Research Article

Multiple Classifier System Based Lung Disease Prediction with Spark Framework in Big Data

V.Durgadevi[a], **and R.Priya**[b]

[a] Research Scholar, department of Computer Applications, VISTAS ,Pallavaram.
[b]Professor, Department of Computer Applications, VISTAS, Pallavaram.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Technologies related to Big Data are potentially effective for transforming healthcare information and have developedseveral industries. Moreover, as cost is reduced, numerous lives are saved and the results are improved. Lung disease causes more death worldwide. The death rate can be reduced when detection is done at early stages but signs and symptoms are not clear in Lung disease at that stage. Hence, preventing or predicting is relatively difficult. This paper focuses on developing a prediction model for diagnosing lung disease which employs multi-structure integrated dataset.With the big data framework for healthcare and accurately predicting lung problems at the earlier stages, using machine learning approaches are considered to be the best. In this research work, a sequence of machine learning methods along with apache spark architectureis proposed for effective data classification and predicting therisk level of disease appropriately. The proposed algorithm is named as Spark framework with Multiple Machine Learning Classifier Algorithm (SMMLCA) which uses Naïve Bayes and J48 Classifier. The proposed approach is compared with two standard methods namely Convolutional Neural Network based Multimodal Disease Risk Prediction (CNN-MDRP) algorithm and Recurrent Machine Learning (RML)-based prediction modelsinterms of accuracy, precision, recall, F1-meassure, ROC and AUC. It is found that the proposed SMMLCA achieves 85.4% of accuracy, 84.2%of precision, 74.2% of recall, 71.4% of F1-measure, 75%of AUC and 61.4% of ROC.

Keywords: Lung disease, big data, classification, prediction, voting process.

1. Introduction

Healthcare, an information intensive industry, is changing at a higher rate. Several processes are undergoing within health sectors. These processes have a greater impact on not only the health of individuals but even assist medical physicians. Big data analytics has an extremely power in processing these data and hence expandingrapidly [1]. These technologies play a vital role in the developing healthcare sectors. The vast healthcare data are combined and structured using the tools of big data. Analytical models help in analyzing these data and predicting diseases or improving healthcare processes. Several major challenges in medical healthcareapplicationsbased on big data analytics are described in [2]. A comprehensive overview of various research areais presented in [3]. Several powerful tools and useful techniques are applied to improve thefacilities of existing healthcare services [4]. Based on several comprehensive widespread architectures which uses open sources like Apache Strom and Hadoop, big data analytics can be developed [5]. Integrating throughput, real time computing ability along with storage capacity effectively handles vast amount ofhealthcare data at faster rate.

One area where the existing methods can influence healthcare when employing Big Data analytics is lung disease which causes more death worldwide. The inability of the lung to circulatesufficient blood to the body tissues is considered as a lung disease [6]. Even though there are several improvements in the providing treatment to cardiac disorders, HF is still the major cause for death globally and the difficult situations faced in healthcare system [7]. In 2015, a survey from American Lung Association (AHA) revealed that [8] round 17.3 million individuals passed away per year due to the failure of lungs and predicted that it may be 23.6 million by 2030. By the statistics given by World Health Organization (WHO) in 2010 [9], 42% of death in Kingdom of Saudi Arabia (KSA) was due to lung disorders. This lung disease which is a complex and heterogeneous disease is very difficult to detect as several unusual symptoms are found [10]. Few risk factors experienced due to lung disease are dyspnea, breathing, fatigue, loss of appetite, sleeping difficulties, memory loss, cough with phlegm or mucus foam, diabetes, hypertension, hyperlipidemia, medication, anemia, smoking and family history. Detection of Lung disease and its failure is based on theperception and experience of the doctor instead of rich information present in the database thus disease diagnosis is not done earlier. Hence, the challenge is to consider and utilize clinical data present in the databases to provide early diagnosis and contribute valuables towards healthcare industry. Prediction done at the early stages eliminates unwanted biases, errors and reduces the costs, improvessurvival rate and provide satisfactory services for patients. Individuals at risk at identified earlier and avoids people becoming critical. Medical data are available in the form of history, test results andcomplex either in the structured, semi-structuredand unstructured form reports [11]. For the risk prediction model, handling structured data is easy. But, in unstructured data, more valuable information are lost as they are discrete, verycomplex, noisy and multidimensional [12].

The major objective of thiswork is to bring out and reveal the valuable information suing medical reports using pulmonologist and designing a model to predict lung disease. The remaining part of this paper is arranged as: related works are discussed in Section II and the proposed architecture is elaborated in Section III with its result analysis in Section IV. The work is concluded with future enhancement in Section V.

