Hybrid K-Mir Algorithm to Predict Type of Lung Cancer Among Stoicism

Venkata Ramana N, Chandra Sekhar Kolli, Ravi Kumar T, P Nagesh

Abstract: ***Health care is the maintenance of health via the prevention, diagnosis, and treatment of disease. The disease that persists over a long period of time is known as Chronic Disease. Chronic diseases may create additional activity restrictions. Common chronic conditions include lung disease, heart stroke, cancer, obesity, and diabetes. Chronic diseases usually show no symptoms and hence not diagnosed in advance. Hence it is necessary to predict the patient-specific chronic diseases in early stage for effective prevention. Machine learning being the vital component of Data Analytics that facilitates the medical domain for malignancy predictions. Patients suffering from misdiagnosed and undiagnosed chronic diseases can be easily recognized with the help of these hospital systems. These systems enable the doctors to take precautionary measures and thereby minimizing the chances of a patient being affected. A new hybrid K-MLR framework using K-means and Multiple Linear Regression has been proposed for diagnosing the type of lung cancer among the patients. As most of the real datasets are high-dimensional, this hybrid framework uses K-Means clustering algorithm that eliminates the noise from the image based dataset at the initial stage. Afterward to reduce the dimensionality it detects the features of nodules in 3D lung CT scans and partitions the data to form the clusters. Finally it reads the new patient data with malignant nodules to predict the type of associated cancer based on the intensity of the nodule features extracted from each CT scan report using Multiple Linear Regression Analysis. Clustering prior to classification makes the hybrid approach beneficial.***

Index Terms: ***Lung cancer, pulmonary nodules, CT scan, Prediction, K-means, and Regression***

I. INTRODUCTION

Lung cancer is identified as the primary cause of cancer deaths in the United States. Approximately 225,000 people are suffering from lung cancer every year and the health care cost also reached to $12 billion this year in United States. Compared over the deaths caused because of colon, prostate and pancreas and breast cancer jointly, the death rate of lung cancer is pretty high. Recent researches concluded that the death rate can be considerably reduced with the help of chest CT (computed tomography) scan. The nodules in the lung can be recognized utilizing a chest CT scan. A pulmonary nodule is a single round or oval growth in the lung that is also known as a coin lesion. Nodule causes no symptoms but usually
found during a chest X-ray or CT. Based on the imaging characteristics, the probability that a nodule is malignant is assessed. This insisted the researchers to build various image analysis research tools like nodule classifiers and nodule detectors to aid the radiologists in generating accurate assessments of the cancer risk. It is necessary to anticipate efficient algorithm that can detect whether the lesions in the lungs are cancerous and give patients the best chance to survive and recover with a proper diagnosis to prevent the cancer [1]. In this work, a new machine learning algorithm has been proposed to predict the risk with lung cancer by utilizing CT scan reports. The proposed structure detects a nodule in prior considering its features as well as image area around the nodule. The proposed algorithm proved to perform better in terms of accuracy, complexity and error rate.

II. LITERATURE REVIEW

Some of the research studies on lung cancer includes the following:
Wang et al. deliberated malignant nodules in 185 patients, 171 benign and texture analysis on NCE CT in 2 patients and observed that there is a differentiation amid benign and malignant nodules with respect to entropy and sum entropy ($P<0.05$) [2].
Gibbs et al. presented a recent study in which he utilized MRI texture analysis to exhibit high entropy in malignant lesions. The experimental results showed that this technique succeeded in distinguishing a malignant from benign lesions with 0.81±0.07 accuracy [3].
Cavouras et al. applied texture analysis merged with CT density matrix analysis on 51 nodules confirmed by NCE CT and achieved 90.2% accurate results in differentiating amid malignant from benign lung lumps [4].
McNitt-Gray et al. used NCE CT to analyze 32 lung nodules that includes 14 benign and 14 malignant. This technique employed four textural features along with sum entropy to yield an area under the ROC curve (Az) of 0.992 [5].
Dujardin et al. did a lot of research on borderline tumors and benign and achieved same kind of outputs. The authors assessed that the distinction may present because of unclear textural differences amid malignant and benign lesions [6].
Chae et al. presented the research done on texture analysis. This paper proposed a technique which excludes CE CT in differentiating amid invasive and pre-invasive lung adeno carcinomas [7]. Ganeshan et al. studied texture analysis technique that includes NE CT alone and proposed technique to assess the association of glucose metabolism with non-small cell lung cancer [8]. The same authors came up with a latest publication in which they considered the same patients and did texture analysis on both NCE CT as well as CE. This novel technique also successfully assessed the relationship between texture in non-small cell lung cancer and histological markers [9].

