iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68691

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Detection of Lung Cancer Using Histopathology Based on CNN

P. Thirumalesh[1], V. V. Sai Dinesh[2], C. Vardhan Kumar[3], A. Rahul[4], Dr. Munipraveena Rela[5], Prof. M. E. Palanivel[6]
[1,2,3,4]CSE- Artificial Intelligence, Sreenivasa Institute of Technology and Management Studies in Chittoor, India
[5]Associate Professor, [6]Professor, CSE- Artificial Intelligence, Sreenivasa Institute of Technology and Management Studies in Chittoor, India

Abstract: ***Early diagnosis of lung cancer is a challenge, making it difficult to treat successfully. This paper discusses the creation of a Convolutional Neural Network (CNN)-based system for lung cancer prediction from Biopsy Scan. Deep learning holds great promise for medical image processing, and CNNs are particularly good at learning complex patterns. Our CNN system will seek to scan Biopsy scans and detect insidious differences consistent with early-stage lung cancer. It can potentially:***
Early detection of lung cancer can greatly enhance the chances of effective intervention. Automation of image analysis using CNNs saves medical professionals' time for diagnosis and treatment planning. CNNs show high accuracy in image recognition, which results in more accurate predictions.
It helps advance personalized medicine by facilitating earlier and more accurate diagnoses. In addition, it aids in targeted screening programs and improved allocation of resources, hence favorably influencing population health policies.
We assess our CNN system's functionality and consider its potential to revolutionize lung cancer diagnosis and treatment. By incorporating this technology into practice, we hope to enhance patients' outcomes as well as ensure an improved responsive and effective health system.

I. INTRODUCTION

Lung cancer is a major global cause of cancer mortality, and early detection and diagnosis are important for successful treatment and survival. Lung cancer was responsible for 1.8 million deaths in 2020, according to the World Health Organization (WHO), and it is the most frequent cause of cancer mortality worldwide.

A Biopsy scan is an x-ray test that uses a computer to produce clear images of the inside of your body. It takes images in several positions. The computer combines them to create a 3 dimensional (3D) picture. Doctors may employ a Biopsy scan to search for lung cancer. It may assist them to:

a. Diagnose and stage lung cancer.

b. Find out if cancer has spread to the liver, adrenal glands, lower neck or lymph nodes in the chest.

c. Learn about having a Biopsy scan.

d. PET-Biopsy scan: A PET-Biopsy scan is a combination of a Biopsy scan and a PET scan. PET stands for positron emission tomography. The PET scan uses a mildly radioactive medicine to highlight areas of your body where cells are more active than usual.

A Biopsy scan involves x-rays to produce detailed cross-sectional images of your body. A Biopsy scanner takes numerous pictures, not 1 or 2, like an ordinary x-ray, and then a computer reconstructs them to reveal a slice of the portion of your body that is being examined.

A Biopsy scan is more likely to detect lung tumors than standard chest x-rays. It is also able to determine the size, shape, and location of any lung tumors and can assist in locating enlarged lymph nodes that may contain cancer that has spread. This test can also be utilized to search for masses elsewhere in the body that may be caused by the spread of lung cancer.

A Biopsy scan for lung cancer can assist to demonstrate:

1) Where the cancer is in your lung.
2) If it has spread elsewhere in the body and to lymph nodes in the chest.
3) How aggressive (metabolic active) is the cancer.
4) Determine which is the most suitable treatment for your cancer.
5) Check if your cancer has recurred and plan radiotherapy treatment.

6) How well a cancer treatment is functioning.

II. LITERATURE SURVEY

1) Title: A Review of recent Lung Cancer Detection Methods based on Machine Learning Authors: Nawazat ahmed, Year: 2021 Summary: Many classifiers Have been used by the researchers in the literature such as: multi-layer perceptron (MLP),SVM, Navie Bayes, Neural Network, and Gradient, Boosted Tree, Decision Tree,k-nearest neighbors,multinomial random forest classifier.

