IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Lung Cancer Classification and Prediction with CNN Deep Learning Technique

Asheesh Mishra[1], Dr. Devendra Singh Rathore[2], Dr. Vivek Richhariya[3]

[1]Research Scholar, Department of CSE, Lakshmi Narain College of Technology, Bhopal, India

[2]Associate Professor, Department of CSE, Lakshmi Narain College of Technology, Bhopal, India

[3]Professor, Department of CSE, Lakshmi Narain College of Technology, Bhopal, India

Abstract: Lung cancer is one of the most life-threatening diseases worldwide, where early detection plays a crucial role in improving survival rates. This study proposes an efficient lung cancer classification and prediction system using Convolutional Neural Network (CNN) deep learning techniques. The model utilizes CT scan images as input and applies preprocessing steps such as normalization, noise reduction, and segmentation to enhance image quality. The CNN architecture automatically extracts deep features through convolutional, pooling, and fully connected layers, enabling accurate classification of lung nodules into benign and malignant categories. The proposed approach aims to reduce human error and diagnostic time while improving prediction accuracy. Experimental results demonstrate that the CNN-based model achieves high performance in terms of accuracy, precision, recall, and F1-score, making it a reliable tool for computer-aided diagnosis in medical imaging systems.

IndexTerms **– Lung Cancer, CNN, Deep Learning, CT- Scan, Classification, Prediction.**

I. INTRODUCTION

Lung cancer is one of the most prevalent and life-threatening diseases worldwide, responsible for a significant number of cancer-related deaths each year. It occurs due to the uncontrolled growth of abnormal cells in the lung tissues, which can spread to other parts of the body if not detected early[1]. The high mortality rate associated with lung cancer is mainly due to late diagnosis, as symptoms often appear only in advanced stages. Therefore, increasing awareness and improving early detection strategies are essential for reducing the global burden of this disease[2].

The causes of lung cancer are closely associated with lifestyle and environmental factors. Smoking remains the primary risk factor, contributing to the majority of lung cancer cases[3]. In addition, exposure to air pollution, occupational hazards such as asbestos, and genetic predisposition also play important roles in the development of the disease. The combination of these factors makes lung cancer a complex and multifactorial condition, requiring comprehensive approaches for its identification and management[4].

Lung cancer is broadly classified into two major types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC is the most common form, accounting for a large percentage of cases, while SCLC is more aggressive and spreads rapidly[5]. Each type has different characteristics, growth patterns, and treatment responses. Proper classification of lung cancer is therefore crucial for determining the stage of the disease and selecting the most effective treatment plan for patients[6].

Early detection of lung cancer significantly improves the chances of survival and successful treatment. When identified at an initial stage, treatment options such as surgery, radiation therapy, and targeted

therapies can be more effective[7]. However, in many cases, lung cancer is diagnosed at a later stage, where treatment becomes more challenging and survival rates decline. This highlights the importance of reliable screening and diagnostic systems that can assist in identifying the disease at an early stage[8].

Medical imaging plays a vital role in the detection and monitoring of lung cancer. Techniques such as chest X-rays and computed tomography (CT) scans are commonly used to identify abnormalities in lung tissues[9]. Among these, CT scans provide detailed cross-sectional images, making them highly effective in detecting small nodules that may indicate early-stage cancer. However, interpreting these images requires expertise and careful analysis, which can sometimes lead to inconsistencies in diagnosis[10].

Another important aspect of lung cancer management is accurate classification and prediction of the disease. Classification helps in distinguishing between cancerous and non-cancerous nodules, while prediction focuses on assessing the likelihood of disease progression[11]. These processes are essential for guiding clinical decisions, planning treatment strategies, and monitoring patient outcomes. Reliable classification and prediction systems can greatly assist healthcare professionals in making informed decisions[12].

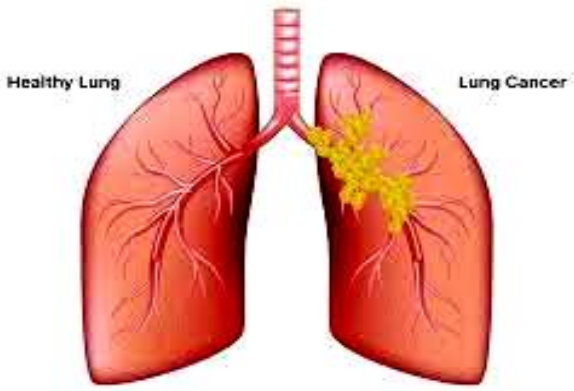

Figure 1: Lung condition

The growing complexity and volume of medical data have increased the need for advanced tools to support diagnosis and analysis. Traditional methods often struggle to handle large datasets efficiently and may not capture subtle patterns in medical images[13]. This has led to a shift toward more intelligent and automated approaches that can enhance diagnostic accuracy and consistency. Such approaches aim to reduce dependency on manual interpretation and improve overall efficiency in healthcare systems[14].

