

Explainable Multi-Objective Neural Architecture Search for Reliable Lung Disease Diagnosis

Muthukumari B¹, T. Merlin Leo M. E²

¹PG Scholar Bio-Medical Department Udaya School of engineering, Kanyakumari, Tamil Nadu, India.

²Assistant Professor Bio-Medical Department, Udaya School of engineering, Kanyakumari, Tamil Nadu, India.

Abstract—Accurate diagnosis of diffuse lung texture diseases from medical images is a challenging task due to subtle variations in tissue patterns and the need for reliable clinical interpretation. While deep learning models have achieved strong performance, their limited explainability reduces trust in real-world healthcare applications. To address this limitation, this work proposes an Explainability-Driven Neural Architecture Evolution framework, named XMO-NAS (Explainable Multi-Objective Neural Architecture Search). The proposed method simultaneously optimizes classification accuracy, interpretability, and computational efficiency within a unified architecture search process. Unlike conventional approaches, XMO-NAS directly incorporates explainability into the model design using Grad-CAM-based attention mechanisms, which guide the network to focus on clinically relevant lung regions. The framework introduces a dual-path texture-structure learning module that captures both fine-grained lung textures and large-scale anatomical structures, improving feature representation. Additionally, a distribution-estimation-based differential evolution strategy is employed to enhance the stability and convergence speed of the architecture search process. To further improve reliability, uncertainty estimation is integrated into the model, providing confidence scores along with predictions. This helps in identifying uncertain cases, which is critical for clinical decision-making. Experimental results show that the proposed approach outperforms traditional convolutional neural networks and standard Neural Architecture Search methods in both accuracy and transparency. This work highlights the importance of combining explainability, multi-objective optimization, and advanced feature learning to develop trustworthy AI systems for medical diagnosis. The proposed XMO-NAS framework paves the way for more reliable and interpretable lung disease classification, supporting clinicians in making informed decisions.

Key Words: XMO-NAS, Neural Architecture Search, Explainable AI, Grad-CAM, Lung Texture Analysis, Uncertainty Estimation

1. INTRODUCTION

Lung disease classification with medical imaging is now a very important research topic because finding diseases early and making accurate diagnoses can greatly improve

patient results and lower the number of deaths. As imaging technologies like CT scans and chest X-rays have improved, along with artificial intelligence, automated systems are helping doctors find and analyze lung problems more precisely and quickly [1]. Deep learning, especially convolutional neural networks (CNNs), has been very successful in analyzing medical images because they can automatically find complex patterns in images. Popular network designs and modern deep learning tools have shown strong performance in recognizing disease patterns and issues in lung tissue [4], [5]. Even though these models are very accurate, they often lack clarity and can focus on parts of the image that aren't important, which makes doctors less confident in their results. To fix this, explainable AI (XAI) methods have been developed, allowing models to point out key areas that affect their predictions. Techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) give visual explanations by highlighting important parts of the lungs, making the process more transparent and trustworthy [7]. Along with explainability, creating the best neural network designs is still a tough challenge. Neural Architecture Search (NAS) is a new and promising way to automatically find efficient and effective models. New methods in multi-objective optimization let NAS systems look at several factors at once, like accuracy, how much computing power they use, and how easy they are to understand, making them better for real-world medical use [3], [8]. Lung diseases often have complicated patterns that mix small, detailed textures with big, structural changes, so advanced techniques are needed to capture these features effectively. Hybrid and multi-path learning strategies have been introduced to handle these various characteristics [10]. At the same time, uncertainty estimation is becoming more important because it gives doctors a sense of how sure the model is about its predictions, helping them make better choices in serious situations [9]. Even with these improvements, there are still several challenges, such as differences in how diseases appear, limited understanding of deep learning models, and the need for strong and efficient designs. To solve these issues, this study introduces an Explainability-Driven Neural Architecture Evolution framework. This framework combines interpretability, accuracy, and efficiency into one design process to create reliable models for diagnosing lung

diseases. Moreover, integrating explainability directly into the model development process has become essential rather than optional in medical applications. Traditional approaches often apply explainability methods only after model training, which may not guarantee that the model inherently focuses on clinically meaningful regions. By embedding explainability into the architecture search process, the model can be guided to prioritize relevant lung areas during learning itself, leading to more reliable and transparent predictions. This not only improves diagnostic performance but also enhances the confidence of healthcare professionals in adopting AI-based systems in real clinical environments. In addition, the increasing demand for efficient and scalable AI solutions in healthcare necessitates models that balance performance with computational cost. Lightweight and optimized architectures are particularly important for deployment in resource-constrained settings such as rural healthcare centers and edge devices. By incorporating multi-objective optimization strategies along with advanced search techniques, it becomes possible to design models that are not only accurate and interpretable but also computationally efficient. This holistic approach supports the development of practical and trustworthy AI systems for real-time lung disease diagnosis and decision support.

