A Review of the Comparative Analysis of Machine Learning Techniques for the Detection and Classification of Lung cancer

1st Suchita Meshram
Department of Computer Science and Engineering,
Jhulelal Institute of Technology
Nagpur, Maharashtra, India
suchitameshram05@gmail.com

2nd Nisha balani
Department of Computer science and engineering
Jhulelal Institute of Technology
Nagpur, Maharashtra, India
n.balani@jit.org.in

3rd Imran Khan
Department of Computer science and engineering
Jhulelal Institute of Technology
Nagpur, Maharashtra, India
m.imran@jitnagpur.edu.in

***Abstract*— Lung cancer is one of the significant reasons for death among India.This research article aims to assess and comparation the effectiveness of different machine learning methods in terms of accuracy.In order to evaluate the various classifiers' degrees of accuracy, we took into account the various models that researchers have employed as well as their drawbacks and restrictions. We make that certain classifiers consume little accuracy and others have advanced accuracy but are not nearly at 100% after doing a thorough analysis of the literature. Therefore, we need to take a more calculated approach to correctly categorise lung cancer nodules. A comprehensive analysis of the literature showed that the little accuracy levels were caused by improper handling of the Dicom photos. We found that the collaborative classifier outpaced the other machine after a thorough analysis. As a result, after accounting for every classifier, we discovered that the main machine learning methods provided accuracy that was far below 90%. Additionally, we discovered that in instruction to recover the accuracy level and draw conclusions for the tumour diagnosis that accurately reflected our growing understanding of how lung cancer is classified, a better model had to be employed. Additionally, the revised model required to be relevant and trustworthy. Ultimately, a thorough investigation into the discipline of oncology was required to improve the classification of benevolent and malignant tumours.**

Keywords*— *lung cancer, back-propagation algorithm, classification, machine learning, and ensemble learning.

I. INTRODUCTION

One of the main reasons of decease universal is lung cancer. Lung cancer claims the exists of additional people each year than all other cancers combined. The same deadly illness affects not just men but also women. The patient's life expectancy after lung cancer diagnosis is extremely short. Early diagnosis increases the patient's balances of survival, which is necessary to discover cancer as soon as possible and raise the patient survival rate. As a result, we may use modern methods in the image processing and machine learning domains to get correct and timely results. Increasing the number of replicas utilized in the process will improve the accuracy. Accurate diagnosis and early prognosis of cancer can increase survival rates. The earlier methods include the analysis of pictures from attractive character imaging, computed imaging scans, and mammography. By using competence, the trained doctors diagnose the illness and establish the cancer's phases. A few medical operations, chemical actions to destroy or prevent malignant cells from replicating, radiotherapy, and targeted therapy are all part of the treatment. These examines are very laborious, costly, and tender for the bodily portion involved. In order to shorten this procedure, different image processing algorithms are used. Hospitals offer blood tasters and CT scan images. Reports from Onscreen Imaging are quieter than those from MRIs and X-rays. The most common cancer in the world, lung cancer, books for approximately 13.9 million cases and 8.29 million deaths [1] due to early detection of abnormalities. In 2018, there were 1,682,210 new cases of cancer in the Unified Situations, with 595,690 fatalities. Lung cancer accounts for one of the highest proportions of all cancer cases, with 158,080 fatalities and 224,391 new cases in 2016. Because lung cancer cells can be identified early and treated effectively, their survival rate is suggestively lower than that of other cancer types. A handout from Globe Wellness Company, the American Lung Connotation, and the American Cancer Cell Group states that when cancer cells are restricted and found at onset, the existence rate increases to 54.4% from 17.7%. Approximately 16% of cancer cases are recognized at start.

