



OPEN

Optimizing non small cell lung cancer detection with convolutional neural networks and differential augmentation

Vahiduddin Shariff¹✉, Chiranjeevi Paritala² & Krishna Mohan Ankala¹

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, with early detection being critical to improving patient outcomes. Recent advancements in deep learning have shown promise in enhancing diagnostic accuracy, particularly through the use of Convolutional Neural Networks (CNNs). This study proposes the integration of Differential Augmentation (DA) with CNNs to address the critical challenge of memory overfitting, a limitation that hampers the generalization of models to unseen data. By introducing targeted augmentation strategies, such as adjustments in hue, brightness, saturation, and contrast, the CNN + DA model diversifies training data and enhances its robustness. The research utilized multiple datasets, including the IQ-OTH/NCCD dataset, to evaluate the proposed model against existing state-of-the-art methods. Hyperparameter tuning was performed using Random Search to optimize parameters, further improving performance. The results revealed that the CNN + DA model achieved an accuracy of 98.78%, outperforming advanced models like DenseNet, ResNet, and EfficientNetB0, as well as hybrid approaches including ensemble models. Additionally, statistical analyses, including Tukey's HSD post-hoc tests, confirmed the significance of the model's superior performance. These findings suggest that the CNN + DA model effectively addresses the limitations of prior works by reducing overfitting and ensuring reliable generalization across diverse datasets. The study concludes that the novel CNN + DA architecture provides a robust, accurate, and computationally efficient framework for lung cancer detection, positioning it as a valuable tool for clinical applications and paving the way for future research in medical image diagnostics.

Keywords Lung Cancer prediction, Convolutional neural networks, Differential augmentation (DA), Memorization overfitting, Medical imaging, Diagnostic robustness, EfficientNetB0

Innovative lung cancer prognosis relies on cutting-edge machine learning algorithms applied to a wealth of patient data. By incorporating various elements like imaging, genetics, and medical records, this model demonstrates remarkable precision in anticipating potential occurrences. Optimizing early intervention and customizing treatments, it holds the potential to elevate patient outcomes in the challenging realm of lung cancer. Early detection is vital for successful treatment¹ and goal is to authenticate our Lung Cancer Prediction CNN, initially trained on US screening data, through an autonomous European multicenter dataset². This research explores early-stage lung cancer diagnosis using CNNs, aiming to enhance diagnostic accuracy and save lives by improving early detection through reliable machine learning models³. While CNN models have demonstrated high accuracy on controlled datasets, their applicability to real-world scenarios with diverse patient populations remains a critical challenge. Addressing this requires validation across multiple datasets and strategies to ensure robustness against unseen variations in imaging data. Introducing a method utilizing Convolutional Neural Networks (CNN) to categorize tumors in lung disease screening through tomography filters, harnessing spatial invariance for effective feature extraction⁴. Utilizing end-to-end learning CNN, this method automatically extracts self-learned features, comparing favorably with conventional approaches and traditional computer-aided diagnosis systems⁵.

¹Department of Computer Science and Engineering, UCEK, Jawaharlal Nehru Technological University Kakinada, Kakinada, Andhra Pradesh-533003, India. ²Department of Computer Science and Engineering, Amrita Sai Institute of Science and Technology, Bathinapadu, Paritala, Andhra Pradesh-521180, India. ✉email: shariff.v@gmail.com

Developing a cutting-edge 1D CNN model for NSCLC staging and grading, leveraging insights from the latest TCIA NSCLC Radio genomics Collection. Hybrid MSER-SURF model integrates tumor features with clinical data for comprehensive analysis⁶. Despite the highly imbalanced and variegated nodule detection challenge, our method excels, achieving effective lung nodule detection and cancer prediction. Using a candidate proposal approach and a 3D CNN, achieve near-human performance⁷ and to introduce a swift CNN-RNN model for NSCLC AJCC staging, surpassing traditional ANNs through deeper layers. Comparison involves standard ML algorithms and related studies⁸. Focusing on individuals with lung cancer, our focus lies in crafting advanced deep learning models for predicting survival outcomes, addressing both classification and regression challenges. Feature importance analysis elucidates relevant factors impacting survival periods⁹. Introducing LungNet, an innovative hybrid model blending a 22-layer Convolutional Neural Network (CNN) with wearable sensor-based MiOT data and CT scan for heightened diagnostic precision¹⁰.

Leveraging CNN's disease detection, this research explores the fusion of blockchain for secure data in enhancing food safety and lung cancer prediction¹¹. This preliminary study underscores CNNs' feasibility and promising efficacy in evaluating lung cancer's T-parameter, offering rapid T1-T2 or T3-T4 probabilities. Ongoing research aims at robust TNM algorithm development¹².

Introducing an innovative approach, employs a 2D CNN with Taguchi optimization for automated lung cancer recognition from CT images, as outlined in reference¹³. Through 36 experiments and 8 control factors, it meticulously refines CNN parameters, significantly improving classification accuracy. Additionally, accuracy was enhanced by integrating convolutional and bidirectional recurrent neural networks into a novel deep learning model, leveraging the NSCLC Radiogenomics dataset with 211 subjects as detailed in reference¹⁴. Our intelligent medical system, detailed in reference¹⁵, incorporates CNN-assisted diagnosis and decision-making, utilizing sensors to analyze NSCLC patient's records and ensuring accuracy through transfer learning and dynamic sampling techniques.

The research structure unfolds with Section 1 delivering into the introduction. Section "Literature review" navigates the scholarly landscape in the literature review. Section "Proposed methodology" intricately unfolds the intricacies of the proposed methodology, offering a detailed roadmap. Section "Result and discussion" unveils the results, igniting insightful discussions around their implications. The "conclusion" section encapsulates profound insights, culminating in a comprehensive closure that synthesizes the study's findings and underscores their significance in the broader context of the research domain.

Literature review

In 2024, Islam et al.¹⁶ provides a comprehensive evaluation of the latest advancements in using deep learning to enhance medical imaging data. The course covers a wide range of applications, popular datasets, evaluation criteria, and both traditional and innovative augmentation techniques. To assess the current condition of the subject and identify its primary challenges, the study analyses prior research findings. Proposed are some avenues for future research that could address these concerns. This review is an invaluable resource for academics and experts in the field of medical image analysis.

In 2024, Saha et al.¹⁷ presents VER-Net, a transfer learning model for CT scan lung cancer detection. Combining VGG19, EfficientNetB0, and ResNet101 transfer learning techniques, this model four lung cancer classifications are produced via picture preparation, data augmentation, and hyperparameter tuning of the model. The results imply that VER-Net can be used to other diseases found on CT scans and is quite accurate in spotting lung cancer. This work highlights the efficient application of transfer learning in the domain of medical image processing, especially in increasing diagnostic accuracy.

In 2024, Riku Klén et al.¹⁸ examine the influence of several image augmentation techniques on the precision of a Convolutional Neural Network (CNN) that has been trained for binary classification. The CNN is trained using 11 medical datasets, mostly comprising X-rays, ultrasound (US) images, PET scans, and MRI scans. The databases centre on cancer and lung diseases. Specifically analysing seven extensively used augmentation techniques, this article compares CNN predictions with and without augmentation. The results imply that for both the US and PET datasets, the application of augmentation methods does not produce statistically significant variations. Gaussian blur was determined to be the most successful augmentation technique for X-rays and MRI images, therefore enhancing the model performance. These results show that the kind of imaging technology applied affects the effectiveness of augmentation strategies.

In 2024, S. Kukreja and M. Sabharwa et al.¹⁹ improving the survival rates of lung cancer, a major and maybe deadly disease, depends on early identification. In order to classify three different histological images: benign, adenocarcinoma, and squamous cell carcinoma, in this paper proposes using a Convolutional Neural Network (CNN). The aim of the project is to evaluate the accuracy of the Convolutional Neural Network (CNN) model against other techniques thereby enhancing the diagnosis of lung cancer. Convolutional neural networks (CNNs) have not been used for the categorisation of these specific images so our method is novel. By means of accurate and timely classification of cancers, this approach may improve the efficacy of therapy and possibly avert death of course. By means of convolutional neural networks (CNNs), the diagnostic process can be simplified, costs can be lowered, and general accuracy in lung cancer detection can be improved.

In 2024, Zhang et al.²⁰ demonstrates the application of Convolutional Neural Networks (CNNs) to Computed Tomography (CT) scans has led to notable progress in automating the diagnosis of lung cancer, resulting in enhanced accuracy in the detection and analysis of this disease. Notwithstanding these progressions, there are still obstacles to overcome, such as the restricted comprehensibility, variability of data, and difficulties in generalising. This paper proposes a new method which integrates the CNN with the DenseNet by using data fusion and mobile edge computing to enhance the identification and classification of lung cancer. Data fusion improves the reliability of models by integrating data from multiple sources where mobile edge computing hastens the model's processing time to instant. Improve DenseNet's classification accuracy further, the approach

introduced an enhanced predictive sparse decomposition (PSD) method to acquire sparse feature from medical images, which also promoted the improvement of the model.

In 2023, Lulu Gai et al.²¹ demonstrated Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) test lung cancer identification using CT data. In this paper used a collection of 212 medical photos and to get beyond dataset restrictions, both models exploited self-supervised learning, transfer learning, and sharpness-aware minimiser. Their training consisted in no lung nodule markings. Convolutional Neural Networks (CNNs) outperform Vision Transformers (ViTs) in self-supervised learning, with an AUC of 98.1% and a recall rate of 93.4%. While both CNNs and Vision Transformers (ViTs) showed potential, CNNs outperformed ViTs, especially with smaller datasets. The study shows that CNNs can identify early-stage lung cancer, supporting their clinical utility.

In 2023, Syeda Reeha et al.²² proposes an ensemble model combining several deep learning architectures, including BEiT, DenseNet, and Sequential CNN, so addressing the immediate demand for efficient lung cancer detection and classification. To improve prediction accuracy the combined strategy uses AND, OR, Weighted Box Fusion, and Boosting. The research shows using the Chest CT-Scan Images Dataset that the ensemble model much beats single-model techniques. The study highlights how effectively combining various classifiers overcomes particular limitations and improves general performance.

In view of the available data and resources, this study provides a reasonable response by demonstrating the potential of ensemble methods as a potent tool for enhancing the precision and efficiency of lung cancer diagnosis.

In 2023, R. Raza et al.²³ based on the EfficientNet architecture with extra top layers for better performance; this work presents Lung-EffNet, a new transfer learning-based model for lung cancer classification. Lung-EffNet sorts lung scans into benign, malignant, or normal categories after five EfficientNet variations (B0–B4) are evaluated on the IQ-OTH/NCCD dataset. Lung-EffNet, especially with EfficientNetB1, shown better accuracy, efficiency, and faster training times than other pre-trained CNN architectures. Large-scale clinical implementation and automated lung cancer diagnosis from CT images fit this paradigm really nicely.

In 2023, Nandita Gautam et al.²⁴ Essential for early detection of lung cancer, this work provides a new ensemble deep learning model that correctly labels the degree of lung nodules from CT scan images. Three advanced convolutional neural network (CNN) models are ResNet-152, DenseNet-169, and EfficientNet-B7. The weight optimisation approach used in this system combines ROC-AUC and F1-scores to improve accuracy. It surpassed recent methodologies and successfully reduced the frequency of false negatives. This approach demonstrates promise in significantly improving lung cancer diagnosis and patient outcomes.

In 2022, Dritsas and Trigka et al.²⁵ used machine learning, the study crafted effective models for early detection of high-risk individuals prone to lung cancer, enabling timely interventions to mitigate long-term complications. Emphasizing the efficacy of Rotation Forest, the article highlighted its robust performance, subjecting it to thorough evaluation through established metrics such as precision, recall, F-Measure, accuracy, and area under the curve.

In 2021, Tsou et al.²⁶ introduced a machine learning model that precisely detected lung cancer based on the analysis of participants' exhaled breath, presenting a non-invasive and radiation-free diagnostic system. The research successfully showcased a novel diagnostic approach by integrating deep learning algorithms and VOC analysis, minimizing environmental interference. Ongoing efforts focused on developing standardized, automated breath sampling protocols, aiming to simplify the collection process and ensure sample quality.

While prior studies focus on augmentation and model optimization, this research uniquely addresses memory overfitting using Differential Augmentation, demonstrating superior performance in lung cancer detection.

Problem statement

The research problem focuses on enhancing the accuracy and efficiency of lung cancer detection using deep learning techniques. Although improved, current models struggle with generalisation, data variability, and interpretability across imaging modalities. Optimising model performance, especially early detection, while addressing dataset variety and augmentation limits is difficult.

Generalizability challenges

Although the CNN + DA model performs well on the controlled dataset, its generalizability to real-world scenarios remains to be validated. Challenges include unseen variations in imaging protocols, patient demographics, and artifacts in clinical datasets. Future studies will explore strategies such as transfer learning and fine-tuning on multi-center datasets to mitigate these challenges and ensure applicability in diverse clinical settings. To increase medical image analysis models' resilience, scalability, and clinical utility, data fusion, mobile edge computing, and ensemble methods must be explored. Table 1 summarizes the objectives and limitations of previous research, highlighting both their goals and shortcomings.

Research gap

While deep learning has revolutionized numerous domains, including medical imaging for lung cancer detection, a persistent research gap remains concerning memory overfitting. Despite various efforts to address general overfitting, the nuanced challenge of models memorizing training data rather than learning to generalize has not been comprehensively addressed. Many existing solutions focus on broader overfitting issues, often overlooking the subtle yet critical aspect of memory overfitting. This gap underscores the need for dedicated research into strategies and methodologies to combat memory overfitting, ensuring that models remain robust and reliable, especially in critical applications like medical diagnostics.

