CViTLNN: A Hybrid Approach based on Vision Transformer and Liquid Neural Network for COVID-19 Detection

Muhammad Waqas
School of Artificial Intelligence and Computer Science, Xian University of Science and Technology, China
waqasmuhammad223@yahoo.com (corresponding author)

Florentin Smarandache
Mathematics, Physics, and Natural Science Division, University of New Mexico, USA
smarand@unm.edu

Muhammad Yasir
School of Computer and Information Engineering, Henan University, China
yasirchuadhry@henu.edu.cn

Farrukh Arslan
Riphah School of Computing and Innovation, Riphah International University, Lahore Campus, Pakistan
farrukh.arslan@riphah.edu.pk

Anum Ali
Hareemtech LLC, Sherman Oaks, California, USA
dranum.ali@hareemtech.co

Received: 26 February 2025 | Revised: 23 March 2025 | Accepted: 2 April 2025

 DOI: https://doi.org/10.48084/etasr.10735

ABSTRACT

The COVID-19 pandemic has underscored the need for accurate and rapid diagnostic tools to assist clinical decision-making. Conventional deep learning models for COVID-19 detection in Chest X-Ray (CXR) images face challenges in poor generalization across imaging conditions and high computational demands. To address these issues, this study proposes CViTLNN, a novel hybrid model combining Vision Transformers (ViTs) and Liquid Neural Networks (LNNs) to improve feature extraction and classification. Specifically, CViTLNN employs a ViT with 24 transformer encoder blocks for efficient extraction of spatial features. The self-attention mechanism of ViTs effectively captures global and local dependencies in CXR images. Furthermore, it incorporates a four-layer LNN for dynamic refinement of features for decision-making. Experimental results demonstrate a test accuracy of 94%, a precision of 95%, and a recall of 94% on a COVID dataset of 5228 CXRs, minimizing false negatives and ensuring high sensitivity. The proposed model provides an efficient and scalable AI-driven diagnostic solution, making it highly suitable for real-world clinical applications, especially in resource-constrained settings.

Keywords-COVID-19 detection; Vision Transformers (ViTs); Liquid Neural Network (LNN); chest X-ray analysis; medical imaging

I. INTRODUCTION

COVID-19 has significantly disrupted global economies, healthcare systems, and daily life, creating an urgent need for accessible and efficient diagnostic tools [1]. Traditional diagnostic methods such as RT-PCR are often slow, costly, and limited in availability, particularly in resource-constrained settings. Chest X-Ray (CXR) images are widely used to detect respiratory diseases, but their interpretation requires expert radiologists, which is a major challenge in underserved regions with limited healthcare infrastructure. Manual analysis of CXR images is time-consuming and prone to human error, underscoring the need for automated diagnostic systems to

improve both speed and accuracy [2]. Recent advances in medical imaging with the integration of deep learning have significantly improved diagnostic accuracy, especially in the context of CXR analysis for pulmonary diseases such as pneumonia, tuberculosis, cancer, and COVID-19 [3]. Several methods have been proposed for automated COVID-19 detection, including a hybrid Convolutional Neural Network (CNN) architecture combined with ResNet50 [4], a transfer learning approach using MobileNetV2 [5], a hybrid of VGG19 with DenseNet [6], a CNN-based DarkCovidNet [7], and an ensemble learning method combining InceptionV3, EfficientNet, and ResNet [8]. Additionally, hybrid CNN and LSTM [9], ResNet50 with InceptionV3 [10], EfficientNet [11], SqueezeNet [12], and Neutrosophic set-based deep transfer learning [13] have been employed with varying degrees of success. Traditional machine learning techniques, such as support vector machines and random forests, depend on manually crafted features and often struggle to handle the complexity of medical images, leading to suboptimal performance [14]. Sequence-based models, such as Recurrent Neural Networks (RNNs), excel at capturing temporal dependencies but fail to effectively model the spatial hierarchies and intricate visual patterns found in medical imaging. Deep learning architectures, particularly CNNs, have shown strong performance in medical image analysis. However, CNNs are limited by their inability to capture global contextual information, which is crucial for accurate disease detection [15]. The introduction of Vision Transformers (ViTs) has addressed this limitation, achieving high accuracy by leveraging self-attention mechanisms that can identify significant patterns across the entire image, making them highly suitable for complex computer vision tasks [16]. Moreover, Liquid Neural Networks (LNNs) offer additional advantages as they dynamically adapt their computations, enabling more flexible and efficient learning from complex data. Their ability to handle temporal and spatial variations in input data makes them ideal for medical image analysis, where conditions can change over time or vary across patients.

