SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

LUNG DISEASE PREDICTION WITH SPARK FRAMEWORK USING OPTIMAL MODULAR NEURAL NETWORK IN BIG DATA

V Durga Devi
Research Scholar, Department of Computer Applications, VISTAS, Pallavaram,
vdurga2216@gmail.com

R Priya
Professor, Department of Computer Applications, VISTAS, Pallavaram,
priyaa.research@gmail.com

Abstract

The advancements in big data analytics helps healthcare professional in identifying high-risk patients and informed them the status of disease to the patients on time. Moreover, doctors can provide efficient treatment and the cost of extending improved care is also reduced. The major cause for chronic lung disease is smoking. Tobacco smoke either in the form of solid, liquid or gas contains particles comprising of thousands of chemical components which include several toxins and carcinogens. In big data, Artificial Intelligence (AI) is a promising tool which helps in predicting the disease. Under the artificial intelligence the role of Convolution Neural Network (CNN) has many benefits; thus allied in small as well as large scale datasets. However, it is computationally more expensive while the original data is trained which also consumes more time for a complex model even when most powerful GPU hardware is employed. To overcome this issue, this paper proposes MobileNet V2 with Modular Neural Network (MobileNet V2-MNN) for analyzing big data. This model comprises of feature selection and classification stages. During feature selection, the significant features are selected using grasshopper optimization algorithm which consumes very less time. During classification, Neural Network is constructed which uses these selected features for classifying patients resulting in improved accuracy while detecting lung disease; moreover false positive rate is also minimized. This goal is consistently achieved and thus the proposed model outperforms few standard approaches which are considered like DeepRisk, Optimized Dual Attention Neural Network (ODANN), Recurrent Convolutional Neural Network (RCNN). It is observed that the proposed MobileNet V2-MNN achieves 97.1% of accuracy, 96.2% of precision, 89.8% of recall,84.82% of F1-score, 54.66% of AUC and 45.92% of ROC.

Keywords-Big data, lung disease, neural network, preprocessing, spark, optimization, classification.

Introduction

The development in Information Technology has increasingly improved the quality of living in several ways and medical science is not an exception to this persistent advancements. The medical science and profession of a doctor are adopted by Internet and telemedicine but Artificial Intelligence (AI) cannot cope up with the intelligence of an experienced Doctor [1]. Data Analytics has proved its excellence as a high reliable source of information in science,

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438
and as every data is treated identical by AI, particularly diagnosing health conditions and epidemics using Big Data is important and made possible. According to this, several research works are carried out in finding appropriate data mining technique to use Big Data for predicting diseases [2]. It is believed that analyzing larger datasets will find the causes leading to lung disease and allergies in future.

Respiratory disease, termed as lung disease, is due to the issues on the airways and lung structures [3]. Pneumonia, tuberculosis and COVID-19 (Coronavirus 2019) are few examples of lung disease. As stated by International Respiratory Societies nearly 334 million individuals are affected due to asthma. Moreover, 1.4 million and 1.6 million people die dur to tuberculosis and lung cancer every year. Following which pneumonia also takes aways he lives of million people. The pandemic situation across the world due to COVID-19 has affected several million people which is a terrible burden for healthcare systems [4]. Obviously, lung disease is one which causes death and disabilities all over the world. Recovery from lung disease and improving survival rates can be achieved only if the disease is detected at their early stages. Generally, lung disease are detected by performing tests in skin, blood and sputum samples, and moreover chest X-ray and computed tomography (CT) scan are also examined. In the recent years, deep learning approaches have proved its excellence and potentially detect diseases from medical data. Deep learning, a class of machine learning, supports in identifying, quantifying and classifying patterns from clinical data [5]. This is possible as deep learning techniques are capable of merely learning features from data. Due to the improved performance of deep learning approaches, these are considered as standard techniques in several medical applications. The advancements in these techniques help physicians to efficiently detect and classify medical data and conditions. Moreover, a big data processing system Hadoop along with machine learning platform Mahout is the suitable choice to work on large scale datasets. An emerging Apache Spark, due to it nature of processing in-memory, is considered as the second generation processing engine for big data [6], as it is much faster than Hadoop, particularly while performing iterative operations in machine learning applications. Spark MLlib and Spark Streaming used correspondingly for machine learning and handling data streams are the libraries integrated in Spark [7,8].

People in this modern world are very much concerned about their health. We have witnessed advancements in healthcare applications are witnessed not only diagnosing diseases or providing treatment, but even in the prediction of possibility of diseases and its risks or earlier detection of the presence of disease using the medical data from healthcare sectors [9]. There is a difference between electronic medical records and data obtained from wearable health sensors. Devices used to monitor blood sugar or blood pressure which were available only in laboratories are now made available at homes. With this motivation, the advanced deep learning approach is developed to predict the status of lung disease with reasonable accuracy. The contribution of this work are as follows,

To construct MobileNet V2 with Modular Neural Network (MobileNet V2-MNN) for enhancing the classification accuracy by maintaining the previous timestamp data. Moreover, the information pertaining to current state obtained from weight optimization can make the

DOI: 10.5281/zenodo.7157438

model robust.

The organisation of this paper is: section 1 describes the background of big data application for classifying the disease, role of deep learning in big data in attacks, motivation and contribution of the work. In section 2, the literature on various techniques disease classification in big data are discussed. Section 3 explains the proposed MobileNet V2-MNN for effective classification of disease In section 4, the experimental analysis is given with graphs by comparing with three standard methods. Finally, the paper ends with section 5 presenting conclusion and future work.

