Optimizing the compact convolution transformer for enhanced pneumonia detection

Muhammad Munsarif [a,1,*], Norshuhani Zamin [b,2], Muhammad Sam'an [a,3]

[a] Universitas Muhammadiyah Semarang, Jl. Kedungmundu Raya No.18, Semarang, Indonesia
[b] College of Computing and Information, Institute of Computing in Energy (IICE), Universiti Tenaga Nasional (UNITEN), Malaysia
[1] m.munsarif@unimus.ac.id; [2] norshuhani@uniten.edu.my; [3] muhammad92sam@unimus.ac.id
* corresponding author

ARTICLE INFO

Article history
Received October 19, 2024
Revised November 26, 2025
Accepted December 10, 2025
Available online May 31, 2026

Keywords
Pneumonia detection
Compact convolution transformer
CNN
CCT
X-Ray Image

ABSTRACT

Pneumonia detection through medical imaging, especially using CT scans or X-rays, presents notable challenges due to the subtle and often unclear signs of the disease. This paper introduces a novel neural network model, the Compact Convolutional Transformer (CCT), designed to address these challenges by optimizing detection accuracy. The CCT model incorporates configuration dropout in its convolutional layers to enhance both robustness and precision. Experiments conducted on a dataset of 5,856 chest X-ray images from pediatric patients aged one to five years demonstrated the model's effectiveness, achieving a remarkable 97% accuracy, 97% recall, 98% precision, and an F1-score of 98%. When compared to state-of-the-art models like DarkNet-53 and VGG-19 + GradCAM, which achieved F1-scores of 97.3% and 95.61% respectively, the CCT model consistently matched or outperformed them, particularly when dealing with smaller and more complex datasets. Even models such as CNN + Bayesian Network, which used larger datasets, only reached an F1-score of 96.3%. These results underscore the superior efficiency and accuracy of the CCT model, highlighting its potential for broader applications in medical diagnostics and image analysis, especially in pneumonia detection.

1. Introduction

Pneumonia remains one of the leading causes of death among children under five years of age, with a global mortality count of 725,557 cases. Although this number reflects a significant reduction of 54% compared to the 1,590,874 deaths recorded in 2000, the disease continues to pose a major public health concern [1]. Globally, the incidence of pneumonia exceeds 1,400 cases per 100,000 children, with the highest rates observed in South Asia (2,500 cases per 100,000) and West and Central Africa (1,620 cases per 100,000). UNICEF emphasizes the need for innovative approaches, including artificial intelligence, to improve early detection and clinical management of pneumonia.

Pneumonia is a respiratory infection that requires antibiotic treatment for bacterial cases, while viral pneumonia is managed through supportive care. Preventive strategies include proper hand hygiene, a balanced diet, regular physical activity, and vaccination. Diagnostic assessment often involves a chest X-ray or CT imaging, with X-ray preferred due to its lower cost and accessibility. Despite advancements in global healthcare, pneumonia remains a major challenge in regions with limited diagnostic resources. The growth of computer vision has significantly contributed to medical image analysis. Convolutional

neural networks have demonstrated strong capabilities in extracting complex radiological features, making them essential for automated pneumonia detection.

Previous research has reported strong performance across various Convolutional Neural Networks (CNN) architectures. VGG-16, VGG-19, ResNet50, and InceptionV3 achieved accuracies between 71% and 88% on a dataset of 624 X-ray images [2]. Dey achieved 98% accuracy, 95% precision, and an F-score of 96% using 1,650 images [3]. Brunese reached 96.2% accuracy on 6,523 images [4]. CSDB CNN achieved F1-scores between 94% and 98% [4]. DenseNet and CapsNet obtained 96% recall and 90.7% accuracy on pneumonia datasets [5]. DenseNet-201 consistently achieved 95% accuracy and 90% precision on 6,000 images [6]. Mask R-CNN, CovXNet, and residual CNNs also demonstrated strong results [7], [8], [9], [10]. Additional techniques, such as prior-attention learning and rank-based pooling, achieved high performance [11], [12], [13]. The RSNA dataset enabled an accuracy of 94.62% using VGG [14]. Ensemble models achieved accuracy between 94% and 99% on 16,700 images [15]. A pediatric-focused model reached 97.2% accuracy, 97.3% recall, 97.4% precision, and an AUC of 0.982 [16]. Vision Transformer models have also been used in pneumonia detection. One model achieved 96.45% accuracy [17], [18], while others reported around 94% accuracy [19]. PneuNet achieved 94.96% accuracy [11]. The Swin Transformer demonstrated strong multi-stage feature learning [20]. However, Vision Transformers rely heavily on large datasets and have limited inductive bias, which reduces their effectiveness on smaller medical datasets.

This study introduces the Compact Convolutional Transformer as an approach that combines convolution-based local feature extraction with global feature representation. The Compact Convolutional Transformer employs a convolutional tokenizer that provides strong inductive bias, making it more effective at detecting local radiological patterns such as opacity and infiltration. This study also presents a systematic analysis of five dropout levels, which has not been conducted in previous research on hybrid CNN–Transformer models. This study offers three primary contributions. First, it develops a Compact Convolutional Transformer optimized for pediatric pneumonia detection using convolutional tokenization. Second, it introduces a comprehensive evaluation of five dropout configurations to assess their influence on robustness, stability, and generalization. Third, it presents a computationally efficient model that achieves performance comparable to larger Vision Transformer models while requiring fewer computational resources. From a theoretical perspective, Compact Convolutional Transformer is advantageous for chest X-ray imaging because its convolutional tokenizer effectively extracts local radiological cues that are crucial for pneumonia identification. Vision Transformers lack this mechanism, which results in lower performance on small datasets. Compact Convolutional Transformer combines the strengths of convolutional and transformer layers to produce more stable and informative feature representations even when data availability is limited.

2. Method

This section outlines our approach to pneumonia detection through the analysis of X-ray images. The workflow depicted in Fig. 1 highlights the steps involved in the implemented pipeline. Following a description of the dataset used, we will provide a detailed explanation of each key component that forms the foundation of the proposed method.

