ZANCO Journal of Pure and Applied Sciences

TYPE Original article
DOI http://dx.doi.org/10.21271/ZJPAS.38.2.13

OPEN ACCESS

*Corresponding author
Aiman Abd Saeed*
aiman.abd789@gmail.com

RECEIVED :30 /06 /2025
ACCEPTED :22/09/ 2025
PUBLISHED :30/ 04/ 2026

KEYWORDS:
Deep learning, Convolutional neural networks (CNNs), Transfer learning, Gradient-weighted class activation mapping (Grad-CAM), COVID-19.

Beyond Black-Box AI: A Quantitative Grad-CAM Analysis of Convolutional Neural Network Interpretability in COVID-19 Chest X-Ray Classification

Aiman Abd Saeed* Rasber Dhahir Rashid

Department of Computer Science and Information Technology, College of Science, Salahaddin University-Erbil, Erbil, Kurdistan Region, Iraq.

ABSTRACT

Modern AI models use deep architectures that obscure how predictions are made. Without understanding how models reach their predictions, it becomes difficult to verify reasoning, identify biases, or trust their reliability in high-stakes domains like healthcare. Many COVID-19 chest X-ray (CXR) studies report high accuracy and present qualitative gradient-weighted class activation mapping (Grad-CAM) heatmaps, providing no quantitative evidence of alignment with lung anatomy and relying on manual, subjective inspection. We introduce an automated quantitative pipeline that converts interpretability into objective, anatomy grounded metrics between Grad-CAM heatmaps and lung masks. We evaluate six convolutional neural networks (CNNs): VGG16, VGG19, ResNet-101, NASNet-Mobile, NASNet-Large, and Xception, for both classification performance and anatomical interpretability in COVID-19 CXR detection. Classification accuracies ranged from 90% to 96%, with Xception achieving the highest accuracy (95.90%) and a balanced precision, recall, and F1-score of 95.92%. NASNet-Large and VGG19 followed at 94.87%, with VGG19 reaching the highest precision (98.89%). To assess model transparency, we automated interpretability analysis by thresholding the Grad-CAM outputs and comparing them to radiologist-annotated lung masks using Intersection-over-Union (IoU) and Dice score metrics.
Our results reveal that while Xception had the strongest anatomical alignment (mean IoU = 0.4015, Dice = 0.5682), VGG19's high accuracy was misleading, as it showed near-zero alignment with lung regions. These results demonstrate that accuracy alone can be misleading. The proposed automated pipeline offers a practical, reproducible way to verify whether models base their decisions in relevant regions, supporting the development and selection of safer AI models.

1.Introduction

Artificial intelligence (AI) has revolutionized decision-making in all sectors providing data analysis, automation, and predictive modeling. In the medical field, AI-based systems aid in diagnosis of diseases, personalized approaches, and improved disease prediction (Tahir and Hamarash, 2025). However as these models become more advanced, they also become harder to understand. In healthcare, it is not enough for a model to make accurate decisions, it is also important to understand how those decisions were made and which features influenced them the most. When transparency is missing, it raises serious concerns especially in a field where both trust and accuracy are critical. (Hassija et al., 2024). Explainable AI (XAI) methods address this challenge by making model decisions interpretable, enabling clinicians and researchers to validate predictions and refine systems for real-world use.

To investigate these challenges and limitations, we use COVID-19 detection in the medical domain as a case study. With the emergence of the SARS-CoV-2 virus, which caused the global pandemic, the healthcare systems of most countries were overwhelmed, and necessitating the development of rapid and accurate diagnostic tools. Although reverse transcription polymerase chain reaction (RT-PCR) is still considered the gold standard, its shortcomings such as supply shortages, delays in processing and false negatives have led researchers to explore alternative methods (Wang et al., 2020). Medical imaging, particularly chest X-rays (CXR) and computed tomography (CT) scans, emerged as a viable complement, revealing lung abnormalities associated with COVID-19. However, manual interpretation of these images is time-consuming and subject to human error since the manifestations of the disease are similar to other respiratory diseases. To address these issues, deep learning for computer-aided diagnosis (CAD) systems have shown great potential in supporting medical decision making. Specifically, convolutional neural networks (CNNs) have proven to be effective in identifying patterns in medical images, which would allow the classification of COVID-19 cases with high accuracy (Rojgar Qarani and Haval Abduljabbar, 2025, Nayla Faiq and Shahab Wahhab, 2025). The models have been found to be helpful in the quick and automated diagnosis, particularly in situations where RT-PCR testing is either not available or delayed. To further enhance CNN performance, researchers tend to use transfer learning, which is a method of transferring a pre-trained model trained on one task to a new and more specific task. Transfer learning minimizes the requirement of large amounts data and considerably decreases training time by building on knowledge acquired from large datasets. In the medical field where collecting data is expensive and time-intensive, transfer learning can be valuable. However, one major issue remains: these models often operate as black boxes. In the clinical setting, an AI model can only be effective when it not only provides a diagnosis but also explains how the decisions were made. Unless one is familiar with how these models make their predictions, there is a risk that they may be relying on unintended artifacts or irrelevant patterns in the data (Majeed et al., 2020).

Explainability methods such as gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) can make AI decisions more transparent by generating heatmaps to illustrate which regions of an image had the greatest influence on the model prediction. Among available XAI options (Grad-CAM, Grad-CAM++, Score-CAM, SHAP), Grad-CAM offers broad adoption, low computational cost, and clinician-readable spatial maps. We employ Grad-CAM for post-hoc explainability and define interpretability as overlap between saliency maps and lung masks. We therefore use Grad-CAM for comparability and replace manual inspection with a dataset-level quantitative assessment. This is particularly helpful in the COVID-19 detection where doctors can see whether the AI is focusing on lung abnormalities or is being distracted by irrelevant features. This level of transparency is necessary not just for building trust, but also for improving models, refining training data, and making AI tools safe for real-world healthcare use. In this study, we explore how Grad-CAM improves the interpretability of CNN models for

COVID-19 detection, review existing methods, highlight their strengths and weaknesses, and show how explainability can increase trust and model performance. Using COVID-19 as an example, we highlight the critical need for transparent AI in medicine. Accordingly, this study: 1. Trains six CNNs in transfer learning mode on a local CXR dataset for COVID-19 detection. 2. Analyzes their classification performance. 3. Applies Grad-CAM to explain each model's decisions. 4. Automates Grad-CAM evaluation by thresholding heatmaps and measuring their overlap with lung masks by computing Intersection-over-Union (IoU) and Dice score across the entire test set to gain insights into each model's overall decision-making behavior.