Related works

In [10], an DeepRisk end-to-end model was developed for risk prediction of cardiovascular diseases based on attention and deep neural network (DNN) mechanisms which automatically learnt the high-quality features from electronic health records (EHRs) as well as integrated heterogeneous and time-ordered medical data efficiently. In [11], the focus was on mining and analyzing information from large scale medical sports data. Improved deep learning approach based on convolutional neural network (CNN), resampling technique with self-adjusting function and tensor convolution self-coding mechanism were involved to effectively produce prediction accuracy and risk assessment results for sports related diseases. Ural network model was developed which helped in analyzing multi-dimensional sports medical data. At last, for constructing an intelligent medical data system for sports related data, a novel cloud-based hardware-in-the-loop simulation system was designed. In [12], a hybrid deep-learning Optimized Dual Attention Neural Network (ODANN), integrated with data assimilation and natural language processing (NLP) feature extraction technique was developed to process large volumes COVID-19 records. The overall performance of ODANN summarized its competitive nature in predicting the increasing rate of COVID-19 across the world. In [13], a new Recurrent CNN (RCNN)-based multimodel for disease risk assessment was constructed which processed structured as well as unstructured text data from clinical sectors. Here, convolutional layer became as a bidirectional RNN as intra-layer recurrent connection was utilized. In the convolutional layer, every neuron received recurrent and feedforward inputs from the neighbor and previous units respectively. Moreover, the region of context capturing operation besides step-by-step recurrent operation increased the efficiency of generating fine-grained features. In [14], the architecture of recursive CNN was analyzed and their weights were connected with the layers to assess the contribution of feature maps, number of layers, and parameters. In [15], developed a novel Network in Network based on deep network where micro neural networks were used rather than linear filter. Moreover, non-linear activation was employed in receptive field and the classification ability of the model was improved. In [16], context-DNN model was designed using multiple-regression to predict depression risk. The input for DNN was the context information related to predictor depression variables and the output was the variable for depression prediction. The regression analysis is used for depression risk which was potential in prediction.

From the above mentioned methods it is noted that the models involved variety of errors due to classifiers. The classifiers involved must exhibit low correlation which ensures that the errors

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

of these classifiers vary. Moreover, the classifiers are complement to each other producing better classification with identical features. If so, then the correlation error will be high. To overcome this problem, this paper concentrates on developing MobileNet V2-MNN for better classification.

System model

The overall architecture for lung disease classification in big data is as shown in figure-1. It consists of three layers such as big data layer, data processing layer and deep learning layer. Under the big data layer all the databases are available with the specifications about lung disease. The data processing layer helps to access the data from database and hence spark streaming based preprocessing is done followed by feature selection. In the third layer, the deep learning process done by getting the optimal features which tends in classification of lung disease

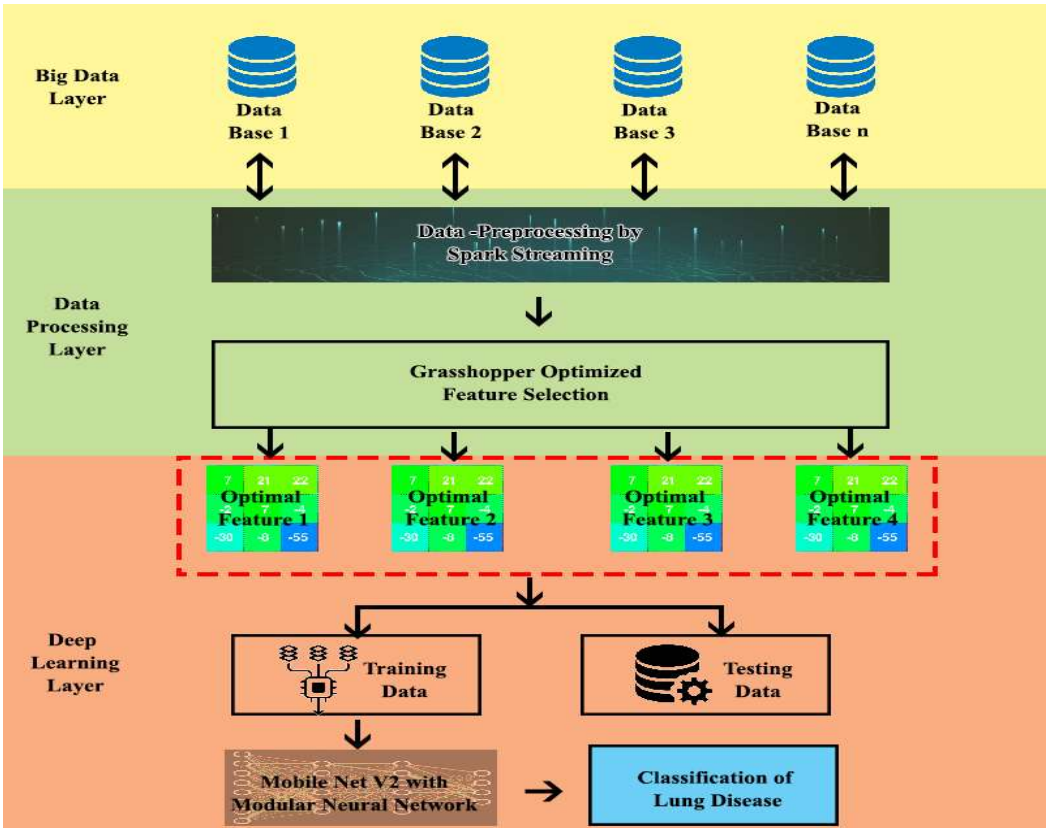


Figure-1- System model for lung disease classification in big data

Data acquisition

Kaggle data from Lung Disease labeled Dataset of UCI Artificial Intelligence repository is used for experiment to train and test the developed deep learning algorithm for lung disease prediction. This dataset is mainly used as it is reliable, publicly available and are used in several research works based on deep learning techniques. This dataset contains 14 attributes namely sex, age, type of pain, fasting blood sugar, blood pressure, serum cholesterol, resting ECG result, breathing trouble, wheezing, cough, mucus production, hypoxemia, swallowing and class label. Among these, class label with a value 0 indicates the absence of the disease and

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

values 1 to 4 indicate its presence. But this works considers only values 0 and 1 indicating the absence and presence of disease respectively. Values 2 to 4 are replaced by 1; thus this dataset is turned as binary dataset.

