# CANCER ANALYZER

[1]Aarav Sharma, [2]Ena Tandon, [3]Kushaagra Mehta, [4]Lakksh Bhardwaj, [5]Dr. D Ganesh Gopal

Department of Computer Science and Engineering
Faculty of Engineering and Technology
SRM Institute of Science and Technology, NCR Campus, Delhi-NCR Campus,
Delhi-Meerut Road, Modinagar, Ghaziabad, UP, India.

*Abstract* - **Cancer detection in critical organs such as the lungs, colon and brain are essential for early diagnosis and effective treatment. We propose a cancer detection system in this paper using Deep Learning. In the proposed system, histopathological images for lung cancer, colon cancer and CT scans for brain cancer are processed using techniques like segmentation and enhancement to highlight key features. A convolutional neural network (CNN) is trained to classify cancerous and non-cancerous tissues, enabling accurate detection. Leveraging transfer learning and large datasets, the system achieves high precision, sensitivity, and specificity. The system also shows promising results in terms of real-time performance, with an accuracy of 92.15%, 89.43% and 96.39% for lung cancer, brain tumour and colon cancer respectively. The proposed system aids healthcare professionals in early cancer detection and timely interventions.**

*Keywords:* **Cancer detection, Convolutional Neural Network, Deep learning, Healthcare, Real-time performance.**

## I. INTRODUCTION

Cancer remains one of the leading causes of death globally, with lung and brain cancers being particularly lethal due to their often late-stage diagnosis and aggressive nature. Early detection significantly increases the chances of successful treatment and survival. However, traditional diagnostic methods, such as manual analysis of medical imaging (CT or histopathological scans), are time-consuming, prone to human error, and require specialized expertise. These challenges highlight the need for automated, efficient, and accurate diagnostic systems.

Recent advancements in computer vision and deep learning have revolutionized medical imaging analysis, enabling the development of models capable of detecting cancerous lesions with high accuracy. Deep learning techniques, particularly convolutional neural networks (CNNs), have shown tremendous success in image classification tasks, making them well-suited for detecting and analysing abnormalities in medical scans. This research aims to harness these technologies to create an automated system for the detection of lung, colon and brain cancer from medical images.

In this paper, we propose a cancer detection system which focuses on automating the diagnostic process by processing CT and histopathological scans through a series of advanced image processing techniques. Preprocessing steps like image segmentation, noise reduction, and contrast enhancement are applied to isolate regions of interest and remove irrelevant information, ensuring that the deep learning model focuses on the most critical features. These preprocessing techniques not only improve image quality but also enhance the model's ability to identify cancerous tissues with precision.

At the core of the proposed system, there are convolutional neural networks (CNNs), a type of deep learning model known for their exceptional performance in image recognition tasks. CNNs are employed to learn patterns from medical images and differentiate between cancerous and non-cancerous regions. The model architecture is designed to extract both low-level features, such as edges and textures, and high-level features, such as tumour shapes and structures. By training the

CNN on large, annotated datasets of lung, colon and brain scans, the system can achieve high accuracy in classifying tumour types, stages, and locations.
The proposed model works on real-time data and has high accuracy, hence is reliable. The F1 score and recall of lung cancer model, brain tumour model and colon cancer model turned out to be (0.924,0.921), (0.906, 0.905) and (0.964, 0.964) respectively.
One potential limitation of the proposed system is the reliance on large and high-quality datasets of medical images, including histopathological scans for lung cancer, colon cancer and CT scans for brain cancer. These datasets must be well-annotated with detailed labels to train the deep learning models effectively. The quality, diversity, and volume of these images are crucial for training a robust model capable of generalizing well to real-world clinical scenarios.
Finally, this proposed system is to provide a reliable, efficient tool for early cancer detection, assisting healthcare professionals in making timely and accurate diagnoses. By automating the analysis of CT and histopathological scans, the system reduces the burden on radiologists and lowers the risk of misdiagnosis due to human error. Ultimately, the integration of computer vision and deep learning in medical imaging has the potential to significantly improve patient outcomes by enabling early detection and providing a scalable solution to cancer diagnosis.

## III. System Implementation

### *A.* EXISTING SYSTEM

Existing systems for cancer detection have evolved significantly with advancements in technology, particularly in the areas of medical imaging, artificial intelligence (AI), and machine learning. Several systems, both research-based and commercially available, have been developed to assist in the early detection and diagnosis of cancer.