Author	Method	Aim	Drawback
Islam et al. ¹⁶	Discussion on the application of various deep learning algorithms in generating and applying data augmentations. This may involve methods like Generative Adversarial Networks (GANs) or other neural network architectures that create realistic variations in medical images.	Identification of field-based difficulties includes constraints in present techniques or problems with data quality and suggested future research paths to handle these difficulties.	Advanced data augmentation techniques can raise the complexity of the model and the training process, thereby perhaps making it more difficult to validate the predictions of the model. More broad and sophisticated augmentations can cause the training process to lengthen and become more resource-intensive, therefore affecting the general efficiency of model development.
Saha et al. ¹⁷	The paper introduces a new transfer learning model called VER-Net, which combines three distinct transfer learning models to improve the diagnosis of lung cancer from CT scan pictures.	Using the benefits of transfer learning and combining several models helps one to increase the accuracy of lung cancer classification. The work also seeks to show that VER-Net can be useful for other medical imaging chores in addition to surpassing current models in performance.	Stacking several transfer learning models under the VER-Net model can complicate the system. This complexity might make model interpretability difficult and raise computational needs.
Riku Klén et al. ¹⁸	A convolutional neural network (CNN) was trained to classify medical images into two categories: impacted, non-affected instances and Augmentation Methods.	Evaluate the influence of various image augmentation methods on the accuracy of a Convolutional Neural Network (CNN) trained to classify medical images into two categories. Identify the most efficacious augmentation methods for various medical imaging modalities (X-rays, ultrasound, positron emission tomography, magnetic resonance imaging) to enhance classification accuracy.	The efficacy of augmentation approaches can vary considerably depending on the nature of the medical imaging data. The study's findings may not have universal applicability across all modalities of medical imaging or across various clinical states.
S. Kukreja and M. Sabharwa et al. ¹⁹	The work uses CNN architecture to categorise lung cancer histological pictures into benign, adenocarcinoma, and squamous cell carcinoma.	To divide histological pictures of lung cancer into three types: benign, adenocarcinoma, and squamous cell carcinoma with accuracy. CNN Model.	This work focuses on CNN and does not compare its performance with a greater spectrum of machine learning or deep learning approaches outside of those already looked at.
Zhang et al. ²⁰	Convolutional Neural Network (CNN) with DenseNet to improve the ability to extract features and learn for the purpose of classifying lung cancer.	Developing and assessing an advanced system stands as the main goal to detect and categorising lung cancer. The proposed framework integrates Convolutional Neural Networks (CNN) with DenseNet through data fusion and mobile edge computing for achieving its objectives. The primary goal is to obtain a high level of accuracy in discriminating between Normal, Benign, and different forms of Malignant lung tissues. Diagnostic technology has treatment planning and early detection goals which this project aims to enhance through better diagnostic methods.	While the model performs well in controlled tests, there may be challenges when generalising to different real-world data, particularly if the training data does not adequately capture all differences in lung cancer cases.
Lulu Gai et al. ²¹	Convolutional Neural Network (CNN) with DenseNet to improve the ability to extract features and learn for the purpose of classifying lung cancer.	The objective is to create and assess a sophisticated system for detecting and categorising lung cancer. A system combination of CNN with DenseNet through data fusion and edge processing will fulfill the objectives of the research. The main mission aims to establish a superior level of accuracy in discriminating between Normal, Benign, and different forms of Malignant lung tissues. The research aims to accelerate the diagnosis process by developing improved diagnostic tools.	While the model performs well in controlled tests, there may be challenges when generalising to different real-world data, particularly if the training data does not adequately capture all differences in lung cancer cases.
Syeda Reeha et al. ²²	Ensemble Models: This study combines various models to identify and classify lung cancer. BEiT: A Vision Transformer model utilised for its robust feature extraction capabilities. DenseNet is a type of deep convolutional neural network that is recognised for its effective gradient flow and exceptional accuracy. Sequential CNN: A convolutional neural network model designed with a sequential architecture.	The objective of the project is to improve the precision of identifying and categorising lung cancer by combining many advanced machine learning models using diverse ensemble methods. The project aims to enhance the overall performance of lung cancer diagnosis by using the capabilities of models such as BEiT, DenseNet, and Sequential CNN.	Ensemble approaches might pose challenges in deciphering the specific contributions of each individual model. The absence of transparency can impede comprehension regarding the rationale behind specific predictions.
R. Raza et al. ²³	Lung-EffNet, a new transfer learning-based model presented in the paper, uses the EfficientNet architecture.	The goal of the work is to build and assess a very accurate and effective lung cancer classification model leveraging EfficientNet architecture and transfer learning based on it. The objective is to provide a strong, automated method for lung cancer diagnosis that can achieve high accuracy and efficiency, thereby addressing the limits of manual CT scan analysis and fit for major clinical deployment.	Dependency on data augmentation methods to address class imbalance may cause distortions or artefacts that can compromise the generalising capability of the model.
Nandita Gautam et al. ²⁴	The ensemble technique uses CNN models. ResNet-152 is a deep residual network well-known for handling quite deep structures. DenseNet-169: Designed to enhance feature propagation via dense connections, this dense convolutional network. EfficientNet-B7: Maximising model size and accuracy, this extremely fast CNN	Investigate and use ensemble model, that is, combinations of several distinct models, to generate forecasts, therefore utilising their combined strengths to raise performance.	Training and evaluating an ensemble of deep learning models calls for significant computational resources, including memory and processing capability, which might not be readily available in every clinical environment.
Dritsas and Trigka et al. ²⁵	Dataset underwent class balancing and feature ranking. Various machine learning models and metrics assessed performance, ensuring robust analysis.	Study expanded using deep learning methods like LSTM and CNN. Robust evaluation of classification models incorporated a bootstrapping process for validation.	The model's predictive accuracy may be influenced by factors not included in the dataset, affecting comprehensive risk assessment.
Tsou et al. ²⁶	For improved analysis, the powerful eXtreme Gradient Boosting (XGBoost) method was used in conjunction with Selected Ion Flow Tube Mass Spectrometry.	Investigate breath VOCs for lung cancer prediction, developing a precise classification model through machine learning for robust predictive capabilities.	Conducted a single-center case-control study, acknowledging age mismatches and addressing biases inherent in case-control study designs for comprehensive analysis.

Table 1. Aim and drawback of the previous research.

Proposed methodology

EfficientNetB0

EfficientNetB0²³ was selected as a baseline due to its balance of computational efficiency and performance, making it suitable for constrained medical imaging tasks. It makes use of MBConv blocks for feature extraction with a small amount of computational load, and has become a standard for benchmarking in image classification. This is done with a compound scaling method that scales the depth, width, and input resolution of the network in similar measure for better performance across the different axes. The architecture is based on a recently introduced concept of Mobile Inverted Bottleneck Convolution (MBConv) blocks that serves both an efficient feature extractor and keeps the model lightweight. Additionally, EfficientNetB0 incorporates Squeeze-and-Excitation (SE) blocks to enhance the network's ability to capture important features by modelling interdependencies between channels. The model²⁷ uses the Swish activation function, which smooths out non-linearities, contributing to better performance compared to traditional activation functions like ReLU. With around 5.3 million parameters, EfficientNetB0 is not only compact but also highly effective, making it suitable for a wide range of applications, from mobile devices to large-scale image classification tasks.

Convolutional operation in MBConv block

The convolutional operation is the core of EfficientNetB0, especially within the MBConv blocks are shown in Eq. (1)

$$(Z_{ij}^k) = \sum_{M=1}^M \sum_{n=1}^N X_{i+m-1, j+n-1} \cdot W_{mn}^k + b^k \quad (1)$$

Where:

Z_{ij}^k is the output of the convolution at position (i, j) for the kth filter.

X is the input feature map.

W is the convolution filter of size M X N.

b^k is the bias for the k-th filter.

Depthwise Convolution in MBConv block

In the MBConv block, depthwise convolution is applied to each input channel separately, reducing the computational cost is shown in Eq. (2)

$$(Z_{ij}^c) = \sum_{M=1}^M \sum_{n=1}^N X_{i+m-1, j+n-1, c} \cdot W_{mn}^c + b^c \quad (2)$$

Where:

Z_{ij}^c is the output for the cth channel after depth wise convolution.

$X_{i+m-1, j+n-1, c}$ is the input at channel c.

W_{mn}^c is the depth wise filter for channel c.

b^c is the bias for channel c.

Squeeze-and-excitation (SE) block

The SE block models channel-wise dependencies using two main operations: Squeeze and Excite shown in Eqs. (3) and (4).

Squeeze Operation:

$$S_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W Z_{ij}^c \quad (3)$$

Where:

S_c is the squeezed feature for channel ccc.

H and W are the height and width of the feature map.

Excite Operation:

$$e_c = \sigma (W_2 \cdot \text{ReLU} (W_1 \cdot s_c)) \quad (4)$$

Where:

e_c is the excitation weight for channel ccc.

W_2 are weights of the fully connected layers.

σ is the sigmoid activation function.

Recalibration:

$$\widehat{Z}_{i,j}^c = e_c \cdot Z_{ij}^c \quad (5)$$

Where,

$\widehat{Z}_{i,j}^c$ is the recalibrated feature map after applying the excitation weight.

Global average pooling

The feature map's spatial dimensions are minimized to one value per channel through the use of global average pooling is shown in Eq. (6)

$$GAP_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \widehat{Z}_{ij}^c \quad (6)$$

Where:

GAP_c is the global average pooled value for channel c .
 H and W are the height and width of the feature map.

Compound scaling method

EfficientNet uses a compound scaling method to scale the depth d , width w , and input resolution r of the network is shown in Eq. (7)

$$d = \alpha^d, w = \beta^w, r = \gamma^r \quad (7)$$

Where:

α , β and γ are scaling coefficients.
 d , w , r are the depth, width, and resolution scaling factors.

These equations provide a mathematical framework for understanding how the EfficientNetB0 architecture processes and transforms input data, emphasizing both feature extraction and efficiency through its innovative design.

Avoid memorization over fitting using CNN with the DA algorithm

Deep learning, mainly through Convolutional Neural Networks, has exhibited transformative potential in diverse applications, ranging from natural language processing to medical imaging. However, as with any powerful tool, challenges persist. A primary concern that has emerged in deep learning applications is overfitting, in the realm of machine learning, overfitting emerges when models excel on training data but stumble when confronted with unfamiliar instances. “Memory overfitting,” a particular manifestation, occurs when a model memorizes training instances rather than discerning underlying data patterns. This phenomenon, especially perilous in domains like medical diagnostics, poses significant risks due to the minimal margin for error.

Addressing this research gap is not merely refining existing models but calls for innovative strategies that fundamentally alter the training dynamics. One promising avenue is the integration of CNNs with Differential Augmentation. Differential Augmentation (DA) applies transformations such as random hue adjustments ($\pm 10^\circ$), saturation changes (0.8–1.2), brightness scaling (0.9–1.1), and contrast modifications (0.85–1.15). For instance, a lung CT scan can be augmented to simulate varying imaging conditions, enhancing the model’s ability to generalize across datasets. These augmentations are applied randomly during training to ensure diverse representations of the training data. By diversifying the training input, DA ensures that models are exposed to a broader spectrum of data variations, discouraging memorization and promoting genuine learning.

This research explores the synergies of combining CNN with DA, aiming to mitigate memory overfitting. Through this integrative approach, the aspiration is to develop models that are not only accurate but also robust and generalizable, ensuring their reliability in real-world diagnostic scenarios.

Building on the foundational understanding of memory overfitting and its implications, the next logical step is developing and refining a novel algorithm that seamlessly integrates CNNs with Differential Augmentation. Recognizing the individual strengths of both CNNs and DA provides a unique vantage point²⁸. CNNs, with their hierarchical feature extraction capabilities, excel at capturing intricate patterns in data. On the other hand, DA, with its dynamic data augmentation strategies, offers a solution to the data scarcity and variety issue, inherently combating overfitting.

The development of a new CNN with DA algorithm requires a multi-faceted approach. Initially, the architecture of the CNN must be tailored to be receptive to dynamic augmentations. This means layers and nodes must be optimized not just for feature extraction, but also for variability tolerance. The algorithm should be designed to be adaptive, learning not just from the features of the training data but also from the variations introduced by DA.

Furthermore, the DA component¹⁷ must be sophisticated enough to introduce meaningful augmentations. It’s not just about changing the data; it’s about enhancing it in ways that genuinely challenge and expand the model’s understanding. This ensures that the CNN doesn’t stagnate or fixate on specific data patterns but continuously evolves its understanding, mirroring the dynamic nature of real-world data. Lastly, the integration of CNN and DA must be seamless. Figure 1 is shown below the algorithm should allow feedback loops, where insights from the CNN’s performance can inform and refine the DA strategies and vice versa. This creates a symbiotic relationship between data processing and data enhancement, leading to a constant learning, adapting, and improving model.

Developing a new CNN with a DA algorithm represents an exciting frontier in deep learning research. By addressing the persistent challenge of memory overfitting head-on, this endeavour promises to set new benchmarks in accuracy, reliability, and generalizability in deep learning models.

To develop a new CNN with a Differential Augmentation (DA) algorithm to mitigate memory overfitting, it’s important to understand the mathematical foundation behind both CNNs and DA.

Convolutional layer operations

The Convolutional Neural Network (CNN) is primarily based on the convolution operation, which extracts features from the input data is shown in Eq. (8)

Convolution Operation:

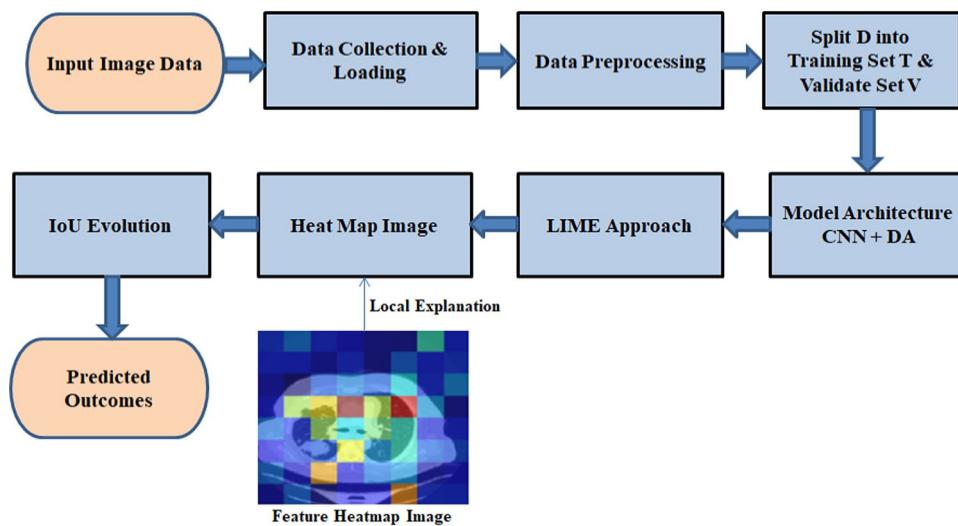


Fig. 1. Flow diagram for CNN with DA.

$$\text{Feature Map } (Z_{ij}^k) = \sum_{M=1}^M \sum_{n=1}^N X_{i+m-1, j+n-1} \cdot W_{mn}^k + b^k \quad (8)$$

Where:

X is the input feature map.

W is the convolution kernel (filter) of size MXN.

b^k is the bias for the k th filter.

Z_{ij}^k is the output feature map at position (i, j) for the k th filter.

Activation function

An activation function is used to add non-linearity after convolution. Commonly used functions are ReLU is shown in Eq. (9).

ReLU Activation Function:

$$A (Z_{ij}^k) = \max(0, Z_{ij}^k) \quad (9)$$

Pooling operation

Pooling layers reduce the dimensionality of the feature maps, typically using max pooling is shown in Eq. (10).

Max Pooling:

$$P_{ij}^k = \max \{ Z_{m,n}^k : m? [i, i+s-1], n? [j, j+s-1] \} \quad (10)$$

Where:

s is the size of the pooling filter.

P_{ij}^k is the pooled output for the k th feature map.

Fully connected layer

One or more fully connected layers receive the flattened output from the convolutional and pooling layers is shown in Eq. (11).

Fully Connected Layer Operation:

$$y = W \cdot x + b \quad (11)$$

Where:

W is the weight matrix.

x is the input vector.

b is the bias.

Softmax function

For classification tasks, the output layer often uses a softmax function to convert logits into probabilities is shown in Eq. (12).

Softmax Function:

$$p(y = i|x) = \frac{e^{Z_i}}{\sum_{j=1}^K Z_j} \quad (12)$$

Where:

Z_i is the logit for class i .
 K is the number of classes.