2.1. Dataset

The dataset used in this study originates from a deep learning competition on Kaggle and contains 5,856 chest X-ray images of children aged one to five years [21], consisting of 4,273 pneumonia cases and 1,583 healthy lung images. All images were obtained from the Guangzhou Women and Children's Medical Center and verified by medical professionals as part of routine clinical assessment. This study employed data augmentation using a Generative Adversarial Network, specifically a Deep Convolutional Generative Adversarial Network, which was trained for 200 epochs with a batch size of 32 and produced 1,000 synthetic images for the minority class to address class imbalance. These synthetic samples were used exclusively during the training phase and were not included in the validation or test sets to ensure

that no synthetic data leaked into the evaluation process. Each generated image was manually inspected to prevent the inclusion of artifacts that could adversely affect model learning.

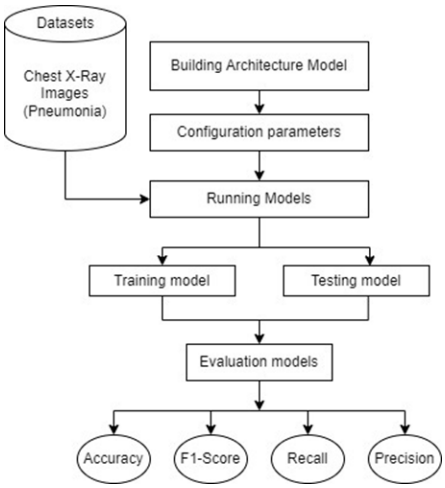


Fig. 1. The workflow of the proposed pipeline for detecting pneumonia.

The original X-ray images had resolutions ranging from 1346 × 1044 pixels to 2090 × 1858 pixels, and all images were resized to 224 × 224 × 3 to match the input requirements of the CNN architectures and the model used in this study. Sample images with their corresponding ground truth labels are shown in Fig. 2.

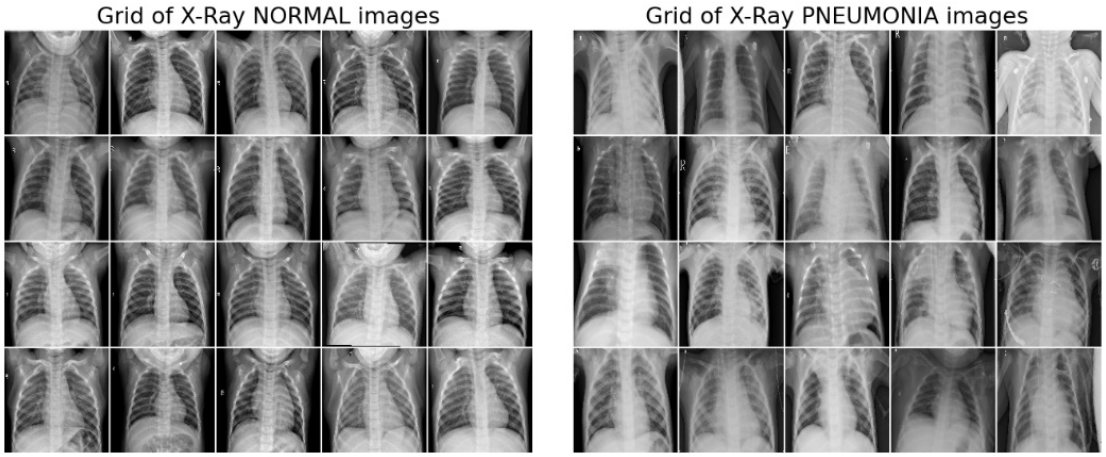


Fig. 2. Examples of input images and their ground truth are difficult to distinguish by the eyes of a non-expert.

2.2. The Architecture Used

2.2.1. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are widely favored in deep learning, particularly in the field of computer vision [22]. The core component of CNN-based architectures is convolution, which is a mathematical linear operation applied to matrices [23]. CNNs have demonstrated great success in pattern recognition, especially when working with visual data [24] (O'Shea & Nash. The pioneering work of Krizhevsky et al revolutionized computer vision [25], addressing challenges in tasks such as medical image analysis, face recognition, image classification, object detection, and semantic segmentation [26], [27], [28].

A CNN-based model typically includes convolutional, pooling, and fully connected layers, as shown in Fig. 3. This structure is designed to classify lung X-ray images into healthy or unhealthy samples. The convolution layer applies a kernel to the input data, producing feature maps for subsequent pooling operations. CNNs exhibit properties like translation equivariance and translational invariance, allowing them to understand the natural statistics of the input image. Integration of sparse interaction, weight

sharing, and equivariant representations enhances efficiency and reduces computational costs [29]. The advantage of CNN lies in its translation equivariance and translational invariance, enabling a deep understanding of natural image statistics while maintaining computational efficiency.

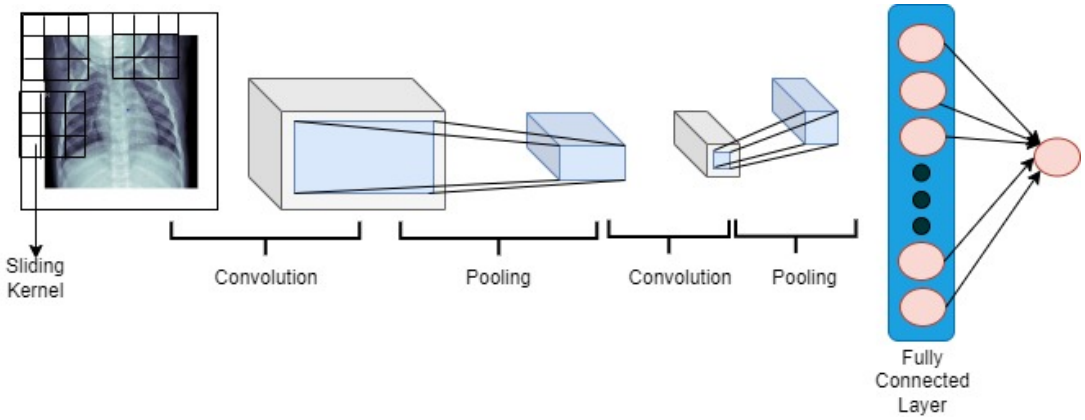


Fig. 3. The general architecture of a CNN-based model.