One example of an existing system is the use of CAD (Computer-Aided Detection) Systems. Computer-Aided Detection (CAD) systems are a significant technological development in the field of medical imaging. CAD systems use advanced algorithms to analyse medical images such as X-rays, mammograms, CT scans, histopathological images and MRIs to detect abnormalities that could indicate the presence of cancer or other diseases. The primary goal of CAD is to assist radiologists and other medical professionals by acting as a "second set of eyes," increasing the likelihood of early and accurate diagnosis.

Another example of an existing system, IBM Watson for Oncology is an AI-driven system designed to assist oncologists in diagnosing and developing treatment plans for cancer patients. Watson uses natural language processing to analyse vast amounts of medical literature and clinical trial data, offering personalized treatment options. While it’s not primarily focused on detection, Watson helps in diagnosing cancer based on patient history and pathology reports, as well as recommending optimal treatments.

The above existing system have some limitations, particularly traditional CAD, still suffer from high false-positive rates, which can lead to unnecessary tests or procedures. Moreover, deep learning models trained on specific datasets may not generalize well to new or diverse patient populations, which limits their applicability in different healthcare settings.

In contrast, the proposed system using Convolution Neural Network along with deep learning has the advantage of being able to provide high accuracy and accountability. Due to the use of large, real-time dataset the limitation of generalization is also resolved with the proposed system.
Additionally, the system's high accuracy and real-time performance make it a valuable addition to existing systems aimed at aiding the healthcare system.

## *B*. PROPOSED SYSTEM

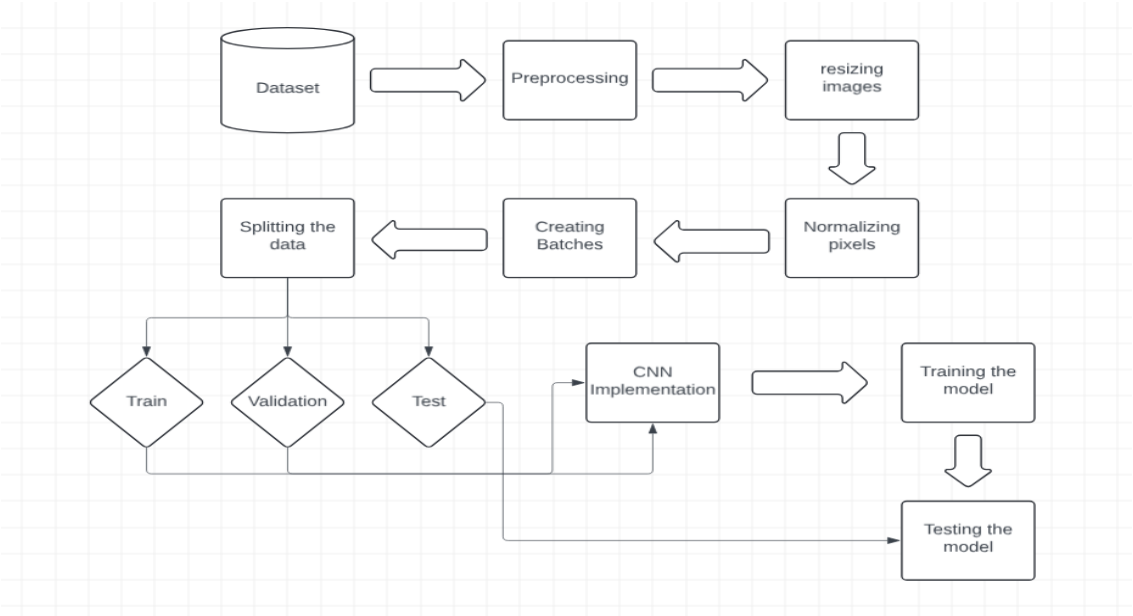
The proposed system is a Cancer Analyser using convolutional neural network (CNN) and deep learning. The system is designed to analyse images and detect the possibility of cancer (brain, colon or lung) using CT scans for brain tumour and histopathological scans for lung, colon cancer. These are one of the leading causes of death due to their late-stage diagnosis. The system uses a CNN model trained on real-time dataset. The dataset contains over 7249 images for brain tumour, 14167 images for lung cancer and 10000 images for colon cancer scans, making it an ideal dataset for training cancer detection models. The pre-trained model is then fine-tuned on a custom dataset of cancerous images. The custom dataset consists of images of cancerous and non-cancerous histopathological and CT scans from various sources, including publicly available data, clinic and hospital data.

