scientific reports

OPEN

Metaheuristic optimizers integrated with vision transformer model for severity detection and classification via multimodal COVID-19 images

V. Padmavathi & Kavitha Ganesan

This study introduces a novel hybrid framework for classifying COVID-19 severity using chest X-rays (CXR) and computed tomography (CT) scans by integrating Vision Transformers (ViT) with metaheuristic optimization techniques. The framework employs the Grey Wolf Optimizer (GWO) for hyperparameter tuning and Particle Swarm Optimization (PSO) for feature selection, leveraging the ViT model's self-attention mechanism to extract global and local image features crucial for severity classification. A multi-phase classification strategy refines predictions by progressively distinguishing normal, mild, moderate, and severe COVID-19 cases. The proposed GWO_ViT_PSO_MLP model achieves outstanding accuracy, with 99.14% for 2-class CXR classification and 98.89% for 2-class CT classification, outperforming traditional CNN-based approaches such as ResNet34 (84.22%) and VGG19 (93.24%). Furthermore, it demonstrates superior performance in multi-class severity classification, especially in differentiating challenging cases like mild and moderate infections. Compared to existing studies, this framework significantly improves accuracy and computational efficiency, highlighting its potential as a scalable and reliable solution for automated COVID-19 severity detection in clinical applications.

Keywords COVID-19 severity classification, Hybrid vision transformer (ViT) model, Chest X-ray (CXR) and CT imaging, Grey Wolf Optimizer (GWO), Particle Swarm Optimization (PSO), Multi-phase classification framework

The COVID-19 pandemic had a thoughtful impact on healthcare systems worldwide, as its unpredictable progression and severe impact on lung function continue to challenge clinicians. Advances in diagnostic tools, including molecular testing and imaging modalities, have been crucial in assessing the severity of the disease, guiding treatment decisions, and predicting patient outcomes. While RT-PCR tests remain essential for the diagnosis of COVID-19, imaging techniques such as chest X-ray (CXR) and computed tomography (CT) have proven vital in evaluating lung involvement and determining the severity of the disease[1].

CXR is a frequently used imaging modality due to its availability and ease of use, especially in resource-constrained environments or during pandemic surges. CXR reveals key radiological features associated with COVID-19, which includes bilateral lung infiltrates, ground-glass opacities (GGOs), and consolidations, which correlate with the severity of lung damage[2]. Its portability and accessibility make it particularly valuable in intensive care units (ICUs), where mobile CXR can be used for rapid bedside evaluations[3]. Several scoring systems have been developed to quantify lung involvement in COVID-19 patients via CXR. For example, the Brixia score is used to evaluate the presence of opacities in different lung zones, whereas the Airspace Opacity Severity Score (ASOS) is used to assess the percentage of lung affected by airspace opacification. These scoring systems provide standardized methods to determine disease severity, help track disease progression and guide treatment decisions[4].

However, despite its utility in critical settings, CXR has limitations, particularly in the detection of early or subtle lung changes, making it less useful in the identification of mild or moderate cases of COVID-19.

Department of Biomedical Engineering, CEG Campus, Anna University, Chennai 600025, India. email: kavithag_mit@annauniv.edu

Nonetheless, in settings where rapid assessments are necessary, especially for critically ill patients, CXR remains an indispensable diagnostic tool.

Compared with CXR, CT imaging particularly high-resolution computed tomography (HRCT), offers a more detailed and sensitive approach for evaluating lung damage caused by COVID-19. CT scans are critical for detecting early lung changes such as GGOs, crazy paving patterns, and consolidations. These features help clinicians differentiate among mild, moderate, and severe COVID-19 cases[5]. The CT severity score (CT-SS) has been widely adopted to evaluate the volume and density of GGOs and consolidations, providing a quantitative assessment of lung involvement and correlating with disease severity.

CT is beneficial in early-stage COVID-19, where patients may present with mild symptoms but have significant lung involvement not visible on CXR. CT can also detect complications associated with severe COVID-19, such as pulmonary embolism and fibrosis, which are critical for managing patients with worsening respiratory function[6]. Both CXR and CT imaging play crucial roles in assessing COVID-19 severity, although their applications differ on the basis of the clinical context. CT is the preferred modality for early detection and comprehensive assessment of lung abnormalities because of its high sensitivity and detailed imaging[7]. However, CXR remains essential in critical care settings due to its accessibility and portability, making it a practical solution for rapid evaluations, particularly in ICUs[8].

Recent advancements in AI-driven systems have enhanced the efficiency and precision of CT severity scoring, offering reliable and objective assessments of lung involvement. Deep learning models have been integrated with high resolution CT to automate lung segmentation and scoring, offering rapid assessments of disease severity in clinicalpractice[9].

Artificial intelligence (AI) is increasingly important in COVID-19 severity analysis because it integrates imaging techniques such as CXR and CT. AI-based tools for segmentation and classification have been applied to automate the detection of infection patterns, including GGOs and consolidations, which help clinicians quantify lung involvement and predict disease progression[10]. These tools enable faster and more accurate assessments, reduce the diagnostic burden on radiologists and improve healthcare resource management.

In recent studies, AI-driven systems have demonstrated high accuracy in automating CT and CXR severity scoring, through better correlations with traditional radiologist assessments[11]. AI models for CT severity scoring can automatically detect and quantify GGOs and consolidations, providing real-time evaluations of disease progression and outcomes. AI has also been used to integrate clinical data with CXR findings to predict ICU admissions and patient mortality, making it an invaluable tool in managing COVID-19 patients.

A comparative study revealed that while CT excels at detecting of early-stage lung changes, CXR remains crucial to monitor the disease progression in critically ill patients. In situations where CT isn't feasible, CXR serves as a dependable alternative to assess lung involvement[12]. Both CXR and CT are key to evaluate the severity of COVID-19, which offers distinct advantages based on clinical needs. The integration of AI into imaging has further improved diagnostic speed and accuracy, leads to better patient management and outcomes.

In addition to enhancing traditional imaging modalities, artificial intelligence has revolutionized medical imaging analysis by applying Vision Transformers (ViT). ViT leverages self-attention mechanisms to extract intricate features from imaging data, outperforming traditional convolutional neural networks (CNNs) in capturing subtle abnormalities. In recent studies, ViT have been employed for COVID-19 classification and severity assessment using both CXR and CT images, achieving state-of-the-art accuracy and interpretability. Notable models include COVID-ViT for CT-based COVID-19 classification[13], xViTCOS for explainable COVID-19 screening via CXR[14], and LT-ViT for multi-label classification of normal, COVID-19, and other pneumonia cases[15].

Optimizers play a crucial role in deep learning by adjusting model parameters to minimize loss and improve performance[59]. Ko et al. (2024) evaluated six different optimizers (Adam, AdamW, NAdam, RAdam, SGDW, and Momentum) on three ViT architectures such as ViT, FastViT and CrossViT using a CXR dataset and identified strategies to enhance accuracy for lung disease prediction, emphasizing the role of optimization techniques[16]. Park et al. (2021) demonstrated the efficacy of ViT for COVID-19 detection using a chest X-ray feature corpus, showcasing the model's capability to generalize across diverse datasets. These models demonstrate the adaptability of ViT to diverse datasets and their potential to generalize effectively across medical imaging tasks[17]. To achieve optimal performance with ViT, fine–tuning and selection of hyperparameters, such as batch size, layer depth, attention heads, and training epochs is required. Overall, these studies establish ViT as a powerful alternative to CNNs, offering improved accuracy and explainability in classifying and detecting COVID-19 and pneumonia from medical images.

Existing studies primarily rely on CNN-based approaches for COVID-19 severity classification. Challenges persist in capturing long-range dependencies in medical images and robust optimization techniques for hyperparameter tuning and feature selection are often lacking. In contrast, the proposed method utilizes ViT, which excels at modeling both global and local features through self-attention mechanisms, effectively addressing these limitations. Additionally, the integration of metaheuristic algorithms, such as the Grey Wolf Optimizer (GWO) and Particle Swarm Optimization (PSO), sets this framework apart by optimizing hyperparameters and identifying significant features, resulting in enhanced accuracy and computational efficiency.

The contributions of the proposed work for COVID-19 severity analysis (mild, moderate, severe) using CXR and CT images are as follows:

1. The Grey Wolf Optimizer (GWO) is used to optimize Vision Transformer (ViT) hyperparameters, including batch size, epochs, layer depth, and attention heads, for improved severity classification performance.
2. The optimized ViT model is trained and employed to extract deep features from CXR and CT images, effectively representing the severity levels.

3. Particle Swarm Optimization (PSO) is applied to select the most significant features from the extracted ViT features, reducing redundancy and enhancing classification accuracy.
4. A Multi-Layer Perceptron (MLP) classifier, with hyperparameters specifically tuned, is trained on the optimized features to accurately classify COVID-19 severity levels into mild, moderate, and severe categories.
5. The final model is evaluated using metrics such as accuracy, precision, recall, and F1-score, along with Grad-CAM visualizations, to validate the effectiveness of the combined GWO, PSO, and MLP approach.

Related work

The COVID-19 pandemic has spurred major progress in the use of medical imaging methods, such as CXR and CT, for early diagnosis and assessment of severity. Machine learning (ML) and deep learning (DL) models, particularly convolutional neural networks (CNNs), have become critical tools for automating the diagnostic process, especially in settings where medical resources are limited[58]. Early detection of COVID-19 using pre-trained CNN models, such as ResNet50, VGG16, and InceptionV3, have proven highly effective in distinguishing COVID-19 from other conditions[18]. Leveraging transfer learning and binary cross-entropy loss, these models have been applied successfully to CXR and CT scans, particularly in the absence of widespread RT-PCR testing[19].

However, despite the success of CNNs, manual interpretation of CXR images remains time-consuming and prone to error, necessitating the development of automated, multi-stage AI systems.[20] proposed a multi-stage ensemble learning system to classify COVID-19 severity and localize infections, which has improved diagnostic accuracy. Similarly, segmentation models such as U-Net have been utilized to extract features such as ground-glass opacities and other infection markers from CT images, aiding in the severity classification of COVID-19 cases[21].

Even with the advances in CNNs, they have limitations due to their small receptive fields, restricting their ability to capture global image features. These shortcomings are critical to analyze the complex structures such as the lungs. The ViT model have emerged as a solution, that offers several advantages over CNNs, particularly to process CXR and CT scans for lung disease detection. The ViT model uses self-attention mechanisms to model long-range dependencies, which effectively capture both local and global features[22]. This global context is essential to identify abnormalities in lung images, demonstrated that ViT model outperform CNN to quantify lung infection severity. Similarly,[23] reported that the ViT model was superior to detect COVID-19 severity, especially when it was combined with attention visualization techniques such as Grad-CAM.

Another key advantage of the ViT model is its reduced reliance on deep convolutional layers, which minimizes the need for complex feature hierarchies and lowers the risk of overfitting. This is particularly relevant in medical imaging, where data are often limited.[24] established that ViT-based models outperformed CNNs, even on smaller datasets, especially for the diagnosis of post-COVID-19 pulmonary conditions.

Recent research has explored hybrid models that combine CNNs and ViT to further increase diagnostic accuracy.[25] developed a fusion model integrating ResNet andViT for 3D CT image classification, to exploit CNN strength in capturing local details and ViT model ability to model the global context. This approach has proven more accurate to find conditions such as COVID-19 and tuberculosis[26]. Similarly,[27] introduced a two-stage ViT model that outperformed the traditional CNN in COVID-19 severity classification by detecting subtle patterns across lung regions.