2.2.2. Vision transformer

Transformer-based models, initially designed for Natural Language Processing (NLP), have gained significant attention from computer vision researchers due to their remarkable performance across various NLP tasks like machine translation [30], question answering [31], text classification [32] and sentiment analysis [33], [34].Introducing Vision Transformers (ViTs) for image classification marked a creative application of transformers to visual data. Fig. 4, illustrates a general procedure in ViT-based models, where an image transforms patches, with each patch representing a specific region. This process enables interpreting an image as sequential data, a prevalent data type in NLP tailored for transformers.

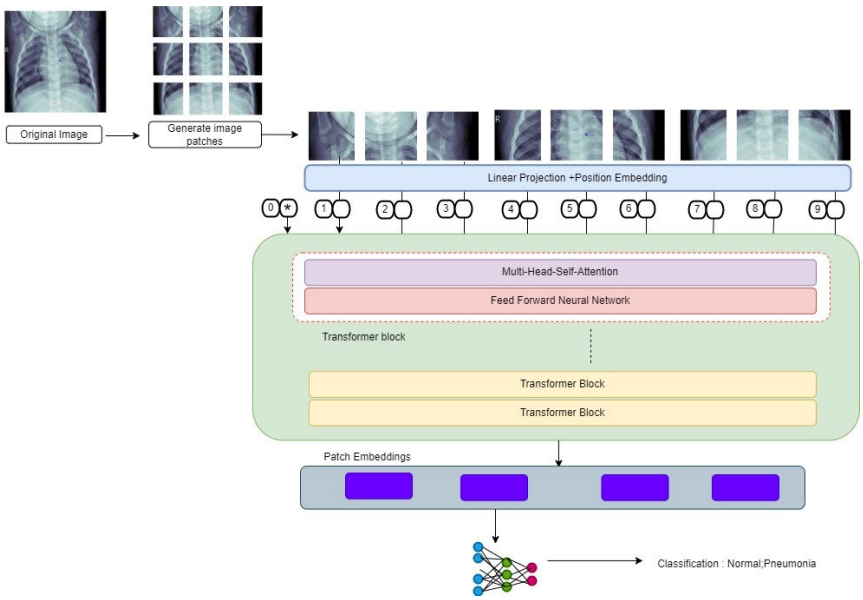


Fig. 4. The general architecture of a ViT-based model.

ViT flattens these patches and processes them through a trainable linear projection layer to standardize their dimensionalities. Since ViT does not inherently recognize the spatial hierarchy of the input image or the specific locations of each patch, position embeddings are introduced to resolve this limitation. These position embeddings ensure the model accounts for the spatial arrangement of patches in the original image. Next, the transformer encoder block integrates these patches, their positional information, and an additional classification token called the CLS token. The transformer encoder

incorporates multi-head attention layers capable of learning various self-attention states. Finally, the outputs from all existing heads are combined and fed into the Multi-Layer Perceptron (MLP).

2.2.3. Compact Convolutional Transformers as Proposed Method

This section introduces our proposed algorithm, which involves various stages in Compact Convolutional Transformers (CCT). Fig. 5, illustrates the overall structure of the CCT architecture. This approach synergistically combines elements from CNN and ViT, creating an optimal framework for image analysis.

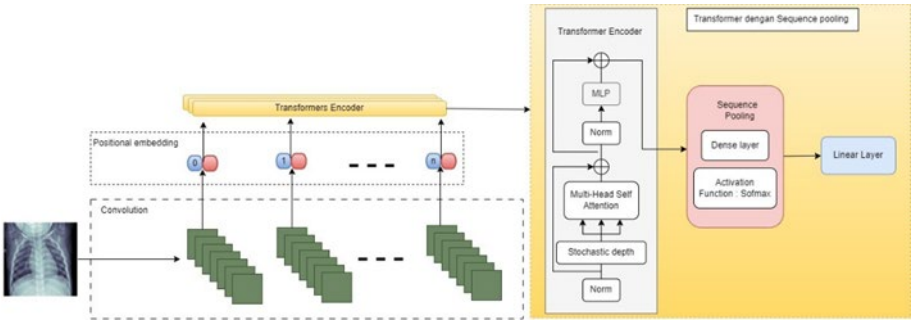


Fig. 5. An overview of the proposed architecture.

Compact Convolutional Transformer (CCT) is a modern and highly efficient transformer-based model tailored for image processing. Its main advantage is its ability to effectively learn from limited data, distinguishing it from the data-hungry nature of traditional Vision Transformer (ViT) models. ViT models often need help to outperform convolutional-based models when many samples are unavailable. Despite efforts by advanced models such as DeiT, ConViT, and Compact Vision Transformers to mitigate data scarcity issues in ViTs, CCT has consistently shown superior performance across various resolutions and benchmark datasets, including FashionMNIST, MNIST, CIFAR-10, CIFAR-100, ImageNet, and Flowers-102. CCT builds on the Compact Vision Transformers (CVT) framework, using a convolutional tokenizer to preserve local information and generate richer tokens, which has proven more effective in encoding connections between patches than standard ViTs. The following discussion will detail the components that make up compact transformers. The design of the CCT model is inspired by the original Vision Transformer and the Transformer [34], [35]. Its encoder consists of transformer blocks containing an MLP block and a Multi-Head Self-Attention (MHSA) layer. Fig. 5 illustrates the patchification, flattening, and linear projection of the input image patches, followed by the addition of positional embeddings. These embeddings are processed by multiple transformer encoders, which include normalization layers and the MHSA module. The encoder employs residual connections, dropout, GELU activation, and Layer Normalization. The design maintains the compactness and simplicity of vision transformers, with variants like ViT-Lite utilizing as few as two layers, two heads, and 128-dimensional hidden layers. Tokenizers are adapted according to image resolution, creating architectures comparable to ViT. The traditional approach in ViT and many transformer-based classifiers, influenced by BERT [36], typically uses a learnable class or query token that passes through the network before reaching the classifier, ultimately converting sequential outputs into a single class index. In contrast, CCT introduces an innovative attention-based technique that performs pooling over the output token sequence. This method, distinct from the learnable token, retains substantial information encompassing diverse aspects of the input image, resulting in enhanced efficiency. The approach enables the network to establish correlations across the input data and weigh the sequential embedding of the transformer encoder's latent space. Moreover, creating CVT involves substituting SeqPool with the conventional class token in ViT-Lite.

