Decision Tree Over KNN For Lung Cancer Detection To Increase Accuracy

N Charan¹, S.Parthiban²

¹Research scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105.

^{2*}Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105.

ABSTRACT:

Aim: Lung cancer detection using Decision Tree over KNN to improve accuracy. **Materials and Methods:** Detection of Lung cancer using Decision Tree over KNN to improve accuracy. Here the G-power analysis was carried out with 80% and the sample size for the two groups are 40. **Results and Discussion:** Detection of Lung cancer using Decision Tree over KNN to improve accuracy of 98.06% and 90.73% respectively. The decision tree achieved a mean of 98.06%, standard deviation of 0.75218 and the standard error mean of 0.23786. For the KNN the mean is 90.73% standard deviation is 0.67100 and standard error mean is 0.21219. The 2-tailed significance value smaller than 0.001 (p<0.05) showed that our hypothesis holds good. **Conclusion:** Decision Tree performs significantly better than KNN. **Keywords:** Lung cancer, SCLC, NSCLC, Novel Decision Tree, KNN, Detection, ID3, Machine Learning.

INTRODUCTION:

World Cancer Research Fund lung cancer statistics shows that Hungary is at an ace rate of cancer. Malignant growth is a gathering of infections described by the uncontrolled development and spread of strange cells (Podolsky et al. 2016). In the event that the spread isn't controlled, it can bring about death. Cellular breakdown in the lungs was the most well-known malignant the growth in world. contributing 2,093,876 of the complete number of new cases analyzed in 2018.e principal normal. We can fix cellular breakdown in the lungs provided that you distinguish the yearly stage. Lung cancer is classified into two types (Flehinger et al.

1984). Habitats for Medicare and Medicaid Services to give Medicare inclusion to cellular breakdown in the lungs screening has made ready for cross country cellular breakdown in the lungs separating the USA (Gunaydin, Gunay, and Sengel 2019). Striking auxiliary lung cancer is immediately ranked on the beam spam sort, SCLC and NSCLC (Gu et al. 2019).

In this research of detection of lung cancer, the number of papers published in the research gate are 1240 and the number of papers published in google scholars are 1150. AI is a technique that permits PCs to figure out how to anticipate specific results. In cellular breakdown in the lungs, scientists are utilizing PC calculations to make PC supported programs that are better ready to distinguish disease in CT checks than radiologists or pathologists (Makaju et al. 2018). Machine learning is creating advancements attributable to its algorithms used for creating predictions for both SCLC and NSCLC (Hirsh and Melosky 2018). NCI-financed specialists are attempting to propel how we might interpret how to forestall, recognize, and treat cellular breakdown in the lungs. There has been a lot of headway made, for researchers are distinguishing a wide range of hereditary changes that can drive cellular breakdown in the lungs development. In calculating effort using any of the higher than mentioned techniques, the requirement of massive datasets with most info concerning the project (Ali et al. 2006). Designated therapies distinguish and go after specific kinds of disease cells with less damage to typical cells. As of late, many designated treatments have opened up for cutting edge cellular breakdown in the lungs and more are being developed. This analysis paper makes use of machine learning to predict the hassle of knowing all the certain parameters. Two styles of models of predictions are there, one is algorithmic and therefore the alternative is nonalgorithmic et al. 2022). (Tjong varieties Algorithmic have outlined formulae for detection and calculation whereas non-algorithmic varieties have no outlined formulae for this NSCLC purpose.(Bhavikatti et al. 2021; Karobari et al. 2021; Shanmugam et al. 2021; Sawant et al. 2021; Muthukrishnan 2021; Preethi et al. 2021; Karthigadevi et al. 2021: Bhanu Teja et al. 2021: Veerasimman et al. 2021; Baskar et al. 2021)

The limitation is major the supposition of linearity between subordinate factors and free factors. It not just gives a proportion of how proper an indicator is yet additionally its heading of affiliation. It does not rely on any Machine learning model that works inside and makes predictions. Novel Decision tree is one more regulated learning calculation under order which goes method (Thamilselvan and Sathiaseelan 2016). It has three estimating boundaries that are data gain, gain proportion and gini list in the SCLC and NSCLC. KNN calculation is a straightforward directed AI used to settle both order and relapse examination. Classification of lung cancer through the decision tree and KNN and their accuracy in this paper profoundly propels every one of the patients and the specialists to utilize these proposed strategies based on SCLC and NSCLC within the healthy industry (Yu et al. 2019). The objective of this research is to detect lung cancer using Decision Tree over KNN to improve accuracy and perform comparative studies to know the best methodologies.