To the best of our knowledge, this is the first study to explore the combined potential of ViTs and LNNs for COVID-19 detection, introducing a novel hybrid approach that integrates the strengths of both architectures. The primary contributions of this work are as follows:

- Introduces a ViT architecture incorporating 24 encoder blocks to extract deep features from chest X-ray images, leveraging self-attention mechanisms to capture both local and global representations for COVID-19 detection.
- Builts a liquid neural network consisting of four layers to refine extracted features, enabling dynamic adaptation to input variability and enhancing the model's ability to learn intricate patterns for accurate and efficient classification.
- Finally, develops CViTLNN, a novel hybrid architecture that combines vision transformer-based deep feature extraction with the adaptability of liquid neural networks to achieve automated COVID-19 detection, specifically designed for resource-constrained environments and underserved regions.

II. MATERIALS AND METHOD

Figure 1 illustrates the proposed framework. CViTLNN integrates a ViT and an LNN to address both spatial and temporal complexities in medical image analysis. The process begins with the input CXR image, which undergoes preprocessing steps, such as downscaling, normalization, random rotations, flips, and cropping. The image is then divided into patches (16×16), and each is embedded in a higher-dimensional space, along with position embeddings to preserve spatial relationships. The sequence length of 197 patches is processed, including the addition of a classification (CLS) token that helps in the final decision-making process. The ViT component consists of 24 transformer encoder blocks that leverage multihead self-attention mechanisms to capture both local and global dependencies within the image. This allows the model to recognize fine-grained details throughout the image, which is essential to distinguish complex classes, such as COVID-19, pneumonia, and normal cases in CXR. The attention mechanism examines the relationships between all parts of the image, enabling the model to recognize patterns that would be difficult for traditional CNNs to capture. After extracting ViT features, the four-layer LNN component refines the learned features, adapting its processing dynamically based on input variations. The LNN is designed to handle complex and evolving data, enhancing the robustness and adaptability of the model. Finally, a classifier layer with linear layers, dropout, GELU activation, and normalization processes the refined features, making predictions about the presence of COVID-19, normal conditions, or pneumonia in the CXR image.

A. Dataset

This study utilized the publicly available COVID dataset [17], which contains a large number of annotated CXR images. This dataset was chosen due to its comprehensive and diverse image collections. The data are categorized into three classes: COVID-19, pneumonia, and non-pneumonia (normal) cases. The COVID-19 images were specifically sourced from datasets dedicated to COVID-19, while pneumonia and normal cases were randomly selected from standard CXR datasets. The dataset includes 1626 COVID-19 images, 1800 pneumonia images, and 1802 normal images, ensuring a broad and varied representation of cases. This distribution enhances the model's ability to distinguish COVID-19 from other respiratory diseases, reflecting real-world clinical scenarios. Additionally, the dataset includes patients across different age groups and genders to reduce potential biases during the training phase, promoting more robust and generalizable model performance.

B. Preprocessing Pipeline

Several preprocessing procedures were applied to the dataset to stabilize the input to the ViT model. All images were scaled down to a standard patch embedding dimension, normally to a measure of 224×224 as used in the blue picture. Normalization was applied to bring the pixel intensity values in the images into a common range to have standard input for the model and increase training robustness. Methods such as random rotations, flips, and cropping were used to improve data variety and reduce the risk of overfitting.