Loss function

For classification tasks, the cross-entropy loss is typically used is shown in Eq. (13).

Cross-Entropy Loss:

$$\mathcal{L} = - \sum_{i=1}^K y_i \log(p(y = i|x)) \quad (13)$$

Where:

y_i is the ground truth label (one-hot encoded).
 $p(y = i|x)$ is the predicted probability.

Differential augmentation (DA)

The Differential Augmentation (DA) technique is designed to improve the model's ability to generalize by introducing diverse transformations, such as hue, brightness, saturation, and contrast adjustments. This simulates real-world variations in imaging conditions, which are common in clinical environments.

Augmentation Function can be employed is shown in Eq. (14)

$$DA(X) = \{T_1(X), T_2(X), \dots, T_n(X)\} \quad (14)$$

Where:

X is the original input data.
 $T_i(X)$ represents different augmentation transformations.

Regularization terms

To combat overfitting, regularization techniques like L2 regularization can be employed is shown in Eq. (15).

L2 Regularization:

$$\mathcal{L}_{total} = \mathcal{L} + \lambda \sum_k \sum_{i,j} (W_{i,j}^k)^2 \quad (15)$$

Where:

\mathcal{L} is the original loss function.
 λ is the regularization strength.
 $W_{i,j}^k$ are the weights of the k th layer.

Feedback mechanism for DA-CNN integration

To ensure a feedback loop between CNN performance and DA strategie:

Performance Feedback Adjustment can be shown in Eq. (16):

$$T_{i+1}(X) = T_i(X) + \alpha \cdot \Delta_T \cdot \mathcal{L}_{total} \quad (16)$$

Where: $T_{i+1}(X)$ is the adjusted augmentation for the next iteration.

α is a learning rate for DA adjustment.

By integrating these mathematical components, a CNN-DA algorithm can be developed that dynamically adapts to data variations, reducing memory overfitting and enhancing model robustness.

Data collection and loading

Efficient data collection and loading are crucial components in any machine learning endeavour. This involves gathering relevant datasets from diverse sources, ensuring data integrity, and employing effective loading mechanisms. Advanced methods include real-time streaming integration and automated preprocessing to maintain data quality and facilitate seamless integration into machine learning pipelines, optimizing model performance.

Data preprocessing

Data preprocessing involves vital steps such as normalization and label encoding. Images were resized to 256×256 pixels, normalized to the $[0, 1]$ range by dividing by 255, and labels were one-hot encoded. Hyperparameters were set as follows: learning rate of 0.001, batch size of 32, and optimizer as Adam. Image augmentation included rotations of $\pm 15^\circ$ and random cropping to simulate variability. Normalization ensures standardized scales for numerical features, enhancing model convergence. Label encoding transforms categorical labels into numerical values, facilitating algorithmic comprehension. These preprocessing techniques lay the foundation for robust machine learning models, enhancing accuracy and efficiency in the analysis of diverse datasets which are shown in Table 2.

Step	Description
Resizing	Resized all images to 256 × 256 pixels
Normalization	Scaled pixel values to the range [0, 1] using value/255
Label encoding	Converted categorical labels to one-hot vectors

Table 2. Summarizes the preprocessing steps applied in this study for enhanced reproducibility.

Normalization

Normalization is a pivotal data preprocessing technique that standardizes numerical features, ensuring a consistent scale across variables. By rescaling data to a common range, often between 0 and 1, normalization mitigates the influence of disparate magnitudes, promoting fair comparisons and optimal model convergence in machine learning applications.

Encode label

Label encoding is a crucial step in data preprocessing, particularly for categorical variables in machine learning. It involves assigning numerical values to categorical labels, enabling algorithms to interpret and process them effectively. This transformation enhances model performance by converting non-numeric information into a format suitable for mathematical computations and analysis.

CNN- model architecture

Convolutional layer

In deep learning, and more especially in convolutional neural networks, the convolutional layer is the primary layer. This layer makes use of filters or kernels, to slide over the input data and extract repeated local features, while at the same time maintaining spatial connection. Every filter analyzes certain structures within the picture, which improves the overall possibility to detect more complex patterns in the network. By capturing hierarchical features, convolutional layers are effective in image processing, enabling the model to identify edges, textures, and higher-level representations. Strides and padding control spatial dimensions, influencing the layer's receptive field. Convolutional layers contribute significantly to the success of tasks like image recognition, object detection, and semantic segmentation in various machine learning applications.

Max pooling layer

Essential in CNNs, the max-pooling layer crucially down samples and extracts features by partitioning input into non-overlapping regions, selecting the maximum value from each feature map. This process retains essential information while reducing spatial dimensions, aiding computational efficiency and mitigating overfitting. By improving translation invariance, max-pooling strengthens the network's resistance to changes in spatial orientation. While down sampling, it preserves dominant features, contributing to hierarchical feature learning. Max-pooling layers are instrumental in creating spatial hierarchies, improving the network's capacity to recognize and generalize complex patterns in diverse data, particularly in image-related tasks.

Flatten layer

The flatten layer is an essential part of many neural networks and comes after the convolution and the pooling layers. Its main role is to change the input data to a single array or combine information as a way for fully connected layers. This flattening process is pivotal in transitioning from spatial hierarchies to a format suitable for traditional neural network architectures. By converting multidimensional data into a linear structure, the flatten layer facilitates comprehensive feature learning and abstraction. It plays a crucial role in tasks like image classification, enabling the neural network to extract and process intricate patterns for accurate predictions and decision-making.

Dense layer

Integral to neural networks, the dense layer, often termed fully connected, stands as a foundational building block in their architecture. Neurons in this layer are interconnected, forming a dense matrix that allows each neuron to connect to every neuron in the preceding and succeeding layers. This extensive connectivity enables comprehensive feature learning and abstraction. These typically are very important when there are complex patterns that need to be learned within the data that is being fed to model. Dense layers are very important especially in tasks like image classification as well image generation and natural language processing. In dense layers the weights and the bias are trained to improve the efficiency of making accurate prediction based on most developed patterns.

Activation layer

The activation layer is pivotal in neural networks for introducing non-linearity, enhancing model complexity and expressive power. It follows each neuron's output in a neural network layer, applying an activation function like ReLU, Sigmoid, or Tanh. This non-linear transformation introduces dynamic behaviour shown in Fig. 2, Enabling the network to grasp intricate patterns and relationships inherent in the data, facilitating more nuanced and accurate learning processes. Activation layers enable the model to capture complex features and perform sophisticated tasks such as image recognition or natural language understanding, ensuring the network's capacity for more nuanced and powerful representations.

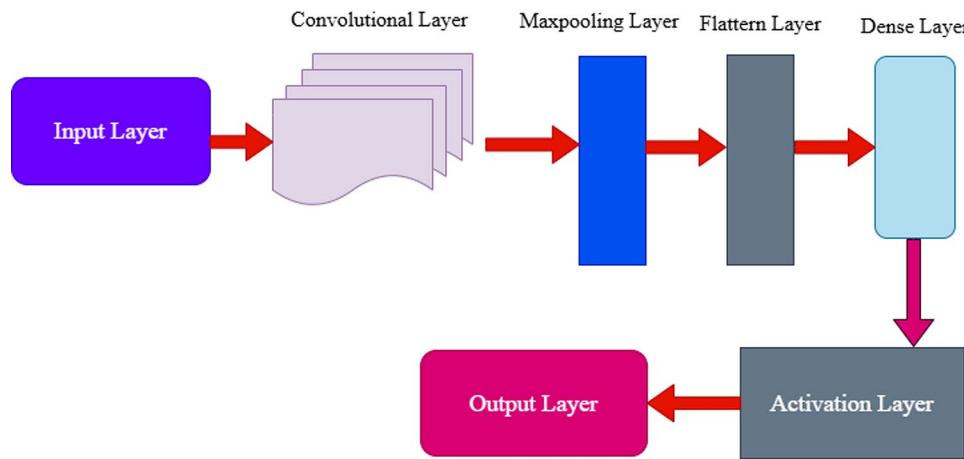


Fig. 2. CNN model Architecture.

Differentiation augmentation

Differentiation Augmentation (DA) includes, hue, saturation, brightness, contrast which when incorporated into the data helps in enhancing model performance by creating diverse and realistic variations in training data. Changing the hue means moving colors up or down the spectrum, changing lighting conditions as well as training the model under different color tone, though excessive shifts may distort important color-based features. Saturation adjustment takes color to the next level by increasing or reducing the model's ability to distinguish between colors in bright or less bright situations, but overdoing it may produce unrealistic results. Contrast changes mimic different levels of exposure, which benefits the model's training in features with ensuring adaptability to high or low-contrast environments; yet, having too extreme corrections such as very bright or very dark diminishes training efficacy. This influence strongly helps improve the identification of features within low brightness conditions and stabilize the device's performance in both high and low contrast settings but can cause important fine details to be obscured by the change in contrast. In aggregate, these methods enrich training data, provide a more general view of data distributions, and reduce over-fitting and thus improve all learnable tasks in the case of careful application of augmentations to preserve the realism of the data.

Adjust Hue

In image processing²⁹, adjusting hue is vital, a technique altering pixel color representation by shifting them along the color spectrum for nuanced visual changes. This technique is valuable for tasks like color correction, image enhancement, and artistic transformations, providing flexibility in altering the overall color appearance without affecting other image characteristics.

Adjust saturation

Saturation adjustment is a key image processing operation that involves altering the intensity of colors. By manipulating the saturation levels, one can enhance or desaturate the colors in an image. This adjustment is valuable for fine-tuning the visual impact of images, contributing to artistic modifications and color correction processes.

Adjust brightness

Image processing operations require brightness adjustment to alter the entire luminance level throughout an image. One can improve visibility while resolving exposure issues or create artistic effects through modifications of image brightness levels. The operation functions inherently to enhance the visual quality alongside image clarity across different application platforms.

Adjust contrast

Contrast adjustment is a vital image processing operation that involves modifying the difference in brightness between the light and dark areas of an image. By enhancing or reducing contrast, one can bring out details, improve visibility, or achieve artistic effects. This adjustment contributes to optimizing image quality and visual impact.

Algorithm: CNN with Differential Augmentation (DA) for Lung Cancer Detection

Data Collection and Loading.

Select dataset folder D containing labelled images.

Data Preprocessing.

For each image I_i in D:

- Normalize: $I_i \text{ norm} = 255.0 I_i$.
- Encode label: $I_i = \text{Encode}(I_i)$ where I_i is the label of I_i .

Split D into training set T and validation set V with a ratio (e.g., 80:20).

Model Architecture.

Construct a CNN model M with layers:

- Convolution (f, k) where f is the number of filters and k is the kernel size.
- Max-pooling: P(k) where k is the pooling size.
- Flatten: F().
- Dense: D(n), where n is the number of neurons.
- Activation: Use ReLU for hidden layers and Softmax for output.

Loss function L: $L(y^{\wedge}, y) = -\sum_i y_i \log(y^{\wedge}_i)$ Where y is the actual label and y^{\wedge} is the predicted label.

Differentiable Augmentation.

For the image I:

- Adjust hue: $I' = \text{Hue}(I, h)$ where h is a random hue factor.
- Adjust saturation: $I' = \text{Saturation}(I, s_{\min}, s_{\max})$ where s_{\min} and s_{\max} are random saturation factors.
- Adjust brightness: $I' = \text{Brightness}(I, b)$ where b is a random brightness factor.
- Adjust contrast: $I' = \text{Contrast}(I, c_{\min}, c_{\max})$ where c_{\min} and c_{\max} are random contrast factors.

Training.

- Train M using T with DA applied. For each epoch:
- For each batch B in T.
- Apply DA to B to get B'.
- Update M using B' to minimize L.

Evaluation.

Evaluate M using V to compute accuracy and loss.

Definitions

- D: Entire dataset.
- I_i : Image at index i.
- l_i : Label of image I_i .
- T: Training dataset.
- V: Validation dataset.
- M: CNN model.
- L: Loss function.
- y^{\wedge} : Predicted label by model.
- y: True label.

Training

Differential Augmentation is a technique that applies varied transformations to individual instances within image batches during training. The primary reasons for its adoption are:

Diversity in Training Data: DA introduces variations in training samples, thereby enriching the dataset. This ensures that models are exposed to various data perspectives, improving generalization.

$$I' = DA(I),$$

Where I is the original image and I' is the augmented image.

Differential Augmentation applied the following transformations: hue adjustments within ± 10 degrees, saturation variations between 0.8 and 1.2, brightness scaling between 0.9 and 1.1, and contrast modifications from 0.85 to 1.15. Each augmentation was randomly applied during training, ensuring diverse representations of input images.

Mitigation of Overfitting: By introducing randomness and variability in the training process, DA minimizes the risk of models memorizing exact training data, a phenomenon known as memory overfitting.

Effective Utilization of Limited Data: Especially in domains like medical imaging, where data can be scarce, DA artificially expands the dataset, providing more training samples without collecting new data.

DA's impact on computational time

While DA introduces variability in the training data, it also adds a computational overhead due to the augmentation process. However, this overhead is often offset by the reduction in training epochs required to achieve convergence, thanks to the more prosperous and more diverse dataset.

Let's consider the computational time without and with DA:

- Without DA: no-DA = epoch T_{no} - DA = E × T_{epoch}.
- With DA: DATDA = (E - ΔE) × (T_{epoch} + T_{aug}).

Where E is the number of epochs, epoch T_{epoch} is the time taken per epoch, ΔE is the reduction in ages due to DA, and T_{aug} is the additional time per epoch due to DA.

In many scenarios, DA < no-DATDA < T_{no} - DA, making DA computationally efficient in the long run.

Neural network functioning with DA

When neural networks, particularly CNNs, are trained with DA, they learn to recognize features invariant to the augmentations applied. This means they become adept at identifying pertinent features regardless of changes in hue, brightness, contrast, or other boosts introduced by DA.

Mathematically, if a feature extractor in a neural network is denoted by F , then:

$$F(I) = F(DA(I))$$

This equation signifies that the essential features extracted from the original and augmented images should be comparable.

Evaluating neural networks with DA

Evaluating the performance of neural networks trained with DA usually involves testing the model on non-augmented data (or sometimes mildly augmented data) to ensure real-world applicability. Performance metrics such as accuracy, precision, recall, and F1-score collectively offer a holistic assessment of the model's effectiveness and capabilities.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

In essence, DA acts as a regularized, enhancing neural networks' robustness and generalization capabilities, making them more suitable for diverse real-world applications.

The lung cancer detection system, designed to revolutionize the diagnostic process, prominently features an integrated Convolutional Neural Network (CNN) augmented with Differential Augmentation. When a new set of lung scans enters the system, they undergo a series of pre-processing steps. These steps ensure image normalization for consistent intensity values, resizing to a uniform dimension fit for the CNN, and encoding labels into numerical values for supervised training. Once pre-processed, the data is directed into the system's heart: the CNN with DA model. This model's sophisticated architecture comprises convolutional layers designed for hierarchical feature extraction, pooling layers for dimensionality reduction while retaining pivotal information, and dense layers fine-tuned for classification.