Preprocessing by spark streaming

Before analyzing and detecting lung disease, data are preprocessed in order to extract more reliable information. This preprocessing undergoes many processes and the output of one stage is the input for the next one and the output is used for further processing and classification. An online pipeline has been implemented for preprocessing stage on Apache Spark platform using Pandas-UDF mechanism in Spark where the conventional data mining tasks are executed. At last, the pipeline is constructed with the use of spark structured stream processing engine which executes test data. For the given large dataset having N records and K local learners, initially in Adaptive Data Slicing, global data slicing is performed; thus global data slice comprising of N/K rows are uniformly allocated among the available local learners thereby load is balanced and learning process is consequent. Then, the apt size of every local learner is obtained by computing Appropriate Learning Size (ALS) value. For efficient learning, on every local learner, global data slice is split to several local slices of ALS. For K Local Learners, Nd data slices is sent to every local learner. Nd is estimated using Equation (1):

$$\text{Nd} = \frac{N}{K*ALS} \quad (1)$$

Grasshopper Optimized Feature Selection (GOFS)

GOFS initially generates random population having significant number of particles or individuals obtaining potential results for optimization. For every particle which is represented by a vector, the fitness value is calculated using Fitness Function (FF) [17]. In (*xa, pbesta, va*), *xa* and *pbesta* speciffy the position and personal optimum position for particle *a* respectively and *va* is its velocity. In every iteration of GOFS, *va* and *xa* change based on the Eq. (2) as given below:

$$\text{xa (z+1)} = \text{f(x)} = \begin{cases} 0 & if\ rand \geq sig\ (va) \\ 1 & otherwise \end{cases} \quad (2)$$

Where, $sig\ (va)$ represents the sigmoid transfer function which specifies the possibility where bit *b* ranges from 0 or 1 as indicated in equation (3):

$$Xi = Si + Gi + Ai \quad (3)$$

Where Xi, Si and G specify the location, social interaction among grasshoppers and its gravitational force respectively. A specifies the wind direction. Using equation(4), the random behavior is created where r varies at random between 0 and 1.

$$Xi = r1Si + r2Gi + r3\ Ai \quad (4)$$

Si is computes as indicated by equation (5), where d_{ij} is the distance between grasshoppers i and j which is computed as $d_{ij} = |Xj - Xi|$.

$$\text{Si} = \sum_{j=1}^{N} s\ (dij) \quad (5)$$

S is a mapping function for d_{ij} which is obtained using equation (6):

$$\text{s (r)} = \text{fe-e}^{-r} \quad (6)$$

here f is the gravitational intensity. The general formula is as in equation (7):

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

$Xi = \sum_{j=1}^{N} s(xj - xi).\frac{xj-xi}{dij}$-geg'+uew' (7)

Here, N represents the number of grasshoppers. As grasshoppers move on the ground, their location has to be less than the threshold. Hence, equation(7) is modified as represented in equation (8):

$Xi(d) = c(\sum_{j=1}^{n} c\frac{ubd-lbd}{2}\ s\left(xj(d) - xi(d)\right)\frac{xj-xi}{dij} + T(d)$ (8)

The grasshopper's location is specified according to its current location, target location, and location of every other grasshopper. In equation (8), reduction factor C is considered as one important parameter of grasshopper optimization technique, which affects the safe, repulsion and gravity regions. This parameter is updated using equation (9)

$C=c(max)\text{-}l\frac{c(\max-\min T)}{L}$ (9)

The two particles are neighbors if a link exists between them. For example, *link*(*xa,xb*) indicates that two particles *xa* and *xb* are neighbors [17]. The neighbors of every particles is represented by matrix *M* with order *N*N* where *N* is the total particles in the population. The value 0 and 1 of *M* (*a, b*) indicates that the particles *xa* and *xb* are neighbors or not. Link function is expressed by multiplying *'a'* and *M* with its column *b* as indicated by equation (10):

$LINK\ (xa,xb) = \sum_{k=1}^{n} M(a,k) * M(k,b)$ (10)

the rank values of every pair of particles is additionally computed using *link*(*xa, xb*). The particle with more neighbors has higher rank values else has only low values. The particle with several positions having *dim* dimension is represented as a value 0 or 1. Every dimension deal with only one feature and the particle *X* having *dim* (which is 9 here) feature is declared with a value 0 or 1. When b is the value at position, then feature *b* has been selected else not. Mean Absolute Difference (MAD) is the objective function of the proposed technique which is as indicated by equations (11) and (12):

$MAD\ (Sa) = \frac{1}{Sa}\sum_{k=1}^{b} x(a,k - xa')$ (11)

Where,

$Xa' = \frac{1}{Sa}\sum_{k=1}^{b} x(a,k)$ (12)

here *sa* represents the total selected features, *xa'* indicates the mean of the vector and the weights of feature *k* are given by *a* and *b.*