MATERIALS AND METHODS

The experiment was conducted in Machine Learning Laboratory, Saveetha School Of Engineering, Saveetha Institute Of Medical And Technical Sciences. In this research 2 groups were taken, 1 group refers to Decision Tree and the group 2 refers to KNN.The system, using the clincalc website, calculates the sample size by setting G-power 80% and measuring it as 40 samples (Group 1=20, Group 2=20).

The sample preparation group 1 is done for the Novel Decision Tree Algorithm, Novel Decision tree is a treelike design, in which the interior hub signifying a large set of branches addresses the results of test sets and leaf node hubs contain the result of main node marks. Many researchers have put a lot of effort into improving these weaknesses, but the research still needs to be improved. This article discusses the basic principle, model, advantages and disadvantages of the Decision tree algorithm.

The sample preparation group 2 is done for the KNN algorithm. The term KNN is used in artificial intelligence and machine learning to clear classification and regular patterns; we group elements that are highly connected together and quickly uncover shared properties and correlations including SCLC and NSCLC. By doing KNN, it is possible to perform further data mining activities such as clustering, classification, and other data mining tasks.

The Google Colab with Windows 10.0 system has been used to develop the prediction of Lung Cancer. The system uses two groups Novel Decision Tree and KNN algorithm technique where this algorithm is fitted into the dataset of SCLC and NSCLC which is then tested and trained for the process of estimation and detection to improve accuracy, the sample dataset is 40.

The data collection is taken from the open source access website IEEEdataport.org and datascience.com that is used for detection of Lung Cancer using Novel Decision Tree over KNN technique. The open data set contains 7 columns and 60 rows.The dataset here the images which are specifically CT images that are used here are analyzed by the algorithms and the algorithms specify the image in its own way and chooses the best path fastly and shortest whereas Naive bayes algorithm split and find the best solution of the images.

Decision Tree

Novel Decision tree is one more administered learning calculation which under characterization strategy. goes Novel Decision tree is a plan, where every inside center point meaning a test set branch tends to the consequences of nodes and test leaf center points contain the aftereffect of the class marks. It has three estimating boundaries that are data gain, gain proportion and gini file (Kubík and Polák 1986). Data gain can be determined in light of the entropy (the proportion of how much vulnerability in the informational index) SCLC, subsequently data gain can be characterized as the contrast between unique data required (entropy of quality prior to parting) NSCLC and new necessity (entropy of property in the wake of parting) (Bajre et al. 2017). In light of the greatest entropy that is prior to parting the property is taken as the root hub and in the wake of estimating the data gain characteristic with most noteworthy data gain is chosen as parting quality. Data gain goes under ID3 (Iterative Dichotomies 3) calculation.

Pseudocode for Decision tree

Step 1:from sklearn.datasets import make_classification

Step	2: from			
sklearn.model_selection	import			
train_test_split				

Step 3: from sklearn.metrics import accuracy_score

Step 4:from sklearn.linear_model import DecisionTreeClassifier

Step 7:xtrain, xtest, ytrain, ytest =
train_test_split(x, y,test_size=0.2,
random_state=32) Step 8:model =
DecisionTreeClassifier()
Step 9:model.fit(xtrain, ytrain)
Step

10:print(accuracy_score(ytest, model.predict(xtest)))

K-Nearest Neighbour (KNN):

The KNN can rival the most reliable models since it makes exceptionally precise forecasts (Lu, Zhu, and Gu 2014). Consequently, you can involve the KNN calculation for applications that require high precision yet that don't need a comprehensible model. KNN might be utilized to foresee chances of fostering guaranteed software based on observed characteristics of the given project which includes SCLC and NSCLC. The KNN algorithm at the planning stage just stores the dataset and when it gets new data, then, at that point, it organizes that data into a class that is similar to the new data.

