


Comparative Evaluation of Deep Learning Models for Diagnosis of COVID-19 Using X-ray Images and Computed Tomography

Aroldo Ferraz   [Federal Technological University of Paraná | ferraz.2020@alunos.utfpr.edu.br]

Roberto Cesar Betini  [Federal Technological University of Paraná | betini@utfpr.edu.br]

 Universidade Tecnológica Federal do Paraná, Av. Sete de Setembro, 3165 - Rebouças, Curitiba - PR, 80230-901, Brazil.

Received: 30 April 2023 • **Accepted:** 29 October 2024 • **Published:** 20 February 2025

Abstract (1) Background: The COVID-19 pandemic is an unprecedented global challenge, having affected more than 776.79 million people, with over 7.07 million deaths recorded since 2020. The application of Deep Learning (DL) in diagnosing COVID-19 through chest X-rays and computed tomography (CXR and CT) has proven promising. While CNNs have been effective, models such as the Vision Transformer and Swin Transformer have emerged as promising solutions in this field. (2) Methods: This study investigated the performance of models like ResNet50, Vision Transformer, and Swin Transformer. We utilized Bayesian Optimization (BO) in the diagnosis of COVID-19 in CXR and CT based on four distinct datasets: COVID-QU-Ex, HCV-UFPR-COVID-19, HUST-19, and SARS-COV-2 Ct-Scan Dataset. We found that, although all tested models achieved commendable performance metrics, the Swin Transformer stood out. Its unique architecture provided greater generalization power, especially in cross-dataset evaluation (CDE) tasks, where it was trained on one dataset and tested on another. (3) Results: Our approach aligns with state-of-the-art (SOTA) methods, even in complex tasks like CDE. On some datasets, we achieved exceptional performance metrics, with AUC, Accuracy, Precision, Recall, and F1-Score values of 1. (4) Conclusion: Results obtained by the Swin Transformer go beyond what is offered by current SOTA methods and indicate actual feasibility for application in medical diagnostic scenarios. The robustness and generalization power of the Swin Transformer, demonstrated across different datasets, encourage future exploration and adoption of this approach in clinical settings.

Keywords: Swin Transformer, COVID-19, Chest X-ray, Deep Learning, Bayesian Optimization

1 Introduction

The COVID-19 pandemic, as documented by the World Health Organization [WHO, 2024], has resulted in over 776.79 million cases and 7.07 million fatalities globally by November 2024, highlighting the critical need for rapid and accurate diagnostic methods in managing highly contagious diseases. The standard diagnostic method, RT-PCR [Castro *et al.*, 2020], though reliable, is time-consuming, creating bottlenecks for timely patient management. While the number of COVID-19 cases has decreased, the pandemic underscored the importance of developing and maintaining advanced diagnostic technologies to prepare for future health crises. Moreover, these innovations have broader applications in improving overall healthcare system efficiency. This challenge extends beyond COVID-19 to other respiratory diseases, emphasizing the ongoing need for faster, more efficient diagnostic tools to optimize resource allocation and improve patient outcomes across various health conditions.

Deep Learning (DL) models, particularly those based on Convolutional Neural Networks (CNN), have shown promise in expediting COVID-19 diagnosis using radiographs, including X-rays and CT scans. However, these models often overlook the global context of the images, which is crucial for accurate diagnostics [Hassan *et al.*, 2022].

This study delves into the Swin Transformer model [Liu *et al.*, 2021], a recent innovation in DL that captures global

contextual information more effectively than conventional CNN and Vision Transformers (ViT) [Dosovitskiy *et al.*, 2020]. This research's potential is to assess whether the Swin Transformer can surpass existing state-of-the-art DL models in classifying X-ray images for the diagnosis of COVID-19, mainly focusing on cross-dataset evaluation.