The proposed model achieves high accuracy in detecting cancerous tissues. Accuracy of lung cancer model, brain tumour model and colon cancer model are 92.15%, 89.43% and 96.39% respectively. The system's performance is calculated using the F1 score and recall of each model during training and testing.

The F1 score is the harmonic mean of a model's precision and recall. A harmonic mean is a type of average used in mathematics and statistics. It is calculated by summing the reciprocal of each value in a data set and then dividing the number of values in the dataset by said sum. Recall, or true positive rate (TPR), is a machine learning metric that measures how well a model can correctly identify positive instances in a dataset. It's calculated by dividing the number of correctly predicted positive instances by the total number of actual positive instances

The proposed system also shows promising results with F1 score and recall of lung cancer, brain tumour and colon cancer being (0.924,0.921), (0.906, 0.905) and (0.964, 0.964) respectively.

The proposed system automatically learns relevant features from raw data during training, without the need for manual feature extraction. This allows CNNs to learn much more complex patterns and make finer distinctions between cancerous and non-cancerous tissues, leading to higher accuracy, reduced false positives, and fewer missed diagnoses (false negatives). The system can be integrated with existing systems such as CAD, leveraging the strengths of both. The combination can capitalize on the expertise built into existing systems and the advanced pattern-recognition capabilities of the proposed system, creating a system that is more accurate, reliable, and efficient.



## IV. MODULES

### Module 1: Data Preprocessing

The data collection and pre-processing module is an essential component of the proposed Cancer analyser using Convolution Neural Network and Deep Learning.
The module's main purpose is to collect histopathological and CT scan images and prepare them for use in training the CNN model.

Data collection involves obtaining images from various sources. The images for preprocessing in a cancer detection system are typically sourced from public medical datasets like LIDC-IDRI for lung cancer and TCGA for brain cancer, which provide large collections of annotated CT, MRI, histopathological and X-ray scans. Additionally, images can be collected from hospital or clinical databases, where real patient data, such as CT and histopathological scans, are archived. These datasets offer a rich variety of cancer cases, allowing for robust model training and testing while maintaining privacy and ethical standards.

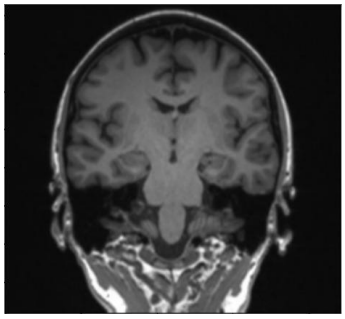
The dataset must next be divided into training and validation sets once the photos have been annotated. A typical split ratio is 80:20, where 80% of the data is used for training and 20% for validation.

Pre-processing the collected data involves various steps though in the proposed model, since Conv2D, is implemented which makes pre-processing easier and efficient. Moreover, medical image datasets consist of lung and colon tissue scans and brain tumours hence there is highly sensitive information and tumours exist in very small sizes. So, there are high chances if we performed too much preprocessing then it could hamper the accuracy of the model.

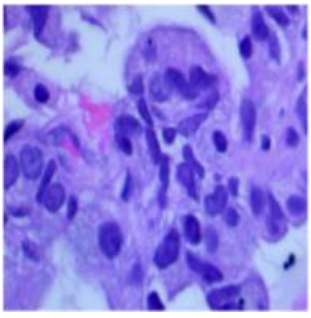
Using NumPy and Pandas, the images and metadata are first loaded and transformed. NumPy is used for handling the raw pixel data of images as arrays, allowing for efficient numerical operations such as resizing, normalization, and conversion of image formats. Pandas can be used to organize and manage related metadata (e.g., patient information, labels for cancer types) for structured analysis.

The ImageDataGenerator module from tensorflow.keras.preprocessing.image is then used for augmenting the dataset, which enhances the model's ability to generalize by applying transformations like rotation, scaling, flipping, and zooming to the images. This also helps deal with the typically limited amount of medical imaging data. The module performs on-the-fly augmentation while loading the images, reducing memory load and improving processing efficiency. These preprocessing steps, including normalization and augmentation, ensure that the data is more uniform, diverse, and ready for input into CNNs for accurate cancer detection.