$$X_0 = MaxPool(ReLU(conv2d(x))) \quad (1)$$

For the final step in designing CCT, a simple convolutional block is used as a substitute for the patch and embedding in ViT-Lite and CVT. This block consists of a single convolution, a ReLU activation, and a max pool, providing the model with more flexibility than ViT. The model is no longer constrained

by the input resolution that must be divisible by the predetermined patch size, allowing CCT to be independent of image resolution. This is achieved through a convolutional tokenizer, whose representation is shown in Eq. 1, Sequence Pooling, and the encoder transformer. The feature map is extracted to represent local features. Eq. 1 indicates that CCT maintains the locality of information from the data through the convolutional block.

The model was trained using the Adam optimizer due to its adaptive capability in adjusting weight updates based on the first and second moments of the gradients, which provides greater stability when dealing with medical datasets that often exhibit high variability. The initial learning rate was set to 0.001 and gradually reduced through a step-based learning rate scheduling strategy to prevent oscillations during the convergence phase of optimization. A batch size of 32 was selected to maintain an optimal balance between GPU memory usage and gradient estimation stability throughout training. Although the training was configured for a maximum of 200 epochs, an early stopping mechanism was employed to prevent overfitting by continuously monitoring the validation loss. Training was automatically halted when no improvement in validation loss was observed for ten consecutive epochs, ensuring that the model learned only the consistent and meaningful patterns present in the data. The categorical cross-entropy function was used as the loss function because it mathematically quantifies the divergence between the predicted probability distribution and the actual class distribution, making it highly suitable for binary classification tasks such as pneumonia detection. In addition to dropout, L2 weight decay regularization was applied to discourage excessive weight magnitudes and reduce the likelihood of the model memorizing the training data, particularly given the relatively limited size of the pediatric X-ray dataset. This regularization contributed to a more stable training curve and improved generalization performance on unseen data. The combination of these hyperparameter configurations provides a robust foundation for effective learning, maintains training stability, minimizes the risk of overfitting, and ensures that other researchers can reliably replicate the training procedure.

2.3. Evaluation Performance

The proposed classifier's performance is evaluated using specific metrics, primarily focusing on the Confusion Matrix (CM). This matrix comprises four crucial components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In the case of a binary classifier, the arrangement and details of the CM can be observed as depicted in Fig. 6. The Confusion Matrix is used to calculate accuracy, recall, precision, and F1-score, which are formulated as follows:

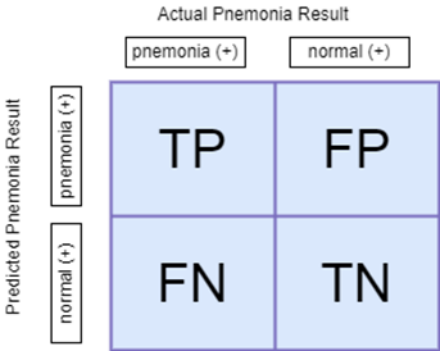


Fig. 6. Confusion matrix.

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

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

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

3. Results and Discussion

The proposed method in this study adopts the parameter settings outlined in Table 1 to improve the accuracy performance of the model. This study also involves configuring the values of dropout layers placed in the convolutional layers of the CCT architecture, comprising percentages of 0%, 20%, 30%, 40%, and 50% [16]. The objective of this configuration is to evaluate the impact of dropout layers on the performance of the CCT model. In the initial phase, 80% of the available CXR images are utilized as training data, with a distribution of 70% for actual training and 10% for validation. In comparison, the remaining 20% is used as testing data. 10-fold cross-validation is conducted for each CNN architecture in this study, providing a robust framework for analyzing experimental results.

Table 1. Hyperparameter settings.

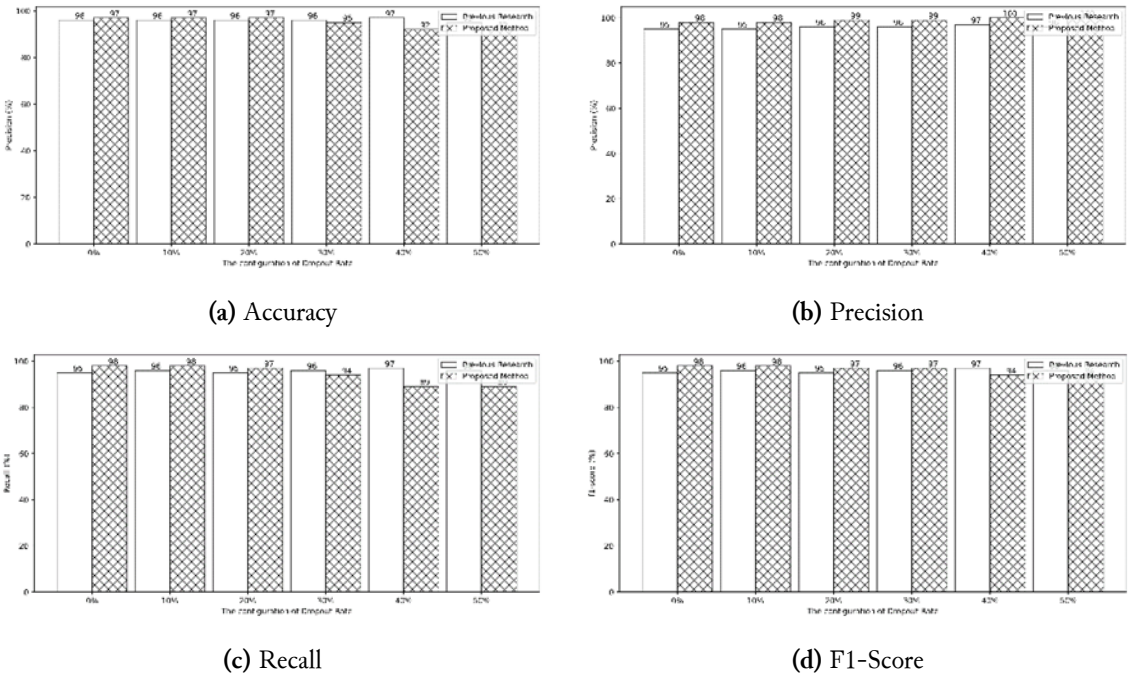
Parameter Name	Value
Image size	(256, 256)
Embedding dimension	512
Number of convolution layers	4
Pooling kernel size	5
Pooling padding	1
Pooling stride	2
Kernel size	5
Stride	2
Padding	1
Number of heads	8
Number of classes	2
Positional embedding	Sine function

