Lung Cancer Prediction Using Machine Learning Algorithms

Kanchan S. Pradhan*, Dr. Priyanka Chawla*
**School of Computer Science & Engineering*
Lovely Professional University,
Phagwara, India

Abstract:

Lung cancer is a major cause of cancer-linked death all over the world. The premature detection of cancer has shown to be significantly useful to cure the disease in an efficient manner. In this paper, the information related to lung cancer, machine learning algorithms, with Internet of Things (IoT) is seen. In the literature review section, the data collected from various papers are analyzed with different machine learning algorithms are used for detecting lung cancer. The review on several machine learning algorithms in order to detect different kinds of diseases is seen in the review section. In research gaps and challenges, the drawbacks related to the existing machine learning algorithms for detecting the disease and the information related to the further development of models for detecting various diseases in an efficient manner is presented.

Keywords: Lung Cancer Detection, Machine Learning Algorithms, Literature Review, Gaps and Challenges

1. Introduction

Cancer is taken as the genetic disorder with indefinite reasons and techniques in more scenarios. Lung cancer is the significant killer disease among different types of cancer in all over the world [14]. The occurrence of lung cancer on compared to other cancers are mainly 25%, all over the world that is more than the other cancer diseases together like prostate, colorectal, and breast cancer. Although, the genetic factors related to lung cancers are still but for being understand clearly as lung cancer is a complicated genetic disease that is introduced by the consensus of several genetic alterations. This disease is categorized into two types such as, Non-Small Cell Lung Cancer (NSCLC), and Small Cell Lung Cancer (SCLC). The NSCLC is mostly seen in the patients affecting from lung cancer i.e., approximately 80%. It is also sub-divided as Large Cell Carcinoma (LCC), Adenocarcinoma (AC), and squamous cell carcinoma.

In the recent times, smart medicine is found to be an interdisciplinary by combining different mathematical methods into the medication. The key intention of this medical field is to provide personalized and invasive healthcare and medical services to the people. For example, “computer aided diagnosis and decision making for the smart medicine” is smart Chinese medicine that tries for producing the quick syndrome differentiation, and recommend treatment by implementing artificial intelligence methods together using edge/cloud computing for existing Chinese medicine medical practice [1]. Moreover, the computing method consists of big data analytics, cloud computing, IoT, and artificial intelligence [13] that is majorly contributing to the development of smart medicine. In particular, few IoT models are combined into medicine, generating medical IoT, for creating and gathering a huge amount of medical information and also represented as medical big data [15]. It provides a huge count of training instances for training “artificial intelligence models, specifically deep learning methods and deep reinforcement learning approaches”. Moreover, the renowned deep learning techniques are evaluated fruitfully for computer aided diagnosis in current health care sector [11] [12].

2. Literature Review

2.1 Related Works

In 2019, Liu *et al*. [1] have suggested many deep reinforcement learning methods, which have the probability for utilizing in detecting lung cancer precisely. Moreover, the usual kinds of lung cancer, and further, the significant features of each and every type were summarized. At last, the disadvantages were pointed out and feasible upcoming future directions of implementing deep reinforcement learning for detecting lung cancer that was likely to support the further development of smart medicine using IoT.

In 2017, Lynch *et al*. [2] have introduced numerous SVM methods are implemented for the "SEER database" to categorize the lung cancer patients regarding survival, consisting of Decision Trees (DT), linear regression, SVM, custom ensemble, and Gradient Boosting Machines (GBM). On implementing these approaches, the key data attributes such comprised of tumour size, gender, stage, age, number of primaries, and tumour grade by aiming for enabling comparison of analytical power among different approaches. The outcomes have revealed that the forecasted observations were similar to original observations for low to moderate survival times that represented the significance of information. Finally, from the results it has been confirmed that custom ensemble algorithm was superior to other algorithms in predicting the lifetime of the lung cancer patients.