Despite these benefits, the ViT model presents challenges due to its high computational costs. Researchers have addressed this issue to develop more efficient ViT model variants.[28] explored the ViT model with fewer trainable parameters for COVID-19 severity assessment, demonstrating that these models can maintain high diagnostic accuracy with reduced computational demands.[29] proposed an efficient feature extraction framework for pneumonia detection via the ViT model, to reduce computational costs while retains its strong performance.[30] demonstrated optimized pre-trained ViT models for COVID-19 classification using Stochastic Configuration Networks (SCNs), to improve performance and avoid overfitting, which is particularly important in data-scarce medical applications[56].

Emerging ViT-based models such as PneuNet and IEViT have shown significant promise in CXR image classification. PneuNet improves COVID-19 pneumonia diagnosis by applying multi-head attention to channel patches, whereas IEViT enhances sensitivity, precision, and generalizability with input optimization[31]. These models offer effective solutions to automate diagnoses and reduce dependence on radiologists, particularly in remote or underserved areas[32]. ViT model frameworks efficiently handle diverse image resolutions, and are well-suited for diagnosis such as pneumonia and COVID-19 cases, improving healthcare accessibility and outcomes[33].

Existing studies on COVID-19 severity detection typically rely on CNN, which excel at capturing local features but struggles with long-range dependencies, limiting their application to complex medical images such as CXR and CT scans. CNN-based approaches, such as ResNet50, VGG16, and InceptionV3, are reviewed, highlighting their successes in early diagnosis but also their limitations, including small receptive fields and the inability to capture global image features. The emergence of ViT is introduced as a solution, leveraging self-attention mechanisms to model both global and local features effectively, addressing CNN shortcomings. Hybrid models that integrate ViT and CNNs are explored, demonstrating improved diagnostic accuracy, along with innovations to reduce ViT computational demands through lightweight variants and optimized architectures. The review identifies research gaps, including a predominant focus on binary classification, limited exploration of multi-class severity classification, and underutilization of multimodal data. The proposed framework integrates ViT with metaheuristic algorithms, such as GWO for hyperparameter tuning and PSO for feature selection, addressing these gaps and offering a scalable, robust solution for COVID-19 severity detection.

Datasets	Images	Normal	COVID-19	Non-COVID-19	Total
COVID-QU-Ex [34]	CXR	10,701	11,956	11,263	33,920
CT scans for COVID-19 Classification [35]	CT	9,979	4,001	5,705	19,685

Table 1. Summary of COVID-19 imaging datasets used for classification.

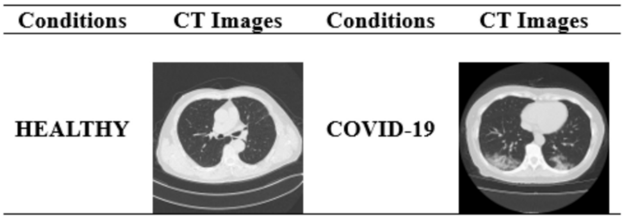


Fig. 1. Sample CT images from the datasets.

Materials and methods

Datasets

This study utilizes two primary datasets. First, the COVID-QU-Ex dataset created by Tahir et al. at Qatar University, includes a detailed set of 33,920 chest X-ray images comprising 11,956 COVID-19 cases, 11,263 cases of non-COVID infections (such as viral or bacterial pneumonia), and 10,701 normal cases[34]. Each image is accompanied by its corresponding ground-truth lung segmentation and infection masks, making it the largest publicly accessible dataset for lung and infection segmentation related to COVID-19[35]. This study uses lung and infection segmentation masks, along with data from the Montgomery and Shenzhen datasets, to detect and quantify lung infections. These tools provide crucial support to assess the severity and to monitor the progression of COVID-19. Table 1 describes the complete dataset details.

The second dataset consists of CT images gathered from Union and Liyuan Hospital. These images are divided into three categories: (i) 5705 non-informative CT images with no visible lung parenchyma for assessment, (ii) 4001 positive CT images displaying evident signs of COVID-19, and (iii) 9979 negative CT images with no infection related to COVID-19[35]. The CT dataset consists of a total of 8,000 images in the proposed work, which are evenly divided into normal and COVID-19 cases. Figure 1 presents sample CT scan images corresponding to two different conditions: Healthy and COVID-19. The left side displays a CT image of a healthy lung, characterized by clear lung fields without visible opacities. The right side shows a CT image of a COVID-19-affected lung, exhibiting ground-glass opacities and abnormalities indicative of infection. This visual comparison highlights the structural differences in lung appearance due to COVID-19. The training set includes 6,400 images, with 3,200 from each class, whereas the testing set comprises 1,600 images, again evenly distributed with 800 images for each class.

Pre-trained models

- In this study, a straightforward CNN deep network is applied to capture essential features from specific CXR biomarkers, enabling differentiation between healthy and COVID-19-affected CXR images. CNN-based architectures optimized for image recognition and classification are used, eliminating the reliance on manually designed features[36,57]. This end-to-end CNN model is built with key elements: convolutional layers paired with batch normalization and ReLU activation functions, max pooling layers to reduce spatial dimensions and remove redundancy, and a fully connected layer to merge the extracted features. Additionally, the model includes a softmax layer for output normalization, and a final classification layer that assigns probability scores for accurate image classification.
- AlexNet uses convolutional layers to extract spatial features from input images by applying learnable filters over theimage, and performing element-wise multiplications and summations. These operations compute output feature maps for all other channels over spatial dimensions from loading activations. It consists of convolutional layers with different filter sizes and channel counts for learning multi-level image information[37]. The ReLU activation function is used at the end of each convolutional or fully connected layer and its role is to prevent the vanishing gradient problem. Spatial dimensions are reduced by max-pooling layers, which provide computational efficiency and translation invariance. Finally, image classification is performed on the learned features via fully connected layers.
- The VGG-19 architecture is a deep CNN designed for image classification tasks. It features a series of 3 × 3 convolutional layers, each with a fixed receptive field, followed by max-pooling layers that reduce the spatial dimensions of the feature maps. The VGG-19 model contains of 19 layers in overall, which includes 16 convolutional layers with three fully connected layers, making it a deeper version of the original VGG-16 design[38]. This additional depth enhances the model's ability to learn more complex features from input images, contributing to improved classification performance.

- The ResNet-34, CNN architecture uses 34 layers designed to address the problem of vanishing gradients during training via residual learning. It incorporates skip connections, allowing the network to bypass one or more layers, effectively learning residual functions that ease the training of very deep networks. The architecture consists of multiple stacked residual blocks, each containing convolutional layers, batch normalization, and ReLU activations. This design enables the model to be both deep and computationally effective, for large-scale image classification tasks. The unique residual blocks of ResNet-34 enhance feature extraction, improving model accuracy while reducing training complexity. It is widely used as a backbone for transfer learning in various applications[39].

Vision transformer

The Vision Transformer (ViT) is a transformer-based deep learning architecture designed for image classification tasks. It employs self-attention (SA) mechanism and allows different parts of an image, represented as tokens, to capture relationships among various regions. These token representations are processed by the transformer encoder, which generates encoded outputs that are further refined through a feed-forward network to extract high-level features from the image representation. Finally, the processed outputs pass through a classification head to categorize the image based on its class[40,41].

A key advantage of the ViT model lies in its ability to interpret image data without relying on predefined spatial hierarchies or hand-crafted features. However, the ViT model is limited by its large image sizes[42], for methods such as patch overlapping and hybrid architectures that integrate CNNs with transformers have been introduced to improve ViT model efficiency on larger images[43]. Figure 2 illustrates the comprehensive approach used for COVID-19 image classification with the ViT model.

ViT Self-attention mechanism

The self-attention mechanism in ViT plays a crucial role by allowing the model to weigh the importance of different regions within an image, which enables it to capture long-range dependencies. This is in contrast to traditional convolutional layers, which typically focus on local features. By dividing the image into patches and treating these patches as tokens, ViT allows the model to capture global relationships and spatial hierarchies, even in larger images. It processes the input image by dividing it into non-overlapping patches, and then projects each patch linearly into a compact, lower-dimensional token representation[44].

If the dimensions of the image are $H \times W$ and the patch size is $P \times P$, the number of tokens can be expressed as $N = \left(\frac{H}{P}\right) \times \left(\frac{W}{P}\right)$. These tokens are subsequently projected into a higher-dimensional space to reserve intricate visual features. To capture global dependencies and connections between tokens, the SA mechanism is combined with a feed-forward neural network, which is applied to the token embeddings. Given an input of token embeddings X, the SA approach calculates attention weights and uses them to generate context-aware representations.

The SA mechanism involves query (Q), key (K), and value (V) linear transformations expressed as:

$$[q, k, v] = xU_{qkv} \tag{1}$$

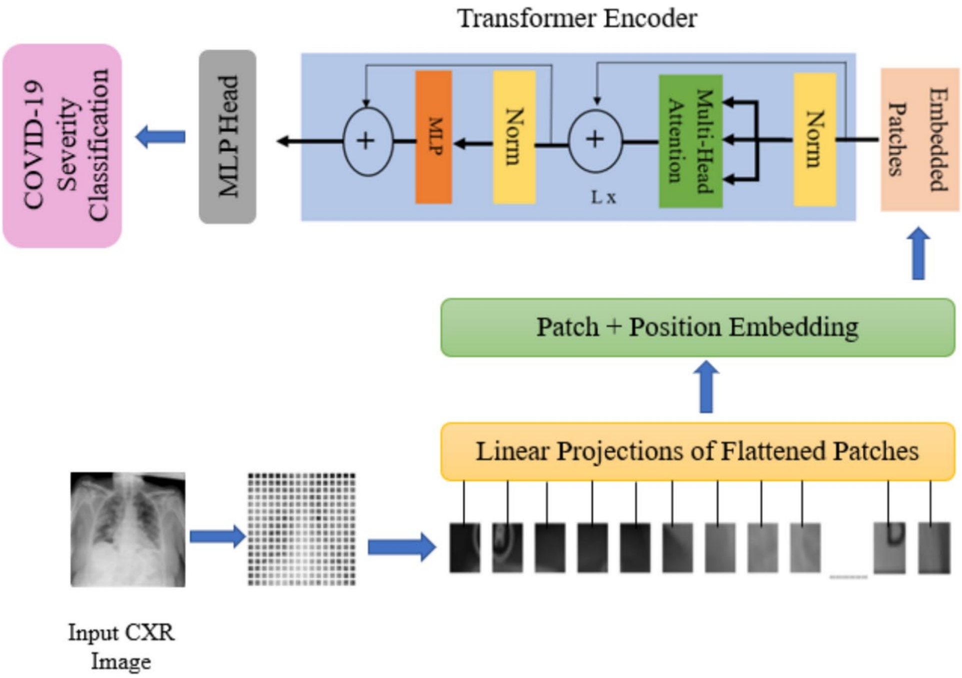

Transformer Encoder
COVID-19 Severity Classification
MLP Head
MLP
Norm
Multi-Head Attention
Norm
L x
Embedded Patches
Patch + Position Embedding
Linear Projections of Flattened Patches
Input CXR Image

Fig. 2. Classification of COVID-19 images using ViT model.

where $U_{qkv} \in R^{D\times 3D}{}_{h}$, and $x \in R^{N\times D}$ denotes the projected token embedding, where D is the dimensionality of the estimated space. The weighted sum is then calculated for all values in order. The attention weights (A_{ij}) depend on the pairwise relationship of the sequence's components, specifically the query (q_i) and key (k_j) representations. These weights are calculated via the softmax function of the scaled dot product between the queries and keys:

$$A = softmax\left(\frac{qk^T}{\sqrt{D_h}}\right) A \in R^{N\times N} \tag{2}$$

This process is carried out individually for each of the multiple attention heads, and the outputs are concatenated and linearly projected to provide the final attention output:

$$SA\left(z\right) = Av \tag{3}$$

The multi-head self-attention (MSA) mechanism is an expansion of single-head SA, where multiple SA procedures operate in parallel, and their results are concatenated:

$$MSA\left(z\right) = \left[SA_1\left(z\right); SA_2\left(z\right); \ldots\ldots; SA_k\left(z\right)\right] U_{msa} \tag{4}$$

The output from the transformer layers is fed into a classification head, generally comprising a linear transformation followed by softmax activation. This classification head yields the final probability scores for each class of the input[45].

