_w and SG_b earlier to bias correction and $SG_{w^{corrected}}$ and $SG_{b^{corrected}}$ done with bias correction ,on iteration *i*, calculate the results of weights: w & bias: b using current mini-batch.

c. Lastly, update parameters on joining information from first step and second step

Initialize G_w , SG_w, G_b and SG_b to zero

Update G_w and G_b like momentum

$$G_w = \beta_1 * G_w + (1-\beta_1) * w \quad (8)$$

$$G_b = \beta_1 * G_b + (1-\beta_1) * b \quad (9)$$

Update SG_w and SG_b like Rmsprop

$$SG_w = \beta_2 * SG_w + (1-\beta_2) * w^2 \quad (10)$$

$$SG_b = \beta_2 * SG_b + (1-\beta_2) * b^2 \quad (11)$$

d. In Adam optimization technique, implement bias correction

$$G_{w^{corrected}} = \frac{G_w}{\left(1-\beta_{1i}\right)} \quad (12)$$

$$G_{b^{corrected}} = \frac{G_b}{\left(1-\beta_{1i}\right)} \quad (13)$$

$$SG_{w^{corrected}} = \frac{S_w}{\left(1-\beta_{2i}\right)} \quad (14)$$

$$SG_{b^{corrected}} = \frac{S_b}{\left(1-\beta_{2i}\right)} \quad (15)$$

e. Update parameters w and b

$$w = w - LR * \left(G_{w^{corrected}} \Big/ \sqrt{(SG_{w^{corrected}} + \varepsilon)} \right) \quad (16)$$

$$b = b - LR * \left(G_{b^{corrected}} \Big/ \sqrt{(SG_{b^{corrected}} + \varepsilon)} \right) \quad (17)$$

Where epsilon ε is a very small number to avoid dividing by zero, β_1 and β_2 are hyper parameters that

regulates two exponentially weighted averages, here the default values for $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

$$-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + (1-y_i)\log\left(1-\hat{y}_i\right)\right] \quad (18)$$

3)Fit the network: Building the Op-DNN classifier model with preparation of training and tuning the lung cancer dataset. For training the data, fit method implied again in step 2 involving best batch size and Epoch that fits the Novel Op-DNN classifier model.

In former steps we have shown how optimizing the given weights helps to improve model efficiency, to do this updating of weights are done by the concept of batch size which specifies the extent of observation required for updating the weights that includes number of training samples in one forward or backward pass, the greater the batch size, the more the memory space required. Epoch specifies one forward pass as well as one backward pass of all the training samples, done through the total number of iterations, specifying number of passes, done using batch size.

$$1\ epoch = \frac{total\ training\ samples}{Batch\ Size} \quad (19)$$

Best value of batch size and epoch is 10 and 50 as acknowledge by the Best Parameter Selector Engine from the steps 2, 3 and step 4 passes on the next step 5 for the evaluation and prediction of the test data set. In this stage evaluation metrics is used to know, is Op-DNN model finest, free from under fitting and over fitting for the trained data. The Table 5, Table 6 and graph shown below in Figure 6, depicts the novel Op-DNN classifier model for the training data set size of 2452 out of 3065 lung cancer data set, performs well with reduced error loss and maximum accuracy, when compared with existing DNN model graph Figure 7, in each iteration of epochs vs. evaluation Metrics.

Grid Search Cross Validation (CV) Algorithm, implements a "fit" and a "score" method. It also implements estimator prediction, as a result of OP-DNN classifier crops to best parameter combination of batch size: 50, Epochs: 50 and optimizer: Adam. The parameters of the estimator used to apply prediction methods are optimized by cross-validated grid-search over a parameter grid. Comparison of Op-DNN Classifier results vs DNN Classifier results for trained data are shown in Table 1 and Figure 4.

Table 1. Op-DNN Classifier results vs DNN Classifier results for trained data.