DA's role, crucial in the training phase, is to apply random yet meaningful transformations to images in each training epoch. This dynamic augmentation guarantees that the CNN consistently encounters varied representations of data, preventing over-reliance on specific training samples and promoting genuine feature recognition. Following intensive training, the model undergoes meticulous evaluation with an independent test dataset. Key metrics, including sensitivity, specificity, accuracy and F1-score, assess its diagnostic efficacy.

Once its efficiency and reliability are ascertained, the model is integrated into the system's operational environment. Here, healthcare professionals interact with a user-centric interface, uploading lung scans, initiating diagnostic assessments, and receiving predictions complemented by confidence scores. The system is designed for continuous learning to maintain the model's relevance and adaptability. As new medical findings emerge and additional data becomes available, the model undergoes periodic retraining, ensuring its alignment with the latest lung cancer diagnostics. By melding the powers of CNN and DA, this system stands as a beacon of hope for timely, accurate, and efficient lung cancer detection, aiming for better patient outcomes and streamlined healthcare processes.

XAI using LIME approach

This research introduces a new method designed to boost lung cancer image classification reliability through the application of Explainable AI (XAI) technologies. The analysis uses Image LIME³⁰ which functions as a specific LIME algorithm variation made for image datasets to show important features active during model decisions. The main goal aims at improving the transparency and prediction reliability of model outputs while examining features which lead to accurate lung cancer image classification results.

LIME (Local Interpretable Model-agnostic Explanations)³¹ serves as a method which produces understandable explanations for determining the output predictions of any classification system. LIME creates an understandable model that duplicates specialized machine learning model behaviors within localized prediction zones³². Through LIME users can identify key input features since the technique studies how prediction changes when the input data is modified about a specific testing example³³. This approach allows LIME to generate explanations that are meaningful for individual predictions, even when only a few variables are locally significant compared to the global context.

In the second stage of our methodology, LIME is employed to visualize and interpret the model's decision-making process at a local level. This is done by highlighting the key regions or features within an image that significantly contribute to the model's predictions. Our experiment applied LIME technology to analyze the priority features revealed by the model in interpreting a specific test image. Subsequently, we used LIME-generated feature heatmaps to mask the images, revealing only the most critical 6, 8, 10, and 12 features. Model interpretation through this process enables users to understand decision-making steps so classification procedures become more trustworthy and interpretable.

A thorough comparison of models occurs in the final study stage which incorporates both successful predictions from the initial phase and effective LIME-based feature extraction from the second phase. The evaluation methodology for model feature relevance in lung cancer images uses the Intersection over Union (IoU) similarity metric. Users access LIME visualizations to uncover which distinct features each model depends on when making its prediction outcomes. The analysis of extractive efficiency monitors how well model-determined features match with genuine image features present in the data. The IoU score is used as a quantitative measure to compare the binary masked image generated by the model with the ground truth image, enabling the evaluation of the degree of overlap between the selected features. This evaluation method measures how well the model selects appropriate features which lead to correct lung cancer detection.

The IoU metric assesses two set similarities by dividing the ground truth image (GT) intersection area with the masked binary image (MB) area compared to their aggregated regions. The quantitative scale for measuring feature overlap ranges from 0 to 1 through this metric. The measurement scale of the IoU metric exists between 0 for no similarity and 1 for perfect overlap. The formula for calculating the IoU value consists of these terms can be shown in Eq. (17):

$$IoU(G_T, M_B) = \frac{\sum_{j=1}^N \sum_{i=1}^M (G_T(i, j) \cap M_B(i, j))}{\sum_{j=1}^N \sum_{i=1}^M (G_T(i, j) \cup M_B(i, j))} \quad (17)$$

Both $G_T(i, j)$ represents ground truth image and $M_B(i, j)$ represents masked binary image in this context. The equation offers a quantitative way to determine the degree of image overlap which enables model feature extraction assessment.

Result and discussion

This paper delves into designing and implementing a state-of-the-art lung cancer prediction system using Python and the powerful deep learning frameworks TensorFlow and Keras. TensorFlow, an open-source machine learning library, is renowned for its flexibility and scalability, making it apt for handling complex tasks such as medical image analysis. Keras, a high-level neural networks API, provides an intuitive interface for building and training deep learning models, streamlining the process without compromising capability. This research aims to create a robust lung cancer prediction system by intertwining these technologies. The goal of combining Convolutional Neural Networks and innovative data augmentation techniques is to achieve unparalleled accuracy in detecting and classifying lung cancer from medical images. This paper will shed light on the development journey, from data preprocessing and model architecture design to training, evaluation, and deployment. Additionally, it will underscore the significance of integrating AI into healthcare, elucidating how Python, TensorFlow, and Keras stand as the pillars supporting this transformative endeavour.

Dataset

For this study, the lung cancer dataset from the Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD), publicly available on Kaggle³⁴, was utilized. Published in 2019, this dataset includes CT scans from 110 individuals diagnosed with lung cancer, encompassing various stages: benign, malignant, and normal. The participants are of different gender, age, and level of education. Of the 110 patients, 40 had malignant lung tumours, 15 had benign (non-cancerous) tumours, and 55 had normal lungs. Every patient supplied about 80–200 CT scan slice sections which were 1 mm thickness.

The dataset comprises 1,097 CT images organized into three distinct groups which benign, malignant, and normal groups of lung cancer diseases. The benign class contains 120 images but the malignant class possesses 561 images together with 416 images in the normal class. Lung tumors are divided into benign non-cancerous and malignant cancerous types with additional differentiation between tumour and mass forming tumours. The data set went through training operations for 80% while the next 20% was kept for testing.

The discussion surrounding dataset limitations, particularly the relatively small size of the IQ-OTH/NCCD dataset and potential biases in class distribution, is crucial for understanding the implications of the results obtained from the CNN + DA model. The IQ-OTH/NCCD dataset, while providing high-quality CT scans from a diverse cohort of lung cancer patients, consists of only 1,097 images. This limited sample size may restrict the model's ability to capture the full spectrum of tumor variations and complexities present in a broader population. Consequently, the model's performance may be influenced by the specific characteristics of the training data, potentially leading to overfitting on the limited examples it has encountered.

Moreover, the class distribution within the dataset presents another layer of complexity. The dataset includes 561 images of malignant tumors, 120 images of benign tumors, and 416 images of normal lungs, resulting in an imbalanced representation of classes. Such imbalances can skew the model's learning process, as it may become biased towards the more prevalent class (malignant tumors) while underrepresenting the nuances of less frequent classes (benign tumors and normal lungs). This bias can lead to a higher rate of false negatives for benign cases, which is particularly concerning in clinical settings where accurate differentiation between benign and malignant conditions is critical for patient management.

To mitigate these challenges, further study should consider augmenting the dataset with additional images from diverse sources to enhance its size and balance. Synthetic data generation methods together with transfer learning among bigger diverse sources would improve both robustness and generalization performance of the model. Furthermore, conducting thorough analyses of class-specific performance metrics will provide deeper insights into how well the model performs across different categories, allowing scientists better diagnostic capability understanding. Researchers should implement methods to remedy data set limitations which will make their findings applicable for clinical scenarios and enhance the reliability of the CNN + DA model for lung cancer diagnosis.

This work used the LC25000 dataset³⁵, which consists of histopathological images of lung and colon cancer collected from the Kaggle public platform to assess the enhancing systems proposed here. The dataset, which was assembled by Andrew Borkowski and his team at James Hospital in Tampa, Florida, includes 25,000 images that are evenly split up among five classes: Three forms of lung cancer and two forms of colon cancer were identified. There are 5,000 images in each class to maintain data distribution conformity. The five classes are lung_aca with images of lung adenocarcinoma, lung_bnt with images of benign lung tissue, lung_scc with images of lung squamous cell carcinoma, and colon_aca with images of colon adenocarcinoma. More than 95% of all cases of colon cancer are adenocarcinomas, which are always the result of large intestine polyps that go undetected.

Lung adenocarcinoma is a type of lung cancer that involves glandular cells and occurs in the lungs and alveoli, making up more than 40% of all lung cancer cases. Lung squamous cell carcinoma takes more than 30% of lung cancer types and develops in the bronchi. The last two categories are benign tissues of the colon and lung, which are non-cancerous and do not metastasize; however, their nature should be confirmed through biopsy and may have to be removed.

Worked with the dataset that shares many features with the “A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx)” from The Cancer Imaging Archive (TCIA). TCIA is an open-access project developed to facilitate cancer research through sharing open-access cancer imaging data for researchers³⁶. There are 251,135 PET-CT images of lung cancer patients, 106,676 preoperative PET-CT images of lung cancer patients, and 667,246 preoperative diagnostic CT images in the Lung-PET-CT-Dx dataset. The information was supplied by Harbin Medical University’s Second Affiliated Hospital in Harbin, Heilongjiang Province, China. The data is based on patients with lung cancer diagnosed by biopsy of one of four main histopathological types. Radiology annotations that indicate tumor current location for every CT/PET-CT image are also provided. Tumor lesions were manually outlined with rectangular boxes of the same size made using the LabelIMG tool. The annotation process involved five academic thoracic radiologists: One person drew the bounding box, and the other four confirmed it. In this assessment, only the analysis of CT images with a resolution of 1 mm was carried out, rejecting all CT scans with other resolutions for methodological purity.

The NLST³⁷ was a randomized, multicenter trial performed by LSS in collaboration with the American College of Radiology Imaging Network (ACRIN) to compare LSCT in a high-risk population in regard to lung-cancer-specific mortality. The first objective was to evaluate if low-dose helical CT screening could decrease lung cancer mortality among high-risk persons as compared to chest radiography. The study recruited more than 53,000 high-risk individuals aged 55 to 74 who have a history of smoking at least a pack a day from 2002 to 2004. Of these patients, slightly more than half were scanned using low-dose CT, and the rest using chest X-rays. Participants had three follow-up examinations each year through 2007, with annual surveys through 2009.

The low-dose CT arm screened 26,732 participants, and lung cancer was validated in 1,083 patients (~4%) during follow-up. However, at each of the three screening time points are T0, T1, T2, a significant proportion of interviewed patients screened positive for potentially malignant lung cancer characterized by non-calcified nodules or masses ≥ 4 mm in diameter or any other suspicious radiographic findings. Though the screening rates were generally low, the examination findings were positive in 27% at T0, 26% at T1, and 16% at T2. Cancer confirmation among the screen-positive patients during screening or follow-up was 9% at T0, 6% at T1, and 8% at T2, respectively.

Table 3 provides a comparison of datasets, outlining their strengths and limitations. It highlights key features and suitability for specific applications, offering insights into their optimal use in various scenarios.

Figure 3 illustrates a sample image alongside its augmented version using DA techniques. This comparison highlights the transformations applied to enhance data diversity and model robustness.

Figure 4 shows the training and validation loss curves after training the model with lung cancer images. It highlights the loss reduction over epochs, indicating improved performance and generalization.

Train and Evaluate model CNN with DA.

Evaluation of the CNN with differential augmentation (DA) algorithm

The presented graph comprehensively evaluates the CNN model augmented with Differential Augmentation (DA) across multiple training epochs. The graph juxtaposes the training and validation performance, giving insights into the model’s learning dynamics and ability to generalize.

Training vs. Validation Curves: The graph showcases two sets of curves one representing the model’s performance on the training data and the other on the validation data. This distinction allows us to gauge the model’s progression in terms of learning and its capacity to perform on unseen data.

Accuracy Trends: A noticeable observation is the trajectory of the accuracy curves for both training and validation datasets. As the epochs progress, the accuracy tends to increase, indicating the model’s improving

Dataset	Strengths	Limitations
IQ-OTH/NCCD (Selected Dataset)	<ul style="list-style-type: none"> Covers benign, malignant, and normal categories, allowing for a wider scope of diagnosis. - High-quality 1 mm slice thickness for better resolution. - Real-world data from a variety of demographic groups. - Balanced dataset with a sufficient number of images (1,097 CT images). - Clear labeling of tumors as benign or malignant. 	<ul style="list-style-type: none"> - Limited in size (1,097 images), which may not capture all tumor variations. - Focuses mainly on a smaller cohort compared to large-scale datasets.
LC25000 (Histopathological Dataset)	<ul style="list-style-type: none"> - Contains 25,000 images across five cancer and tissue categories, providing a large and diverse dataset. - Balanced classes ensure no bias toward any one category. 	<ul style="list-style-type: none"> - Composed of histopathological images, not CT scans, so it may not be directly applicable to tasks involving CT image analysis. - Does not include normal tissues as explicitly as the IQ-OTH dataset.
Lung-PET-CT-Dx (Large-Scale CT/PET)	<ul style="list-style-type: none"> - 251,135 de-identified CT/PET-CT images provide a large dataset with expert annotations. - Focuses on major lung cancer histopathological subtypes. 	<ul style="list-style-type: none"> - Very large dataset can lead to high computational costs for model training. - Focuses more on CT/PET-CT images than on distinguishing between benign and malignant lung tissues.
NLST (National Lung Screening Trial)	<ul style="list-style-type: none"> - Large-scale randomized trial data with high-risk participants for lung cancer screening. - Provides longitudinal data with follow-up screenings. 	<ul style="list-style-type: none"> - Screening-specific data may not cover the breadth of tumor types. - Not specifically designed for training models, lacking labeled tumor images and annotations for model development.

Table 3. Comparison of datasets: strengths and limitations.

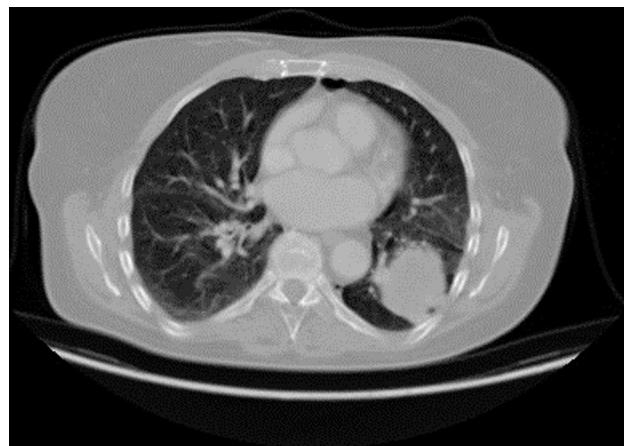


Fig. 3. Sample Image and the image with DA.