Grasshopper Optimized Feature Selection algorithm

Input-preprocessed data (p)
Output- Feature selected data (F)
Initialization of population ={p1,p2,p3….pn}
Particles deployment ={pa1,pa2,pa3,….pan}
Velocity of grasshopper particle (a)
Vector (a)←[v(p1),v(p2),….v(pn)]
Best value← *pbesta,bbesta*
Iteration (I) start
K=I+1
Update the position (pos)

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

$Pos \leftarrow K$

Update the gravity (Xi)

$Xi = \sum_{j=1}^{N} s(xj - xi).\frac{xj-}{dij}$-geg'+uew'

Compute the link function (L)

$L(a,b) \leftarrow Xi$

MobileNet V2 with Modular Neural Network (MobileNet V2-MNN) based classification

Fm × Fm is the size of the feature vector map and Fs × Fs is the filter size. p is the input variable where q represent the output variable. Figure-2 illustrates the architecture of the proposed MobileNet V2-MNN model developed for classification.

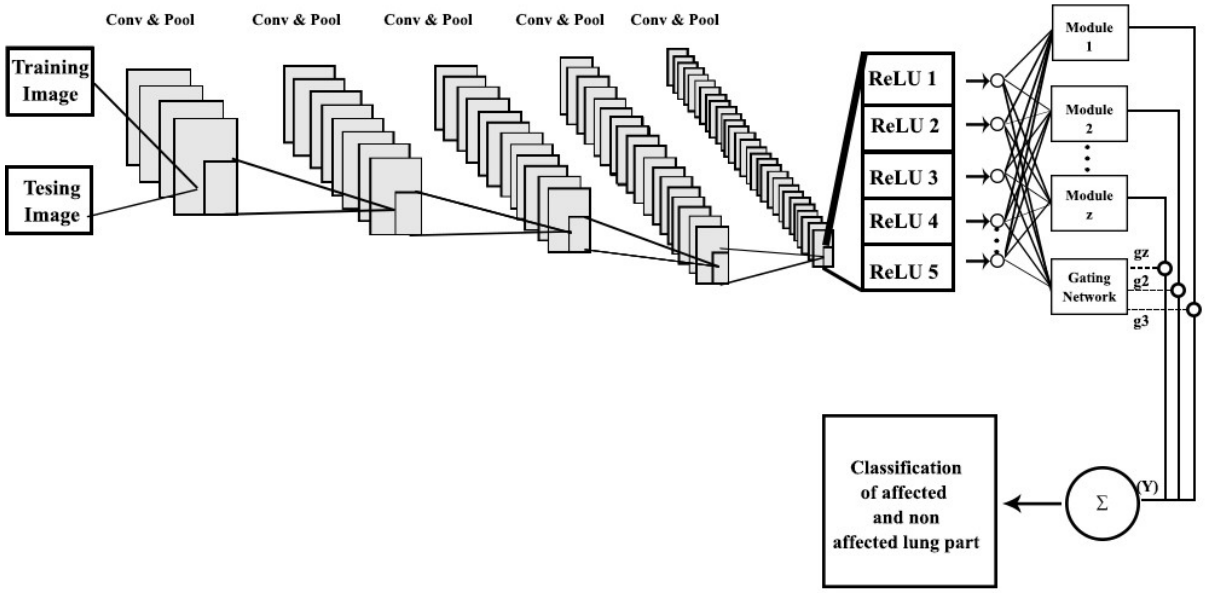


Figure-2-Architecture of MobileNet V2-MNN for classification

In the architecture, the overall computation efforts for core abstract layers are denoted using the variable ce which is estimated using Equation (13):

$$ce = Fs.\ Fs.w.\ \alpha Fm\ .\ \alpha Fm + w.\Omega.\alpha Fm.\ \alpha Fm \quad (13)$$

The value used to multiply w is context-specific, and while classifying lung disease [18], it ranges between 1 and n. The value of the variable resolution multiplier a is considered as 1. The variable cost helps to recognize the computational efforts and is computed using Equation (14):

$$coste = Fs.Fs.\ w.\Omega.Fm.\ Fm \quad (14)$$

The depth-wise as well as point-wise convolutions are incorporated by this proposed model which is limited to depletion variable identified by d which is approximated as in Equation (15):

$$d = \frac{Fs.Fs.w.\alpha Fm\ .\alpha Fm + w.\Omega.\alpha Fm.\alpha Fm}{Fs.Fs.w.\Omega.Fm.Fm} \quad (15)$$

The width as well as resolution multipliers are the two hyper-features which help in adjusting

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

the optimum size window for predicting accurately on the basis of context. The size of the input data in the proposed model is 224 × 224 × 3 where the first two indicate the in-built attributes of data and which are always higher than 32 while the third one indicates that the model has 3 input channels. For the training set with *p* pairs represented as (**x**1, **y**1) . . . , (**x***p*, **y***p*), the entire network is constructed using many sub-networks [19] where each receives the input **x** generating y as output with probability as indicated in equation (16):

$$P\ (Y/X,\mu i) \tag{16}$$

where μi specifies the parameter vector of sub-network *i*. The output of every sub-network is weighted by a gating subnetwork which examines **x** producing *m* expert sub-networks with set of gating values *g*1, *g*2, . . . , *gm*. At this moment, *F*1 Layer takes binary input vectors all having threshold 2 then is passed to F2 layer. Then, the attention unit is turned off only by the *F*2 layer. In *F*2 layer, when a unit is fired, the attention unit is turned off by the negative weight. Moreover, in *F*2 layer, the winning unit returns 1 via through the link between *F*1 and *F*2 layer. Now, in *F*1 layer, every unit takes the respective input vector **x** and weight vector **w** as input. The unit *I* in *F*1 layers compares *xi* and *wi* where the output obtained is *xiwi*. This information along with the elements of **x** weighted by ρ is sent to the reset unit where ρ is the attention parameter. It is computed as indicated in equation (17):