1.1Related Work

In this section, we review key studies that apply CNN-based deep learning models to detect COVID-19 from CXR images. We start by examining research focused on classification accuracy, particularly those using popular CNN architectures and transfer learning techniques. Next, we look at studies that introduced explainability tools like Grad-CAM to make model predictions more interpretable. Our aim is to highlight common methodologies, point out limitations, and clarify how our work builds upon and contributes to this growing field.

(Panwar et al., 2020) developed a custom CNN model called nCOVnet that builds on the VGG16 architecture for feature extraction along with custom built classification layers. The model achieved a sensitivity of 97.62% and an accuracy of 88.10% on a dataset comprising 284 chest X-ray images (142 COVID-19 positive and 142 normal). However, the dataset used is very small. The reported metrics may not generalize to a broader population.

(Sethy et al., 2020) explored an alternative approach to diagnosing COVID-19 using deep learning and machine learning techniques. Instead of relying solely on CNNs for classification, they extracted deep features from CXR images using a pre-trained deep learning model. These extracted features were then fed into a Support Vector Machine (SVM) classifier to distinguish COVID-19 cases from normal and all other pneumonia cases. They evaluated 13 pre-trained CNN models for deep feature extraction on a dataset containing 381 images. ResNet-50 had the best results obtaining the accuracy of 95.33%. The evaluation of 13 pretrained backbones with an SVM classifier are noteworthy. However the dataset is small.

(Apostolopoulos and Mpesiana, 2020) assessed five CNN architectures: VGG19, MobileNet v2, Inception, Xception, and Inception ResNet v2. They used transfer learning using transfer learning on two public CXR collections: Dataset A (224 COVID-19, 700 bacterial pneumonia, 504 normal) and Dataset B (224 COVID-19, 714 bacterial/viral pneumonia, 504 normal). Reporting the highest accuracy (96.78%), Sensitivity (98.66%), specificity (96.46%) with MobileNetV2 in distinguishing COVID-19 cases from others. One limitation is that the classes are imbalanced with relatively few COVID-19 cases (n=224) compared with pneumonia (700–714) and normal (504), increasing the risk of a biased decision boundary.

(Halgurd et al., 2021) explored deep learning and transfer learning for diagnosing COVID-19 using CXR and CT scans. Due to limited datasets, they compiled images from various sources to create a small dataset. They experimented two models: a custom CNN and a modified pre-trained AlexNet. Their results showed that AlexNet achieved 98% accuracy, while the proposed CNN reached 94.1%. Notably, the study relied on a small, aggregated dataset.

Similarly, (Khurana and Soni, 2022), evaluated four CNN architectures: ResNet-50, EfficientNetB0, VGG16, and a custom CNN, to classify images as COVID-19 positive or negative. They used two public datasets: COVID-19 Radiography (X-ray) and HUST-19 (CT) downsampled to 4,000 images. The ResNet-50 model demonstrated the highest performance, achieving an accuracy of (98.7% on CXR, 98.9% on CT scans).

(Narayan Das et al., 2022) applied transfer learning using the Xception model, obtaining significant performance improvements compared to several models achieving the accuracy of 97.4%.

(Hariri and Avşar, 2023) proposed a lightweight convolutional neural network designed to classify images into four categories: Healthy, COVID-19, viral pneumonia, and bacterial pneumonia. The model was compared with nine CNNs using transfer learning, including: MobilenetV2, InceptionResNetV2, ResNetV2, EfficientNet B2, EfficientNet B0, NasNet-Mobile, InceptionV3, VGG16 and VGG19. The evaluation was conducted on a dataset containing 9207 CXR images (3269 normal, 1281 COVID-19, 3001 bacterial pneumonia and 1656 viral pneumonia. Then data augmentation was applied to the training set. The proposed lightweight model outperformed these benchmarks, achieving an accuracy of 89.89%. They also reported difficulty across all models in distinguishing viral pneumonia from bacterial pneumonia, which impacted overall precision. The class distribution is imbalanced, with fewer COVID-19 cases than the other categories.

(El Houby, 2024) experimented with two pre-trained CNNs, namely **VGG19** and **EfficientNetB0**. The models were trained on the **COVID-19 Radiography Dataset**, which included both full and segmented X-ray images enhanced through preprocessing techniques to improve diagnostic accuracy. VGG19 demonstrated superior performance when trained on enhanced full X-ray images, achieving **95% accuracy, 96% sensitivity, 94% specificity, 94.12%.**

Earlier researches achieved strong classification performance, but often failed to explain the reasoning behind the model's predictions. Recognizing the critical need for interpretability in medical decision-making, (Majeed et al., 2020) conducted a comprehensive analysis of CNNs combined with transfer learning. They evaluated 12 popular pre-trained CNN models: AlexNet, VGG16, VGG19, ResNet-18, ResNet-50, ResNet-101, GoogleNet, InceptionV3, SqueezeNet, Inception-ResNet-v2, Xception, and DenseNet201 as well as introduced a custom shallow CNN trained from scratch across three publicly available datasets. Among the evaluated models, VGG19, DenseNet201, and Xception stood out, achieving the highest specificity of 100%. However, despite impressive accuracy metrics, qualitative analyses using class activation maps (CAMs) indicated that these models frequently relied on irrelevant regions within the X-ray images. Consequently, the authors emphasized that CNN-derived predictions should only be trusted when clinicians can verify that the models focus explicitly on relevant anatomical areas.