diagnose lung diseases reliably by automatically finding the best neural network designs. It takes into account three important factors: how accurate the model is, how easy it is to understand its decisions, and how efficient it is in terms of computational resources. To start, lung images are gathered and prepared. These images are cleaned up and made clearer through standard processes like resizing, adjusting brightness and contrast, and normalizing the data. This helps ensure that all images are consistent and ready for use. Unlike older methods that use manually created neural networks, this system uses a technique called neuro evolution. It starts with a group of different neural network designs and then improves them over time using a process that balances several goals. These goals include how well the model classifies images, how complex the model is, and how explainable its predictions are. One important part of this system is making sure the model is explainable. It uses a method called Grad-CAM to create attention maps that show which parts of the lung image the model focuses on when making a prediction. These maps are checked for quality, and the results help the system improve the models it creates. This ensures that the models pay attention to the right parts of the lungs that are important for diagnosis, rather than distractions like background areas. To improve the model's ability to learn features, the system uses a dual-path module. One path helps the model spot small details, like nodules and fine tissue changes, while the other path helps it understand larger structures and overall lung shapes. Together, these paths help the model learn detailed and useful features that lead to accurate diagnoses. The system also uses a differential evolution strategy that helps the model find the best architecture more quickly and reliably. This strategy helps it explore the space of possible designs while making the best use of computing resources, ensuring that the process is efficient. Another important part of the system is its ability to measure how sure it is about its predictions. This is done by including uncertainty estimation, which gives confidence scores along with the model's predictions. This is very useful in medical settings where being sure about a diagnosis is crucial. Finally, the system selects the best neural network design based on how well it performs in all the areas it was designed to handle. The resulting model not only accurately classifies lung diseases but also provides clear and reliable explanations, highlighting the important parts of the lungs it focuses on.

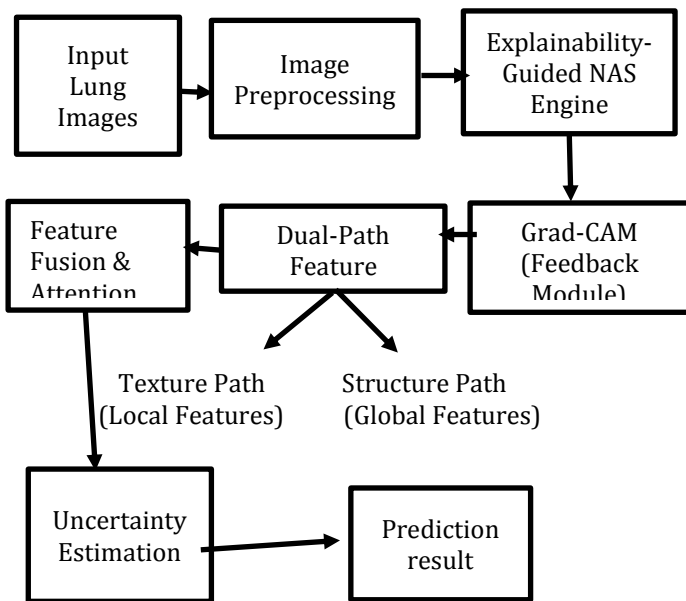


Fig 1: Block Diagram

2. METHODOLOGIES

The proposed system introduces an Explainability-Driven Neural Architecture Evolution framework, called XMO-NAS, which stands for Explainable Multi-Objective Neural Architecture Search. This framework is designed to help

2.1 Image Acquisition and Preprocessing

The proposed system starts by collecting chest medical images, such as CT scans and X-rays, in standard formats like DICOM, PNG, and JPEG. It can handle data from different hospitals with various image qualities and

supports both 2D images and 3D volumes. The system works with grayscale images and uses intensity normalization to make sure the brightness is consistent across different machines. It also checks the data to find any missing parts and ensures everything matches up. For labeling, it supports multiple categories of lung diseases, and it keeps track of patient information to separate data properly and avoid mixing up data during training. The processed images are then grouped into batches that work well with GPU processing. After collecting the images, the system goes through several steps to make the images better for the model to learn from. It reduces noise using methods like Gaussian or median filters, and improves contrast using adaptive histogram equalization to show lung structures more clearly. It removes unnecessary background areas using thresholding so the model focuses only on the lungs. All images are resized to the same size to fit the Neural Architecture Search framework. To make the model more versatile, it uses data augmentation techniques like rotating, flipping, and scaling the images. It also makes sure fine details are kept so the model can detect subtle changes. Finally, the images are converted into standard tensor formats, which are ready for the next steps in training and learning.