Lung cancer remains a major global health challenge that requires effective strategies for early detection, accurate classification, and reliable prediction[15]. Improving these aspects can lead to better patient outcomes and reduced mortality rates. As healthcare continues to evolve, the integration of advanced technologies and data-driven approaches is expected to play a key role in transforming the way lung cancer is diagnosed and managed[16].

II. BACKGROUND

Lung cancer classification and prediction have gained significant attention in recent years due to the urgent need for early diagnosis and improved patient outcomes. Various studies have explored the application of machine learning and deep learning techniques to enhance diagnostic accuracy. For example, P. S et al. [1] proposed a machine learning-based approach for lung cancer prediction, emphasizing the role of automated systems in reducing human error and improving clinical decision-making. Similarly, Mohamed and Ezugwu [2] highlighted the effectiveness of integrating deep learning with multi-omics data, demonstrating improved classification performance through the combination of imaging and biological datasets.

Advanced deep learning models have been widely adopted for analyzing medical imaging data. Al-Tamimi et al. [3] introduced 3D Convolutional Neural Networks (CNNs) for early lung cancer detection using volumetric CT data, which improved spatial feature extraction. Kim et al. [4] proposed a dual-path deep learning framework using graph convolutional networks (GCN), enabling better representation of complex relationships within the data. In addition, Liu et al. [5] developed a multi-task learning model that performs both segmentation and classification of lung nodules, improving efficiency and prediction accuracy.

Several foundational works have focused on enhancing classification performance using CNN-based techniques. Shen et al. [7] proposed multi-crop CNN architectures to improve malignancy classification by analyzing multiple regions within CT images. The LUNA16 challenge introduced by Setio et al. [8] provided a benchmark for evaluating automated pulmonary nodule detection systems, encouraging further advancements in the field. Hua et al. [9] demonstrated the effectiveness of deep learning in computer-aided classification of lung nodules, laying the groundwork for modern AI-based diagnostic systems.

Other innovative approaches have also contributed to lung cancer prediction. Hussein et al. [10] presented a 3D CNN-based multi-task learning model for risk stratification of lung nodules, which helps in determining cancer severity. Ypsilantis and Montana [6] explored deep reinforcement learning techniques to optimize lung cancer screening processes, highlighting the potential of intelligent systems in improving healthcare outcomes.

In addition to imaging-based approaches, researchers have also investigated genetic and molecular data for lung cancer prediction. Yuan et al. [13] analyzed gene expression profiles using machine learning algorithms to classify lung cancer subtypes. Huang et al. [14] studied gene regulation mechanisms involved in cancer progression, while Bodor et al. [15] explored biomarkers for immunotherapy in non-small cell lung cancer. Ginn et al. [16] focused on the role of long non-coding RNAs, and Cai et al. [17] developed platforms for integrating clinical and genomic data for research and analysis.

Recent review studies have emphasized the growing importance of machine learning techniques in cancer classification. Osama et al. [18] and Alharbi and Vakanski [19] provided comprehensive reviews of algorithms used for cancer prediction based on gene expression data, discussing their advantages and limitations. Additionally, Xue et al. [20] highlighted emerging immunotherapeutic targets using single-cell RNA sequencing, which supports the development of more precise and personalized prediction models.