'Adam'	**batch_size = 50 epochs = 50**	
Train data =2452		
	Op-DNN Classifier	***DNN Classifier***
units/step	60us/step	116us/step
loss	0.2569	0.3389
accuracy	0.91353	0.87645
f2_score	0.8503	0.8552
precision	0.9702	0.92653
recall	0.93442	0.919032

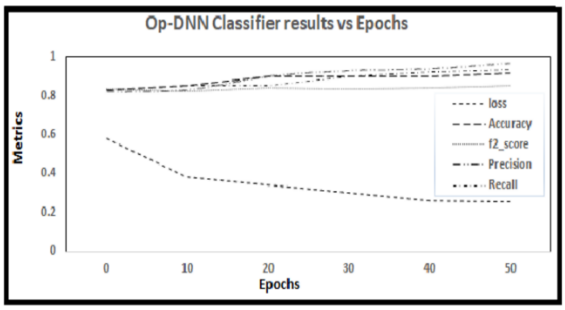

Figure 4. Graph of Op-DNN Classifier results vs. epochs for each iteration

Stating to the conclusion for the trained data, novel Op-DNN model is best when compared to the regular DNN model with the reduced error loss and accuracy. Evaluation and prediction of the results need to be done for the test data set which is of size 613 in step 4.

4) Evaluation and prediction: Novel Op-DNN model is set to be identified by the performance level calculated by different types of evaluation metrics as shown below.

a) *Confusion Matrix:* Provides a matrix as output and defines the thorough performance on testing model on 613 of dataset count showing the predicted class that the lung cancer patient will survive more than one year or not verses to the actual class is shown in Table 2.

Table 2. Confusion matrix of the test data obtained by Novel Op-DNN classifier model

Test data count =613	**Predicted class= dead or alive**		
Actual class dead or alive		***Class = Yes***	***Class = No***
	Class = yes	TP=454	FN=83
	Class = No	FP=36	TN=40

Where FP=False Positives, TP =True positive, FN=False Negatives, TN=True Negatives.

Based on the confusion matrix, metric scores are calculated leading to the performance of the Novel OP-DNN classifier model. The following are the evaluation metrics considered.

Accuracy, Sensitivity, Specificity, Precision, False Positive Rate (FPR), F2-Score

The survivability prediction and recommendation of the treatment to the patients is the main objective of this research work. To attain the objective, the task is to know the efficiency of predicting lung cancer survivability with application of NBs method. The different algorithms performance like OP-DNN, SVM-Linear, Naive Bayes (NBs) and classification trees (C4.5) are compared to know the precision, accuracy and the AUC.

Table 3. describes that the accuracy and precision provide is high (91.35% and 97.02 %) for the proposed novel Op-DNN classifier model when compared with the existing methods of DNN, SVM- Linear, C4.5, NBs.

Table 3. Lists the performance of algorithms

	Accuracy	Precision	Recall	specificity	FPR	fmeasure
OP-DNN	0.91353	0.970204	0.934424	0.709096	0.290913	0.951504
DNN	0.87645	0.926532	0.919032	0.697434	0.302521	0.922816
C4.5	0.72727	0.827273	0.689394	0.784091	0.215909	0.752063
NBs	0.56181	0.661818	0.551515	0.526971	0.422727	0.601612
SVM-linear	0.42830	0.549091	0.457576	0.250825	0.620001	0.499175

The performance evaluation graphs for the proposed and existing methods are shown in Figure 5.

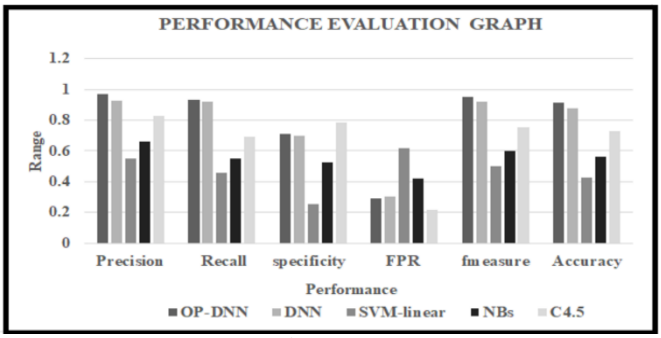

Figure 5. Performance evaluation graph

D. *Results and discussion of Novel Op-DNN Classifiaction technique*

The novel Op-DNN classifier model is built using jupyter notebook with skit learn, keras and Tensorflow. Performance evaluation is measured with output as "Less" i.e., the patient will survive for 1 month to 1-year range or "More" i.e., the patient will survive for more than 1 year. The plan for treatment action for the lung cancer patient will be prepared consequently by the referring pathologist. Op-DNN classifier survivability prediction outputs are shown in Table 4 and Table 5 shows the lung cancer survivability status prediction considering only patients who are alive