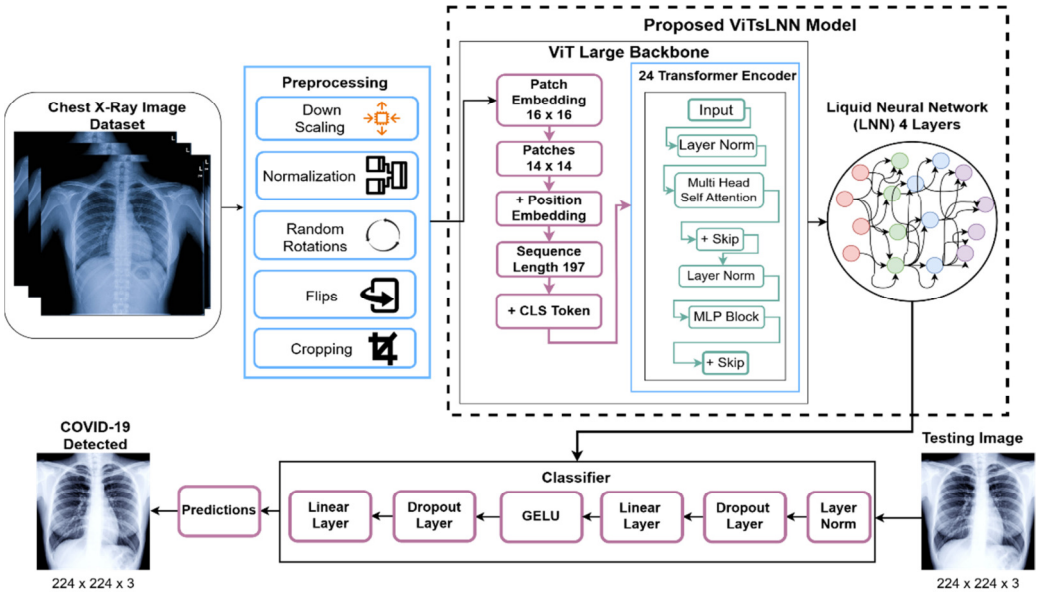

Fig. 1. The framework of the proposed approach (CViTLNN) for COVID-19 detection.

C. Patch Generation Process

Figure 2 shows the patch generation process for a CXR input image with dimensions 224×224×3. This process begins by dividing the input image into smaller square patches of fixed size, such as 16×16 pixels. Given the image dimensions (224×224), it can be divided into a grid of 14 patches along the height and 14 patches along the width, resulting in a total of 14×14 = 196 patches. Each patch retains the three RGB color channels. Each patch, initially a 16×16×3 tensor, is flattened into a one-dimensional vector of size 768 (calculated as 16×16×3). These flattened patch vectors are then passed through a learnable embedding layer, which transforms each vector into a fixed-size embedding to match the model's input requirements.

D. Integrated Vision Transformer (ViT) Backbone and Liquid Neural Network Model (ViTLNN)

Figure 3 shows the ViTLNN process, which begins by dividing the input image into fixed-size patches that are flattened into vectors (encoded patches, $X1$, $X2$, ..., $X196$). A Class (CLS) token is added to the sequence for the global image summary. These encoded patches and the CLS token are input into the transformer encoder with positional encodings that preserve spatial relationships. The ViT processes these embeddings by passing them through transformer layers with self-attention mechanisms, resulting in a feature tensor of shape (batch_size, seq_length, embed_size). The LNN further refines these features. The LiquidLayer function iteratively updates the features by generating a time embedding and applying a state network to refine the current state. The refined features are normalized after each update. This process produces an output tensor of the same shape as the input. Finally, the classification head extracts the CLS token from the refined features, normalizes it, and passes it through a feedforward network to produce classification logits, which represent the final predictions. The main workflow in Figure 1 shows this sequence, using ViT for initial feature extraction, LNN for dynamic refinement, and the Classification Head to output predictions. This hybrid design combines ViT's spatial reasoning with LNN dynamic adaptability, enhancing the model's ability to make accurate classifications.