In pursuit of this goal, we will compare the performance of the Swin Transformer against other leading models to validate its effectiveness in this domain. Ultimately, this work seeks to establish the Swin Transformer's enhanced capability for cross-dataset generalization, which is essential for real-world medical applications [Liu *et al.*, 2017].

This study aims to provide insights into these models' applicability and effectiveness by comparing them against other state of the art deep learning models. It aims to pave the way for enhanced diagnostic processes that integrate speed with accuracy, which is crucial for managing pandemics effectively.

1.1 Justifying the Use of Computer Vision in COVID-19 Diagnostics

Assessing the existence of COVID-19 by analyzing chest radiographs is laborious and time-consuming. Computer vision models can significantly accelerate and automate this crucial activity [Castro *et al.*, 2020; Hassan *et al.*, 2022]. In the scenario brought about by the coronavirus pandemic, the

rapid and accurate diagnosis is critical for the preservation of lives and for containing community contagion [WHO, 2024].

However, the surge in suspected new cases of COVID-19 often leads to a rapid saturation of the analysis capacity of radiologists [WHO, 2024]. These professionals, who tirelessly analyze each radiograph to determine whether the case is COVID-19, could greatly benefit from a DL system trained to assist the diagnosis of this disease. Such a system could provide invaluable support [Hassan *et al.*, 2022], potentially revolutionizing how we approach COVID-19 diagnosis.

The ideal implementation would consist of a two-stage process: the first stage involves classification using DL, where the images are processed through the computer vision model for initial screening. Subsequently, in the second stage, these results are validated by a senior specialist in radiology [Hassan *et al.*, 2022; Liu *et al.*, 2017]. This routine has the potential to blend the pattern detection capabilities of computer models with the specialized diagnostic expertise of radiologists, thereby speeding up the diagnostic process and enhancing the overall accuracy of COVID-19 diagnostics [Cireşan *et al.*, 2012].

According to [Liu *et al.*, 2017], there have been significant advances in the application of computer vision models in the medical field. This inflection occurred in 2012, after the first publications on using CNN in that same year [Cireşan *et al.*, 2012].

According to [Luján-García *et al.*, 2020], since 2012, several studies have appeared that employ CNN models for image classification tasks. Among the works we can mention: VGG-16 [Geng *et al.*, 2019], Inception-v3 [Szegedy *et al.*, 2016], Residual Networks v1 and v2 ResNet1 and ResNet2 [Jung and Chi, 2020] and [He *et al.*, 2016], Xception Depth Separable Convolution Networks [Chollet, 2017], Densely Connected Networks, DenseNet [Yao *et al.*, 2020], among others. These CNN-based models are often used to implement systems for Computer Vision (CV) tasks and for Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) [Bakator and Radosav, 2018].

The systematic review work by Hassan [Hassan *et al.*, 2022] specifically studied the use of CNN models for the classification, detection, and segmentation of radiographic images regarding the presence or absence of COVID-19. In addition, he mapped datasets that were used in these works. As for the classification task, it was pointed out that [Zhao *et al.*, 2020] obtained an AUC of 0.98, while [Castiglione *et al.*, 2021] obtained an accuracy of 0.999. These impressive results were achieved using the SARS-COV-2 CT-Scan dataset, which contains high-quality CT scans of COVID-19 positive and negative cases.

1.1.1 Use of ResNet in COVID-19 Diagnostics

The ResNet50 model was chosen from among the many available CNNs in the present study due to a combination of factors. Figure 1 shows the details of the ResNet architecture. Firstly, ResNet50 is a deep CNN with fifty layers, representing a significant evolution over previous models with fewer layers, enabling it to capture more complex features in the data. Consequently, it has the potential for superior performance in image classification. A relevant aspect is

ResNet50's fundamental feature: the introduction of the concept of residual connections, also known as "shortcuts" or "skip connections", which allows the model to effectively circumvent the "vanishing gradient" problem, facilitating learning and improving performance.