$$\beta \sum_{i=1}^{n} xi - X.W \geq 0 \tag{17}$$

This corresponds to the test as indicated in equation (18):

$$\frac{x.w}{\sum_{i=1}^{n} x(i)} \tag{18}$$

When the input is only out of attention cone of the winning unit, the reset unit is fired. F2 layer receives the reset signal whereas only the winning unit is reserved. Following this, the attention unit is then activated and then new calculations are performed. the nexdt step during training is the supervised adjustment of *zi* values in the presence of input vector **x,** when the hidden unit *i* fires, approximation quadratic error is given as indicated in equation (19):

$$E = \frac{1}{2}(zi\text{-}f(x))^2 \tag{19}$$

It has to be noted that every output of hidden unit is normalized where the output is divided by the sum of the output of all other hidden units. At last, fine tuning operation is performed based on BP model. Classification of clinical data comprises of two stages. Initially, the output layer is set at the top of DNN. Then, DL model uses *N* input neurons and triple hidden layers. In training phase, weight is enhanced and BP model is initialized with weights obtained during pre-training. A low error value is computed as a consequence of leveraging,; but high classification accuracy is obtained as optimized weights are used which is as indicated by equation (20):

$$F\ (v,h) = -\sum_{k-1}^{K}\sum_{l-1}^{L} Wkl.vk.hk - \sum_{k-1}^{K} ak.vk - \sum_{i-1}^{L} \beta l.hl \tag{20}$$

here, Wkl indicates the interaction between visible unit vk and hidden unit hl. The bias terms are given by α and β and the number of visible and hidden units is represent by K and L respectively. For a preparation vector, the corresponding log likelihood reminds the effects of inconsistency.

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

Performance analysis

The experimental result is using the parameters such as, accuracy, precision, recall, F1-score, AUC and ROC. These parameters are compared with three state of art methods such as DeepRisk, Optimized Dual Attention Neural Network (ODANN), Recurrent Convolutional Neural Network (RCNN) with the proposed MobileNet V2 with Modular Neural Network (MobileNet V2-MNN)

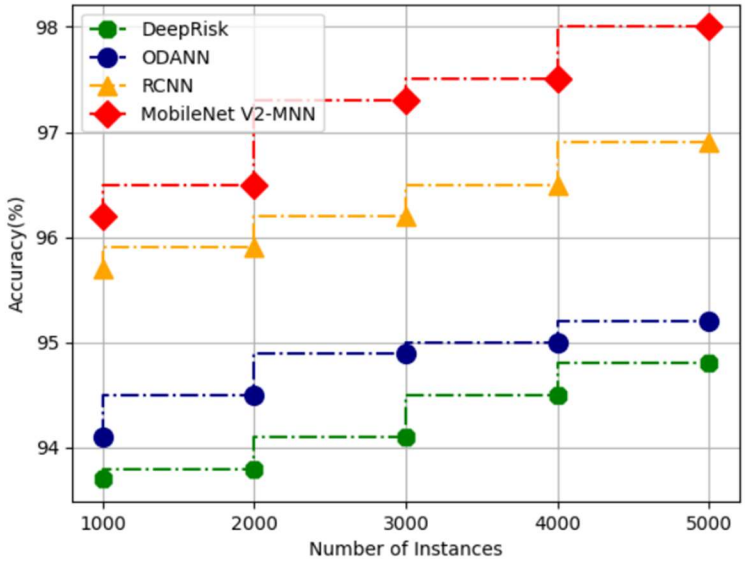
Accuracy presents the ability of the overall prediction produced by the model. True positive (TP) and true negative (TN) provides the capability of predicting the absence and presence of disease. False positive (FP) and false negative (FN) presents the false predictions made by the used model. The formula for accuracy is given as in equation (21):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{21}$$

Table-1. shows the comparison of accuracy between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method.

Table 1. Comparison for accuracy

Number of instances	DeepRisk[10]	ODANN[12]	RCNN[13]	MobileNet V2-MNN (proposed)
1000	93.7	94.1	95.7	96.2
2000	93.8	94.5	95.9	96.5
3000	94.1	94.9	96.2	97.3
4000	94.5	95.0	96.5	97.5
5000	94.8	95.2	96.9	98



SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

Figure-3 Comparison of accuracy

Figure 3 illustrates the comparative analysis of accuracy between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method where X axis shows the number of instances and y axis shows the accuracy in %. The existing DeepRisk, ODANN, RCNN methods achieve 94.18%, 94.74%and 96.24%while the proposed MobileNet V2-MNN method achieves 97.1%which is 3.08% better than DeepRisk,3.64% better than ODANN and 1.14% better than RCNN.

- **Precision** measure the success of the disease classification model. Precision is the ability of the classifier predicting positive results in the presence of disease. This is termed as true positive (TP) rate which is estimated as indicated in equation (22):

$$Precision\ (P) = \frac{TP}{TP+FP} \tag{22}$$

Table-2 shows the comparison of precision between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method.