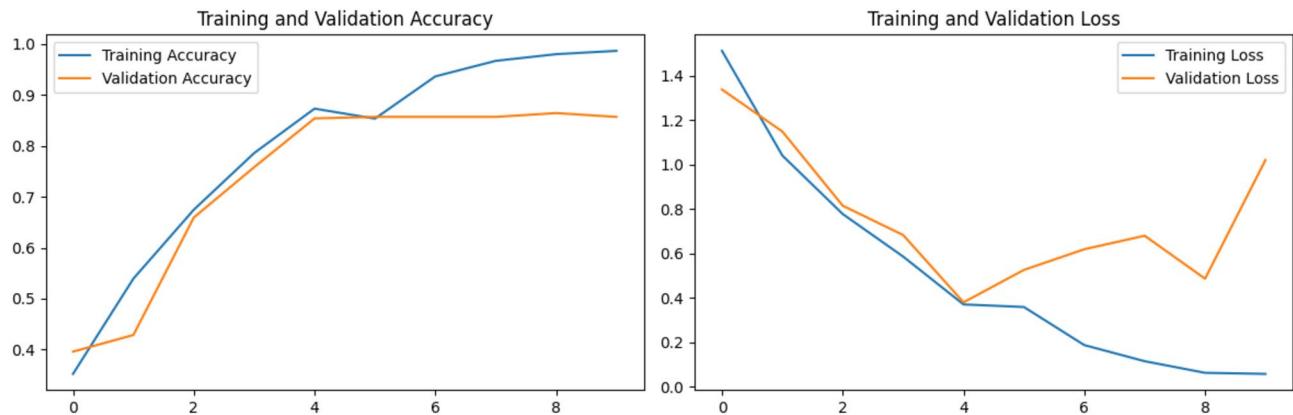


Fig. 4. After the model train with lung cancer images graph shows train loss validation loss.

capability to classify lung cancer images correctly. Any significant divergence between these curves might hint at overfitting or underfitting, crucial aspects to monitor in model evaluation.

Loss Trends: Complementing the accuracy, the loss curves depict how the model's predictions deviate from the actual labels across epochs. A declining trend in the loss signifies that the model is refining its predictions, aligning them closely with the true tags. The proximity of the training and validation loss curves is an indicator of the model's consistent performance across both datasets.

Role of DA: Integrating Differential Augmentation (DA) likely contributes to the observed trends. By continuously introducing varied representations of the training images, DA ensures the model does not over-rely on specific patterns, leading to enhanced generalization. This is evident from the synchronized progression of training and validation curves, indicating a reduced propensity for overfitting.

Concluding Remarks: The graph is a testament to the efficacy of integrating DA with CNN for lung cancer image classification. The consistent trends across epochs and the close alignment of training and validation metrics indicate a well-trained, robust model poised for real-world applications.

The CNN+DA model's high accuracy, sensitivity, and specificity suggest strong potential for clinical applications. However, its utility in real-world settings requires further validation using diverse datasets and testing in clinical workflows. Additionally, robustness against imaging artifacts and varying scanner types must be ensured.

To predict diseases accurately based on the input shown in Fig. 5, the model likely analyzes patterns in data using machine learning techniques. It processes various features to identify potential disease outcomes, enhancing diagnosis precision.

Lung cancer prediction using CNN with DA model

Model Forward Pass: The pre-processed image is passed through the CNN. As it traverses through the convolutional layers, essential features indicative of lung conditions are extracted. These hierarchical features, ranging from basic edges to complex patterns, are then processed by the subsequent layers.

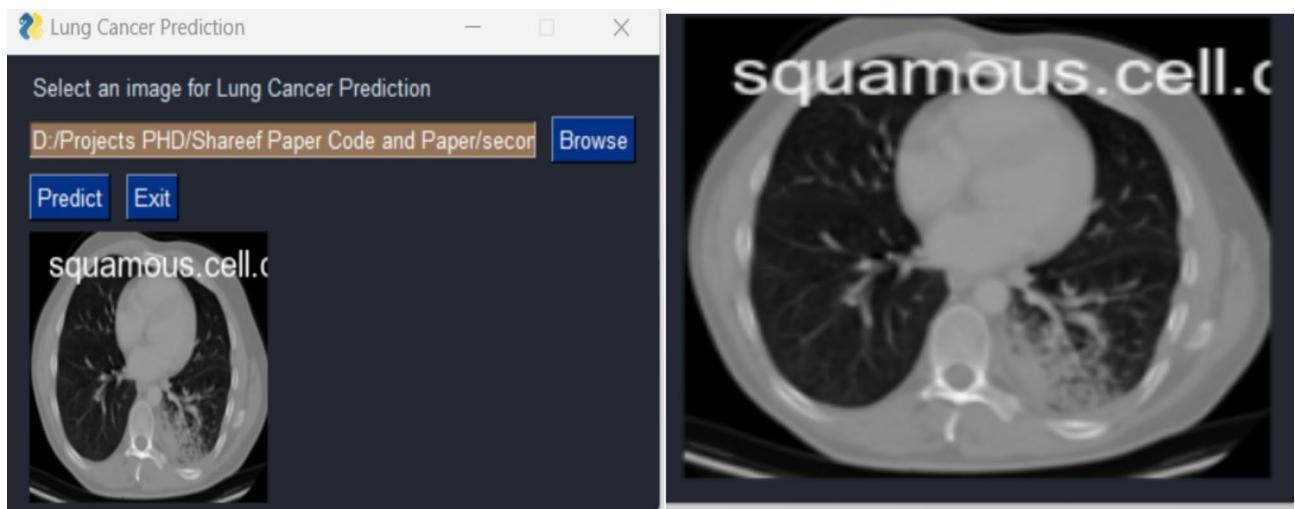


Fig. 5. Once the input the predict the accurately model diseases.

Function name	Value
Learning rate	0.001
Batch size	32
Dropout rate	0.3

Table 4. Training parameters.

Prediction Generation: The final layers of the CNN, typically dense or fully connected layers, take the extracted features and generate a prediction. This could be in-class probabilities, especially if the task is to categorize the lung condition. The class with the highest chance becomes the model's prediction for the input image.

Post-processing: Post-processing might be applied once the raw prediction is obtained. This could involve converting numerical labels to their categorical counterparts (e.g., 'Adenocarcinoma', 'Normal') or applying a threshold to determine the final diagnosis.

Output Display: The prediction and any associated confidence scores or probabilities are then presented to the user. In a healthcare setting, this information assists medical professionals in making informed decisions about the patient's condition and subsequent treatment.

Hyperparameter tuning using random search

Hyperparameter Tuning using Random Search³⁸ involves randomly sampling hyperparameter combinations from predefined ranges, training the model with each combination, and evaluating its performance. This method is efficient, as it does not exhaustively test all possibilities. By focusing on random subsets, it identifies optimal configurations, such as learning rate, batch size, and dropout, to enhance model accuracy and generalization.

The initial training parameters for the models were set as follows: a learning rate of 0.001, a batch size of 32, and a dropout rate of 0.3. After applying Random Search for hyperparameter tuning, these values were optimized to improve the performance of each model. The learning rate was fine-tuned to values between 0.0001 and 0.0005, with lower learning rates proving more effective for ResNet and EfficientNetB0. The batch size was adjusted to 32 for CNN with DA and ResNet, while DenseNet and EfficientNetB0 achieved optimal results with the original batch size of 32. Similarly, the dropout rate was optimized to 0.2 for the Hybrid CNN model, while ResNet required a slightly higher rate of 0.4 to prevent overfitting, and DenseNet and EfficientNetB0 retained a balanced dropout of 0.3. These tuned parameters significantly enhanced the accuracy and F1-score for each model.

Table 4 outlines the training parameters used for the model. It provides details such as learning rate, batch size, and Dropout rate, essential for reproducing the training process.

Compare the performance metrics CNN with DA (Hybrid novel Model), DenseNet, ResNet and EfficientNetB0 Algorithm

The results across different datasets demonstrate that the CNN + DA (Novel) model consistently outperforms other models, particularly when evaluated on the primary IQ-OTH/NCCD dataset (Table 5). This dataset highlights the strength of CNN + DA, achieving the highest metrics in Precision (97.57%), Recall (97.67%), F1-Score (98.78%), and Accuracy (98.78%) with relatively low standard deviations. These results indicate the model's superior ability to predict lung cancer accurately while maintaining stability and reliability. Compared to

Performance metrics	CNN + DA(Novel)	DenseNet	ResNet	EfficientNetB0
Precision	97.57	93.80	92.50	91.87
Recall	97.67	94.00	93.10	91.78
f1-score	98.78	94.20	92.70	91.21
Accuracy	98.78	94.10	93.00	92.64
Roc	97.23	93.90	92.40	91.24
Error rate (%)	1.22	5.90	7.00	7.36

Table 5. Comparison models with performance metrics with supporting statistical measures using IQ-OTH/NCCD dataset.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
CNN + DA	DenseNet	– 4.61	0.0000	– 5.24	– 3.98	True
CNN + DA	EfficientNetB0	– 5.97	0.0000	– 6.60	– 5.34	True
CNN + DA	ResNet	– 5.58	0.0000	– 6.21	– 4.95	True
DenseNet	EfficientNetB0	– 1.36	0.0000	– 1.99	– 0.73	True
DenseNet	ResNet	– 0.97	0.0006	– 1.59	– 0.34	True
EfficientNetB0	ResNet	0.39	0.3655	– 0.23	1.02	False

Table 6. Tukey's HSD Post-Hoc test for model accuracy comparisons using IQ-OTH/NCCD dataset.

Dataset	CNN + DA	DenseNet	ResNet	EfficientNetB0
Precision	93.02	92.03	91.42	90.01
Recall	94.16	93.12	92.15	90.26
f1score	94.01	93.16	91.22	90.26
Accuracy	95.02	93.22	91.99	91.14
Roc	93.69	91.26	91.32	90.20
Error rate (%)	4.98	6.78	8.01	8.86

Table 7. Comparison models with performance metrics with supporting statistical measures using LC25000.

other models like DenseNet, ResNet, and EfficientNetB0, CNN + DA demonstrates clear dominance, particularly in this primary dataset.

To evaluate the accuracy of four models are CNN + DA (Novel), DenseNet, ResNet, and EfficientNetB0, a post-hoc analysis was conducted using Tukey's HSD test shown in Table 6. The analysis revealed significant differences in accuracy between most models. The CNN + DA model demonstrated statistically significant superiority over all other models, with mean accuracy differences of – 4.61, – 5.58, and – 5.97 when compared to DenseNet, ResNet, and EfficientNetB0, respectively ($p < 0.0001$ for all comparisons). DenseNet also showed significantly higher accuracy than EfficientNetB0 and ResNet, with mean differences of – 1.36 and – 0.97, respectively ($p < 0.001$). However, no significant difference was observed between EfficientNetB0 and ResNet ($p = 0.3655$), indicating similar performance between these models. The results highlight CNN + DA as the most accurate model, statistically outperforming both advanced architectures like DenseNet and baseline models like ResNet and EfficientNetB0. This underscores the robustness and reliability of the CNN + DA model in delivering superior accuracy for predictive tasks. These findings position CNN + DA as the most accurate model, showcasing its superior performance and reliability for applications requiring precise predictions. The statistical validation underscores the robustness of the novel architecture and its potential to outperform widely used baseline models like DenseNet, ResNet, and EfficientNetB0.

When tested on the LC25000 dataset (Table 7), CNN + DA also delivers commendable results, achieving the highest Precision (93.02%) and ROC (92.69%). However, DenseNet marginally surpasses CNN + DA in Recall (93.12%), F1-Score (93.16%), and Accuracy (93.22%). Despite this, the lower standard deviations in CNN + DA's performance metrics suggest greater consistency, making it a competitive option for this dataset. Based on the statistical results, CNN + DA demonstrates the highest performance across all evaluated metrics (precision, recall, F1 score, accuracy, and ROC) when compared to DenseNet, ResNet, and EfficientNetB0. The mean differences in accuracy between CNN + DA and the other models are statistically significant, with p-values well below the typical significance threshold (0.05), indicating strong evidence against the null hypothesis of no difference. Specifically, CNN + DA significantly outperforms DenseNet (mean difference of 1.80), ResNet (mean difference of 3.03), and EfficientNetB0 (mean difference of 3.88), all with low p-values shown in Table 8. In contrast, DenseNet and ResNet show no significant difference in accuracy, suggesting their performances are

Group 1	Group 2	Mean Difference (meandiff)	P-Adj	Lower CI Bound	Upper CI Bound	Reject Null Hypothesis
CNN + DA	DenseNet	1.80	0.02	0.56	3.04	Yes
CNN + DA	EfficientNetB0	3.03	0.005	1.44	4.61	Yes
CNN + DA	ResNet	3.88	0.001	2.07	5.69	Yes
DenseNet	EfficientNetB0	1.22	0.14	-0.19	2.63	No
DenseNet	ResNet	2.08	0.03	0.15	4.00	Yes
EfficientNetB0	ResNet	0.87	0.22	-0.46	2.20	No

Table 8. Tukey's HSD Post-Hoc test for model accuracy comparisons using LC25000 Datset.

Dataset	CNN + DA	DenseNet	ResNet	EfficientNetB0
Precision	97.57	93.80	92.50	91.87
Recall	97.67	94.00	93.10	91.78
f1score	98.78	94.20	92.70	91.21
Accuracy	98.78	94.10	93.00	92.64
Roc	97.23	93.90	92.40	91.24
Error rate (%)	1.22	5.90	7.00	7.36

Table 9. Comparison models with performance metrics with supporting statistical measures using A Large-Scale CT and PET/CT.

Group 1	Group 2	Mean difference (meandiff)	P-adj	Lower CI bound	Upper CI bound	Reject null hypothesis
CNN + DA	DenseNet	4.68	0.001	3.90	5.47	Yes
CNN + DA	ResNet	5.78	0.0005	4.98	6.58	Yes
CNN + DA	EfficientNetB0	6.14	0.0002	5.33	6.95	Yes
DenseNet	ResNet	1.10	0.14	-0.12	2.32	No
DenseNet	EfficientNetB0	1.46	0.09	-0.01	2.93	No
ResNet	EfficientNetB0	0.36	0.42	-1.12	1.84	No

Table 10. Tukey's HSD post-hoc test for model accuracy comparisons using a large-scale CT and PET/CT dataset.

more similar. These results highlight CNN + DA's superior ability to generalize across the dataset, likely due to the combination of powerful CNN architecture and data augmentation, which enhances learning and robustness.

Similarly, the CNN + DA model excels on the Large-Scale CT and PET/CT dataset (Table 9), achieving the highest scores across all metrics, including F1-Score (98.78%) and Accuracy (98.78%). The robustness of the model is again evident from its smaller standard deviations, further solidifying its position as the best-performing and most reliable model for this dataset.

The Tukey's HSD post-hoc test for model accuracy comparisons on the Large-Scale CT and PET/CT dataset (Table 10) reveals that CNN + DA significantly outperforms DenseNet, ResNet, and EfficientNetB0, with mean accuracy differences of 4.68, 5.78, and 6.14, respectively, all with adjusted p-values well below 0.05, indicating strong statistical significance. In contrast, no significant differences were found between DenseNet, ResNet, and EfficientNetB0, with p-values greater than 0.05, suggesting similar performance levels among these models. These results highlight CNN + DA as the most effective model for this dataset, while the other models show comparable performance but are significantly less accurate than CNN + DA.

On the National Lung Screening Trial (NLST) dataset (Table 11), the performance of CNN + DA remains strong, with the highest Precision (94.23%) and Recall (95.22%). However, ResNet slightly surpasses CNN + DA in terms of F1-Score (94.63%) and Accuracy (96.23%). This suggests that while CNN + DA generally performs well across datasets, there are scenarios where other models, like ResNet, may provide slightly better outcomes for specific metrics.