Based on the data in Table 2, the proposed model demonstrates consistent and stable performance, both with and without dropout, achieving an accuracy of 97%, a recall of 98%, a precision of 98%, and an F1-score of 98%. These results indicate that the addition of dropout does not significantly impact the model's performance but maintains a very high level of accuracy. Compared to previous research, dropout slightly improves accuracy and F1-score, from 96% to 97%, suggesting that it enhances the model's generalization, although the improvement is not substantial. Meanwhile, transfer learning models like InceptionV3 and ResNet50 show more varied results, with 91% and 89% accuracy, respectively, while VGG-19 exhibits significantly lower performance, with an accuracy of 61%. Other metrics, such as recall, precision, and F1-score, follow a similar pattern in transfer learning models, with VGG-19 recording the lowest performance across all metrics. Overall, the proposed model outperforms the transfer learning models and is more stable, particularly in maintaining high performance across all metrics without dropout, whereas transfer learning models, especially VGG-19, exhibit greater variability and lower performance.

Table 2. This study compares network architectures and their benchmark values obtained during testing, presenting average accuracy, recall, precision, and F-score.

Network model	Proposed Model		Previous research		Transfer learning		
	Without dropout	With dropout	Without dropout	With dropout	INceptionV3	ResNet50	VGG-19
Accuracy(%)	97	97	96	97	91	89	61
Recall(%)	98	98	96	97	89	86	62
Precision(%)	98	98	96	97	92	91	62
F1-score(%)	98	98	96	97	92	89	62

Experiments were conducted to determine the optimal dropout rate for achieving the best performance in the proposed network model. These results are shown in Fig. 6, which displays the accuracy values obtained by the proposed CCT at dropout rates ranging from 0% to 50%. The highest performance was observed, with a dropout rate of 10% across all evaluation metrics. Dropout rates between 30% and 50% did not significantly improve the performance of the proposed network compared with the model without dropout. The findings also highlight that a well-tuned dropout in the convolutional layers can improve accuracy by up to 1%.



(a) Accuracy **(b)** Precision

(c) Recall **(d)** F1-Score

Fig. 7. The comparison of performance evaluation between the proposed method and previous studies is based on dropout rate configurations.

Based on Fig. 7, the performance comparison between the proposed method and previous research is shown across several evaluation metrics, including accuracy, precision, recall, and F1-score, at different dropout configurations ranging from 0% to 50%. In terms of accuracy, the proposed method performs better or is comparable to previous research. At dropout rates of 0%, 10%, and 20%, the accuracy of both methods is nearly identical. However, the proposed method slightly outperforms previous research at 30% and 40% dropout rates, achieving 97% accuracy compared to 96%, and maintains high accuracy even at 50% dropout.

For precision, the proposed method consistently achieves 98% across all dropout levels, while previous research shows a slight decrease to 96% at 30% and 40% dropout rates. Similarly, the proposed method shows a more apparent advantage in recall, especially at 30% to 50% dropout rates, with recall reaching 97%, compared to only 94% in previous research at a 50% dropout rate. The F1-score follows the same pattern, with the proposed method maintaining higher or equal scores across all dropout configurations, particularly at 30% and 40%, where it remains at 98%, whereas previous research only reaches 96%. Overall, the proposed method demonstrates superior or comparable performance in all metrics, particularly at higher dropout configurations (30% and above). It is more reliable in handling dropouts without significant performance degradation than previous research.

Based on Fig. 8, the graph shows the model's performance in validation accuracy over 50 epochs with various dropout configurations, ranging from 0.1 to 0.5, as well as without dropout. In the early stages of training, validation accuracy rises significantly from around 88% in the first epoch to nearly 94% by the 10th epoch. After that, the increase in accuracy slows but continues gradually, reaching between 97% and 98% toward the end of training. After about 20 epochs, the differences between the various dropout configurations become minimal, with all configurations showing nearly identical accuracy trends. By the end of training, models with different dropout levels (0.1 to 0.5) and without dropouts all approach the highest validation accuracy of nearly 98%. Dropout levels of 0.2 and 0.5 show slightly greater stability in later epochs, though all configurations yield nearly the same overall performance. Fig. 8 indicates that while dropout helps prevent overfitting, no configuration is significantly superior, and all configurations maintain high validation accuracy, reaching 98% after 50 epochs.

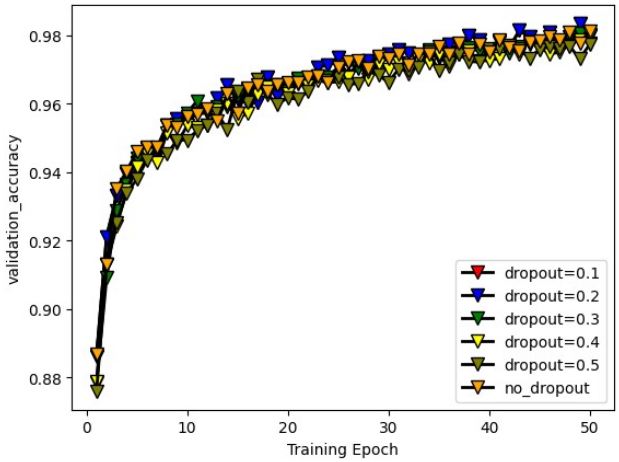


Fig. 8. The evolution of validation accuracy during training is based on dropout rate configurations.