Similarly, (Chow et al., 2023) provided both quantitative and qualitative insights by evaluating 18 state-of-the-art CNN models with transfer learning for COVID-19 detection from CXRs. In their study, VGG16, ResNet-101, VGG19, and SqueezeNet all achieved over 90% accuracy, with VGG16 reaching the highest accuracy and F1-score at 94.3%. NasNet-Mobile, NasNet-Large, and Xception showed the weakest performance. The authors then selected the top four and bottom three models for qualitative analysis via Grad-CAM, reviewed by radiologists. Despite lower accuracy than VGG16, SqueezeNet's activation maps aligned best with radiologist assessments. A brief summary of the related works is provided in Table 1.

Motivated by these findings, our study focuses on testing and evaluating six different CNN models on a local CXR dataset to detect COVID-19 cases. Unlike prior studies that primarily relied on manual, visual inspection for Grad-CAMs for qualitative analysis, we develop a streamlined, automated process to evaluate Grad-CAMs. This setup not only saves effort and time but also offers consistent and reliable insights. As a result, our work doesn't just deliver new accuracy metrics, it also showcases a practical way to ensure deep learning tools are focusing on medically meaningful areas inside the images. Prior COVID-19 CXR studies commonly use transfer learning and report high accuracies across various backbones. However, many rely on small or class imbalanced datasets and qualitative Grad-CAM visualizations. As a result, there is limited, reproducible evidence that model attention aligns with lung anatomy at a dataset level. To address this gap, we evaluate six CNNs under a unified pipeline and introduce an automated quantitative assessment that thresholds Grad-CAM maps and measures their overlap with radiologist annotated lung masks

using IoU and Dice score, providing objective, anatomy-grounded interpretability alongside standard classification metrics.

Table 1: Summary of related COVID-19 CXR classification studies, sorted by Grad-CAM usage and then publication year.

Study	Grad-CAM used	Best model	Headline metric	Key notes
(Panwar et al., 2020)	No	Custom model	Accuracy: 88.10%	Very small dataset
(Sethy et al., 2020)	No	Deep learning + SVM	Accuracy: 95.33%	Very small dataset
(Apostolopoulos and Mpesiana, 2020)	No	MobileNetV2	Accuracy; 96.78%	Class imbalance
(Halgurd et al., 2021)	No	AlexNet	Accuracy: 98%	Small dataset
(Khurana and Soni, 2022)	No	ResNet-50	Accuracy: 98.7% (CXR), 98.9% (CT)	Dataset truncation
(Narayan Das et al., 2022)	No	Xception	Accuracy: 97.4%.	Small dataset
(Hariri and Avşar, 2023)	No	Custom model	Accuracy: 89.89%.	Difficulty in classifying viral vs bacterial images
(El Houby, 2024)	No	VGG19	Accuracy: 95%	No XAI analysis reported
(Majeed et al., 2020)	Yes	VGG19, DenseNet201, and Xception	Specifity: 100% (for all three models)	Qualitative maps reveal off-lung focus
(Chow et al., 2023)	Yes	VGG16	Accuracy: 94.1%	Top accuracy did not produce the best Grad-CAM

2.Materials and Methods

2.1Model Selection and Architectures

We selected six CNN architectures for this study: VGG16, VGG19, ResNet-101, NASNet-Mobile, NASNet-Large, and Xception, to investigate both their predictive performance and interpretability in diagnosing COVID-19 from CXR images. VGG16, VGG19, and ResNet-101 were chosen for their reported strong performance (Chow et al., 2023). On the other hand, NASNet-Mobile, NASNet-Large, and Xception were included precisely because they showed comparatively weaker results in that same research. By including both best performing and worst performing models allowed us to examine not just classification accuracy, but also how architectural differences might affect interpretability when using Grad-CAM visualizations.

2.2Overview of Selected Architectures

VGG16 and VGG19 come from the Visual Geometry Group (VGG) series known for their simple architecture. They rely on stacking multiple 3×3 convolutional layers with max-pooling layers inserted at regular intervals. The number represents the total number of layers in each architecture where VGG16 has 16 layers and VGG19 has 19 layers.

ResNet-101 belongs to the family of Residual Network (ResNet), the first network architecture that introduced skip (residual) connections to reduce the vanishing gradient issue in deep networks. The number following the ResNet is the number of the layers of weights. With 101 layers, it can capture complex features while still maintaining manageable training stability.

NASNet (Neural Architecture Search Network) models are designed via an automated search algorithm to identify optimal building blocks for convolutional architectures. NASNet models have several hundred layers when counting each convolution, batch normalization, and activation layer. The exact total depends on the implementation because of the searching

algorithm. NASNet-Mobile is the lighter-weight variant optimized for mobile or resource-limited scenarios, while NASNet-Large is a more computationally intensive version.

Xception (Extreme Inception) replaces the standard Inception modules with depthwise separable convolutions. This architecture reduces the number of parameters while retaining representational capacity.

These six models span a range of design philosophies from straightforward (VGG) to residual-based (ResNet) to automated search (NASNet) to depthwise separable convolutions (Xception). Studying such diversity helps generalize findings on which architectural features might yield more meaningful Grad-CAM heatmaps.

All models are widely used and have ImageNet pre-trained weights readily available. This streamlines transfer learning to the COVID-19 classification task and ensures consistency across different architectures.

2.3Transfer Learning Setup

To benefit from the rich visual features learned from large-scale ImageNet training while avoiding overfitting on our smaller COVID-19 dataset, we employed a straightforward transfer learning approach on all six selected models (VGG16, VGG19, ResNet-101, NASNet-Mobile, NASNet-Large, and Xception). This approach treats each pre-trained model as a fixed feature extractor. All the convolutional layers (and their learned weights) remained frozen during training. We removed the original fully connected layers and added a global average pooling layer to shrink high-level feature maps, followed by a fully connected layer having an output size of a single output neuron with a sigmoid activation function. A threshold of 0.5 is used for the activation function to classify COVID-19 or normal.

By freezing all but the final classification layers, we effectively strike a balance between powerful pre-trained representations and customizing the model to accurately detect COVID-19 in CXR images. The next sections include the dataset used, data preprocessing steps, training protocols, and the Grad-CAM workflow used to evaluate each architecture's interpretability.