Feature extraction using optimized vision transformer using Grey Wolf Optimization (GWO)

New findings in neural networks, especially in transformer architectures, indicate that performance tends to improve as the number of model parameters increases. However, it remains uncertain whether these findings hold in low-resource environments, as they are primarily based on data-rich scenarios. The efficiency of the model can be increased by reducing the transformer layer count. Therefore, to increase the performance of the transformer architecture, it is crucial to optimize hyperparameters such as the patch size, the number of attention heads, the number of encoder layers, the dimension of the feed-forward network, and the embedding dimensions. To optimize the hyperparameters of a transformer, the GWO algorithm is presented.

GWO is a nature-inspired optimization algorithm that emulates the hunting mechanism and social leadership ranking of grey wolves in the wild. It was introduced in[46] and has since become a popular algorithm for solving various optimization problems due to its simplicity and effectiveness in balancing exploration and exploitation. Grey wolves are considered apex predators and exhibit a well-defined social hierarchy that helps them effectively organize group hunting. GWO simulates this social behavior by categorizing wolves into four distinct classes: alpha (α), beta (β), delta (δ), and omega (ω), where each group plays aspecific role in the hunting process.

Tunning of ViT hyperparameters

Hyperparameter tuning is a critical process for optimizing the performance of ViT. GWO is utilized to explore and fine-tune essential parameters of the ViT architecture. Key parameters include patch size, embedding dimensions, the number of encoder layers, attention head count, and feed-forward layer dimensions. This systematic optimization enhances model efficiency and classification accuracy. The various hyperparameters which were considered for the lung images are shown in Table 2.

The optimization process involves initializing the positions of the grey wolves, evaluating their fitness based on the model's performance, and iteratively updating the positions to find the optimal hyperparameter set that minimizes classification error. The following steps outline the procedure for hyperparameter optimization using GWO.

1. Initialization of grey wolf positions

The positions of the grey wolves are initialized randomly within the search space defined by the ranges of these hyperparameters. Each grey wolf's position represents a candidate solution for the set of hyperparameters. For a given grey wolf i, its position vector can be defined as:

$$X_i = \left[P_i, H_i, L_i, E_i, F_i\right] \tag{5}$$

Hyperparameters	Range
Patch Size	[8, 16, 32]
Embedding dimensions	[128, 256, 512,768, 1024]
Number of layers in encoder	[6, 8, 10, 12, 14]
Count of attention heads	[8, 10, 12, 14, 16]
Dimension of feed forward layer	[128, 256, 512, 768, 1024, 2048]

Table 2. The range of hyperparameters of ViT model.

where P_i is the patch size, L_i is the number of layers in the encoder, H_i is the count of attention heads, F_i is the dimension of the feed-forward layer and E_i is the embedding dimension for the *i*-th wolf.

2. Fitness evaluation

The fitness function evaluates a given set of hyperparameters. The ViT fitness function can be based on cross-entropy loss. The fitness function f{X}_i) for a grey wolf is defined as:

$$f\left(X_i\right) = Loss_{val}\left(X_i\right) = -\frac{1}{N}\sum_{j=1}^{N} y_j \log(\widehat{y_{\text{j}}}) \tag{6}$$

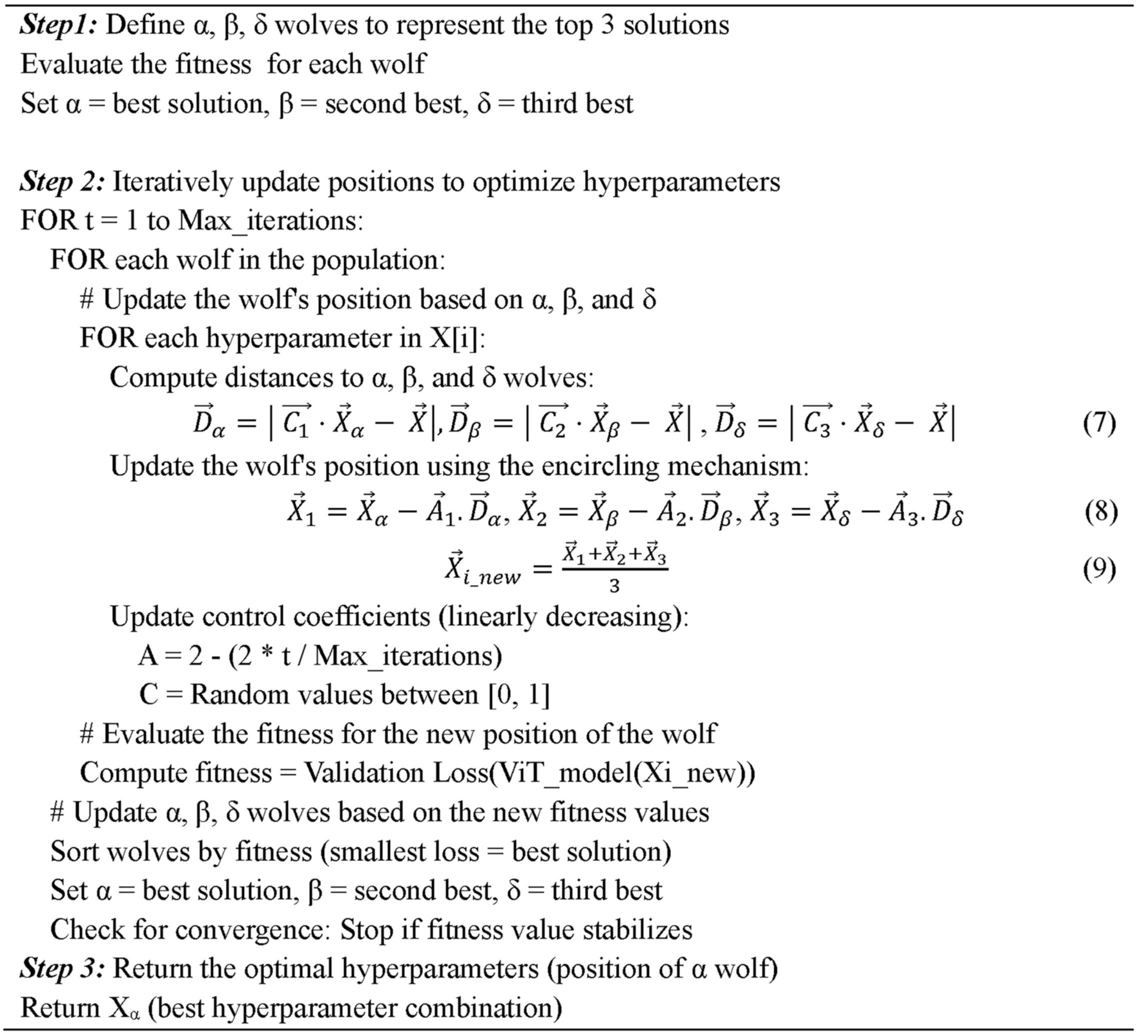
where N is the number of samples in the validation set, y_j is the true label for the *j*-th sample and $\widehat{y_{\text{j}}}$ is the predicted probability of the correct class for the *j*-th sample. The main objective of GWO is to minimize this fitness function, finding the hyperparameters that result in the lowest classification error.

3. GWO algorithm steps for ViT hyperparameters optimization

Table 3 shows the effective optimization of VIT hyperparameters using GWO algorithm to improve the classification performance.

Feature selection using Particle Swarm Optimization (PSO)

In this work, Particle Swarm Optimization (PSO) is employed to select the most significant features extracted from the ViT model. PSO, inspired by the collective behavior of swarms in nature, identifies optimal solutions by iteratively refining the feature set based on individual and shared experiences of agents[47]. This approach enhances the model's performance by focusing on the most relevant features, to reduce dimensionality, and to



```
Step1: Define α, β, δ wolves to represent the top 3 solutions
Evaluate the fitness  for each wolf
Set α = best solution, β = second best, δ = third best

Step 2: Iteratively update positions to optimize hyperparameters
FOR t = 1 to Max_iterations:
    FOR each wolf in the population:
        # Update the wolf's position based on α, β, and δ
        FOR each hyperparameter in X[i]:
            Compute distances to α, β, and δ wolves:
                $\vec{D}_\alpha = |\vec{C_1} \cdot \vec{X}_\alpha - \vec{X}|, \vec{D}_\beta = |\vec{C_2} \cdot \vec{X}_\beta - \vec{X}|, \vec{D}_\delta = |\vec{C_3} \cdot \vec{X}_\delta - \vec{X}|$    (7)
            Update the wolf's position using the encircling mechanism:
                $\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 . \vec{D}_\alpha, \vec{X}_2 = \vec{X}_\beta - \vec{A}_2 . \vec{D}_\beta, \vec{X}_3 = \vec{X}_\delta - \vec{A}_3 . \vec{D}_\delta$    (8)
                $\vec{X}_{i\_new} = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3}$    (9)
            Update control coefficients (linearly decreasing):
                A = 2 - (2 * t / Max_iterations)
                C = Random values between [0, 1]
        # Evaluate the fitness for the new position of the wolf
        Compute fitness = Validation Loss(ViT_model(Xi_new))
    # Update α, β, δ wolves based on the new fitness values
    Sort wolves by fitness (smallest loss = best solution)
    Set α = best solution, β = second best, δ = third best
    Check for convergence: Stop if fitness value stabilizes
Step 3: Return the optimal hyperparameters (position of α wolf)
Return $X_\alpha$ (best hyperparameter combination)
```

Table 3. Hyperparameter Optimization of ViT using GWO.

minimize redundant information. The combination of PSO with the optimized ViT model ensures efficient and accurate classification, to improve the overall robustness of the model.

Swarm optimization is a metaheuristic approach inspired by the collective behaviors seen in social insects, such as ants and bees, to address complex problem-solving. This technique relies on collaboration and data-sharing among individual agents to explore the search space and identify optimal solutions[48]. Agents continuously refine their positions based on both personal and collective experiences, resulting in intelligent group dynamics. Each agent adjusts its direction and speed to pursue the best solution encountered to date. This approach allows for simultaneous searches, helps to prevent entrapment in local optima, and demonstrates resilience by efficiently managing noise and uncertainties in challenging environments[49,50].

In PSO, each particle, or agent, corresponds to a unique position within the feature space. If there are n parameters requiring optimization, a particle's position can be defined as:

$$X_i = [x_{i1}, x_{i2}, \ldots\ldots, x_{in}] \tag{10}$$

The population size, represented by m, indicates the total number of candidate solutions, as follows:

$$\mathrm{X} = [X_1, X_2, \ldots\ldots, X_m] \tag{11}$$

As each particle navigates over the exploration space to find the optimum solution, it calculates its path and continually updates its position as follows:

$$X_i\,(t+1)\ =\ V_i\,(t) + V_i\,(t+1) \tag{12}$$

Here, V_i is the velocity vector and t represents the iteration number, guiding the particle's movement within the search space.