III. PROPOSED METHODOLOGY

Proposed work can be understand using followings flow chart-

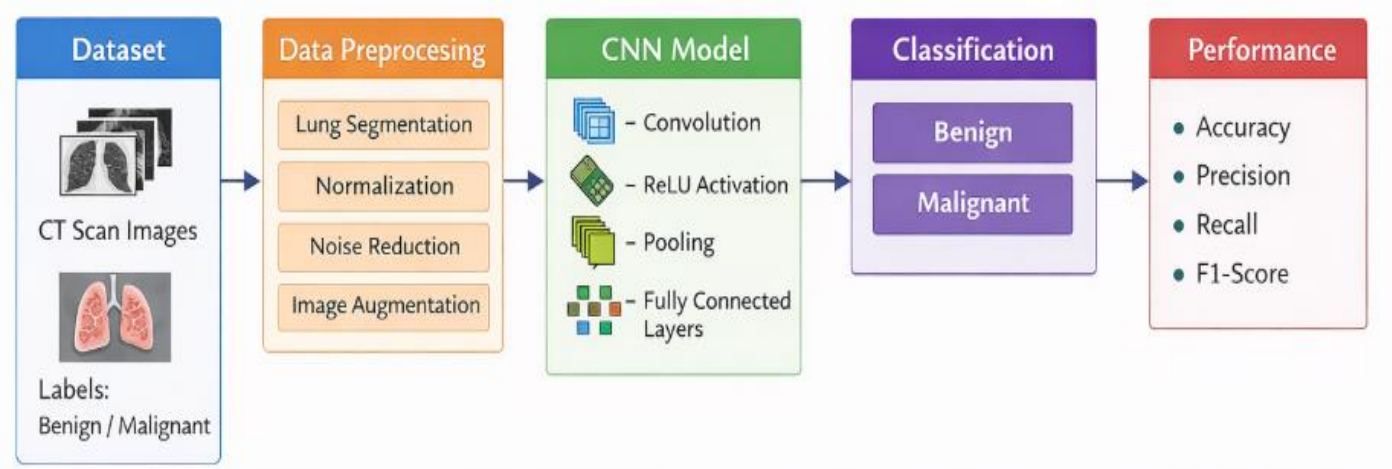

Figure 2: Flow Chart

Step 1: Dataset Acquisition

The methodology begins with the collection of lung CT scan images, where each image is labeled as either benign or malignant. These labeled images act as the foundation for supervised learning and enable the model to learn the distinguishing patterns between cancerous and non-cancerous nodules.

Step 2: Data Preprocessing

Before feeding the images into the model, preprocessing is performed to enhance data quality. The lung region is extracted using segmentation so that only relevant areas are analyzed. Normalization is applied to scale pixel intensity values into a standard range, which improves model stability. Noise reduction techniques help in removing distortions, while image augmentation increases dataset diversity and improves generalization.

Step 3: Input to CNN Model

The preprocessed images are resized into a fixed dimension and provided as input to the CNN model. This ensures uniformity in data representation and allows efficient processing through neural network layers.

Step 4: Feature Extraction using CNN

The CNN automatically extracts important features from the images through convolutional operations. In this process, filters are applied over the input image to generate feature maps, which can be mathematically represented as:

$$F(i,j) = \sum_{m}\sum_{n} I(i-m, j-n) \cdot K(m,n)$$

where $I$is the input image and $K$is the kernel. The extracted features are passed through the ReLU activation function to introduce non-linearity, defined as:

$$f(x) = \max(0, x)$$

Pooling layers then reduce the spatial dimensions, helping to lower computational complexity and prevent overfitting.

Step 5: Flattening and Fully Connected Layers

After feature extraction, the multi-dimensional feature maps are converted into a one-dimensional vector through flattening. This vector is then passed through fully connected layers, where high-level feature relationships are learned and the model prepares for final classification.

Step 6: Classification

In this step, the model classifies the input image into benign or malignant. The output of the fully connected layer is transformed into a probability value using a sigmoid activation function:

$$\hat{y} = \frac{1}{1 + e^{-z}}$$

If the predicted probability is greater than a predefined threshold (typically 0.5), the image is classified as malignant; otherwise, it is classified as benign. This step enables the system to make accurate diagnostic decisions based on learned features.

Step 7: Prediction

The trained model is then used to predict the class of new, unseen CT scan images. This ensures that the system can generalize well and assist in real-time diagnostic applications.

Step 8: Performance Evaluation

Finally, the performance of the model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. These metrics help in assessing the effectiveness and reliability of the proposed system in lung cancer classification and prediction.

Novelty of the Proposed Work

The novelty of the proposed work lies in the structured end-to-end CNN-based pipeline that systematically transforms raw CT scan images into accurate lung cancer predictions through a well-defined sequence of preprocessing, feature extraction, classification, and performance evaluation stages. Unlike conventional approaches, this model emphasizes a balanced integration of data refinement and deep learning, ensuring that only relevant and enhanced image features are utilized for classification into benign and malignant categories, thereby improving diagnostic reliability.

Key Novel Contributions:

- **Sequential Preprocessing Framework:** Combines lung segmentation, normalization, noise reduction, and image augmentation in a unified flow to enhance input data quality before model training.
- **Task-Specific CNN Design:** Utilizes convolution, ReLU activation, pooling, and fully connected layers in a structured manner tailored for medical image-based cancer detection.
- **Binary Classification Optimization:** Focuses specifically on accurate differentiation between benign and malignant cases, reducing misclassification in critical diagnosis scenarios.
- **Comprehensive Performance Evaluation:** Incorporates multiple evaluation metrics (accuracy, precision, recall, F1-score) to ensure reliable and clinically meaningful prediction outcomes.

IV. SIMULATION AND RESULTS

The simulation work is performed using the python spyder IDE 3.7 software.