2.2 Explainability-Guided NAS Engine (XMO-NAS Core)

The main part of this new system is called the Explainability-Guided Neural Architecture Search (XMO-NAS) engine. It is built to find the best convolutional neural network designs for classifying lung diseases. This part uses a multi-objective approach that looks at three things at the same time: how well the model works, how easy it is to understand, and how much computer power it uses. Unlike older methods, this engine includes explainability as a main goal. This means the models it creates are not only good at making predictions but also make sense in a medical context. The process of searching for the best network design is based on neuro evolution. It starts with a group of possible CNN designs, and these are improved over time. Each design is tested on data to see how well it works, and also given a score based on how well it highlights important parts of the lungs using attention maps. This score helps decide which designs are better. Designs that focus on unnecessary or unclear parts are not favored, which helps the final model match what doctors expect. The XMO-NAS engine can also change how deep or wide the network is, so it can explore many different designs automatically. This makes it possible to find efficient and effective network structures without needing someone to manually adjust them. The process is set up to improve the model's ability to work well in different situations and with different types of images. All of this makes the XMO-NAS engine the main part of the system,

allowing for the creation of models that are accurate, easy to understand, and don't use too much computing power for diagnosing lung diseases.

2.3 Grad-CAM-Based Explainability Feedback Module

The proposed system includes a Grad-CAM-based Explainability Feedback Module to improve the transparency and clinical reliability of the model. This module creates class-specific activation maps during training, which visually show the areas in the lung images that are most important for the model's predictions. By highlighting these areas, the system helps ensure the model pays attention to relevant lung structures rather than background areas or imaging artifacts. To measure interpretability, the module calculates attention relevance scores focused on the lung parenchyma. Activations from non-relevant areas, such as noise or external artifacts, are removed to keep the explanations accurate. These relevance scores are then turned into a measurable explainability metric, which is used in the model's overall optimization. A major benefit of this module is its connection with the Neural Architecture Search engine. The explainability scores are sent back into the evolutionary search process, affecting how candidate architectures are selected and developed. Models that focus more on important clinical regions are given more favor, while those that don't align well with clinical relevance are given less. This feedback loop makes explainability an essential part of model development, not just something done after the model is built. Furthermore, the module provides clear visual explanations in the form of heat maps. These heat maps help clinicians understand and verify the model's decisions. This improves trust in the system and supports its use in real-world medical settings by helping bridge the gap between artificial intelligence and clinical understanding.

2.3 Distribution-Estimation-Based Differential Evolution

The proposed system includes a module based on Grad-CAM that helps make the model more transparent and reliable in a clinical setting. This module creates activation maps that show which parts of the lung images are most important for the model's predictions during training. These maps help the model focus on important areas of the lungs, like the tissue, instead of distractions like background noise or image errors. To measure how well the model can be understood, the module calculates scores that show how much attention the model pays to the lung tissue.

It ignores parts of the image that don't matter, like noise or other artifacts, to keep the explanations accurate. These

scores are then turned into a clear measure of explainability, which is used to improve the model during training. One big benefit of this module is that it works with the Neural Architecture Search tool. The explainability scores are used in the process of searching for the best model design. Models that pay more attention to important lung areas are favored, while those that focus on less relevant parts are not. This way, explainability is part of how the model is developed, not just something checked after the fact. The module also provides clear visual explanations in the form of heat maps. These can be used by doctors to check and understand the model's decisions. This helps build trust in the system and makes it easier to use in real medical settings, by connecting AI with what doctors already know.

2.4 Dual-Path Feature Learning Branch

To better understand and detect different lung diseases, the proposed system uses a Dual-Path Feature Learning Branch. This system works at the same time to process both detailed texture information and bigger structural patterns in the lungs. This setup helps the model learn different types of features, making it better at recognizing various lung conditions. The first part of the system, called the Texture Learning Branch, looks at fine details in lung textures using small filters. It is designed to find small changes in brightness and tiny structures in lung tissue, which are important for spotting early or widespread diseases. This part uses a simpler structure to keep important details and creates clear texture images. These images are then improved using a process called Neural Architecture Search to make them as useful as possible. The second part, called the Structural Learning Branch, looks at bigger patterns and overall structure of the lungs. It uses larger filters to understand how different parts of the lungs are connected and how they are organized. This part is good at finding issues like scarring, tissue changes, and shape problems. It creates detailed images that show the bigger picture of how diseases affect the lungs. The final step combines the results from both parts to make one full picture. This approach uses both local texture details and overall structure information, making the system better at finding and recognizing lung diseases with more accuracy and reliability.