2. Related works

In [13], Min Chenet al., modified the prediction model which used real-time clinical data obtained from central China. With the incomplete data, difficulties were faced and to overcome these difficulties latent factor model was employed for reconstructing the data that were missing. A convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm was designed which used structured as well as unstructured data. This was the first work which concentrated on both data types in medical big data analytics. The prediction accuracy obtained was 94.8% with faster speed than CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm. In [14], Joo et al., analyzed the Korean National Health Insurance Service–National Health Sample Cohort (KNHSC) data and examined the features of ML and big data to predict CVD risk. Particularly, efficiency of different ML methods was analyzed in predicting CVD risks like atrial coronary artery disease,fibrillation, strokes and lung failure. Medical data, questionnaire results, comorbidities, and past medical information were considered to develop this Recurrent Machine Learning-based prediction models using deep neural networks (DNN), logistic regression (LR), random forests (RF), and LightGBM. The performance was validated with metrics like receiver operating characteristic (ROC) curves, precision, recall, specificity, and F1 measure. This approach was better than the baseline approach. In[15], Murillo et al.,coined Structural cooccurrence matrix (SCM)classification model to detect the disease as benignor malignant. The features from the image were extracted using GLCM approach. For better performance of classification, Gaussian, Laplace, and Sobel filters were integrated with SCM. SVM classifier was better than Decision-based,MLP, andANN classifiers. In [16], Usama et al. developed new recurrent convolutional neural network (RCNN)-based model to assess the risk using structured as well as unstructured clinical text information. For recurrent operation, the ROI was increased, thereby feature extraction was simplified. Further, data parallelism approach was employedfor training and testing the modelwhich had fast conversion speed. In [17], Verma et al., developed a hybrid model which integrated feature selection approach and Correlation for diagnosing coronary artery disease (CAD). In [18], short text was automatically learned to diagnose disease with the help of machine learning approaches. In [19], Nalband et al. employed a feature selection and classification technique for diagnosing knee joint disorders using vag signal. In [20], Chang et al., employed semi-supervised multi-label feature selection technique for larger datasets. In [21], Zhu et al., coined a relational regularization model to select features by embedding relational information and classifying Alzheimer disease. In [22], feature selection approaches were examined for multi-label classification for detecting chronic diseases. Other relevantinvestigation on chronic diseases was conducted. In [23], Kim et al., Decision Tree and Fuzzy Logic model was developed for predicting the risk of coronary lung disease. In [24], Shi et al., designed a uniform model which had the ability to assess the risks of multiple diseases.Here, CNN was employed for extracting features from unstructured data. In [25], Moral et al., presented an automatic CNN-based method for feature extraction using electronic health records and diagnosis was based on multilabel learning.

3. System model

Supervised learning classification process is employed for the prediction of given input and classifying with certain labelledclass. In classification, the novelty is based on the function used for mapping the input to a certain output. Learning classifiers utilized here are Naïve Bayes andJ48 Classifier. In the decision support system, dataset containing images of various diseases are involved and algorithms are applied for training the dataset. This process is illustrated in figure-1. User dataare gathered which are given as the input to the model for processing on the server where the diagnosis is made and results are predicted. This paper focuses on the concept of novel Machine Learning approach to diagnose lung disease dataset in order to produce better accuracy rate.