1.1 CAUSES OF CANCER

The popular of incidences of lung cancer are produced by smoke, which is also the main risk factor. Over time, lung tissue is damaged by the cancer-causing compounds in tobacco smoke, which can result in the development of cancer. Furthermore, there is a large increase in risk associated with secondhand smoke, radon gas, asbestos, and other environmental contaminants. The chance of getting lung cancer is also influenced by genetics and a domestic history of the disease. Although these cases are less frequent, factors like air contamination and work-related exposures can cause lung cancer in non-smokers. The prognosis is much improved by early lung nodule diagnosis, which may even save lives and lower death rates. CT scans can detect dense forms called lung nodules, which have a size range of 2.9 to 30 mm. Compared to chest X-rays, CT has a 20% lower incidence of lung cancer and a lower 5-year mortality rate, according to the Lung Cancer Cells Screening Routes (LCST). When it comes to spotting worrisome lesions on CT scans, particularly tiny nodules, radiologists are extremely important. Computer-Aided Detection (CADe) methods have been created to help radiologists by improving accuracy and minimizing interpretation time, given the complexity of lung nodule detection[1]. CADe technology is future to help in the early discovery of lung cancer by detecting possible lesions with more automation, decreased false positives, and increased sensitivity [2].

II. MACHINE LEARNING APPROACHES

Probably unity of the most inclusive methods for gathering data from many data bases [3, 4] using formularies. There are many different types of information bases, such as private or public centers, the Lung Photos Source Company, the Early Cancer Cells Action Strategy, the Japanese Culture of Radiological

Modern Technology (JSRT) data sources, the Automatic Blemishes Sighting 2009 (anode09), the Lung Photo File Consortium, and the Picture Data Source Initiative (lidc idri). Recently, Cade and Cadx have been searching for and providing services for the greatest important notes related to finding lung cancer cells. These parts are currently being hauled around for different tasks, though. The radiotherapist's uniqueness of growths is not providing by cad strategies, and cad approaches also do not find flaws or have very high mechanization levels. This is the reason why scientific approaches have not yet made extensive use of these strategies. In order to increase the level of automation, it is necessary to create a superior method for the documentation and medical judgement of lung lesions on CT scans by arranging the lesions into a only telephone system for credentials and characterisation. The blog post also offers recompence using the landmark and pie diagram of targeted slope approaches (hog) for lung blemish eradication specifically and for separating the reachable blemishes from certain other structures. The likelihood of hatred providing far more assistance in the radiologists' decision-making is the basis for the medical diagnosis. For the purpose of eliminating false positives, a novel classifier and an artificial semantic network (ANSM) have also been employed [4]. The relational database used in this particular analysis has 520 examples that were randomly selected from various datasets in the community area. The technique below is a division with a 97% precision rate combined with an outstanding system discovery formula that has a 94.4% sensitivity level and 7.04 false positives per scenario.

III. MOVITATION

According to 10,000 studies conducted in 2017 alone, 18.1 million uncommon new bags of cancer were reported worldwide, accounting for 15.6% of all fatalities [4]. There were 520 people with one to eight lung nodules. We can positively state that 64% of nodules are rationally unlikely of existence spiteful, 149 have nodules with undetermined malevolent cells, 78% have nodules [4] that are abstemiously doubtful of being malicious, and 62% have nodules that are very doubtful of being malevolent. Of them, 31 have nodes that are very doubtful of being malignant.

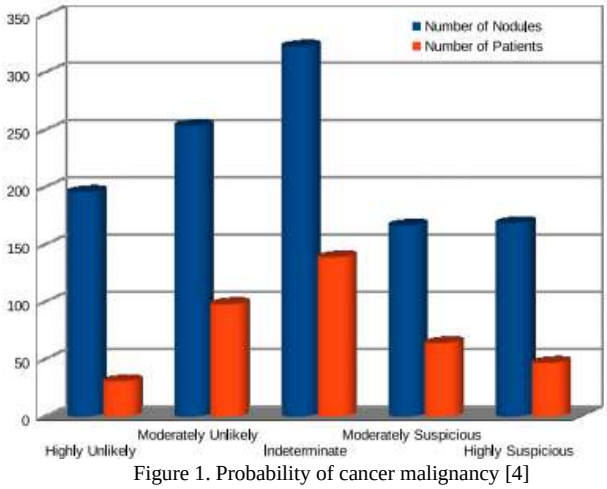

Figure 1. Probability of cancer malignancy [4]

The "Number of Nodules" and "Number of Patients" are compared in a bar chart across five nodule suspicion levels, ranging from "Highly Unlikely" to "Highly Suspicious." There is doubt regarding the risk of malignancy, as seen by the majority of nodules and patients falling into the "Indeterminate" group. The largest counts are found in the "Moderately Unlikely" and "Indeterminate" categories, whereas the "Highly Unlikely" and "Highly Suspicious" categories have comparatively lower values. This indicates that a sizable fraction of nodules are difficult to classify, underscoring the difficulty in determining the risk of nodule malignancy.