Revised Manuscript Received on 8 February 2019.

Venkata Ramana N Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
Chandra Sekhar Kolli, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
Ravi Kumar T, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
P.Nagesh, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

V. Krishnaiah et.al., proposed an automatic technique for tumor delineation by using the single frame of images of the PET lung. SVM and K-nearest neighbor classifiers are used to discover the performance of the topography. This technique enables the physicians in minimizing the radiotherapy treatment and to forecast lung tumors [10].

Zakaria Suliman Zubi proposed various segmentation and improvement techniques to get accurate results. The proposed method uses MATLAB to generate each measure. The methods such as image pre-processing, processing, feature extraction and segmentation were discussed in image processing techniques in detail. To improve the quality of the image the techniques like Fast Fourier transform methods, auto improvement and Associate Gabor filter are being used. In the segmentation stage, the Watershed and Thresholding Segmentation is castoff and the contrast has been made [11].

K. Balachandran proposed a novel method named FIS (Fuzzy Inference System) distinguish between benign, malicious and advanced lung nodes. FIS uses artificial neural networks to make these differentiations. The author also suggests to perform linear filtering as a pre-processing step to eradicate the disturbances and to improve the quality of the images [12].

Tariq et.al presented a neuro-fuzzy classifier to identify the lung nodules by assessing the CT scan images. The technique includes feature classification, extraction, lung division and enrichment. Threshold segmentation is being used to extract the nodules from the image and to remove contextual smear. This method enables in identifying even the tiny nodules in lungs at an early stage [13].

Ada et al., introduced a technique to identify tumors in lungs based on the images. This tool also enables the physicians to easily differentiate amid normal and abnormal lungs as well as to predict the probability of the survival of a sufferer so that the cancer patients' lifetime can be increased [14].

Dechang Chen et al., proposed a novel clustering algorithm EACCD. It works in two steps. At first stage, partitioning among medoids is considered to calculate a different measure. Secondly, the dissimilarity found is analyzed to attain patients' clusters. These patients' cluster arrangement is a basis of a predictive system. [15].

III. RESEARCH METHODOLOGY

A data set including thousands of high-quality lung scans has been provided by National Cancer Institute. A Lung nodule can either be a cancerous or noncancerous. A noncancerous nodule is known as benign where as a cancerous lung nodule is considered as a malignant. Lung nodules are usually about 5 to 30mm in size. Larger lung nodules above 30 mm have a higher chance of being cancerous.

HU, Hounsfield unit is calculated by considering regions of interest (ROI). ROI is usually drawn around the lesion on every image of the CT scan. Hounsfield measurement of a pulmonary nodule can be used to reliably identify benign or malignant nodules [16]. Based on the Lesion diameter, tumor size is characterized and intensity of cancer is identified accordingly. Pathological invasiveness for pulmonary nodules is predicted by considering both Tumor size as well as CT attenuation into account.

The stages of cancer are labeled as Adeno carcinoma, squamous cell carcinoma, Metastatic carcinoma, bronchoalveolar carcinoma and Mucoepidermoid carcinoma.

A. Proposed Methodology:

1. Capture the cross-sectional images of the lung tissue and detect the abnormalities.
2. Preprocess the dataset to eliminate the noise and transform into consistent unlabeled data.
3. Extract the relevant features like lungareapx, lungareamm2, lungvolumefraction, lungmeanhu, lungpd95hu, and lungpd05huthat help to identify the disease with which the patient is suffering.
4. Using segmentation technique like K-mean clustering, detect the different types of lung diseases. (Adenocarcinoma, squamous cell carcinoma, Metastatic carcinoma, bronchoalveolar carcinoma and Mucoepidermoid carcinoma)
5. Classify the records of the patient test data to predict the dependent variable, type of disease based on the extracted features or say, independent variables using Multiple Linear Regression Analysis.