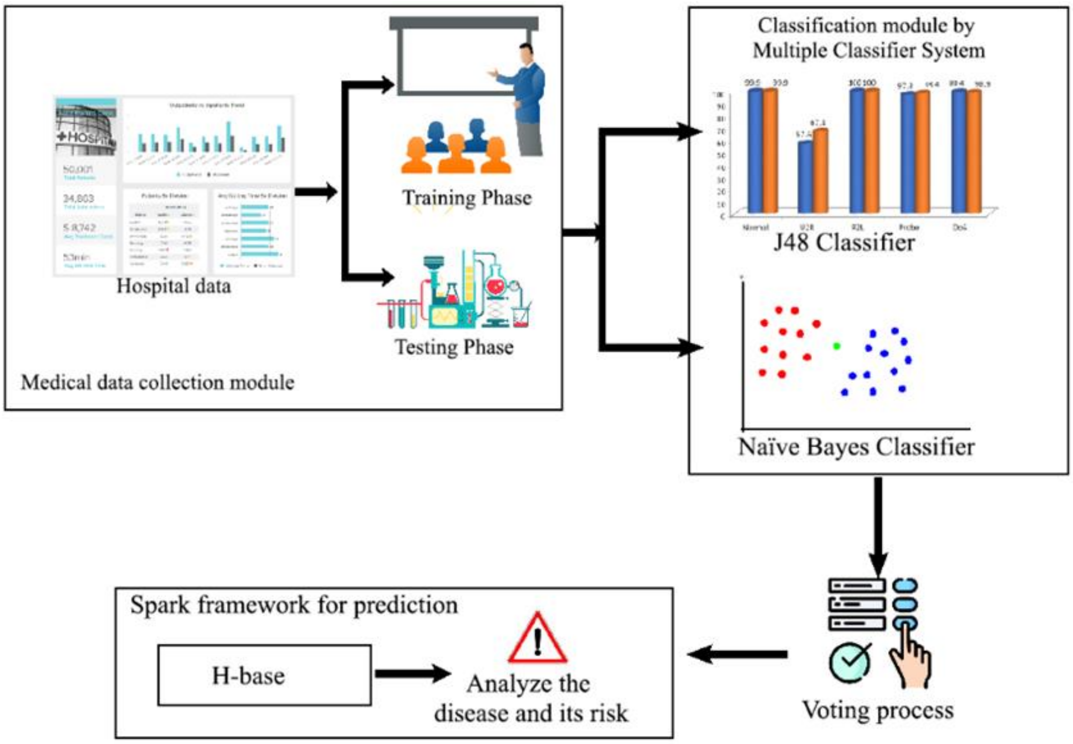

Figure1:System architecture of the proposed lung disease prediction method

3.1 Heterogeneous classifier system:

A set of classifiers termed asMultiple Classifier System (MCS)combines every prediction of each classifier for classifying new samples. Classifier ensemble models using multiple learning algorithms are involved to achieve better prediction results than a single learning algorithm. Combining different type of classifiers yields better results. The criteria to be followed is that the classifiers must possess a considerabledisagreement level by providing independent errors among themselves. The errors identified by the classifiers should independent and every classifier has to perform well than guessing atrandom. When MCS is involved, local different behaviors of every classifier are reduced where the average of the results of every classifier is considered.

3.1.1 J48 Classifier

J48 represents c4.5 in weka tool of java. Decision tree concept is implementedto determine the solution for the problem. Leaf nodes of the tree represent the class labels while the internal nodes define the attributes. Here, the process of selecting attributes is performed usinginformation gain and gain index. Based on information gain and its importance, classification is performed using decision tree. The information gain for an attribute X of a node is computed by:

Information gain (n,x) =entropy (n) - $\sum_{x}^{n} \frac{n}{n'}$entropy (n)

here nand n′ indicates the set of instancesof a specific node andcardinality respectively

Entropy of n is computed by:

Entropy(n)=$\sum_{i=1}^{n} -pi\ log2pi$

3.1.2 Naïve Bayes Classifter

This classifier, one among probabilistic classifier, has strong independent assumption between features. Naive Bayes classifier works with the principle of bayes Theorem and uses bayesian network with a posterior maximum decision rule in a Bayesian Set up. By using this classier, features classified are independent with each otheralways. If x and y indicate the dependent feature vector and a class variable respectively, then.

$$Y=\text{argmax } p\ (y) \prod_{i=1}^{n} p\ (\frac{xi}{y})$$

P(y) and P($\frac{xi}{y}$) denote the class probability andconditional probability. Bayesian probability is given by

Posterior=$\frac{\text{Prior}*\text{Likelihood}}{evidence}$

3.2 Voting system

Initially the assumption is done, that there are M classes C1,C2,C3….Cn in the dataset, with every Ci representing the ith class. For K classifiers, Ek denotes the kthclassifier. The confusion matrix PTk is achieved using Ek for classifying thetesting samples. For Ek classifier with PTk , the possibilities that suggests Ci = 1,2…M are true when an event occurs.

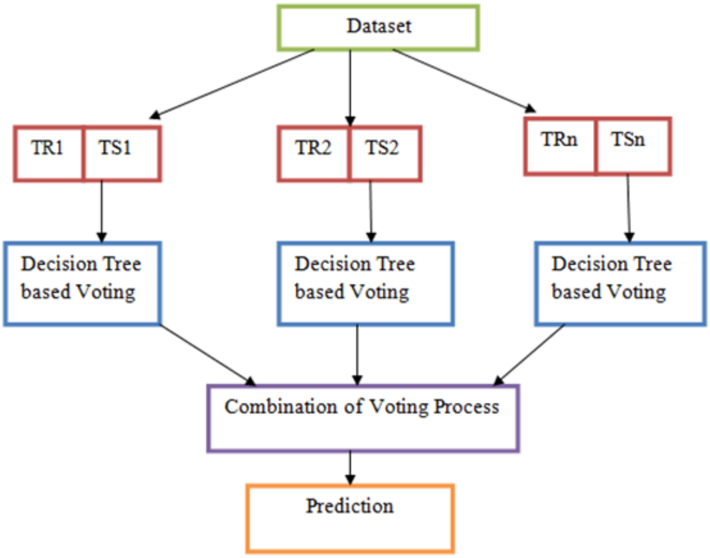

Figure-2Role of voting system in prediction process

Further,the computations on preprocessed data are performed using trained classifier module. For ensemble training,training dataset which is a labeled one is employed. Once training every group in the model, trained classifiers are independently aggregatedto a suitable combination method.The Weighted Majority Voting (WMV) ensemble mechanism sorts out unlabeled instances to a classgets most common votes or the highest number of voting. The WMV ensemble mechanism is generally denoted as Plurality Vote (PV) approach.Most often, WMV mechanism is applied for equating the performance of various methods. Mathematically

$$\text{class}(x) = \arg max_{c_i \in dom\ (y)} (\sum_k g(y_k(x), c_i))$$

here classification of kth classifier is given as $y_k(x)$ and $g(y,c)$ gives the index function and is demonstrated as

$$g(y,c) = \begin{cases} 1 & y = c \\ 0 & y \neq c \end{cases}$$

If the probabilistic classifier is utilized, the classification $y_k(x)$ is obtained using

$$\text{class}(x) = \arg max_{c_i\ \in dom\ (y)} PM_k(y = c_i|x)$$

Where M_k is applied to demonstrate the classifier k and $PM_k(y = c_i|x)$ represents about the probability of class c for x. Each voting process presents a varying weight for every base classifier. This weight relies on the accuracy rate provided by the classifier.For each bug report, a severity class is predicted by every n base classifier.