2.4Dataset Description and Preparation

For this study, we utilized the local dataset KURD-COVID, which was introduced and collected in the study by (Hamad and Majeed, 2022). The dataset consists of 1300 CXR images and their corresponding masks, including 613 images from patients confirmed to have COVID-19 and 687 images from individuals with normal (healthy) lungs. All the images in the dataset have the resolution of 512×512 pixels. Figure 1 presents an example of the dataset, illustrating both the original CXR and its corresponding ground truth mask for normal and COVID-19 cases. We split the 1300 CXR- mask pairs of the KURD-COVID dataset into 910 images for training (70%), 195 for validation (15%), and 195 for final testing (15%). No data augmentation was applied so that every heatmap-mask comparison would reflect the original radiographic appearance. The dataset split is detailed in Table 2.

To ensure compatibility across architectures, each required different input dimensions, the images and corresponding lung masks were resized accordingly to maintain compatibility with the respective models. The resize configurations are summarized in Table 3.

Table 2: Dataset split details across training, validation, and test sets.

Class	Train	Validation	Test	Total
Normal	471	119	97	687
Covid	439	76	98	613
Total	910	195	195	1300

Table 3: Input size requirements for selected CNN architectures.

Model	**Input size**
VGG16	224 × 224
VGG19	224 × 224
ResNet-101	224 × 224
NASNet-Mobile	224 × 224
Xception	299 × 299
NASNet-Large	331 × 331

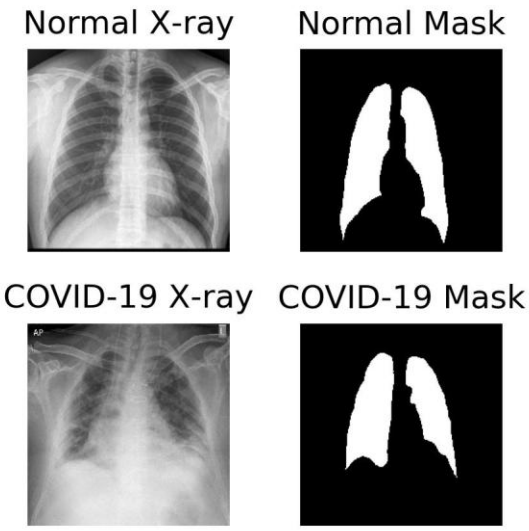

Figure 1. Sample images from the KURD-COVID dataset, displaying original chest X-ray images alongside their corresponding ground truth masks for both normal and COVID-19 cases.

2.5Training Protocol

All experiments were conducted on a machine equipped with an AMD Ryzen7 7435HS processor, 16 GB RAM, and NVIDIA RTX 4060 GPU. The models were implemented in TensorFlow version 2.10. Each model was trained for 20 epochs with a batch size of 32, using the Adam optimizer with a default learning rate of 0.001. The CXR images are used to train the models, and the masks are used to evaluate the quality of the decision making process.

2.6Classification Performance Evaluation

To evaluate the predictive performance, we computed standard classification metrics derived from the confusion matrix:

- True Positives (TP): COVID-19 cases correctly classified as positive.
- True Negatives (TN): Normal cases correctly classified as negative.
- False Positives (FP): Normal cases incorrectly classified as COVID-19.
- False Negatives (FN): COVID-19 cases incorrectly classified as normal.

These values were used to calculate the following metrics:

1. Accuracy: Measures overall correctness of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2. **Precision:** Measures how accurate the model's positive predictions are.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

3. **Recall (Sensitivity):** Indicates the model's ability to detect all positive cases.

$$Recall/Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

4. **F1-Score:** Harmonic mean of precision and recall, balancing both metrics.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

2.7Qualitative Analysis using Grad-CAM

While quantitative metrics (accuracy, precision, recall, F1-score) are necessary for evaluating model performance, they do not reveal how a model arrives at its decisions. To understand which part of the image is considered important, we can use Grad-CAM, an explainability technique that generates a heatmap highlighting the image regions most influential to a model's prediction. Grad-CAM works by computing gradients of the target class score with respect to the final convolutional layer's feature maps during backpropagation, then creating a weighted combination of these maps to produce a coarse heatmap that is upsampled to match the input image size. After the grayscale Grad-CAM has been produced, a colormap is applied to the grayscale image where different colors represent varying levels of influence on the prediction. Red regions show the most critical regions, intermediate colors (like yellow, green) reflect areas of moderate influence, and blue signifies areas with very small or no influence.

Previous methods required manual inspection to check how well Grad-CAM heatmaps aligned with clinically relevant anatomical regions, which was a slow, subjective process. To address these issues, we created an automated system that first converts raw Grad-CAMs into binary maps (using a 0.5 threshold), separating grayscale activation patterns into preserving only

the important regions while filtering out areas with little to no influence (Figure 2). Then we measure how well these binary maps align with actual lung masks using two standard metrics: IoU, originally developed for object detection tasks, and the Dice score, commonly used in medical image analysis. These metrics provide an objective, quantitative measure of how well the model's attention (captured by the binary Grad-CAM) corresponds to the actual lung regions in the images. The formulas of IoU and Dice score are given in Equation (5) and Equation (6).

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

$$Dice\ score = \frac{2|A \cap B|}{|A| + |B|} \quad (6)$$

Where A represents the binary Grad-CAM activations and B is the ground-truth lung mask. This automated approach provides objective, reproducible measures of how well each model's attention corresponds to clinically relevant anatomical regions.

Figure 3 shows the overview of training, classification and the Grad-CAM evaluation pipeline.

