Deep Learning Technique for Detecting NSCLC

Bhargav Hegde, Dayananda P, Mahesh Hegde, Chetan C

Abstract**:** ***The lung cancer is one of the major cancers in the world. In lung cancer we have two main types. They are small cell lung cancer and non-small cell lung cancer. In this paper we mainly concentrated on the detection of non-small cell lung cancer. There are several types in NSCLC and we have several stages in NSCLC. The flow of proposed paper consist the following steps: (1) Background: Here we describe the different types of lung cancer and mainly about NSCLC; (2) Methods: To find the NSCLC, we are using the Recurrent Neural Network (RNN); (3) Results: After the training and prediction of the model, we will get the final result as weather the given patient suffering from NSCLC or not; and (4) Conclusions: The given model is working for all the possible datasets and the training accuracy is 88%. The accuracy of the model is mainly depends on the epoch value. For ideal epoch value the accuracy of the model is high.***

Dataset: The datasets are taken from the NCBI website. We have used the nucleotide datasets of the NCBI website. The datasets are open source and easily accessible. We have used the DNA sequence data of the human genome data. All the NSCLC patients data are taken as positive data and human reference gene data are taken as negative data.

Keywords **:** ***Non-small cell lung cancer (NSCLC); Recurrent Neural Network (RNN); NCBI website.***

I. INTRODUCTION

In recent years, there has been an immense discussion about how to cure the lung cancer in early stages. It needs well engagement with the patient. Patient engagement is consider as a high-quality object in practice and physical condition concern organizations. Persistent commitment is a vital methodology for accomplishing the point of medicinal services, improving the patient experience. Offering significance to quiet commitment can improve productivity, diminish out-movement and lessen the general expense of patient consideration.

Now a day we can see two main types of lung cancer: SCLC and NSCLC. These 2 types are treated in a different way. Here are some details about NSCLC. NSCLC starts when fine lung cells alter and grows uncontrollable, forms an accumulation in lung. This accumulation is called as lung nodule and it can start at any part of the lung. These lung nodules can drop the cells to the blood flow. The lung nodule is in small size, structured wise it is bean that helps to battle disease. These nodes are positioned in the lungs and any other place in the human body. These nodes will stream to the center of the chest, so lung cancer frequently infects the center of the chest first. The travelling of the cancer cell around the body is called as metastasis. NSCLC starts from the weak cells of the lung. It can also be explain based on the nature of these weak cells of the lung where the cancer begins. There are three main types of the non-small cell lung cancer. They are adenocarcinomas, sqaumous cell carcinomas and the large cell carcinomas These types are essential for oncologists to differentiate among lung carcinomas that start in the sqaumous cells and lung carcinomas that start in other cells. Using these details oncologists finds the treatment methods. In this project, we try to predict the non-small cell lung cancer of the given patient

II. RELATED WORK

Here we discuss about the different paper and their work towards proposed topic. This section contains the survey of the several papers which are already in the different journals. The authors [1] mainly concentrated about the prediction of enhancer-promoter interaction (EPI) in genome. The EPI has the important role in transcriptional regulation.

Here author used the genome data and applied that data to convolutional neural network (CNN) and with recurrent neural network (RNN) to have highest performance from the given dataset.

The authors [2] used the DNA sequence data and using encoder to change the sequence data to encode array data. And applied this data to the CNN algorithm to get the maximum performance from the given dataset. The author used different methods of CNN, like DeepBind, DeepSEA and reverse complement etc... After all methods the result of the CNN methods for DNA sequence data increases by 5.7% which is high compare to other normal methods. The authors [3] used the RNA-sequence data of the pan cancer patients. They applied several methods on 33 patient’s gene expression data. After that they applied these data to CNN method and got 95% result which is much more than GA/KNN method. They are the first to apply CNN method for pan-cancer atlas classification.

Manuscript published on 30 September 2019
* Correspondence Author

Bhargav Hegde* , JSS Academy of Technical Education, Bengaluru, India. Email: bhargavhegde660@gmail.com

Mahesh Hegde, JSS Academy of Technical Education, Bengaluru, India. Email: maheshhegde113@gmail.com

Chetan C, department, , JSS Academy of Technical Education, Bengaluru, India. Email: chetanchannappagol@gmail.com

Dayananda P*, department, , JSS Academy of Technical Education, Bengaluru, India. Email: dayanandap@gmail.com

The authors [4] used different machine learning and deep learning algorithm to detect and diagnose several diseases, like Rheumatoid Arthritis, Cancer, Lung Diseases, Heart Diseases, Diabetic Retinopathy, Hepatitis Disease, Alzheimer’s disease, Liver Disease, Dengue Disease and Parkinson Disease.