3.3 Spark framework

Let WT be the weight of vector which separates the training set from labels given as input to WTW. Assume $\Sigma_{(xi\ ,yi)}\beta$as the total slack required to attain marginal constraint.βi denotes a slack variable which solves the optimization problem of label classification. Theoptimization ofactual label from other ones uses 1-βiwhich is positive (βi> 0). Here, taking K labels as input, weights is given as nK. With the help of the kernel function α(*xi*), training dataset x is mapped to a high dimensionalplane. The corresponding functionargmax($wTm\alpha(x) + bm$) gives higher decision valueto the class. For initializing the marginal value and maximizing it, the following is used.

Ωi = (W.Xj+b)yj

Where,

$\Omega = \frac{1}{\sqrt{w}}$

Xj = xj’ + μ $\frac{w}{\sqrt{w}}$

W = 1(default)

Forlinearly separable cases, maximizing the marginal value is given by,

$$mx\ (W.Xj+b)yj>\Omega$$

Generally, this is used to setup a margin setup. Here, in the proposed multi classification model, the linear problem is solved using a slack variable which is given as:

$$\min W.W + C\sum_{k=j}^{n} \forall j(W.Xj + b)YJ \geq 1 - \beta J, \forall J \geq 0$$

To handle massive data with higher dimension to be efficient, in apache sparkarchitecture is proposed which function in a better way with distributed dataprocessing in python. With this perception, impacts of multi-classclassificationafter the process of map-reduce processing, features extraction and generation of CSV file is based on its properties. The CSV file which is generated moves toapache spark architecture where the code is compiled obtaining a byte code in which dynamical interfacing is allowed. As the dataset used is high dimensional and allows categorical values, the process of spark uses map-reduce and SQLdatabase for data analysis.In several servers, Kafka acts as cluster and resilient to store a sequence of records consisting of a key, value and timestamp.

3.4 Algorithm: Spark framework with Multiple Machine Learning Classifier Algorithm (SMMLCA)

```
Algorithm 1: Proposed SMMLCA
Input-
        Learning Rate (LR) =β
        β i=1,2,3….n regularization constant;
        N denotes the maximum number of iterations;
        P(u) and Q(u) represent the initialization of Puand qv respectively;
Output-
        Predicted disease ⇠y
Start
Training datasets ($tr_d$), testing datasets($te_d$) ⇠D(t)
for
        Retriving the attributes (X)
        Calculate gain (n,x) ⇠z
        Z ⇠ entropy (n)
        Compute Y and posterior probability
        Assumption of classes C= C1,C2,C3….Cn
        K= E1,E2,E3…..Ek
End for
K ⇠M
Computation of Plurality Vote (PV)
        $class(x) = arg\ max_{c_i \in dom(y)}(\sum_k g(y_k(x), c_i))$
Estimating the probabilistic unit for slacking process
        βi> 0
        Ωi = (W.Xj+b)yj
Maximize Ωi
                If
                        Repeat = 'data'_id
                        Neglect or minimize Ωi
                Else
                        Repeat
End if
Finding the minimal content min W with slacking
```

4. Performance analysis

Comparison of the proposed Spark framework with Multiple Machine Learning Classifier Algorithm (SMMLCA) is done with the existing methods such asconvolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm and Recurrent Machine Learning (RML)-based prediction models for evaluating the performance. *TP*, *FP*, *TN* and *FN* represent true positive (total relevant instances predicted correctly), false positive(totalrelevant instances predicted incorrectly),true negative (totalirrelevant instances predicted correctly) and false negative (total irrelevant instancespredicted incorrectly), respectively. Then, four prametric measures namely accuracy, precision, recall andF1-measure are considered:

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

Precision $=\frac{TP}{TP+FP}$

Recall $=\frac{TP}{TP+FN}$

F1-Measure $=\frac{2\times \text{Precision} \times \text{Recall}}{Precision+recall}$

Besides the evaluation criteria mentioned above, ROC (Receiver Operating Characteristics) curve and AUC (Area Under Curve) were employed to evaluate performance of theclassifier. ROC curve illustrates the trade-off between true positives and false positives.

The model is considered as a better one, when it is observed that the ROC curve is closer to the upper left corner of thegraph andif the area is closer to 1. When dealing with medicaldata, more attention has to be given on recall instead of accuracy. When recall is high, rick factor for a lung disease patient is low.

The table 1 shows the comparison of existing CNN-MDRP and RML with proposed SMMLCA method for accuracy.