Table 2. Comparison for precision

Number of instances	DeepRisk[10]	ODANN[12]	RCNN[13]	MobileNet V2-MNN (proposed)
1000	92.2	93.1	94.3	95.7
2000	92.6	93.4	94.5	95.8
3000	93.1	94.1	94.8	96.2
4000	93.4	94.4	95.2	96.5
5000	93.8	94.6	95.4	96.8

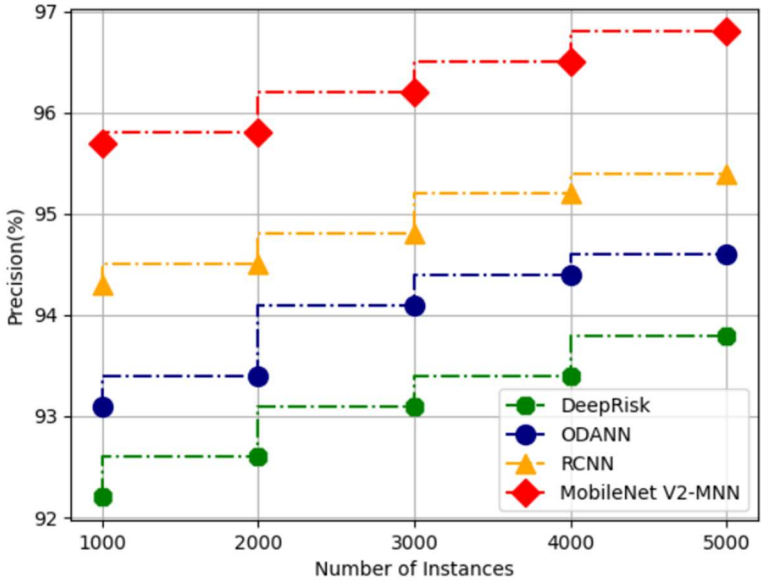


Figure-4 Comparison of precision

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

Figure 4 illustrates the comparative analysis of precision between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method where X axis shows the number of instances and y axis shows the precision in %. The existing DeepRisk, ODANN, RCNN methods achieve 93.02%, 93.92%and 94.84%while the proposed MobileNet V2-MNN method achieves 96.2%which is 3.22% better than DeepRisk,3.72% better than ODANN and 2.44% better than RCNN.

- **Recall** is the ability of the classifier to predict negative results in the absence of traffic which is also termed as true negative (TN) rate. This is estimated as indicated in equation (23):

$$Recall(R) = \frac{TP}{TP+} \quad (23)$$

Table-3 shows the comparison of recall between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method.

Table 3. Comparison for recall

Number of instances	DeepRisk[10]	ODANN[12]	RCNN[13]	MobileNet V2-MNN (proposed)
1000	84.4	86.1	87.9	89.2
2000	84.8	86.4	88.1	89.4
3000	85.4	86.9	88.6	89.8
4000	85.7	87.1	89.2	90.1
5000	85.9	87.5	89.5	90.5

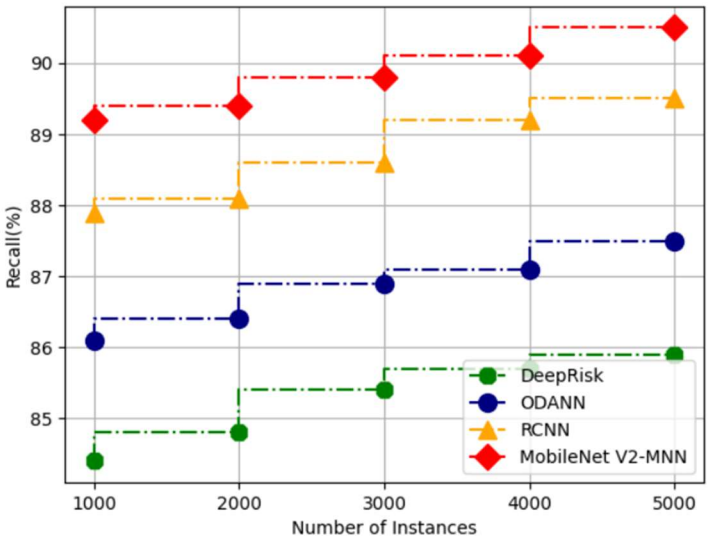


Figure-5 Comparison of recall

Figure 5 illustrates the comparative analysis of recall between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method where X axis shows the number

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

of instances and y axis shows the recall in %. The existing DeepRisk, ODANN, RCNN methods achieve 85.24%, 86.8%and 88.66%while the proposed MobileNet V2-MNN method achieves 89.8%which is 4.44% better than DeepRisk,3% better than ODANN and 1.72% better than RCNN.

F1- Score is utilized to determine the prediction performance. It is the harmonic mean of precision and recall. A value of 1 is assumed to be the best while 0 indicates the worst. F1-Score is estimated as indicated in equation (24):

$$F1 - Score = \frac{2*P*R}{P+R} \tag{24}$$

Table-4 shows the comparison of F1-Score between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method.

Table 4. Comparison for F1-Score

Number of instances	DeepRisk[10]	ODANN[12]	RCNN[13]	MobileNet V2-MNN (proposed)
1000	79	81.1	82.5	84.2
2000	79.1	81.3	82.9	84.5
3000	79.5	81.5	83.1	84.9
4000	80.4	82.5	83.4	85.1
5000	80.6	82.7	83.6	85.4