Table 4. Op-DNN Classifier Survivability Prediction Output

SMOKER					
Lung Cancer Stage		***2nd stage***		***3rd stage***	
Gender		***Male***	***Female***	***Male***	***Female***
NO DIABETES	Survived Count	98	34	154	40
	Dead Count	0	0	0	0
T1D	Survived Count	257	65	427	205
	Dead Count	0	10	10	0
T2D	Survived Count	79	61	130	52
	Dead Count	22	11	80	27
NON- SMOKER					
Lung Cancer Stage		***2nd stage***		***3rd stage***	
Gender		***Male***	***Female***	***Male***	***Female***
NO DIABETES	Survived Count	8	15	50	24
	Dead Count	0	0	0	0
T1D	Survived Count	257	57	278	141
	Dead Count	0	11	0	8
T2D	Survived Count	90	31	141	88
	Dead Count	10	16	46	32

Table 5. Op-DNN Classifier lung cancer Survivability status prediction

Study id	Gender	Age	smoker	Tumor location	T stage	N stage	Stage	Timing	Diabetes	Status	Meal cal	Wt loss	Ph ecog	ph_ karno	pat_ karno	Survived Status Prediction
2	2	66	1	1	4	3	2	3	2	2	1225	15	0	90	90	1
3	1	58	2	4	4	1	2	3	2	1	423	15	0	90	90	1
5	1	59	2	6	2	3	1	3	2	2	654	0	0	100	90	1
6	1	49	1	1	3	3	1	3	2	1	513	0	1	50	80	1
7	2	65	2	6	2	3	1	3	2	2	384	10	2	70	60	1

Output

E. *Op-DNN Regression*

A large set of labeled data are trained by using Deep learning models based on neural network architectures which learns features right from the data without the need for manual feature extraction. With this regard out of 3065 lung cancer data set, the patients who have been survived i.e. alive has to be considered. From the data set 2773 patient's data set count founded with survived status as alive and 292 patients having the survival status dead, were found out. Considering only the survived patients data set, the train data and test data are considered in the ratio of 80% and 20%, so that, train data count is 2218 and test data count is 555.the following are the steps considered for deep neural regression

The proposed research work helps to find out best Optimal-DNN (Op-DNN) regression model by evaluation and performance of Wider and Deeper Regression-Deep Neural Network (WDR-DNN) model. WDR-DNN model is built adding a greater number of hidden layers, and also by doubling the number of neurons in the hidden layers as shown in the Figure 6.

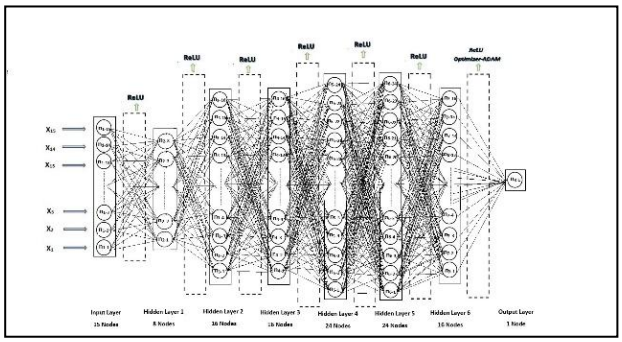

Figure 6. WDR-DNN model

a) The input data consisting of 15 lung cancer attributes textual data is fed to 8 nodes of dense layer, then passed on to 6 hidden layers, where first hidden layer consists till sixth hidden layers consists of 8,16,16,24,24 and 16 units, with one output node at end for prediction, comprising a total of 128 nodes, 1928 edges, having the Dropout rate as 0.2.
b) *Optimizer***:** In this research work to minimize loss Adaptive Moment Estimation (Adam) one of the Gradient descent optimization algorithm is used, which computes adaptive learning rates for each parameter in the DNN.
c) After building, compiling and training the WDR-DNN model. Keras scikit-learn metrics evaluation helps to get Metric values that are recorded at the end of each epoch on the training dataset [26]. The metric evaluation results show that MSE Score and MAE Score is less, increase in r2 Score as shown in the Table 6, when compared with the previous three Regression based DNN models. This indicates that WDR-DNN model is best optimal fine-tuned model built for the lung cancer survivability prediction in Months using regression approach.