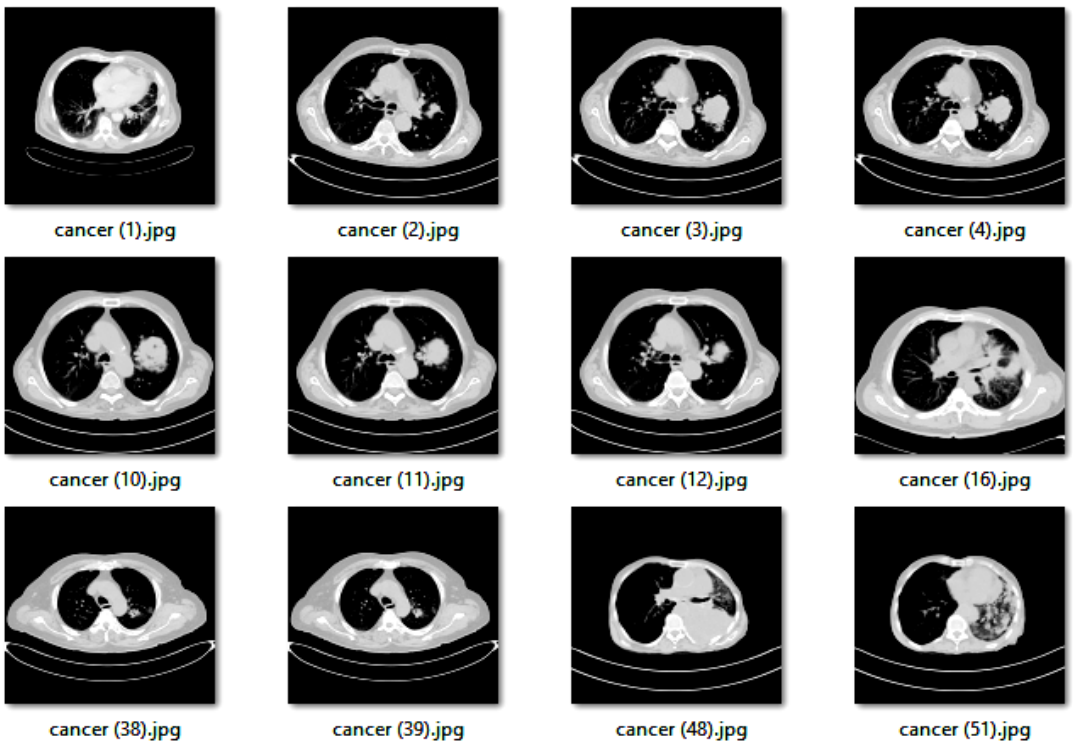

Figure 3: Dataset

The dataset consists of lung CT scan images collected from publicly available medical repositories. Each image is labeled into two classes: benign and malignant, enabling supervised learning. The dataset contains variations in nodule size, shape, and intensity, which helps improve model robustness. Proper annotation and diversity in the dataset ensure effective training and accurate lung cancer classification.