The ResNet architecture has also demonstrated excellent results in various medical image classification problems, establishing itself as a reliable and solid choice for similar tasks. Meedeniya *et al.* [2022] highlighted that this architecture is most likely to be utilized by researchers who apply CNN models to the current field of medical image research. Its popularity means there is a broad research community with abundant resources, tutorials, and code examples that can assist in implementation and problem-solving.

1.1.2 Use of Vision Transformer in COVID-19 Diagnostics

However, recently, a new technique has shown promise in computer vision tasks: the Vision Transformer (ViT) [Dosovitskiy *et al.*, 2020]. Details of the ViT architecture are demonstrated in Figure 2. Although ViT is based on the Transformer model created for Natural Language Processing (NLP) tasks, as described by [Wolf *et al.*, 2020], ViT has shown promise for applications in the image domain.

In our literature review on the ViT model, we found that several authors have demonstrated that the ViT model has the potential to outperform CNN models in image classification tasks for diagnosing COVID-19, among which the following stand out: [Mehboob *et al.*, 2022; Park *et al.*, 2022b; Chetoui and Akhloufi, 2022; Park *et al.*, 2022a; Jiang *et al.*, 2022; Mondal *et al.*, 2022; Konwer and Prasanna, 2022; Aytekin *et al.*, 2022; Krishnan and Krishnan, 2021; Zhang and Wen, 2021; Al Rahhal *et al.*, 2022; Li *et al.*, 2021; Dehkordi *et al.*, 2021; Than *et al.*, 2021; Balderas *et al.*, 2023; Khobragade and Manthalkar, 2024].

1.1.3 Use of Swin Transformer in COVID-19 Diagnostics

Even more recent than ViT is the Swin Transformer model [Liu *et al.*, 2021], a ViT variant. The Swin Transformer will be used in the present work to conduct experiments on classifying X-ray images to detect cases of COVID-19. This work will also test this model's ability to provide better ranking performance metrics than the CNN and ViT models.

In light of the evolving landscape of DL models applied to medical image classification, particularly in detecting and diagnosing COVID-19, this study aims to conduct an in-depth comparative analysis of three distinct models: ResNet50, ViT, and Swin Transformer. While the ResNet50 has been a prevalent choice among researchers for its deep architectural features and robustness in image classification, the emergent ViT model has begun challenging the supremacy of CNN in visual domains. Furthermore, the Swin Transformer, a recent variant of ViT, has shown promise in delivering potentially superior ranking performance metrics (architecture details shown in Figures 3, 4, and 5). By comparing these three models on common grounds, we intend to evaluate their respective merits and limitations in classifying X-ray

images for COVID-19 detection. This comparative assessment will shed light on the effectiveness and adaptability of these models, contributing valuable insights to the ongoing pursuit of accurate and efficient medical image analysis methodologies.

This article is organized as follows: Section 2 deals with Materials and Methods where the research method is described, explaining the process used to experiment in addition to carrying out a literature review on the subject; Section 3 discusses the Results where the performance metrics obtained by the employed model are listed, in general terms; Section 4 brings off the Discussion on the topic from the results obtained and the comparison of these results with those of the researched works in the literature review and that represent state of the art. Also, in this section, open problems and directions for future research are presented. Finally, section 5 brings the Conclusion, that is, it is the finalization of the article explaining what it was intended to do, what was done, and what are the results obtained; limitations and threats related to the results obtained in the research work are also indicated.

2 Materials and Methods

The ResNet architecture, introduced by He *et al.* [2016], is characterized by “residual blocks”, which contain shortcuts or skip connections that enable the gradient to be backpropagated to earlier layers. These shortcuts or residual connections alleviate the vanishing gradient problem, allowing intense networks to be trained effectively. The critical element of a residual block is the “shortcut connection” that skips one or more layers. Figure 1 illustrates the architecture of the ResNet network. Residual networks have proven that networks can indeed benefit from being deeper. He *et al.* [2016] successfully trained ResNets with up to 152 layers while improving performance without increasing computational complexity compared to shallower networks.