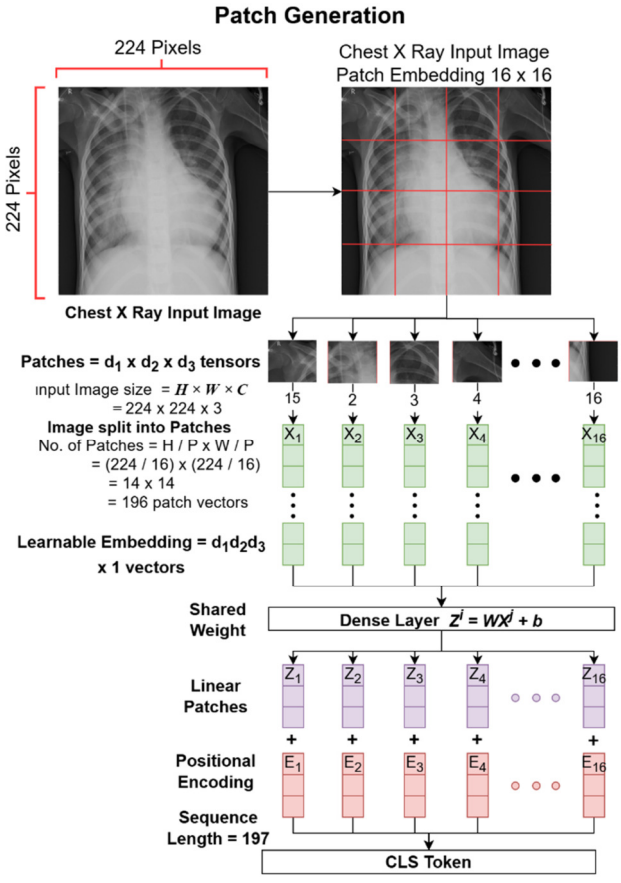

Fig. 2. Step-by-step illustration of the patch generation process.

TABLE I. DATASET DISTRIBUTION BY TYPE AND TOTAL SAMPLES

Dataset type	COVID-19 cases	Pneumonia cases	Normal cases	Total samples
Training Set	1300	1440	1441	4181
Validation Set	164	180	181	522
Testing Set	162	180	180	525
Total	1626	1800	1802	5228

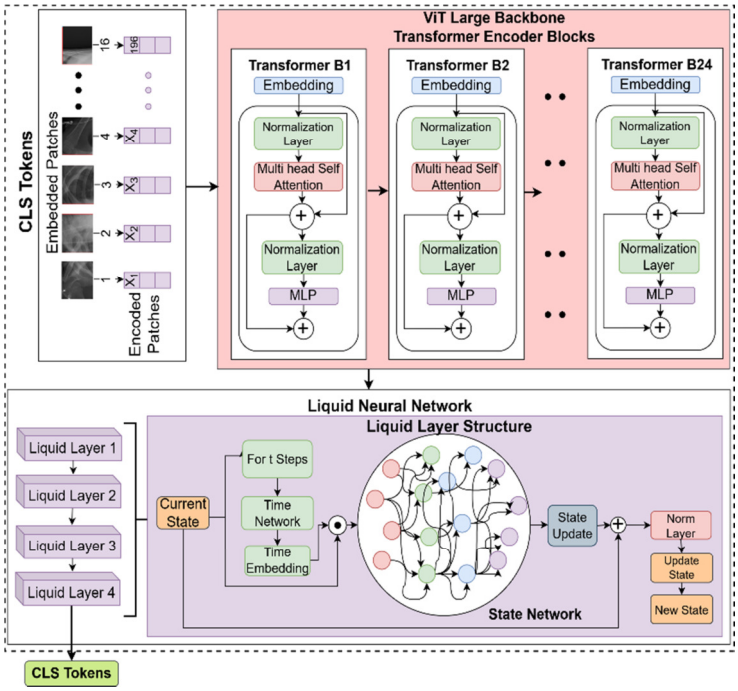

Fig. 3. The architecture of hybrid ViT and LLN.

E. Transformer Encoder Blocks

The ViTs backbone in Figure 3 consists of 24 transformer encoder blocks, each designed to handle deeper aspects of the features. The Multi-head Self-Attention (MHSA) mechanism calculates relationships between all patches of the image simultaneously, capturing both local and global context. This enables the model to understand complex dependencies across the entire image, unlike traditional convolutional methods that focus primarily on local features. Each block also contains Feedforward Networks (MLP), which consist of two linear layers with a GELU activation function. The first layer increases the input dimensions to capture more intricate patterns, while the second layer reduces the dimensions back to the original size, refining the features extracted by MHSA. Layer Normalization (LayerNorm) is applied to stabilize training and improve convergence, normalizing the input across feature dimensions to mitigate issues like exploding or vanishing gradients. Additionally, residual connections are used to allow information to bypass certain layers, preserving original features and enhancing performance. After processing through all 24 blocks, the CLS token contains a condensed representation of the image, which is then used for downstream tasks such as classification.