Table 1: Analysis of accuracy

Number of datasets	**CNN-MDRP(%)**	**RML(%)**	**SMMLCA(%)**
1000	70	75	80
2000	75	78	84
3000	79	82	86
4000	82	85	88
5000	85	86	89

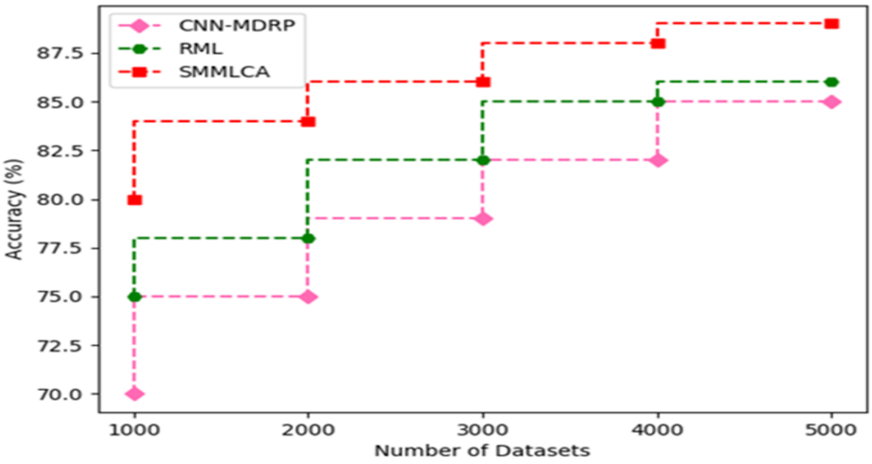

Figure 3: comparison of accuracy

Figure 3 plots the accuracy of the of existing CNN-MDRP and RML with proposed SMMLCA method. X axis represents the number of datasets while the Y axis provides the obtained accuracy values in percentage. When compared, existing method achieves 78.2% and 81.2% while the proposed method achieves 7.6% better than CNN-MDRP and 4% better than RML.

The table 2 shows the comparison of existing CNN-MDRP and RML with proposed SMMLCA method for precision.

Table 2: Analysis of precision

Number of datasets	**CNN-MDRP(%)**	**RML(%)**	**SMMLCA(%)**
1000	72	74	78
2000	76	77	81
3000	80	81.5	85
4000	83	84.2	87
5000	86	85	90

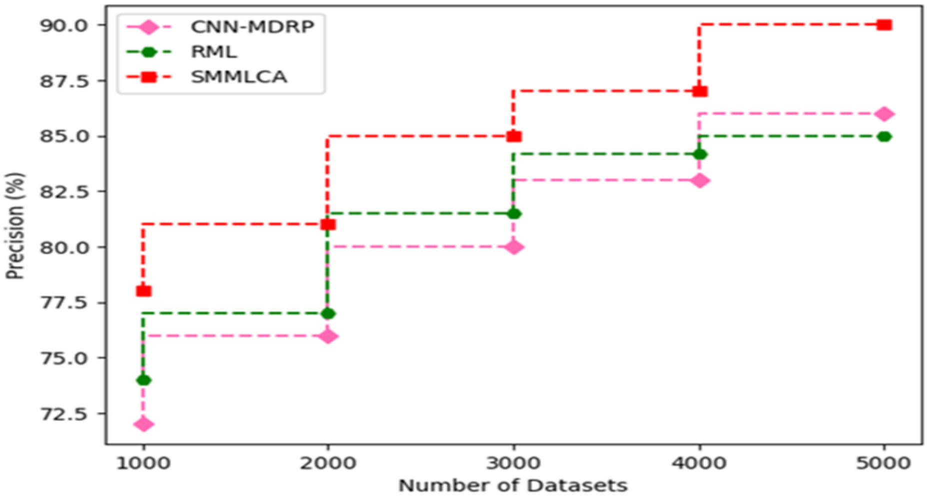

Figure 3: Comparison of precision

Figure 3 plots the precision of the of existing CNN-MDRP and RML with proposed SMMLCA method. X axis represents the number of datasets while the Y axis provides the obtained precision values in percentage. When compared, existing method achieves 79.4% and 80.34% while the proposed method achieves 5% better than A CNN-MDRP and 4% better than RML.

The table 3 shows the comparison of existing CNN-MDRP and RML with proposed SMMLCA method for recall.

Table 3: Analysis of recall

Number of datasets	CNN-MDRP(%)	RML(%)	SMMLCA(%)
1000	62	64	68
2000	66	67	71
3000	70	72.5	75
4000	73	75.2	77
5000	76	79	80