2) Title: A Comprehensive Review on Lung Cancer Detection with Machine Learning Techniques: Systematic Study Authors: Debnath Bhattacharya,Year: 2020 Summary: The Main Obejective of this research paper is to investigate the accuracy levels of various machine learning algorithms. To find out the accuracy levels of various classifiers Based on the detection velocity for lung cancer using CT is 2.6 ten times greater than utilizing analogue radiography

A. Role Of Convolutional Neural Network

Convolutional Neural Network (CNN) is a specific deep learning technique well adapted for processing visual data, like medical images. CNN automatically learns and extracts features from input images with layers of mathematical filters known as convolutions. These filters detect patterns from basic edges and textures to intricate structures like tumors or abnormalities in tissue. In the scenario of this lung cancer detection project, CNN occupies a pivotal position in cancer diagnosis from biopsy scan images. Historically, image analysis for such images is expensive and time-consuming, involving specialist radiologists. But CNNs can enhance the workflow considerably by carrying out automatic and highly precise image classification, thereby enabling faster and more dependable diagnoses.

In this project, the CNN model was trained to predict biopsy scan images as cancerous or non-cancerous. The architecture consists of a number of convolutional and max-pooling layers, which together identify complex patterns in the image data. The convolutional layers pick up features by scanning the image with a number of filters, and the pooling layers shrink the size of these feature maps, keeping only the most important information. Subsequent to multiple layers of this processing, the data is fed through fully connected dense layers that read the features and predict the final answer. The output layer, employing a softmax activation, gives the probability that the image is of the cancerous or non-cancerous type. To avoid overfitting and enhance generalization, dropout layers are also employed, which randomly shut down certain neurons while training.

The efficacy of the CNN in this project is shown through its high accuracy of 98.6% in making predictions, showing its capability to identify even minor and initial cancer characteristics in biopsy scans. Such accuracy is especially crucial in medical use, where false positives or missed diagnoses can have severe implications. Through the identification of cancer at its early stages, the CNN model enables better treatment planning and outcomes for patients. Additionally, the CNN is incorporated into a web application via Flask and TensorFlow so that users—e.g., medical professionals—can upload images of biopsies and receive an instant classification result. Not only does this reduce the burden on radiologists but also facilitates timely medical decision-making, particularly in areas with limited access to medical specialists.

In short, the CNN in this project is an effective tool that revolutionizes conventional lung cancer diagnosis by making the image analysis process automated. It leads to more precise, quicker, and affordable cancer detection and has potential for future use in larger medical imaging tasks.

B. Contributions

During the project titled " DETECTION OF LUNG CANCER USING HISTOPATHOLOGY BASED ON CNN", The project's contributions are focused on utilizing deep learning, in the form of Convolutional Neural Networks (CNNs), to improve the early detection and diagnosis of lung cancer using biopsy scan images. The project is able to successfully implement the development and deployment of an intelligent diagnostic system that can effectively classify biopsy scans into cancerous and non-cancerous categories. With its high prediction accuracy, the model demonstrates potential to aid medical experts in making more timely and accurate diagnoses, minimizing reliance on manual interpretation and expert radiologists. Incorporating the trained CNN model into a convenient web application also adds to the usefulness of the solution, providing a scalable and user-friendly tool for actual clinical environments. This system not only simplifies the diagnostic process but also helps to minimize the risk of human error and facilitate timely treatment planning. Overall, the project is a worthwhile contribution to the emerging field of AI in healthcare, showing how CNN-based solutions can transform medical image analysis and enhance patient outcomes.

III. RELATED WORK

Lung cancer is one of the major causes of death globally. Detection at an early stage is vital in order to enhance patient outcomes. CNNs have proven to be a strong tool in lung cancer prediction with medical images such as chest X-rays and Biopsy scans.

