Smart Lung Cancer Detection & Prediction Using Deep Learning

Sandeep Kaur[1],Dr.Jaspreet Singh[2]
[1]Assistant Professor, Department of Computer Engineering and Technology,
Guru Nanak Dev University, Amritsar
[2]Associate Professor ,Department of Computer Science And Engineering,
GCET,Kahnpur Khui(Anandpur Sahib)

ABSTRACT:

The uncontrollable proliferation of cells in our lungs normally contributes to lung cancer in both men and women. The disorder inhale and exhale part of the chest causes a severe respiratory problem. Cigarette and passive smoking are the major causes in the World Health Organisation to lung cancer. In younger and older people as compared with other cancers, the death risk from lung cancer is rising every day. While high-tech medical facilities are available to diagnose carefully and successfully manage the death risk is still not properly regulated. In the early stages, however, early steps are strongly important so that improved diagnosis can be accomplished from signs and results. The high compute capacity to detect diseases early on and reliably interpret data makes computer training nowadays a significant effect on medical services. We also analyzed different methods for learning the machine to interpret available knowledge on lung cancer. In this paper we will speak about the profound learning approach and application of cancer detection.

Keyword: Lung Cancer, Deep learning, Classifiers, ML.

I. INTRODUCTION

The world's leading cause of death from lung cancer. In the final point the signs of lung cancer become evident. So in the early stage it is very difficult to classify. Therefore, in contrast to all other forms of cancer, the mortality risk for lung cancer is very high[24]. Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC)[9] are the two forms of lung diseasing that arise and spread spontaneously. The lung disease process refers to the degree of expansion in the lung. This is one of the deadliest diseases; in developmental countries, the mortality rate is dropping just 17 per cent of individuals that have been determined that the lung tumor persists 5 years after identification. Local binary pattern [10] operator converts an image into an array or image of integer labels representing the tiny image presence [25]. Those markers are then used for further image analyzes (most commonly the histograph). The LBP texture operator has become a common approach in numerous applications thanks to its discriminative capacity and computational simplicity[22].
The prevalence of prostate or breast cancer in the last three decades was greatest among both men and woman patients, but in patients with cancer lung cancer it was highest[13]. One major cause is the more advanced and systemic prognostic models of prostate cancer and breast cancer than lung cancer[20]. An appropriate early-stage lung cancer forecast model must therefore be developed as a matter of urgency. In linear and nonlinear issues[11] SVM has superior predictive efficiency and is commonly used in numerous fields[12] and [14], including in the area of medical care [14]. Although SVM is a superior classifier, it is relatively immature with the field of cancer prognosis models [14].

Many advanced techniques in the field of lung cancer prediction were introduced in the last decades. Mutation testing [15] has become an important tool in clinical test to decide the correct treatment options for patients.

In the last decades a lot of new methods have been implemented in the area of lung cancer prediction. Mutation test[15] has become an important way of evaluating the best therapeutic choices for patients in the clinical research. Direct sequencing is an additional, undisclosed mutations screening process. Mutation testing Epidermal Growth Factor Testing (EGFR) [16] is developed in the field of lung cancer gene mutation studies. Several computer-based research methods were developed to provide a comparison of ensembles and their non assortment variants between two forms of classificators: the artificial neural system (ANN, for its Spanish initial application) [19] [23] and supporting vector maker[21] [17]. The weight of the misjudgement is greater than that of the minority class, and this leads to a high chance of confusion. Traditional algorithms of classification are not efficient and fine. Though improvement in diagnosis and treatment methods has been made lately, NSCLC patients' prognosis is low and is primarily attributed to a shortage of early diagnostic instruments.

II. LITERATURE REVIEW

In 2019, Daoudy et al.[1] analyzed the possibility that an Adaboost mixture will use decision stumps to predict early lung cancer (machinery learning ensemble). For illustrations, 9 trace elements of 122 urine specimens were analyzed with a dataset. The Kennard and Stone (KS) algorithm in conjunction with alternative re-sampling to do the sample collection partitioning. The whole data set was split into equally large training and test sets that were tweeted to establish a further case A and case B.The predicted results were close to those in an Adaboost-oriented discrimination analysis by Fisher (FDA). In both cases, the final Adaboost Classification System was completely sensitive: 93.8% accuracy, 95.7% accuracy and overall accuracy of 95.1%, 96.7% for cases A and B respectively. In any event, Adaboost is still stronger than FDA and less prone to the structure of the training set than FDA. Adaboost is clearly superior to the FDA in today's challenge. In medical practice, adaboost and trace electrolyte urine analysis combined to diagnose early lung cell cancer.