Table 6. Metric Evaluation for WDR-DNN model vs Normal ANN

Metric Evaluation of Different Regression models			
Error Loss Regression Models	***MSE Score***	***MAE Score***	***r^2 Score***
ANN	15.84215478	125.57	0.72
WDR-DNN model	2.92077542	81.20	0.81

d) This predictor engine uses the optimal Wider and Deeper regression Deep neural network WDR-DNN model built from the previous step. The fine-tuned WDR-DNN model undergoes training, so it is set to make predictions on the test data as well as the real time data. i.e. how many months the lung cancer patient will survived, a random sample is extracted from predictor engine sample output is shown in the Table 7.

IV. RESULTS AND DISCUSSION OF NOVEL OP-DNN REGRESSION TECHNIQUE

The output taken from the op-DNN classifier model which predicted the how many of lung cancer patients with the T1-diabetes and Type 2 diabetes will be survived or dead. The WDR-DNN model takes Op-DNN classifier inputs of only survived lung cancer patient's data.

The predictor engine model i.e. WDR-DNN model developed based on regression technique known as Optimized-Deep Neural Network Regressor (Op-DNN Regressor) with minimum error loss having greater accuracy based in the r^2 score helps the doctors to predict survived lung cancer patient's survivability in terms of Model months, so that further treatment of chemotherapy can be recommended.

Table 7. Predicter model of op-DNN Regression sample output

Study Id	**Actual survived month**	**Predicted survived months**
900	6.47	6.01
1567	4.21	4.00
665	11.28	12.06
2301	8.41	8.91
209	1.88	1.03
2052	3.22	3.01
1148	9.83	9.83

model engine, Evaluate and performance engine and predictor engine which undergoes the following steps in each phase.

V. CONCLUSION AND FUTURE SCOPE

In this paper we proposed an efficient methodology which combines the optimized classification and regression approach based on batch size, epoch, activation and optimizer along with the DNN concept to find the lung cancer survivability prediction, dependent upon three conditions diabetes, smoker and gender. The new methodology architecture resemble the DNN architecture but with optimized conditions and takes a suitable time of processing for large textual data, a new Op-DNN framework is proposed but with optimized conditions which takes suitable time to process large textual data. In addition, using the Op-DNN classification and regression shows high prediction rate compared to other Machine learning regression techniques.

REFERENCES

[1] K, Saravanan, and Sasithra S. "Review on Classification Based on Artificial Neural Networks.", The International Journal of Ambient Systems and Applications, vol. 2, no. 4, 2014, pp. 11–18.
[2] Jiang, Fei, et al. "Artificial Intelligence in Healthcare: Past, Present and Future." , Stroke and Vascular Neurology, vol. 2, no. 4, 2017, pp. 230–243.