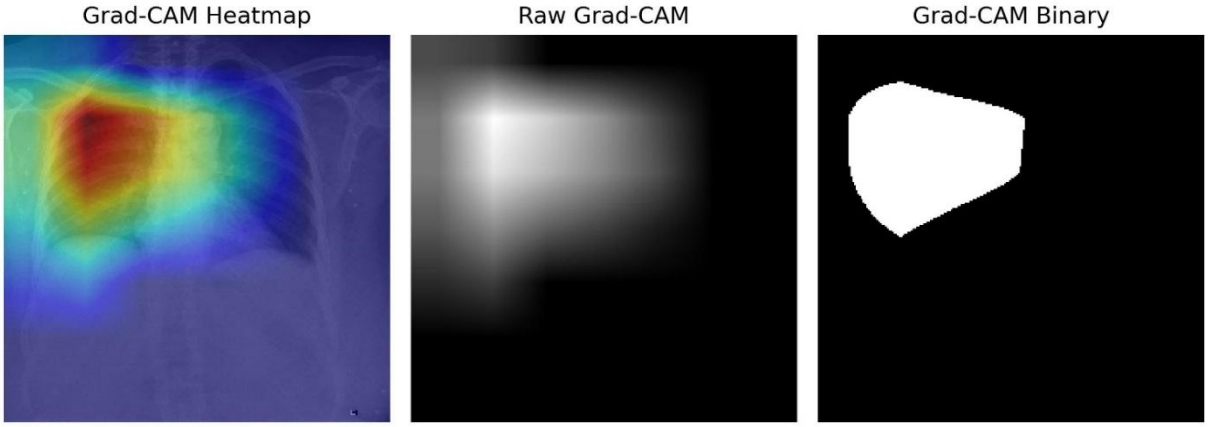

Figure 2. Illustration of the Grad-CAM binarization process: converting raw grayscale activation maps into binary masks using a fixed threshold of 0.5.

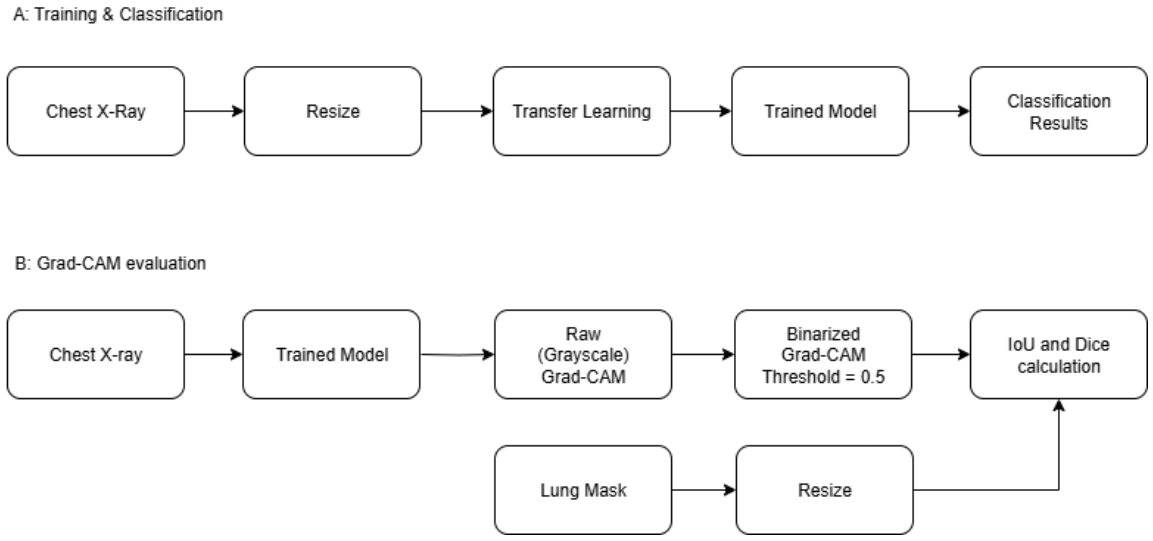

Figure 3. Summary of the pipelines used in this study

3.Results

All six CNNs were trained for 20 epochs under a unified schedule (no early stopping). To keep the section concise, we report complete metrics on the held-out test set. The classification results of all six models are summarized in Table 4, which includes accuracy, precision, recall, and F1-score, sorted by descending accuracy. All CNN architectures demonstrated strong diagnostic performance, with accuracy ranging from 90% to 96%. Xception achieved the highest accuracy (95.90%) along with balanced precision, recall, and F1-score (95.92%). NASNet-Large and VGG19 tied for the accuracy of 94.87%, with VGG19 showing near-perfect precision (98.89%). VGG19 demonstrated near-perfect precision of 98.89, while VGG16 showed the lowest performance, with an accuracy of 90.77% and the lowest recall value of 84.69%.

Although the models achieved high classification accuracy, these results alone cannot validate whether models focused on clinically relevant lung regions or confounding image artifacts. To assess the interpretability of these models, Grad-CAM was applied. Example heatmaps for true positive, true negative, false positive, and false negative cases are illustrated in Figure 4. Visual inspection of these heatmaps revealed distinct attention patterns among the models. Xception, NASNet-Large, and NASNet-Mobile consistently focused on the central regions of the image with partial overlap on the lungs. In contrast, ResNet-101 produced asymmetrical activation patterns, with activations concentrated in the upper left lung and towards the center of the image which appeared less consistent. VGG16 and VGG19 often focused on irrelevant regions, such as the corners and borders of the image, indicating a very weak interpretability. In addition, Figure 5 shows Grad-CAM visualizations with the lowest IoU and Dice scores for each classification case (TP, TN, FP, FN) across the six evaluated models, highlighting worst aligned attention patterns.

To further analyze the Grad-CAM heatmaps numerically and understand the overall model's region of focus, quantitative evaluation of the Grad-CAM heatmaps was conducted using IoU and Dice score on the entire test set. The distribution of scores across ten decile intervals is presented in Table 5 for IoU and Table 6 for Dice, while the average values are summarized in Table 7. Xception demonstrated the highest anatomical alignment, with the majority of samples falling within the 0.4–0.5 IoU and 0.5–0.7 Dice intervals. NASNet-Mobile and NASNet-Large followed behind, demonstrating moderate localization performance. ResNet-101 showed limited alignment capability. VGG16 and VGG19 had the poorest localization, with nearly all scores concentrated in the lowest range of 0–0.1.

Table 4: Classification metrics on test set

CNN model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Xception	**95.90**	95.92	**95.92**	**95.92**
NASNet-Large	94.87	97.83	91.84	94.74
VGG19	94.87	**98.89**	90.82	94.68
ResNet-101	94.36	96.77	91.84	94.24
NASNet-Mobile	93.33	94.74	91.84	93.26
VGG16	90.77	96.51	84.69	90.22

Table 5: Distribution of Grad-CAM heatmap alignment using Intersection-over-Union (IoU) scores across ten bins for six CNN models, sorted by classification accuracy.