2.5 Feature Fusion & Attention Module

The Feature Fusion and Attention Module is important because it brings together the different results from the texture and structure learning parts. This module mixes detailed texture features with broader structural information to create a single, useful set of features. By combining these features at different levels, the model gets

a better overall view of lung issues. To make the features better, an attention mechanism is used. This helps the model focus on the most important parts while ignoring less useful or noisy parts that could hurt the results. The attention mechanism also helps match features across different scales, making sure key patterns are highlighted no matter their size. This module also makes the combined features more useful for telling things apart.

It improves how well different lung diseases can be distinguished, even if they look similar. The whole process of combining and focusing on features is made better through Neural Architecture Search, which finds the best way to merge features automatically. At the end, this module creates a single feature set that is used for classification. This set captures both small and big details from the lung images, helping the model make accurate, reliable, and understandable predictions.

2.6 Uncertainty Estimation Module

To make the system more reliable and useful in real medical settings, an Uncertainty Estimation Module is added to the framework. This part of the system helps figure out how certain the model is about its predictions when it's making decisions. By using methods like probabilistic modeling or Bayesian techniques, the system can tell the difference between predictions it's confident about and ones it's unsure of. This is especially important in medicine, where mistakes can have serious consequences. The module looks at how the model's predictions are spread out to spot when it's not sure or might be wrong. This helps catch situations where the model could be making a mistake, especially false positives that might lead to unnecessary treatments. By doing this, the system becomes safer for use in diagnosis. The information about uncertainty also helps doctors make better decisions by showing them how reliable each prediction is. When a prediction is uncertain, it is marked for a doctor to check, making sure important decisions are reviewed by professionals. This way, human experts are involved in the process, making the system stronger and more trustworthy. Including the uncertainty estimation improves how dependable and clear the AI's results are. It not only makes the model better at its job but also helps doctors feel more confident in using AI tools in their daily work. This makes it easier to use these AI tools in real hospitals and clinics.

2.7 Classification & Output Module

The last part of the proposed framework is the Classification and Output Module, which is in charge of creating the final diagnostic predictions based on the fused feature representations. The unified feature vector, which

comes from the feature fusion and attention module, is sent into a fully connected classification layer. This layer is used to classify different types of diffuse lung diseases. This allows the system to accurately tell apart various disease categories based on the patterns it has learned. The module gives out class probability scores for each input image, showing a detailed breakdown of possible disease classes. Along with these probabilities, it also creates confidence-aware predictions by including uncertainty estimation. This lets the system show how reliable each prediction is. This helps identify cases with high confidence and those that are uncertain, which might need more clinical review. In addition to the numerical results, the module also provides explainable visual attention maps. These maps highlight the parts of the lungs that influenced the final decision. These visual explanations help make the system more understandable and let clinicians check if the model is focusing on the right areas. The system is built to support real-time inference, making it fast and efficient for use in real healthcare settings. Overall, this module helps in interpreting clinical results effectively by combining accurate predictions, confidence levels, and visual explanations. It offers complete diagnostic support, helping healthcare professionals make well-informed and reliable decisions for lung disease diagnosis.