[3] Hughes, Mark, et al. "Medical text classification using convolutional neural networks.", Stud Health Technol Inform 235 2017, pp. 246-250.
[4] Kollias, Dimitrios, et al. “Deep Neural Architectures for Prediction in Healthcare”, Complex & Intelligent Systems, vol. 4, no. 2, 2017, pp. 119–131.
[5] Chon, Albert, Niranjan Balachandar, and Peter Lu. "Deep convolutional neural networks for lung cancer detection." , Tech. rep., Stanford University,2017.
[6] Sarwar, Abid, and Vinod Sharma. “Comparative Analysis of Machine Learning Techniques in Prognosis of Type II Diabetes.” Ai & Society, vol. 29, no. 1, 2013, pp. 123–129.,
[7] Rodrigo, Hansapani, and Chris P. Tsokos. “Artificial Neural Network Model for Predicting Lung Cancer Survival” , Journal of Data Analysis and Information Processing, vol. 05, no. 01, 2017, pp. 33–47.
[8] Kuan, et al. “Deep Learning for Lung Cancer Detection: Tackling the Kaggle Data Science Bowl 2017 Challenge.”, ArXiv.org, 26 May 2017
[9] Schmidhuber, and Juergen. “Deep Learning in Neural Networks: An Overview.” ArXiv.org, 8 Oct. 2014
[10] Agrawal, Shikha, and Jitendra Agrawal. “Neural Network Techniques for Cancer Prediction: A Survey.” Procedia Computer Science, vol. 60, 2015, pp. 769–774.
[11] Burt, Jeremy R., et al. "Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks.", The British journal of radiology 91.1089 ,2018, p.20170545.
[12] Bengio, Yoshua, and Yann LeCun. "Scaling learning algorithms towards AI.", Large-scale kernel machines vol.34, no.05, 2007, pp.1-41.
[13] Huerta, E. A., et al. "Real-time regression analysis with deep convolutional neural networks.", arXiv, 2018.
[14] S. Yan, H. Wang, X. Tang, and T. S. Huang, “Learning auto structured regressor from uncertain non negative labels,” in CVPR, 2007, pp. 1–8.
[15] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." ,In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.
[16] Stephane Lathuiliere, Pablo Mesejo, Xavier Alameda-Pineda, Member IEEE, and Radu Horaud, “A Comprehensive Analysis of Deep Regression”, arxiv:1803.0845. vol, 22 Mar 2018.
[17] Bae, K.T., Kim, C.J., “An Agricultural Estimate Price Model of Artificial Neural Network by Optimizing Hidden Layer”, Journal of Intelligent Information Systems, vol,12, 2016, pp.161-169.
[18] Ki-Young Lee[1], Kyu-Ho Kim[1,*], Jeong-Jin Kang[2], Sung-Jai Choi[3], Yong-Soon Im[4], Young-Dae Lee[5], Yun-Sik Lim[6], “Comparison and Analysis of Linear Regression & Artificial Neural Network”, International Journal of Applied Engineering Research ISSN 0973-4562, Vol 12, no. 20 , 2017, pp. 9820-9825.
[19] Loprinzi, C., Laurie, J., Wieand, H., Krook, J., Novotny, P., & Kugler, J.,” Prospective evaluation of prognostic variables from patient-completed questionnaires”, North Central Cancer Treatment Group. Journal of Clinical Oncology, Vol 12, no. 03, 1994, pp. 601-607.
[20] Pradeep KR, Naveen NC.,” Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4. 5 and Naive Bayes Algorithms for Healthcare Analytics”, Procedia computer science,132, 31 Dec 2018, pp.412-420.
[21] Cox, Victoria. “Exploratory Data Analysis.” SpringerLink, Apress, Berkeley, CA, 2017.
[22] Ketkar N,” Introduction to Keras. In: Deep Learning with Python. Apress”, Berkeley, CA, 2017.
[23] Schmidt-Hieber, J., “Nonparametric regression using deep neural networks with ReLU activation function”, arXiv,2018
[24] Bahar, Parnia, et al. “Empirical Investigation of Optimization Algorithms in Neural Machine Translation.” The Prague Bulletin of Mathematical Linguistics, vol. 108, no. 1, 2017, pp. 13–25.
[25] Kingma, Diederik, and Jimmy Ba, "Adam: a method for stochastic optimization”, arXiv 2015.
[26] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1, 2014, pp. 1929-1958.

Authors Profile

Mr. Pradeep K R pursed Bachelor of Engineering from Vivesvaraya Technological University, Belagavi, Karnataka in 2006 and Master of Technology from Vivesvaraya Technological University, Belagavi, Karnataka in year 2010. He is currently pursuing Ph.D. in Vivesvaraya Technological University, Belagavi, Karnataka and currently working as Assistant Professor in Department of Computer Science and engineering in K.S.Institutte of technology ,Bengaluru,karnataka since 2012. He is a member of ISTE & MIE since 2011. He has published more than 10 research papers in reputed international journals (SCI) and conferences including IEEE and it’s also available online. His main research work focuses on Big Data analytics, Data Mining, Machine learning and artificial intelligence. He has 9 years of teaching experience and 3 years of Research Experience.

Dr. Naveen N C pursed Bachelor of Engineering from Bangalore University, Bengaluru , Karnataka in 1994 and Master of Engineering from Bangalore University, Bengaluru Karnataka in year 1999. He has completed Ph.D. from SRM University,Chennai in 2013 and currently working as Professor and HOD,Department of CSE, JSS Academy of Technical Education, Bengaluru. since 2018. He is a member of IEEE & IEEE computer society since 2013,. He has published more than 35 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it’s also available online. His main research work focuses on Big data Analytics, Image processing, Data mining, Machine Learning and artificial intelligence.. He has 20 years of teaching experience and 4 years of Research Experience.