Taking these parameters into account, the updated velocity for a particle can be expressed as

$$V_i\,(t+1) = \omega V_i\,(t) + C_1\,(P_i - X_i\,(t))\,R_1 + C_2\,(g - X_i\,(t))\,R_2 \tag{13}$$

In this equation, ω denotes the inertia weight, P_i signifies the optimal solution (local best) that an individual particle has identified up to this point, and g represents the best solution (global best) identified across the entire swarm. The acceleration coefficients usually fall anywhere between $\mathrm{C}_1 = 0$ and $\mathrm{C}_2 = 4$, and R_1 and R_2 are diagonal matrices containing random values between 0 and 1. This setup directs particle movement along a trajectory shaped by their own experiences and the collective knowledge of the swarm.

Multi-Phase framework for classification

The proposed methodology adopts a multi-phase classification strategy to improve accuracy in differentiating COVID-19 severity levels by segmenting the task into three manageable phases. This approach addresses the challenges in distinguishing mild and moderate cases, which exhibit subtle radiological differences, while leveraging well-defined distinctions for severe cases.

1. Phase-wise Classification Approach

Phase 1: Binary classification to separate Normal cases from COVID-19 cases (including mild, moderate, and severe).

Phase 2: Binary classification to distinguish Mild and Moderate cases from Severe cases.

Phase 3: Fine-grained classification to differentiate Mild cases from Moderate cases.

2. Severity Quantification using Extent of Infection (EI)

The Extent of Infection (EI) ratio is used to quantitatively define severity levels based on the area of infection relative to the total lung area:

$$\mathrm{EI} = \frac{\text{Infected area (pixels)}}{\text{Total Lung area (pixels)}} \times 100 \tag{14}$$

The severity thresholds are:

o **Mild**: $0 \le \mathrm{EI} < 25\%$
o **Moderate**: $25 \le \mathrm{EI} < 50\%$
o **Severe**: $\mathrm{EI} \ge 50\%$

3. Framework implementation:

The infected and total lung area are computed from dataset's infection masks and ground truth masks. These masks are used to calculate the EI metric, which is the basis for distinguishing the severity classes. The calculated EI values are integrated into the multi-phase classification framework to guide the classification process.

Feature Extraction: Features are extracted using an optimized ViT model.

Hyperparameter Optimization: GWO is applied to tune ViT hyperparameters, including batch size, learning rate, and attention heads.

Feature Selection: PSO is employed to identify significant features for each classification phase.

Classification: The optimized features are input into a tunned MLP for severity classification across all phases.

Proposed model

The proposed model in Fig. 3 aims to classify multiple classes effectively from CXR and CT images via an integrated approach involving the ViT model, GWO, PSO, and Multi-Layer Perceptron (MLP) classifier. This process is carried out through the steps outlined in Algorithm 1.

- The vision transformer model hyperparameters such as the patch size, embedding dimensions, number of layers in the encoder, number of attention heads, dimensions of feedforward layers are optimized via the Grey Wolf Optimizer (GWO) to determine the optimal configuration for severity classification.
- The ViT model is trained with the best hyperparameters and used to capture features from CXR and CT COVID-19 images, which represent different severity levels (mild, moderate, and severe).
- Particle Swarm Optimization (PSO) selects the most significant features from the ViT output, to reduce redundancy and improve model efficiency.
- A Multi-Layer Perceptron (MLP) classifier, adapted through tuning is trained on the PSO-optimized features to classify severity levels, and the model performance is evaluated via metrics such as accuracy, precision, recall and F1-score.

Evaluation metrics

The evaluation metrics include accuracy, which indicates the percentage of precise predictions within all predictions, recall (sensitivity), which reflects how well the model identifies actual positive samples. Precision reflects the proportion of true positives among the predicted positives, whereas the F1-score balances precision and recall, offering a comprehensive performance measure[52]. The ROC curve analyses the trade-off between the true positive rate (TPR) and false positive rate (FPR) across different thresholds. Finally, the AUC reflects the model's ability to differentiate between positive and negative classes, with higher values indicating superior performance.

Results and discussion

Dataset preparation and preprocessing

In this study, the COVID-QU-Ex dataset is specifically designed to support infection segmentation tasks from chest X-ray images (CXR), particularly to focus on COVID-19 cases. The dataset includes COVID-19 infection and ground-truth segmentation masks. These segmentation data are crucial for models aiming to localize and quantify lung infection areas.

The provided segmentation masks enable direct analysis of infections without the need for additional segmentation models. Using these masks, the Extent of Infection (EI) can be calculated to assess the severity of COVID-19 infection. The infection mask identifies the infected area, whereas the lung mask provides the total lung area. The ratio of the infected area calculates the EI value to the total lung area, quantifying the extent of lung infection as shown in Eq. 14. Figure 4 demonstrates the EI for a CXR image.

The EI value is then used to classify the severity of the patient's condition into three categories: mild, moderate, and severe. Specifically, an EI of 0–25% indicates mild infection, 25–50% specifies moderate infection and greater than 50% is classified as severe. This approach offers an efficient and reliable method to assess

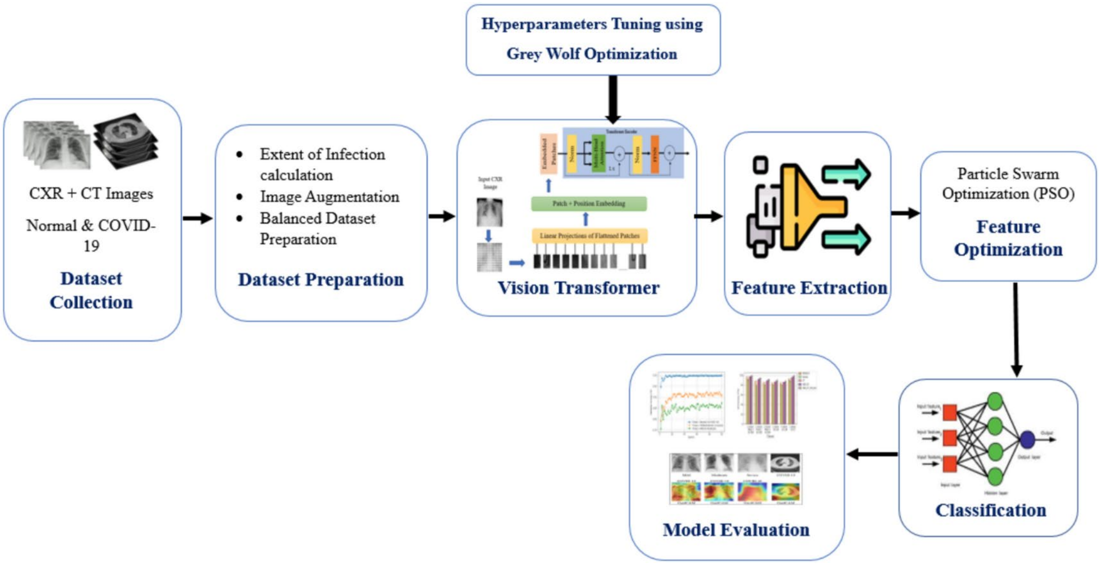
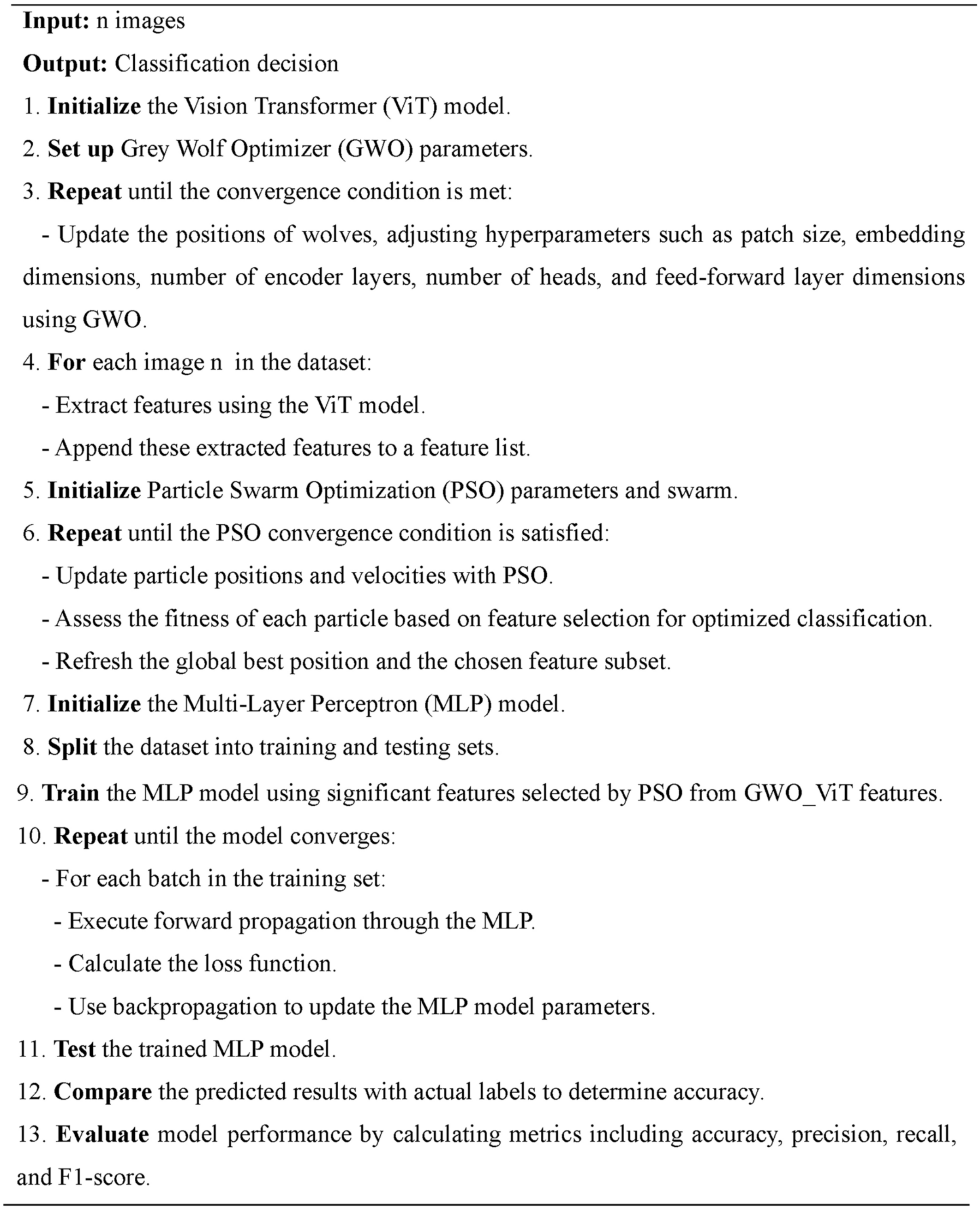


Fig. 3. Proposed framework.



```
Input: n images
Output: Classification decision
1. Initialize the Vision Transformer (ViT) model.
2. Set up Grey Wolf Optimizer (GWO) parameters.
3. Repeat until the convergence condition is met:
   - Update the positions of wolves, adjusting hyperparameters such as patch size, embedding dimensions, number of encoder layers, number of heads, and feed-forward layer dimensions using GWO.
4. For each image n  in the dataset:
   - Extract features using the ViT model.
   - Append these extracted features to a feature list.
5. Initialize Particle Swarm Optimization (PSO) parameters and swarm.
6. Repeat until the PSO convergence condition is satisfied:
   - Update particle positions and velocities with PSO.
   - Assess the fitness of each particle based on feature selection for optimized classification.
   - Refresh the global best position and the chosen feature subset.
7. Initialize the Multi-Layer Perceptron (MLP) model.
8. Split the dataset into training and testing sets.
9. Train the MLP model using significant features selected by PSO from GWO_ViT features.
10. Repeat until the model converges:
   - For each batch in the training set:
     - Execute forward propagation through the MLP.
     - Calculate the loss function.
     - Use backpropagation to update the MLP model parameters.
11. Test the trained MLP model.
12. Compare the predicted results with actual labels to determine accuracy.
13. Evaluate model performance by calculating metrics including accuracy, precision, recall, and F1-score.
```

Algorithm 1. Proposed Model.