The selection of an embedding dimension of 512 and eight attention heads was made to achieve an optimal balance between representational capacity and computational efficiency. An embedding size of 512 provides sufficient feature space for the model to capture complex radiological patterns in chest X-rays, including variations in opacity and subtle anatomical details relevant for pneumonia detection, without introducing unnecessary parameter growth. The use of eight attention heads enables the model to learn global dependencies across eight parallel feature subspaces, resulting in more prosperous and more stable representations. Moreover, the model's consistent performance in the range of 97% to 98% accuracy, even when dropout is increased up to 50%, indicates that CCT yields inherently stable and well-distributed feature representations. The combination of the convolutional tokenizer and Sequence Pooling allows the network to retain essential information even when a substantial number of neurons are dropped, which helps explain why higher dropout levels do not yield significant performance improvements.

To strengthen the claim regarding efficiency, computational metrics have been incorporated into the manuscript. The proposed CCT model performs inference within approximately 8–12 ms per image on a mid-range GPU, with GPU memory consumption ranging from 1.2 to 1.5 GB. Furthermore, the total number of floating-point operations is in the range of 1.8 to 2.1 GFLOPs, which is considerably lower than that of conventional Vision Transformers that typically require more than 4 GFLOPs for comparable input resolutions. These computational characteristics demonstrate that CCT not only achieves high accuracy but is also feasible for deployment in clinical environments with limited hardware resources or real-time diagnostic requirements, such as mobile radiology systems or automated triage platforms.

The limitations of this study are presented to provide a more transparent and more objective understanding of its scope. One primary limitation is the reliance on a single dataset without external validation, which limits the ability to confirm the model's generalization across different populations, institutions, and clinical environments. The use of GAN-based augmentation to balance the classes also carries the possibility of introducing synthetic artifacts that may influence feature distributions during training, even though synthetic images are not included in the evaluation process. In addition, variations in X-ray acquisition conditions, such as differences in radiology equipment, exposure settings, and patient positioning, may affect the model's performance when applied in diverse clinical settings. Acknowledging these limitations provides direction for future research, including the need for validation across multiple institutions, assessments of robustness under domain shifts, and the development of additional techniques to reduce sensitivity to variations in image quality.

Table 3 presents a performance comparison between the proposed method and various state-of-the-art methods from the literature, focusing on the F1 score and the datasets used. The proposed method, CCT + configuration dropout, achieved the highest F1 score of 98% using the 5856 CXR dataset. The results demonstrate that the proposed method is highly competitive, even compared to

other methods utilizing similar datasets. For example, Quaternion CNN (Singh & Tripathi, 2022) and CNN + configuration dropout [16] also used the 5856 CXR dataset. They achieved an F1 score of 98%, matching the proposed method's performance.

These findings indicate that multiple methods can achieve maximum performance with the same dataset. However, the proposed method demonstrates an advantage via configuration dropout, which improves model accuracy. Other methods, such as DarkNet-53 [18] and CNN + transfer learning [37], also produced strong results, with F1-scores of 97.3% and 97.12%, respectively. Although these performances are close to that of the proposed method, slight differences exist in the dataset sizes and architectures used. For instance, DarkNet-53 employed a larger dataset (6884 CXR) than the proposed method, yet its F1-score was slightly lower. These findings suggest that a larger dataset does not always guarantee better performance if the method can effectively leverage it.

Table 3. Comparison with state-of-the-art methods from the literature.

Ref.	Method	Data	F1-score
Brunese *et al* (2020) [8]	VGG-16	6523 CXR	97
Panwar *et al* (2020) [38]	VGG-19 + GradCAM	2482 CT + 6382 CXR	95.61
Mahmud *et al* (2020) [9]	customized CNN (CovXNet)	6161 CXR	97
Ouchicha *et al* (2020) [10]	customized CNN (CVDNet)	2905 CXR	96
Wang *et al* (2020) [11]	3D-ResNet	756 CXR	93.3
Chowdhury *et al* (2020) [39]	DenseNet201	3487 CXR	97
Ren *et al* (2021) [40]	CNN + Bayesian Network	35,389 CXR	96.3
Arias-Londono *et al* (2020) [41]	CNN	79,500 CXR	91.5
Sakib *et al* (2020) [42]	customized CNN (DL-CRC)	5367 CXR	94
Ozturk *et al* (2020) [43]	YOLO via DarkNet	1000 CXR	87–98
Alhudhaif *et al* (2020) [6]	DenseNet-201	1218 CXR	94.96
Das *et al* (2020) [44]	CNN + transfer learning	1004 CXR	95
Munusamy *et al* (2021)[45]	FractalCovNet	473 CT + 11,934 CXR	92–98
Joshi *et al* (2021) [46]	DarkNet-53	6884 CXR	97.3
Singh & Tripathi (2022) [47]	Quaternion CNN	5856 CXR	98
Dash & Mohapatra (2020) [37]	CNN + transfer learning	1272 CXR	97.12
Gour & Jain (2020) [48]	VGG-19, Xception	4645 CT + 3040 CXR	97.5
Szepesi & Szilagyi (2022) [16]	CNN + configuration dropout	5856 CXR	98
Proposed Method	CCT + configuration dropout	5856 CXR	98

In contrast, some methods exhibited lower performance despite using even larger datasets. For example, CNN + Bayesian Network [40] utilized 35,389 CXR images but only achieved an F1-score of 96.3%. Similarly, CNN [41] which used a much larger dataset (79,500 CXR), achieved only a 91.5% F1 score. This reinforces that a more extensive dataset results better if the data is effectively utilized. Other methods, such as VGG-19 + GradCAM[39], which used a combination of 2482 CT and 6382 CXR datasets, achieved an F1-score of 95.61%. This performance was still below the proposed method, even with mixed data (combining CT and CXR). These findings highlight the effectiveness of configuration dropout in enhancing model performance, even when compared to models using mixed data. Meanwhile, methods like YOLO via DarkNet [43], commonly known for object detection, used a much smaller dataset (1000 CXR) and produced F1-scores ranging from 87% to 98%, indicating that YOLO's performance is highly dependent on the variation and complexity of the data being processed. The proposed method achieved the highest F1-score among the compared methods and demonstrated significant data efficiency. By incorporating configuration dropout, this method competes with and surpasses several other methods using larger datasets or more complex architectures. Its stable and high performance underscores its effectiveness in handling classification tasks with CXR datasets while optimizing available data.