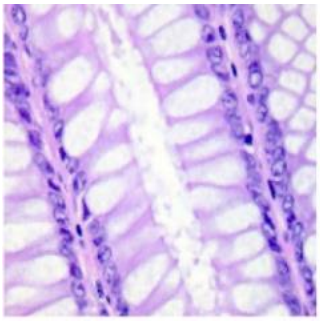
**BRAIN TUMOUR**



**LUNG CANNCER**



**COLON CANCER**



**Module 2: Model Training**

Model training is a critical step in the development of the proposed Cancer Analyser using CNN and Deep Learning. The main objective of model training is to teach the algorithm to detect cancerous areas in real-time accurately.

In the proposed system, training the model effectively is crucial to achieving high accuracy and generalization. The first step in model training is to initialize the weights using the training dataset. This step is important as it allows the model to leverage the knowledge learned from the model to detect cancerous tissue images.

The process then typically begins by defining the model architecture, where GlobalAveragePooling2D is used to reduce the spatial dimensions of feature maps after convolution layers. Instead of using dense layers with many parameters, GlobalAveragePooling2D averages the spatial data, reducing overfitting and the computational burden. This step is critical in medical image analysis, as it extracts meaningful global features from CT or histopathological scans without overcomplicating the model, allowing it to focus on high-level patterns

In the proposed model, overfitting is prevented by integrating dropout layers into the model. Dropout randomly deactivates a portion of neurons during each training step, ensuring that the model does not rely too heavily on specific neurons. By applying Dropout after pooling or dense layers, the model becomes more generalizable and less likely to overfit the training data, thus improving its performance on new, unseen images.

Further, while training the proposed model, the Adam optimizer is used for updating the model's weights. Adam adjusts the learning rates for each parameter based on the gradient's momentum and variance, allowing the model to converge quickly and efficiently. Adam's adaptive nature ensures that the training process is smooth, avoiding large oscillations or overshooting, which is crucial when distinguishing subtle features in medical images like tumours. As the proposed model trains, Early Stopping is employed to monitor the validation loss. If the model's performance plateaus or begins to degrade, Early Stopping halts the training early, saving time and preventing overfitting.

To further refine the learning process, the proposed model uses ReduceLROnPlateau to dynamically reduce the learning rate when the validation performance stops improving. This allows the model to make more fine-tuned updates to its weights, helping it settle into an optimal solution without large weight fluctuations.

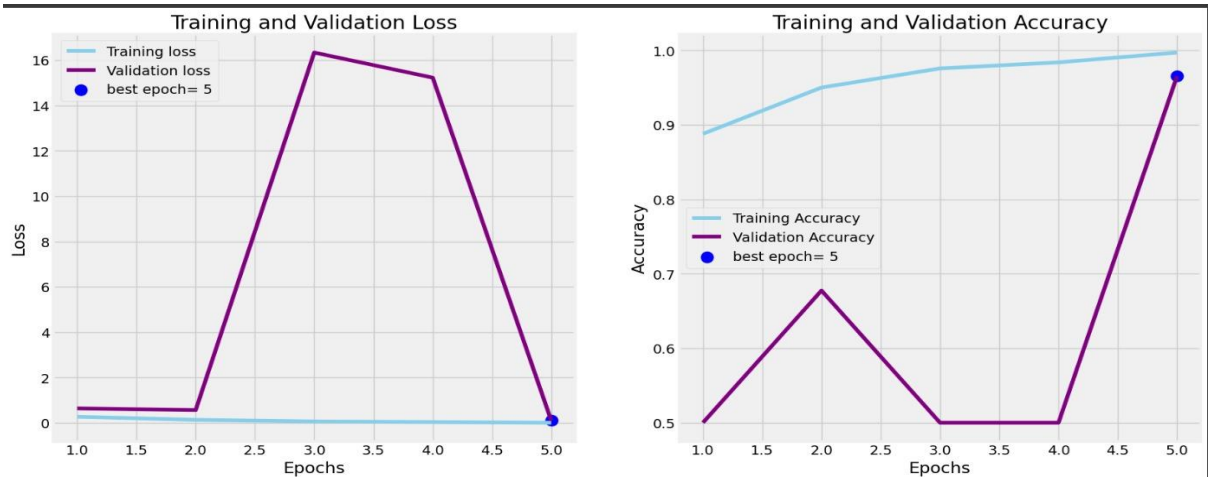
**Module 3: Prediction of output**

The output of the proposed model is typically a probability score for each input image, indicating the likelihood of the presence of cancer in lung, colon or brain. These probability scores are then converted into binary predictions using a threshold. The model's predictions can then be compared to the true labels from the validation or test dataset to assess its performance.