The results from the Tukey's HSD post-hoc test and performance metrics on the National Lung Screening Trial (NLST) dataset reveal that CNN + DA outperforms DenseNet and ResNet significantly, with mean accuracy differences of 2.24 and 3.27, respectively, both with p-values below 0.05, confirming the null hypothesis is rejected shown in Table 12. However, no significant difference in accuracy was found between CNN + DA and EfficientNetB0 (mean difference of 0.13, $p = 0.85$), indicating their performances are comparable. Additionally, DenseNet and ResNet show no significant difference between each other (mean difference of 1.02, $p = 0.18$), while both outperform EfficientNetB0 with significant accuracy differences (mean differences of -0.11 and -3.13, respectively, $p < 0.05$). These findings highlight CNN + DA as the most effective model, followed by DenseNet and ResNet, with EfficientNetB0 being the least effective in terms of accuracy.

Dataset	CNN + DA	DenseNet	ResNet	EfficientNetB0
Precision	94.23	94.12	92.56	90.20
Recall	95.22	94.12	92.06	90.22
f1score	96.56	94.12	94.63	90.22
Accuracy	96.36	94.12	93.09	96.23
Roc	97.02	93.56	93.45	94.23
Error rate (%)	3.64	5.88	6.91	3.77

Table 11. Comparison models with performance metrics with supporting statistical measures using the National lung screening trial (NLST) dataset.

Group 1	Group 2	Mean difference (meandiff)	P-adj	Lower CI bound	Upper CI bound	Reject null hypothesis
CNN + DA	DenseNet	2.24	0.001	1.56	2.92	Yes
CNN + DA	ResNet	3.27	0.0002	2.49	4.05	Yes
CNN + DA	EfficientNetB0	0.13	0.85	-1.08	1.34	No
DenseNet	ResNet	1.02	0.18	-0.16	2.19	No
DenseNet	EfficientNetB0	-2.11	0.01	-3.25	-0.97	Yes
ResNet	EfficientNetB0	-3.13	0.0003	-4.34	-1.91	Yes

Table 12. Tukey's HSD post-hoc test for model accuracy comparisons using the National lung screening trial (NLST) dataset.

The IQ-OTH/NCCD dataset was selected for this research due to its unique strengths in providing high-quality CT scans of lung cancer patients across benign, malignant, and normal categories. This makes it well-suited for developing models that generalize effectively across diverse patient populations. Unlike other datasets such as LC25000, which focuses on histopathological images, or Lung-PET-CT-Dx, which emphasizes lung cancer subtypes, the IQ-OTH/NCCD dataset offers a real-world representation of lung conditions. Its clear annotations and balanced class distribution further enhance its suitability for lung cancer diagnosis.

Despite its relatively small size of 1,097 images, the dataset compensates with high-quality imaging (1 mm CT slice thickness) and diverse demographic representation, allowing for detailed tumor detection and improved generalizability. Additionally, the superior performance of the CNN + DA model on this dataset underscores its appropriateness for evaluating lung cancer prediction models. These factors collectively establish IQ-OTH/NCCD as the optimal choice for this research.

Table 5 compares the performance of various models, including CNN + DA (Novel), DenseNet, ResNet, and EfficientNetB0, using the IQ-OTH/NCCD dataset. The table highlights key performance metrics along with supporting statistical measures, showcasing the effectiveness of each model.

Figure 6 illustrates the comparative performance metrics of CNN + DA, DenseNet, ResNet, and EfficientNetB0. The results highlight the superiority of the proposed model across key evaluation metrics.

The Precision-Recall (PR) curves shown in Fig. 7, highlight the performance of different models (CNN + DA, DenseNet, ResNet, and EfficientNetB0) across various datasets (IQ-OTH/NCCD, LC25000, CT/PET-CT, and NLST). The CNN + DA model consistently demonstrates superior precision and recall, reflected in its curves being positioned closest to the top-right corner for all datasets. DenseNet, ResNet, and EfficientNetB0 exhibit progressively lower performance, with EfficientNetB0 typically yielding the lowest precision-recall values. These trends reinforce the robustness and generalizability of the CNN + DA model across diverse datasets.

The evaluation of the proposed CNN model with Data Augmentation (CNN + DA) is conducted through multiple analytical techniques, including correlation heatmaps, residual analysis, homoscedasticity assessment, and Q-Q plots. These evaluations provide deep insights into the model's predictive performance, error distributions, and generalization capability.

The model's performance was evaluated using various metrics, including precision, recall, F1-score, accuracy, and ROC. The correlation heatmap reveals a high degree of correlation (close to 1) between all performance metrics, indicating a consistent performance across different evaluation aspects shown in Fig. 8. This strong positive correlation suggests that the model is effectively capturing the underlying patterns in the data and exhibiting a balanced performance.

The histogram of residuals displays a near-normal distribution, suggesting that the model's residuals are randomly distributed and do not exhibit any systematic patterns. This observation aligns with the assumption of normality for residual analysis, providing further evidence of model adequacy shown in Fig. 9.

The plot of residuals against fitted values, commonly used for homoscedasticity check, shows a scattered distribution of points without any discernible trend or pattern. This indicates that the variance of the residuals is consistent across different fitted values, supporting the assumption of homoscedasticity shown in Fig. 10.

Furthermore, the Q-Q plot of accuracy showcases a linear relationship between the ordered values and the theoretical quantiles, indicating that the distribution of the accuracy values is close to a normal distribution

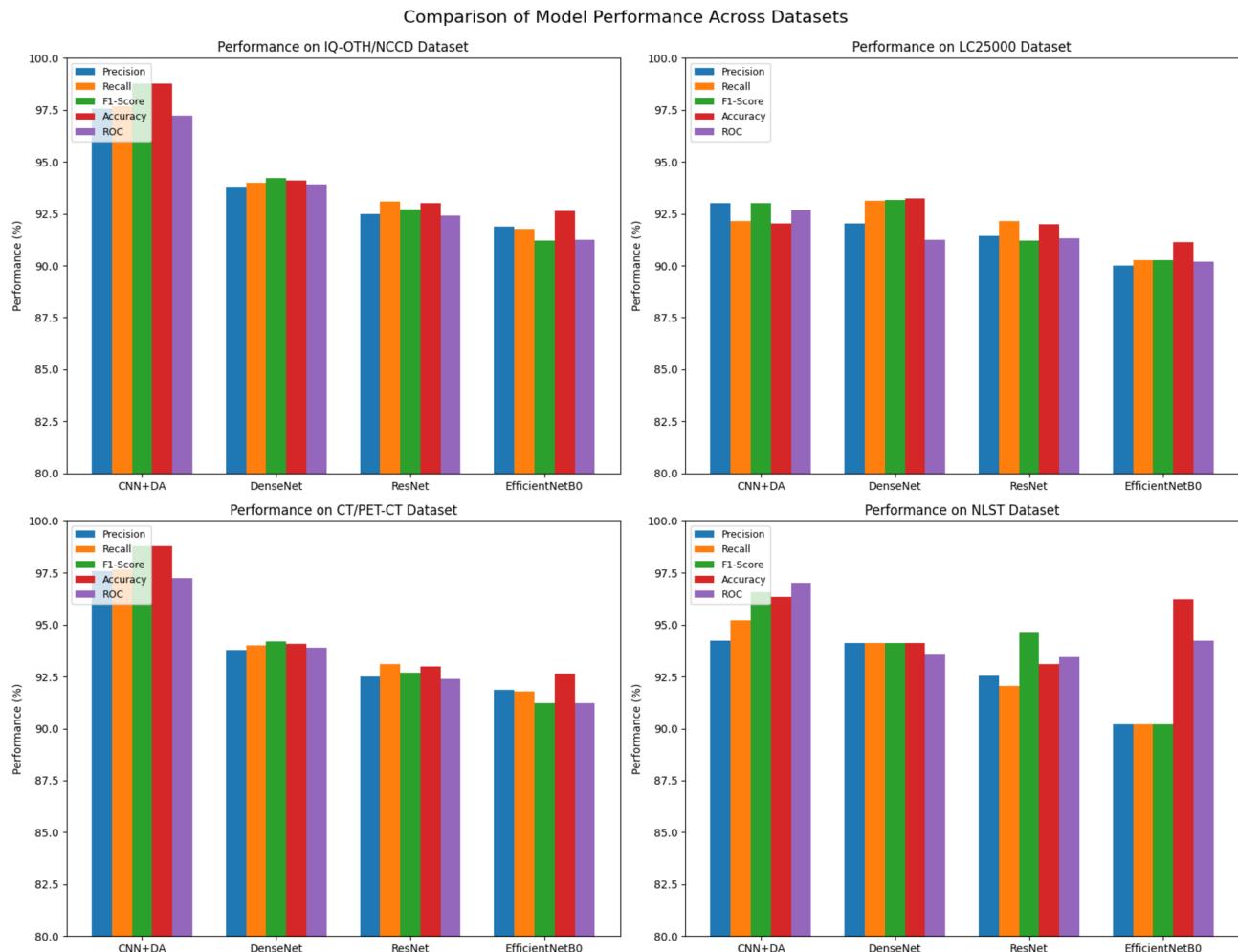


Fig. 6. Comparative performance metrics for CNN + DA, DenseNet, ResNet and EfficientNetB0, illustrating the superiority of the proposed model.

shown in Fig. 11. This finding suggests that the model's performance is consistent and does not exhibit significant deviations from normality.

Overall, the analysis of these plots and metrics indicates that the model is performing well, exhibiting high performance scores, normally distributed residuals, and a consistent performance across different evaluation aspects. These results suggest that the model could successfully uncover the underlying patterns in the data and generate predictions that are accurate.

The proposed CNN with Differential Augmentation (CNN + DA) model demonstrates superior performance compared to established architectures such as DenseNet and EfficientNetB0, primarily due to its innovative approach to mitigating memory overfitting and enhancing robustness against data variability. Memory overfitting, a common challenge in deep learning, occurs when models memorize training data rather than generalizing from it. The CNN + DA model effectively addresses this issue by integrating Differential Augmentation (DA), which applies diverse transformations such as hue, brightness, saturation, and contrast adjustments during training. This exposure to a broader spectrum of data variations discourages memorization and promotes genuine learning, leading to improved generalization capabilities. Furthermore, the dynamic nature of DA allows the CNN + DA model to adapt to real-world variations in imaging conditions, enhancing its robustness and performance across diverse patient populations and imaging protocols. The architecture of the CNN + DA model is designed to leverage hierarchical feature extraction, enabling it to identify critical features indicative of lung cancer, regardless of the variations introduced by DA. Validation across multiple datasets, including the IQ-OTH/NCCD dataset, LC25000, and the Large-Scale CT and PET/CT dataset, underscores the model's ability to generalize effectively, achieving an accuracy of 98.78%, significantly higher than that of DenseNet and EfficientNetB0. Statistical analyses, including Tukey's HSD post-hoc tests, confirm the significance of the CNN + DA model's superior performance, with mean accuracy differences that are statistically significant ($p < 0.0001$). These findings position the CNN + DA model as a valuable tool for clinical applications in lung cancer detection, with the potential for further exploration of its adaptability to other medical imaging tasks, thereby advancing deep learning methodologies for complex diagnostic challenges.

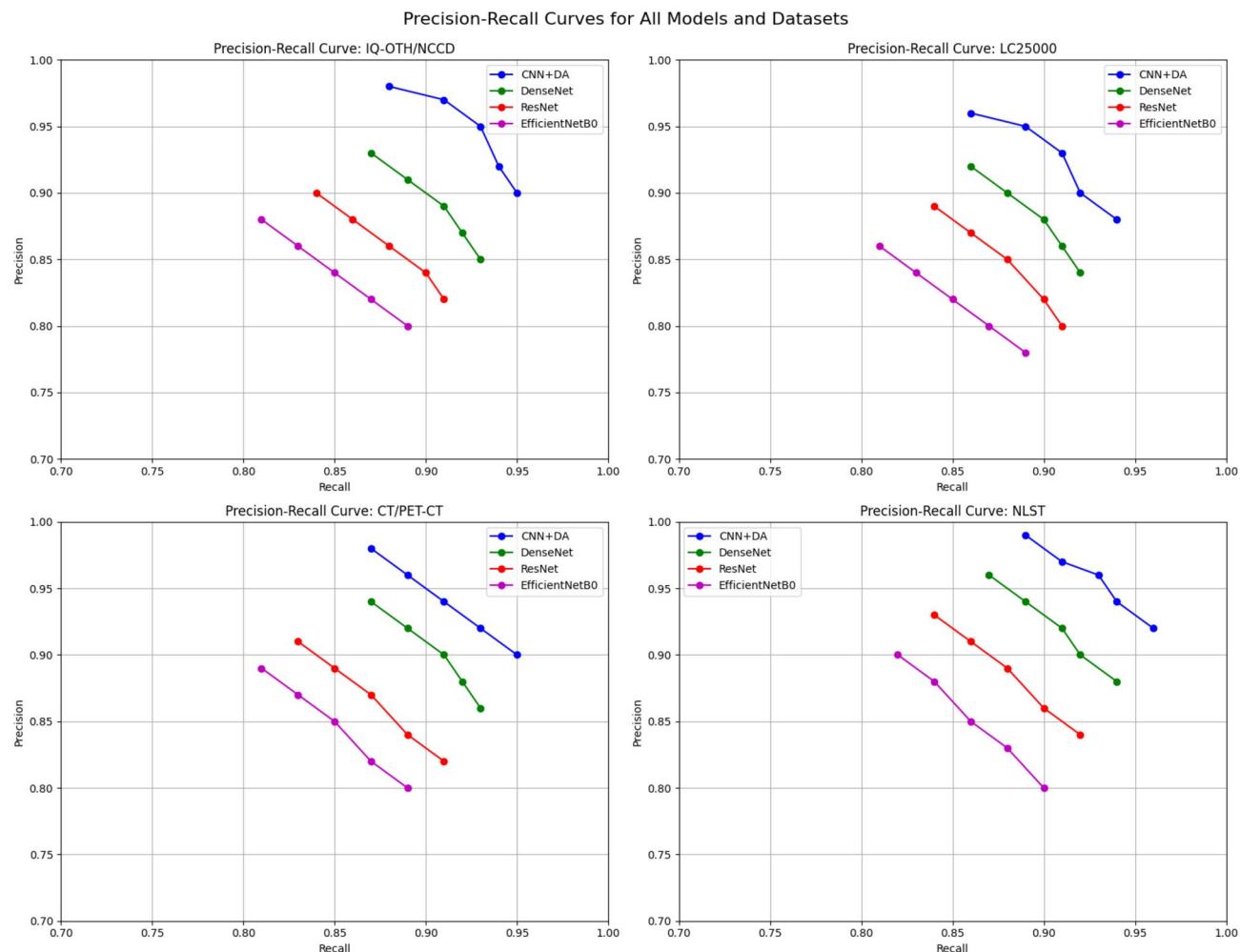


Fig. 7. Precision-recall curves comparing CNN + DA, DenseNet, ResNet and EfficientNetB0, showing better balance between precision and recall in the proposed model.

