Lung Cancer Prediction Using SHAP and Deep Learning Techniques

1st Shubhamkumar Pandit
Department of Computer Engineering
Marwadi University
Rajkot, India
panditshubham283@gmail.com

2nd Neeharika Joshi
Department of Computer Engineering
Marwadi University
Jamnagar, India
joshineeharika981@gmail.com

3rd Shailendrasinh Chauhan
Department of Computer Engineering
Marwadi University
Rajkot, India
Shailendrasinh.chauhan@marwadieducation.edu.in

***Abstract*— The most common cause of cancer-related deaths in the world is lung cancer; this is mainly because of late diagnosis and the little available means of early diagnosis. The present research offers a multimodal, explainable AI-based model of predicting and classifying lung cancer. There were 5,000 clinical records of 18 medical attributes and CT scan image of benign, malignant and normal tumors were used. Random Forest and Logistic Regression machine learning algorithms were used with clinical data, and deep learning networks, like DenseNet121, were used with images. The models had great predictive power, with DenseNet121: the accuracy was 99% and the precision and recall were 99.45 and 99.44 respectively. To overcome the absence of transparency in the traditional methods, SHAP and Grad-CAM were combined to offer explainable feature-level features and heatmap visualization, which guaranteed clinical trust. The findings indicate that multimodal data with explainable AI is highly promising in regards to improving early diagnosis, reliability, and usability in healthcare settings.**

Keywords— multimodal mental health detection, BERT, CNN-LSTM, YOLOv8n, facial expression analysis, acoustic analysis.

I. INTRODUCTION

One of the worldwide most common and deadliest cancers is lung cancer which takes the lives of more than 2.2 million people each year in the form of new cases and 1.8 million in the form of deaths. The latter is viewed as a major cause of the high death rate since the majority were diagnosed only at the advanced stage of the disease when no chances at treatment remained and the survival rates decreased drastically. Early decision/early identification is thus essential in enhancing patient outcomes, but there are a number of limitations to the current diagnostic methods. Conventional approaches such as manual interpretation of CT-images and rule-based evaluation of clinical risk factors are time-consuming, error-prone and extremely reliant on the experience of clinicians and radiologists.

Recent developments in the field of artificial intelligence (AI) and machine learning (ML) have shown considerable potential in medical diagnostics, providing a high-quality predictive projection, and capability to have a large-scale analysis of clinical and imaging data. There is however a major weakness among most of the current AI models: due to their lack of transparency, which is a major problem given their tendency to be black box, and their inability to effectively combine multimodal data sources. These drawbacks inhibit their clinical use since healthcare professionals are not only expected to make accurate predictions but also to obtain interpretable and actionable insights.

To overcome these difficulties, this paper proposes a multimodal lung cancer prediction model consisting of a mixture of clinical data and CT scan images through state-of-the-art machine learning and deep learning algorithms. Besides the high classification performance, the system has explainable AI (XAI) methods like SHAP (SHapley Additive Explanations) on clinical features and Grad-CAM on CT scans, which have explanations that are transparent, feature-level, and visual explanations of predictions. The method fills the gap between predictive performance and interpretability and allows increased trust and usability in clinical settings. Such combination of multimodal data, explainability, and real-time deployment makes this framework a developmental tool to assist in early diagnosis and informed medical decision-making in lung cancer care.

Even though the current machine learning methods promise to yield good results, various researchers have noted their flaws. Most of these works are based on small or unbalanced sets of data and restrict the extrapolation of findings and frequently result in overfitting. Moreover, the greater part of the previous studies has tested the models in a controlled experimental setting, whereas they have not been tested in clinic conditions. Although powerful, imaging-based models tend to ignore contextual patient data including smoking history, age, and family history, among others, which is crucial to determining risk. On the other hand, the text-based clinical models do not take note of morphological patterns that can be seen in CT scans, which leads to lack of complete diagnostic information.