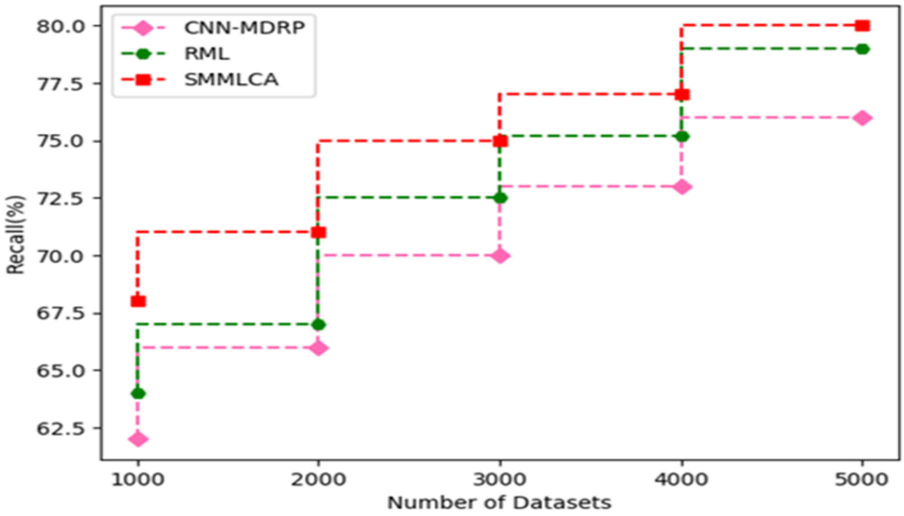

Figure 5: Comparison of recall

Figure 5 plots the recall of the of existing CNN-MDRP and RML with proposed SMMLCA method.X axis represents the number of datasets while the Y axis provides the obtained recall values in percentage. When compared, existing method achieves 69.4% and 71.54% while the proposed method achieves 5.6% better than CNN-MDRP and 3.2% better than RML.

The table 4 shows the comparison of existing CNN-MDRP and RML with proposed SMMLCA method for F1measure.

Table 4: Analysis of F1 measure

Number of datasets	CNN-MDRP(%)	RML(%)	SMMLCA(%)
1000	60	62	65
2000	62	65	68

3000	66	69	71
4000	70	72	74
5000	72	75	79

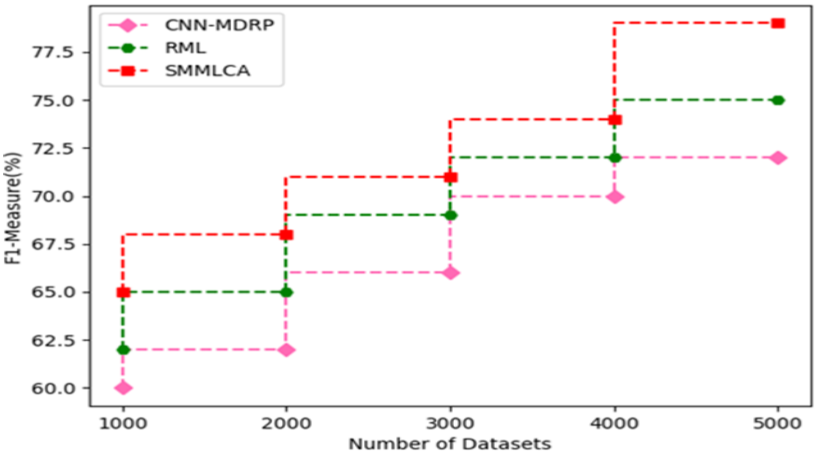

Figure 6: comparison of F1 measure

Figure 6chows the F1 measure of the existing CNN-MDRP and RML with proposed SMMLCA method.X axis represents the number of datasets while the Y axis provides the obtained F1-measure values in percentage. When compared, existing method achieves 66% and 68.6% while the proposed method achieves 5.4% better than CNN-MDRP and 3.2% better than RML.

The table 5 shows the comparison of existing CNN-MDRP and RML with proposed SMMLCA method for AUC.

Table 5: Analysis of areaunder curve (AUC)

Number of datasets	CNN-MDRP(%)	RML(%)	SMMLCA(%)
1000	63	65	69
2000	65	69	72
3000	69	71	75
4000	75	75	79
5000	78	80	83

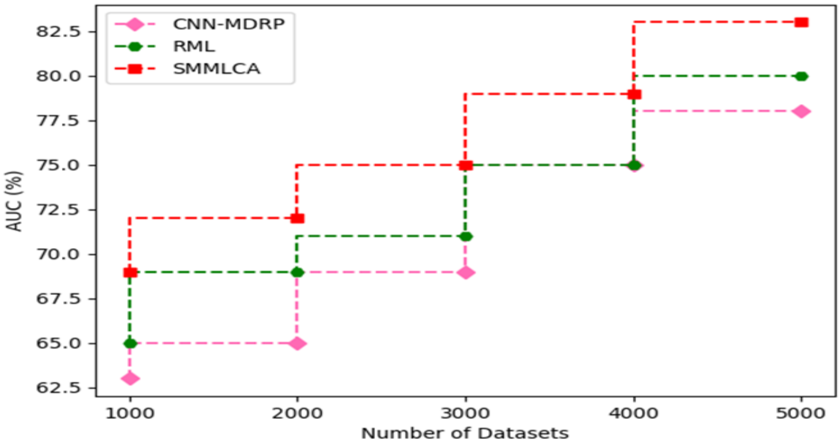

Figure 7: comparison of AUC

Figure 7 plots the AUC of the of existing CNN-MDRP and RML with proposed SMMLCA method.X axis represents thenumber of datasets and Y axis provides the obtained AUC values in percentage. By comparison, existing approach achieves 70% and 72% whereas the proposed approach achieves 5.8% better than CNN-MDRP and 3.8% better than RML

The table 6 shows the comparison of existing CNN-MDRP and RML with proposed SMMLCA method for ROC.