Pseudocode for KNN

Step 1:from sklearn.datasets import make_classification

Step2:fromsklearn.model_selectionimporttrain_test_split

Step 3:from sklearn.metrics import accuracy_score

Step 4:from sklearn.neighbors import KNeighborsClassifier

Step 5:nb_samples = 1050

Step6:x,y=make_classification(n_samples=nb_samples,n_features=2,n_informative=2,

n_redundant=0, n_clusters_per_class=1) **Step 7:**xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size=0.2, random_state=32)

Step 8:model

=KNeighborsClassifier()

Step 9:model.fit(xtrain, ytrain) print(accuracy_score(ytest, model.predict(xtest)))

In the proposed system the training and testing of the data is made in the google colab and using the spss software to predict the graph and also the accuracy of the corresponding Novel Decision tree and KNN in the process of detecting Lung Cancer. The accuracy of the Novel Decision tree should be greater than KNN.

Statistical Analysis

The analysis was performed using IBM SPSS version 21. Independent variables project, p_name, year_end etc and dependent variables are age, smokes, area, alkhol. The proposed algorithm iterations were done with some samples and existing algorithm iterations were done with a sample of 20 and for every iteration the resultant accuracy was used for analyzing final accuracy (Bhuvaneswari and Brintha Therese 2015). Independent Sample t-test was performed with the values obtained from the iterations. There is no significant difference in accuracy of the adopted algorithms with p-value = 0.001 (<0.05).

RESULT

In Table 1, the novel decision tree algorithm is better than the KNN algorithm. From the lung cancer dataset, it is verified that the accuracy of the decision tree was much greater than the k-nearest neighbor algorithm.

In Table 2, The decision tree was achieved with the mean of 98.06%, standard deviation of 0.75218 and the standard error mean of 0.23786. For the KNN the mean is 90.73%, standard deviation is 0.67100 and standard error mean is 0.21219. In Table 3, The 2-tailed significance value is smaller than 0.001 (p<0.05), proves that our hypothesis is best.

Figure 1, Represents mean accuracy of novel decision tree algorithm and KNN technique. Decision tree model obtained 98.06% accuracy and the KNN obtained 90.73% accuracy. The Decision tree technique achieved better performance than KNN.

DISCUSSION

The Novel Decision Tree has accuracy of 98.06% which is higher than the KNN which has 90.73% (Blandin Knight et al. 2017). There is a 2-tailed significant difference in the accuracy for the decision tree and KNN is 0.001 (p< 0.05) by calculating independent sample ttests (Pastorino et al. 2003). The dataset is taken from the open source access website kaggle that is used for lung cancer detection along with SCLC and NSCLC using Novel Decision Tree over KNN to improve accuracy. The open access dataset consists of 60 rows and 7 columns.

In the existing system, the accuracy of Decision Tree and KNN are 96.32% and 89.12% respectively (Pradhan and Chawla 2020). In the proposed system the testing of the data along with the training is made in the google colab and using the SPSS software to predict the graph and also the accuracy of the corresponding Novel Decision Tree and KNN. The accuracy of Naive Bayes has to be greater when compared to Logistic Regression (Wajid et al. 2016).

This analysis paper makes use of machine learning to predict the accuracy of knowing all the certain parameters. The limitation is that the real time dataset with more parameters in SCLC and NSCLC gives the results of accuracy. In the future work, integration of our estimation model to any application. It will help better to the user to predict the accuracy of the corresponding algorithm. Factors affecting the algorithms are sample size of the dataset and the test size of the dataset. Although our proposed algorithm is faster and better efficient than other algorithms, the reports suffer with a large prevalence of lack of methodology without reviewed literature reaching the NSCLC threshold and to support utilization in SCLC clinical practices. Further, this research work can be improved by deploying a model that increases robustness.