The other essential gap is the inability in the current AI systems to be explained. As with conventional deep learning and ensemble models, high accuracy may not be able to support its choices to clinicians. In sensitive domains like healthcare, such opacity reduces trust and hinders adoption. The necessity of explainable AI (XAI) is demonstrated by the fact that a diagnosis with no interpretable evidence may prompt unwillingness to make treatment decisions. Additional explanations related to clinical data importance of features or region-based heatmap of CT scans can bring significant interpretation, as doctors can test the predictions against medical experience and develop trust in AI-based diagnostic assistance.

These are the constraints that we will seek to address in this work through building a multimodal and explainable AI-based lung cancer prediction model. The system utilizes a clinical data of 5,000 patient records comprising of 18 attributes and CT scan images that were classified as benign, malignant and normal. Clinical data is used with random Forest and Logistic Regression, whereas DenseNet121, the

state-of-the-art CNN architecture, is used with imaging data. SHAP values can be relied on to determine the contribution made by each clinical feature whereas Grad-CAM indicates suspicious areas on the lung images. This framework will provide a practical, reliable, and transparent instrument to detect early lung cancer due to its combination of predictive accuracy with interpretability and implementation of the solution using a web application based on FastAPI.

II. Literature Review/Survey

The use of machine learning and artificial intelligence in lung cancer prediction has received more and more attention over the last few years, and various works have investigated algorithms to analyse clinical and imaging data. Murthy Nimmagadda et al. (2024) used machine learning classifiers to predict the lung cancer risk, but the authors showed a better accuracy but only based on clinical datasets with low interpretability. On the same note, Bhuiyan et al. (2024) explored predictive modelling as one of their applications in the field of public health, noting that AI may effectively forecast early-stage diagnosis, but their model did not multimodally combine clinical and imaging data, limiting the scope of diagnosis.

Pathan et al. (2024) presented explainable ai models of lung cancer prediction and used interpretability methods to predict lung cancer on clinical data. Although this work covered transparency, it did not go beyond explaining image-based predictions, so there is a gap in multimodal interpretability. Chaturvedi et al. (2021) concentrated on comparative studying of machine learning models in classification tasks, but the studies were constrained by the imbalance of the datasets, the lack of feature-level explanations, and the unwillingness to deploy them in the clinical setting.

Ensemble and optimization methods have been discussed by other researchers in order to enhance model accuracy. An example is that Bortty et al. (2024) used sophisticated optimization methods to predict risks, and their models remained more of black boxes with no clinical-focused insights. Ensemble-based prediction systems were developed by Mamun et al. (2022), who, however, were restricted to textual data and failed to use CT scans or explainability frameworks. On the same note, Moon and Jetawat (2024) tried KNN to classify lung cancer, yet the approach was not scaled well and did not work well with high-dimensional data.

Kadir and Gleeson (2018) identified machine learning in combination with advanced imaging techniques to detect lung cancer in the field of imaging. Although they were effective in the detection of the malignancies, their method was not integrated with the clinical records, which is a critical point in diagnosis in a holistic manner. Agarwal et al. (2022) and Yunanto (2020) provided comparative studies of the ML algorithms, and they were more of an academic activity that did not show clinical viability or implementation in real-world systems.

Altogether, existing literature demonstrates the promise of AI in predicting lung cancer but notes that it is battery-based, lacks strong explainability, is overfitting because of small or skewed datasets, and has few avenues to clinical implementation. These deficiencies highlight why a multimodal, interpretable, and comprehensive system is needed that can be highly accurate and assure clinical trust-which is the specific purpose of the proposed invention.