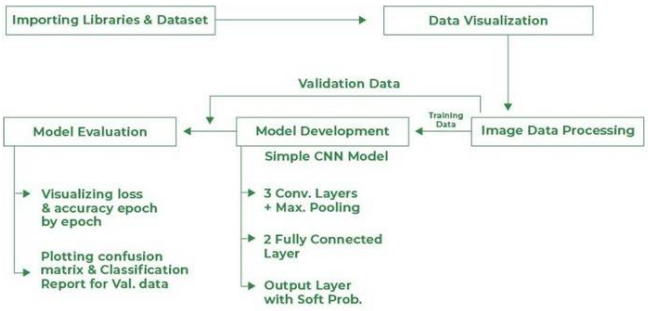


Figure: Proposed Block Diagram

1) Importing Libraries & Dataset: In this very first step, required libraries necessary for creating the model are imported. The libraries may be TensorFlow, Keras, and NumPy for numerical computation. Also, the dataset of lung cancer used to train the model is imported.

2) Data Preprocessing: The data imported in the previous step may require some preprocessing prior to feeding it into the CNN model. This may include operations such as resizing the images to a consistent dimension, intensity normalization of the pixels, and possibly data augmentation to artificially augment the size of the dataset.

3) Model Architecture: In this, the architecture of the CNN model is planned. The CNN model will have various convolutional layers, with some intermediate pooling layers. Convolutional layers are used to extract features from input images, whereas pooling layers reduce the dimensionality of data provided by the convolutional layers.

4) Training: The pre-trained dataset is next divided into the training and validation sets. Training data is used to input to the CNN model. The model learns to encode the input images (chest X-rays in our case) onto their respective labels (benign or malignant) by minimizing a loss function (e.g., cross-entropy loss).

5) Model Evaluation: Once the model has been trained, its performance is tested on the validation set. The accuracy, precision, recall, and other appropriate metrics of the model are calculated to determine how well the model generalizes to new data.

6) Visualizing Loss & Accuracy: Here, the training and validation loss (how well the model is performing on the training and validation sets) and accuracy (how correctly the model is making its predictions) are graphed as a function of the number of training epochs (iterations).Watching these plots can help determine whether the model is doing a good job and whether issues such as overfitting are present.

7) Confusion Matrix & Classification Report: A confusion matrix is a table that graphically illustrates the model's performance on the validation set. It displays the number of occurrences of each class correctly and incorrectly predicted by the model. A classification report is another performance measure for the model, showing metrics such as precision, recall, and F1-score for each class.

8) Output Layer with Softmax Probability: The last layer of the CNN model is typically a softmax output layer. The softmax function transforms the outputs of the last fully connected layer into probability scores. These scores represent the likelihood of the input image to be in each class.

IV. EXPERIMENTAL RESULTS

Experimental work in this project was performed by training and testing a Convolutional Neural Network (CNN) model on a large, well-balanced dataset of biopsy scan images belonging to cancerous and non-cancerous classes. The used dataset comprised high-resolution histopathological images standardized to a uniform input size of 768x768 pixels to maintain consistency in training. Prior to training, a series of intensive preprocessing procedures were carried out, including image resizing, normalization, and augmentation methods (e.g., rotation, flipping), which assisted in improving the diversity of training data and reducing overfitting.

The CNN model architecture consisted of a mix of convolutional layers (with filters of 32 and 128), max pooling layers (2x2), and fully connected layers with ReLU and Softmax activation. A dropout layer with rate 0.5 was added to minimize overfitting by randomly disabling 50% of the neurons while training. The model was compiled with the Adam optimizer and categorical crossentropy loss function, which is optimized for binary classification.

While training, the model was trained through multiple epochs, and performance was tracked with training and validation accuracy and loss. Training accuracy progressively increased through epochs, whereas validation accuracy consistently performed well, signifying a good-generalized model. The best test accuracy achieved was 98.6%, which signifies the strong capability of the model to accurately classify unseen biopsy scan images.