CONCLUSION

Accuracy of Novel Decision Tree (98.06%) performs better than the accuracy of KNN (90.73%) in the detection of lung cancer. Thereby the study focused more on the machine learning algorithm of Decision tree.

DECLARATION Conflict of Interests

No conflict of interest in this manuscript.

Authors Contribution

Author NC was involved in data collection, data analysis, manuscript writing, Author PS was involved in conceptualization, data validation and critical review of manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1.Awata Software Systems Pvt.Ltd.

2.Saveetha University

3.Saveetha Institute of Medical and Technical Sciences.

4.Saveetha School of Engineering.

REFERENCES

- Ali, Iqbal Unnisa, Zhen Xiao, Winfred Malone, Mylinh Smith, Thomas P. Conrads, Timothy D. Veenstra, Peter Greenwald, Brian T. Luke, and Jerry W. McLarty. 2006. "Plasma Proteomic Profiling: Search for Lung Cancer Diagnostic and Early Detection Markers." Oncology Reports 15 (5): 1367–72.
- 2. Bajre, M. K., M. Pennington, N. C. Woznitza. Beardmore. M. Radhakrishnan, R. Harris, and P. McCrone. 2017. "Expanding the Role Radiographers in Reporting of Suspected Lung Cancer: A Cost-Effectiveness Analysis Using a Decision Tree Model." Radiography 23 (4): 273–78.
- Baskar, M., R. Renuka Devi, J. Ramkumar, P. Kalyanasundaram, M. Suchithra, and B. Amutha. 2021. "Region Centric Minutiae Propagation Measure Orient Forgery Detection with Finger Print Analysis in Health Care Systems." *Neural Processing Letters*, January. https://doi.org/10.1007/s11063-020-10407-4.
- 4. Bhanu Teja, N., Yuvarajan Devarajan, Ruby Mishra, S. Sivasaravanan, and D.

Thanikaivel Murugan. 2021. "Detailed Analysis on Sterculia Foetida Kernel Oil as Renewable Fuel in Compression Ignition Engine." *Biomass Conversion and Biorefinery*, February. https://doi.org/10.1007/s13399-021-01328-w.

- 5. Bhavikatti, Shaeesta Khaleelahmed, Mohmed Isaqali Karobari, Siti Lailatul Akmar Zainuddin, Anand Marya, Sameer J. Nadaf, Vijay J. Sawant, Sandeep B. Patil, Adith Venugopal, Messina. Pietro and Giuseppe Alessandro Scardina. 2021. "Investigating the Antioxidant and Cytocompatibility of Mimusops Elengi Linn Extract over Human Gingival Fibroblast Cells." International Journal of Environmental Research and Public Health 18 (13).https://doi.org/10.3390/ijerph18137162
- Bhuvaneswari, P., and A. Brintha Therese. 2015. "Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm." *Procedia Materials Science*. https://doi.org/10.1016/j.mspro.2015.0 6.077.
- 7. Blandin Knight, Sean, Phil A. Crosbie, Haval Balata, Jakub Chudziak, Tracy Hussell, and Caroline Dive. 2017.
 "Progress and Prospects of Early Detection in Lung Cancer." *Open Biology* 7 (9). https://doi.org/10.1098/rsob.170070.
- Flehinger, B. J., M. R. Melamed, M. B. Zaman, R. T. Heelan, W. B. Perchick, and N. Martini. 1984. "Early Lung Cancer Detection: Results of the Initial (prevalence) Radiologic and Cytologic Screening in the Memorial Sloan-Kettering Study." *The American Review of Respiratory Disease* 130 (4):

555-60.