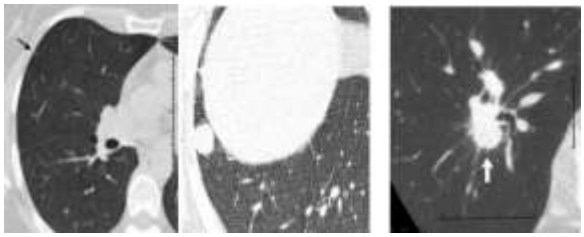
Figure 2. CT scan segment of lung nodule

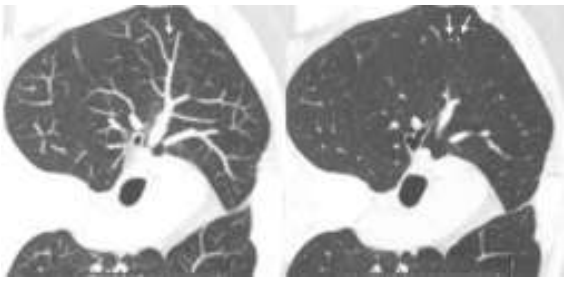
Figure 3. 1.25 mm weighty transverse CT

The thickness of the CT slice is 1.23 mm, and it displays a nodule that is around 2.6 mm long. Larger malignant tumors and nodules can both seem normal on a conventional CT scan, especially if the scan is from a young adult. It is clear from Figure 2 that a nodule measuring 2.6 mm in length may be important for diagnosis. Documented evidence supports the idea that this nodule may spread to other parts of the body or may suggest new findings in a patient [5]. Consequently, it might not be possible to definitively test for lung cancer with just one imaging. After comparing the two pictures, I can see that the CT scan's left side's 12.9 mm lung images show a patient with lung cancer. The pulmonary blood arteries are readily visible in the right side nodule [6]. This comparison helps us distinguish between a typical lump nodule and an unhealthy growth of a nodule, even though the image on the left is much smaller.
Figure 3 shows 2.8 mm round opaqueness that appear to be lung nodes in a 1.27 mm CT section taken at an oblique angle. The results of an 8 mm steady TS MIP placed on the oblique section directly directly above show that the front clarity relates to a normal blood vessel, and only the later of the two opaqueness represents a lung nodule (shown by the missile in the bottom image). While TM is effective in identifying lung nodules, it also mistakenly identifies a large number of blood vessels as nodules [6, 7].

IV. LITERATURE SURVEY

As of right now, the imaging modality best suited for lung cancer investigations for early detection is computed tomography (CT). High spatial determination, high temporal resolve, and high difference resolve of the chest's anatomical features are all providing by CT scans. In this manner, tiny nodules that would be difficult to see on traditional radiography can be shown [8]. Awai et al. [9] report that CT has a 2.6–10 times higher lung

cancer detection rate than analog radiography. Though, CT produces a lot of medical images, and when joint with radiologists' workload, this could lead to misinterpretation or erroneous detection (cancer not being detected) (inability to appropriately diagnose a tumor). As a result, computer technologies are essential for helping radiologists make decisions.

Several research, like Armato et al. [10], have suggested using CADe systems to detect lung nodules in the literature. They created a CADe system with a compassion of 70% and 9.6 FP per case using linear discriminant analysis. With 187 nodules this approach has been validated. Using 121 nodules

(solitary, juxtapleural, juxtavascular, and ground-glass nodules), Suzuki et al. [11] created a design credit technique based on an artificial neural network for a CADe organization, dubbed MTANN, and got a sensitivity of 80.3% with 4.8 FP per case. A CADe method with a sensitivity of 82.66% was published by Messay et al. [12] . Three FP per case were validated using 143 nodules (solitary, juxtapleural, juxtavascular, and ground-glass nodules).