COVID-19 severity directly from CXR images using the given infection and lung masks, which eliminates the need for additional segmentation algorithms.

Table 4 and Fig. 5 effectively convey the dataset distribution and COVID-19 severity level of infection and the normal class. Class imbalance in the COVID-QU-Ex dataset requires actual supervision to prevent the model from supporting the majority class. Image augmentation techniques such as rotation, flipping, and scaling create diverse training samples, which improves model robustness. These strategies enhance ViT model performance with more balanced input data, to reduce overfitting, and promote better generalization.

Multi-phase framework for COVID-19 severity classification

Clinical practice indicates that to identify severe COVID-19 cases is relatively straightforward. However, to differentiate mild and moderate cases presents a greater challenge. A multi-phase classification strategy is implemented to increase classification accuracy across different severity levels.

Phase 1: Distinguishes between normal patients and those with mild, moderate, or severe COVID-19

Phase 2: Separates mild and moderate cases from severe cases of COVID-19

Phase 3: Further differentiates between mild and moderate COVID-19 cases

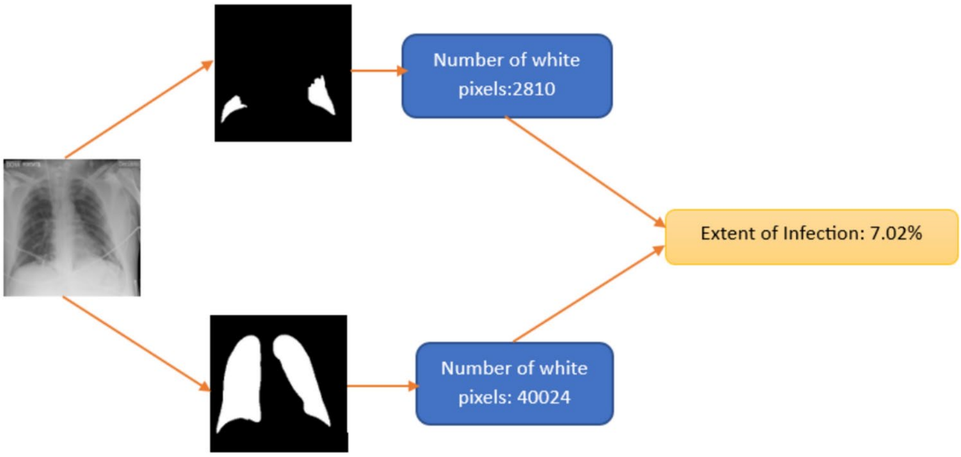

Number of white pixels:2810
Extent of Infection: 7.02%
Number of white pixels: 40024

Fig. 4. Proposed framework to determining the extent of infection in CXR images.

Dataset	Normal	Mild	Moderate	Severe	Total
Training	932	810	632	888	3262
Testing	291	203	158	222	872
Total	1223	1013	790	1110	4136

Table 4. Dataset distribution of CXR images.

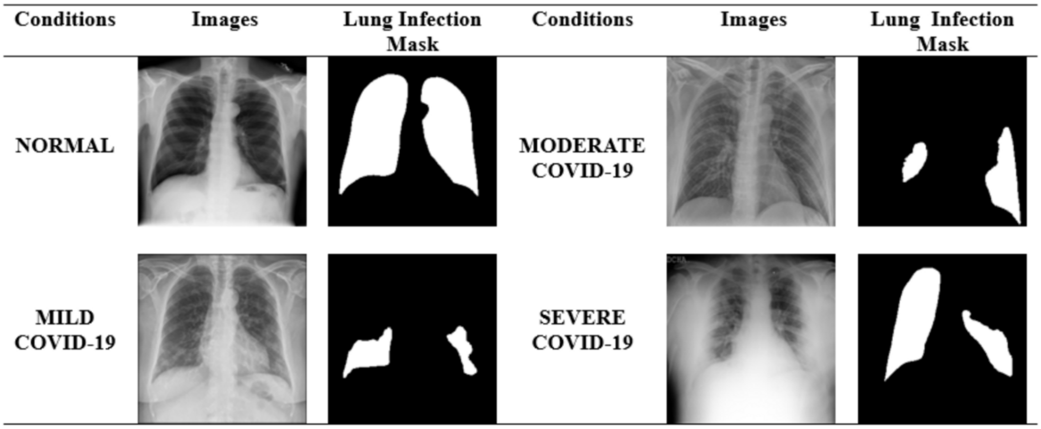


Fig. 5. COVID-19 Severity Levels in CXR Images with Infection Masks.

The framework is presented in Fig. 6. This multi-phase approach promotes more specialized classification, to improve both the generalizability and robustness of the model. This hierarchical strategy optimizes diagnostic performance, to support precise clinical decisions and effective patient management.

Optimized ViT model with PSO based feature selection

Table 5 summarizes the feature dimensions extracted from a ViT model applied to multi-class classification tasks, with and without optimization. The standard ViT generates 768-dimensional features, whereas optimization via GWO reduces the feature size to 512, which improves efficiency. Further refinement with a combined GWO and PSO approach results in smaller feature dimensions, ranging from 221–256 in all classes, which depend on the dataset and task. The best feature size was selected after 5 independent trials, to ensure consistent performance and reliability. PSO enhances the optimization process by effectively exploring the feature space to identify optimal subsets, and to check a compact representation to preserve classification accuracy. These optimizations improve computational efficacy and reduce the risk of overfitting, particularly in multi-class classification.

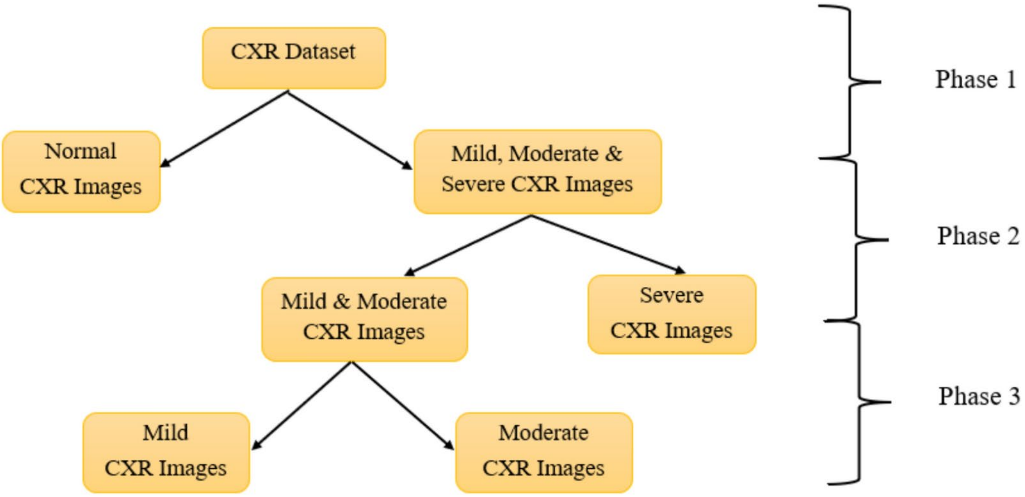


Fig. 6. Multi-Phase framework for COVID-19 severity detection.

Dataset	Classes	Features Dimension of Models		
		ViT	GWO_ViT	GWO_ViT_PSO
CXR Images	2 Classes	768	512	224
	3 Classes	768	512	238
	4 Classes	768	512	256
CT Images	2 Classes	768	512	221

Table 5. Dimensions of the extracted features from the optimized ViT model.

Approaches	Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Pre-trained Models	CNN	83.87	83.88	83.87	83.65
	ResNet34	84.22	84.26	84.22	84.36
	VGG19	82.64	82.68	82.64	82.77
	AlexNet	79.77	79.79	79.77	79.25
Vision Transformer Models	ViT	84.88	84.82	84.88	84.67
	GWO_VIT	86.33	86.28	86.33	86.39
Proposed Models	VIT_PSO_MLP	85.87	85.85	85.87	85.58
	VIT_PSO_SVM	85.05	85.17	85.05	85.02
	GWO_VIT_PSO_MLP	87.84	87.81	87.83	87.77
	GWO_VIT_PSO_SVM	87.27	87.14	87.15	87.19

Table 6. Model performance of the four classes of CXR images.

Results of the COVID-QU-Ex images dataset

Classification results for the four classes

Table 6 presents the performance metrics of various models applied to classify CXR images into four classes: normal, mild, moderate, and severe COVID-19 cases. The enhanced performance is driven by a sequential process: GWO optimizes ViT for better feature extraction, PSO selects the most significant features, and MLP makes the final decision, to ensure precise and efficient classification. These models are evaluated via key metrics such as accuracy, precision, recall, and F1 score, which provide a complete understanding of each model's effectiveness. The first group, pre-trained Models, includes CNN-based architectures such as CNN, ResNet34, VGG19, and AlexNet. Among these, ResNet34 performs the best with an accuracy of 84.22%, whereas AlexNet shows the lowest performance with 79.77% accuracy. CNN and VGG19 yield comparable results, with accuracy of 83.87% and 82.64%, respectively. The second category focuses on ViT models. The basic ViT model achieves an accuracy of 84.88%, which is slightly better than those of the pre-trained CNN models. When optimized via the GWO, the model accuracy improved to 86.33%, which represents a significant increase in performance through hyperparameter optimization. Among the proposed models, the integration of PSO with ViT and an MLP architecture achieved an accuracy of 85.87%, whereas the combination of ViT, PSO, and Support Vector

Approaches	Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Pre-trained Models	CNN	83.08	83.11	83.08	83.16
	ResNet34	84.18	84.21	84.19	84.07
	VGG19	82.16	82.19	82.16	82.24
	AlexNet	80.20	80.28	80.22	80.17
Vision Transformer Models	ViT	85.68	85.66	85.68	85.67
	GWO_VIT	87.55	87.54	87.55	87.52
Proposed Models	VIT_PSO_MLP	86.49	86.55	86.48	86.45
	VIT_PSO_SVM	86.19	86.20	86.19	86.16
	GWO_VIT_PSO_MLP	89.74	89.75	89.74	89.70
	GWO_VIT_PSO_SVM	89.11	89.04	89.12	89.04

Table 7. Model performance of three classes of CXR images.