Here they used Artificial Neural Network to find the disease of the patient. The authors [5] mainly concentrated about the lung cancer, breast cancer and melanoma. Here they applied normalized Gradient Descent and an Artificial Neural Network for breast cancer prediction, artificial neural network coupled with image processing to detect melanoma and Support Vector Machine algorithm incorporating image processing to detect lung cancer. They got the good 90-95% Result for these methods.

The author [6] presents a study of feature selection methods effect, using a filter approach, on the accuracy and error of supervised classification of cancer. The classification is done using the KNN and SVM method. Here we can find that SVM method has the highest accuracy. The authors [7] proposed the different classifier to colon cancer gene expression data set. In paper [8], a work on the microarray data of the human genome sequence of the colon cancer and leukemia patient. The implementation and testing of the datasets are done using four different methods; they are KNN, RF, SVM and Neural network. The comparison study of these methods results is done and the maximum result is considered. The authors [9] studies the gene expression microarray data of the lung cancer is taken, and prior knowledge method is applied to these datasets. And there is 98- 99% is found is testing dataset. The authors [10] used the non-invasive diagnostic method to find the DNA –sequence from the sample of the non-small cell lung cancer patient. The authors [11] has exclusively reviewed various classification by machine learning algorithms for analysis of human genome data.

III. DATA DESCRIPTION

We used the datasets from the NCBI website. The datasets are in .fasta or .gb format. We have converted this data into .csv format. All the patient data are consider as positive data and human reference gene data are taken as negative data. We have taken human reference gene data hg38, in that data we have considered the KRAS gene data. The data sequence is in ATGC format and targets are either 0 or 1.

IV. METHODS

Deep learning algorithm run information through a few "layers" of neural system calculations, every one of which passes a rearranged portrayal of the information to the following layer. Most AI calculations function admirably on datasets that have up to a couple of hundred highlights, or sections. Nonetheless, an unstructured dataset like one from a picture has such countless that this procedure winds up bulky or totally unfeasible. A solitary 800-by-1000-pixel picture in RGB shading has 2,400,000 highlights – extremely numerous for customary AI calculations to deal with, which attempt and take in all the data on the double. Deep learning algorithm adapt logically progressively about the picture as it experiences each neural system layer. Early layers figure out how to recognize low-level highlights like edges, and consequent layers join highlights from prior layers into a progressively all-encompassing portrayal. For instance, a center layer may recognize edges to distinguish portions of an article in the photograph like a leg or a branch, while a profound layer will almost certainly identify the full item like a canine or a tree.

Recurrent Neural Network: Recurrent Neural Networks (RNN) are a stunning and solid kind of neural frameworks and have a spot with the most reassuring computations out there right now since they are the primary ones with an inward memory. RNN's are commonly old, similarly as other significant learning estimations. They were at first made during the 1980s, in any case, can simply show their certified potential since a few years, because of the extension unavailable computational power, the enormous proportions of data that we have nowadays and the development of LSTM during the 1990s. In light of their inside memory, RNN's can remember huge things about the data they got, which engages them to be definite in envisioning what's coming straightaway. This is the inspiration driving why they are the favored figuring for progressive data like time plan, talk, content, cash related data, sound, video, atmosphere and considerably more since they can shape and significantly further cognizance of a gathering and its exceptional circumstance, diverged from various counts. The figure 4.1 will show the RNN-LSTM model for proposed training

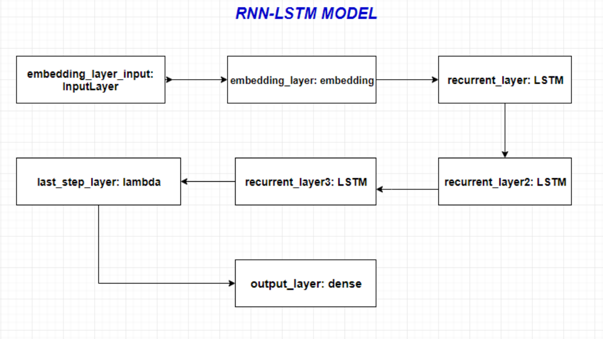

Figure 1: RNN-LSTM model

V. RESULTS

We have applied several data for the prediction of the non-small cell lung cancer. The datasets are taken from the NCBI website. We took the non-small cell patients and human reference genome data of the KRAS gene from the NCBI website for training and testing purpose. For prediction purpose we took the following data from the NCBI website: JB834943.1, JB834944.1, JB834945.1, JB834946.1, and JB834947.1. We also used the NVIDIA GPU for the training of data. Since our laptop does not support for a large number of epochs, we used GPU for training with large epochs.

The below table 5.1 shows the result of the training score and accuracy for different epochs and with or without GPU.