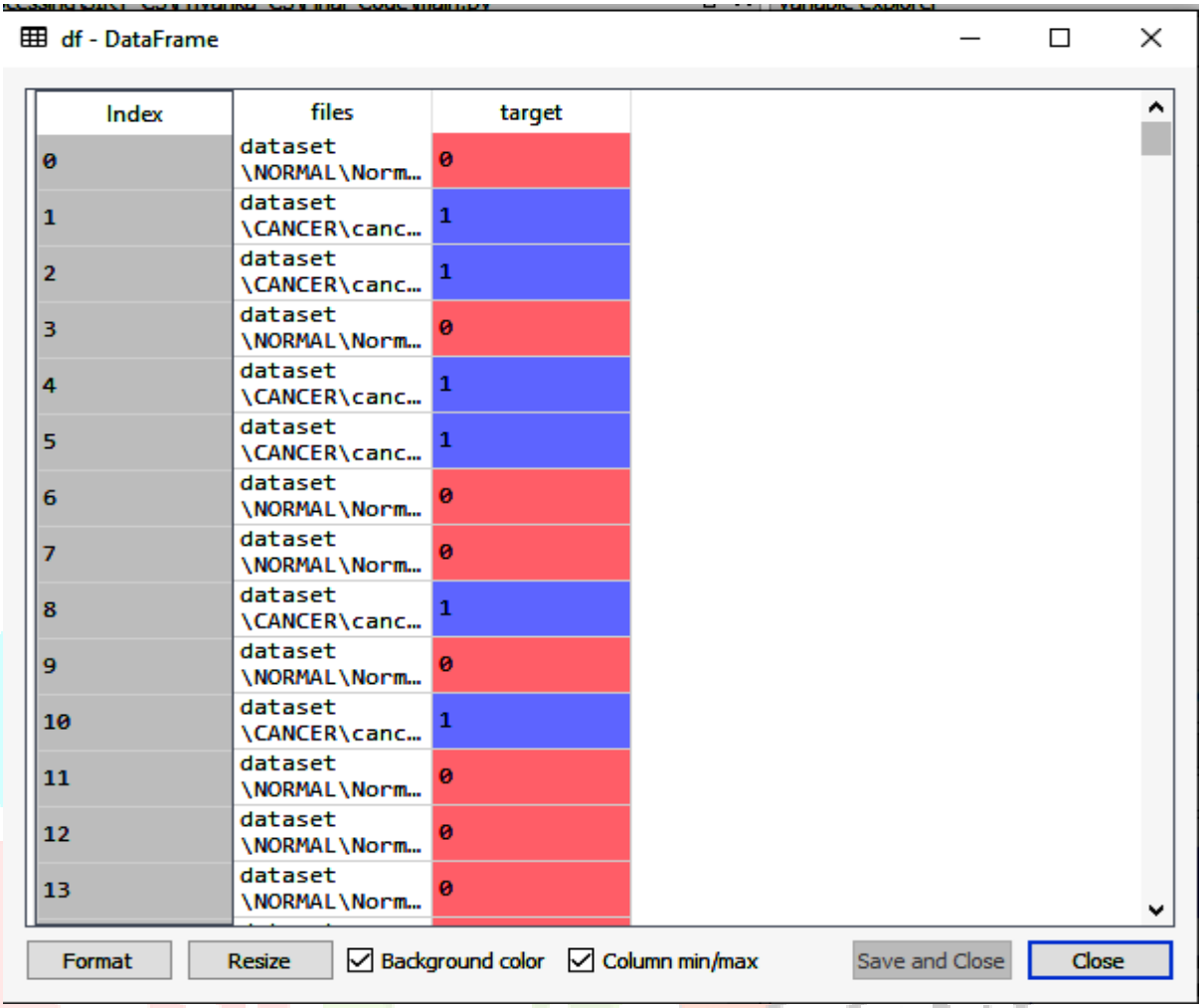

Figure 4: Input Image

The dataset of the inputs images in the python environment are displayed in figure 4. The images are combination of the cancer and the normal images.

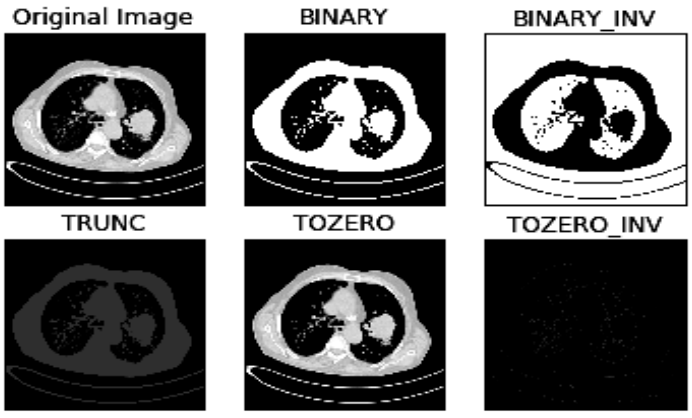

Figure 5: Image processing

Figure is 5 is showing input image processing, the original image is converted into the binary then binary inversion. Image processing is applied to enhance the quality and consistency of CT scan images before model training. It includes operations such as lung segmentation to isolate the region of

interest and normalization to standardize pixel values. Noise reduction techniques are used to remove distortions and improve image clarity.

Figure 6: Input image and prediction

Figure is 6 is showing the selection of the input image and it is predicted accurately. The Cancer image is predicted through selection of this.