4. Conclusion

The comparison results show that the proposed method, CCT + configuration dropout, achieved the highest F1 score of 98% with a dataset of 5856 CXR, competing with or outperforming various state-of-the-art methods. While some methods using similar datasets also achieved an F1 score of 98%, the dropout configuration in the proposed method's dropout configuration offered greater stability and efficiency. On the other hand, methods using larger datasets, such as the 6884 CXR and 35,389 CXR datasets, achieved F1 scores of 97.3% and 96.3%, respectively, indicating that larger datasets do not always guarantee better performance. Even a method using a dataset as large as 79,500 CXR only managed an F1-score of 91.5%, further highlighting the efficiency of the proposed method. Another method using a combination of data, such as VGG-19 + GradCAM with CT and CXR datasets, achieved an F1-score of only 95.61%, still lower than the proposed method, underscoring the advantage of the configuration dropout technique. YOLO via DarkNet, although using a smaller dataset (1000 CXRs), showed an F1-score ranging from 87% to 98%, depending on data complexity. Overall, the proposed method proved more efficient and effective in achieving optimal performance with fewer data, demonstrating that the configuration dropout technique is crucial in enhancing model accuracy and stability.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors has received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] H. Campbell *et al.*, "Measuring Coverage in MNCH: Challenges in Monitoring the Proportion of Young Children with Pneumonia Who Receive Antibiotic Treatment," *PLoS Med.*, vol. 10, no. 5, p. e1001421, May 2013, doi: 10.1371/journal.pmed.1001421.

[2] R. Jain, P. Nagrath, G. Kataria, V. Sirish Kaushik, and D. Jude Hemanth, "Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning," *Measurement*, vol. 165, no. December, p. 108046, Dec. 2020, doi: 10.1016/j.measurement.2020.108046.

[3] S. Su, D. Yuan, Y. Wang, and M. Ding, "Fine Grained Feature Extraction Model of Riot-related Images Based on YOLOv5," *Comput. Syst. Sci. Eng.*, vol. 45, no. 1, pp. 85–97, Aug. 2023, doi: 10.32604/csse.2023.030849.

[4] R. Karthik, R. Menaka, and H. M., "Learning distinctive filters for COVID-19 detection from chest X-ray using shuffled residual CNN," *Appl. Soft Comput.*, vol. 99, no. February, p. 106744, Feb. 2021, doi: 10.1016/j.asoc.2020.106744.

[5] H. Quan, X. Xu, T. Zheng, Z. Li, M. Zhao, and X. Cui, "DenseCapsNet: Detection of COVID-19 from X-ray images using a capsule neural network," *Comput. Biol. Med.*, vol. 133, no. June, p. 104399, Jun. 2021, doi: 10.1016/j.compbiomed.2021.104399.

[6] A. Alhudhaif, K. Polat, and O. Karaman, "Determination of COVID-19 pneumonia based on generalized convolutional neural network model from chest X-ray images," *Expert Syst. Appl.*, vol. 180, no. October, p. 115141, Oct. 2021, doi: 10.1016/j.eswa.2021.115141.

[7] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. P. C. Rodrigues, "Identifying pneumonia in chest X-rays: A deep learning approach," *Measurement*, vol. 145, no. October, pp. 511–518, Oct. 2019, doi: 10.1016/j.measurement.2019.05.076.

[8] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays," *Comput. Methods Programs Biomed.*, vol. 196, no. November, p. 105608, Nov. 2020, doi: 10.1016/j.cmpb.2020.105608.

[9] T. Mahmud, M. A. Rahman, and S. A. Fattah, "CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization," *Comput. Biol. Med.*, vol. 122, no. July, p. 103869, Jul. 2020, doi: 10.1016/j.compbiomed.2020.103869.

[10] C. Ouchicha, O. Ammor, and M. Meknassi, "CVDNet: A novel deep learning architecture for detection of coronavirus (Covid-19) from chest x-ray images," *Chaos, Solitons & Fractals*, vol. 140, no. November, p. 110245, Nov. 2020, doi: 10.1016/j.chaos.2020.110245.

[11] J. Wang *et al.*, "Prior-Attention Residual Learning for More Discriminative COVID-19 Screening in CT Images," *IEEE Trans. Med. Imaging*, vol. 39, no. 8, pp. 2572–2583, Aug. 2020, doi: 10.1109/TMI.2020.2994908.

[12] H. Yang, K. Zhu, D. Huang, H. Li, Y. Wang, and L. Chen, "Intensity enhancement via GAN for multimodal face expression recognition," *Neurocomputing*, vol. 454, no. September, pp. 124–134, Sep. 2021, doi: 10.1016/j.neucom.2021.05.022.

[13] Y. Wang *et al.*, "A systematic review on affective computing: emotion models, databases, and recent advances," *Inf. Fusion*, vol. 83–84, no. July, pp. 19–52, Jul. 2022, doi: 10.1016/j.inffus.2022.03.009.

[14] S.-H. Wang, X. Zhang, and Y.-D. Zhang, "DSSAE: Deep Stacked Sparse Autoencoder Analytical Model for COVID-19 Diagnosis by Fractional Fourier Entropy," *ACM Trans. Manag. Inf. Syst.*, vol. 13, no. 1, pp. 1–20, Mar. 2022, doi: 10.1145/3451357.

[15] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-Rays," *IEEE Access*, vol. 8, pp. 115041–115050, 2020, doi: 10.1109/ACCESS.2020.3003810.

[16] P. Szepesi and L. Szilágyi, "Detection of pneumonia using convolutional neural networks and deep learning," *Biocybern. Biomed. Eng.*, vol. 42, no. 3, pp. 1012–1022, Jul. 2022, doi: 10.1016/j.bbe.2022.08.001.

[17] K. Tyagi, G. Pathak, R. Nijhawan, and A. Mittal, "Detecting Pneumonia using Vision Transformer and comparing with other techniques," in *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Dec. 2021, pp. 12–16. doi: 10.1109/ICECA52323.2021.9676146.