F. Experimental Settings

All experiments were executed on an RTX 4060 GPU with 32GB RAM and an Intel i9 13th generation processor. The model was trained for 100 epochs using the Adam optimizer with a learning rate of 0.0001. The batch size was set to 32, and a dropout rate of 0.1 was applied to prevent overfitting during training. The GELU activation function was used in the feed-forward layers, and the positional encoding type was learnable, allowing the model to retain spatial information in the input patches. The cross-entropy loss function with label smoothing was used to train the model. This loss function is well-suited for multiclass classification problems and combines the softmax operation with the negative log-likelihood of the predicted probabilities. To prevent the model from becoming overly confident in its predictions and to improve generalization, label smoothing was applied with a smoothing factor of 0.1. This means that the true class label is assigned a probability of 0.9, and the remaining 0.1 is evenly distributed among the other classes.

G. Evaluation Metrics

Well-known metrics, namely accuracy, precision, and recall, were employed to assess the performance of the proposed approach.

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

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

$$Recall = \frac{TP}{TP+FP} \quad (3)$$

where TP, TN, FP, and FN denote true positives, true negative, false positive, and false negative predictions, respectively.

III. RESULTS AND DISCUSSION

The proposed approach achieved an overall classification accuracy of 94%, a precision of 95%, and a recall of 94%. Figure 4 shows the progression of precision and recall over 100 epochs of training for the proposed model. Both metrics show consistent improvement, with precision (blue line) and recall (red line) steadily increasing throughout the training process. Early epochs exhibit lower values, but both precision and recall stabilize near 0.95 after 100 epochs, indicating that the model has effectively learned to differentiate between classes. The close alignment of the precision and recall curves suggests balanced performance, with the model maintaining high accuracy in both correctly identifying positive instances (precision) and detecting all positive cases (recall).

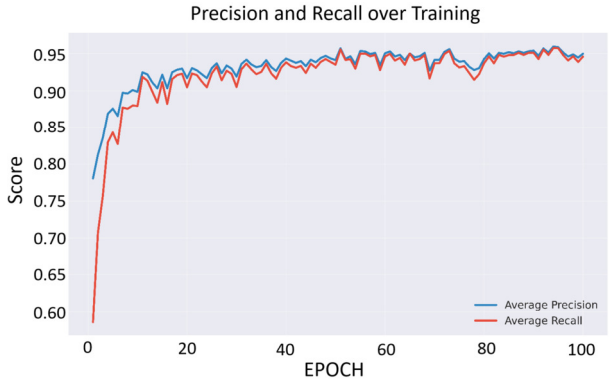

Fig. 4. Precision and recall curves over 100 epochs.

Figure 5 shows the training, validation, and test accuracy curves of the proposed approach across 100 epochs. The blue curve represents the training accuracy, the red curve represents the validation accuracy, and the green curve represents the test accuracy. At the beginning of training, all accuracy curves start relatively low. As training progresses, the training accuracy steadily increases, surpassing 90% by epoch 30, and then stabilizes at a high level close to 97%. This rapid improvement in training accuracy indicates that the model is effectively learning and fitting the training data. The validation and test accuracy curves show a similar trend, but their growth is slightly more gradual than the training accuracy. Both curves follow the same pattern, with validation accuracy (red) and test accuracy (green) nearing 90% toward the end of the training. The proximity of the three curves suggests that the model is not overfitting, as the validation and test accuracies remain close to the training accuracy, indicating generalization to unseen data. Overall, the results show that the CViTLNN approach achieves robust performance across all datasets (training, validation, and test), with the test accuracy stabilizing at 93.76%, which reflects the model's ability to generalize well while avoiding overfitting.