IoU Range	Xception	NASNet-Large	VGG19	ResNet-101	NASNet-Mobile	VGG16
[0 - 0.1]	0	1	195	6	2	195
[0.1 - 0.2]	5	12	0	84	14	0
[0.2 - 0.3]	15	70	0	97	75	0
[0.3 - 0.4]	60	94	0	8	80	0
[0.4 - 0.5]	99	18	0	0	24	0
[0.5 - 0.6]	16	0	0	0	0	0
[0.6 - 0.7]	0	0	0	0	0	0
[0.7 - 0.8]	0	0	0	0	0	0
[0.8 - 0.9]	0	0	0	0	0	0
[0.9 - 1.0]	0	0	0	0	0	0

Table 6: Distribution of Grad-CAM heatmap alignment using Dice scores across ten bins for six CNN models, sorted by classification accuracy.

Dice Range	Xception	NASNet-Large	VGG19	ResNet-101	NASNet-Mobile	VGG16
[0 - 0.1]	0	0	195	1	1	195
[0.1 - 0.2]	1	1	0	6	1	0
[0.2 -0.3]	2	7	0	47	9	0
[0.3 - 0.4]	7	30	0	102	34	0
[0.4 - 0.5]	27	88	0	34	80	0
[0.5 - 0.6]	80	66	0	5	63	0
[0.6 - 0.7]	74	3	0	0	7	0
[0.7 - 0.8]	4	0	0	0	0	0
[0.8 - 0.9]	0	0	0	0	0	0
[0.9 - 1.0]	0	0	0	0	0	0

Table 7: Mean Intersection-over-Union (IoU) and Dice scores for Grad-CAM alignment with lung masks across six CNN models, sorted by classification accuracy.

CNN Model	Mean IoU	Mean Dice Score
Xception	**0.4015**	**0.5682**
NASNet-Large	0.3053	0.4635
VGG19	0.0007	0.0014
ResNet-101	0.2063	0.3389
NASNet-Mobile	0.3032	0.4597
VGG16	0.0002	0.0004

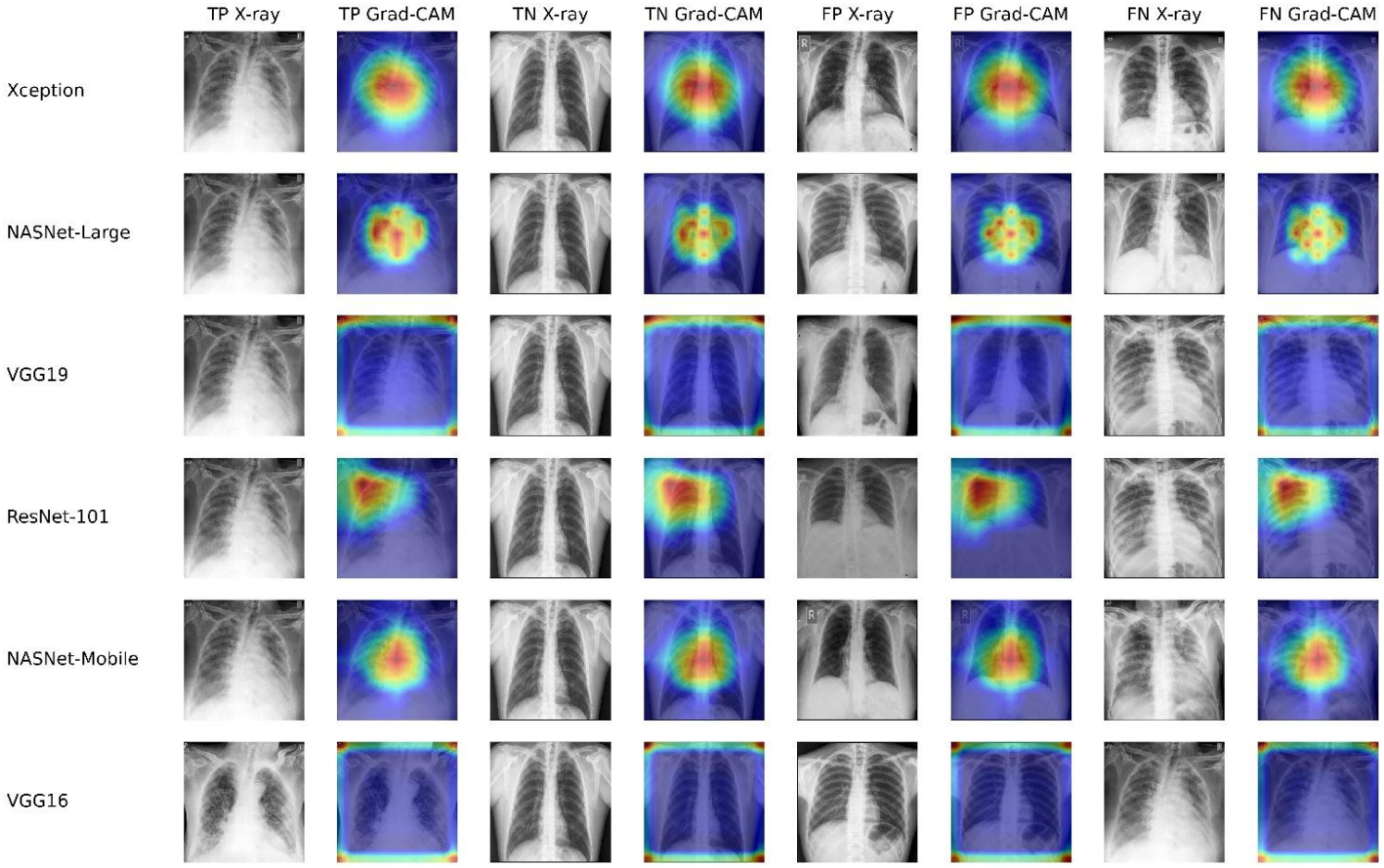

Figure 4. Grad-CAM visualizations for four prediction scenarios (TP, TN, FP, FN) across six CNN models, sorted by descending classification accuracy.

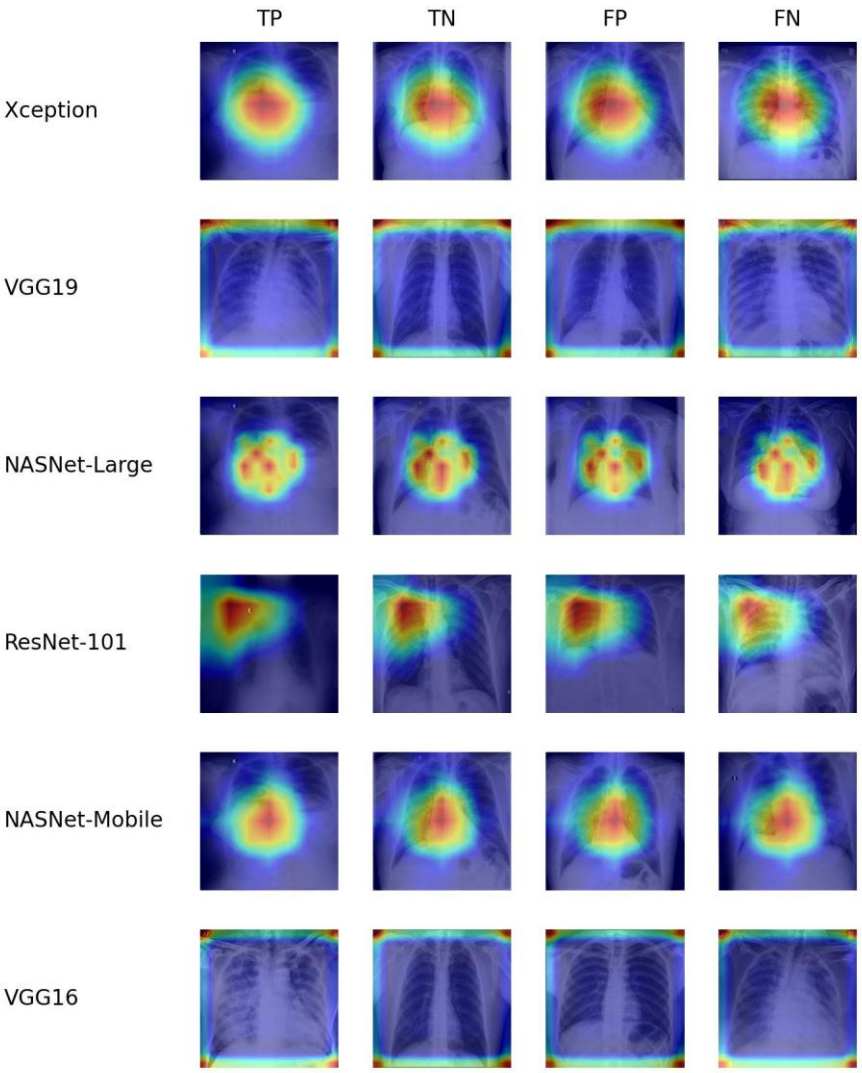

Figure 5. Grad-CAM visualizations showing examples with the lowest IoU and Dice scores for each classification case (TP, TN, FP, FN) across the six evaluated models.

4.Discussion

Although Xception achieved the highest classification accuracy and the strongest alignment with lung regions, the distribution of IoU and Dice scores indicates that attention was still incomplete in most cases. Irrelevant regions were frequently included, and lung regions were often partially covered. This suggests that even top-performing models may consistently fail to base their decisions on medically relevant features when making predictions. The other models performed worse in terms of localization, and in particular, VGG19 demonstrated the highest precision (98.89%) in classification but produced Grad-CAM visualizations with virtually no alignment to the lung regions, indicating that its predictions were based on entirely non-clinical features. These results reinforce that accuracy alone is insufficient for evaluating CNNs in medical imaging and motivate reporting it alongside quantitative, anatomy-grounded evidence of model attention. Our test accuracy is in line with transfer-learning reports in the literature, yet many prior studies either emphasize accuracy alone or rely on qualitative Grad-CAM visualizations. In contrast, we provide dataset-level, quantitative evidence of attention–anatomy alignment, revealing cases such as VGG19 where strong accuracy coexists with poor lung alignment. Methodologically, two choices strengthen the credibility and usefulness of our findings. A key strength is the unified training

protocol across six backbones, enabling fair comparison under the same preprocessing and optimization settings. More importantly, we replace case by case visual inspection with an automated quantitative pipeline that converts interpretability into objective metrics by computing IoU and Dice score overlap between heatmaps and lung masks across the full test set. Reporting performance alongside anatomical alignment offers a practical basis for selecting safer AI models, rather than defaulting to the numerically most accurate model.

Compared with earlier studies (Majeed et al., 2020, Chow et al., 2023) which primarily relied on visual Grad-CAM assessments, our approach standardizes interpretability analysis using spatial metrics across the entire test set. This improves evaluation consistency and uncovers when high accuracy models may not be trustworthy.

Despite these contributions, this study has limitations. Firstly, the analysis was conducted on a single local COVID-19 CXR dataset. Future research should explore larger datasets, other clinical areas, and domains outside of medical fields, to better understand model behavior across varied scenarios. Secondly, the evaluation of Grad-CAM alignment depended on annotated lung masks, a resource that may not be readily available in other domains or datasets. Building on these limitations, several future directions are recommended. There is a pressing need for the development of CNN architectures that are explicitly designed to extract and attend to semantically relevant features, particularly within medical images, to improve the interpretability and trustworthiness of model predictions. Moreover, dataset quality plays a pivotal role in shaping model behavior, and future research should focus on careful inspection and preprocessing of datasets to eliminate confounding elements. Radiographic artifacts, embedded text labels, and irrelevant background regions should be removed where possible to eliminate the risk of these elements being mistakenly extracted as informative features. Finally, we recommend the standardized implementation of IoU and Dice metrics in the Grad-CAM evaluation pipeline. These metrics provide a scalable, quantitative approach for evaluating spatial alignment between model attention and relevant image regions and should be incorporated as a standard component of interpretability assessment pipelines.

5.Conclusion

This study presented an automated framework to evaluate the anatomical focus of six CNN models, including VGG16, VGG19, ResNet-101, NASNet-Mobile, NASNet-Large, and Xception, for COVID-19 detection in CXR images. We used an annotated private dataset of 1,300 CXRs to assess each model's classification performance and how well their Grad-CAM outputs aligned with annotated lung regions.

The classification performance results revealed that model accuracies ranged between 90%–96%, with Xception as the top performer, achieving the highest accuracy of 95.90% and a perfect balance between precision and recall, both reaching 95.92%. NASNet-Large and VGG19 followed closely, both achieving 94.87% accuracy, with VGG19 attaining the highest precision of 98.89%.

Our automated Grad-CAM analysis, based on binary overlap metrics (IoU and Dice score), revealed that even top-performing models lacked consistent focus on clinically meaningful regions. Xception demonstrated the best anatomical focus, achieving a mean IoU of 0.4015 and a mean Dice score of 0.5682. However, these modest values suggest that the model's attention either captures non-lung features or fails to fully focus on the lung regions. VGG19's high precision masked a complete disregard for lung anatomy (mean IoU = 0.0007, mean Dice score = 0.0014), indicating that the model learned to exploit image artifacts rather than pathological features.

Most importantly, these discrepancies were undetectable through traditional evaluation metrics, revealing a significant limitation in current AI validation approaches. Our findings indicate that CNN models can unintentionally learn to depend on artifacts, embedded text, or irrelevant background features during training, which can compromise the clinical reliability of their predictions despite high reported performance. This underscores the need for

interpretability validation as an essential component of model evaluation, as accuracy alone cannot expose reliance on non-clinical features. Among the tested models, Xception stands out as the most promising option for clinical use due to its balanced classification performance and partial anatomical focus. However, further architectural innovations are needed to achieve truly trustworthy AI diagnostics.

Acknowledgement

The authors did not receive any external support and have no additional acknowledgements to declare.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

APOSTOLOPOULOS, I. D. & MPESIANA, T. A. 2020. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine,* 43**,** 635-640.

CHOW, L. S., TANG, G. S., SOLIHIN, M. I., GOWDH, N. M., RAMLI, N. & RAHMAT, K. 2023. Quantitative and Qualitative Analysis of 18 Deep Convolutional Neural Network (CNN) Models with Transfer Learning to Diagnose COVID-19 on Chest X-Ray (CXR) Images. *SN Computer Science,* 4**,** 141.

EL HOUBY, E. M. F. 2024. COVID-19 detection from chest X-ray images using transfer learning. *Scientific Reports,* 14**,** 11639.

HALGURD, S. M., ARAS, T. A., KAYHAN ZRAR, G., ALI SAFAA, S., SEYEDALI, M. & MUHAMMAD KHURRAM, K. Diagnosing COVID-19 pneumonia from x-ray and CT images using deep learning and transfer learning algorithms. Proc.SPIE, 2021. 117340E.

HAMAD, Z. H. & MAJEED, T. F. 2022. Lung Region Segmentation Using Modified U-Net Architecture. *EURASIAN JOURNAL OF SCIENCE AND ENGINEERING,* 8**,** 25-38.

HARIRI, M. & AVŞAR, E. 2023. COVID-19 and pneumonia diagnosis from chest X-ray images using convolutional neural networks. *Network Modeling Analysis in Health Informatics and Bioinformatics,* 12**,** 17.

HASSIJA, V., CHAMOLA, V., MAHAPATRA, A., SINGAL, A., GOEL, D., HUANG, K., SCARDAPANE, S., SPINELLI, I., MAHMUD, M. & HUSSAIN, A. 2024. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation,* 16**,** 45-74.

KHURANA, Y. & SONI, U. 2022. Leveraging deep learning for COVID-19 diagnosis through chest imaging. *Neural Computing and Applications,* 34**,** 14003-14012.

MAJEED, T., RASHID, R., ALI, D. & ASAAD, A. 2020. Issues associated with deploying CNN transfer learning to detect COVID-19 from chest X-rays. *Physical and Engineering Sciences in Medicine,* 43**,** 1289-1303.

NARAYAN DAS, N., KUMAR, N., KAUR, M., KUMAR, V. & SINGH, D. 2022. Automated Deep Transfer Learning-Based Approach for Detection of COVID-19 Infection in Chest X-rays. *IRBM,* 43**,** 114-119.

NAYLA FAIQ, O. & SHAHAB WAHHAB, K. 2025. Enhancing Brain Tumor Classification Accuracy Using Deep Learning with Real and Synthetic MRI Images. *Zanco Journal of Pure and Applied Sciences,* 37**,** 126-149.

PANWAR, H., GUPTA, P. K., SIDDIQUI, M. K., MORALES-MENENDEZ, R. & SINGH, V. 2020. Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. *Chaos, Solitons & Fractals,* 138**,** 109944.

ROJGAR QARANI, I. & HAVAL ABDULJABBAR, S. 2025. Sequential Hybrid Integration of U-Net and Fully Convolutional Networks with Mask R-CNN for Enhanced Building Boundary Segmentation from Satellite Imagery. *Zanco Journal of Pure and Applied Sciences,* 37**,** 157-171.

SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D. & BATRA, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV), 22-29 Oct. 2017 2017. 618-626.

SETHY, P. K., BEHERA, S. K., RATHA, P. K. & BISWAS, P. 2020. Detection of Coronavirus Disease (COVID-19) Based on Deep Features and Support Vector Machine. *Preprints.* Preprints.

TAHIR, Y. M. & HAMARASH, I. I. 2025. Enhanced Human Activity Recognition (HAR) with IMU Sensors in Smartphones: Insights from Machine Learning Models. *Zanco Journal of Pure and Applied Sciences,* 37**,** 101-110.

WANG, L., LIN, Z. Q. & WONG, A. 2020. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports,* 10**,** 19549.