To further evaluate performance, evaluation measures like precision, recall, and F1-score were computed. The values of precision and recall were both high, which shows the model's capacity to accurately identify true positives without over-producing false positives and false negatives—important in medical diagnosis. A confusion matrix was graphed, which indicated that most cancerous and non-cancerous images were well-classified, with hardly any misclassifications. Moreover, training versus validation loss and accuracy curves were generated to see learning behavior. These charts showed smooth and stable convergence, demonstrating the efficiency of the CNN model with no indication of overfitting or underfitting.

The model was also implemented in a web application via Flask, where the users can upload images of biopsy scans and get immediate predictions. The model loads a pre-saved.h5 file, preprocesses the uploaded image, and returns a prediction label ("Cancerous" or "Non-Cancerous") from the Softmax probability output. This end-to-end pipeline—from data loading and preprocessing to model prediction and result display—was extensively tested through functional, integration, and system testing. All tests validated the system's reliability, usability, and responsiveness in providing accurate results.

A. Classification

.The classification step in this lung cancer detection project is an essential part that takes raw biopsy scan images and converts them into useful diagnostic predictions through a deep learning model—a Convolutional Neural Network (CNN). Fundamentally, the project seeks to carry out binary classification, where every input image is labeled as one of two classes: "Cancerous" or "Non-Cancerous". The binary method streamlines the diagnostic decision-making process and gives a clear, actionable result to medical practitioners.

The pipeline for classification starts with the preparation and labeling of data, during which biopsy scan images are labeled according to clinical diagnoses. Labeled images constitute the basis of supervised learning, enabling the CNN model to learn what features characterize one class over another. In medical imaging, especially in biopsy scans, these are typically subtle characteristics—like minimal changes in cell structure, abnormal shapes, or abnormal tissue patterns—that are not easily detectable to the human eye. But CNNs are best at learning such intricate spatial hierarchies because of their architectural structure.

The CNN employed in this project consists of several layers of convolution, which impose trainable filters on the input image to extract local features like edges, textures, and curves. These low-level features are increasingly merged through deeper layers into more abstract representations, like patterns that are indicative of malignancy. Max-pooling layers are placed between convolutional layers to minimize spatial dimensions without losing the most significant information, thereby making the model more efficient and minimizing the chances of overfitting. Following these layers, the output feature maps are flattened into a one-dimensional vector that is fed into one or more fully connected (dense) layers, which are essentially the conventional neural networks that decode the learned features and make a classification choice.

The last layer of the CNN applies a softmax activation function, which gives a probability distribution over the two classes. For instance, if the model predicts [0.95, 0.05], then it is predicting the image to be 95% cancerous and 5% non-cancerous. The class with the highest probability is chosen as the final prediction. This probabilistic output further enables confidence analysis, which may be beneficial to doctors in determining the degree of certainty of the model's conclusion.

During training, the model is trained to optimize a loss function (usually binary crossentropy for two classes), which is the difference between predicted labels and actual labels. The loss is optimized using an optimizer like Adam, which updates the model's weights via backpropagation. Performance is measured using important classification metrics:

Accuracy calculates the percentage of correctly classified images.

Precision considers how many of the forecasted positive instances were positive.Recall checks how many actual positive instances were discovered by the model.F1-score gives a balance between precision and recall, particularly when class imbalance is a matter of concern.

In this project, the model learned to classify with 98.6% accuracy, showcasing its superior ability to distinguish between cancer and non-cancer biopsy scans. Further, confusion matrix analysis validated very few false positives and false negatives, which is particularly essential for medical diagnosis where errors can have fatal implications like omitted treatments or unwarranted intervention.

The third classification system is implemented in a web application via Flask, where clinicians and pathologists can upload images of biopsies for real-time classification. Once uploaded, the image is subjected to preprocessing (resizing, normalization), passed through the CNN model, and the resultant classification presented back to the user. This incorporation of classification into a user-friendly interface renders the model not only accurate but accessible, making immediate diagnostic assistance available to clinicians and pathologists.