Architecture of the Proposed System

B. K-MLR(K-Means and Multiple Linear Regression) Algorithm:

Step1: Initial guesses are made for the means m1, m2, mk
Step2: Until there are no changes in any mean
Step3: Samples are classified into clusters utilizing the estimated means
Step4: For i from 1 to k
Replace mi with the mean of all of the samples for cluster
end for
Step 5: $y=\beta 0 + \beta 1\ x1+ \beta 2x2+.....+ \beta ixn$ for i = 1…n.
Step 6: Make Predictions
end until.

IV. EXPERIMENTAL RESULTS

C# language built on .NET framework; WinForms, Visual Studio 2013- an IDE and SQL Server, a SQL-based relational database management system and a graphical GUI technology is used to implement the system.

National Cancer Institute provided a real time lung cancer dataset. K-means Clustering algorithm has been applied on dataset of 267 patient records with six appropriate characteristics of malignant solitary pulmonary nodules to create accurate and distinct clusters of patients with chronic lung disease.

lungareapx	lungareamm2	lungvolumefraction	lungmeanhu	lungpd95hu	lungpd05hu
57342	40318.59	0.218742	-644.751	-368	-798

95955	67468.36	0.366039	-720.043	-544	-847
60782	42737.34	0.231865	-616.723	-294	-777
52852	37161.56	0.201614	-664.036	-267.55	-851
81260	57135.94	0.309982	-675.348	-295	-842
52456	36883.13	0.200104	-596.439	-356	-749
59331	41717.11	0.22633	-625.406	-392	-777
75752	53263.13	0.288971	-641.878	-356	-813
75298	52943.91	0.287239	-675.118	-340	-841

Table1: Sample Lung Cancer Dataset with six appropriate features

lung area px	lung area mm 2	Lung volume fraction	lung mea nhu	lung pd95 hu	lung pd05 hu	Cluster
57342	40318.59	0.218742	-644.751	-368	-798	4
95955	67468.36	0.366039	-720.043	-544	-847	3
60782	42737.34	0.231865	-616.723	-294	-777	4
52852	37161.56	0.201614	-664.036	-267.55	-851	1
81260	57135.94	0.309982	-675.348	-295	-842	3
52456	36883.13	0.200104	-596.439	-356	-749	4
59331	41717.11	0.22633	-625.406	-392	-777	4
75752	53263.13	0.288971	-641.878	-356	-813	5
75298	52943.91	0.287239	-675.118	-340	-841	5

Table2: Result of K-means Clustering with 5 clusters each of size 41, 55, 95, 29, 47

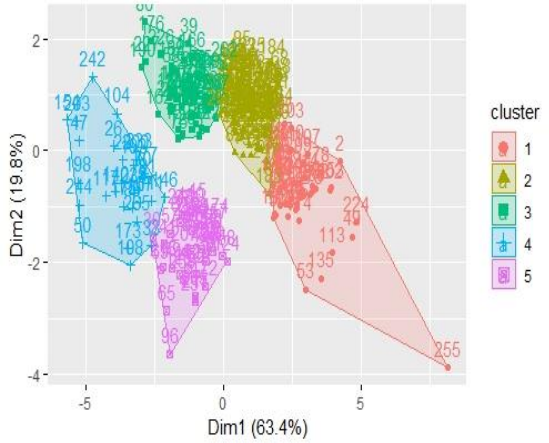

Figure1: Plotting of cluster data by K-Means Algorithm

Disease Name	Cluster	Record Count
Adenocarcinoma	1	47
Squamous cell carcinoma	2	95
Bronchoalveolar carcinoma	3	38
Meta stastic carcinoma	4	53
Mucopidermoid carcinoma	5	29

Table 3: Specification of Diseases are specified based on the six independent variables