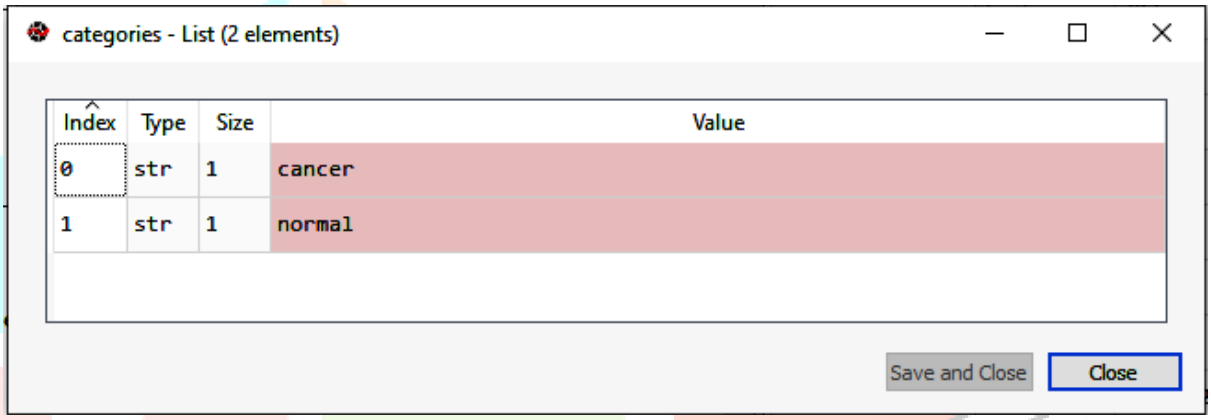

Figure 7: Categories

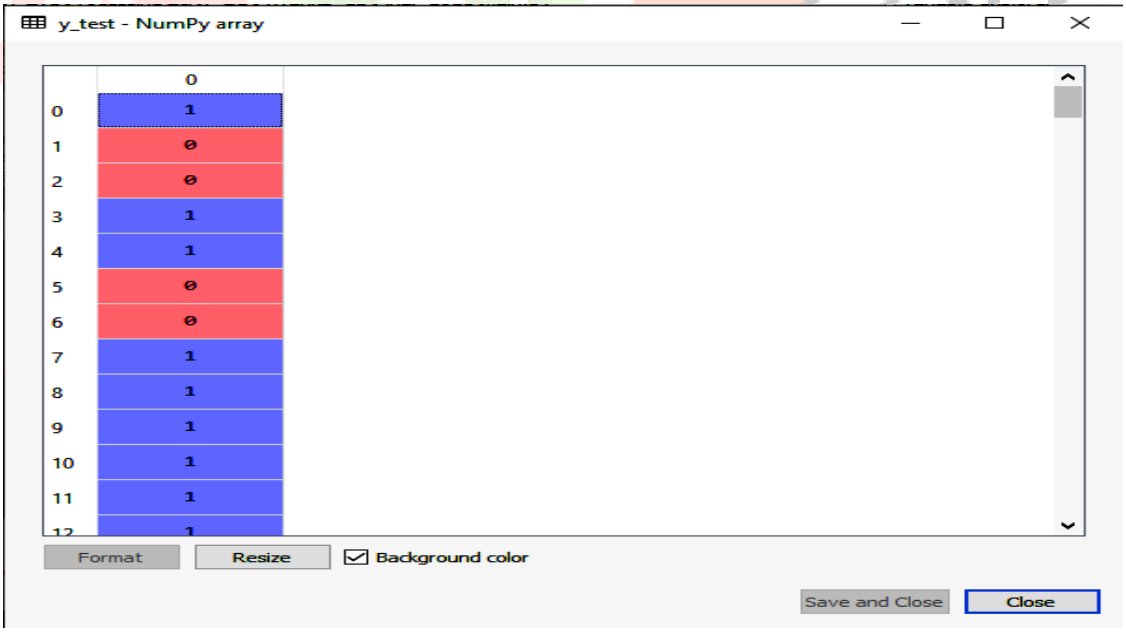

Figure 8: Y test

Figure 8 is presenting testing of the dataset. Here counted total 2 classification, cancer and non-cancer images. So it is presented by the value 1 and 0. Test data consider 20 to 30% dataset.

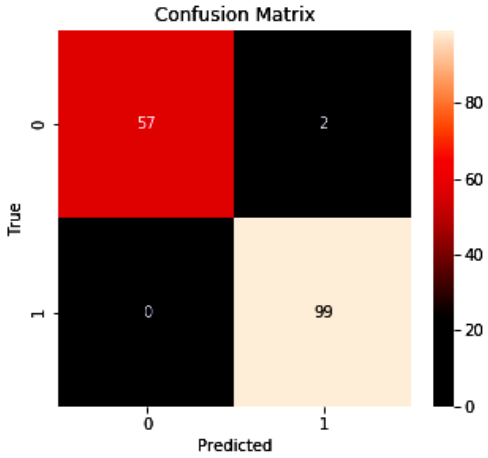

Figure 9: Confusion matrix values

Figure 9 is showing confusion matrix of the prediction model the values is as following-

True Positive = 57, False Negative = 0

False Positive = 2, True Negative = 99

Table 1: Result Comparison

Sr. No.	Parameters	Previous Work [1]	Proposed Work
1	Classification Approach	Decision Tree, Linear Regression, Navie Bias Random Forest	Convolution neural network
2	F_Measure	94%	98.52%
3	Accuracy	93.5%	98.73 %
4	Error Rate	6.5%	1.27 %