A multi-class data pathway behavior transformation strategy called the analyz-of-variance feature set was introduced in 2015 by Engchuan and Chan[4] (AFS). The findings of classification by using the proposed approach pathway operation indicate strong classification intensity in triple cross-validation and robustness in the entire data collection validation for all four lung cancer datasets.

Azzawi et al.[5] suggested a GEP model in 2016 to forecast lung cancer from the details on microarray. In order to isolate major genes associated with lung cancer, authors are using two gene selection approaches, and thus give multiple GEP prediction models. Predictive performance assessments and comparisons were extensively carried out in true micromic array lung cancer data sets between the authors' GEP models and three representative machine learning approaches, vector support, multi-lag perceptron and radial function neural network. The validation of the cross-data collection was tested for reliability. The experimental results show that, in terms of precision, sensitivity, specificity and area under the receiver operating characteristic curve, the GEP model using feature less genes outperforms other model. The GEP model has been concluded to be a better approach to prediction problems in lung cancer.

In 2016, a number of Bayesian dynamics networks (DBN) were developed and analyzed by Petousis et al.[6] to provide insights into how longitudinal evidence can better guide decisions on the diagnosis of lung cancer. The LDCT arm of the NLST dataset is used to build five DBNs for high-risk entities. The reverse building has been used to design three of these DBNs and two have been constructed through structural approaches. Both models are based on demographic statistics, cancer diagnosis, family history of lung cancer, risk factors for exposure, lung cancer co-orbidities and LDCT screening data.A lung cancer test model was used to differentiate individuals' cancer status over time due to confusion with the lung cancer screening. Models and experiments have been tested to resolve data inequality and overcompatibilities on healthy exercise and on cancer and non- cancer cases. The findings were contrasted by expert judgments. The total area under the curve (AUC) for both models was greater than 0.75 for the 3 intervention points of the NLST analysis.The assessment model was shown to be effective for generalization on the entire NLST Data Sets LDCT Arm (N = 25,486). DBNs are preferable to compare models such as logistical regression and naive Bayes. The lung cancer screening of DBNs has demonstrated substantial discrimination and predictive potential in the majority of cancers and non-cancers.

Lynch etc.[7] used the SEER data to classify 2017 lung cancer patients, including linear regression, decision trees, GBM, vector support machines (SVM) as well as custom set up machines (customized ensemble).The key characteristics of data in applying these methods include the ranking of the tumours, tumour, sex, age, measures and number of primaries, in order to comparison the predictive power of various methods. The forecast was not split into categories, but rather as an ongoing goal to boost longevity. These results indicate that the predicted values follow the true values, which are mainly the numbers, during the low to moderate survival intervals. The best efficiency technique was a custom ensemble with a 15.05 Root Mean Square Error (RMSE).In the custom ensemble GBM was the most important model. Owing to the minimal number of discrete results Decision trees cannot be included. In comparison, the findings reveal that GBM with 15.32 RMSE was one of the five versions that was manufactured more precisely. While SVM has an RMSE value of 15.82 less, it is just a statistical analysis that provides a distinctive performance.The results were consistent with the classic Cox proportional hazard model, which was used as a reference method. It was eventually found that the SEER database can be used to estimate the survival time of the patient with the ultimate objective to guid patient care decisions through the use of certain managed learning techniques to data on lung cancer and that their performance is comparable to the traditional with a particular data collection.

In 2019, Petousis et al.[8] merged a variety of learning machine approaches to learn a partially measurable decision-making method by Markov, and at the same time enhances diagnosis of lung cancer and improves research specificity. A complex Bayesian network was trained as an analytical model using the NLST data and used reverse strengthening learning to define a reward role based on the decisions of experts. The resulting predictive model lowered the false positive rate while retaining a high true positive rate at a human expertly comparable level. In[29], forecasts on breast cancer survival were implemented using broad data sets using two renowned data-mining methods such as ANN and DT, and also a conventional mathematical method was used. LR. The proposed model also described a variety of lung cancers previously.

Table 1: Features and challenges of existing lung cancer prediction models

Author [citation]	Methodology	Features	Challenges
Tan *et al*. [1]	Adaboost	Has attained high sensitivity and best performance. It is very simple to implement.	It is very sensitive to noisy data.
Kim *et al*. [2]	Decision Tree (DT)	These are simple to interpret. It should be taken as the minimal decision standard of work-relatedness for lung cancer.	They suffer from overfitting.
Zieba *et al*. [3]	Boosted SVM	It is used in medical application for predicting post-operative life expectancy in lung cancer patients. It is used to solve the imbalanced data problems.	The running time of training algorithms do not scale well with the size of the training set.
Engchuan, and Chan [4]	SVM	It is used to build n- hyperlanes and n-features for dividing each different class apart from maximal margin. It improves the	Many parameters need to be set accurately for attaining the best results.