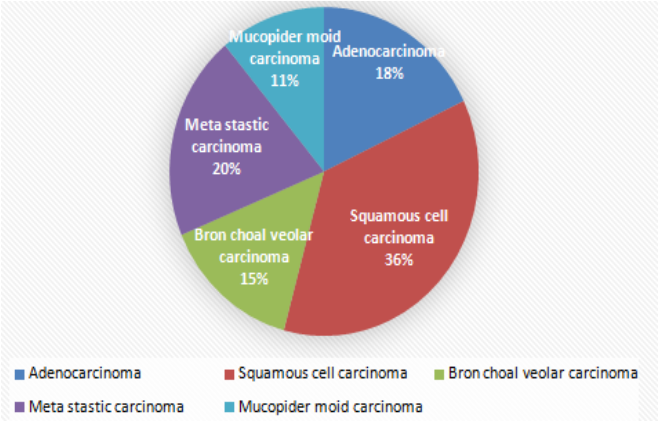

Figure2: Patient Ratio based on Lung Disease Type

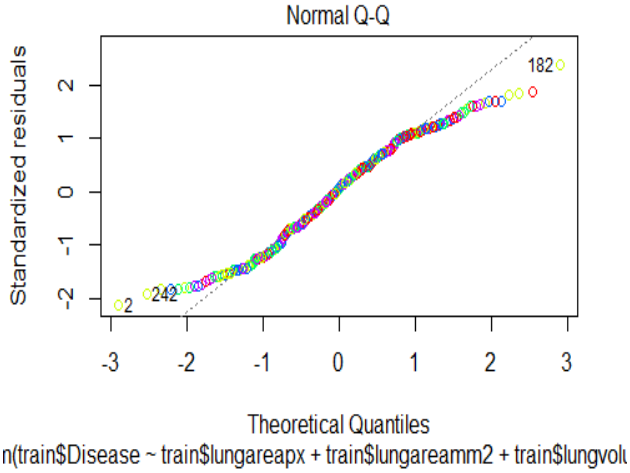

Figure 3: Q-Q Plot of standardEized residuals vs theoretical quantities

Algorithm	No Of Records	Time in ms	Space in Kb
kNN	267	4256	8955
MLR	267	3456	6895
KMLR	267	1218	3458

Table 4: Comparison of algorithms with the proposed one in terms of time and space complexities

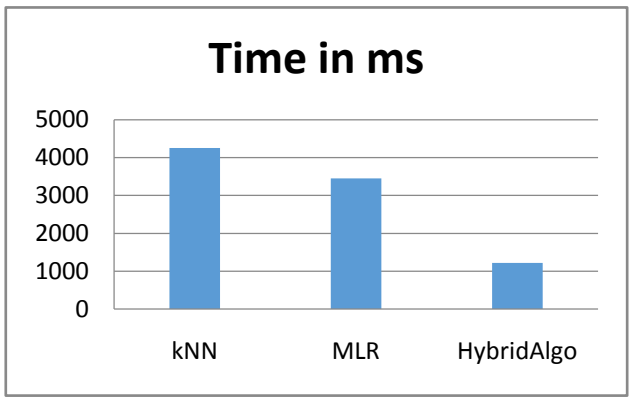

Figure 4: Comparison graph showing the time complexities of existing and proposed algorithm

Figure 5: Comparison graph showing the space complexities (kb) of existing and proposed algorithm

Algorithm	No Of Records	Correct Instances	Incorrect Instances	Prediction Accuracy
kNN	267	245	22	91.27%
MLR	267	236	31	88.36
HybridAlgo	267	256	11	95.88

Table 5: Correct and Incorrect classified instances of different algorithms

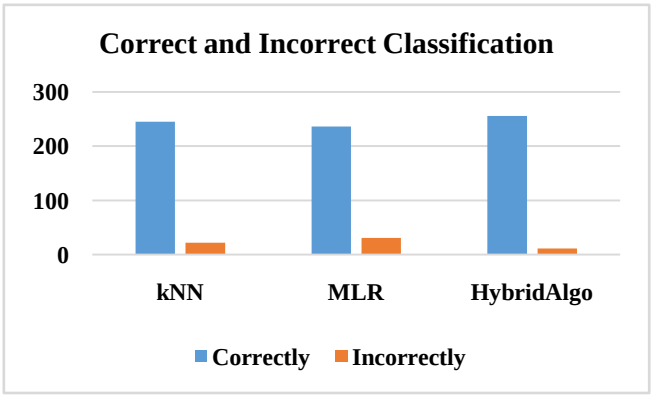

Figure 6: Correct and incorrect classified instances by kNN, MLR with respect to KMLR Algorithm