He *et al.* [2016] reports experimental results in several benchmark competitions and image recognition challenges. ResNets set record results in many of these challenges, attesting to the effectiveness of their architecture. In the ImageNet challenge, for instance, a 152-layer ResNet won the competition with a top-5 error rate of 3.57%, a record at the time. This paper has significantly impacted DL and CNN, as the ResNet architecture has become one of the primary choices for many computer vision tasks. It paved the way for even deeper networks and other shortcut connection architectures.

It has been shown that the ViT model can overcome a limitation of CNN in computer vision applications that typically require the integration of the global relationship between pixels. This overcoming is possible because the ViT architecture [Vaswani *et al.*, 2017] was proposed to model the short-range and long-range dependency between pixels through the self-attention mechanism. The ViT model achieved excellent results, comparable to the state-of-the-art in the image classification task [Dosovitskiy *et al.*, 2020].

The ViT model divides an input image into some patches (squares of a grid or square portions of the input image), analogous to the word embedding sequence used when a Trans-

former is applied to texts, typically known as NLP [Wolf *et al.*, 2020], and predicts the image class labels directly in the output.

One of the premises for achieving DL models suitable for use in practical applications is to have large volumes of data for training and testing [Raghu *et al.*, 2021], which is not always possible, either for reasons of secrecy and sensitivity information or other limitations. In the case of data from the medical field, issues such as the protection of sensitive personal data are involved, which often limits the availability of this data for studies that are not restricted to researchers from the institution that owns the data.

A technique widely used to overcome the problem of a database shortage is Transfer Learning (TL), which consists of using models trained in much larger databases and even in other domains. When used in medical imaging, TL can achieve higher predictive performance metrics, even on small datasets, by leveraging previously acquired learning on larger datasets. Along these lines, several authors have already demonstrated that this technique obtains good practical results [Cha *et al.*, 2021; Chouhan *et al.*, 2020; Hashmi *et al.*, 2020; Jain *et al.*, 2020; Liang and Zheng, 2020; Rahman *et al.*, 2020; Luján-García *et al.*, 2020].

Recently, the Swin Transformer [Liu *et al.*, 2021] was presented (architecture shown in Figure 5). It is a ViT-based DL model with state-of-the-art performance for CV tasks. Unlike the ViT [Dosovitskiy *et al.*, 2020] that precedes it, the Swin Transformer is highly efficient and has greater accuracy [Liu *et al.*, 2021]. Due to these desirable properties, the Swin Transformer can be the backbone of many CV model architectures.

Swin Transformer solved the problems inherent to the original ViT model using hierarchical feature maps and offset window multi-head self-attention (MSA) [Liu *et al.*, 2021].

In the present work, a Swin Transformer, a ViT, and a ResNet50 models were implemented with the TL of the learning performed in the ImageNet-1K dataset [Liu *et al.*, 2021]. The word Swin (in Swin Transformer) is an acronym that stands for Shifted window (Figure 4). This hanging window concept is not new to the research community. It has been used in CNN for many years. One of CNN’s features made it stand out in computer vision, as it brought excellent efficiency. However, it had not yet been used in ViT. Figure 3 compares how the patches are segmented in the ViT and Swin Transformer models.

Similar to the ViT model, Swin Transformer uses patches; however, instead of using a fixed size, such as $16 \times 16px$, Swin Transformer starts with small patches ($4 \times 4px$) on the first Transformer layer and changes the size of patches in deeper layers. The model merges these smaller layers into more extensive layers as it delves deeper into the network architecture. It takes an image and splits it into $4px$ by $4px$ patches. Each patch is a color image with three channels in Red, Green and Blue (RGB) patterns. Thus, a patch has a total of 48 resource dimensions. That is $4 \times 4 \times 3 = 48$. It is then linearly transformed into a dimensionality called C , of your choice. The image patches are smaller compared to ViT up to this point. The value, C , determines the size of your Transformer model. It should be noted that although ViT is computationally more expensive than CNN when applied to

large datasets, it tends to perform better than CNN models [Dosovitskiy *et al.*, 2020; Tuli *et al.*, 2021; Wei *et al.*, 2022]. Another limitation of ViT is that its computational complexity is quadratic to the image size or to the number of patches, making it unsuitable for high-resolution images.