III. RESEARCH GAP

Lung cancer is the second major disease occurred in humans and it mainly leads to cancer mortality in the entire world. The whole 5-year survival rate of patients with lung cancer is not beyond 14% that is drastically less than the patients suffering from cancer in other organs like breast, cervix, bladder, prostate, and colon [27] [28]. Thus, early prediction of lung cancer is very important for the appropriate treatments for decreasing the deaths. In big data, healthcare is one of the significant sources. Accurate examination of healthcare information is mostly demanded for detecting lung cancer in an early stage. Multiple researches are being designing newly to recognize lung cancer with more quality using big data. Still, there is a necessity to classification approach for improving the detection accuracy with respect to time. In addition, machine learning techniques are modelled for enhancing the detection accuracy in big data. Specifically, lung cancer is not well known that means which kind of approaches will give high detection data and which data attributes must be employed for the detection purpose.

In [29], forecasts on breast cancer survival were implemented using broad data sets using two renowned data-mining methods such as ANN and DT, and also a conventional mathematical method was used. LR. The proposed model also described a variety of lung cancers previously. For measuring the unbiased assessment of three detection models, ten-fold cross validation mechanisms were used for the performance comparison. The outcomes have proven that DT was the well performing classifier for predicting the disease with an accuracy of 93.6% on the holdout sample; ANN was standing the second-best position with an accuracy of 91.2%. Similarly, logistic regression has attained the accuracy of 89.2%. A research was done in [30] for developing detection techniques to know the survivability of prostate cancer, using SVM along with that three methods that were mentioned earlier. Here, the outputs have revealed that the singled out SVM acquired high accuracy, next to that ANN and DT attained more accuracy. In addition, ANNs, DTs and LR approaches were tested for the survival of the prostate cancer[31]. The SEER colon cancer case datasets for estimating survival rates have been contrasted with several methods[32], recognizing that NNs are better used to estimate survival rates.
Ensemble voting of three outperformed classifiers present in [33] was resulted in optimal prediction, and AU-ROC curve to colon cancer survival rate. In some researches, the survival of lung cancer patient was examined by evaluating the SEER database using machine learning algorithms, consisting of SVM, LR [34], unsupervised approaches [35], and clustering-based techniques [36]. In [37], data classification approaches were assessed for finding the chances of patients with definite indications for the growth of lung cancer. The performance of DT and NB classifiers were compared in [38], and implemented for lung cancer data acquired from SEER database. This attained approximately 90% precision in detecting the survival of patient. Ensemble voting of five DTs and meta-classifiers existing in [40] [39] was resolute for acquiring the best prediction survival rate of lung cancer regarding precision and AU-ROC curve.

IV. RESEARCH METHODOLOGY

The main intent of this proposal is to introduce a novel methodology for lung cancer prediction using the patient's health record. The proposed model involves three main phases: (a) Data Gathering, (b) Feature extraction, and (c) Prediction. Initially, the data gathering is done for collecting diverse benchmark datasets from Google datasets, which involves the attribute information of different patients in the form of heath record. Further, feature extraction will be performed using two well-performing techniques like t-Distributed Stochastic Neighbor Embedding (t-SNE) and Principal component analysis (PCA). As an innovative contribution, a novel optimized weighted feature extraction will be performed, in which the a parameter of weight computation will be optimized or tuned by the new meta- heuristic algorithm called Self Adaptive Sea Lion Optimization Algorithm (SA-SLnO) [26]. Further, the extracted features will be subjected to a deep learning algorithm termed as Recurrent Neural Network (RNN). In RNN, the number of hidden neurons will be optimized by the same SA-SLnO. The experimental results show that the proposed model proves the enhancement of the proposed model by comparing over conventional methods. The diagrammatic representation of the proposed lung cancer prediction is given in Fig. 1.