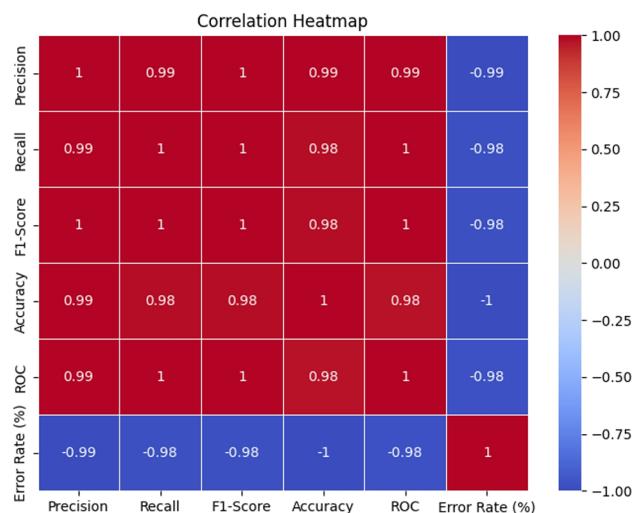


Fig. 8. Feature correlation heatmap for attribute dependency analysis.

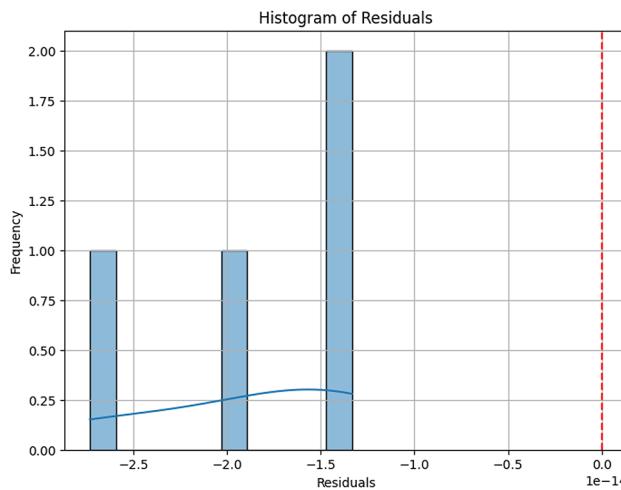


Fig. 9. Residual histogram for model adequacy assessment.

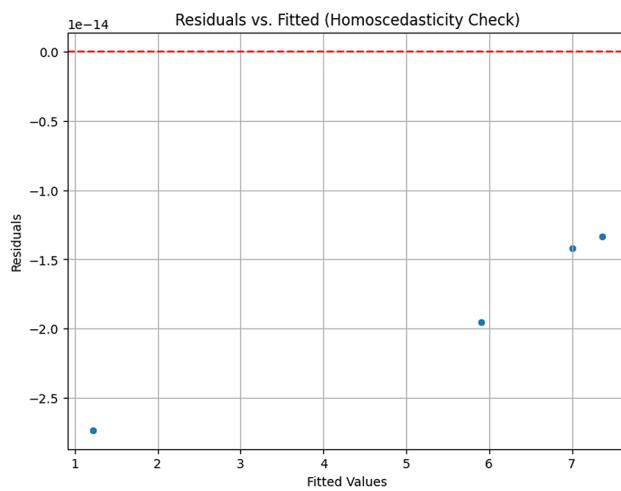


Fig. 10. Residuals vs. fitted values plot for homoscedasticity check.

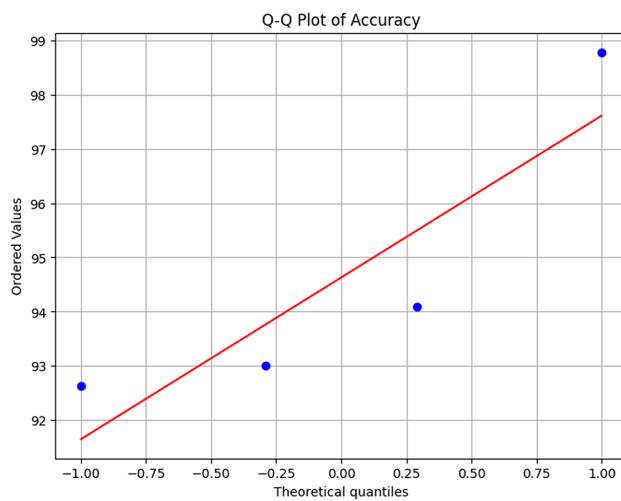


Fig. 11. Q-Q plot of accuracy for normality assessment.

CNN models	Intersection over union (IoU) score				
	6 features	8 features	10 features	12 features	Average IoU score
DenseNet	0.33	0.36	0.39	0.43	0.37
ResNet	0.29	0.31	0.35	0.39	0.33
EfficientNetB0	0.31	0.37	0.41	0.43	0.38
CNN + DA	0.35	0.40	0.43	0.47	0.41

Table 13. IoU score of the selected pretrained models for Benign case.

CNN models	Intersection over union (IoU) score				
	6 features	8 features	10 features	12 features	Average IoU score
DenseNet	0.28	0.32	0.35	0.38	0.33
ResNet	0.25	0.28	0.31	0.34	0.29
EfficientNetB0	0.27	0.30	0.33	0.36	0.31
CNN + DA	0.33	0.37	0.40	0.44	0.38

Table 14. IoU score of the selected pretrained models for malignant case.

Error analysis on four datasets

Possible errors are also characterized by misclassifications in terms of error rate, which gives a clear understanding of the weak sides of the model. Others with relatively higher error rates in our experiments (e.g., 7.36% for EfficientNetB0 on the IQ-OTH/NCCD dataset) have higher misclassification rates, especially for those images that are just slightly different from one another. This means that with similar appearances, their differentiation can be a problem for the model, for example in the case of early stage diseases or noisy data. However, in the CNN + DA proposed with the lowest error rate to 1.22%, the model still occasionally misclassifies subjects. This is common, particularly in individual-specific scenarios that the identification algorithm may not recognize well enough. Also, different sets of error rates are observed, which is true with EfficientNetB0 model having slightly low error rate at 3.77% when tested on the NLST dataset but high error rate of 8.86% on the LC25000 dataset to suggest that the architecture seems to have difficulty in generalizing across different data distribution or image qualities. High error rates as results from some problems like overlapping of classes, low quality input data or imbalanced data where minority classes are misclassified. This is has many advantages like enhanced accuracy of the diagnosis in various conditions, increased ability to generalize over various situations and scenarios and hence, increased patient care by minimizing errors. Specific characteristics, for instance improving the quality of data input, controlling the imbalance in the number of instances in classes and modifying model structures can help reduce errors to acceptable levels to make a model useful in practical applications.

Ethical and practical implications

The use of CNN with Differential Augmentation (DA) model for the detection of lung cancer comes with some certain unethical and practical issues. Data privacy and security is an important factor to consider since such data needs to be shared, encrypted etc. To reduce bias and unfairness to patients, bias from multiple sets of data is prevented; Mitigating bias through diverse datasets promotes fairness, while explainable AI enhances transparency for clinicians. Technical considerations are checking the actual performance of the proposed model in different real-world datasets and handling the differences in imaging to have general applications. seamless integration into clinical workflows with user-friendly interfaces is crucial for adoption. These steps acts as crucial so that the integrated CNN + DA model can be transformed into a reliable and accurate diagnosis tool which would assist in early diagnosis and better treatment, all the while, trying to prevent ethical and practical issues in its implementation.

Results for XAI using LIME

The evaluation of Convolutional Neural Network (CNN) models using the Intersection over Union (IoU) metric is presented in two contexts: benign and malignant cases. Table 13 (for the benign case) and Table 14 (for the malignant case) compare the performance of four models: DenseNet, ResNet, EfficientNetB0, and CNN + DA (CNN with Data Augmentation). In Table 13, CNN + DA achieves the highest average IoU score of 0.41, indicating superior performance in identifying relevant features across different feature sets (6, 8, 10, and 12 features). EfficientNetB0 follows with an average IoU of 0.38, while DenseNet and ResNet score 0.37 and 0.33, respectively. Similarly, in Table 14 (malignant case), CNN + DA again leads with an average IoU of 0.38, demonstrating consistent feature identification. EfficientNetB0 and DenseNet follow with average IoU scores of 0.31 and 0.33, respectively, while ResNet trails with 0.29. Figures 12 and 13 likely illustrate the architectures or efficiency of these models, visually complementing the tabular data. The results highlight CNN + DA as the most effective model in both benign and malignant cases, followed by EfficientNetB0 and DenseNet, due to their ability to consistently align predictions with ground truth across varying feature sets.

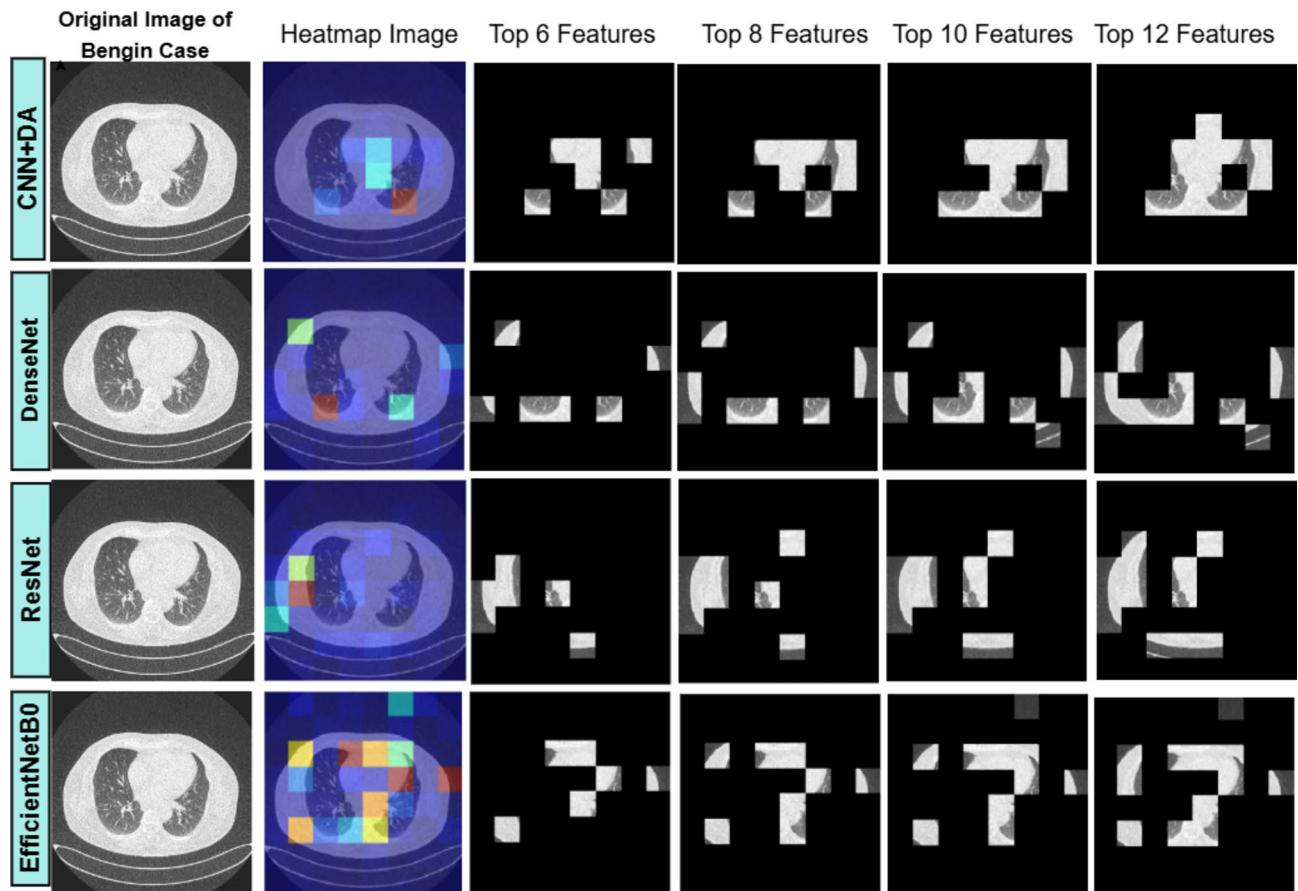


Fig. 12. Visualization of lung image analysis: original image of benign case, LIME heatmaps, and feature-based masking for 6, 8, 10, and 12 important features.

Performance comparison of existing works with proposed work

The proposed CNN with Differential Augmentation (CNN + DA) model achieves a groundbreaking accuracy of 98.78%, outperforming existing state-of-the-art approaches in lung cancer diagnosis. Prior works have explored various deep learning models and augmentation techniques, yet gaps in addressing memory overfitting and achieving robust generalization remain. Critically, a study by Islam et al. (2024)¹⁶ developed a combination of GANs and VAEs with 94% accuracy and Saha et al. (2024)¹⁷ used the VER-Net method, a transfer learning-based method, with 91% accuracy.

Traditional Convolutional Neural Networks (CNNs) by Rainio and Klén (2024)¹⁸ and Kukreja and Sabharwal (2024)¹⁹ demonstrated accuracies of 92.6% and 96.11%, respectively, with no explicit focus on mitigating overfitting.

Advanced ensemble and hybrid models have shown promise, with Quasar et al. (2023)²⁰ achieving 96.34% accuracy using a combination of BEiT, DenseNet, and Sequential CNN, while Raza et al. (2023)²¹ and Zhang et al. (2024)²² reported accuracies of 96.10% and 96% using Lung-EffNet and DenseNet-CNN Integration, respectively. Similarly, Gautam et al. (2023)²³ utilized an ensemble of ResNet-152, DenseNet-169, and EfficientNet-B7, achieving 97.23%, and Dritsas and Trigka (2022)²⁴ employed Rotation Forest for a comparable accuracy of 97.1%. However, these models either relied on complex architectures or lacked tailored strategies to prevent memory overfitting.

In contrast, the CNN + DA model integrates Differential Augmentation to introduce diverse transformations, such as hue, brightness, and contrast adjustments, directly addressing overfitting and enhancing generalization. Compared to other works, the CNN + DA model exhibits superior accuracy, proving its reliability and robustness for clinical applications in lung cancer detection. Its innovative approach bridges the gaps left by prior models, solidifying its contribution as a novel and highly effective diagnostic tool. Table 15 compares the performance of existing methods with the proposed approach, highlighting improvements in accuracy, efficiency, or other key metrics.