Table 6: Analysis of receiver operating characteristic (ROC) curve

Number of datasets	CNN-MDRP(%)	RML(%)	SMMLCA(%)
1000	50	52	55
2000	52	55	58
3000	56	59	61

4000	60	62	64
5000	62	65	69

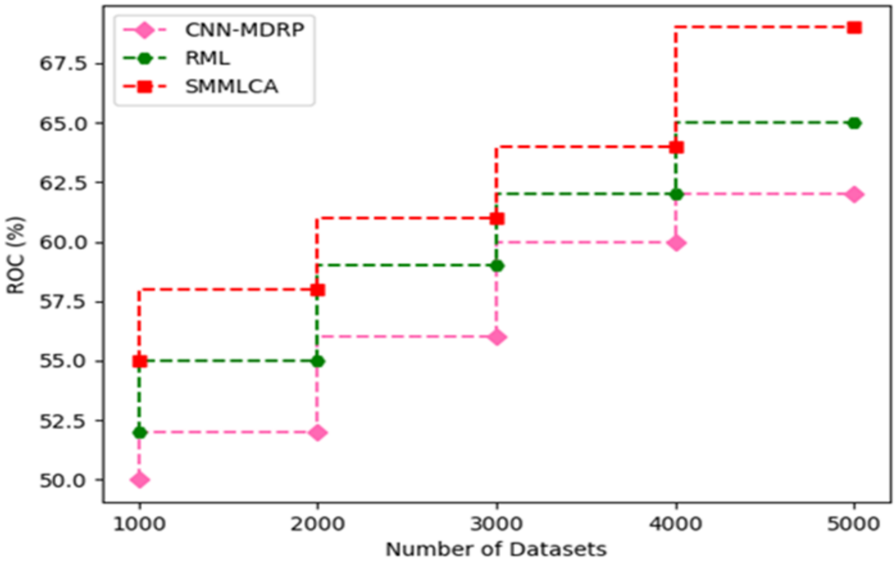

Figure 8: comparison of ROC

Figure 7 plots the ROC of the of existing CNN-MDRP and RML with proposed SMMLCA method .X axis and Y axis shows that number of datasets and the ROC values obtained in percentage respectively. When compared, existing method achieves 56% and 58.6% while the proposed method achieves 5.4% better than CNN-MDRP and 3.2% better than RML.

The table 7 shows the overall comparison of existing CNN-MDRP and RML with proposed SMMLCA

Table 7: Overall comparative analysis

Parameters	**CNN-MDRP(%)**	**RML(%)**	**SMMLCA(%)**
Accuracy	78.2	81.2	85.4
Precision	79.4	80.34	84.2
Recall	69.4	71.54	74.2
F1 measure	66	68.6	71.4
AUC	70	72	75.8
ROC	56	58.6	61.4

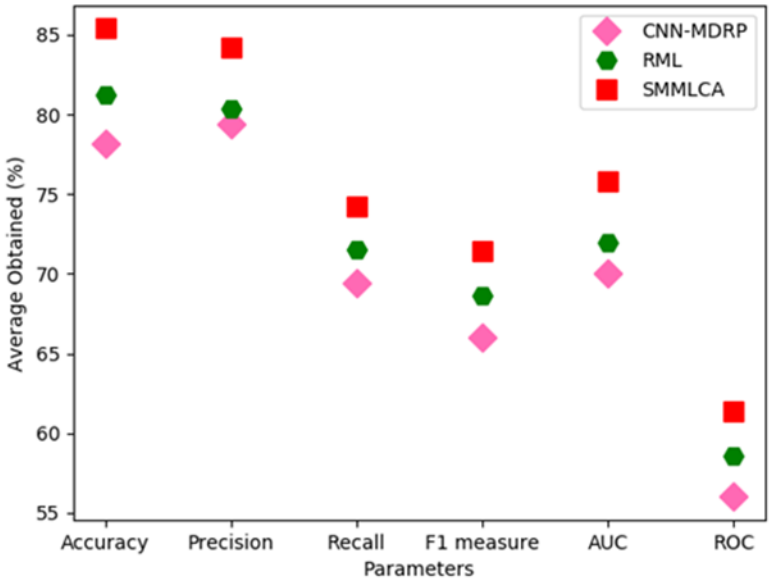

Figure 9: Overall analysis of between existing and proposed method

The figure 9 compares the values achieved for the parameters. Xaxis represents parameters considered for analysis and and Y axis values obtained in percentage respectively.

5. Conclusion

Big Data analytics provides a systematic way to produce better results like affordability and availability of healthcare service to everyone. Non-Communicable lung disease is one of the major hazardsrelated to health all over the world. After transforming numerous health records of the patients affected withlungdisease into useful result, the patientsare aware of the complications that can occur. In this work, data classification played a vital role in performing multi-structured datasets. The objective here is to deal with lung disease predication in healthcare using Big Data analytics technique. This designed predictive analysis model has provided enhanced data analytics for better outcomes towards healthcare. The proposed Spark framework with Multiple Machine Learning Classifier Algorithm (SMMLCA)improved the performance of the classifier. The results of SMMLCA are better than other works and as achieves 85.4% of accuracy, 84.2%of precision, 74.2% of recall, 71.4% of F1 measure, 75%of AUC and 61.4% of ROC.