- Gunaydin, Ozge, Melike Gunay, and Oznur Sengel. 2019. "Comparison of Lung Cancer Detection Algorithms." 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). https://doi.org/10.1109/ebbt.2019.8741 826.
- 10. Gu, Qianbiao, Zhichao Feng, Qi Liang, Meijiao Li, Jiao Deng, Mengtian Ma, Wei Wang, Jianbin Liu, Peng Liu, and Pengfei Rong. 2019. "Machine Learning-Based Radiomics Strategy for Prediction of Cell Proliferation in Non-Small Cell Lung Cancer." *European Journal of Radiology* 118 (September): 32–37.
- Hirsh, Vera, and Barbara Melosky. 2018. Update on the Treatment of Metastatic Non-Small Cell Lung Cancer (NSCLC) in New Era of Personalised Medicine. Frontiers Media SA.
- 12. Karobari, Mohmed Isaqali, Syed Nahid Basheer, Fazlur Rahman Sayed, Sufiyan Shaikh, Muhammad Atif Saleem Agwan, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "An In Vitro Stereomicroscopic Evaluation of Bioactivity between Neo MTA Plus, Pro Root MTA, BIODENTINE & Glass Ionomer Cement Using Dye Penetration Method." Materials 14 (12).

https://doi.org/10.3390/ma14123159.

13. Karthigadevi, Guruviah, Sivasubramanian Manikandan, Natchimuthu Karmegam, Ramasamy Subbaiya, Sivasankaran Chozhavendhan, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2021. "Chemico-Nanotreatment Methods for the Removal of Persistent Organic Pollutants and Xenobiotics in Water -A Review." *Bioresource Technology* 324 (March): 124678.

- 14. Kubík, A., and J. Polák. 1986. "Lung Cancer Detection. Results of a Randomized Prospective Study in Czechoslovakia." *Cancer* 57 (12): 2427–37.
- 15. Lu, Chunhong, Zhaomin Zhu, and Xiaofeng Gu. 2014. "An Intelligent System for Lung Cancer Diagnosis Using a New Genetic Algorithm Based Feature Selection Method." *Journal of Medical Systems* 38 (9): 97.
- 16. Makaju, Suren, P. W. C. Prasad, Abeer Alsadoon, A. K. Singh, and A. Elchouemi. 2018. "Lung Cancer Detection Using CT Scan Images." *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2017.12 .016.
- 17. Muthukrishnan, Lakshmipathy. 2021.
 "Nanotechnology for Cleaner Leather Production: A Review." *Environmental Chemistry Letters* 19 (3): 2527–49.
- 18. Pastorino, Ugo, Massimo Bellomi, Claudio Landoni, Elvio De Fiori, Patrizia Arnaldi, Maria Picchio, Giuseppe Pelosi, Peter Boyle, and Ferruccio Fazio. 2003. "Early Lung-Cancer Detection with Spiral CT and Positron Emission Tomography in Heavy Smokers: 2-Year Results." *The Lancet* 362 (9384): 593–97.
- 19. Podolsky, Maxim D., Anton A. Barchuk, Vladimir I. Kuznetcov, Natalia F. Gusarova, Vadim S. Gaidukov, and Segrey A. Tarakanov. 2016. "Evaluation of Machine Learning Algorithm Utilization for Lung Cancer Classification Based on

Gene Expression Levels." Asian Pacific Journal of Cancer Prevention: APJCP 17 (2): 835–38.

- 20. Pradhan, Kanchan, and Privanka Chawla. 2020. "Medical Internet of Machine Things Using Learning Algorithms for Lung Cancer Detection." Journal of Management Analytics. https://doi.org/10.1080/23270012.2020 .1811789.
- 21. Preethi, K. Auxzilia, K. Auxzilia Preethi, Ganesh Lakshmanan, and Durairaj Sekar. 2021. "Antagomir Technology in the Treatment of Different Types of Cancer." *Epigenomics.*

https://doi.org/10.2217/epi-2020-0439.

- 22. Sawant, Kashmira, Ajinkya M. Pawar, Kulvinder Singh Banga, Ricardo Machado, Mohmed Isaqali Karobari, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "Dentinal Microcracks after Root Canal Instrumentation Using Instruments Manufactured with Different NiTi Alloys and the SAF System: A Systematic Review." NATO Advanced Science Institutes Series E: Applied Sciences 11 (11): 4984.
- 23. Shanmugam, Vigneshwaran, Rhoda Afriyie Mensah, Michael Försth, Gabriel Sas, Ágoston Restás, Cyrus Addy, Qiang Xu, et al. 2021. "Circular Economy in Biocomposite Development: State-of-the-Art, Challenges and Emerging Trends." *Composites Part C: Open Access* 5 (July): 100138.
- 24. Thamilselvan, P., and J. G. R. Sathiaseelan. 2016. "Detection and Classification of Lung Cancer MRI Images by Using Enhanced K Nearest Neighbor Algorithm." *Indian Journal*

of Science and Technology. https://doi.org/10.17485/ijst/2016/v9i4 3/104642.