III. Proposed Methodolody

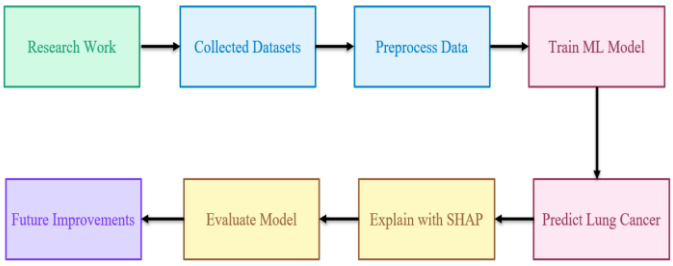

Fig. 1 Work Flow

A. Research and Data Collection

The study started with a review of the literature to determine gaps like the use of data that are only in one modality and the inability to interpret it. In order to address them, two datasets were gathered, including a clinical dataset of 5,000 records of 18 attributes that included demographics, lifestyle factors, medical indicators, and symptoms, and a CT scan dataset categorized into benign, malignant, and normal cases. The case data collected was clinical data that presented the structured information about the patients whereas the CT data given presented a visual presentation of the disease. Both datasets were quality- and balance-pre-processed to build a multimodal basis and train machine learning and deep learning models to predict lung cancer accurately and explainable.

B. Data Preprocessing and Visualization

Data sets collected were preprocessed to enhance quality and consistency. In the clinical data, missing values were treated, categorical attributes coded, and numerical ones normalized in order to format the records to machine learning models. In the case of the CT scan images, resizing, normalization and augmentation were used to conduct preprocessing to enhance robustness and minimize overfitting.

Visualization methods were used to gain a better insight into the datasets. Visualization of clinical data identified relationships between smoking and pulmonary disease, risk distribution by age, and the presence of such symptoms as breathing-related problems and sore throat. In the case of CT scans, sample images were visualized by the benign, malignant, and normal groups, which give the clear insights into the dataset diversity and allow the design of CNN-based classification models. These preprocessing and visualization procedures laid a solid ground in making sound and interpretable predictions.

Also, the significance of clinical attributes was subsequently confirmed using SHAP-based feature plots and heat maps, whereas Grad-CAM visualizations revealed suspicious areas of CT scans. These not only contributed to better model interpretability, but also to clinical relevance, providing greater overall reliability to the proposed framework.

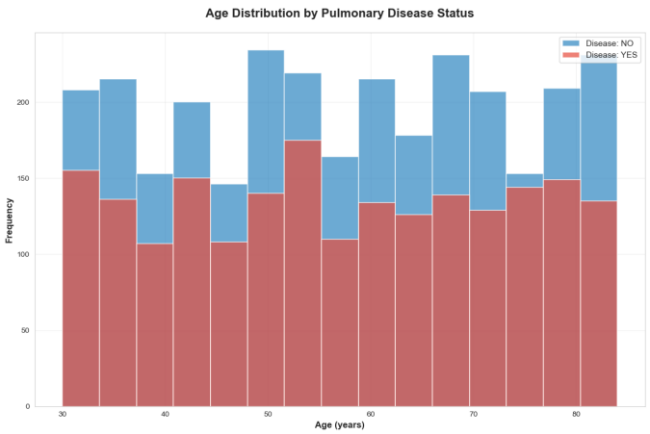

Fig. 2 Age Distribution by Pulmonary Disease Status (Textual)

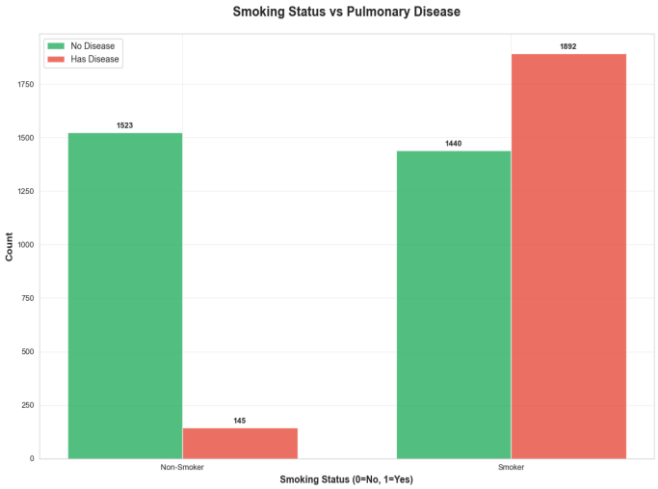

Fig 3 Smoking Status vs Pulmonary Disease (Textual)