According to [Sun *et al.*, 2022], in ViT, the $\text{Att}(\cdot)$ self-attention module learns three weights: W_Q , W_K , and W_V . Based on these weights, x is projected into the query (Q), key (K), and value (V). An attention matrix A is usually computed by a similarity function $S(\cdot)$ over queries and keys. In standard self-attention, $S(\cdot)$ is softmax normalization. The outputs of the self-attention module are, hence, $O = \text{Att}(x) = AV$, where $A \in \mathbb{R}^{N \times N}$ suffers from a space-time quadratic complexity concerning the patch number N . Therefore, in this case, the theoretical computation complexity is $\mathcal{O}(N^2d) = \mathcal{O}\left(\frac{H^2W^2}{p^4}d\right)$. Consequently, the self-attention module becomes sensitive to the image size, suffering from the increase in height (H) and width (W) and the patch size p . Thus, the quadratic complexity of the ViT model is, by definition, its main computational bottleneck.

The approach employed by Swin Transformer is known as shifted window [Liu *et al.*, 2021]. It consists of computing self-attention within local windows instead of computing within a global receptive field, as ViT does. An offset window contains non-overlapping patches of $M \times M$ (where $M = 7$ is the window size), and self-attention is calculated on that window. Figure 4 illustrates how the shifted window technique operates.

As a result, the computational complexity of the original ViT Multi-Head Self-Attention (MSA), which is quadratic for the patch number and $H \times W$, drops dramatically for the Swin Transformer case because the W-MSA based on the Swin Transformer window is linear. Comparatively, the equation for the ViT is $\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C$, while for the Swin Transformer is $\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC$. The equations demonstrate that the computational cost of the Swin Transformer is much lower. The MSA represents the ViT self-attention mechanism in the equations, while the (W-MSA) Window Multi-Head Self-Attention represents the Swin Transformer.

As can be seen, the Swin Transformer model adds a linear computational complexity to the size of the input image. It calculates self-attention only within the local window, not globally, as with the ViT model. This feature allows the model to perform dense recognition tasks and is used for more general-purpose computer vision tasks with larger color images.

In Swin Transformer, the output of a layer is merged by a Merging Layer, which concatenates the vectors of patch groups from the neighborhoods in the image each time the attention window changes from the previous layer. For example, if attention is limited to the neighborhood of these regions in the first layer, regions are shifted (as in the stridden convolution) in the next layer.

Patches that arrived in separate windows on the first layer and could not communicate can do so on layer two. The blend layer blends these resulting patches. This process is repeated depending on the number of layers chosen.

In our work, we used the tiny version of Swin Transformer

(Swin Transformer Tiny also known as SwinT). The model starts by splitting the input RGB image into non-overlapping patches, such as ViT. Each patch is treated as a “token”, and its resource is defined as a concatenation of the RGB values of the originating pixel.

In summary, Swin Transformer integrates CNN’s characteristic advantages in CV with the efficient and robust architecture of ViT, as highlighted by [Liu *et al.*, 2021]. This is achieved because the hierarchical representation can achieve scale invariance, and the shifted windows approach can efficiently transmit information within the local window. The Merging Layer is in charge of integrating the global information of the pixels.

2.1 Literature Review

In our literature review, we selected 33 papers (shown in Table 1) where it was evident that the CNN, ViT, and SwinT models applied to the binary classification of COVID-19 and NORMAL showed promising results in terms of the absolute values of the performance metrics obtained. However, direct comparisons between models and even outcomes from the same model should be cautiously approached. The nuances of the datasets used in each study and differences in approach and network architectures are limitations that prevent metric-to-metric comparison from definitively determining if one model is superior to another.