In 2014, Zieba *et al*. [3] have recommended boosted Support Vector Machine (SVM) for resolving the excessive data issues. The developed model merged the advantages of making use of ensemble classifiers for irregular information as one with cost-sensitive SVMs. Moreover, the oracle-based method was introduced to extract decision rules from the boosted SVM. After that, the quality of the suggested technique was evaluated with techniques in terms of performance that dealt with uneven information. In the end, the proposed model was utilized for clinical application in predicting the life of patients affected with lung cancer.

In 2016, Petousis *et al*. [4] have developed Dynamic Bayesian Networks (DBN) and analysed for producing nearby into how longitudinal data was utilized for assisting lung cancer screening decisions. "The Low-Dose Computed Tomography (LDCT) arm of the

National Lung Screening Trial (NLST) dataset" was utilized for creating and investigating five DBNs for extreme risk individuals. By using backward construction procedure, three of them were built, and the other two were built by structure learning approaches. The entire approaches utilized the information "such as smoking status, family lung cancer history, comorbidities, exposure risk factors, demographics, and cancer history related to lung cancer and LDCT screening" The approaches were validated on training and test datasets of cancer and non-cancer patients for dealing with uneven data, and over fitting information. The results have shown that the suggested model has high inequity and prediction control with both cancer and noncancer cases.

In 2018, Shukla and Aggarwal [5] had acquired tachogram from the observed electrocardiogram of 104 lung cancer data and 30 healthy controls for extracting Heart Rate Variability (HRV) indices. The attained outcomes have recommended lowered HRV observations from Eastern Cooperative Oncology Group (ECOG) 1 to ECOG 4. The control females had more HRV measures, when compared over the control males. There was no connection between the age and the HRV measures. The prediction evaluation by Support Vector Machine (SVM) and Artificial Neural Network (ANN) has attained maximum accuracy with non-linear HRV indices as the input and therefore recommended those algorithms as a tool for prognostic significance.

In 2018, Xiao *et al.* [6] have introduced a "semi-supervised deep learning mechanism named Stacked Sparse Auto-Encoder (SSAE)-based classification" in order to predict cancer with the help of RNA-seq data. Moreover, the suggested model makes use of the greedy layer-wise pre-training and a sparsity penalty term for supporting capturing and extracting significant data from the huge size data and later categorized the samples. The experimental analysis was performed on three public RNA-seq datasets of three types of cancers and contrasted the performance of detection over multiple regularly utilized classification approaches. The outcomes have proven that the developed model was performed well for all the three datasets in different measures.

In 2016, Azzawi *et al*. [7] have suggested a Gene Expression Programming (GEP)-based approach for forecasting the lung cancer from the collected microarray data. Here, two gene selection approaches were used for extracting the important lung cancer related genes, and consequently suggested various GEP-based prediction approaches. The prediction performance was evaluated and compared with the three machine learning approaches, such as "SVM, Multi-Layer Perceptron (MLP), and radial basis function neural network". The results have shown that the proposed GEP model outperformed the other algorithms for predicting the lung cancer related problems.

In 2019, Balakrishnan *et al*. [8] have suggested a new method for defending the privacy breach, introduced with the improvement of deep learning methods named Deep Belief Network (DBN). The intrusion detection technique examined the malicious activity, which is active in the network, and the one who tried to get entry permission. The improvement of security network of DBN was compared with the benchmark Domain Generation Algorithms (DGAs) and Intrusion Detection System (IDS) methods, and the outputs were evaluated.

In 2017, Chen *et al*. [9] have simplified machine learning techniques to detect the chronic disease outbreak effectively from the gathered dataset. The tests were performed on real-life data collected from central China in 2013-2015. In order to overwhelm the problem of imperfect data, 'latent factor' was employed for reconstructing the lost data. Later, the experiments were performed on regional chronic disease of cerebral infarction. The results have shown that the suggested model was acquired the best accuracy with convergence speed that was quicker.