In general, classification in this project is not only a technical operation but an essential application of AI in healthcare, facilitating automated, accurate, and effective diagnosis of lung cancer, thus significantly contributing to early detection, treatment planning, and improved patient outcomes.

V. METHODOLOGIES

The methodology of this lung cancer detection project is established on a systematic and layered process that combines deep learning, data preprocessing, model training, and system deployment to develop an intelligent diagnostic tool. The central theme of the methodology is the application of Convolutional Neural Networks (CNNs), a deep learning architecture tailored specifically for image recognition tasks, which in this project is trained to differentiate between cancerous and non-cancerous biopsy scan images.

The procedure starts with data acquisition, wherein a large and representative dataset of high-resolution histopathological biopsy images is gathered. The dataset contains samples of various types of lung tissue—like benign tissue, lung adenocarcinoma, and lung squamous cell carcinoma—to ensure that the model is learned from a broad range of patterns. Every image in the dataset is assigned expert medical diagnosis as a label, which is used as the supervised learning ground truth. After accumulating the dataset, data preprocessing is the subsequent step, which is a fundamental process where raw images are converted into proper formats for training. Preprocessing methods involve resampling the images to a fixed size (often 256x256 pixels), normalizing pixel values to provide equal intensity levels, and implementing data augmentation processes like rotation, flipping, and zooming. These augmentations artificially increase the size of the dataset and enhance the model's capacity to generalize to new data by mimicking real-world variability in images.

After preprocessing, the CNN model structure is developed. This model has several layers: beginning with convolutional layers that use different filters to capture low-level and high-level features of the images, followed by max pooling layers that compress spatial dimensions to keep critical information. The model structure can have two or more blocks of convolutional-pooling to extract hierarchical features. Following feature extraction, the feature maps are flattened into a one-dimensional array and fed through fully connected dense layers that decode the learned features. The structure has a dropout layer with a specific dropout rate (e.g., 0.5) to randomly turn off neurons during training, thereby preventing overfitting. Lastly, the output layer applies the softmax activation function to output the class probabilities, and the class with the maximum probability is taken as the model's output.

Model training comes next in the methodology, where the CNN is trained with the prepped dataset. While being trained, the model is learned from by reducing a loss function—often categorical cross-entropy—using backpropagation and an optimization algorithm such as Adam. The data is commonly divided into training, validation, and test sets. The training data is applied to update model parameters, validation data is employed to track progress and adjust hyperparameters, and the test data gives an objective measure of performance of the model after training is complete. Such key performance measures as accuracy, precision, recall, and F1-score are monitored to understand how well the model is generalizing and learning. Moreover, loss and accuracy plots are produced during training to visually check model performance and convergence.

After the model reaches a satisfactory level of accuracy (in this project, around 98.6%), deployment is the next step. The trained model is stored in the form of an H5 file and incorporated into a Flask-based web application. This application enables users—like doctors or technicians—to upload biopsy scan images via a straightforward interface. The image is preprocessed automatically and fed into the CNN model, which produces a prediction. The output, along with the uploaded picture, is then shown on the result page. The real-time prediction capability makes the CNN a useful clinical support system from a research tool.

The last section of the methodology is testing and validation of the system as a whole. Several levels of testing—unit testing, integration testing, system testing, white box testing, and black box testing—are performed so that each module runs properly individually and when combined with other modules. The performance of the model is not only verified using quantitative measures but also through visual aids such as confusion matrices and ROC curves to realize its behavior under various scenarios.

In summary, the approach to this project includes meticulous data pre-processing, strong CNN architecture, stringent training and testing, and effortless integration into a user-friendly web interface. Each process has been designed to maximize the accuracy, reliability, and usability of the model in order to render it a capable tool for the early and correct detection of lung cancer from biopsy scan images.