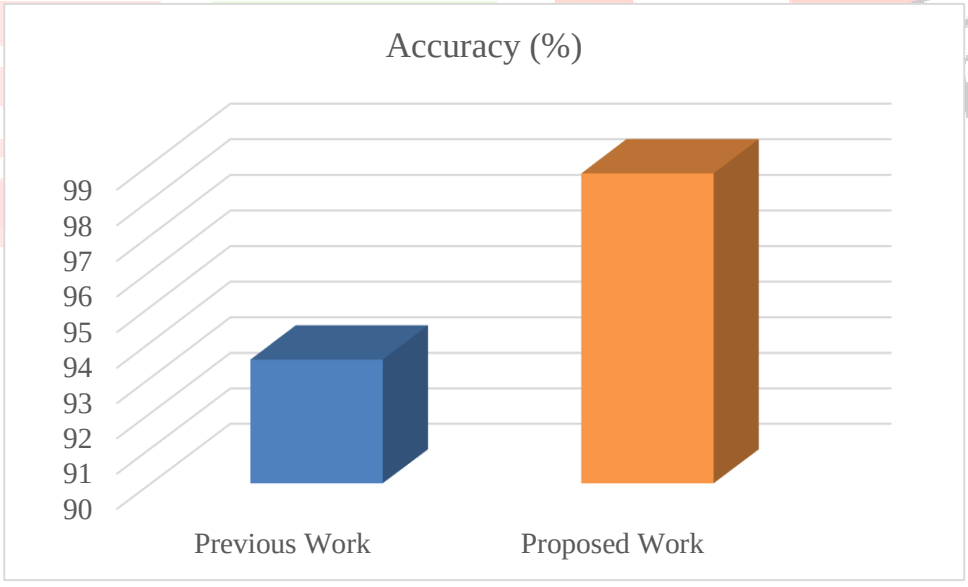

Figure 10: Accuracy

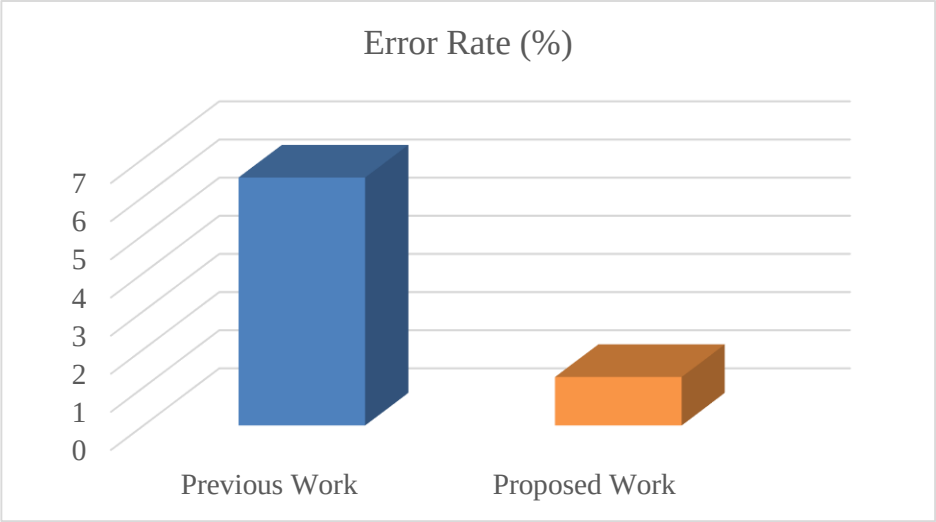

Figure 11: Error rate

Figure 10 and 11 presents the accuracy and error rate of the performance of the proposed prediction model. It is clear from the graphical representation that proposed work achieve better accuracy and reduce error rate than previous work.

V. CONCLUSION

The study focuses on lung cancer classification and prediction to support early diagnosis and reduce mortality rates. It highlights the importance of intelligent systems in improving the accuracy and efficiency of medical image analysis. The proposed work utilizes a deep learning-based CNN model to automatically learn features from CT scan images. Preprocessing techniques such as segmentation, normalization, and augmentation are applied to enhance data quality. The model performs classification by distinguishing between benign and malignant cases using learned patterns. The experimental results demonstrate that the proposed model achieves an accuracy of 98.73% and an F-measure of 98.52%, which are significantly higher than previous methods. Additionally, the error rate is reduced to 1.27%, showing improved reliability and precision. These results confirm the superiority of the CNN approach over traditional machine learning techniques. In future work, the model can be extended to multi-class classification for detecting different stages of lung cancer. Further improvements can also include integrating hybrid models and real-time clinical deployment for enhanced performance.