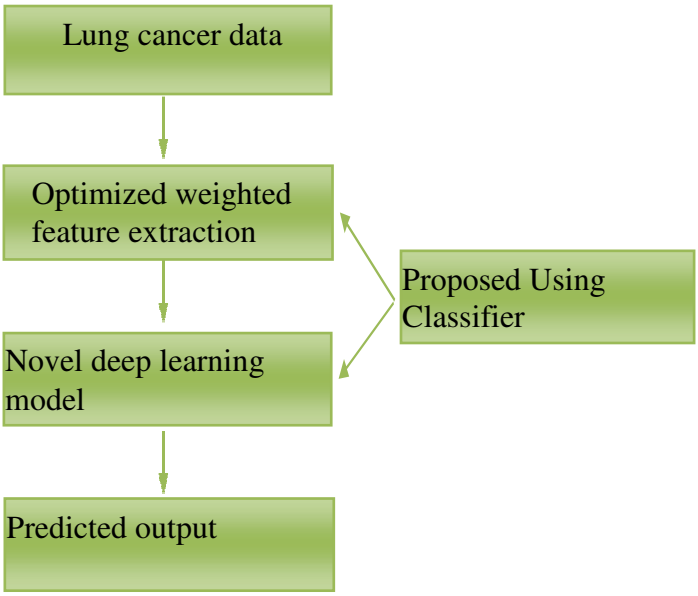

Figure 1: Block diagram of proposed lung cancer prediction

V. CONCLUSION:

In this paper we have discuss the proposed system-based classifier which able to give better accuracy on lung cancer data for detecting lung cancer. In the study the precision can be further improved with the required method of choosing functions, and an adaptive approach to other supervised learning methods.

REFERENCES

1. ChaoTan, HuiChen, and ChengyunXia, "Early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm", Journal of Pharmaceutical and Biomedical Analysis, vol.49, no.3, pp.746-752, 5 April 2009.
2. Tae-WooKim, Dong-HeeKoh, and Chung-YillPark, "Decision Tree of Occupational Lung Cancer Using Classification and Regression Analysis", Safety and Health at Work, vol.1, no.2, pp.140-148, December 2010.
3. MaciejZięba, Jakub M.Tomczak, MarekLubicz, and JerzyŚwiątek, "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients", Applied Soft Computing, vol.14, pp.99-108, January 2014.
4. WorrawatEngchuan, and Jonathan H.Chan, "Pathway activity transformation for multi-class classification of lung cancer datasets", Neurocomputing, vol.165, pp.81-89, 1 October 2015.
5. H. Azzawi, J. Hou, Y. Xiang and R. Alanni, "Lung cancer prediction from microarray data by gene expression programming," IET Systems Biology, vol. 10, no. 5, pp. 168-178, 10 2016.
6. PanayiotisPetousis, Simon X.Han, DeniseAberle, and Alex A.T.Bui, "Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network", Artificial Intelligence in Medicine, vol.72, pp.42-55, September 2016.
7. Chip M.Lynch, BehnazAbdollahi, Joshua D.Fuqua, Alexandra R.de Carlo, James A.Bartholomai, Rayeanne N.Balgemann, Victor H.van Berkel, and Hermann B.Frieboes, "Prediction of lung cancer patient survival via supervised machine learning classification techniques", International Journal of Medical Informatics, vol.108, pp.1-8, December 2017.
8. P. Petousis, A. Winter, W. Speier, D. R. Aberle, W. Hsu and A. A. T. Bui, "Using Sequential Decision Making to Improve Lung Cancer Screening Performance," IEEE Access, vol. 7, pp. 119403-119419, 2019.
9. Krishnaiah, G. Narsimha, C. Subhash, "Diagnosis of lung cancer prediction system using data mining classification techniques,"in International Journal of Computer Science and Information Technologies, Vol. 4, issue 1, pp. 39-45, December 2013.
10. T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971-987, July 2002.