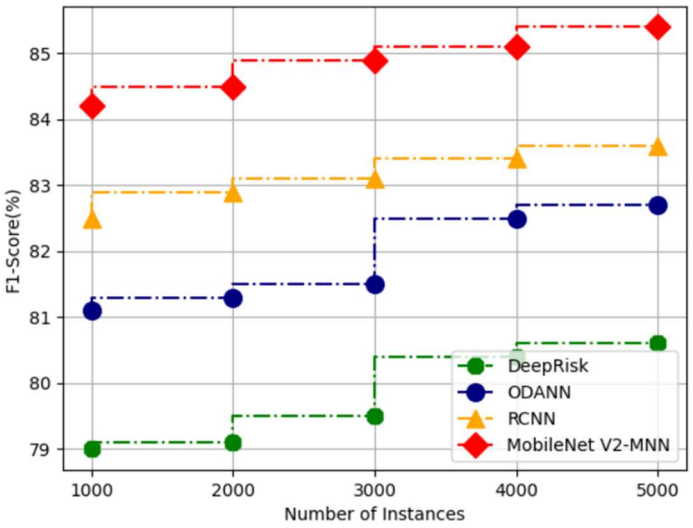


Figure-6 Comparison of F1-Score

Figure 6 illustrates the comparative analysis of F1-Score between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method where X axis shows the number of instances and y axis shows the F1-Score in %. The existing DeepRisk, ODANN, RCNN methods achieve 79.72%, 81.82%and 83.1%while the proposed MobileNet V2-MNN method

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438

achieves 84.82% which is 5.1% better than DeepRisk,3% better than ODANN and 1.72% better than RCNN.

Along with these evaluation parameters, ROC (Receiver Operating Characteristic) curve and AUC (area under curve) are also used for evaluating the performance of the classifier. ROC curve is computed by considering TPR (True Positive Rate) and FPR (False Positive Rate) which are defined as in equations (25) and (26):

$$TPR=\frac{TP}{TP+FN} \quad (25)$$

$$TFR=\frac{FP}{FP+TN} \quad (26)$$

The model is better when ROC curve is towards upper left corner. With AUC, the model is better if the area is close to 1. When processing medical data, more attention is given to recall instead of accuracy. For higher recall, risk of a patient with lung cancer is low.

Table-5 presents the comparison of AUC between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method.

Table 5. Comparison for AUC

Number of instances	DeepRisk[10]	ODANN[12]	RCNN[13]	MobileNet V2-MNN (proposed)
1000	50.1	49.9	52.4	53.9
2000	50.4	51.1	52.7	54.1
3000	50.6	51.4	52.9	54.6
4000	51.1	51.9	53.1	55.1
5000	51.3	52.1	53.6	55.6

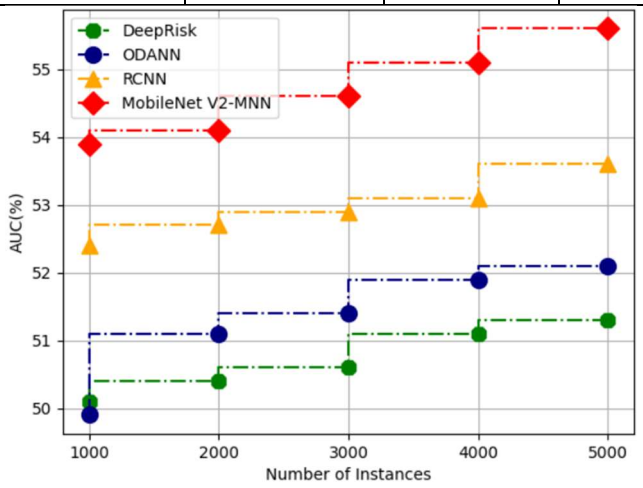


Figure-7 Comparison of AUC

The figure 7 shows the comparison of AUC between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method where X axis represents the number of instances while the y axis represents AUC in %. The existing DeepRisk, ODANN, RCNN methods achieve 50.7%, 51.28%and 52.94%while the proposed MobileNet V2-MNN method achieves 54.66%which is 4.16% better than DeepRisk,3.42% better than ODANN and 2.32%

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7157438
better than RCNN.

Table-6 shows the comparison of ROC between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method.

Table 6. Comparison for ROC

Number of instances	DeepRisk[10]	ODANN[12]	RCNN[13]	MobileNet V2-MNN (proposed)
1000	39.7	42.1	44.1	45.4
2000	40.1	42.4	44.4	45.7
3000	40.5	42.7	44.6	45.9
4000	41.1	42.8	44.9	46.1
5000	41.6	43.2	45.1	46.5

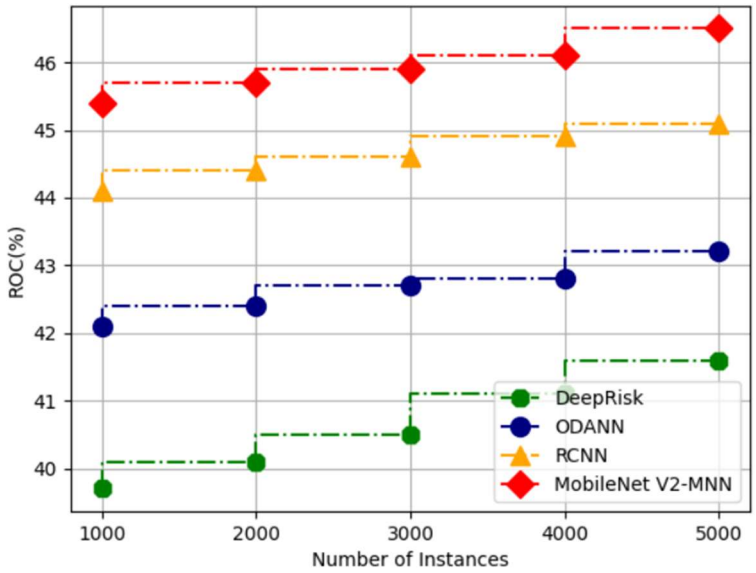


Figure-8 Comparison of ROC

The figure 8 shows the comparison of ROC between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method where X axis represents the number of instances while y axis represents ROC in %. The existing DeepRisk, ODANN, RCNN methods achieve 40.6%, 42.64%and 44.62%while the proposed MobileNet V2-MNN method achieves 45.92%which is 5.32% better than DeepRisk,3.32% better than ODANN and 1.3% better than RCNN.