To measure the effectiveness of the proposed model, recall is particularly important in cancer detection, as it measures the proposed model's ability to correctly identify cancerous cases. High recall ensures that the model minimizes false negatives, which is critical in medical diagnosis where early detection can significantly improve outcomes.

The F1 Score provides a balanced evaluation of the model by combining both precision and recall. The F1 Score is the harmonic mean of precision and recall, giving a more comprehensive view of the proposed model's accuracy, especially in cases where the data is imbalanced. A high F1 Score ensures that the model not only detects cancer cases accurately but also avoids incorrectly labelling healthy scans as cancerous, thus balancing the trade-off between false positives and false negatives.

These metrics provide a clearer picture of the proposed model's diagnostic capabilities, allowing for fine-tuning if necessary, ensuring the model is reliable and ready for aiding the healthcare system.

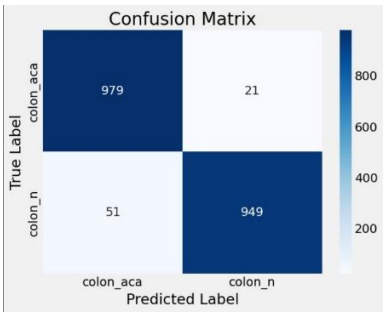



## V. RESULTS

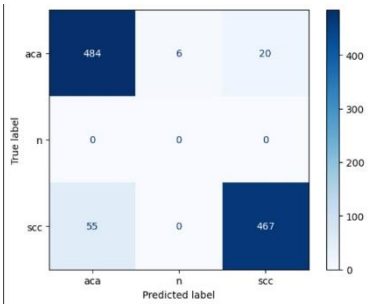
The proposed cancer analyser model using convolution neural network and deep neural network has shown promising results in highlighting cancerous and non-cancerous areas. The model was evaluated on a custom dataset of histopathological and CT scan images obtained through the data collection and pre-processing module.

Finally, the proposed cancer analyser has an accuracy of 92.15%, 89.43% and 96.39% for lung cancer, brain tumour and colon cancer respectively with F1 score and recall being (0.924,0.921), (0.906, 0.905) and (0.964, 0.964) of lung cancer, brain tumour and colon cancer respectively.

**COLON CANCER**




**LUNG CANCER**




## VI. CONCLUSION

In conclusion, we have presented a cancer analyser using Convolution Neural Network and Deep Learning, which is a real-time and high-accuracy cancer detection model. The proposed model demonstrates a powerful and efficient approach to diagnosing cancer through medical imaging. It effectively balances learning from complex image features while preventing overfitting, ensuring robustness and generalization to new data, leading to high accuracy in distinguishing between cancerous and non-cancerous scans. The use of evaluation metrics confirms the proposed model's clinical viability. High recall ensures that the model captures the majority of cancerous cases, minimizing the risk of false negatives, while the F1 Score balances the need for both precision and sensitivity, reducing the likelihood of false positives. Together, these metrics highlight the model's strong potential for accurate and reliable cancer detection, making it a valuable tool in the early diagnosis and treatment of cancer. With further validation and refinement, this proposed model can serve as a critical asset in medical diagnostics, supporting healthcare professionals in improving patient outcomes, and we believe that this work will inspire further research and development in the field of cancer detection systems.

**REFERENCES:**


- A deep convolutional neural network for the detection of polyps in colonoscopy images by Tariq Rahim, Syed Ali Hassan.
- Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images by Ramin Ranjbarzadeh, Abbas Begherian Kasgari.
- Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method by Akitoshi Shimazaki, Daiju Ueda, Antoine Choppin, Akira Yamamoto, Takashi Honjo, Yuki Shimahara & Yukio Miki.