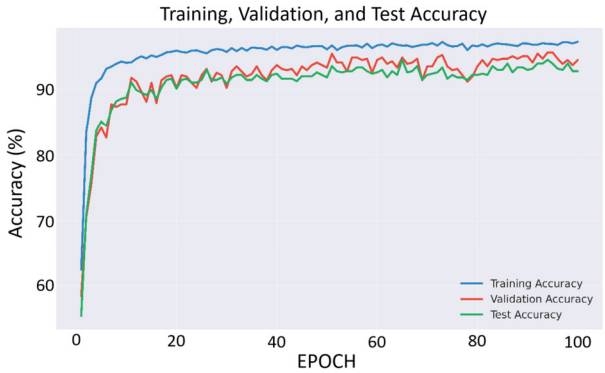

Fig. 5. Training, validation, and test accuracy over 100 epochs.

Figure 6 shows the confusion matrix of the CViTLNN predictions for three classes: COVID, Normal, and Pneumonia. The diagonal elements of the matrix indicate the number of correct predictions, with 162 COVID, 177 Normal, and 160 Pneumonia instances correctly classified. The off-diagonal elements reveal misclassifications.

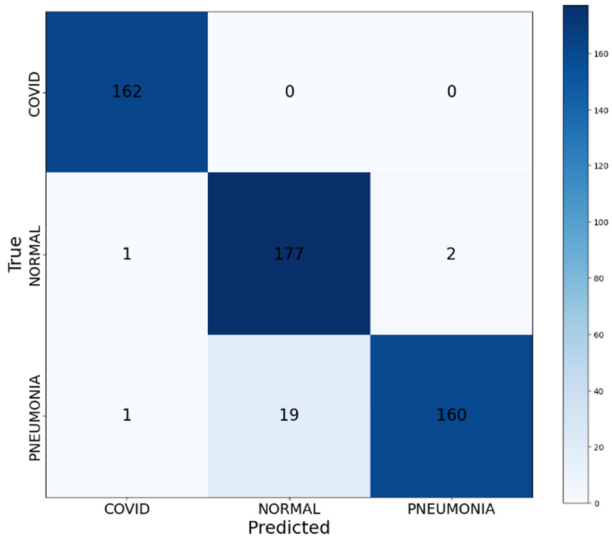

Fig. 6. Confusion matrix of CViTLNN predictions, with correct classifications on the diagonal and misclassifications off it.

Table II presents a comparison of the proposed approach with existing methods, including hybrid CNN-ResNet50, transfer learning with MobileNetV2, VGG19-DenseNet, DarkCovidNet-CNN, ensemble approach using InceptionV3, Xception, and ResNet, CNN-LSTM, Capsule Network (CapsNet), Efficient Net, Squeeze Net and CNN on wavelet transform, all focusing on the detection of COVID-19, pneumonia, and normal cases. The CViTLNN model delivers state-of-the-art performance, achieving a test accuracy of 94%, a precision of 95%, and a recall of 94%, surpassing most existing approaches.

TABLE II. COMPARISON WITH STATE-OF-THE-ART METHODS

Ref	Method	Accuracy (%)	Precision (%)	Recall (%)
[4]	Hybrid (CNN-ResNet50)	92	94	95
[5]	Transfer Learning with MobileNetV2	89	90	91
[6]	VGG19-DenseNet	90	91	92
[7]	DarkCovidNet-CNN	87	89	88
[8]	Ensemble(InceptionV3, Xception, ResNet)	93	94	93
[9]	CNN-LSTM	91	92	91
[10	ResNet50-InceptionV3	91	92	91
[11]	EfficientNet	89	90	89
[12]	SqueezeNet	88	89	88
This	**CViTLNN**	**94**	**95**	**94**

IV. CONCLUSION

This work introduced CViTLNN, a hybrid architecture based on ViT and LNN for COVID-19 detection from CXR images. The ViT architecture employed 24 encoder blocks to extract deep features from chest X-ray images, using self-attention mechanisms to capture both local and global representations. Subsequently, the LNN was integrated with ViT for dynamic feature refinement for COVID-19 detection. Experimental results demonstrated that CViTLNN achieved superior performance by surpassing existing methods, achieving a competitive test accuracy of 94%, a precision of 95%, and a recall of 94%, highlighting its robustness and reliability in detecting COVID-19 from CXR images. This work is particularly significant for deployment in low-resource settings, where access to expert radiologists and advanced diagnostic tools is limited. The proposed architecture provides an efficient and effective solution for COVID-19 detection, offering the potential to assist healthcare professionals in underserved populations. Future work will involve expanding the application of the ViTLNN model to other biomedical image analysis tasks, such as lung cancer detection.