The preprocessing stage of machine learning models depends heavily on the developments in image segmentation approaches, which Tripathi, Tyagi, and Nath (2019) discussed. In tasks like object identification and medical imaging, their work highlights the significance of good segmentation for accurate classification and subsequent processing[13]. The application of Decision Trees, Logistic Regression, and Support Vector Machines (SVMs) in classification problems was investigated by Radhika, Nair, and Veena (2019). These algorithms have been very helpful in solving binary and multiclass classification issues, especially SVMs. Their study demonstrated the approaches' advantages in terms of readability, ease of use, and precision across a range of datasets[14]. The use of SVMs and ensemble approaches, in particular Random Forest, were examined by Roy et al. (2019). By averaging the predictions of several decision trees, Random Forest, an ensemble of decision trees, provides resilience and lessens overfitting. On the other hand, SVMs are well-known for working well in high-dimensional spaces, which makes them appropriate for a range of classification tasks[15].

A number of academics have showed extensive study on the request of support vector machines (SVMs) in various domains. These researchers include Abbas Ali et al. (2018), Makaju et al. (2018), and Song et al. (2012). Abbas Ali et al. (2018) focused on the application of SVMs in medical diagnostics, while Makaju et al. (2018) explored their use in image classification tasks[16]. Artificial Neural Networks (ANNs) and Naive Bayes classifiers were used in various classification problems, with Günaydin et al. (2019) and Dash et al. (2014) investigating their use.

Authors and Year	Title	Journal/Conferece	Techniques Used	Dataset	Findings/Conclusions	Research Gap
Zhang et al.(2020)	Lung Cancer Detection Using Deep Learning menthod	IEEE Transactions on Medical Imaging	CNN, Transfer Learning	LIDC-IDRI	Demonstrated that deep learning techniques, particularly CNNs, significantly recover the accuracy of lung cancer detection. Transfer learning was effective in improving model performance with limited labeled data.	Limited generalizability due to reliance on a single dataset. Need for diverse datasets to validate the model across different populations.
Khan et al. (2021)	Machine Learning Systems for Lung Cancer Prediction : A Review	IEEE Access	SVM, Random Forest, k-NN	Multiple Public Datasets	Reviewed various ML techniques, with SVM and Random Forest being the most effective. Emphasized the need for large, high-quality datasets	Lack of emphasis on deep learning models. The review did not explore the potential of combining ML and DL techniques for better performance.
Singh & Gupta (2022)	Automated Lung Cancer Detection Using Hybrid Machine Learning	Journal of Biomedical Informatics	Hybrid models (SVM + CNN, Random Forest + ANN)	CT Scan Images (LUNA16)	Hybrid models combining SVM with CNN and Random Forest with ANN showed improved detection accuracy. Highlighted the potential of hybrid approaches in medical image analysis.	Need for real-time implementation and validation in clinical settings. The study lacked external validation on independent datasets.
Sharma et al.(2023)	Deep Learning-Based Framework for Primary	International Conference on Machine Learning	RNN, CNN, LSTM	Private Dataset	Proposed a deep learning framework combining CNN and LSTM, achieving high compassion and specificity in early-	The study used a private dataset, limiting the ability to compare results with other research. A lack

	Detection of Lung Cancer				stage lung cancer detection. Stressed the importance of temporal information in improving prediction accuracy.	of explainability in the model was also noted.
Li et al.(2024)	Advances in Machine Learning for Lung Cancer Screening	Journal of Thoracic Oncology	GANs, XGBoostEnsemble Learning	NLST Dataset	Discussed the latest advancements in ML techniques like GANs and ensemble methods. XGBoost and ensemble models outperformed traditional methods in lung cancer screening. Identified challenges like data imbalance and interpretability.	he study identified interpretability as a major challenge. There is also a need for addressing data imbalance and expanding research to multi-modal data integration.

V. COMPARATIVE STUDY AND PERFORMANCE METRICS

Table 1. Various Machine Learning Algorithms deployed for classification

Algorithm	Accuracy	Sensitivity	Specificity
Image Segmentation Technique	59.12%	65.4%	65.4%
Decision Trees	58.11%	72.67%	72.67%
Logistic Regression	65.4%	84.34%	96.66%
Random Forest	62.5%	64.28%	64.28%
Support Vector machines	62.5%	88.5%	88.5%
Artificial Neural Network	65.4%	65.4%	65.4%
Naive Bayes	62.5%	62.5%	62.5%
Ensemble Classifier	88.5%	72.67%	65.4%
Decision Fusion	72.67%	74.78%	62.5%
Linear Discriminator Analysis	74.78%	62.5%	91.66%
Back Propagation Network	72.67%	72.67%	72.67%
K-Nearest	72.24%	94.12%	53.78%

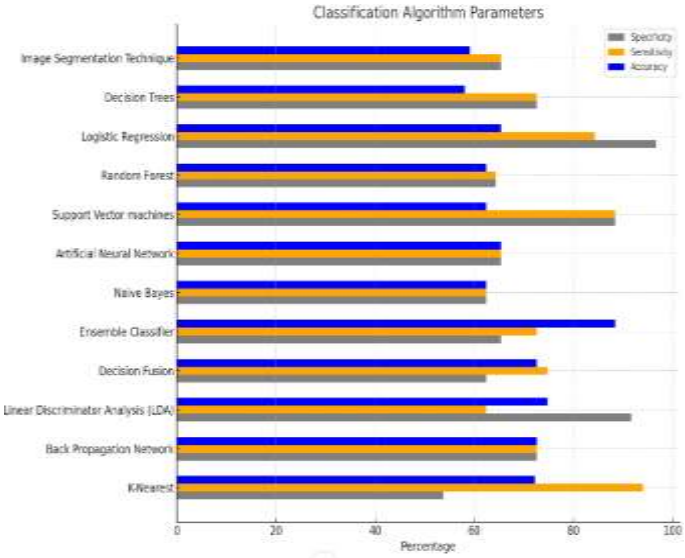

Figure 3. The accuracy levels of various machine learning algorithms for lung cancer classification

The bar graph that contrasts different categorization algorithms' sensitivity, specificity, and accuracy. Each algorithm's performance measures are reflected in the data, which makes it easy to see how effective each method is across these three criteria.

VI. RESULTS DISCUSSIONS

Table 2 displays the performance metrics of different machine learning method based on characteristics such as accuracy, sensitivity, specificity, ROC, AUC, and standard deviation. We decided after leading extensive research that noneof the classifiers is even close to having 100% accuracy. The classifier ensemble fared well overall. Additionally, Figure 11 displays the effectiveness and performance metrics of the dissimilar machine algorithms.

VII. CONCLUSION

This work offers a thorough analysis of several machine learning method used in the documentation and categorization of lung cancer. A thorough examination of numerous research publications revealed that there is still considerable difficulty in correctly differentiating between benign and malignant tumors. There are three main breaks in the literature at the moment. First,

the articles' classifiers' frequently insufficient accuracy emphasizes the need for more accurate and dependable models. Secondly, new methods designed for medical imaging, such MRI (Magnetic Resonance Imaging) and DICOM (Digital Imaging and Communications in Medicine), are required, especially for efficient noise reduction. Third, because they frequently work with raw images, classic machine learning algorithms like SVM (Support Vector Machine), Naive Bayes, K-Nearest Neighbors, and Decision Trees may miss little patterns that are essential for a precise diagnosis. Because medical informatics is such a delicate field, even small mistakes can have serious repercussions, which highlights the need for sophisticated classifiers with better algorithmic methods to increase accuracy. Thus, it is essential to conduct rigorous and ongoing research in cancer diagnosis in order to lower the disease's death rate.