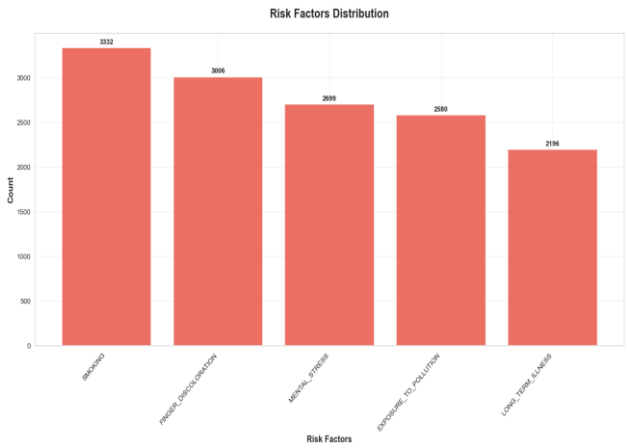

Fig 4 Risk Factor Distribution (Textual)

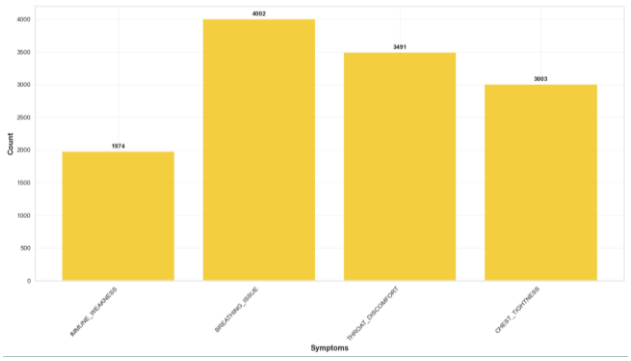

Fig 5 Symptoms Distribution (Textual)