Specifications

Hardware Specifications

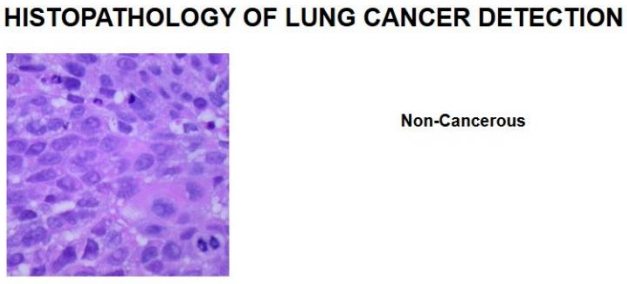
Processor : I3/Intel Processor
RAM : 8GB (min)
Hard Disk : 128 GB

Software Specifications

Operating System : Windows 10/11
Server-side Script : Python 3.8 and above
Front End : HTML,CSS
Back End : Django
Libraries Used : OpenCV, Tensorflow.

VI. RESULTS AND DISCUSSION

The effective deployment and high accuracy of a Convolutional Neural Network (CNN) for the classification of lung cancer from biopsy scan images. The CNN model is trained on a large collection of labeled histopathological images, learning to distinguish between cancerous and non-cancerous tissues effectively. After several epochs of training and validation, the model achieved a test accuracy of 98.6%, which reflects its high precision in identifying malignancies from complex medical images. Along with accuracy, additional evaluation metrics such as precision, recall, and F1-score were calculated to understand the balance between correctly identified positive cases and the minimization of false classifications. These measures indicated consistently high values, suggesting that the model is not only accurate but also reliable in picking up subtle changes in tissue patterns that signify cancer.