VIII. FUTURE ENHANCEMENTS

In order to improve methods for lung cancer diagnosis and prediction, more research needs to focus on a few important areas. Firstly, generalizability and robustness of the model will be improved by growing and diversifying datasets across various populations and medical contexts. For practical use, hybrid models that combine deep learning and conventional machine learning approaches must be implemented in real-time and validated clinically. Model accuracy will increase by addressing data imbalance through methods like multi-modal data integration and data augmentation. Furthermore, improving the interpretability and explainability of models using techniques such as LIME and SHAP will encourage adoption and build trust in therapeutic settings. Models should be talented to adapt and learn continuously in instruction to remain relevant when new data becomes available. Last but not least, in order to guarantee adherence to medical standards and enable the effective application of AI-based solutions in healthcare, ethical and regulatory issues need to be explored.

IX. REFERENCES

[1] Syed, A.A., Fatima, W., Wajahat, R. (2018). Comparative analysis of learning algorithms for lung cancer identification. Indian Journal of Science and Technology, 11(27)

[2] Cruz, C.S.D., Tanoue, L.T., Matthay, R.A. (2011). Lung cancer: Epidemiology, etiology, and prevention. Clinics in Chest Medicine, 32(4): 605-644

[3] Dash, J.K., Mukhopadhyay, S., Garg, M.K., Prabhakar, N., Khandelwal, N. (2014). Multi-classifier framework for lung tissue classification. 2014 IEEE Students' Technology Symposium, Kharagpur, India, pp. 264-269.

[4] Firmino, M., Angelo, G., Morais, H. (2016). Computeraided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. BioMedical Engineering Online, 15(1): 1-17.

[5] Messay T, Hardie RC, Rogers SK. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Med Image Anal. 2010;14(3):390–406.

[6] Günaydin, Ö., Günay, M., Şengel, Ö. (2019). Comparison of lung cancer detection algorithms. 2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, Istanbul, Turkey, pp. 1-4.

[7] Kadir, T., Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. Translational Lung Cancer Research, 7(3): 304-312.

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

[9] Li Q. Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. Comput Med Imaging Graph. 2007;31(4–5):248–57.

[10] Kazuo A, Kohei M, Akio O, Masanori K, Haruo H, Shinichi H, Yasumasa N. Pulmonary nodules at chest ct: effect of computer-aided diagnosis on radiologists' detection performance. Radiology. 2004;230:347–52.

[11] Armato SG, Gieger ML, Moran CJ, Blackburn JT, Doi K, Macmahan H. Computerized detection of pulmonary nodules on CT scans. Radiographics. 1999;19(5):1303–11.

[12] Suzuki K, Armato III SG, Li F, Sone S, Doi K. Massive training artificial neural network (mtann) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. Medical Physics. 2003;30(7):1602–17.

[13] Tripathi, P., Tyagi, S., Nath, M. (2019). A comparative analysis of segmentation techniques for lung cancer detection. Pattern Recognition and Image Analysis, 29(1): 167-173.

[14] Radhika, P.R., Nair, R.A.S., Veena, G. (2019). A comparative study of lung cancer detection using machine learning algorithms. 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, pp. 1-4

[15] Roy, K., Chaudhury, S.S., Burman, M., Ganguly, A., Dutta, C., Banik, R. (2019). A comparative study of lung cancer detection using supervised neural network. 2019 International Conference on Opto-Electronics and Applied Optics (Optronix), Kolkata, India, pp. 1-5

[16] Song, D., Zhukov, T.A., Markov, O., Qian, W., Tockman, M.S. (2012). Prognosis of stage I lung cancer patients through quantitative analysis of centrosomal features. 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, pp. 1607-1610.

[17] Richter, A.N., Khoshgoftaar, T.M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artificial Intelligence in Medicine, 90: 1-14.

[18] Nishio, M., Sugiyama, O., Yakami, M., Ueno, S., Kubo, T., Kuroda, T., Togashi, K. (2018). Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning. PLoS ONE, 13(7): e0200721.

[19] Kadir, T., Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. Transl Lung Cancer Res, 7(3): 304-312.