CNN models yielded promising results, achieving high accuracy and precision despite discrepancies in the findings across various studies. The performance variation among CNN models can be attributed to differences in model architectures, data preprocessing techniques, and the datasets used.

The ViT and SwinT models were also examined. Some studies reported favorable outcomes, though not all provided comprehensive details on every metric evaluated. Due to the inconsistent performance metrics available across the studies, it is not feasible to directly compare CNN and ViT models.

From the literature reviewed, performance metrics documented for CNN models typically emphasized accuracy, focusing less on AUC and recall (Sensitivity). This trend is also evident in research concerning ViT models, although some studies have slightly increased focus on AUC and Sensitivity. The same pattern is observed in studies on the SwinT model, with a more pronounced emphasis on AUC and Sensitivity in some research.

In summary, all three models demonstrate potential for binary classification, but there’s a need for standardized and more detailed information on performance metrics for proper comparison. It’s vital to consider the variation in outcomes across studies and to conduct additional research to confirm and compare the models’ efficacy on different datasets.

With this observation in mind, experiments in this study aimed to explore the most suitable model for medical image classification. Therefore, the same datasets and execution conditions were adopted to ensure a fair evaluation as much as possible.

2.2 Datasets

Regarding the dataset, nine were used: four unique datasets, five datasets generated from the CDE strategy, and a combination of X-ray and CT datasets (hybrid or mixing approach).

The first dataset used was the COVID-QU-EX Dataset, employed initially in Raman *et al.* [2021] and enhanced by Ning *et al.* [2020]. It is accessible on Kaggle. A team of researchers from the University of Qatar in Doha, Qatar, created this dataset. Researchers compiled a database of CXR images for positive COVID-19 cases and images of regular and viral pneumonia. We utilized 33,920 images from this dataset, including 11,956 COVID-19 images, 11,263 non-COVID-19 pneumonia images (not used in this study), and 10,701 healthy lung images.

The second dataset was the HCV-UFPR-COVID-19 dataset Luz *et al.* [2021], comprising 281 X-ray images from COVID-19-infected individuals and 232 images from individuals who tested negative for the disease. All images have three 8-bit color channels (RGB), and their resolution varies from 2974×2612 to 4248×3480 pixels. Images are labeled into two classes, COVID-19 and non-COVID, and there are no annotations regarding the image angle view. The HCV-UFPR-COVID-19 X-ray dataset is made available to researchers upon request on a case-by-case basis. This dataset was created by the Red Cross Hospital, which received and documented some COVID-19 cases, in collaboration with the Federal University of Paraná (UFPR), both located in Curitiba, Paraná, Brazil.

The third dataset was a Brazilian SARS-COV2 CT dataset Mehboob *et al.* [2022], accessible on Kaggle. The SARS CT dataset consists of 2,481 CT scans from 120 patients, including 1,252 CT scans from 60 infected patients (COVID) and 1,229 CT scans from 60 non-infected patients (non-COVID). The data was collected from actual patients in São Paulo, Brazil hospitals. The CT images vary in size, with the smallest image being 104×153 and the most prominent image being 484×416 . The dataset contains heterogeneous CT images with a low number of instances. In CT images, an instance refers to a single image acquired during the CT scan. Additionally, the CT images have different contrasts and resolutions.

The fourth dataset was the HUST-19 dataset, utilized in Ning *et al.* [2020] and available on the National Genomics Data Center website. This database consists of 4,001 positive CT slices (pCT) and 9,979 negative CT slices (nCT), randomly selected from 61 patients with COVID-19 pneumonia and 43 patients without pneumonia.

In addition to the four original datasets, we created two more by mixing the four datasets, as shown in Table 2. This strategy evaluated the models in a scenario with mixed data containing X-ray and CT images.