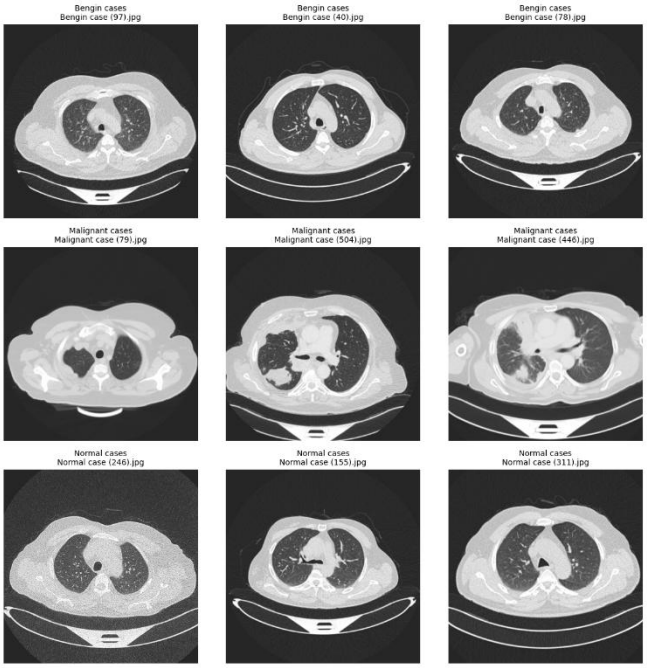

Figure 5 Sample Data of CT-Scan based image dataset

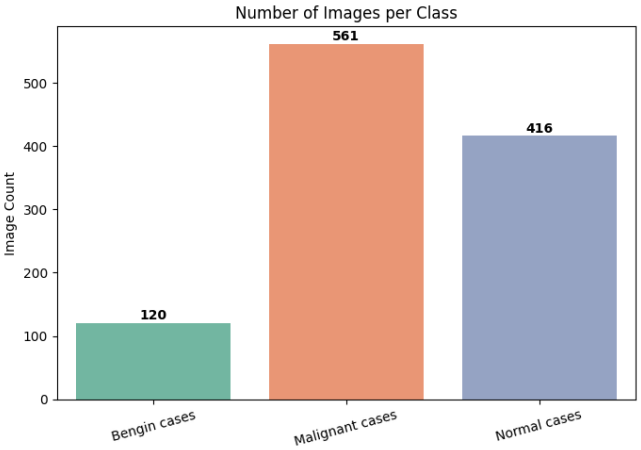

Figure 6 Number of Images per Class

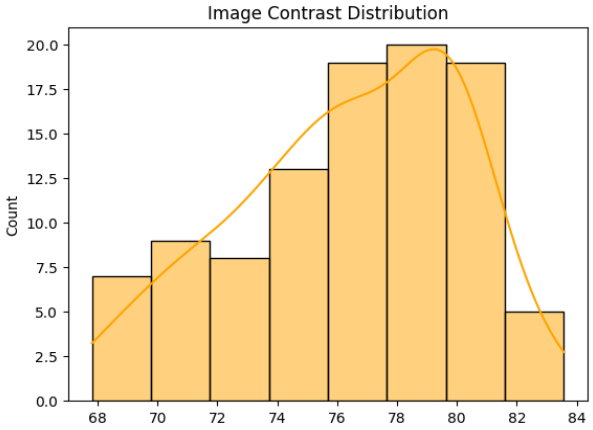

Figure 7 Image Contrast Distribution

C. Model Selection and Implementation

In cases of clinical data, several machine learning algorithms were tested to determine the most useful models to be used in classification. Algorithms like Random Forest, Logistic Regression, Gradient Boosting, AdaBoost, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes and Decision trees were tested and compared in terms of important performance metrics (accuracy, precision, recall and F1-score). Random Forest was the most successful of these, offering balanced performance on all measures and being able to cope with feature variability.

In case of imaging data, the use of advanced Convolutional Neural Networks (CNNs) was chosen because it has been demonstrated in the field of medical image classification. DenseNet121, ResNet50, and EfficientNet-B0, were trained on the CT scan dataset. As was found in comparative analysis, DenseNet121 provided the best recurring performance with the highest accuracy, 99% and a precision and recall of 99.45 and 99.44 respectively, thus making it the choice of deployment.

The implementation was carried out using TensorFlow/Keras for deep learning and scikit-learn for machine learning models. To ensure scalability and real-time usability, the models were integrated into a FastAPI backend and connected with a web-based frontend for clinical deployment. This selection and implementation strategy ensured that the system not only achieved high accuracy but also maintained robustness, interpretability, and practical applicability in healthcare environments.

D. Model Evaluation

The performance of the proposed models was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. These metrics are defined as follows:

Accuracy:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots\ldots (1)$$

Precision:

$$precision = \frac{TP}{TP+FP} \quad \ldots\ldots (2)$$

Recall:

$$precision = \frac{TP}{TP+FN} \quad \ldots\ldots (3)$$

F1-score:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \ldots\ldots (4)$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

In the case of the clinical dataset, the best balance between these metrics was obtained with the help of the Random Forest as it diagnosed both disease-positive and disease-negative cases and minimized the number of misclassifications. The confusion matrix showed the good results, and SHAP analysis demonstrated that the most significant predictive features were smoking history, discomfort in the throat, and pollution exposure.