- Michael C., Danielle 25. Tiong, S. Bitterman, Kristen Brantley, Anju Nohria, Udo Hoffmann, Katelyn M. Atkins, and Raymond H. Mak. 2022. "Major Adverse Cardiac Event Risk Prediction Model Incorporating Baseline Cardiac Disease, Hypertension, and Logarithmic Left Anterior Descending Coronary Artery Radiation Dose in Lung Cancer (CHyLL)." *Radiotherapy* and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology, February. https://doi.org/10.1016/j.radonc.2022.0 2.010.
- 26. Veerasimman, Arumugaprabu, Vigneshwaran Shanmugam, Sundarakannan Rajendran, Deepak Joel Johnson, Ajith Subbiah, John Koilpichai, and Uthayakumar Marimuthu. 2021. "Thermal Properties of Natural Fiber Sisal Based Hybrid Composites A Brief Review." *Journal of Natural Fibers*, January, 1–11.
- 27. Wajid, Summrina Kanwal, Amir Hussain, Kaizhu Huang, and Wadii Boulila. 2016. "Lung Cancer Detection Using Local Energy-Based Shape Histogram (LESH) Feature Extraction and Cognitive Machine Learning Techniques." 2016 IEEE 1.5th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC). https://doi.org/10.1109/icci-

cc.2016.7862060.

28. Yu, Lingming, Guangyu Tao, Lei Zhu, Gang Wang, Ziming Li, Jianding Ye, and Qunhui Chen. 2019. "Prediction of Pathologic Stage in Non-Small Cell Lung Cancer Using Machine Learning Algorithm Based on CT Image Feature Analysis." *BMC Cancer.* https://doi.org/10.1186/s12885-019-5646-9.

TABLES AND FIGURES

Table 1. Comparison of prediction of accuracy between Decision Tree and KNN. Decision Tree achieved an accuracy of 98.06% compared to KNN, having 90.73%.

Execution	Decision tree	KNN		
1	98.06	90.74		
2	98.42	90.24		
3	98.45	90.42		
4	97.33	91.33		
5	97.21	90.11		
6	99.35	91.29		
7	98.44	92.21		
8	98.20	90.22		
9	97.20	90.45		
10	97.30	90.37		

Table 2. Mean value of the Decision Tree is 98.0600 and KNN mean value is 90.7380. The table below shows the Decision Tree obtained standard deviation (0.75218) and standard error means (0.23786). KNN has standard deviation (0.67100) and standard error mean (0.21219).

Accuracy	Algorithm	Ν	Means	Std deviation	Std error means	
	DECISION TREE	10	98.0600	0.75218	0.23786	
	KNN	10	90.7380	0.67100	0.21219	

Table 3. Independent Samples Test. The accuracy increases and the error rate decreases. The 2-tailed significance is less than 0.001.

	Levene fo Equa Varia	e's Test or lity of ances	t-test for Means			or Equality of			
	f	sig	t	df	Sig		Std.	95% Confidence Interval of the Difference	
					(2tailed)	Mean Diff.	Differen ce	Lower	Upper
Equal variances assumed	0.529	0.477	22.971	18	0.001	7.32200	0.31875	6.65233	7.99167
Equal variances not assumed			22.971	17.770	0.001	7.32200	0.31875	6.65171	7.99229

GRAPH





Fig. 1. Prediction accuracy for the two algorithms. The accuracy of the Decision Tree is better than the accuracy of KNN. X Axis: Naive Bayes vs KNN Y Axis: Mean accuracy of detection \pm 1SD.