Table- I Results of the training

Epoch value	**Score (%)**	**Accuracy (%)**	**Time taken/steps**
(with GPU)			
100	45.51	**83.00**	4ms
150	35.95	**88.00**	4ms
200	100	**79.66**	4ms
(without GPU)			
5	43.34	**83.99**	50ms
10	34.62	**85.99**	52ms
20	35.92	**87.99**	55ms

From the above table 5.1, we can find the perfect epoch value for the training. If the epoch value is too less the model will under fit or if the epoch value is too large the model will over fit. For 200 epochs the model will over fit and the accuracy will be very less. The time taken for per epoch is very less when we are using the GPU and it will increase to above 50 when we are not using the GPU. We will get highest accuracy, when we have the epoch value near 150. So we will highly recommend GPU for the training to get highest accuracy.

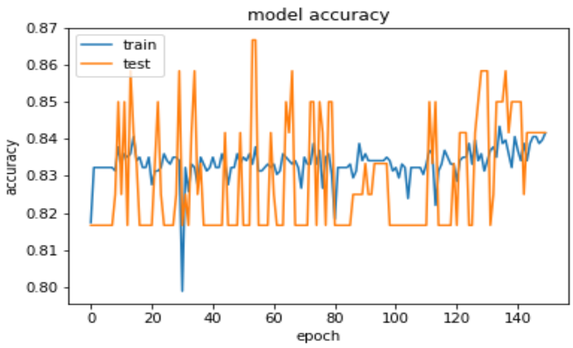

Figure 2: Accuracy graph for epoch 150 with GPU

V1. CONCLUSION

quantity of non-small cell lung cancer survivors keeps on expanding as a result of both advances in early location and treatment and maturing and development of population. For general prosperity gatherings to all the more likely serve these survivors, the American Cancer Society, and National Cancer Institute cooperate to measure the nature of present and future non-little cell lung disease survivors using data from the Surveillance, Epidemiology and End results from harm vaults. In spite of the fact that there are a developing number of instruments that can help patients, parental figures, and oncologist in exploring the different periods of malignant growth survivors. As per the studies and observation made, it is very clear that RNN-LSTM algorithm perform better and better as the variety of the data set and instances of the dataset increases.

ACKNOWLEDGMENT

We would like to thank JSSATE, Bangalore for their kind support to carry out this work . We would like to specially thank NVIDIA by providing GPU Quadro P5000 for carrying out the project.

REFERENCES

1. Zhong Zhuang, Xiaotong, Wei Pan. "A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data", Bioinformatics Oxford University Press, 2019.
2. Zhen Cao and Shihua Zhang. "Simple tricks of convolutional neural network architecture improve DNA-protein binding prediction", Bioinformatics Oxford University Press, 2018.
3. Boyu Lyu and Anamul Haque. "Deep learning based tumor type classification using gene expression data", bioRxiv, 2018.
4. Dinu A.J, Ganesan R, Felix Joseph, Balaji V. "A study on deep learning algorithm for diagnosis of diseases", International Journal of Applied Engineering Research (ISSN0973- 4562), 2017.
5. Mohnish Chakravarthi. "A comprehensive study on the applications of machine learning for diagnosis of cancer", arXiv, 2015.
6. Sara Haddou Bouazza, Nezha Hamdi, Khalid Auhmani. "Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers", 2015 IEEE, 2015.
7. Sara Tarek, Reda Abd Elwadab, Mahmoud Shoman. “Gene expression-based cancer classification”, Egyptian Informatics Journal, 2016.
8. Jing sun, Kalpdrum and Chakresh Kumar Jain. “Improved microarray data analysis using feature selection methods with machine learning methods”, IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016.
9. Peng Guan, Desheng Huang, Miao He,and baosen Zhou. “Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method”, Journal of experimental and clinical cancer research, 2009.
10. Vamsidhar Velcheti, Nathan A. Pennell. “Non-invasive diagnosticplatforms in management of non-small cell lung cancer: opportunities andchallenges”, Annals of Translational Medicine, 2017.
11. Neelambika. B. Hiremath, & Dayananda P. (Jan 2019). Machine Learning Techniques for Analysis of Human Genome Data. International Journal of Smart Education and Urban Society, volume 10(1),

AUTHORS PROFILE

Bhargav Hegde, JSS Academy of Technical Education, Bangalore India, Research area of Interest in Machine Learning.

Dr,Dayananda P, JSS Academy of Technical Education, Bangalore India Research area of Interest in Machine Learning. Currently working as Associate Prof and HOD in Dept of ISE.

Mahesh Hegde, JSS Academy of Technical Education, Bangalore India, Research area of Interest in Machine Learning.

Chetan C , JSS Academy of Technical Education, Bangalore India, Research area of Interest in Machine Learning.