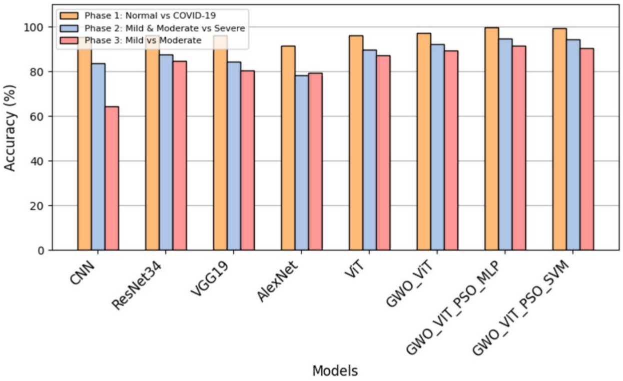


Fig. 7. Comparative analysis of model performance across COVID-19 severity phases.

Machine (SVM) yielded slightly lower performance, with an accuracy of 85.05%. The most successful model was the GWO-optimized ViT combined with PSO and MLP, which attained the highest accuracy of 87.84% a precision of 87.81%, a recall of 87.83%, and an F1-score of 87.77%, followed closely by the GWO_ViT_PSO_SVM model.

Classification results for the three classes

Table 7 shows the classification results for the three-class CXR image dataset (mild, moderate, and severe COVID-19 cases), which reveals strong performance across different models. Among the pre-trained CNNs, ResNet34 achieved precision, recall, and F1 score of approximately 84%, followed by VGG19. The standard CNN model and AlexNet showed slightly lower effectiveness, to performance metrics. The ViT outperforms most CNN models, which achieves balanced scores of about 85% across accuracy, precision, recall, and F1. Optimizing the ViT using the GWO led to notable improvements, which reached around 87% across all key metrics, demonstrating the effectiveness of the GWO in increasing the model performance.

Among the proposed models, combining ViT with PSO and MLP yielded robust results, with an accuracy of 86.49%. A similar combination with SVM performed correspondingly. However, the GWO_ViT_PSO_MLP model achieved the highest overall performance, with accuracy of 89.74%, precision of 89.75%, recall of 89.74%, and F1 score of 89.70%. The GWO_ViT_PSO_SVM model also performed impressively, with an accuracy of 89.11%. These results highlight the efficacy of metaheuristic algorithms such as GWO and PSO in combination with deep learning models to improve three-class classification performance on CXR images.

Classification results for the two classes

The classification results for the two-class CXR images are presented across three phases: Phase 1 (normal vs. COVID-19), Phase 2 (mild & moderate vs. severe), and Phase 3 (mild vs. moderate) as shown in Fig. 7 and Table 7. Each phase evaluates a range of models, which includes pre-trained models, ViT, and proposed models that combine ViT with optimization techniques.

Phase 1 (Normal vs. COVID-19): In this phase, the pre-trained models demonstrate strong performance with ResNet34 and attain an accuracy of 96.03%, followed closely by VGG19 and CNN. AlexNet has a slightly lower accuracy of 91.44%. The ViT model outperforms the pre-trained models with an accuracy of 96.25%, and when optimized with GWO, it reaches an accuracy of 97.11%. Among the proposed models, the combination of GWO_ViT with PSO and MLP delivers the highest performance, with an accuracy of 99.79%, precision, recall, and F1 score metrics indicate near-perfect classification in this phase. The GWO_ViT_PSO_SVM model also yields excellent results, with 99.14% accuracy.

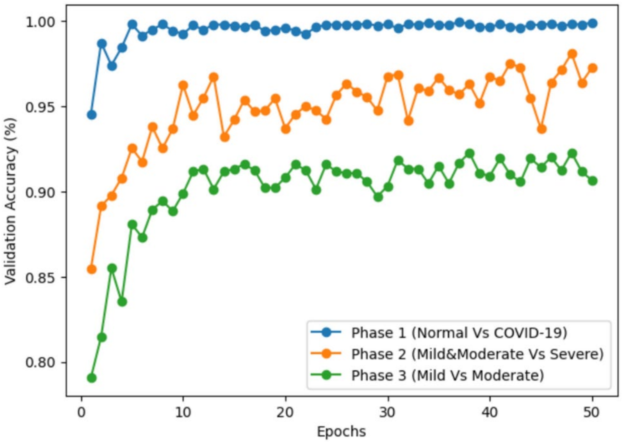


Fig. 8. COVID-19 severity classification: Validation accuracy across multiple phases.

Approaches	Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Pre-trained Models	CNN	91.58	91.62	91.51	91.57
	ResNet34	92.09	92.11	92.10	92.22
	VGG19	93.24	93.22	93.21	93.20
	AlexNet	89.21	89.22	89.20	89.24
Vision Transformer Models	ViT	94.43	94.45	94.44	94.42
	GWO_VIT	96.66	96.65	96.69	96.62
Proposed Models	VIT_PSO_MLP	95.91	95.90	95.92	95.89
	VIT_PSO_SVM	95.33	95.34	95.38	95.31
	GWO_VIT_PSO_MLP	98.89	98.88	98.85	98.92
	GWO_VIT_PSO_SVM	98.04	98.05	98.06	98.02

Table 8. Model performance for two class classification of CT images.

Phase 2 (Mild & Moderate vs. Severe): In Phase 2, the task of classification between mild & moderate vs. severe shows lower accuracy overall than in Phase 1. Among the pre-trained models, ResNet34 performs the best, with an accuracy of 87.47%, whereas AlexNet falls behind at 78.24%. The ViT model improves upon these with 89.77% accuracy, and when optimized with GWO, the accuracy improves to 92.11%. Among the proposed models designed to improve performance, the GWO_ViT_PSO_MLP model achieved the highest accuracy of 94.56%, along with high precision, recall, and F1 score. The GWO_ViT_PSO_SVM model also performed well, with an accuracy of 94.28%.

Phase 3 (Mild vs. Moderate): This phase, which focuses on differentiating between mild and moderate COVID-19 cases, is the most challenging phase. In the pre-trained models, CNN achieves the lowest accuracy at 64.24%, and even ResNet34 reaches 84.67% accuracy. The ViT model provides better results, with 87.02% accuracy, whereas GWO_ViT achieves 89.44%. In the proposed models, GWO_ViT_PSO_MLP leads with an accuracy of 91.45%, followed by GWO_ViT_PSO_SVM, which achieves 90.28% accuracy. These results suggest that optimization algorithms significantly improve classification performance in this challenging phase.

Figure 8 illustrates the validation accuracy over 50 training epochs for three classification phases in the COVID-19 severity classification task. Phase 1 (Normal vs. COVID-19) achieves the highest and most stable accuracy, nearing perfection early in the training process. Phase 2 (Mild & Moderate vs. Severe) demonstrates a steady improvement, stabilizing at a high accuracy level. Phase 3 (Mild vs. Moderate) shows the most gradual increase, reflecting the challenge of distinguishing between these closely related severity levels. The overall trend across all phases indicates consistent improvement in the model's performance with increased training epochs, validating its ability to adapt and generalize effectively for severity classification.

Results of the CT images for the COVID-19 classification dataset

Table 8 illustrates the classification outcomes for the two-class COVID-19 CT scan dataset and reveals the performance of various models based on accuracy, precision, recall, and F1 score. Among the pre-trained models, VGG19 performed the best, with accuracy and precision values of 93.24%. ResNet34 exhibited slightly lower performance, whereas the standard CNN and AlexNet models exhibited comparatively lower results, with an accuracy of 89.21%. The ViT showed significant improvements over the CNN-based models, reaching around 94% across all metrics. After the GWO, the ViT performance increased considerably and achieved 96.66% accuracy, which highlights the advantage of using optimization techniques.

In the proposed models, the integration of ViT with PSO and classifiers further improved the results. The ViT_PSO_MLP model achieved 95.91%, whereas the ViT_PSO_SVM model performed slightly lower at 95.33%. The most successful models were as follows: the GWO_ViT_PSO_MLP model reached the highest performance of 98.89% accuracy, followed by the GWO_ViT_PSO_SVM model at 98.04% accuracy. These results demonstrate the effectiveness of the hybrid approach, ViT, GWO, and PSO combination for the precise classification of COVID-19 cases via CT images.

Grad-CAM visualization of COVID-19 severity

Figure 9 displays CXR and CT images of COVID-19 patients at varying severity stages—mild, moderate, and severe—along with their corresponding Grad-CAM visualizations, which illustrate the areas the deep learning model focused on during predictions. The highlighted regions in red and yellow on the Grad-CAM images represent the area's most relevant to the model's decision-making process. In mild cases, the emphasis is on localized lung regions, while moderate and severe cases exhibit broader highlighted areas, indicating greater lung involvement with increasing severity. These visualizations provide insight into the model's predictions by identifying the lung regions that played a key role in classifying the severity of the infection.

Discussion

Figure 10 illustrates the comparative performance of various deep learning models across multiple classification tasks using CXR and CT images. The models analyzed include ResNet34, AlexNet, ViT, GWO_ViT, and GWO_ViT_PSO_MLP, with accuracy evaluated for 2-class, 3-class, and 4-class classifications. Among these, the GWO_ViT_PSO_MLP model consistently outperforms others, achieving near-perfect accuracy of 99.14% in 2-class CXR classification and 98.89% in 2-class CT classification. These results highlight the effectiveness of hybrid approaches that integrate ViT with optimization techniques like GWO and PSO, which enhance model performance by fine-tuning parameters to achieve maximum accuracy.

Figure 11 provides a comprehensive visual analysis of the model's classification performance through ROC curves, illustrating its effectiveness in distinguishing between COVID-19 severity levels using CXR images. In Fig. 11(a), the model achieves AUC values of 0.99, 0.98, 0.93, and 0.97 for the four severity classes—normal, mild, moderate, and severe, respectively. These results indicate strong classification capabilities, though the moderate class presents some difficulty with a slightly lower AUC. Similarly, Fig. 11(b) evaluates the model in a three-class scenario (mild, moderate, and severe), achieving AUC values of 0.97, 0.89, and 0.98, respectively. While the model performs exceptionally well for the mild and severe classes, the moderate class remains challenging, as

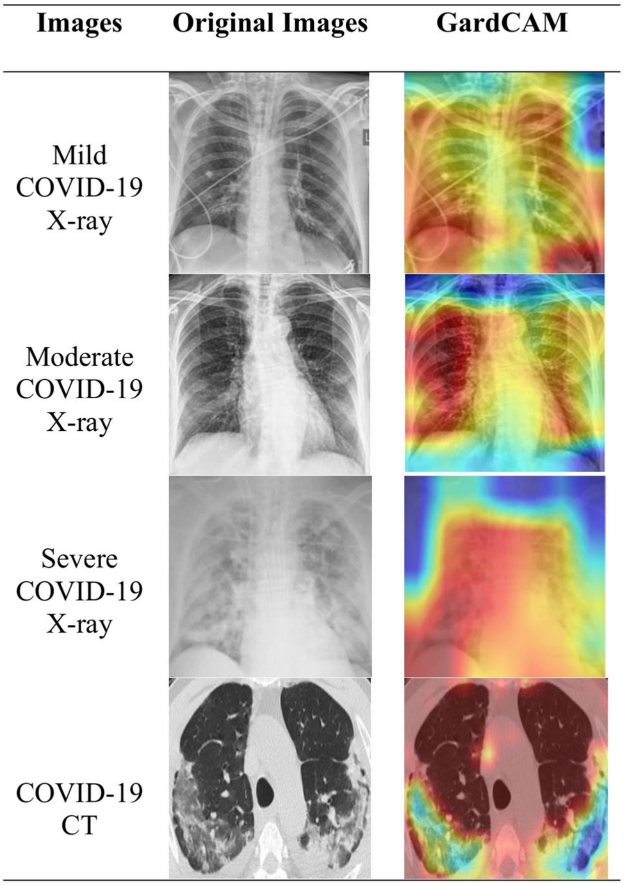


Fig. 9. Grad-CAM visualization of COVID-19 severity in images.