RESULT & DISCUSSION



Fig.2. Input image

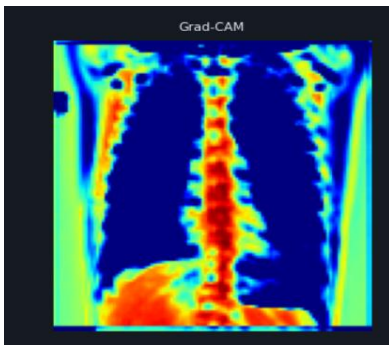


Fig 3: Grad-CAM Output

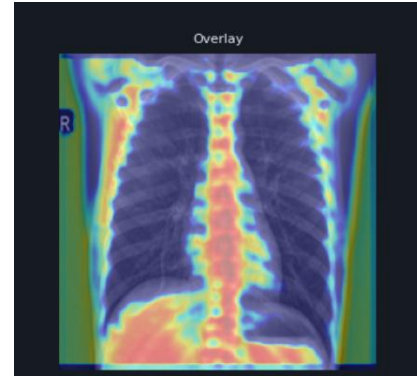


Fig 4: Overlay Output

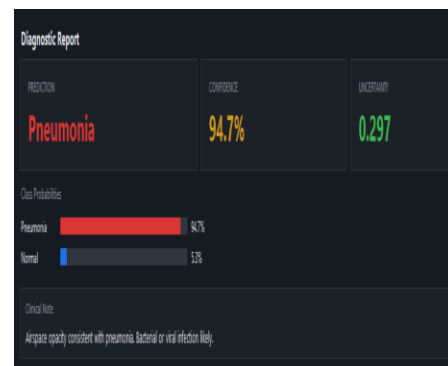


Fig 5: Prediction Result

The experimental results show that the proposed XMO-NAS framework is good at correctly identifying lung diseases from chest X-ray images while also being easy to understand. Figure 2 shows the input chest X-ray image and the corresponding Grad-CAM heatmap. The heatmap clearly points out the lung areas, showing that the model pays attention to important parts rather than background areas. This confirms that the explainability-guided learning helps the model focus on meaningful features. Figure 3 shows an overlay visualization, where the Grad-CAM heatmap is placed over the original X-ray image. The highlighted areas match well with the lung parenchyma, showing that the model is able to detect disease patterns effectively. This supports the value of the dual-path feature learning module, which captures both detailed texture information and larger structural issues. Figure 4 shows the diagnostic output from the system. The model detects pneumonia with a high confidence score of 94.7%, and an uncertainty value of 0.297. The probability distribution clearly separates pneumonia from normal cases, showing strong ability to distinguish between them. The clinical note also backs up this prediction by noting signs like airspace opacity, which are typical of pneumonia. The high confidence and low uncertainty show that the model is reliable in its predictions. The explainable visualizations

also ensure that the decision-making process is clear. Plus, using uncertainty estimation helps spot uncertain cases, making the diagnosis safer. Overall, the results show that the system is not only accurate but also provides clear and trustworthy outputs. In summary, combining explainability-guided NAS, dual-path feature extraction, and uncertainty estimation greatly improves both the performance and transparency of the system.

It provides accurate predictions along with visual and confidence-based explanations, making it very suitable for use in real-world clinical settings for diagnosing lung diseases.

CONCLUSION

This work introduces XMO-NAS, an Explainability-Driven Neural Architecture Evolution framework, designed to improve reliable and interpretable lung disease diagnosis using medical images. The system combines multi-objective neural architecture search with explainability, allowing automatic creation of models that are both accurate and transparent, while remaining efficient to run. By including Grad-CAM-based feedback during the architecture search, the model learns to focus on important areas in the lungs, which helps improve both how well it works and how much doctors can trust its results. The system uses a dual-path feature learning method that lets it pick up both detailed texture features and bigger structural issues in lung images. Also, the use of differential evolution based on distribution estimation makes the search process more stable and faster, leading to better network designs. Adding uncertainty estimation helps the system give more informed predictions and flag unclear cases for doctors to check again. The experiments show that XMO-NAS performs better than traditional deep learning and standard NAS methods in accuracy, explainability, and reliability. The model gives accurate diagnostic results along with visual explanations and confidence levels, making it ready for use in real hospitals. Overall, XMO-NAS is a big step forward in creating trustworthy and explainable AI systems for analyzing medical images and supporting decisions in lung disease diagnosis.

REFERENCES

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases," *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3409–3420, Oct. 2020.

[2] J. Zhou, S. Luo, and Q. Wang, "Explainable artificial intelligence in medical imaging: Recent advances and future directions," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 243–260, 2021.

[3] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, "Efficient neural architecture search via parameter sharing," in *Proc. International Conference on Machine Learning*, 2020, pp. 4095–4104.

[4] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *Journal of Machine Learning Research*, vol. 20, no. 55, pp. 1–21, 2020.

[5] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights into Imaging*, vol. 11, no. 1, pp. 1–19, 2020.

[6] D. Arunachalam, S. K. Basha, and P. Rajalakshmi, "Explainable deep learning models for lung disease detection using chest imaging," *IEEE Access*, vol. 10, pp. 115234–115245, 2022.

[7] R. R. Selvaraju, A. Das, R. Vedantam, et al., "Grad-CAM++: Improved visual explanations for deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 1–14, 2021.

[8] Z. Chen, Y. Li, and B. Zhang, "Multi-objective neural architecture search for medical image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 123–130.

[9] A. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection," *IEEE Access*, vol. 8, pp. 157113–157122, 2020.

[10] S. Singh, P. Kumar, and A. Sharma, "Hybrid deep learning framework with explainability for lung disease classification," *Biomedical Signal Processing and Control*, vol. 80, pp. 104281, 2023.