Three CNN models, namely DenseNet121, ResNet50, and EfficientNet-B0, were compared on the CT scan data. The highest accuracy of 99% and precision of 99.45, recall 99.44, and F1-score of 99.44, showed the best results on denseNet121 where the network identified almost all benign, malignant, and normal cases. The predictions were further supported by the visualizations of grad-CAM, which pointed to the abnormal lung regions in the malignant scans and confirmed the non-abnormal regions in normal cases.

In general, the results of the evaluation reveal that the interaction of Random Forest in the clinical data and DenseNet121 in the imaging data creates an effective multimodal framework that exhibits a better predictive and clinical interpretability capacity.

IV. Results and Discussion

In case of the textual dataset, the confusion matrix (Figure 1) shows that the Random Forest model is effective. Individually, 544 negative(NO) and 361 positive (YES) cases were rightly classified, and only 95 cases were misclassified (49 NO wrongly classified as YES, 46 YES wrongly classified as NO). This shows that there is equal predictive strength in both classes. SHAP analysis (Figure 11) also provided an insight of how each feature contributed to model choices. This was because critical factors like smoking, pollution exposure, discomfort of the throat and breathing problems had significant impacts on predictions but those characteristics like age and family history had moderate impacts. This interpretability provides transparency during the decision making process which is a major condition to clinical adoption.

In the case of the CT-scan dataset, the confusion-matrix (Figure 12) indicates the best performance of the DenseNet121 model that correctly classified 23 benign, 113 malignant, and 82 normal cases with two misclassifications. The model was 100% accurate in the cases of malignancies proving that such a model has a highly discriminative effect with regard to identifying high-risk patients. Figure 13 demonstrated, through visual outputs of Grad-CAM and image classification using CT images, the interpretability of the deep learning model with heatmaps and localized areas showing agreement with radiologically important abnormal areas. Such results do not just confirm the accuracy of predictions but also give visual confirmation to clinical practitioners making them have more trust in the model.