In 2019, Mohan *et al.* [10] have developed a novel approach, which planned in recognizing the important features by implementing machine learning algorithms that results in enhancing the accuracy for predicting the cardiovascular disease. Therefore, the prediction technique was developed using various combinations of characteristics and multiple known classification approaches. The proposed Hybrid Random Forest with Linear Model (HRFLM) has attained the best accuracy in predicting the heart disease effectively.

2.2 Discussion

The features and challenges of several machine learning algorithms for the detection of various diseases in the health care sector is depicted in Table I with its benefits and limitations.

Table 1. Features and challenges of existing cooperative spectrum sensing models

Authors [Citations]	Methodology	Features	Challenges
Liu *et al.* [1]	Deep Reinforcement Learning	• Has acquired huge success in broad range of applications. • Has best performance.	• In few complicated goal-directed works, sparse feedback poses a significant drawback.
Lynch *et al.* [2]	Custom Ensemble	• It is simple to implement, and can easily adapt to models. • Has the ability for predicting the accurate results.	• Need to improve the life time prediction accuracy of the lung cancer patients.
Zieba *et al.* [3]	Boosted SVM	• It is utilized for solving the uneven issues. • It is utilized for extracting the rules.	• In order to acquire the best classification outcomes for any problem, many parameters need to

			be set exactly.
Petousis *et al*. [4]	Dynamic Bayesian Network (DBN)	• Has high performance. • Has the ability to explain and show the contributing factors for prediction.	• It is computationally expensive.
Shukla and Aggarwal [5]	SVM and ANN	• Has high accuracy. • By using both techniques, an automated diagnostic tool is designed.	• Required more different cancer cases for optimizing the model.
Xiao *et al*. [6]	SSAE	• Has the ability to process high dimensional information. • Has high efficiency and accuracy in cancer prediction.	• Requires more training data. • It is time consuming.
Azzawi *et al*. [7]	ANN	• These solve many problems related to classification and pattern recognition. • In order to train, ANNs make use of input variables.	• It consumes more time that causes to inefficient performance.
Balakrishnan *et al*. [8]	Deep Belief Network	• Has the ability to extract features similar to human eye and brain works. • It pre-processes the data for filtering the noises that fits the data.	• Requires to improve the performance.
Chen *et al*. [9]	CNN	• Has high accuracy and convergence speed. • It extracts text characters automatically.	• Has high computational cost.
Mohan *et al*. [10]	HRFLM	• Has high accuracy in predicting the disease.	• In order to predict best performance, this model should

ISSN: 0971-1260
Vol-22-NO-30 Special Issue-December-2019

		• It utilizes all the features of feature selection without any restrictions.	be evaluated on different datasets.

3. Research Gaps and Challenges

Prediction of disease with the help of conventional disease risk system that generally consisting of Machine Learning (ML) and supervised learning algorithm that utilizes training data with the labels in order to train the models. The classification of high-risk and less-risk patients are classified into groups test sets. Yet, these methods are only precious in medical conditions and are broadly investigated. As big data is developed in healthcare and biomedical communities, accurate analysis of clinical information is advantageous in early disease recognition, community services, and patient care. If the medical information quality is imperfect, then the accuracy of the analysis is decreased. In addition, various regions show distinctive forms of assured regional diseases that might produce in fading the prediction of disease occurrence. Data created using IoT devices is massive and thus, the conventional data collection, processing appraoches, and storage might not work at this scale. Moreover, the sheer quantity of data is also utilzed for behaviours, predictions, assessment, and patterns. In addition, the heterogeneity of the data created using IoT generates the other front to the present data processing techniques. Thus, for connecting the value of IoT-generated data, fresh methods are required. ML algorithms were modernized other regions like image recognition. However, the growth from the first perceptron 53, 54 up to latest Deep CNNs was an extreme and twisted procedure. For providing major outcomes in material sciences, one essentially has not only to play for the robustness of published applications are relatively prime in original and the difficulty. They include fitting methods frequently for rigourous training sets or even implementing ML approches for work spaces, which is probably mapped out in many CPU hours. It is possible for utilizing ML techniques because an easy fitting process for less dimensional databases. However, this won't work according to their robustness and doesn't permit for reproducing the accomplishment of ML approaches in another areas. The count of dependent variables taken into consideration is the variance among the existing and the ML mechanisms in disease prediction. Moreover, in existing methodologies, only some variables are taken into consideration like weight, gender, age, etc.,. Consequently, ML is being processed on measuring tools is considered a vast count of variables that results in an improved accuracy related to health care sector information. An assortment of applications like fraud detection, malware detection, computer vision, authentication, speech recognition, and bio-informatics employ ML algorithms. In the recent researches, there are many benefits with the ML algorithms in detecting various kinds of diseases effectively, but there are few disadvantages with ML techniques like it is majorly apprehensive with supervised learning. Similarly, ML is leverage in IoT for offering quick services for lung cancer prediction.

4. Conclusion

This paper has presented the information regarding the prediction of lung cancer disease using machine learning algorithms linked with IoT. The review regarding the lung cancer disease and the different machine learning algorithms were described in literature review. The advantages and disadvantages with the existing ML models were tabulated and described the future gaps for the existing challenges.

References

ISSN: 0971-1260
Vol-22-NO-30 Special Issue-December-2019

[1] Z. Liu, C. Yao, H. Yu, and T. Wu, "Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things", Future Generation Computer Systems, vol. 97, pp. 1-9, 2019.

[2] C. M. Lynch, B. Abdollahi, J. D. Fuqua, A. R. de Carlo, J. A. Bartholomai, R. N. Balgemann, V. H. van Berkel, H. B. Frieboes, "Prediction of lung cancer patient survival via supervised machine learning classification techniques", International Journal of Medical Informatics, vol. 108, pp. 1-8, 2017.

[3] M. Zięba, J. M. Tomczak, M. Lubicz, and J. Świątek, "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] P. Petousis, S. X. Han, D. Aberle, and A. 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.

[5] R. S. Shukla, and Y. Aggarwal, "Nonlinear Heart Rate Variability based artificial intelligence in lung cancer prediction", Journal of Applied Biomedicine, vol.16, no.2, pp.145-155, May 2018.

[6] Y. Xiao, J. Wu, Z. liLin, and X. Zhao, "A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data", Computer Methods and Programs in Biomedicine, vol.166, pp.99-105, November 2018.

[7] 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.

[8] Nagaraj Balakrishnan, Arunkumar Rajendran, Danilo Pelusi, and Vijayakumar Ponnusamy, "Deep Belief Network enhanced Intrusion Detection System to Prevent Security Breach in the Internet of Things", Internet of Things, Available online 17 September 2019.

[9] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," IEEE Access, vol. 5, pp. 8869-8879, 2017.

[10] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019.

[11] Q. Zhang, C. Bai, Z. Chen, P. Li, S. Wang, and H. Gao, "Smart Chinese medicine for hypertension treatment with a deep learning model", Journal of Network and Computer Applications, vol.129, pp.1-8, 1 March 2019.

[12] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep Learning for Health Informatics," IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 4-21, Jan. 2017.

[13] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang and W. Zhao, "A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications," IEEE Internet of Things Journal, vol. 4, no. 5, pp. 1125-1142, Oct. 2017.

[14] W. Engchuan, and J. H.Chan, "Pathway activity transformation for multi-class classification of lung cancer datasets", Neurocomputing, vol.165, pp.81-89, 1 October 2015.

[15] D. V. Dimitrov, "Medical Internet of Things and Big Data in Healthcare", Healthc Inform Res, vol. 22, no.3, pp. 156–163, Jul 2016.