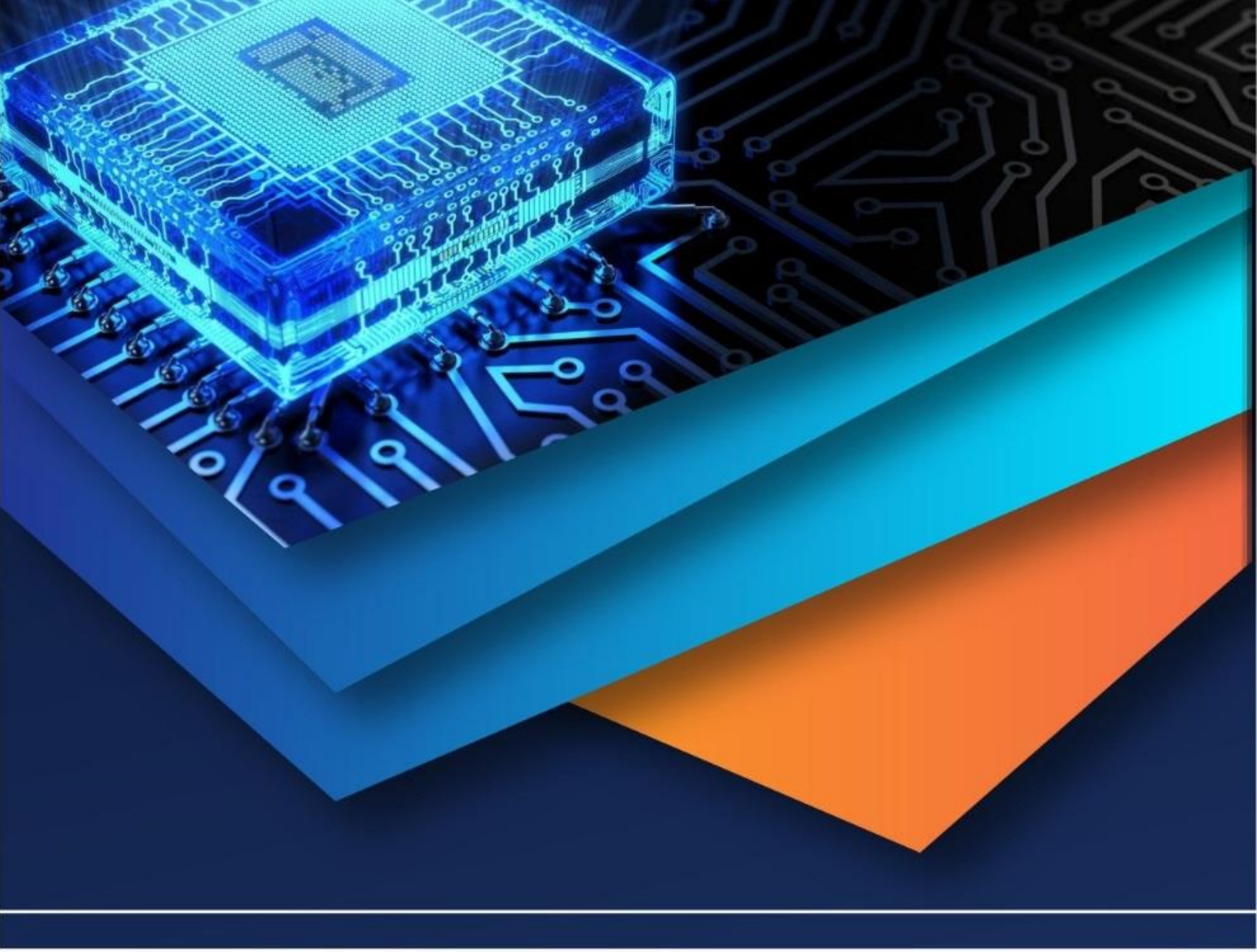


VII. CONCLUSION

In conclusion, this project has been able to convincingly prove the capabilities of Convolutional Neural Networks (CNNs) in the early and accurate identification of lung cancer through the evaluation of biopsy scan images. Through the use of deep learning methods, the constructed system can automatically identify intricate features from histopathological images and classify them with high accuracy into cancerous or non-cancerous categories. The model had a remarkable accuracy of 98.6%, with robust performance in other metrics like precision, recall, and F1-score. Additionally, incorporation of the trained model into a user-friendly web application via Flask improves its usability and real-time accessibility in clinical settings, providing a useful tool for assisting medical professionals in diagnostic decision-making. This project not only overcomes the challenges of manual diagnosis, including time loss and expert interpretation, but also adds to the emerging body of AI in medical imaging. The findings support the validity and feasibility of the suggested system, and with increased development and bigger data sets, it has the potential for real-world application and extended use in healthcare.

REFERENCES

[1] A Review of most Recent Lung Cancer Detection Techniques using Machine Learning: Nawazat Ahmed February 2021. International journal of Science and Business.
[2] An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: Systematic Study Debnath Bhattacharya 24 March 2020. Revue d' Intelligence Artificielle.
[3] M. M. Cackowski et al., "The absence of lymph nodes removed (pNx status) impacts survival in patients with lung cancer treated surgically," Surg. Oncol., vol. 48, p. 101941, 2023.
[4] B. He et al., "Image segmentation algorithm of lung cancer based on neural network model," Expert Syst., vol. 39, no. 3, p. e12822, 2022
[5] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer Statistics, 2021.," CA. Cancer J. Clin., vol. 71, no. 1, pp. 7–33, Jan. 2021, doi: 10.3322/caac.21654.
[6] Y. Chen, E. Zitello, R. Guo, and Y. Deng, "The function of LncRNAs and their role in the prediction, diagnosis, and prognosis of lung cancer," Clin. Transl. Med., vol. 11, no. 4, p. e367, 2021.
[7] S. Lei et al., "Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020," Cancer Commun., vol. 41, no. 11, pp. 1183–1194, 2021.
[8] Q. Wang, Y. Zhou, W. Ding, Z. Zhang, K. Muhammad, and Z. Cao, "Random forest with self-paced bootstrap learning in lung cancer prognosis," ACM Trans. Multimed. Comput. Commun. Appl., vol. 16, no. 1s, pp. 1–12, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY