Applied Computer Science 2025, vol. 21, no. 3, pp. 38–58
https://doi.org/10.35784/acs_8015

Submitted: 2025-06-30 | Revised: 2025-08-04 | Accepted: 2025-08-18

Keywords: lung diseases detection, deep learning, Grad-CAM, ensemble learning, chest X-ray imaging

Patrycja KWAŚNIEWSKA [1], *Grzegorz ZIELIŃSKI* [1], *Pawel POWROZNIK* [1], *Maria SKUBLEWSKA-PASZKOWSKA* [1*]

[1] Lublin University of Technology, Poland, kwasniewska.patrycja01@gmail.com, s95622@pollub.edu.pl, p.powroznik@pollub.pl, maria.paszkowska@pollub.pl

[*] Corresponding author: maria.paszkowska@pollub.pl

Pulmonary diseases identification: Deep learning models and ensemble learning

Abstract

Deep learning models provide tremendous support for medical imaging by understanding lung conditions and indicating multiple lung diseases. Due to the global burden of respiratory diseases, their prevention and control is of great importance. Therefore, this study focuses on the effectiveness of different deep learning architectures in diagnosing lung diseases from chest X-ray images. Five deep convolutional neural networks are involved: VGG16, DenseNet-121, ResNet-50, MobileNet, and Vision Transformers. They are pre-trained using the ImageNet dataset. Both transfer learning and development of custom models based on the above architectures will be applied. The study deals with the determination of the most effective single model for the identification of lung diseases. The gradient-weighted class activation map is used to highlight the key regions that influence model decisions. In addition, soft voting ensemble learning methods are used to improve the performance of lung disease detection. Commonly used metrics are applied to evaluate all models. The results for COVID-19, pneumonia and normal case identification exceeded 95% accuracy, 95% precision, 96% recall and 95% Fβ for individual models. The ViT model outperformed DenseNet-121, achieving 96.66% accuracy. The results for bacterial pneumonia, viral pneumonia, tuberculosis, COVID-19 and healthy case identification exceeded 85% accuracy, 86% precision, 85% recall and 94% Fβ for single models. Ensemble learning further improved performance. These results demonstrate the high potential of deep learning and ensemble approaches to support accurate and efficient diagnosis of lung diseases using chest X-rays. The deep learning models provide promising decision support tools for this type of healthcare diagnosis.

1. INTRODUCTION

Machine learning and deep learning algorithms are widely used to develop effective support tools for pulmonologists, reducing the workload in healthcare. The deep learning models provide tremendous support for medical imaging by understanding the lung conditions and indicating the multiple lung diseases (Hamza et al., 2022; Ismael & Şengür, 2021; Jasmine Pemeena Priyadarsini et al., 2023). For high-speed identification, the chest chest X-ray are one of the most important and common radiological examinations and therefore have a major impact on patient care (Abbas et al, 2021; Kong & Cheng, 2022; Pereira et al, 2020; Rahaman et al., 2020).

Convolutional Neural Networks (CNNs) provide effective image understanding by extracting features that result in accurate image-based identification (Fan, 2023; Jogin et al., 2018; Powroźnik et al., 2024; Skublewska-Paszkowska et al., 2024). Single-backbone CNNs have gained much interest in lung disease detection, where deep multilayer neural networks are applied to reveal complex patterns and features (Jogin et al., 2018; Kesuma et al., 2023). Vision Transformer (ViT) is another approach to image-based identification. This technique reveals global relationships from non-overlapping image patches. ViT incorporates a self-attention mechanism that provides it with a comprehensive understanding of images, allowing it to detect broader dependencies than in the case of CNNs. This architecture has attracted much interest due to its remarkable performance in pattern recognition tasks (Öztürk et al., 2025).

Ensemble learning is a powerful technique in machine learning that combines the predictions of multiple models, also known as base learners, to produce a final prediction that is more reliable and accurate (Akter et

al., 2023; Balan et al., 2024; Kesuma et al., 2023; Rajpoot et al., 2024). The basic idea is that even though individual models may have flaws or inaccuracies, integrating their results can counteract these drawbacks and lead to better overall performance. This idea is especially helpful for complicated problems like image classification, where a variety of visual cues can make it difficult for a single model to generalize effectively. There are several approaches to ensemble learning, one of the most effective being voting-based ensembles (Akter et al., 2023; Kumaran et al., 2024; Mabrouk et al., 2022). The most notable of these is soft voting. Each base model in soft voting produces a probability distribution over all possible classes as an output, rather than just a class label. After averaging these probabilities across all models, the class with the highest mean probability is selected as the final prediction. In the context of image classification, soft voting has been widely used to improve performance in both academic research and real-world applications. For example, in medical imaging applications such as identifying pneumonia, COVID-19, or other conditions from X-rays, models can be tuned separately and then integrated using soft voting. Together, their probabilistic outputs provide a more thorough evaluation of the input image because each model may be particularly good at capturing different visual elements, such as local textures, edge patterns, or global context. This typically results in higher accuracy and lower rates of false positives and false negatives, which is critical in medical diagnostics. However, these methods do not always improve the accuracy of the model (Rajpoot et al., 2024). This is why it is important to start with simple models and then apply advanced modeling techniques.

The use of deep neural networks and ensemble learning very often goes hand in hand with the use of visualization techniques. The Gradient-weighted Class Activation Mapping (Grad-CAM) method is often utilized to highlight the most relevant areas that were utilized for making predictions (Hamza et al., 2022; Kumaran et al., 2024; Umair et al., 2021). This visual interpretability is widely applied to indicate explanations for model decisions.

The main motivations for automated medical diagnosis include:

- The large number of lung disease cases and the increasing workload of pulmonologists,
- The importance of rapid and accurate diagnosis due to the time-consuming nature of medical image analysis,
- The high mortality rate associated with pulmonary disease,
- Limited availability of advanced diagnostic techniques
- Limited access to local specialists, especially in rural areas,
- The need for more accurate diagnosis.

In response to the above issues, this study focuses on the comparison of selected deep learning models for efficient identification of common lung diseases. CNN pre-trained architectures as well as ViT models have been analyzed. The soft voting ensemble learning is used to verify its performance for various lung disease identification. The Grad-CAM heatmaps are chosen to indicate the most relevant image areas on which the decision is made. Two experiments are conducted to determine the most effective models for different types of lung diseases.

The main contributions of this study are as follows:

- Create the new datasets consisting of COVID-19, pneumonia, and normal cases.
- Identify COVID-19, pneumonia and normal cases using the ResNet-50, DenseNet-121 and ViT architectures with high accuracy. The ResNet-50 achieved the highest performance, reaching 97% accuracy.
- Application of soft voting ensemble learning to the ResNet-50, DenseNet-121, and ViT models, which reduced misclassification, especially between normal and pneumonia classes.
- Apply Grad-CAM to localize the most relevant areas for model decision.
- Identification of bacterial pneumonia, viral pneumonia, tuberculosis, COVID-19, and healthy cases using the ResNet-50, DenseNet-121, VGG16, and MobileNet. The VGG16 was found to be the most effective model reaching 89.23% accuracy.
- The soft voting ensemble learning is proposed, which involves different merging of the individual models. Among all tested combinations, the best performance was achieved by the pair VGG16 + MobileNet, which reached an accuracy above 90%.
- Application of Grad-CAM to localize the most relevant areas for the model decision.

The rest of the paper is organized as follows. Section 2 presents an overview of deep neural network architectures used for lung disease identification. In section 3, the material and methods are described. The

obtained results together with the discussion are summarized in section 4. Finally, in section 5, the summary for the conducted studies is presented.

2. RELATED WORKS

2.1. Overview of neural network models in medical diagnostics

In recent years, interest in the application of machine learning to medicine has grown significantly. Particular attention has been paid to the diagnosis of lung diseases. An analysis of scientific publications shows that there is a growing trend of studies devoted to convolutional CNN neural networks in the context of lung diseases. In 2016, 17 articles in this field were reported in the Scopus database, while by 2024, the number of papers has already increased sharply to 4011 publications (Fig. 1).

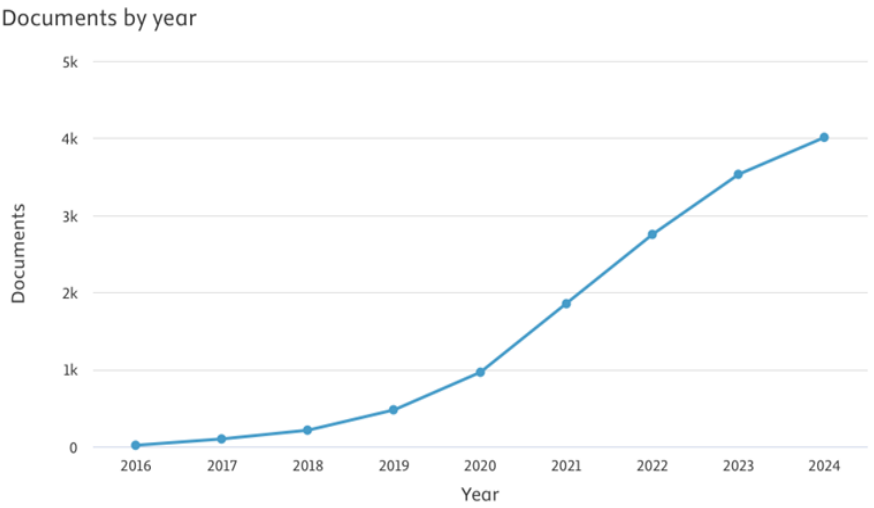


Fig. 1. Number of papers containing the keywords "convolutional neural network", "lungs", "CNN", "disease" and "machine learning" from 2016 to 2024 in the Scopus database

Based on an analysis of publications in the Scopus database from 2014 to 2024, a clear preference in the choice of neural network architecture in medical diagnosis was noted. Among the deep learning models, the most popular are:

- VGG16 was chosen for its easy implementation due to the simplicity and clarity of its architecture (Akter et al., 2023; Jasmine Pemeena Priyadarsini et al., 2023; Jenber Belay et al., 2024; Kong & Cheng, 2022; Kumaran et al., 2024; Marwa & Mohammed, 2024; Moujahid et al., 2022; Odeh & Mustafa, 2024; Öztürk et al., 2025; Umair et al., 2021)
- DenseNet-121 was widely applied due to its structure, which is characterized by direct connections between all layers (Akter et al., 2023; Hadj Bouzid et al., 2024; Naik et al., 2020; Odeh & Mustafa, 2024; Sriporn et al., 2020; Umair et al., 2021)
- AlexNet was utilized due to small number of layers and the quick training (Muntasir et al., 2024; Taha Ahmed & Malallah Kadhem, 2021)
- ResNet-50 was applied due to its high accuracy in complex tasks (Akter et al., 2023; Hadj Bouzid et al., 2024; Kumaran et al., 2024; Odeh & Mustafa, 2024; Sriporn et al., 2020; Umair et al., 2021)
- MobileNet was considered due to its optimization for mobile devices and low computational requirements (Akter et al., 2023; Hamza et al., 2022; Moujahid et al., 2022; Muntasir et al., 2024; Sriporn et al., 2020; Umair et al., 2021)
- InceptionV3 was applied because of its convenient scalability (Akter et al., 2023; Kumaran et al., 2024; Odeh & Mustafa, 2024)
- Xception was chosen because of the high performance achieved through depthwise separable convolution (Akter et al., 2023; Sharma et al., 2024).

Less commonly, but still used in scientific studies, are the EfficientNet (Hadj Bouzid et al., 2024; Vignesh Kumaran & Preethi, 2025) and CapsNet (Bhosale et al., 2024).

2.2. Data foundation

Progress in computer-aided diagnosis is inextricably linked to the quality and quantity of radiographic data. Several of the reviewed studies address this bottleneck. Large, balanced CXR repositories such as COVQU (≈18 k images) (Waheed et al., 2020) and COVIDGR-1.0 with graded severity labels (Tabik et al., 2020) allow for robust training and fair evaluation across disease spectra. Where acquisition is limited, authors generate additional samples - either through classical augmentation pipelines or, more aggressively, through synthetic images generated by an ACGAN in CovidGAN (Waheed et al., 2020). Lung segmentation with U-net variants (Ismael & Şengür, 2021; Waheed et al., 2020) and targeted contrast enhancement (gamma correction was best in (Waheed et al., 2020)) further enrich the effective signal-to-noise ratio before any learning takes place.

2.3. Single-backbone convolutional networks

A substantial body of research shows that carefully tuned, stand-alone CNNs can already achieve strong performance in CXR-based detection tasks. Custom architectures like DarkCovidNet (Ozturk et al., 2020) and CoroNet (Khan et al., 2020) show that domain-specific tuning of the layer depth or receptive field can increase the accuracy above 95% in binary settings. Nevertheless, generic ImageNet backbones remain highly competitive: ResNet families dominate comparative benchmarks (Jain et al., 2020; Nayak et al., 2021; Rahaman et al., 2020) EfficientNet-B4 achieves 99.6% accuracy in binary protocol (Marques et al., 2020) and compact ResNet-18 patch-based pipelines (Ismael & Şengür, 2021) mitigate overfitting when only a few hundred positives are available. In all of these investigations, rigorous hyperparameter search and stratified cross-validation are recurring ingredients of success (Nayak et al., 2021).

2.4. Hybrid and multimodal pipelines

When dataset size or class imbalance threatens to impede end-to-end training, several groups fuse deep visual embeddings with external classifiers. Representative examples include CNN-feature extraction followed by linear or kernel SVMs (Abbas et al., 2021) and Social Mimic Optimisation that distils MobileNetV2/SqueezeNet features before SVM discrimination (Toğaçar et al., 2020). Beyond CXRs, one work explores transfer learning across X-ray, ultrasound, and CT imagery, reporting particularly high accuracy for ultrasound (Horry et al., 2020). These hybrid and multimodal strategies underline the flexibility of deep representations: once distilled, they interface naturally with classical machine-learning or heterogeneous data sources.

2.5. Ensemble and hierarchical strategies

Ensemble learning is currently widely used in the context of research on the application of deep neural networks in medicine. To reduce the variance inherent in individual models and to improve generalization, ensemble paradigms are widely used. In this technique, multiple machine learning models are combined into a single optimal predictive model that leverages the best features of each model. This technique has many advantages, such as: increasing classification accuracy by leveraging the strengths of each model (Abd Elaziz et al, 2022; Akter et al, 2023; Jenber Belay et al, 2024; Kesuma et al, 2023; Kumaran et al, 2024; Mabrouk et al, 2022; Ozturk et al, 2020; Rajpoot et al, 2024) reducing the risk of overfitting and improving the ability to detect different patterns (Abd Elaziz et al, 2022; Akter et al, 2023; Balan et al, 2024; Kumaran et al, 2024) improved detection of comorbidities (Ozturk et al., 2020) and improved accuracy for small datasets (Jenber Belay et al., 2024; Ozturk et al., 2020; Rajpoot et al., 2024). Several approaches are used. The stacking technique combines independently trained models using a metamodel that learns from their outputs (Abd Elaziz et al., 2022; Akter et al., 2023; Imam et al., 2024). Stacking method combining DenseNet-121 and MobileNet achieved 78% accuracy in classifying seven skin cancer types (Rajpoot et al., 2024). It is worth noting that this result was lower than the accuracy of the individual models (88% and 87%). This suggests that the effectiveness of ensemble learning depends on the appropriate choice of underlying models and methods for aggregating results. Voting consists of aggregating predictions from models to obtain a final prediction (Mabrouk et al., 2022; Ozturk et al., 2020). Depends on averaging of model results (Castillo-Barnes et al., 2023). A majority-voting cascade that first separates normal from abnormal radiographs and then discriminates COVID-19 from other pneumonias (Chandra et al., 2021) provides >98% accuracy in the initial stage. Iteratively pruned ensembles of modality-specific networks increase AUC to 0.997 on composite datasets

while keeping inference costs low (Rajaraman et al., 2020). Hierarchical label structures, rather than flat multi-class Softmax, further improve macro-F1 by explicitly reflecting the clinical taxonomy of lung disease (Pereira et al., 2020). Feature concatenation in ensemble learning combines features extracted by different models (Abd Elaziz et al., 2022; Balan et al., 2024; Imam et al., 2024; Rajpoot et al., 2024). The ensemble model consisting of DenseNet-169, ResNet-50, and VGG16 architectures using feature concatenation technique was adopted for COVID-19 detection from X-ray and CT images (Rajpoot et al., 2024). The use of ensemble learning improved accuracy from 98% (for individual DenseNet-169 and VGG16 models) to 99%.

2.6. Model interpretability and feature visualization

Clinical use requires more than raw performance. Radiologists need insight into the image evidence behind each prediction. Saliency-Guided Patch Voting (Ismael & Şengür, 2021) and probabilistic Gradient-Weighted Class Activation Mapping (Grad-CAM) maps localize discriminative regions with minimal parameter overhead. In other words, it generates activation maps that indicate which areas of the images had the greatest impact on the decisions made by the model (Hadj Bouzid et al, 2024; Hamza et al, 2022; Moujahid et al., 2022; Muntasir et al, 2024; Odeh & Mustafa, 2024; Ozturk et al, 2020; Panwar et al, 2020; Rajpoot et al., 2024; Sharma et al., 2024; Shaziya, 2022; Umair et al, 2021; Vignesh Kumaran & Preethi, 2025). This technique has been used to analyze the VGG16 and MobileNet models (Moujahid et al., 2022). It showed that the VGG16 made decisions based on the correct areas of the lungs, while the MobileNet focused on too-small area, which explains its slightly worse results. Attention mechanisms injected into the VGG-16 (Sitaula & Hossain, 2021) not only increase classification accuracy, but also sharpen the spatial focus on pathologically relevant areas. DeTraC's class decomposition layer (Abbas et al., 2021) serves the dual purpose of addressing data irregularities and revealing finer-grained visual patterns that drive the final decision. The Grad-CAM visualization technique was used to increase the confidence in the predictions of four deep neural network models, VGG16, ResNet-50, InceptionV3, and DenseNet-121, for COVID-19 detection (Odeh & Mustafa, 2024).

2.7. Synthesis of effective practices

Taken together, the literature converges on three pillars for reliable lung disease screening with deep learning: curated, balanced data, enhanced by segmentation or GAN-based synthesis; parameter-efficient CNN backbones, fine-tuned under rigorous validation; and diversity, whether through hybrid learning, hierarchical labeling, or lightweight ensembles, to hedge against dataset shift.

An analysis of the average accuracy and F1 scores of the most popular models in lung disease classification has been calculated (Tab. 1). It was found that the DenseNet-121 model achieved the best results in both binary (98%) and multiclass classification (87.5%). Satisfactory results were also obtained by MobileNet , despite its lighter architecture (98% in binary classification and 86.5% in multiclass classification). A literature review confirms that modern deep convolutional networks are effective tools to support medical diagnosis, with a trend of improving results in newer architectures using advanced learning mechanisms.

Tab. 1. The arithmetic mean of accuracy and F1-score from studies concerning VGG16, DenseNet-121, ResNet-50 and MobileNet

AVG (Accuracy, F1-score) [%]					
VGG16	**DenseNet-121**	**ResNet-50**	**MobileNet**	**Type of classification**	**Reference**
84.14	96.75	92.74	96.74	binary	(Umair et al., 2021)
72.00	87.50	81.50	86.50	multiclass	(Akter et al., 2023)
90.95	95.03	94.21	90.24	binary	(Mabrouk et al., 2022)
96.00	98.00	97.00	98.00	binary	(Yu et al., 2021)

3. MATERIALS AND METHODS

3.1. Datasets

Two datasets representing different types of pneumonia diseases are involved in this study. The first one (Dataset#1) is sourced from publicly available repositories and has been compiled and made available on

Kaggle. The original sources are from the other repositories available on Kaggle (Patel, 2019) and github (Cohen et al., 2020a; 2020b). The data set contains a total of 6,432 chest X-ray images of patients who were suffered from COVID-19 and pneumonia or healthy as shown in Figure 2. The training set consists of a total of 5,144 images, with 460 images classified as COVID-19, 1,266 images classified as normal, and 3,418 images classified as pneumonia, which is the most representative group. The test set consists of 1,288 images, with 116 images classified as COVID-9, 317 images classified as Normal, and 855 images classified as Pneumonia.

To increase the size and variety of the data, a data augmentation technique is used. During learning, input images are randomly modified so that each image is used multiple times, increasing generalization and reducing the chance of model overfitting, which is useful with limited data. Using this technique in conjunction with pre-trained models is particularly beneficial because it promotes model adaptation to the unique features of a new image while maintaining the learned representations. Data augmentation techniques applied on Dataset#1:

1. Random Resized Crop: Simulates zooming in and out of objects in an image. The method randomly resizes the image from 80% to 120% of the original image size before resizing to a fixed input size.
2. Random Horizontal Flip: Images may be flipped horizontally with a 50% probability.
3. Random Rotation: Increases the model's resistance to changes in object orientation by randomly rotating the image within ±15 degrees.
4. Random transformation: Geometric distortion to simulate natural zooms and shifts in the field of view. Slight scaling changes from 90% to 110% and o-sets in both directions up to 10%.
5. Color jitter: To mimic changes in color and lighting, brightness and contrast were changed by up to 20% and hue was changed by 10% to simulate real-world situations.
6. Gaussian blur: To simulate low quality or blurry images, a Gaussian blur was randomly applied in the range of 0.1 to 2.0, causing variations in image sharpness.
7. Resizing and normalization: The modified images were normalized to the target size. Finally, the standard mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] values were used to normalize the pixel values.

a) **b)** **c)**

Fig. 2. Example of Chest X-ray, a) Covid-19; b) Normal; c) Pneumonia

The second dataset (Dataset#2) used is called "Lung Disease Dataset (4 types)" and is publicly available on the Kaggle website (Dalvi, 2022). It contains 10113 images in the following formats: .png, .jpg and .jpeg. The data set is divided into three folders: training, validation, and test. The training set contains 6060 images, the validation set contains 2022 images, and the test set contains 2031 images. This gives about 60% for the training and 20% each for the validation and test sets. This split provides enough data for the model to learn patterns and tune hyperparameters. Each folder listed contains subfolders representing diseases: bacterial pneumonia, viral pneumonia, tuberculosis, COVID-19, and healthy lung (Fig. 3). The training set contains approximately 1215 images per class and the validation and test sets contain approximately 405 images per class. The size of all images was set to 224 × 224 pixels.

a) b) c)

Fig. 3. Example of a) bacterial pneumonia; b) viral pneumonia; c) tuberculosis

These datasets include a wide variety of chest X-rays from different publicly available databases to achieve a well-represented distribution of cases in the three categories. This dataset serves as a basis for evaluating the effectiveness of selected neural network tools in classification tasks based on chest X-rays.

The images from the "Lung Disease Dataset (4 types)" were augmented to increase data diversity, improve generalization, and expand the dataset. Augmentation included random rotations up to 20°, horizontal and vertical displacements up to 20%, zooming up to 20%, and random brightness adjustments between 80% and 120%. Images were also flipped horizontally. Vertical flipping was disabled due to the importance of orientation in X-ray images. New pixels created during augmentation were filled using the "nearest" method - using the value of the closest neighboring pixel. These augmentation techniques made the model more robust to displacement, symmetry, scale, and lighting variations. Augmentation was applied only to the training set and performed dynamically during training ("on-the-fly" augmentation). During each of the 50 training and fine-tuning epochs, the model processed 6,060 unique augmented image versions. In total, the model saw approximately 606,000 variations (100 epochs × 6,060 images) distributed across five categories - approximately 121,200 variations per disease class. It's worth noting that since augmentation is random, some versions may have repeated. The physical number of images remained the same - augmentation increased the diversity, not the size of the dataset. The detailed description of the following classes of both datasets can be found in Tab. 2. It should be emphasized that for Dataset #2 the augmentation was done during the training process using the torchvision.transforms tool.

Tab. 2. Datasets used in the study

Dataset	**Covid-19**	**Pneumonia**	**Bacterial pneumonia**	**Viral pneumonia**	**Tuberculosis**	**Healthy**
Dataset#1	Training: 460 Testing: 116	Training: 3418 Testing: 855	-	-	-	Training: 1266 Testing: 317
Dataset#2 before augmentation	Training: 1215 Validating:405 Testing: 405	-	Training: 1215 Validating: 405 Testing: 405	Training: 1215 Validating: 405 Testing: 405	Training: 1215 Validating: 405 Testing: 405	Training: 1215 Validating: 405 Testing: 405
Dataset#2 after augmentation	Training: 121 200 Validating:405 Testing: 405	-	Training: 121 200 Validating: 405 Testing: 405	Training: 121 200 Validating: 405 Testing: 405	Training: 121 200 Validating: 405 Testing: 405	Training: 121 200 Validating: 405 Testing: 405

3.2. Deep learning models

ResNet-50 is a member of the ResNet family of CNN architectures. Its name comes from the number of layers used to build it. It is known for its use of residual connections to solve the problem of vanishing gradients in deep networks. With its fast learning in complex hierarchical features from images, the ResNet-50 has gained notoriety and great momentum in application to a wide range of deep learning tasks. One of the applications has been in medical image classification, where the analysis of chest X-rays from COVID-19 patients has been used very well to perform binary and multi-class classification tasks. Most important in the ResNet-50 is the use of residual blocks. These introduce skip connections, allowing gradients to flow easily

through the network as the model is trained. This design allows the network to learn deep, abstract features, and eliminates the risk of accuracy degradation that often occurs when building deeper networks. ResNet-50 can efficiently extract complex patterns from medical images, which is critical for identifying subtle signs of COVID-19 in chest X-rays. While fine-tuned on COVID-19-specific datasets, the model achieves very good accuracy, especially in binary classification tasks, such as distinguishing COVID-19 from non-COVID-19 cases. Training the model on the test set requires enormous computational resources, making the model unsuitable for resource-constrained environments. Additionally, this means that the model may be prone to overfitting when trained on small datasets. The complexity and depth of the model may indicate this tendency. In cases where the amount of labeled data is limited, this becomes problematic. The solution to this problem is transfer learning, where the model is first initialized with pre-trained weights on large datasets and then fine-tuned on smaller datasets. The ResNet-50 is most commonly used in environments where large datasets are available and large computing resources are available, such as hospitals and research institutions. Although it can be fine-tuned for other classification tasks, such as distinguishing between lung diseases like COVID-19, viral pneumonia or bacterial pneumonia, this does not make it much more suitable for low-resource environments because the requirements are still high. All in all, the best use of the ResNet-50 for automated COVID-19 detection will be in healthcare facilities and research institutions where available resources are not an issue.

DenseNet-121 is a deep learning model belonging to the DenseNet family, designed to improve the flow of information and gradients across the network, solving typical problems in training deep networks. The model is characterized by dense connectivity, where every layer in a dense block is connected to every other layer. This solution helps overcome challenges such as vanishing gradients and promotes effective feature learning, which is particularly useful in image recognition tasks. Unlike traditional CNN networks, where each layer is only connected to the previous layer, the DenseNet-121 ensures that each layer within a block is connected to every other layer, improving the flow of information and gradients during backpropagation. The convolutional layer is the beginning of this DenseNet model architecture, followed by a sequence of dense blocks. Each block contains multiple convolutional layers, and the outputs of each previous layer are combined, resulting in a rich feature representation. As a result, each layer has a rich set of features from the previous layers, resulting in more efficient learning and feature extraction. The transition layers after each dense block reduce the dimensionality of the feature maps and perform downsampling, improving computational efficiency without sacrificing network performance. The final layers of DenseNet-121 include global average pooling, which converts the output feature map into a single vector, followed by one or two dense layers that make the final class prediction. The main advantage of DenseNet-121 is the ability to reuse features across layers. This reuse not only improves the learning of complex representations, but also reduces the number of parameters compared to traditional CNNs. Despite having multiple layers, the DenseNet-121 achieves high performance with fewer parameters than standard CNNs.

The DenseNet-121 shows excellent performance in image recognition and vision related tasks. It excels at large-scale image classification problems, such as those posed by datasets where it classifies images into categories such as objects and animals. The model also performs well in transfer learning, where pre-trained models can be fine-tuned for specific tasks, even with small datasets. This makes DenseNet-121 particularly valuable in scenarios where labeled data is limited. DenseNet-121 represents a significant innovation in deep learning, combining dense connectivity with efficient feature reuse to address the challenges of gradient flow and parameter efficiency in deep networks. Its architecture, which includes dense blocks, transition layers, and global average pooling, allows it to perform well in various image recognition tasks while maintaining computational efficiency. Although training can be time and memory consuming, DenseNet-121's ability to deliver competitive performance with fewer parameters makes it a powerful tool in many computer vision applications.

The VGG16 model is a convolutional neural network architecture developed by the Geometry Group (VGG) at the University of Oxford in 2014 for image classification (Simonyan & Zisserman, 2014). The architecture has 16 layers, 13 of which are convolutional layers with a fixed filter size of 3×3. The remaining 3 layers are fully connected layers. The model requires that the input images have a dimension of (224, 224, 3) - (width, height, number of RGB channels). Each block of the VGG16 model contains a set of convolution layers, ending with a max-pooling layer that reduces the dimensionality of the output data from the convolution layers. After passing through all blocks, the data is flattened and passes through 3 fully connected layers. The architecture is completed with a Softmax layer to classify 1000 classes. The VGG16 has become the basis for much further research (Akter et al., 2023; Kong & Cheng, 2022; Moujahid et al., 2022; Vignesh Kumaran &

Preethi, 2025). It is widely used in image classification, segmentation and object detection tasks (Jasmine Pemeena Priyadarsini et al, 2023; Marwa & Mohammed, 2024; Odeh & Mustafa, 2024; Ozturk et al, 2020; Umair et al, 2021).

The MobileNet model was introduced in 2017 by a team of Google researchers in a research paper titled "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications" (Howard et al., 2017). It is tailored for environments with limited computing resources. Due to its lightweight architecture, it is used to perform fast and accurate predictions (Akter et al., 2023; Hamza et al., 2022; Moujahid et al., 2022; Muntasir et al., 2024; Sriporn et al., 2020). It is applied in medical devices that operate in the field and in mobile applications that support remote health care. It differs from other models by using depth-separable convolutions in its architecture, a technique that splits a standard convolution into two steps. The first applies a single filter to each input channel separately, and the second applies a 1x1 convolution to combine the results of the previous step. It also has two hyperparameters: width and resolution multiplier, which scale the number of filters in each layer and the resolution of the input image, respectively.

The Vision Transformer (ViT) represents a major shift in the approach to computer vision. Dosovitskiy by adapting the Transformer-based architecture, originally designed for (Dosovitskiy et al., 2020). Natural Language Processing (NLP) to visual tasks, represents the most modern approach to image classification and other related visual applications by using self-attention mechanisms instead of convolutional operations. With this architectural change, ViT can capture long-range and global contextual relationships more effectively than traditional convolutional neural networks. Analogous to tokenized words in NLP models, the Vision Transformer processes images by transforming them into sequences of patches. This is achieved by several key steps. Instead of processing the image as a whole, it is divided into fixed-size patches, each of which is projected linearly into a lower-dimensional embedding space, creating a sequence of embeddings. Since transformers do not inherently encode spatial information, positional embeddings are added to the patch embeddings to preserve the spatial structure of the image. The self-attention mechanism allows the model to simultaneously attend to the other parts of the image, capturing local and global dependencies. Multi-head attention improves the model's ability to learn various features from different regions of the image. The sequence of patch embeddings passes through multiple transformer encoder layers, each consisting of multi-head self-attention and feedforward neural networks. The final layer consists of a multi-layer perceptron (MLP) head that processes the encoded representation for classification or other related vision tasks. ViTs can model long-range dependencies as a whole image, improving feature extraction, unlike CNNs that rely on local receptive fields. ViTs can significantly outperform conventional CNNs when trained on large datasets, benefiting from their ability to generalize better. ViTs are inherently parallelized, allowing for more efficient computation on modern hardware such as GPUs or TPUs. Despite their impressive capabilities, ViTs present some challenges. They require large amounts of training data to generalize effectively, unlike CNNs, which can perform well on small datasets. The self-attention mechanism scales quadratically with image resolution, making ViTs computationally expensive compared to CNNs, which use weight sharing for efficiency. Vision transformers have demonstrated significant success in various computer vision tasks, including image classification, autonomous vehicles, segmentation, medical imaging, and object recognition. ViTs often require more time to train and more advanced optimization techniques that add cost. By using benchmark datasets such as ImageNet and adopting object detection frameworks, they achieve state-of-the-art accuracy and provide better precision compared to CNN-based models. They are used to detect anomalies in medical scans by using global feature extraction and for self-driving car systems by improving scene understanding. Introduction Vision Transformers have revolutionized the field of computer vision by providing an alternative to CNN-based architectures. By using self-attention mechanisms and a token-based processing approach, ViTs have achieved unrivaled efficiency in the field. However, their widespread adoption and usability depend on solving problems related to data requirements and computational efficiency. As research on this technology continues, advances in hybrid models, optimized training techniques, and efficient architectures will further strengthen its role in the future.

3.3. Experiments

3.3.1. COVID-19, pneumonia and normal cases identification

In the first experiments, the ResNet-50, DenseNet-121 and ViT models are used to detect COVID-19, pneumonia and normal cases from the first dataset mentioned above. All models were trained using similar

methods and parameters. We use speaker-independent training splits, ensuring no overlap between speakers in the training (80% of the data) and validation (20% of the data) subsets. The proposed model is trained end-to-end using a combination of categorical cross-entropy and a modality-attention entropy regularizer, given by Eq.1.

$$\mathcal{L} = -\sum_{C=1}^{C} y_c \log(\hat{y}_c) + \lambda \sum_{m=1}^{3} \alpha_m \log(\alpha_m) \quad (1)$$

where: y_c and $\hat{y}_c$ – denote the ground truth and predicted probability for class c, respectively, α_m – represents the attention weights across m=1,2,3 modalities, λ – is a regularization coefficient that controls the influence of the entropy term. The second term (entropy regularization) encourages sparsity or interpretability in modality attention weights during cross-modal fusion.

Model optimization is performed using the Adam optimizer with an initial learning rate of $\mathbf{10^{-4}}$, exponential decay scheduling, batch size of 32, maximum 100 epochs early stopping based on validation accuracy. In order to avoid overfitting, standard data enhancement techniques are applied, which include time-shifting noise addition and random masking in the spectrogram domain. To minimize the influence of randomness, ten independent tests are performed. In addition, ensemble learning is used in this study. Each base model in soft voting (weighted ensemble) generates a probability distribution over all possible classes as an output, instead of just producing a class label (Fig. 4). After averaging these probabilities across all models, the class with the highest mean probability is selected as the final prediction. All base classifiers used in this technique must be able to produce probabilistic outputs, such as those from the Softmax layers of deep neural networks. One of the key advantages of soft voting is that it takes into account the confidence of each model's prediction. Soft voting allows this confidence to have a greater impact on the outcome, for example, if one model has a high degree of prediction confidence while others are uncertain. This results in increased stability and often improved performance, especially when there are notable differences in accuracy or behavior between the base models.

The weights used in the ensemble model prediction were taken from the final probability of a given class for each of the three models. It was manually set to determine which method should be more important in a given class, which is a form of tuning. The final average probability is the arithmetic mean of the probabilities of the individual models. The ensemble itself did not have much impact on the higher computational cost, except that the three models had to be trained first, so it took three times as long to generate the ensemble. During prediction, it had to wait for the three models to predict, which took a little longer than usual, but only a few seconds.

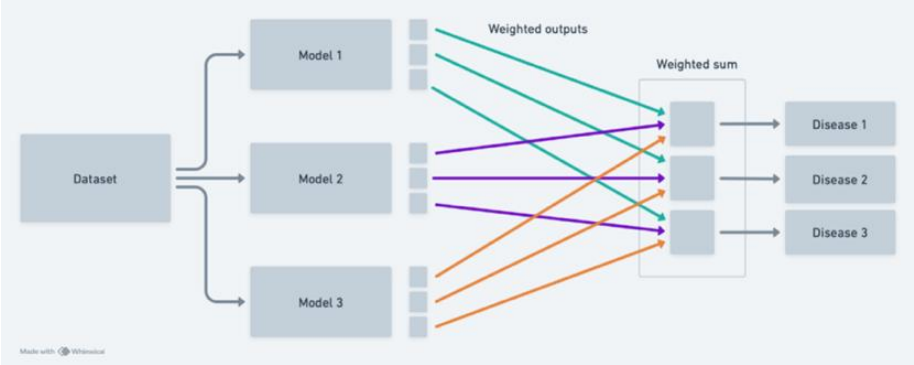


Fig. 4. Diagram of the weighted ensemble learning method

3.3.2. Bacterial pneumonia, viral pneumonia, tuberculosis, COVID-19 and healthy cases identification

In the second experiments, the model development is based on pre-trained CNN architectures including ResNet-50, VGG16, DenseNet-121 and MobileNet for the detection of bacterial pneumonia, viral pneumonia, tuberculosis, COVID-19 as well as healthy cases. The original classification layers of the above models are removed and replaced with custom classification layers. These include a Global Average Pooling layer to reduce dimensionality, fully connected Dense layers with ReLU activation to learn complex patterns, Batch Normalization layers to stabilize training, Dropout layers to reduce overfitting, and a final Softmax layer to generate probability distributions over the target classes. Transfer learning is used by freezing the weights of

the convolutional base during the initial training phase, using the generalized features learned from large datasets such as ImageNet. The model is built using the Adam optimizer and categorical cross-entropy loss functions, which is suitable for multi-class classification. Training was done in two stages. The first stage consisted of 50 epochs where the base model was frozen and only the added layers were trained. In the second stage, called fine-tuning, the last 20 layers of the base model were unfrozen and trained at a reduced learning rate. This was done to allow the model to better adapt to the task while retaining the pre-trained knowledge. Callbacks have been added: early stopping, model checkpointing, and adaptive learning rate reduction. These prevent overfitting and improve learning efficiency. After training, the models were tested on a test dataset. The test was repeated 10 times to ensure model stability and robustness. The interpretability of the model was considered. It was tested using the Grad-CAM technique, which showed which regions of the image had the greatest impact on the models' decisions.

Weighted ensemble learning was used to improve classification performance by aggregating the predictions of multiple deep learning models. The approach began with the preparation of the computational environment and data according to the configuration used to train each model. Dedicated test data generators were prepared to apply model-specific preprocessing steps to ensure compatibility with the input requirements of each architecture. Four pre-trained convolutional neural networks, VGG16, DenseNet-121, ResNet-50, and MobileNet, were used as base models. Each model made predictions on the same test set. Their individual classification accuracies were calculated and used to assign weights, with higher performing models contributing more to the final ensemble output. Each model made predictions on the same test set. Their individual accuracies were calculated and each was given a weight. Models with higher scores will contribute more to the final ensemble learning result, while those with lower scores will contribute less.

3.4. Metrics

The performance achieved by the models in two experiments is monitored using the following metrics: accuracy, precision, sensitivity, and F1-score (Eq. 2-5). These measures were chosen because they provide a comprehensive assessment of the effectiveness of the models.

Accuracy measures the overall correctness of the model, taking into account true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). It measures the proportion of correct predictions out of all predictions. It is useful for balanced data sets (equation 2).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{2}$$

Where TP is a case where a sick patient is correctly diagnosed, TN is a case where a healthy patient is correctly diagnosed, FP is a case where a healthy patient is misdiagnosed as sick, and finally FN is a case where the patient is reported as healthy.

Precision finds false positives (FP). These are situations where a healthy patient is mistakenly thought to be sick. Such cases lead to unnecessary medical procedures and additional costs of treating the misdiagnosed patient (Eq. 3).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3}$$

Recall identifies false negatives (FN). It highlights dangerous cases where a sick patient has been misdiagnosed as a healthy person. This is an important measure because lack of accuracy in this aspect can lead to missed diagnosis or inappropriate treatment (Eq. 4).

$$\text{Recall} = \frac{TP}{TP+FN} \tag{4}$$

The F1-score is the harmonic mean of precision and sensitivity. It takes into account the effect of false positives (FP) and false negatives (FN) (equation 5). It is an important measure because it balances two types of errors: missed disease (FN) and misdiagnosis (FP).

$$\text{F1} - \text{score} = 2 \cdot \frac{Precision \cdot \text{Recall}}{Precision+Recall} \tag{5}$$

Fβ-score is a generalized version of the classic F1-score, which allows more weight to be given to precision or recall, depending on the specific application (Eq. 6). While β is greater than one, the focus is on recall if less precision. The score results will be the same as the classic F1-score while β is equal to one.

$$F_\beta \text{ score } = (1+\beta^2)\cdot\frac{Precision\cdot \text{Recall}}{\beta^2\cdot Precision+Recall} \quad (6)$$

4. RESULTS AND DISCUSSION

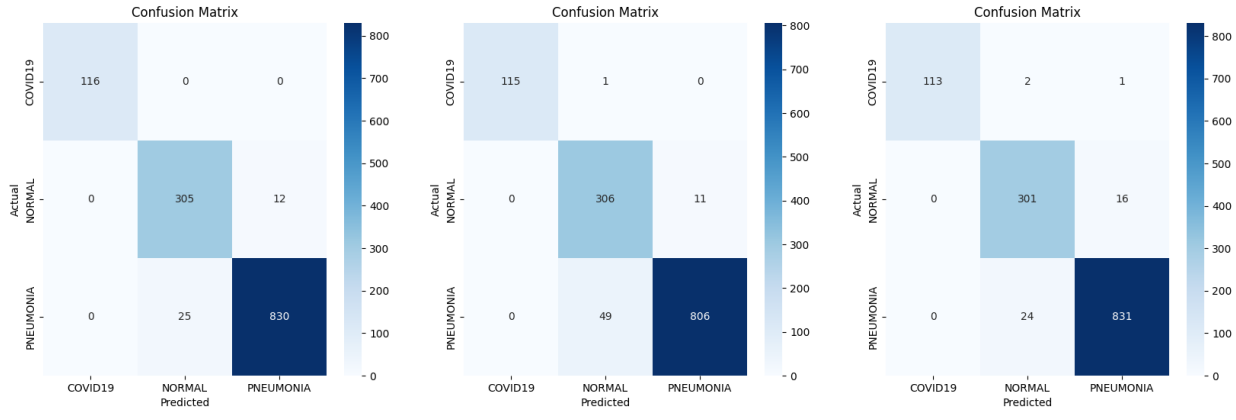
4.1. COVID-19, pneumonia and normal cases detection

The accuracy, precision, recall and F1-score for the detection of COVID-19, pneumonia and normal cases are shown in Tab. 3.

Tab. 3. Comparison of the performance of ResNet-50, DenseNet-121, ViT, and Ensemble models

Model	Accuracy [%]	Precision [%]	Recall [%]	Fβ = 1 [%]	Fβ = 2 [%]	Fβ = 0.5 [%]
ResNet-50	97.13	97.00	97.76	97.38	97.61	97.15
DenseNet-121	95.26	94.87	96.65	95.75	96.29	95.22
ViT	96.66	96.68	96.52	96.60	96.55	96.65
Ensemble	96.19	95.72	96.96	96.34	96.71	95.97

The confusion matrices for ResNet-50, DenseNet-121, and Vision Transformer (ViT), highlighting the classification performance of each model on COVID-19, normal, and pneumonia cases, are shown in Figure 5. ResNet-50 (left) shows excellent performance, correctly identifying 100% of COVID-19 cases and 96.2% of normal cases. It also classifies pneumonia with 97.1% accuracy, with little confusion between normal and pneumonia. DenseNet-121 (middle) also performs well, with 99.1% accuracy for COVID-19 and 96.5% for normal cases. However, it shows slightly lower accuracy for pneumonia (94.3%), with more cases misclassified as normal. Vision Transformer (right) maintains high performance across all classes. It correctly identifies 97.2% of pneumonia cases and 97.4% of COVID-19 cases. While normal classification drops slightly to 95.0%, overall performance remains competitive.



a) b) c)

Fig. 5. Confusion matrices of a) ResNet-50; b) DenseNet-121; c) Vision Transformer

The three models, ResNet-50, DenseNet-121, and Vision Transformer, showed strong performance, achieving test accuracies consistently above 96%, with the best model approaching 97.8%, indicating reliable generalization (Fig. 6).

Fig. 6. Train and test accuracy per epoch of a) ResNet-50; b) DenseNet-121 c) Vision Transformer

The Grad-CAM visualizations, shown in Figures 7-9, provide valuable insight into the decision-making processes of the models. All three architectures show clear attention to lung regions, confirming their focus on medically relevant areas. The ResNet-50 and DenseNet-121 produce concentrated heat maps around areas typically affected by pneumonia and COVID-19, such as localized opacities or infiltrates. These CNN-based models tend to capture strong, localized features due to their limited receptive fields and hierarchical structure. In contrast, ViT shows broader and more diffuse attention maps. This is expected due to its self-attention mechanism, which allows it to capture long-range dependencies and consider global context. As a result, the ViT tends to analyze the entire chest region, potentially picking up subtle, distributed patterns that CNNs may miss. This holistic view may contribute to its strong generalization performance, with test accuracy reaching nearly 97.8%. Overall, the Grad-CAM results confirm that all models attend to meaningful lung regions, with ViT providing a more global perspective that complements the localized focus of the CNNs.

Fig. 7. Grad-CAM for ResNet50 model

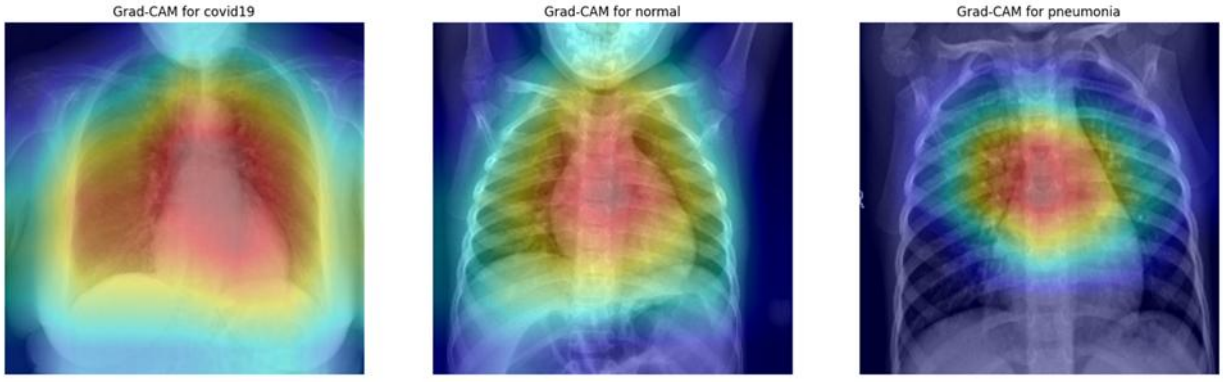


Fig. 8. Grad-CAM for DenseNet-121 model

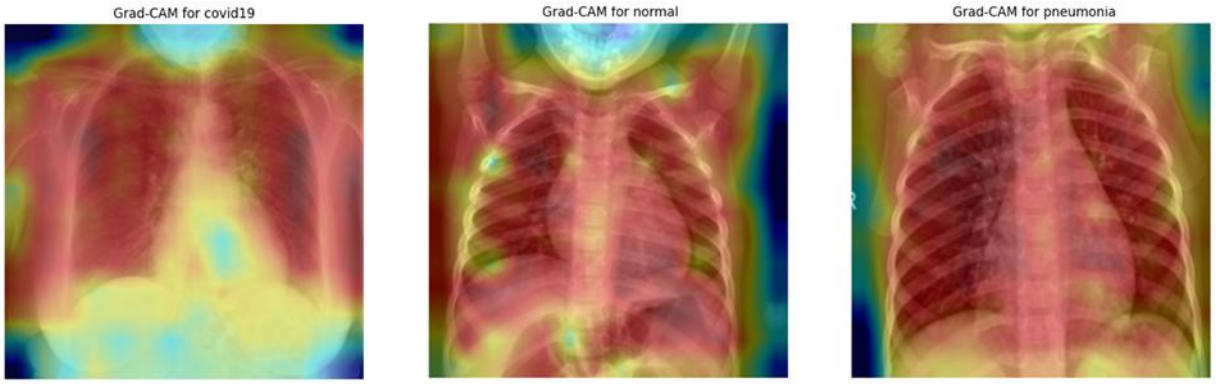


Fig. 9. Grad-CAM for ViT model

The ensemble model, which combines the ResNet-50, DenseNet-121 and ViT models via soft voting, shows a strong and balanced performance across all classes (Fig. 10). The model correctly classified 114 out of 116 COVID-19 cases (98.3%), misclassifying only 2 as normal. For the normal class, 308 of 317 samples (97.2%) were correctly identified, with 9 misclassified as pneumonia. In the pneumonia class, the model achieved 816 correct predictions out of 855 (95.4%), with 39 misclassified as normal. Compared to individual models, the ensemble approach reduces misclassification, especially between normal and pneumonia, where there is often overlap. By leveraging the local feature sensitivity of CNNs and the global attention of ViT, the ensemble achieves higher robustness and generalization. This synergy improves both overall accuracy and class-specific reliability, making the ensemble more suitable for high-stakes medical diagnostics.

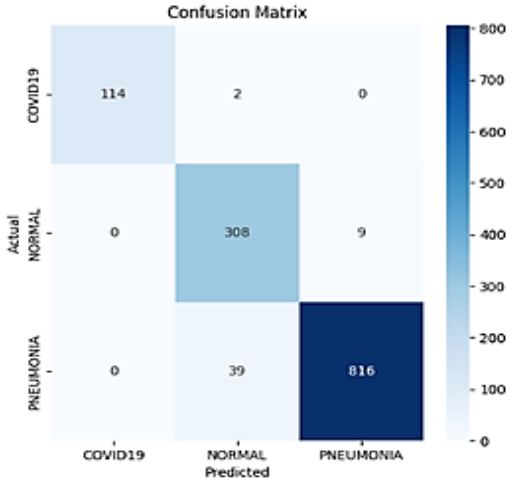


Fig. 10. Confusion matrix for ensemble soft voting model

4.2. Bacterial pneumonia, viral pneumonia, tuberculosis, COVID-19 and healthy cases detection

To evaluate the effectiveness of deep learning architectures in classifying lung diseases based on chest X-ray images, a comprehensive evaluation was performed on four models: VGG16, DenseNet-121, ResNet-50, and MobileNet. The overall performance evaluation of the models is summarized in Table 4. It is clear that the VGG16 model achieved the highest scores in all evaluation metrics: accuracy of 89.23%, precision of 89.01%, recall of 89.23%, and F1-score of 89.03%. In contrast, the ResNet-50 model achieved the lowest performance with an accuracy of 85.31%.

Tab. 4. Comparison of the performance of VGG16, DenseNet-121, ResNet-50, and MobileNet models

Model	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]	AVG (Accuracy, F1-score) [%]
VGG16	89.23	89.01	89.23	89.03	89.13
DenseNet-121	87.16	87.17	87.16	86.88	87.02
ResNet-50	85.31	86.30	85.31	84.69	85.00
MobileNet	89.01	88.98	89.01	88.87	88.94

The bootstrapping method was applied by performing 10 independent evaluation replicates for each model. This was done to assess the stability of the results. Analysis of the 95% confidence intervals confirmed the high stability of all models, with narrow confidence intervals of approximately 2-3 percentage points (Tab. 5).

Tab. 5. 95% Confidence intervals for the performance metrics of the evaluated models

Model	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]
VGG16	87.60 - 90.27	87.63 - 90.42	87.56 - 90.32	87.31 - 90.29
DenseNet-121	86.83 - 88.03	86.74 - 87.60	86.23 - 88.04	86.03 - 87.23
ResNet-50	84.40 - 87.36	84.44 - 87.30	84.44 - 87.31	83.82 - 86.01
MobileNet	87.56 - 90.27	87.23 - 90.27	87.51 - 90.17	87.35 - 90.16

The confusion matrices illustrating the classification results for five lung health conditions: bacterial pneumonia, viral pneumonia, COVID-19, tuberculosis, and healthy lung are shown in Figure 11. The diagonal values in each matrix represent correct classifications. The models perform best in identifying tuberculosis, COVID-19 and healthy lung. In contrast, bacterial pneumonia is often mistaken for viral pneumonia and vice versa. Tuberculosis and COVID-19 have high specificity and are rarely misclassified as other diseases. The models are reliable in identifying lung diseases (Fig. 12).

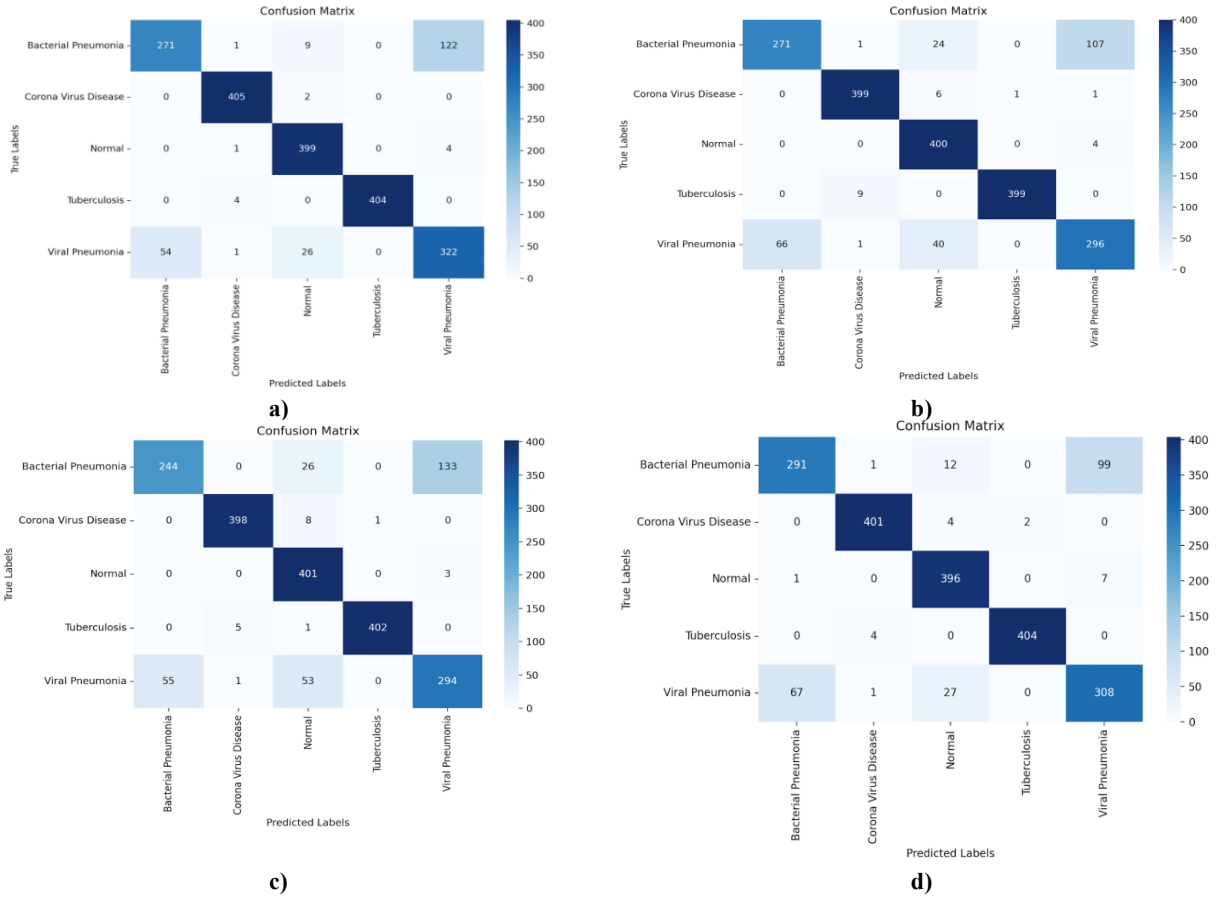


Fig. 11. Confusion matrices of five lung conditions for a)VGG16; b) DenseNet-121; c) ResNet-50; d) MobileNet models

An important aspect is to investigate whether combining multiple models using a weighted ensemble learning approach could improve the classification performance. In this study, the ensemble learning is proposed with different merging of the single models (Tab. 6). Among all tested combinations, the highest performance was achieved by the pair VGG16 + MobileNet, which reached an average accuracy and F1-score of 90.21%. This result is more than one percentage point higher than the best single model, VGG16. On the other hand, the lowest performance was observed for the combination of DenseNet-121 and ResNet-50 with an average score of 89.92%. The best three-model combination was VGG16, DenseNet-121 and MobileNet, with an average accuracy and F1-score of 90.16%. This is the second best result among all configurations tested. In contrast, the lowest performance among three model combinations was recorded for VGG16, DenseNet-121, and ResNet-50 with a score of 88.43%. Combining all four models - VGG16, DenseNet-121, ResNet-50, and MobileNet - resulted in an average accuracy and F1 score of 89.85%. While this is an improvement over individual models, it does not outperform the best performing two- and three-model ensembles.

The Grad-CAM visualizations use color-coded heatmaps to highlight the areas of the image where the model focused its attention. Regions marked in dark blue have the least influence on the model's prediction. This means that they were not considered in the model's decision making process. Green, yellow, and orange areas represent moderate influence - regions that contributed to the prediction but are not critical. Red areas indicate the highest model activation. These regions have the greatest impact on classification. All models: VGG16, DenseNet-121, ResNet-50, and MobileNet correctly identified the disease of the patient shown in Figure 13 as COVID-19.

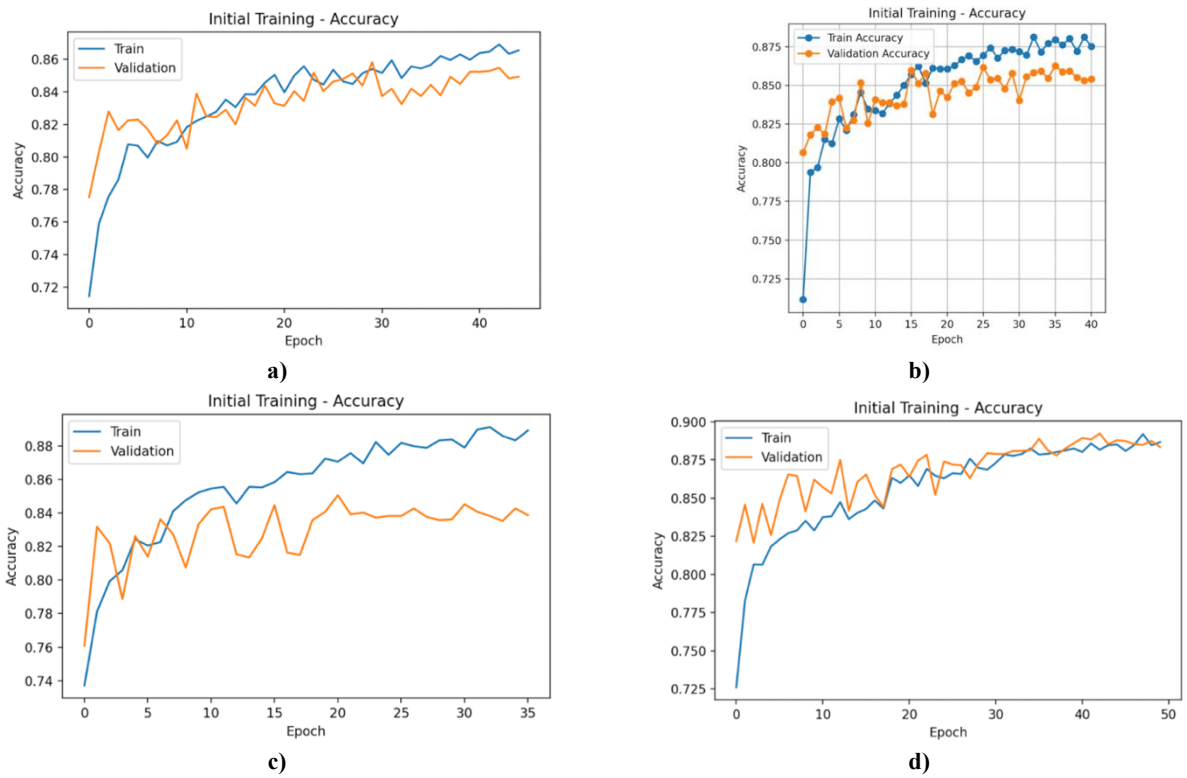


Fig. 12. Train and validation accuracy for a) VGG16; b) DenseNet-121; c) ResNet-50; d) MobileNet models

The VGG16 model shows bilateral involvement of the lower lung, which is consistent with the typical radiologic manifestation of COVID-19. This virus is characterized by bilateral and multifocal activations. The heat map clearly shows red areas in both lower lobes. The detection is accurate and anatomically correct. The DenseNet-121 model focuses primarily on the left lung. The model shows some interest in the right lobe. The ResNet-50 model focuses mainly on the left lower lobe and pays little attention to the right lobe. This is a less complete interpretation than the VGG16 model, although it still identifies significant areas. The MobileNet model shows the greatest activation in the central-upper portion of the lung. It marks a very generalized lung area. The wide area of activation may lead to lower specificity.

Tab. 6. Comparison of the results for all model combinations in weighted ensemble learning

Ensemble	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]	AVG (Accuracy, F1-score) [%]
VGG16+MobileNet	90.28	90.08	90.28	90.13	90.21
ResNet-50+MobileNet	90.08	89.90	90.08	89.89	89.99
VGG16+DenseNet-121	89.53	89.34	89.53	89.35	89.44
DenseNet-121+MobileNet	89.49	89.37	89.49	89.31	89.40
VGG16+ResNet-50	88.05	87.86	88.05	87.72	87.89
DenseNet-121+ResNet-50	87.07	86.80	87.07	86.76	86.92
VGG16+DenseNet-121+MobileNet	90.23	90.06	90.23	90.09	90.16
VGG16+ResNet-50+MobileNet	89.93	89.73	89.93	89.74	89.84
DenseNet-121+ResNet-50+MobileNet	89.49	89.30	89.49	89.30	89.40
VGG16+DenseNet-121+ ResNet-50	88.55	88.31	88.55	88.31	88.43
VGG16+DenseNet-121+ ResNet-50+MobileNet	89.93	89.74	89.93	89.76	89.85

The VGG16, DenseNet-121 and MobileNet models correctly identified the bacterial pneumonia disease shown in Figure 13. The heatmap of the VGG16 model shows that the model is primarily focused on the left lung lobe. It also shows intense activation in the mediastinal region. The area of activation predominantly covers anatomically relevant regions and shows unilateral changes in the lung, consistent with typical bacterial pneumonia. The DenseNet-121 model shows exceptionally localized activation covering a portion of the right lung and the border of the image, with minimal activation in the left lung lobe. The ResNet-50 shows a broad area of activation with intense focus on the left side of the image. The model's attention extends to a portion of the structure outside the lung region and to peripheral areas of the image. In addition, it focuses on the upper lung field on the right side. The model also highlights elements of medical equipment or image artifacts that are not directly related to lung pathology. The MobileNet model generates a heatmap with a clear concentration in the lower part of the left lung field and its outer region. There is also a weaker activation in the interlobar space.

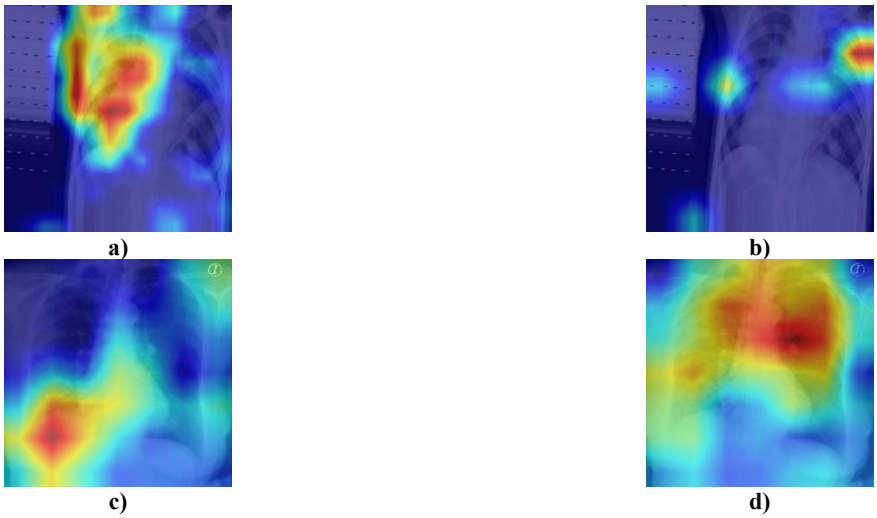


Fig. 13. Visualisation of the Grad-CAM method using an example of an X-ray showing the lungs of a bacterial pneumonia for a) VGG16; b) DenseNet-121; the lungs of a COVID-19 patient for c) ResNet-50; d) MobileNet models

The obtained results are consistent with the results of studies found in the literature. VGG16, DenseNet-121, ResNet-50 and MobileNet models pre-trained on ImageNet are often chosen for disease classification, including COVID-19 (Umair et al., 2021). They achieved very high accuracy in disease detection. The MobileNet and DenseNet-121 architectures proved to be the most suitable for this task. The MobileNet, DenseNet-121, and ResNet-50 models were used to detect lung disease based on X-ray images (Sriporn et al.,

2020). The DenseNet-121 architecture was the best model in terms of accuracy. However, ResNet-50 and MobileNet were slightly worse, by only 5.6 percentage points.

Ensemble learning has been widely used to detect lung disease from X-ray images. The combination of DenseNet-169, ResNet-50, and VGG16 models proved to be highly effective, achieving 99% accuracy (Rajpoot et al., 2024). The study was conducted on the COVIDx and CXR-3 datasets. In Akter et al. (2023) discussed stacked ensemble learning, which combines the results of pre-trained models to improve classification accuracy. The following configurations have been used: Inceptionv3-Inceptionv3, DenseNet-Mobilenet, Inceptionv3-Xception, Resnet-50-VGG16, and a stack combining all six baseline models. The best result was obtained for the stacked DenseNet-121 and MobileNet model, which achieved an accuracy of 78%. This result is lower than the accuracy of the DenseNet-121 and MobileNet models individually. The ResNet-50 and VGG16 ensemble learning models also performed below the accuracy of these models individually.

5. CONCLUSIONS

This study focuses on the detection of the most common lung diseases using deep learning models and ensemble learning. Two experiments were proposed for different types of datasets. The obtained results confirm the effectiveness of convolutional neural networks (CNNs) in classifying lung diseases from chest X-rays. In the first experiments, the highest accuracy was obtained for the ResNet for single model recognition. In the case of soft voting, the ensemble approach reduces misclassification, especially between normal and pneumonia classes.

In the second experiment, the highest accuracy results and F1-scores were obtained by the VGG16 model. This means that this model has a high ability to generalize different classes. This performance can be attributed to its relatively deep yet stable architecture and its suitability for medical imaging tasks. MobileNet also performed competitively. It offered a favorable balance between accuracy and computational efficiency, which may be particularly beneficial in real-time or resource-constrained clinical settings.

In contrast, the ResNet-50 model received the lowest performance in all evaluation metrics. The Grad-CAM visualization was used to verify which areas of the image the model was focusing on. It was found that the model often focused on irrelevant areas of the image, including areas outside the lung and medical device artifacts. Inaccurate areas of interest in an image can lead to inaccurate predictions, lower specificity, and an inability to trust the model's judgment. These observations underscore the importance of interpretive tools to identify model weaknesses beyond numerical performance metrics.

Confusion matrix analysis showed that all models consistently struggled to distinguish between bacterial and viral pneumonia, often confusing the two. However, the models were more accurate in identifying tuberculosis, COVID-19, and healthy lungs, which typically have more distinct radiological patterns.

Ensemble learning proved to be an effective way to improve classification performance. The weighted ensemble learning combination of VGG16 and MobileNet outperformed each of the individual models by more than one percentage point, achieving the best performance of all combinations. It was found that adding more models in the ensemble learning did not always lead to better performance. Ensemble learning of four models performed worse than the best combinations of two and three models. This means that the complementarity of the models and their individual accuracy play a greater role than the number of ensembles.

The Grad-CAM visualizations confirmed the high performance of the VGG16 model, showing that the model focuses its attention on anatomically relevant areas of the lung. The ability to see on which areas the models based their classification decisions is particularly important because it increases the transparency of the neural network's performance.

This study confirms that deep learning models have great potential to support radiological diagnosis of lung disease. However, further work is needed to ensure the robustness, generalizability and clinical applicability of such systems. The performance of other deep learning models should be verified for different pulmonary diseases detection. In addition, other aggregation methods using Chocquet integral as well as fuzzy logic should be considered (Karczmarek et al., 2022; Kiersztyn et al., 2021; Skublewska-Paszkowska et al., 2023). The study considering Generative Adversarial Network for dataset balancing is also highly desirable (Powroźnik et al., 2024).

Conflicts of interest

The authors declare no conflict of interest.

REFERENCES

Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2021). Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Applied Intelligence*, *51*, 854–864. https://doi.org/10.1007/s10489-020-01829-7

Abd Elaziz, M., Mabrouk, A., Dahou, A., & Chelloug, S. A. (2022). Medical image classification utilizing ensemble learning and levy flight-based honey badger algorithm on 6G-enabled internet of things. *Computational Intelligence and Neuroscience*, *2022*(1), 5830766. https://doi.org/10.1155/2022/5830766

Akter, M. S., Shahriar, H., Sneha, S., & Cuzzocrea, A. (2023). Multi-class skin cancer classification architecture based on deep convolutional neural network. *2022 IEEE International Conference on Big Data (Big Data)* (pp. 5404-5413). IEEE. https://doi.org/10.1109/BigData55660.2022.10020302

Balan, D. A., Paul, V., Nair, A. N., N, B. P., & G, R. M. L. (2024). Ensemble learning model by feature fusion from VGG16 and Resnet50 with attention mechanism for pneumonia classification. *SSRN*. https://doi.org/10.2139/ssrn.4853011

Bhosale, Y. H., Patnaik, K. S., Zanwar, S. R., Singh, S. Kr., Singh, V., & Shinde, U. B. (2024). Thoracic-net: Explainable artificial intelligence (XAI) based few shots learning feature fusion technique for multi-classifying thoracic diseases using medical imaging. *Multimedia Tools and Applications*, *84*, 5397–5433. https://doi.org/10.1007/s11042-024-20327-3

Castillo-Barnes, D., Martinez-Murcia, F. J., Jimenez-Mesa, C., Arco, J. E., Salas-Gonzalez, D., Ramírez, J., & Górriz, J. M. (2023). Nonlinear weighting ensemble learning model to diagnose Parkinson's disease using multimodal data. *International Journal of Neural Systems*, *33*(08), 2350041. https://doi.org/10.1142/S0129065723500417

Chandra, T. B., Verma, K., Singh, B. K., Jain, D., & Netam, S. S. (2021). Coronavirus disease (COVID-19) detection in Chest X-Ray images using majority voting based classifier ensemble. *Expert Systems with Applications*, *165*, 113909. https://doi.org/10.1016/j.eswa.2020.113909

Cohen, J. P., Morrison, P., & Dao, L. (2020a). COVID-19 image data collection (Version 1). *ArXiv, abs/2003.11597*. https://doi.org/10.48550/ARXIV.2003.11597

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T., & Ghassem, M. (2020b). COVID-19 image data collection: Prospective predictions are the future. *Machine Learning for Biomedical Imaging*, *1*(December 2020), 1–38. https://doi.org/10.59275/j.melba.2020-48g7

Dalvi, O. M. D. (2022, March 26). *Lungs disease dataset (4 types)*. IEEE DataPort.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale (Version 2). *ArXiv, abs/2010.11929*. https://doi.org/10.48550/ARXIV.2010.11929

Fan, C.-L. (2023). Multiscale feature extraction by using convolutional neural network: Extraction of objects from multiresolution images of urban areas. *ISPRS International Journal of Geo-Information*, *13*(1), 5. https://doi.org/10.3390/ijgi13010005

Hadj Bouzid, A. I., Berrani, S.-A., Yahiaoui, S., Belaid, A., Belazzougui, D., Djouad, M., Bensalah, K., Belbachir, H., Naïli, Q., Abdi, M. E.-H., & Tliba, S. (2024). Deep learning-based Covid-19 diagnosis: A thorough assessment with a focus on generalization capabilities. *EURASIP Journal on Image and Video Processing*, *2024*, 40. https://doi.org/10.1186/s13640-024-00656-x

Hamza, A., Attique Khan, M., Wang, S.-H., Alhaisoni, M., Alharbi, M., Hussein, H. S., Alshazly, H., Kim, Y. J., & Cha, J. (2022). COVID-19 classification using chest X-ray images based on fusion-assisted deep Bayesian optimization and Grad-CAM visualization. *Frontiers in Public Health*, *10*, 1046296. https://doi.org/10.3389/fpubh.2022.1046296

Horry, M. J., Chakraborty, S., Paul, M., Ulhaq, A., Pradhan, B., Saha, M., & Shukla, N. (2020). COVID-19 detection through transfer learning using multimodal imaging data. *IEEE Access*, *8*, 149808–149824. https://doi.org/10.1109/ACCESS.2020.3016780

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications (Version 1). *ArXiv*, *abs/1704.04861*. https://doi.org/10.48550/ARXIV.1704.04861

Imam, Md. H., Nahar, N., Rahman, Md. A., & Rabbi, F. (2024). Enhancing skin cancer classification using a fusion of Densenet and Mobilenet models: A deep learning ensemble approach. *Multidisciplinary Science Journal*, *6*(7), 2024117. https://doi.org/10.31893/multiscience.2024117

Ismael, A. M., & Şengür, A. (2021). Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Systems with Applications*, *164*, 114054. https://doi.org/10.1016/j.eswa.2020.114054

Jain, G., Mittal, D., Thakur, D., & Mittal, M. K. (2020). A deep learning approach to detect Covid-19 coronavirus with X-Ray images. *Biocybernetics and Biomedical Engineering*, *40*(4), 1391–1405. https://doi.org/10.1016/j.bbe.2020.08.008