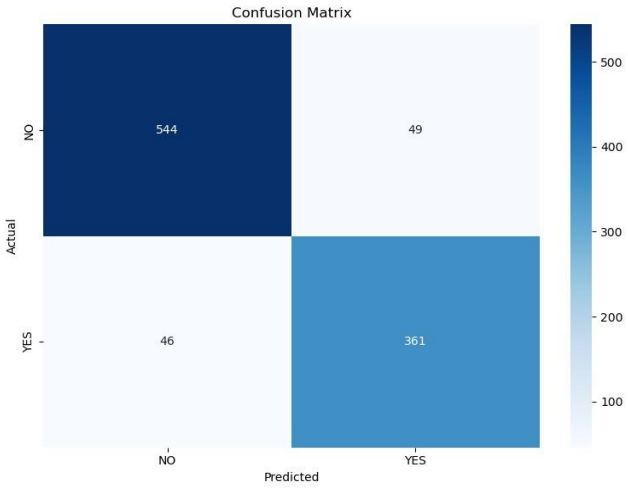

Figure 8 Confusion Matrix of Textual Data

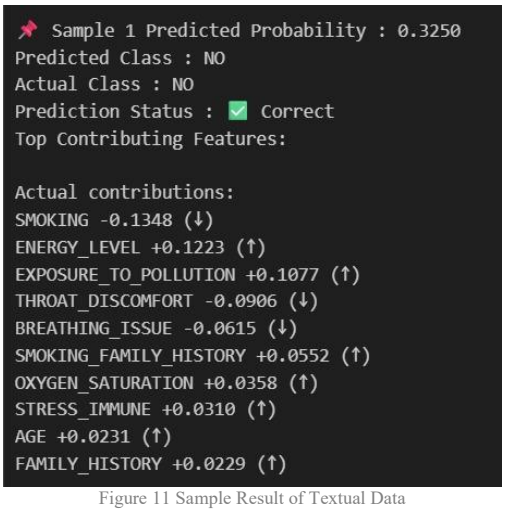

Figure 11 Sample Result of Textual Data

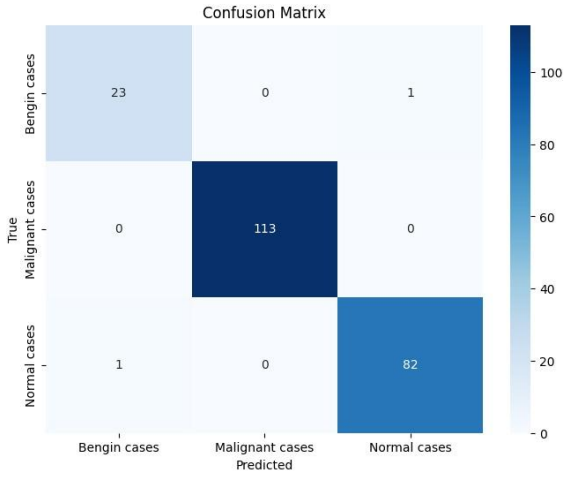

Figure 12 Confusion Matrix of CT-Scan based Data

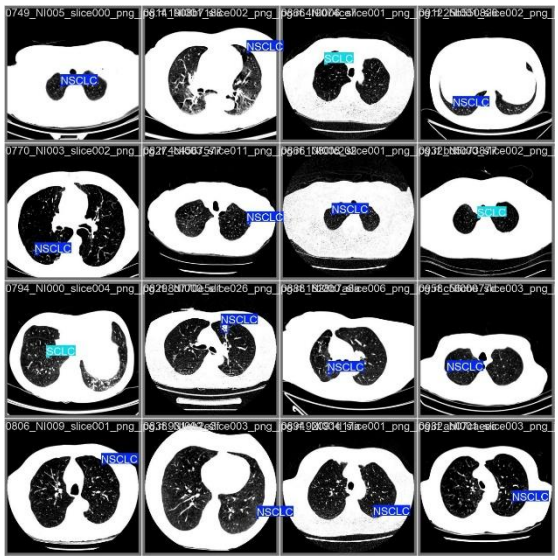

Figure 13 Sample Result of CT-Scan Based Data

V. Future Directions

Despite high accuracy and interpretability of the proposed multimodal framework, there are a number of directions that can be enhanced. A possible line of development would be the creation of a multimodal system that integrates to process clinical data and CT scan images at the same time instead of assessing them individually, to enhance robustness and the reliability of the decisions. Along with this, evaluating the more sophisticated ensemble and hybrid models can also be useful in improving predictive performance through the integration of the benefits of several algorithms.

The other direction that is significant is to extend the dataset by conducting real-life clinical trials and partnerships with healthcare organizations. Such would not only support system validation in different patient populations, but would also overcome dataset imbalance and enhance generalizability. Furthermore, the application to other diseases and cancers of the lungs could be used to apply the framework to a larger range of diseases and cancers in medical diagnostics.

Lastly, attention will be paid to deployment improvements, such as integration with cloud-computing systems, electronic health records (EHR) programs, and telemedicine applications that can make large-scale deployment of the assistive burden in urban hospitals, and in rural healthcare facilities. The use of user-friendly interfaces, real-time monitoring dashboards will enhance clinical utility even further so that, once the system is integrated, it will stop being just a research prototype but a scalable, practical healthcare solution.

VI. Conclusion

The paper is based on a multimodal approach to the early detection and diagnosis of lung cancer in the two forms: clinical and CT scan data. The system was able to make predictions with a very high degree of reliability and high interpretability by combining machine learning (Random Forest) with textual patient data and deep learning (DenseNet121) with image-based classification. Findings indicated that Random Forest yielded good results to detect important risk factors (smoking and exposure to pollution), whereas DenseNet121 produced almost perfect identity of malignant cases, which was confirmed by the use of confusion matrices and Grad-CAM visualizations.