Limitations

The primary dataset used (IQ-OTH/NCCD), consisting of only 1,097 images, is relatively small for training deep learning models, which limits the model's ability to generalize across a broader spectrum of lung cancer variations and patient demographics. This dataset's class distribution is also imbalanced, with more malignant cases than benign or normal ones, potentially biasing the model toward the majority class. While the CNN + DA model

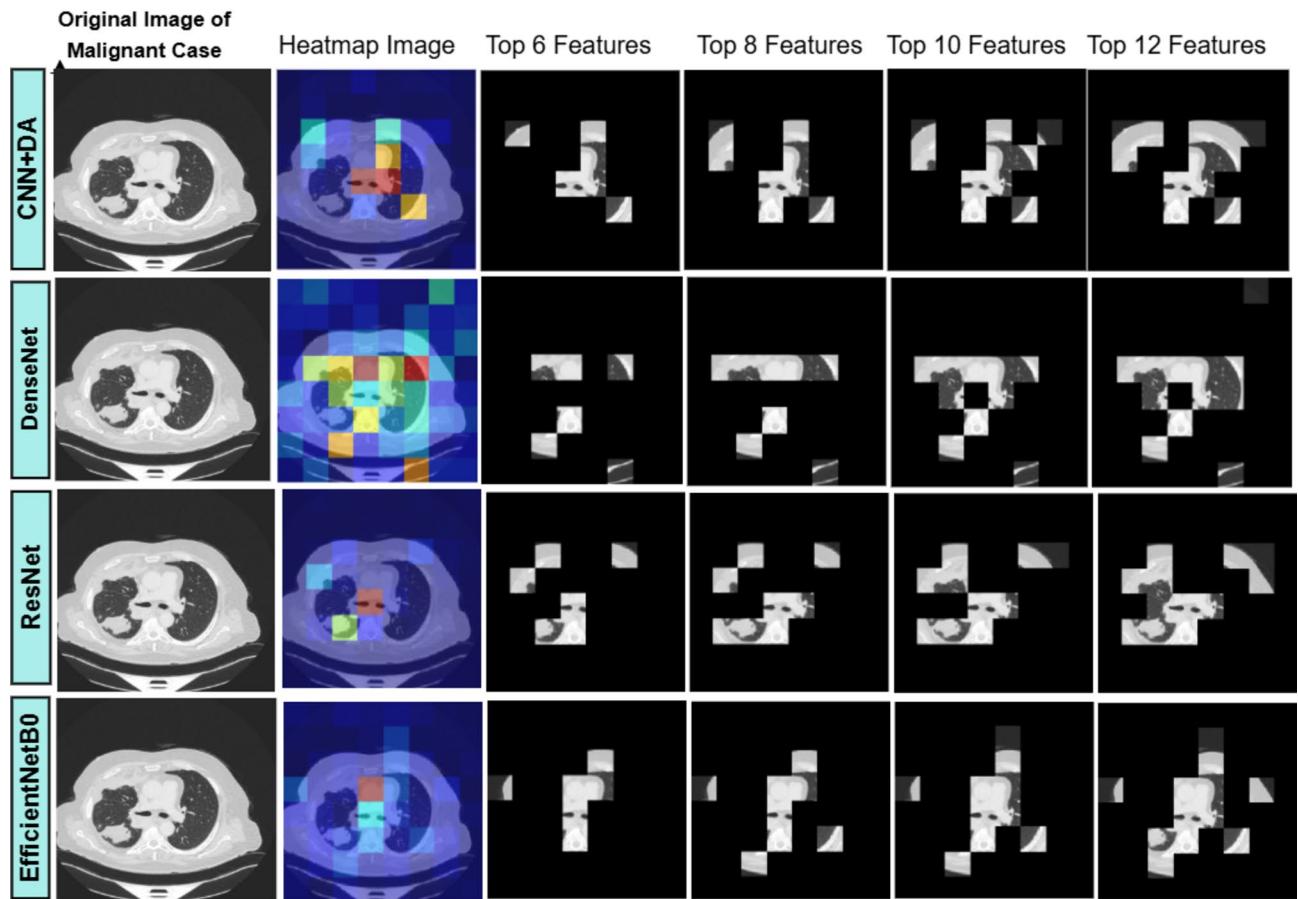


Fig. 13. Visualization of Lung Image Analysis: Original image of Malignant case, LIME Heatmaps, and Feature-Based Masking for 6, 8, 10, and 12 Important Features.

Author(s)	Proposed model	Accuracy
Islam et al. (2024) ¹⁶	Generative adversarial networks (GANs) and Variational Autoencoders (VAEs)	94%
Saha et al. (2024) ¹⁷	VER-Net	91%
Rainio and Klén (2024) ¹⁸	Convolutional Neural Network (CNN)	92.6%
Kukreja and Sabharwal (2024) ¹⁹	Convolutional Neural Network (CNN)	96.11%
Zhang et al. (2024) ²⁰	DenseNet-CNN Integration	96%
Gai et al. (2023) ²¹	Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs)	93.4%
Quasar et al. (2023) ²²	Ensemble Model (BEiT, DenseNet, Sequential CNN with ensemble methods)	96.34%
Raza et al. (2023) ²³	Lung-EffNet (EfficientNet with modified top layers)	96.10%
Gautam et al. (2023) ²⁴	Ensemble (ResNet-152, DenseNet-169, EfficientNet-B7 with weight optimization)	97.23%
Dritsas and Trigka (2022) ²⁵	Rotation Forest	97.1%
Tsou et al. (2021) ²⁶	eXtreme Gradient Boosting (XGBoost)	92%
Our Work	CNN with DA	98.78%

Table 15. Performance comparison of existing works with proposed work.

performs well on the controlled dataset, its performance in real-world clinical environments, which involve diverse imaging protocols, patient demographics, and potential artifacts, has yet to be validated. The model might struggle to handle unseen variations in imaging conditions common in clinical settings. Additionally, the integration of Differential Augmentation (DA) during training introduces extra computational overhead, increasing both training time and resource requirements, which could be a challenge in resource-constrained environments. Finally, the model's performance is highly dependent on the quality of input data; noisy or low-quality images could lead to misclassifications, particularly when distinguishing between benign and malignant tumors.

Conclusion

Lung cancer remains a significant global health challenge, requiring innovative solutions to improve diagnostic accuracy. This study addresses the persistent problem of memory overfitting in deep learning models, which impairs their ability to generalize across diverse datasets. As an effort to improve the quality of lung cancer detection models, this research seeks to design a new method that involves the merge of a Differential Augmentation technique with a Convolutional Neural Network.

The effectiveness of the proposed CNN + DA model has been demonstrated, attaining the prediction accuracy of 98.78% that promoted the model to the highest level of various existing methods, such as DenseNet, ResNet, and EfficientNetB0. By leveraging targeted augmentation techniques such as hue, brightness, and contrast adjustments, the model effectively reduces overfitting, improving its reliability in diverse imaging scenarios. Statistical analyses, including Tukey's HSD post-hoc tests, confirm the model's significance in achieving higher accuracy and generalization.

As shown in this research, incorporating Differential Augmentation with CNNs is a good solution to the problems existing in traditional models. Thus, the conclusions point out the need of developing models that do not only solve the problems precisely, but also are immune to adversarial perturbations and can be easily scaled to tackle a wider range of problems. Future work can further explore the adaptability of this approach to other domains and datasets, expanding its utility in advancing deep learning research for complex diagnostic tasks.

Data availability

The data used to support the findings of created new data set, this study is available from the corresponding author upon request.

Received: 24 September 2024; Accepted: 14 April 2025

Published online: 05 May 2025

References

1. Jaksik, R., Szumała, K., Dinh, K. N. & Śmieja, J. Multiomics-Based feature extraction and selection for the prediction of lung Cancer survival. *Int. J. Mol. Sci.* **25** (7 March 25), 3661. <https://doi.org/10.3390/ijms25073661> (2024).
2. Lasrado, S. A. & Babu, G. N. K. S. Fused feature reduction and selection system for early lung Cancer detection (April 4). <https://doi.org/10.1109/ICONSTEM60960.2024.10568801> (2024).
3. Aggarwal, P., Marwah, N., Kaur, R. & Mittal, A. Lung Cancer Survival Prognosis Using a Two-Stage Modeling Approach. *Multimedia Tools and Applications* (January 31). <https://doi.org/10.1007/s11042-024-18280-2> (2024).
4. Hussain, L., Almaraashi, M. S., Aziz, W., Habib, N. & Abbasi, S. U. R. S. Machine learning-based lungs cancer detection using reconstruction independent component analysis and sparse filter features. *Waves Random Complex. Media* **34**, 226–251. <https://doi.org/10.1080/17455030.2021.1905912> (2021).
5. Mamun, M., Mahmud, M. I. & Meherin, M. & Abdelgawad, A. LCDctCNN: lung Cancer diagnosis of CT scan images using CNN based model. ArXiv (Cornell University). (2023). <https://doi.org/10.48550/arxiv.2304.04814>
6. Moitra, D. & Mandal, R. K. Classification of non-small cell lung cancer using one-dimensional convolutional neural network. *Expert Syst. Appl.* **159**, 113564 (2020).
7. Perez, G. & Arbelaez, P. *Automated Lung cancer Diagnosis Using three-dimensional Convolutional Neural Networks*, 1803–1815 (Medical & Biological Engineering & Computing, 2020).
8. Moitra, D. & Mandal, R. K. Automated AJCC staging of non-small cell lung cancer (NSCLC) using deep convolutional neural network (CNN) and recurrent neural network (RNN). *Health Inform. Sci. Syst.* **7**, 1–12 (2019).
9. Doppalapudi, S., Qiu, R. G. & Badr, Y. Lung cancer survival period prediction and understanding: deep learning approaches. *Int. J. Med. Informatics.* **148**, 104371 (2021).
10. Faruqui, N. et al. LungNet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data. *Comput. Biol. Med.* **139**, 104961. (2021).
11. Aboamer, M. A. et al. An investigation in analyzing the food quality well-being for lung cancer using blockchain through cnn. *J. Food Qual.* (2022).
12. Kirienko, M. et al. Convolutional neural networks promising in lung cancer T-parameter assessment on baseline FDG-PET/CT. *Contrast Media Mol. Imaging* (2018).
13. Lin, C. J., Jeng, S. Y. & Chen, M. K. Using 2D CNN with Taguchi parametric optimization for lung cancer recognition from CT images. *Appl. Sci.* **10** (7), 2591 (2020).
14. Moitra, D. & Mandal, R. K. Prediction of non-small cell lung cancer histology by a deep ensemble of convolutional and bidirectional recurrent neural network. *J. Digit. Imaging.* **33**, 895–902 (2020).
15. Zhan, X. et al. A convolutional neural network-based intelligent medical system with sensors for assistive diagnosis and decision-making in non-small cell lung cancer. *Sensors*, 21(23), p.7996. (2021).
16. Islam, Tauhidul, M. S., Hafiz, J. R., Jim & Mridha, M. F. Md. Mohsin Kabir, and A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions. *Healthcare Analytics* 5June 1, : 100340. (2024). <https://doi.org/10.1016/j.health.2024.100340>
17. Saha, A. et al. VER-Net: A hybrid transfer learning model for lung cancer detection using CT scan images. *BMC Med. Imaging* **24**. <https://doi.org/10.1186/s12880-024-01238-z> (2024).
18. Rainio, O. & Klén, R. Comparison of simple augmentation transformations for a convolutional neural network classifying medical images. *Signal Image Video Process.* <https://doi.org/10.1007/s11760-024-02998-5> (2024).
19. Kukreja, S. & Sabharwal, M. An effective lung Cancer diagnosis model using the CNN algorithm. *EAI Endorsed Trans. Pervasive Health Technol.* **10** <https://doi.org/10.4108/eptpt.10.6805> (2024).
20. Zhang, C. et al. Enhancing lung cancer diagnosis with data fusion and mobile edge computing using DenseNet and CNN. *J. Cloud Comput. Adv. Syst. Appl.* **13**, <https://doi.org/10.1186/s13677-024-00597-w> (2024).
21. Gai, L., Xing, M., Chen, W., Zhang, Y. & Qiao, X. Comparing CNN-based and transformer-based models for identifying lung cancer: which is more effective? *Multim. Tools Appl.* **83**, 59253–59269. <https://doi.org/10.1007/s11042-023-17644-4> (2023).
22. Quasar, S. et al. Ensemble methods for computed tomography scan images to improve lung cancer detection and classification. *Multimedia Tools Appl.* **83** (17), 52867–52897. <https://doi.org/10.1007/s11042-023-17616-8> (2023).
23. Raza, R. et al. Lung-EffNet: lung cancer classification using EfficientNet from CT-scan images. *Eng. Appl. Artif. Intell.* **126** (1), 106902. <https://doi.org/10.1016/j.engappai.2023.106902> (2023).
24. Gautam, N., Basu, A. & Sarkar, R. Lung cancer detection from thoracic CT scans using an ensemble of deep learning models. *Neural Comput. Appl.* **36**, 2459–2477. <https://doi.org/10.1007/s00521-023-09130-7> (2023).

25. Dritsas, E. & Trigka, M. Lung cancer risk prediction with machine learning models. *Big Data Cogn. Comput.* **6** (4), 139. (2022).
26. Tsou, P. H. et al. Exploring volatile organic compounds in breath for high-accuracy prediction of lung cancer. *Cancers* **13** (6), 1431 (2021).
27. Muniasamy, A., Alquhtani, S. A. S., Bilfaqih, S. M., Balaji, P. & Karunakaran, G. Lung cancer histopathology image classification using transfer learning with Convolution neural network model. *Technol. Health Care November.* **9**, 1–12. <https://doi.org/10.3233/THC-231029> (2023).
28. Jothi, J., Sathiya & Chinnadurai, M. A combined deep CNN with a chimp optimization approach for lung cancer diagnosis. *J. Intell. Fuzzy Syst.* **46** (2), 4681–4696. <https://doi.org/10.3233/JIFS-237339> (2024).
29. Garcea, F., Serra, A., Lamberti, F. & Morra, L. Data augmentation for medical imaging: A systematic literature review. *Comput. Biol. Med.* **152** (January 1), 106391. <https://doi.org/10.1016/j.combiomed.2022.106391> (2023).
30. Mathworks Understand Network Predictions Using LIME - MATLAB & Simulink - MathWorks India—in.mathworks.com. (2024). <https://in.mathworks.com/help/deeplearning/ug/understand-network-predictions-using-lime.html>. Accessed 28 May 2024.
31. Kumar, S. V. & Kondaveeti, H. K. Towards transparency in AI: explainable bird species image classification for ecological research. *Ecol. Ind.* **169**, 112886. <https://doi.org/10.1016/j.ecolind.2024.112886> (2024).
32. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should i trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. (2016).
33. Zafar, M. R. & Khan, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Mach. Learn. Knowl. Extr.* **3** (3), 525–541 (2021).
34. Hamdalla, F. The & IQ-OTH/NCCD Lung Cancer Dataset. Kaggle. (2020). <https://www.kaggle.com/datasets/hamdallak/the-iqoth-nccc-lung-cancer-dataset>. Accessed 3 Jan 2023.
35. Lung and Colon Cancer Histopathological Images. <https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images>. Accessed 7 Dec 2024.
36. Clark, K. et al. The Cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging.* **26**, 1045–1057. <https://doi.org/10.1007/s10278-013-9622-7> (2013).
37. Team, N. L. S. T. R. The National lung screening trial: overview and study design. *Radiology* **258** (1), 243–253. <https://doi.org/10.1148/radiol.10091808> (2011).
38. Tirumandham, N. Improving predictive performance in e-learning through hybrid 2-tier feature selection and hyper parameter-optimized 3-tier ensemble modeling. *Int. J. Inf. Technol.* **16**, 5429–5456. <https://doi.org/10.1007/s41870-024-02038-y> (2024).

Acknowledgements

Authors are saying to thanks for supporting of completion of this research work.

Author contributions

All authors reviewed the manuscript.

Funding

Author declared that no funding was received for this Research and Publication.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participant and Animals performed by author.

Additional information

Correspondence and requests for materials should be addressed to V.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025