11. DavidMeyer, FriedrichLeisch, and KurtHornik, "The support vector machine under test", Neurocomputing, vol. 55, no. s 1–2, pp. 169-186, September 2003.
12. L.Demidova, I.Klyueva, Y.Sokolova, N.Stepanov, and N.Tyart, "Intellectual Approaches to Improvement of the Classification Decisions Quality on the Base of the SVM Classifier", Procedia Computer Science, vol. 103, pp. 222-230, 2017.
13. N. Picco, R. A. Gatenby and A. R. A. Anderson, "Stem Cell Plasticity and Niche Dynamics in Cancer Progression," IEEE Transactions on Biomedical Engineering, vol. 64, no. 3, pp. 528- 537, March 2017.
14. Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, Park MY, and Park RW, "Development of novel breast cancer recurrence prediction model using support vector machine", J Breast Cancer, vol. 15,no. 2, pp. 230-238, 2012.
15. Paweł Krawczyk, Tomasz Kucharczyk and Kamila Wojas-Krawczyk, "Screening of Gene Mutations in Lung Cancer for
16. Qualification to Molecularly Targeted Therapies", INTECH Open Access Publisher, 2012.
17. Colquhoun, L. McHugh, E. Tulchinsky, M. Kriajevska and J. Mellon, "Combination Treatment with Ionising Radiation and Gefitinib ('Iressa', ZD1839), an Epidermal Growth Factor Receptor (EGFR) Inhibitor, Significantly Inhibits Bladder Cancer Cell Growth in vitro and in vivo," Journal of Radiation Research, vol. 48, no. 5, pp. 351-360, Sept. 2007.
18. Emmanuel Adetiba, and Oludayo O. Olugbara, "Lung Cancer Prediction Using Neural Network Ensemble with Histogram of Oriented Gradient Genomic Features", The Scientific World Journal, 2015.
19. S. S. Alahmari, D. Cherezov, D. B. Goldgof, L. O. Hall, R. J. Gillies and M. B. Schabath, "Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening," IEEE Access, vol. 6, pp. 77796-77806, 2018.
20. S. Park, S. J. Lee, E. Weiss and Y. Motai, "Intra- and Inter-Fractional Variation Prediction of Lung Tumors Using Fuzzy Deep Learning," IEEE Journal of Translational Engineering in Health and Medicine, vol. 4, pp. 1-12, 2016.
21. A. Raweh, M. Nassef and A. Badr, "A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation," IEEE Access, vol. 6, pp. 15212-15223, 2018.
22. J. Pati, "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach," IEEE Access, vol. 7, pp. 4232-4238, 2019.
23. B. Zhang et al., "Ensemble Learners of Multiple Deep CNNs for Pulmonary Nodules Classification Using CT Images," IEEE Access, vol. 7, pp. 110358-110371, 2019.
24. C. Arunkumar and S. Ramakrishnan, "Prediction of cancer using customised fuzzy rough machine learning approaches," Healthcare Technology Letters, vol. 6, no. 1, pp. 13-18, 2 2019.
25. H. Guo, U. Kruger, G. Wang, M. K. Kalra and P. Yan, "Knowledge-Based Analysis for Mortality Prediction from CT Images," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 2, pp. 457-464, Feb. 2020.
26. J. Yang, N. Li, S. Fang, K. Yu and Y. Chen, "Semantic Features Prediction for Pulmonary Nodule Diagnosis Based on Online Streaming Feature Selection," IEEE Access, vol. 7, pp. 61121-61135, 2019.
27. Raja Mohammad Taisir Masadeh, Basel A. Mahafzah, and Ahmad Abdel-Aziz Sharieh, "Sea Lion Optimization Algorithm", International Journal of Advanced Computer Science and Applications, vol.10, no.5, pp.388-395, May 2019.
28. Z.W. Huang, A. Mcwilliams, H. Lui, D. Mclean, S. Lan, H.S. Zeng, "Near-infrared Raman spectroscopy for optical diagnosis of lung cancer", Int J Cancer, vol.107, no.6, pp.1047-52, 20 Dec 2003.
29. [28]A. Jemal, F. Bray, M. M Center, J. J.Ferlay, E. Ward, and D. Forman, "CA A Cancer Journal for Clinicians, vol.61, no.2, pp.69-90, February 2011.
30. [29]D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", Artif Intell Med, vol. 34, no.2, pp. 113-127, 2005.
31. [30]Dursun Delen, "Analysis of Cancer Data: A Data Mining Approach", Expert Systems, vol. 20, no.1, pp. 100-112, 2009.
32. D. Delen, and N. Patil, "Knowledge Extraction from Prostate Cancer Data", Proceedings of the 39th Annual Hawaii International Conference on, vol.5, 2006.
33. M. Hoogendoorn, L. M. G. Moons, Mattijs E. Numans, and Robert-Jan Sips, "Utilizing Data Mining for Predictive Modeling of Colorectal Cancer Using Electronic Medical Records", International Conference on Brain Informatics and Health BIH 2014: Brain Informatics and Health, pp 132-141, 2014.
34. R. Al-Bahrani, A. Agrawal and A. Choudhary, "Colon cancer survival prediction using ensemble data mining on SEER data," 2013 IEEE International Conference on Big Data, Silicon Valley, CA, pp. 9-16, 2013.

35. C.M. Lynch, V. H. V. Berkel, and H.B. Frieboes, "Application of unsupervised analysis techniques to lung cancer patient data", PLoS One, vol. 12, no.9, 2017.
36. G. Dimitoglou, J. A. Adams, and C. M. Jim, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability", J. Comput, 2012.