Jasmine Pemeena Priyadarsini, M., Kotecha, K., Rajini, G. K., Hariharan, K., Utkarsh Raj, K., Bhargav Ram, K., Indragandhi, V., Subramaniyaswamy, V., & Pandya, S. (2023). Lung diseases detection using various deep learning algorithms. *Journal of Healthcare Engineering*, *2023*(1), 3563696. https://doi.org/10.1155/2023/3563696

Jenber Belay, A., Walle, Y. M., & Haile, M. B. (2024). Deep Ensemble learning and quantum machine learning approach for Alzheimer's disease detection. *Scientific Reports*, *14*, 14196. https://doi.org/10.1038/s41598-024-61452-1

Jogin, M., Mohana, Madhulika, M. S., Divya, G. D., Meghana, R. K., & Apoorva, S. (2018). Feature extraction using Convolution Neural Networks (CNN) and deep learning. *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 2319–2323). IEEE. https://doi.org/10.1109/RTEICT42901.2018.9012507

Karczmarek, P., Dolecki, M., Powroznik, P., Galka, L., Pedrycz, W., & Czerwinski, D. (2022). Quadrature-inspired generalized choquet integral. *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1–7). IEEE. https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882684

Kesuma, L. I., Ermatita, & Erwin. (2023). ELREI: Ensemble learning of ResNet, EfficientNet, and Inception-v3 for lung disease classification based on chest X-Ray image. *International Journal of Intelligent Engineering and Systems*, *16*(5), 149–161. https://doi.org/10.22266/ijies2023.1031.14

Khan, A. I., Shah, J. L., & Bhat, M. M. (2020). CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. *Computer Methods and Programs in Biomedicine*, *196*, 105581. https://doi.org/10.1016/j.cmpb.2020.105581

Kiersztyn, A., Lopucki, R., Kiersztyn, K., Karczmarek, P., Powroznik, P., Czerwinski, D., & Pedrycz, W. (2021). A comprehensive analysis of the impact of selecting the training set elements on the correctness of classification for highly variable ecological data. *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1–6). IEEE. https://doi.org/10.1109/FUZZ45933.2021.9494399

Kong, L., & Cheng, J. (2022). Classification and detection of COVID-19 X-Ray images based on DenseNet and VGG16 feature fusion. *Biomedical Signal Processing and Control*, *77*, 103772. https://doi.org/10.1016/j.bspc.2022.103772

Kumaran S, Y., Jeya, J. J., R, M. T., Khan, S. B., Alzahrani, S., & Alojail, M. (2024). Explainable lung cancer classification with ensemble transfer learning of VGG16, Resnet50 and InceptionV3 using grad-cam. *BMC Medical Imaging*, *24*, 176. https://doi.org/10.1186/s12880-024-01345-x

Mabrouk, A., Díaz Redondo, R. P., Dahou, A., Abd Elaziz, M., & Kayed, M. (2022). Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks. *Applied Sciences*, *12*(13), 6448. https://doi.org/10.3390/app12136448

Marques, G., Agarwal, D., & De La Torre Díez, I. (2020). Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *Applied Soft Computing*, *96*, 106691. https://doi.org/10.1016/j.asoc.2020.106691

Marwa, A. S., & Mohammed, K. (2024). Lung infection detection via CT images and transfer learning techniques in deep learning. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *47*(1), 206–218. https://doi.org/10.37934/araset.47.1.206218

Moujahid, H., Cherradi, B., Al-Sarem, M., Bahatti, L., Bakr Assedik Mohammed Yahya Eljialy, A., Alsaeedi, A., & Saeed, F. (2022). Combining CNN and Grad-Cam for COVID-19 disease prediction and visual explanation. *Intelligent Automation & Soft Computing*, *32*(2), 723–745. https://doi.org/10.32604/iasc.2022.022179

Muntasir, F., Datta, A., & Mahmud, S. (2024). Interpreting multiclass lung cancer from CT scans using Grad-CAM on lightweight CNN layers. *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)* (pp. 208–213). IEEE. https://doi.org/10.1109/ICEEICT62016.2024.10534491

Naik, R., Wani, T., Bajaj, S., Ahir, S., & Joshi, A. (2020). Detection of lung diseases using deep learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3568730

Nayak, S. R., Nayak, D. R., Sinha, U., Arora, V., & Pachori, R. B. (2021). Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study. *Biomedical Signal Processing and Control*, *64*, 102365. https://doi.org/10.1016/j.bspc.2020.102365

Odeh, A. A.-R., & Mustafa, A. (2024). Explaining transfer learning models for the detection of COVID-19 on X-ray lung images. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(4), 4542. https://doi.org/10.11591/ijece.v14i4.pp4542-4550

Öztürk, Ş., Turalı, M. Y., & Çukur, T. (2025). HydraViT: Adaptive multi-branch transformer for multi-label disease classification from Chest X-ray images. *Biomedical Signal Processing and Control*, *100*(Part A), 106959. https://doi.org/10.1016/j.bspc.2024.106959

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Rajendra Acharya, U. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*, *121*, 103792. https://doi.org/10.1016/j.compbiomed.2020.103792

Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. (2020). A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. *Chaos, Solitons & Fractals*, *140*, 110190. https://doi.org/10.1016/j.chaos.2020.110190

Patel, P. (2019). *Chest X-ray (Covid-19 & Pneumonia)*. Kaggle.

Pereira, R. M., Bertolini, D., Teixeira, L. O., Silla, C. N., & Costa, Y. M. G. (2020). COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. *Computer Methods and Programs in Biomedicine*, *194*, 105532. https://doi.org/10.1016/j.cmpb.2020.105532

Powroźnik, P., Skublewska-Paszkowska, M., Nowomiejska, K., Aristidou, A., Panayides, A., & Rejdak, R. (2024). Deep convolutional generative adversarial networks in retinitis pigmentosa disease images augmentation and detection. *Advances in Science and Technology Research Journal*, *19*(2), 321–340. https://doi.org/10.12913/22998624/196179

Rahaman, M. M., Li, C., Yao, Y., Kulwa, F., Rahman, M. A., Wang, Q., Qi, S., Kong, F., Zhu, X., & Zhao, X. (2020). Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. *Journal of X-Ray Science and Technology: Clinical Applications of Diagnosis and Therapeutics*, *28*(5), 821–839. https://doi.org/10.3233/XST-200715

Rajaraman, S., Siegelman, J., Alderson, P. O., Folio, L. S., Folio, L. R., & Antani, S. K. (2020). Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-Rays. *IEEE Access*, *8*, 115041–115050. https://doi.org/10.1109/ACCESS.2020.3003810

Rajpoot, R., Gour, M., Jain, S., & Semwal, V. B. (2024). Integrated ensemble CNN and explainable AI for COVID-19 diagnosis from CT scan and X-ray images. *Scientific Reports*, *14*, 24985. https://doi.org/10.1038/s41598-024-75915-y

Sharma, V., Nillmani, Gupta, S. K., & Shukla, K. K. (2024). Deep learning models for tuberculosis detection and infected region visualization in chest X-ray images. *Intelligent Medicine*, *4*(2), 104–113. https://doi.org/10.1016/j.imed.2023.06.001

Shaziya, H. (2022). Explainable deep learning through Grad-CAM and feature visualization for the detection of COVID-19 in chest X-ray images. In M. A. Chaurasia & S. Mozar (Eds.), *Contactless Healthcare Facilitation and Commodity Delivery Management During COVID 19 Pandemic* (pp. 27–34). Springer Singapore. https://doi.org/10.1007/978-981-16-5411-4_4

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition (Version 6). *ArXiv, abs/1409.1556*. https://doi.org/10.48550/ARXIV.1409.1556

Sitaula, C., & Hossain, M. B. (2021). Attention-based VGG-16 model for COVID-19 chest X-ray image classification. *Applied Intelligence*, *51*, 2850–2863. https://doi.org/10.1007/s10489-020-02055-x

Skublewska-Paszkowska, M., Karczmarek, P., Powroznik, P., Lukasik, E., Smolka, J., & Dolecki, M. (2023). Aggregation of tennis multivariate time-series using the choquet integral and its generalizations. *2023 IEEE International Conference on Fuzzy Systems (FUZZ)* (pp. 1–6). IEEE. https://doi.org/10.1109/FUZZ52849.2023.10309746

Skublewska-Paszkowska, M., Powroznik, P., Lukasik, E., & Smolka, J. (2024). Tennis patterns recognition based on a novel tennis dataset – 3DTennisDS. *Advances in Science and Technology Research Journal*, *18*(6), 159–176. https://doi.org/10.12913/22998624/191264

Sriporn, K., Tsai, C.-F., Tsai, C.-E., & Wang, P. (2020). Analyzing lung disease using highly effective deep learning techniques. *Healthcare*, *8*(2), 107. https://doi.org/10.3390/healthcare8020107

Tabik, S., Gomez-Rios, A., Martin-Rodriguez, J. L., Sevillano-Garcia, I., Rey-Area, M., Charte, D., Guirado, E., Suarez, J. L., Luengo, J., Valero-Gonzalez, M. A., Garcia-Villanova, P., Olmedo-Sanchez, E., & Herrera, F. (2020). COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on xhest X-Ray images. *IEEE Journal of Biomedical and Health Informatics*, *24*(12), 3595–3605. https://doi.org/10.1109/JBHI.2020.3037127

Taha Ahmed, S., & Malallah Kadhem, S. (2021). Using machine learning via deep learning algorithms to diagnose the lung disease based on chest imaging: A survey. *International Journal of Interactive Mobile Technologies*, *15*(16), 95. https://doi.org/10.3991/ijim.v15i16.24191

Toğaçar, M., Ergen, B., & Cömert, Z. (2020). COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. *Computers in Biology and Medicine*, *121*, 103805. https://doi.org/10.1016/j.compbiomed.2020.103805

Umair, M., Khan, M. S., Ahmed, F., Baothman, F., Alqahtani, F., Alian, M., & Ahmad, J. (2021). Detection of COVID-19 using transfer learning and Grad-CAM visualization on indigenously collected X-ray dataset. *Sensors*, *21*(17), 5813. https://doi.org/10.3390/s21175813

Vignesh Kumaran, N., & Preethi, D. M. D. (2025). Intelligent decision support system for lung cancer classification with ensemble inference system using fuzzy. *Biomedical Signal Processing and Control*, *108*, 107958. https://doi.org/10.1016/j.bspc.2025.107958

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., & Pinheiro, P. R. (2020). CovidGAN: Data augmentation using auxiliary classifier GAN for improved Covid-19 detection. *IEEE Access*, *8*, 91916–91923. https://doi.org/10.1109/ACCESS.2020.2994762

Yu, W., Zhou, H., Goldin, J. G., Wong, W. K., & Kim, G. H. J. (2021). End-to-end domain knowledge-assisted automatic diagnosis of idiopathic pulmonary fibrosis (IPF) using computed tomography (CT). *Medical Physics*, *48*(5), 2458–2467. https://doi.org/10.1002/mp.14754