Table 7 shows the overall comparative analysis between existing DeepRisk, ODANN, RCNN methods and proposed MobileNet V2-MNN method.

Table-7 Overall comparison between existing and proposed method

Parameters	DeepRisk[10]	ODANN[12]	RCNN[13]	MobileNet V2-MNN (proposed)
Accuracy(%)	94.18	94.74	96.24	97.1

DOI: 10.5281/zenodo.7157438

Precision (%)	93.02	93.92	94.84	96.2
Recall (%)	85.24	86.8	88.66	89.8
F1-Score (%)	79.72	81.82	83.1	84.82
AUC (%)	50.7	51.28	52.94	54.66
ROC (%)	40.6	42.64	44.62	45.92

Conclusion

A scalable health status prediction model for real time applications was designed, and implemented and tested on data processing layer, where deep learning technique is applied for classifying the data. It involves various stages like data acquisition, streaming echo, optimization based feature selection and classification. Grasshopper Optimization Feature Selection (GOFS) algorithm is applied thereby the convergence rate is increased. This system is developed based on MobileNet V2 and MNN approach proving the efficiency of classifying and detecting lung disease with maximum accuracy. Moreover, this is a computationally effective model and by maintaining the previous timestamp data the accuracy of prediction is improved. The model is robust as the information related to the current state is used through weight optimization. The experimental results are obtained for comparison with parameters such as accuracy, precision, recall, F1-score, AUC and ROC. It is observed that the proposed MobileNet V2-MNN model achieves 97.1% of accuracy, 96.2% of precision, 89.8% of recall,84.82% of F1-score, 54.66% of AUC and 45.92% of ROC. The future work concentrates on extend the framework by including online apache spark streaming process for data-preprocessing the social media data.

Reference

1. Hirak Kashyap, H. A. (2014). Big Data Analytics in Bioinformatics: A Machine Learning Perspective. IEEE.
2. Isaac Triguero, D. P. (2014). A MapReduce solution for prototype reduction in big data classification. IEEE.
3. Rahaman, M.M.; Li, C.; Yao, Y.; Kulwa, F.; Rahman, M.A.; Wang, Q.; Qi, S.; Kong, F.; Zhu, X.; Zhao, X. Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. J. X-Ray Sci. Technol. 2020, 28, 821–839. [CrossRef]
4. Yahiaoui, A.; Er, O.; Yumusak, N. A new method of automatic recognition for tuberculosis disease diagnosis using support vector machines. Biomed. Res. 2017, 28, 4208–4212.
5. Hu, Z.; Tang, J.; Wang, Z.; Zhang, K.; Zhang, L.; Sun, Q. Deep learning for image-based cancer detection and diagnosis-A survey. Pattern Recognit. 2018, 83, 134–149
6. Wu X, Zhu X, Gong-Qing W, Ding W. Data mining with big data. IEEE Trans Knowl Data Eng 2014;26(1):97–107.
7. Gebara F, Hofstee H, Nowka K. Second-generation big data systems. IEEE Comput 2015;48(1):36–41.

DOI: 10.5281/zenodo.7157438

8. Ramesh, S., Nirmalraj, S., Murugan, S., Manikandan, R., & Al-Turjman, F. (2021). Optimization of Energy and Security in Mobile Sensor Network Using Classification Based Signal Processing in Heterogeneous Network. *Journal of Signal Processing Systems*, 1-8.

9. S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

10. An, Y., Huang, N., Chen, X., Wu, F., & Wang, J. (2019). High-risk prediction of cardiovascular diseases via attention-based deep neural networks. *IEEE/ACM transactions on computational biology and bioinformatics*.

11. Ma, H., & Pang, X. (2019). Research and analysis of sport medical data processing algorithms based on deep learning and Internet of Things. *IEEE Access*, *7*, 118839-118849.

12. Chew, A. W. Z., Pan, Y., Wang, Y., & Zhang, L. (2021). Hybrid deep learning of social media big data for predicting the evolution of COVID-19 transmission. *Knowledge-Based Systems*, *233*, 107417.

13. Usama, M., Ahmad, B., Wan, J., Hossain, M. S., Alhamid, M. F., & Hossain, M. A. (2018). Deep feature learning for disease risk assessment based on convolutional neural network with intra-layer recurrent connection by using hospital big data. *Ieee Access*, *6*, 67927-67939.

14. D. Eigen, J. Rolfe, R. Fergus, and Y. LeCun, "Understanding deep architectures using a recursive convolutional network", In International Conference on Learning Representations (ICLR), 2014.

15. M. Lin, Q. Chen, and S. Yan. "Network in network", In International Conference on Learning Representations (ICLR),2014

16. Baek, J. W., & Chung, K. (2020). Context deep neural network model for predicting depression risk using multiple regression. *IEEE Access*, *8*, 18171-18181.

17. Rajasekaran Manikandan, Yassine Abdulsalam, Shamim Hossain M, Alhamid Mohammed F, Guizani Mohsen. Autonomous monitoring in healthcare environment: Reward-based energy charging mechanism for IoMT wireless sensing nodes. Future Generat Comput Syst 2019;98:565–76.

18. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv **2017**, arXiv:1704.04861.

19. Velasco, J.; Pascion, C.; Alberio, J.W.; Apuang, J.; Cruz, J.; Gomez, M.A.; Molina, B.; Tuala, L.; Thio-ac, A.; Jorda, R.J. A Smartphone-Based Skin Disease Classification Using MobileNet CNN. Int. J. Adv. Trends Comput. Sci. Eng. **2019**, 8, 2632–2637.