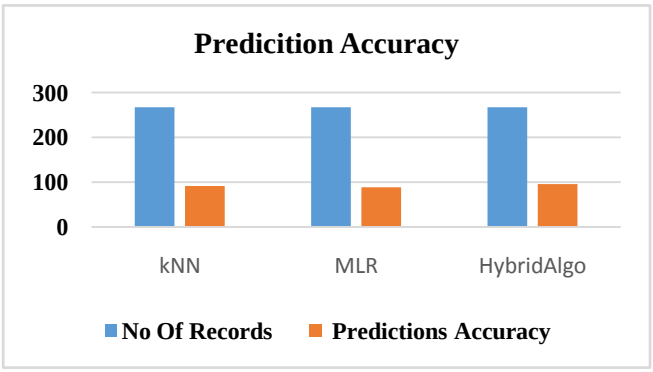

Figure 7: Comparison graph showing the accurate prediction of proposed K-MLR Algorithm

V. CONCLUSION

Today one of the most deadly diseases in the world is Cancer. In India Lung Cancer is identified as the most fatal disease with highest probability of occurrence and death rate. It is very difficult to predict the occurrence of lung cancer as it bases on multiple diagnostic attributes which are very difficult to analyze [17]. This paper presents a novel method to discover the stage of malignant nodules. This paper concentrates on developing a new technique using which clusters can be generated with high accuracy and to assess the stage of the lung cancer patients by taking the diagnostic features from the CT scan reports in to consideration. The experimental results proved that this technique succeeded in generating highest performance ratio over other existing algorithms with respect to accuracy, complexity and error rate.

REFERENCES

1. Rubin, G. D. (2015). Lung nodule and cancer detection in CT screening. Journal of thoracic imaging, 30(2), 130.
2. Wang H, Guo XH, Jia ZW, Li HK, Liang ZG, Li KC, He Q. Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image. Eur J Radiol 2010;74:124-9.
3. Gibbs P, Turnbull LW. Textural analysis of contrast-enhanced MR images of the breast. Magn Reson Med 2003;50:92-8.
4. Cavouras D, Prassopoulos P, Pantelidis N. Image analysis methods for solitary pulmonary nodule characterization by computed tomography. Eur J Radiol 1992;14:169-72.
5. McNitt-Gray MF, Wyckoff N, Sayre JW, Goldin JG, Aberle DR. The effects of co-occurrence matrix based texture parameters on the classification of solitary pulmonary nodules imaged on computed tomography. Comput Med Imaging Graph 1999;23:339-48.
6. Dujardin M, Gibbs P, Turnbull LW. Texture analysis of 3T high resolution T2 weighted images in ovarian cystadenoma versus borderline tumor. Proc Intl Soc Magn Reson Med 2014;22:2218. Available online: http://cds.ismrm.org/protected/14MPresentations/abstracts/2218.pdf
7. Chae HD, Park CM, Park SJ, Lee SM, Kim KG, Goo JM. Computerized texture analysis of persistent part-solid ground-glass nodules: differentiation of preinvasive lesions from invasive pulmonary adenocarcinomas. Radiology 2014;273:285-93.
8. Ganeshan B, Abaleke S, Young RC, Chatwin CR, Miles KA. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging 2010;10:137-43.