The invention suggested does not only exceed the state-of-the-art approaches in accuracy and efficiency but also solves the black-box problem with SHAP and heatmap-based interpretability, which will establish trust among clinicians. Moreover, the system is accessible and expandable to the real-world healthcare setting, and it has a potential of being integrated into diagnostic centres, hospitals, and telemedicine platforms.

Acknowledgment

The authors would like to express their sincere gratitude to their faculty mentors and project supervisors for their continuous guidance, constructive feedback, and encouragement throughout the development of this work. We also acknowledge the support of our department and institution for providing the necessary resources, computational facilities, and academic environment to successfully carry out this research. Special thanks are extended to open-source communities and publicly available datasets that made this study possible. Finally, we are grateful

to our peers and colleagues for their valuable suggestions and moral support during the project.

REFERENCES

[1] S. N. Tisha and S. A. Ani, "Predictive Insights: Empowering Early Detection of Lung Cancer Using Machine Learning Excellence," in 2024 IEEE Region 10 Symposium, TENSYMP 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/TENSYMP61132.2024.10752136.

[2] S. Murthy Nimmagadda, K. Likhitha, G. Srilatha, and S. M. Sree, "Lung Cancer Prediction and Classification Using Machine Learning Algorithms," in Proceedings - 2024 International Conference on Expert Clouds and Applications, ICOECA 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1012–1015. doi: 10.1109/ICOECA62351.2024.00176.

[3] Mohammad Shafiquzzaman Bhuiyan et al., "Advancements in Early Detection of Lung Cancer in Public Health: A Comprehensive Study Utilizing Machine Learning Algorithms and Predictive Models," Journal of Computer Science and Technology Studies, vol. 6, no. 1, pp. 113–121, Jan. 2024, doi: 10.32996/jcsts.2024.6.1.12.

[4] S. P. Maurya, P. S. Sisodia, R. Mishra, and D. P. singh, "Performance of machine learning algorithms for lung cancer prediction: a comparative approach," Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-58345-8.

[5] Joy Chakra Bortty et al., "Optimizing Lung Cancer Risk Prediction with Advanced Machine Learning Algorithms and Techniques," Journal of Medical and Health Studies, vol. 5, no. 4, pp. 35–48, Oct. 2024, doi: 10.32996/jmhs.2024.5.4.7.

[6] K. Moon and A. Jetawat, "Predicting Lung Cancer with K-Nearest Neighbors (KNN): A Computational Approach," Indian J Sci Technol, vol. 17, no. 21, pp. 2199–2206, May 2024, doi: 10.17485/IJST/v17i21.1192.

[7] R. K. Pathan, I. J. Shorna, M. S. Hossain, M. U. Khandaker, H. I. Almohammed, and Z. Y. Hamd, "The efficacy of machine learning models in lung cancer risk prediction with explainability," PLoS One, vol. 19, no. 6 June, Jun. 2024, doi: 10.1371/journal.pone.0305035.

[8] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in 2022 IEEE World AI IoT Congress, AIIoT 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 187–193. doi: 10.1109/AIIoT54504.2022.9817326.

[9] S. Agarwal, S. Thakur, and A. Chaudhary, "Prediction of Lung Cancer Using Machine Learning Techniques and their Comparative Analysis," in 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICRITO56286.2022.9965052.

[10] P. Chaturvedi, A. Jhamb, M. Vanani, and V. Nemade, "Prediction and Classification of Lung Cancer Using Machine Learning Techniques," IOP Conf Ser Mater Sci Eng, vol. 1099, no. 1, p. 012059, Mar. 2021, doi: 10.1088/1757-899x/1099/1/012059.

[11] A. Ampuh. Yunanto, 2020 International Electronics Symposium : September 29-30th 2020, Surabaya, Indonesia. IEEE, 2020.

[12]] T. Kadir and F. Gleeson, "Lung cancer prediction using machine learning and advanced imaging techniques," Jun. 01, 2018, AME Publishing Company. doi: 10.21037/tlcr.2018.05.15.