DATA AVAILABILITY STATEMENT

Dataset: https://data.mendeley.com/datasets/dvntn9yhd2/1

Code: https://github.com/BioAIMLDL/CViTLNN

REFERENCES

[1] N. Naveed *et al.*, "The Global Impact of COVID-19: A Comprehensive Analysis of Its Effects on Various Aspects of Life," *Toxicology Research*, vol. 13, no. 2, Apr. 2024, Art. no. tfae045, https://doi.org/10.1093/toxres/tfae045.

[2] K. Y. Tehseen *et al.*, "Transformative effects of COVID-19 on global economy and internet of medical things (IoMT): current vision, role and applications," *International Journal on Emerging Technologies*, vol. 12, no. 2, pp. 66–76, 2021.

[3] M. Ibrar, M. Asif, M. Kashif, N. Imran, S. Hameed, and M. Ali, "Speech Recognition System (Home Appliances Controller of Local & Remote System) using LPC & HMMs Methodologies," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 3, pp. 2365–2370, Jun. 2021, https://doi.org/10.30534/ijatcse/2021/1211032021.

[4] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Scientific Reports*, vol. 10, no. 1, Nov. 2020, Art. no. 19549, https://doi.org/10.1038/s41598-020-76550-z.

[5] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 635–640, Jun. 2020, https://doi.org/10.1007/s13246-020-00865-4.

[6] E. E. D. Hemdan, M. A. Shouman, and M. E. Karar, "COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images." arXiv, Mar. 24, 2020, https://doi.org/10.48550/arXiv.2003.11055.

[7] 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," *Computers in Biology and Medicine*, vol. 121, Jun. 2020, Art. no. 103792, https://doi.org/10.1016/j.compbiomed.2020.103792.

[8] T. Rahman *et al.*, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Computers in Biology and Medicine*, vol. 132, May 2021, Art. no. 104319, https://doi.org/10.1016/j.compbiomed.2021.104319.

[9] R. Jain, M. Gupta, S. Taneja, and D. J. Hemanth, "Deep learning based detection and analysis of COVID-19 on chest X-ray images," *Applied Intelligence*, vol. 51, no. 3, pp. 1690–1700, Mar. 2021, https://doi.org/10.1007/s10489-020-01902-1.

[10] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," *Pattern Analysis and Applications*, vol. 24, no. 3, pp. 1207–1220, Aug. 2021, https://doi.org/10.1007/s10044-021-00984-y.

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

[12] F. Ucar and D. Korkmaz, "COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images," *Medical Hypotheses*, vol. 140, Jul. 2020, Art. no. 109761, https://doi.org/10.1016/j.mehy.2020.109761.

[13] N. E. M. Khalifa, F. Smarandache, G. Manogaran, and M. Loey, "A Study of the Neutrosophic Set Significance on Deep Transfer Learning Models: an Experimental Case on a Limited COVID-19 Chest X-ray Dataset," *Cognitive Computation*, vol. 16, no. 4, pp. 1602–1611, Jul. 2024, https://doi.org/10.1007/s12559-020-09802-9.

[14] H. Alalawi, M. Alsuwat, and H. Alhakami, "A Survey of the Application of Artifical Intellegence on COVID-19 Diagnosis and Prediction," *Engineering, Technology & Applied Science Research*, vol. 11, no. 6, pp. 7824–7835, Dec. 2021, https://doi.org/10.48084/etasr.4503.

[15] A. Sufian, A. Ghosh, A. S. Sadiq, and F. Smarandache, "A Survey on Deep Transfer Learning to Edge Computing for Mitigating the COVID-19 Pandemic," *Journal of Systems Architecture*, vol. 108, Sep. 2020, Art. no. 101830, https://doi.org/10.1016/j.sysarc.2020.101830.

[16] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021, https://doi.org/10.48550/arXiv.2010.11929.

[17] S. Kumar, "Covid19-Pneumonia-Normal Chest X-Ray Images." Mendeley, Jun. 14, 2022, https://doi.org/10.17632/DVNTN9YHD2.1.