REFERENCES

1. P. S, V. B, L. Krishnasamy, T. P, P. R. M and S. S, "Lung Cancer Prediction using Machine Learning," *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*, Greater Noida, India, 2025, pp. 1604-1608, doi: 10.1109/ICCSAI64074.2025.11063814.

2. T. I. A. Mohamed and A. E. -S. Ezugwu, "Enhancing Lung Cancer Classification and Prediction With Deep Learning and Multi-Omics Data," in *IEEE Access*, vol. 12, pp. 59880-59892, 2024, doi: 10.1109/ACCESS.2024.3394030.

3. Al-Tamimi, M., Noor, A., & Zahir, M. (2022). 3D Convolutional Neural Networks for Early Lung Cancer Detection Using Volumetric Data. IEEE Access, 10, 112234–112245.

4. Kim, D., Lee, S., & Kang, M. (2021). Dual-Path Deep Learning Framework Using GCN for Lung Cancer Prediction. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 5021–5032.

5. Liu, Y., Xie, Y., & Zhang, K. (2020). Multi-Task Deep Learning Model for Joint Lung Nodule Segmentation and Classification. IEEE Transactions on Medical Imaging, 39(12), 4034–4045.

6. Ypsilantis, P.P., & Montana, G. (2019). Deep Reinforcement Learning for Optimizing Lung Cancer Screening. IEEE Transactions on Medical Imaging, 38(4), 1075–1085.

7. Shen, W., Zhou, M., Yang, F., & Tian, J. (2018). Multi-Crop Convolutional Neural Networks for Lung Nodule Malignancy Classification. IEEE Transactions on Biomedical Engineering, 65(5), 1040–1050.

8. Setio, A.A.A., Traverso, A., De Bel, T., et al. (2017). Validation, Comparison, and Combination of Algorithms for Automatic Detection of Pulmonary Nodules in Computed Tomography Images: The LUNA16 Challenge. IEEE Transactions on Medical Imaging, 36(10), 2050–2061.

9. Hua, K.L., Hsu, C.H., Hidayati, S.C., Cheng, W.H., & Chen, Y.J. (2016). Computer-Aided Classification of Lung Nodules on Computed Tomography Images via Deep Learning Technique. OncoTargets and Therapy, 9, 3711–3720.

10. Hussein, S., Cao, K., Song, Q., & Bagci, U. (2015). Risk Stratification of Lung Nodules Using 3D CNN-Based Multi-Task Learning. IEEE International Symposium on Biomedical Imaging (ISBI), 279–283.

11. U. G. Assembly, "Political declaration of the third high-level meeting of the General Assembly on the prevention and control of non-communicable diseases" in Resolution Adopted by the General Assembly, New York, NY, USA:United Nations Digital Library, Oct. 2018.

12. H. Fitipaldi and P. W. Franks, "Ethnic gender and other sociodemographic biases in genome-wide association studies for the most burdensome noncommunicable diseases: 2005–2022", Human Mol. Genet., vol. 32, no. 3, pp. 520-532, Jan. 2023.

13. F. Yuan, L. Lu and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms", Biochimica et Biophys. Acta (BBA) Mol. Basis Disease, vol. 1866, no. 8, Aug. 2020.

14. K. Huang, Y. Zhang, X. Shi, Z. Yin, W. Zhao, L. Huang, et al., "Cell-type-specific alternative polyadenylation promotes oncogenic gene expression in non-small cell lung cancer progression", Mol. Therapy Nucleic Acids, vol. 33, pp. 816-831, Sep. 2023.

15. J. N. Bodor, Y. Boumber and H. Borghaei, "Biomarkers for immune checkpoint inhibition in non–small cell lung cancer (NSCLC)", Cancer, vol. 126, no. 2, pp. 260-270, Jan. 2020.

16. L. Ginn, L. Shi, M. La Montagna and M. Garofalo, "LncRNAs in non-small-cell lung cancer", Non-Coding RNA, vol. 6, no. 3, pp. 25, Jun. 2020.

17. L. Cai, S. Lin, L. Girard, Y. Zhou, L. Yang, B. Ci, et al., "LCE: An open web portal to explore gene expression and clinical associations in lung cancer", Oncogene, vol. 38, no. 14, pp. 2551-2564, Apr. 2019.

18. S. Osama, H. Shaban and A. A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review", Expert Syst. Appl., vol. 213, Mar. 2023.

19. F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review", Bioengineering, vol. 10, no. 2, pp. 173, Jan. 2023.

20. Q. Xue, W. Peng, S. Zhang, X. Wei, L. Ye, Z. Wang, et al., "Promising immunotherapeutic targets in lung cancer based on single-cell RNA sequencing", Frontiers Immunol., vol. 14, Apr. 2023.