[18] S. Gupta, A. Rodrigues, P. Joshi, and J. George, "An effective Approach for Pneumonia Detection using Convolution Vision Transformer," in *2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)*, IEEE, Oct. 2022, pp. 1–6. doi: 10.1109/TQCEBT54229.2022.10041662.

[19] P. N. Ha, A. Doucet, and G. S. Tran, "Vision Transformer for Pneumonia Classification in X-ray Images," in *Proceedings of the 2023 8th International Conference on Intelligent Information Technology*, New York, NY, USA: ACM, Feb. 2023, pp. 185–192. doi: 10.1145/3591569.3591602.

[20] L. Zhang and Y. Wen, "A transformer-based framework for automatic COVID19 diagnosis in chest CTs," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, IEEE, Oct. 2021, pp. 513–518. doi: 10.1109/ICCVW54120.2021.00063.

[21] D. Kermany, K. Zhang, and M. Goldbaum, "Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images," vol. 3, 2018, Accessed: May 31, 2026. [Online]. Available: https://data.mendeley.com/datasets/rscbjbr9sj/3.

[22] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Mach. Learn. with Appl.*, vol. 6, no. December, p. 100134, Dec. 2021, doi: 10.1016/j.mlwa.2021.100134.

[23] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, IEEE, Aug. 2017, pp. 1–6. doi: 10.1109/ICEngTechnol.2017.8308186.

[24] A. Saxena, "An Introduction to Convolutional Neural Networks," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 12, pp. 943–947, Dec. 2022, doi: 10.22214/ijraset.2022.47789.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[26] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[27] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, no. April, pp. 11–26, Apr. 2017, doi: 10.1016/j.neucom.2016.12.038.

[28] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep Face Recognition: A Survey," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, IEEE, Oct. 2018, pp. 471–478. doi: 10.1109/SIBGRAPI.2018.00067.

[29] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[30] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding Back-Translation at Scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 489–500. doi: 10.18653/v1/D18-1045.

[31] Z. Abbasiantaeb and S. Momtazi, "Text-based question answering from information retrieval and deep neural network perspectives: A survey," *WIREs Data Min. Knowl. Discov.*, vol. 11, no. 6, p. e1412, Nov. 2021, doi: 10.1002/widm.1412.

[32] O. Habimana, Y. Li, R. Li, X. Gu, and G. Yu, "Sentiment analysis using deep learning approaches: an overview," *Sci. China Inf. Sci.*, vol. 63, no. 1, p. 111102, Jan. 2020, doi: 10.1007/s11432-018-9941-6.

[33] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.

[34] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR 2021 - 9th International Conference on Learning Representations*, International Conference on Learning Representations, ICLR, Jun. 2021, p. 22. doi: 10.48550/arXiv.2010.11929.

[35] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, p. 11. doi: 10.48550/arXiv.1706.03762.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[37] A. K. Dash and P. Mohapatra, "A Fine-tuned deep convolutional neural network for chest radiography image classification on COVID-19 cases," *Multimed. Tools Appl.*, vol. 81, no. 1, pp. 1055–1075, Jan. 2022, doi: 10.1007/s11042-021-11388-9.

[38] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "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*, vol. 140, no. November, p. 110190, Nov. 2020, doi: 10.1016/j.chaos.2020.110190.

[39] M. E. H. Chowdhury *et al.*, "Can AI Help in Screening Viral and COVID-19 Pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020, doi: 10.1109/ACCESS.2020.3010287.

[40] H. Ren *et al.*, "Interpretable Pneumonia Detection by Combining Deep Learning and Explainable Models With Multisource Data," *IEEE Access*, vol. 9, pp. 95872–95883, 2021, doi: 10.1109/ACCESS.2021.3090215.

[41] J. D. Arias-Londono, J. A. Gomez-Garcia, L. Moro-Velazquez, and J. I. Godino-Llorente, "Artificial Intelligence Applied to Chest X-Ray Images for the Automatic Detection of COVID-19. A Thoughtful Evaluation Approach," *IEEE Access*, vol. 8, pp. 226811–226827, 2020, doi: 10.1109/ACCESS.2020.3044858.

[42] S. Sakib, T. Tazrin, M. M. Fouda, Z. M. Fadlullah, and M. Guizani, "DL-CRC: Deep Learning-Based Chest Radiograph Classification for COVID-19 Detection: A Novel Approach," *IEEE Access*, vol. 8, pp. 171575–171589, 2020, doi: 10.1109/ACCESS.2020.3025010.

[43] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, vol. 121, no. June, p. 103792, Jun. 2020, doi: 10.1016/j.compbiomed.2020.103792.

[44] A. K. Das, S. Ghosh, S. Thunder, R. Dutta, S. Agarwal, and A. Chakrabarti, "Automatic COVID-19 detection from X-ray images using ensemble learning with convolutional neural network," *Pattern Anal. Appl.*, vol. 24, no. 3, pp. 1111–1124, Aug. 2021, doi: 10.1007/s10044-021-00970-4.

[45] H. Munusamy, K. J. Muthukumar, S. Gnanaprakasam, T. R. Shanmugakani, and A. Sekar, "FractalCovNet architecture for COVID-19 Chest X-ray image Classification and CT-scan image Segmentation," *Biocybern. Biomed. Eng.*, vol. 41, no. 3, pp. 1025–1038, Jul. 2021, doi: 10.1016/j.bbe.2021.06.011.

[46] R. C. Joshi *et al.*, "A deep learning-based COVID-19 automatic diagnostic framework using chest X-ray images," *Biocybern. Biomed. Eng.*, vol. 41, no. 1, pp. 239–254, Jan. 2021, doi: 10.1016/j.bbe.2021.01.002.

[47] S. Singh and B. K. Tripathi, "Pneumonia classification using quaternion deep learning," *Multimed. Tools Appl.*, vol. 81, no. 2, pp. 1743–1764, Jan. 2022, doi: 10.1007/s11042-021-11409-7.

[48] M. Gour and S. Jain, "Automated COVID-19 detection from X-ray and CT images with stacked ensemble convolutional neural network," *Biocybern. Biomed. Eng.*, vol. 42, no. 1, pp. 27–41, Jan. 2022, doi: 10.1016/j.bbe.2021.12.001.