9. Ganeshan B, Goh V, Mandeville HC, Ng QS, Hoskin PJ, Miles KA. Non-small cell lung cancer: histopathologic correlates for texture parameters at CT. Radiology 2013;266:326-36.
10. V.Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" International Journal of Computer Science and Information Technologies, Vol. 4 (1), 39 – 45 www.ijcsit.Com ISSN: 0975-9646, 2013.
11. Zakaria Suliman zubi "Improves Treatment Programs of Lung Cancer using Data Mining Techniques" Journal of Software Engineering and Applications, 7, 69-77, February 2014.
12. K. Balachandran "Classifiers based Approach for PreDiagnosis of Lung Cancer Disease" International Journal of Computer Applications® (IJCA) (0975 – 8887), proceedings on National Conference on Emerging Trends in Information & Communication Technology (NCETICT 2013).
13. Anam Tariq, M. Usman Akram and M. Younus Javed, "Lung Nodule Detection in CT Images using Neuro Fuzzy Classifier", Fourth International Workshop on Computational Intelligence in Medical Imaging (CIMI), pp:49-53, 2013.
14. Ada R. Wolfsen, William D. Odell, ProACTH: Use for early detection of lung cancer, The American Journal of Medicine, Volume 66, Issue 5, Pages 765–772, May 1979.
15. Dechang Chen "Developing Prognostic Systems of Cancer Patients by Ensemble Clustering" Hindawi publishing corporation, Journal of Biomedicine and Biotechnology Volume, Article Id 632786, 2009.
16. Vesal, S., Ravikumar, N., Ellman, S., & Maier, A. (2018). Comparative Analysis of Unsupervised Algorithms for Breast MRI Lesion Segmentation. In Bildverarbeitung für die Medizin 2018 (pp. 257-262). Springer Vieweg, Berlin, Heidelberg.
17. Gao, X., Chu, C., Li, Y., Lu, P., Wang, W., Liu, W., & Yu, L. (2015). The method and efficacy of support vector machine classifiers based on texture features and multi-resolution histogram from 18F-FDG PET-CT images for the evaluation of mediastinal lymph nodes in patients with lung cancer. European journal of radiology, 84(2), 312-317.
18. Soni Lanka., Madhavi M. R., Abusahmin, B.S., Puvvada, N., Lakshminarayana, V., (2017), "Predictive data mining techniques for management of high dimensional big-data". Journal of Industrial Pollution Control vol 33, pp 1430-1436.
19. Venkata Ramana N , Seravana Kumar P. V. M , Puvvada Nagesh ." Analytic architecture to overcome real time traffic control as an intelligent transportation system using big data". *International Journal of Engineering & Technology, 7 (2.18) (2018) 7-11*
20. N. VenkataRamana , Puvvada Nagesh , Seravana Kumar P. V. M , U Vignesh "IoT Based Scientific design to conquer constant movement control as a canny transportation framework utilizing huge information available in Cloud Networks ". Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 07-Special Issue, 2018
21. Venkata Ramana N., Nagesh P., Lanka S., Karri R.R. (2019), "Big Data Analytics and IoT Gadgets for Tech Savvy Cities". In: Omar S., Haji Suhaili W., Phon-Amnuaisuk S. (eds) Computational Intelligence in Information Systems. CIIS 2018. Advances in Intelligent Systems and Computing, vol 888. pp 131-144, Springer Nature.
22. U. Vignesh, Sivakumar, N. Venkata Ramana "Survey and implementation on classification algorithms with approach on the environment". International Journal of Engineering & Technology, 7 (2.33) (2018) 438-440
23. Soni Lanka., Madhavi M. R., Abusahmin, B.S., Puvvada, N., Lakshminarayana, V., (2017), "Predictive data mining techniques for management of high dimensional big-data". Journal of Industrial Pollution Control vol 33, pp 1430-1436.

AUTHORS PROFILE

Venkata Ramana N. Currently working as Asst.Professor in K L University, pursuing Ph.D degree in Computer Science and Engineering from Annamalai University, Chidambaram, TN , India. His research interest includes Big Data ,IOT,Digital Image Processing.

Chandra sekhar Kolli . Currently working as Asst.Professor in KL University, pursuing Ph.D degree in Computer Science and Engineering from Gitam University, Visakhapatnam, India. His research interest includes Big Data ,IOT.

Ravi Kumar T. Currently working as Assoc. Professor in KL University, pursuing Ph.D degree in Computer Science and Engineering. His research interest includes Big Data, IOT**.**

Puvvada Nagesh *Currently* working as Asst.Professor in KL University, pursuing Ph.D degree in Computer Science and Engineering from KL University, Vijayawada, India. His research interest includes BigData, IOT, Cloud Computing.