Reference

1. SantoshiKumari, Haripriya.A, Aruna.A, Vidya.D.S, Nithya.M.N, Immunize - Baby Steps for smart healthcare Smart solutions to Child Vaccination, IEEE International Conference on Innovations in Green Energy and Healthcare Technologies(ICIGEHT'17), 2017 IEEE
2. K. Shailaja, B. Seetharamulu, M.A. Jabbar , Prediction of Breast Cancer Using Big Data Analytics, International Journal of Engineering & Technology, 7 (4.6) (2018) 223-226
3. Rati Shukla, Vikash Yadav, Parashu Ram Pal, Pankaj Pathak, Machine Learning Techniques for Detecting and Predicting Breast Cancer, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-7 May, 2019
4. Liu YY, Chen YM, Yen SH, Tsai CM, Perng RP (2002) Multiple primary malignancies involving lung cancer—clinical characteristics and prognosis. Lung Cancer 35(2):189–194
5. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, ''Dropout: A simple way to prevent neural networks from overfitting,'' J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014
6. G. Koulaouzidis, D. K. Iakovidis, and A. L. Clark, "Telemonitoring predicts in advance heart failure admissions". Int J Cardiol, vol. 216, pp. 78–84, 2016.
7. L. Turgeman and J. H. May, "A mixed-ensemble model for hospital readmission.",ArtifIntell Med, vol. 72, pp. 72–82, 2016.
8. Y. Kang, M.D. McHugh, J. Chittams and K.H. Bowles, "Utilizing home healthcare electronic health records for telehomecare patients with heart failure. A decision tree approach to detect associations with rehospitalizations". Comput Inform Nurs, vol. 34 no. 4, pp.175– 182, 2016.
9. M. Panahiazar, V. Taslimitehrani, N. Pereira and J. Pathak, "Using EHRs and machine learning for heart failure survival analysis." Stud Health Technol Inform, vol. 216, pp. 40–44, 2015.
10. M. Saqlain, W. Hussain, N. Saqib and A. Muazzam Khan, "Identification of Heart Failure by Using Unstructured Data of Cardiac Patients.", 45th International Conference on Parallel Processing Workshops, 2016.
11. G. Yang et al., "A heart failure diagnosis model based on support vector machine". 3rd International Conference on Biomedical Engineering and Informatics. 2015;
12. C. Moral, A. Antonio, R. Imbert and J. Ramírez, "A survey of stemming algorithms in information retrieval." Information research, vol. 19 no. 1, March 2014
13. Min Chen, YixueHao, Kai Hwang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities" IEEE Access, vol. 4, pp. 5937–5947, 2016
14. Joo, G., Song, Y., Im, H., & Park, J. (2020). Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea). IEEE Access, 8, 157643-157653.
15. Murillo BR (2018) Health of things algorithms for malignancy level classifcation of lung nodules. IEEE Acces

16. 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.
17. L. Verma, S. Srivastava and P. C. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data", J. Med. Syst., vol. 40, no. 7, pp. 178, 2016.
18. O. Frunza, D. Inkpen and T. Tran, "A machine learning approach for identifying disease-treatment relations in short texts", IEEE Trans. Knowl. Data Eng., vol. 23, no. 6, pp. 801-814, Jun. 2011.
19. S. Nalband, A. Sundar, A. Agarwal and A. A. Prince, "Feature selection and classification methodology for the detection of knee-joint disorders", Comput. Methods Programs Biomed., vol. 127, pp. 94-104, Apr. 2016.
20. X. Chang, H. Shen, S. Wang, J. Liu and X. Li, "Semi-supervised feature analysis for multimedia annotation by mining label correlation", Proc. Adv. Knowl. Discovery Data Mining 18th Pacific-Asia Conf. (PAKDD), pp. 74-85, May 2014.
21. X. Zhu, H.-I. Suk, L. Wang, S.-W. Lee, A. D. N. Initiative and D. Shen, "A novel relational regularization feature selection method for joint regression and classification in AD diagnosis", Med. Image Anal., vol. 38, pp. 205-214, May 2015.
22. Show Context CrossRef Google Scholar
23. X. Zhu, H.-I. Suk, S.-W. Lee and D. Shen, "Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification", IEEE Trans. Biomed. Eng., vol. 63, no. 3, pp. 607-618, Mar. 2016.
24. J. Kim, J. Lee and Y. Lee, "Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree", Healthcare Inform. Res., vol. 21, no. 3, pp. 167-174, Jul. 2015.
25. X. Shi et al., "Multiple disease risk assessment with uniform model based on medical clinical notes", IEEE Access, vol. 4, pp. 7074-7083, 2016.
26. C. Moral, A. Antonio, R. Imbert and J. Ramírez, “A survey of stemming algorithms in information retrieval.” Information research, vol. 19 no. 1, March 2014
27. Murugan, S., Jeyalaksshmi, S., Mahalakshmi, B., Suseendran, G., Jabeen, T. N., & Manikandan, R. (2020). Comparison of ACO and PSO algorithm using energy consumption and load balancing in emerging MANET and VANET infrastructure. Journal of Critical Reviews, 7(9), 2020.