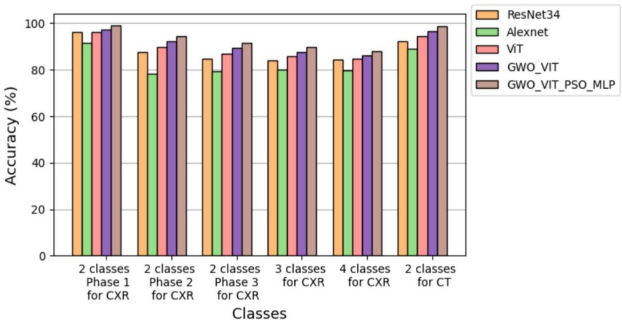
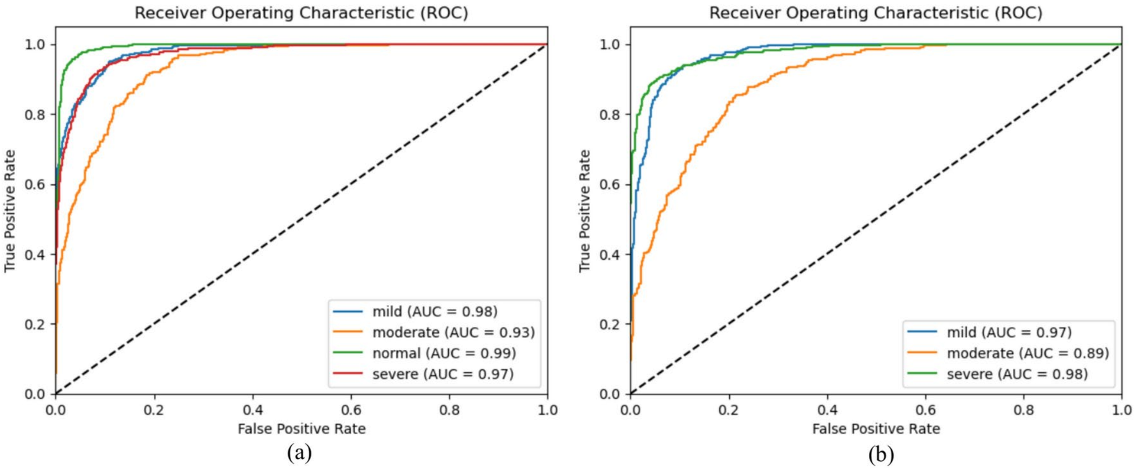


Fig. 10. Performance comparison of deep learning models in multi-Class CXR and CT image classification.



(a) (b)

Fig. 11. ROC curves for 4-class and 3-class severity classification of CXR images.

reflected by its comparatively lower AUC. These findings highlight the robustness of the ViT-based architecture in capturing long-range dependencies and global context within images. Combined with high accuracy metrics, the results validate the model's reliability and clinical relevance, while also identifying the moderate class as an area for further improvement.

Hypothesis, advantages, and limitations of the developed method

The proposed method hypothesizes that integrating ViT with metaheuristic optimization algorithms, such as GWO for hyperparameter tuning and PSO for feature selection, enhances the accuracy and efficiency of COVID-19 severity classification using multimodal medical images (CXR and CT). ViT self-attention mechanism captures both global and local features, addressing CNN limitations, while the multi-phase classification strategy effectively differentiates severity levels, including challenging cases like mild Vs. moderate COVID-19. The results validate this hypothesis, with the GWO_ViT_PSO_MLP model achieving 99.14% accuracy for 2-class CXR and 98.89% for 2-class CT classification, outperforming CNN-based models like ResNet34 and VGG19. The integration of GWO and PSO ensures optimal hyperparameter selection and reduced feature redundancy, boosting efficiency and robustness.

However, the high computational cost and challenges in distinguishing subtle differences, such as between mild and moderate cases, limit its applicability in resource-constrained settings. Future work could address these issues by exploring lightweight ViT variants to improve practicality and scalability in diverse clinical environments. Table 9 provides a detailed comparison of previous works based on techniques, preprocessing steps, datasets, evaluation metrics, and key advantages and disadvantages.

Conclusion

This study proposed a hybrid framework for COVID-19 severity classification using CXR and CT images, integrating Vision Transformers (ViT) with metaheuristic optimization techniques to address limitations in existing approaches. The framework leveraged Grey Wolf Optimizer (GWO) for hyperparameter tuning and Particle Swarm Optimization (PSO) for feature selection, resulting in significant improvements in accuracy and computational efficiency. The GWO_ViT_PSO_MLP model achieved outstanding performance, with 99.14% accuracy for 2-class CXR classification, 89.74% for 3-class CXR classification, 87.84% for 4-class CXR classification and 98.89% accuracy for 2-class CT classification, outperforming traditional CNN-based models

Reference	Techniques Used	Preprocessing	Dataset	Evaluation Measures	Advantages	Disadvantages
Gao et al.[13]	Vision Transformer	Normalization	MIA-COV19 competition	Accuracy: 78.8% F1 score: 76%	State-of-the-art for CT-based classification	Computationally expensive
Mondal et al.[14]	Explainable ViT	Image resizing, normalization	COVID-19 CXR COVIDx-CT-2A	Accuracy: CXR-96% CT-98.1%	Improved explainability for clinicians	Focused on binary classification
Hassan et al[18]	Transfer learning with DCNN and BCE classifier ResNet50	Image resizing, normalization	COVIDx CT	Accuracy: 99.07%	Achieves high accuracy using transfer learning	Limited to binary classification tasks
Singh et al.[19]	CNN + ViT hybrid model	Normalization, patch embedding	Mendeley CXR images	Accuracy: 97.61% F1: 95%	Balances local and global feature extraction	Computational overhead due to a hybrid model
Taye et al.[27]	ViT for COVID-19 severity classification	Data augmentation	COVID-CT SARS-CoV-2 CT-scan COVID-19 CT Lung	Accuracy: 98% Recall: 98%	Effective for severity classification	High computational demand
Abdo et al.[50]	Chi-Square and PSO for feature selection	-	Heart diseases Kaggle dataset	Accuracy: 83.3%	Effective feature selection for small datasets	Limited scalability for large datasets
Balasubramaniam et al.[51]	Ensemble model-SVM, CNN, optimized NN, RF	-	-	Attained higher accuracy in COVID-19 finding	High Efficacy	This method does not prioritise categorical variation
Imagawa et al.[53]	Fine-tuning a pre-trained ViT	Image resizing, normalization	NIH Chest X-ray14 BIMCV	AUC-76.1% GRAD CAM	Leveraging pre-training significantly enhances model accuracy and reliability	Lack of generalization to diverse pathologies
Balasubramaniam et al.[54]	ResUnet + Chef-based optimization	Data augmentation	Retinal fundus images from Kaggle	Accuracy: 90%	High accuracy for blood vessel segmentation	High computational requirements for optimizer
Singh et al.[55]	U-Net, Conv-Caps, and DenseNet201	-	COVID-19 Radiography Dataset	IoU: 0.979 Accuracy:93.98%	High segmentation accuracy Effective classification	High computational cost Limited generalizability
Proposed Work	GWO Optimized ViT	Image resizing, normalization	COVID-QU-Ex, CT scans for COVID-19 Classification	Accuracy CXR-99.14% CT-98.89% GRAD CAM	Improves accuracy with ViT and optimization	High computational cost

Table 9. Techniques, datasets, and performance metrics for multi-disease classification.

such as ResNet34 and VGG19. These improvements can be attributed to the ViT self-attention mechanism, which captures both global and local features, addressing the limitations of CNNs in modeling long-range dependencies and complex structures in medical images.

The integration of ViT with metaheuristic algorithms outperforms existing methods, particularly indistinguishing challenging severity levels, such as mild Vs. moderate COVID-19 cases. Compared to prior approaches relying on CNNs, attention-based hybrid models, or handcrafted feature selection, the proposed framework exhibits superior accuracy by effectively modeling global and local features. Moreover, the multi-phase classification strategy further enhances prediction precision, particularly in distinguishing between closely related severity categories. Overall, the proposed framework provides a scalable and efficient solution for automated COVID-19 severity detection, demonstrating its applicability across both CXR and CT modalities. Future research could focus on reducing the computational complexity of ViT models through lightweight variants and extending the framework to broader medical imaging applications to enhance clinical decision-making.

Data availability

The datasets used in this article are publicly available on Kaggle. The URL of the datasets are https://www.kaggle.com/datasets/anasmohammedtahir/covidqu and https://www.kaggle.com/datasets/azaemon/preprocessed-ct-scans-for-covid19.

Received: 6 November 2024; Accepted: 14 April 2025
Published online: 22 April 2025

References

1. Wong, H. Y. F. et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. *Radiology* **296**, E72–E78 (2020).
2. Carbone, F. et al. Chest X-ray severity score Brixia: From marker of early COVID-19 infection to predictor of worse outcome in internal medicine wards. *Eur. J. Clin. Invest.* **53**(2), e13908. https://doi.org/10.1111/eci.13908 (2023).
3. Antony, M. et al. Challenges of AI-driven diagnosis of chest X-rays transmitted through smartphones: A case study in COVID-19. *Sci. Rep.* **13**, 18102. https://doi.org/10.1038/s41598-023-44653-y (2023).
4. Chamberlin, J. H. et al. Automated diagnosis and prognosis of COVID-19 pneumonia from initial ER chest X-rays using deep learning. *BMC Infect. Dis.* **22**, 637. https://doi.org/10.1186/s12879-022-07617-7 (2022).
5. Salehi, S., Abedi, A., Balakrishnan, S. & Gholamrezanezhad, A. Coronavirus Disease 2019 (COVID-19): A systematic review of imaging findings in 919 patients. *AJR Am. J. Roentgenol.* **215**(1), 87–93. https://doi.org/10.2214/AJR.20.23034 (2020).

6. Altmayer, S. et al. Comparison of the computed tomography findings in COVID-19 and other viral pneumonia in immunocompetent adults: A systematic review and meta-analysis. *Eur. Radiol.* **30**(12), 6485–6496. https://doi.org/10.1007/s00330-020-07018-x (2020).
7. Majrashi, N. A. A. The value of chest X-ray and CT severity scoring systems in the diagnosis of COVID-19: A review. *Front. Med.* **9**, 1076184. https://doi.org/10.3389/fmed.2022.1076184 (2023).
8. Bernheim, A. et al. Chest CT findings in CORONAVIRUS DISease-19 (COVID-19): Relationship to duration of infection. *Radiology* **295**(3), 200463. https://doi.org/10.1148/radiol.2020200463 (2020).
9. Farahat, I. S. et al. An AI-based novel system for predicting respiratory support in COVID-19 patients through CT imaging analysis. *Sci. Rep.* **14**, 851. https://doi.org/10.1038/s41598-023-51053-9 (2024).
10. Fusco, R. et al. Artificial intelligence and COVID-19 using chest CT scan and chest X-ray images: Machine learning and deep learning approaches for diagnosis and treatment. *J. Personal. Med.* **11**(10), 993. https://doi.org/10.3390/jpm11100993 (2021).
11. Guiot, J. et al. Automatized lung disease quantification in patients with COVID-19 as a predictive tool to assess hospitalization severity. *Front. Med.* **9**, 930055. https://doi.org/10.3389/fmed.2022.930055 (2022).
12. Benmalek, E., Elmhamdi, J. & Jilbab, A. Comparing CT scan and chest X-ray imaging for COVID-19 diagnosis. *Biomed. Eng. Adv.* **1**, 100003. https://doi.org/10.1016/j.bea.2021.100003 (2021).
13. Gao, E., et al., COVID-VIT: Classification of Covid-19 from 3D CT chest images based on vision transformer model. 2022 3rd International conference on next generation computing applications (NextComp), Flic-en-Flac, Mauritius 1–4 https://doi.org/10.1109/NextComp55567.2022.9932246(2022)
14. Mondal, A. K., Bhattacharjee, A., Singla, P. & Prathosh, A. P. xViTCOS: Explainable vision transformer based COVID-19 screening using radiography. *IEEE J. Transl. Eng. Health & Med.* **10**, 1100110. https://doi.org/10.1109/JTEHM.2021.3134096 (2021).
15. Umar Marikkar, Sara Atito, Muhammad Awais, and Adam Mahdi. Lt-vit: A vision transformer for multi-label chest x-ray classification. In 2023 IEEE International Conference on Image Processing (ICIP) 2565–2569. IEEE (2023).
16. Ko, J., Park, S. & Woo, H. G. Optimization of vision transformer-based detection of lung diseases from chest X-ray images. *BMC Med. Inform. Decis. Mak.* **24**, 191. https://doi.org/10.1186/s12911-024-02591-3 (2024).
17. Park, S. et al. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. *Med. Image Anal.* **75**, 102299. https://doi.org/10.1016/j.media.2021.102299 (2022).
18. Hassan, E. et al. Detecting COVID-19 in chest CT images based on several pre-trained models. *Multimed. Tools & Appl.* **83**, 65267–65287. https://doi.org/10.1007/s11042-023-17990-3 (2024).
19. Singh, A. K., Kumar, A., Mahmud, M., Kaiser, M. S. & Kishore, A. COVID-19 infection detection from chest X-ray images using hybrid social group optimization and support vector classifier. *Cogn. Comput.* **4**, 1–13. https://doi.org/10.1007/s12559-021-09848-3 (2021).
20. Sahoo, P., Saha, S., Sharma, S., Mondal, S. & Gowda, S. A multi-stage framework for COVID-19 detection and severity assessment from chest radiography images using advanced fuzzy ensemble technique. *Expert Syst. Appl.* **238**, 121724. https://doi.org/10.1016/j.eswa.2023.121724 (2023).
21. Albataineh, Z., Aldrweesh, F. & Alzubaidi, M. A. COVID-19 CT-images diagnosis and severity assessment using machine learning algorithm. *Clust. Comput.* **27**, 547–562. https://doi.org/10.1007/s10586-023-03972-5 (2024).
22. Duong, T. L., Le, N., Tran, T., Ngo, V. & Nguyen, P. Detection of tuberculosis from chest X-ray images: Boosting the performance with Vision Transformer and transfer learning. *Expert Syst. Appl.* **184**, 115519 (2021).
23. Chen, T. et al. A vision transformer machine learning model for COVID-19 diagnosis using chest X-ray images. *Healthc. Anal.* **5**, 100332. https://doi.org/10.1016/j.health.2024.100332 (2024).
24. Mezina, A. & Burget, R. Detection of post-COVID-19-related pulmonary diseases in X-ray images using vision transformer-based neural network. *Biomed. Signal Process. Control* **87**, 105380 (2024).
25. Liu, C. & Sun, C. A fusion deep learning model of ResNet and vision transformer for 3D CT images. *IEEE Access.*https://doi.org/10.1109/ACCESS.2024.3423689 (2024).
26. Kotei, E. & Thirunavukarasu, R. Tuberculosis detection from chest X-ray image modalities based on transformer and convolutional neural network. *IEEE Access.* **12**, 97417–97427. https://doi.org/10.1109/ACCESS.2024.3428446 (2024).
27. Taye, G. D. et al. Thoracic computed tomography (CT) image-based identification and severity classification of COVID-19 cases using vision transformer (ViT). *Discov. Appl. Sci.* **6**, 384. https://doi.org/10.1007/s42452-024-06048-0 (2024).
28. Slika, B. et al. Lung pneumonia severity scoring in chest X-ray images using transformers. *Med. Biol. Eng. Compu.* **62**, 2389–2407. https://doi.org/10.1007/s11517-024-03066-3 (2024).
29. Ukwuoma, C. C. et al. Automated lung-related pneumonia and COVID-19 detection based on novel feature extraction framework and vision transformer approaches using chest X-ray images. *Bioengineering* **9**(11), 709. https://doi.org/10.3390/bioengineering9110709 (2022).
30. Zhu, Z., Liu, L., Free, R., Anjum, A. & Panneerselvam, J. OPT-CO: Optimizing pre-trained transformer models for efficient COVID-19 classification with stochastic configuration networks. *Inf. Sci.* **680**, 121141. https://doi.org/10.1016/j.ins.2024.121141 (2024).
31. Wang, T. et al. F: Deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using vision transformer. *Med. Biol. Eng. Compu.* **61**(6), 1395–1408. https://doi.org/10.1007/s11517-022-02746-2 (2023).
32. Okolo, G. I., Katsigiannis, S. & Ramzan, N. IEViT: An enhanced vision transformer architecture for chest X-ray image classification. *Comput. Methods Programs Biomed.* **226**, 107141. https://doi.org/10.1016/j.cmpb.2022.107141 (2022).
33. Singh, S. et al. Efficient pneumonia detection using Vision Transformers on chest X-rays. *Sci. Rep.* **14**, 2487. https://doi.org/10.1038/s41598-024-52703-2 (2024).
34. Tahir, A. M. et al. COVID-19 infection localization and severity grading from chest X-ray images. *Comput. Biol. Med.* **139**, 105002. https://doi.org/10.1016/j.compbiomed.2021.105002 (2021).
35. Ning, W. et al. iCTCF: An integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia. *Res. Square* https://doi.org/10.21203/rs.3.rs-21834/v1 (2020).
36. Govindarajan, S. & Swaminathan, R. Differentiation of COVID-19 conditions in planar chest radiographs using optimized convolutional neural networks. *Appl. Intell.* **51**, 2764–2775. https://doi.org/10.1007/s10489-020-01941-8 (2021).
37. Chen, H. C. et al. AlexNet convolutional neural network for disease detection and classification of tomato leaf. *Electronics* **11**, 951 (2022).
38. Dinh, P. H. & Giang, N. Medical image fusion based on transfer learning techniques and coupled neural P systems. *Neural Comput. & Appl.* https://doi.org/10.1007/s00521-023-09294-2 (2023).
39. Ortiz, A. et al. Effective deep learning approaches for predicting COVID-19 outcomes from chest computed tomography volumes. *Sci. Rep.* **12**, 1716. https://doi.org/10.1038/s41598-022-05532-0 (2022).
40. Chen, J., He, Y., Frey, E. C., Li, Y., Du, Y. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. arXiv. arXiv:2104.06468 (2021).
41. Al-Hammuri, K., Gebali, F., Kanan, A. & Chelvan, I. T. Vision transformer architecture and applications in digital health: A tutorial and survey. *Visual Comput. Ind. Biomed. & art* **6**, 14 (2023).
42. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. arXiv:2010.11929 (2020).
43. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P. Vision transformers need registers. arXiv. arXiv:2309.16588 (2023).

44. Ramadan, S. T. Y., Sakib, T., Rahat, M. A., Mosharrof, S., Rakin, F. I., Jahangir, R. Enhancing Mango Leaf Disease Classification: ViT, BiT, and CNN-Based models evaluated on CycleGAN-augmented data. In Proc. of the 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh 13–15 December 2023 IEEE: Piscataway, NJ, USA 1–6 (2023).
45. Khan, S. et al. Transformers in vision: A survey. *ACM Comput. Surv.* **54**, 1–41 (2022).
46. Mirjalili, S., Mirjalili, S. M. & Lewis, A. Grey Wolf Optimizer. *Adv. Eng. Softw.* **69**, 46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007 (2014).
47. Garg, H. A hybrid PSO-GA algorithm for constrained optimization problems. *Appl. Math. Comput.* **274**, 292–305. https://doi.org/10.1016/j.amc.2015.11.030 (2016).
48. Gupta, S. L., Baghel, A. S., Iqbal, A. Big data classification using scale-free binary particle swarm optimization. In Proc. of the harmony search and nature inspired optimization algorithms: Theory and applications, ICHSA 2018 Springer: Singapore 1177–87 (2018).
49. Khan, A. & Baig, A. R. Multi-objective feature subset selection using non-dominated sorting genetic algorithm. *J. Appl. Res. & Technol.* **13**, 145–159 (2015).
50. Abdo, A., Mostafa, R. & Abdel-Hamid, L. An Optimized hybrid approach for feature selection based on chi-square and particle swarm optimization algorithms. *Data* **9**(2), 20 (2024).
51. Balasubramaniam, S., Arishma, M., Kumar, S. & Rajesh Kumar, D. A comprehensive exploration of artificial intelligence methods for COVID-19 diagnosis. *EAI Endorsed Trans. Pervasive Health & Technol.* **10**, 5174. https://doi.org/10.4108/eetpht.10.5174 (2024).
52. Puttagunta, M. & Ravi, S. Medical image analysis based on a deep learning approach. *Multimed. Tools & Appl.* **80**, 24365–24398. https://doi.org/10.1007/s11042-021-10707-4 (2021).
53. Imagawa, K. & Shiomoto, K. Evaluation of effectiveness of pre-training method in chest X-ray imaging using vision transformer. *Comput. Methods Biomech. & Biomed. Eng : Imaging & Vis.* https://doi.org/10.1080/21681163.2024.2345823 (2024).
54. Balasubramaniam, S. et al. Res-Unet-based blood vessel segmentation and cardiovascular disease prediction using chronological chef-based optimization algorithm and deep residual networks from retinal fundus images. *Multimed. Tools & Appl.* **83**, 87929–87958. https://doi.org/10.1007/s11042-024-18810-y (2024).
55. Singh, T. et al. COVID-19 severity detection using chest X-ray segmentation and deep learning. *Sci Rep* **14**, 19846. https://doi.org/10.1038/s41598-024-70801-z (2024).
56. Balasubramaniam, S., & Kumar, S. Prediction of breast cancer using ensemble learning and boosting techniques. In 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), 513–519 https://doi.org/10.1109/IC3SE62002.2024.10593047 (2024)
57. Balasubramaniam, S., Kadry, S. & Kumar, S. Osprey Gannet optimization enabled CNN-based transfer learning for optic disc detection and cardiovascular risk prediction using retinal fundus images. *Biomed. Signal Process. Control* **93**, 106177. https://doi.org/10.1016/j.bspc.2024.106177 (2024).
58. Balasubramaniam, S., Nelson, S., Arishma, M., Rajan, A. & Kumar, S. Machine learning-based disease and pest detection in agricultural crops. *EAI Endorsed Trans.Internet Things* **10**, 5049. https://doi.org/10.4108/eetiot.5049 (2024).
59. Balasubramaniam, S. & Kumar, S. Fractional Feedback political optimizer with prioritization-based charge scheduling in cloud-assisted electric vehicular network. *Ad. Hoc. & Sens. Wirel. Netw.* **52**, 173–198 (2022).

Author contributions

V.Padmavathi: Conceptualization, Data Collection, Formal analysis, Validation and Writing original draft. Kavitha Ganesan: Supervision, visualization, Writing – review and editing.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.