Additionally, we applied the CDE strategy between the X-ray datasets, between those composed of CT scans, and finally, between the hybrid datasets, resulting in three more datasets. This approach assessed whether the models could effectively generalize the information in X-ray, CT, and both simultaneously.

It is important to note that we worked with two classes (binary classification) for all nine datasets used.

2.2.1 Dataset Partitioning Strategy

Proper dataset partitioning into training, validation, and test sets is essential for developing robust DL models. This process helps prevent overfitting, ensures realistic performance evaluation, and enables practical hyperparameter tuning. We adopted the following partitioning strategies:

- **Training Set (60-70%):** This most significant portion is used to train the model, adjusting its internal parameters based on the data.
- **Validation Set (15-20%):** This set is utilized to tune model hyperparameters and to mitigate overfitting during the training process, thereby providing essential feedback on model performance.
- **Test Set (15-20%):** This set is reserved to assess the model's final performance after adjustments during training and validation, ensuring its effectiveness on unseen data.

Experimenting with dataset partitioning ratios of 70/15/15 and 60/20/20 enables identifying the optimal balance between training depth and model generalization. The 70/15/15 strategy increases training data, potentially enhancing learning. The 60/20/20 arrangement offers a more extensive validation set, allowing rigorous tuning and validation across unseen data.

Validations used 5-fold cross-validation within the validation set to comprehensively assess performance robustness. This approach ensures representative partitions and minimizes bias for reliable metrics. Repeated training with random splits further verifies model consistency and robustness, essential for developing high-performing, reliable models.

To further optimize the model, Bayesian optimization was employed in conjunction with the 5-fold cross-validation strategy. Bayesian optimization is a powerful technique for efficiently tuning hyperparameters by modeling the objective function with a Gaussian process and adaptively selecting the next hyperparameter configuration to evaluate based on an acquisition function. This Bayesian optimization approach was executed individually for the ResNet, ViT, and SwinT models, helping identify the ideal hyperparameter settings to maximize performance and ensure robust generalization.

2.2.2 Justification for the Cross Dataset Evaluation Data Partitioning Strategy

This data partitioning strategy was carefully designed to assess the generalization ability of the models in a challenging scenario, where the training and validation sets consist of completely distinct data and originate from different datasets compared to the test set. Furthermore, it was ensured that X-ray and computed tomography images were handled separately, without any mixing between them, considering that these types of exams have substantially different characteristics and information sets.

Specifically, smaller datasets, such as HCV-UFPR-COVID-19 and SARS-COV2 CT, were exclusively allocated to the test set, while larger datasets were used for training and validation. This approach was grounded in the need

to simulate an environment where the models would be evaluated on data that is not only unseen but also significantly different from the data used during training and validation. This strategy aims to ensure that the model is tested under conditions that closely resemble a real-world production environment.

For example, smaller patient datasets such as HCV-UFPR-COVID-19 and SARS-COV2 CT were exclusively allocated to the test set due to their distinct characteristics and size, ensuring that the model is evaluated on data that is both unseen and significantly different from the training and validation data (see Tables 3, 4, 5, 6, 7, and 8). This approach also ensures that X-ray and computed tomography images are handled separately, avoiding any mixing between them given their substantially different data characteristics.

Similarly, the DSHybrid1 dataset was utilized for training and validation, while DSHybrid2 was reserved solely for testing. This patient-wise and modality-specific split was carefully designed to prevent any cross-contamination between the training/validation and testing phases, ensuring a rigorous evaluation of the model's generalization capabilities under realistic and challenging conditions.

The rigorous separation between datasets was designed to reflect practical scenarios where, in the inference stage during production, the model frequently encounters data with characteristics distinct from those seen during its development. This evaluation in an adverse scenario enables a more reliable estimation of the model's robustness and generalization capabilities, which are critical aspects for its viability in real-world applications.

Additionally, by using smaller and more specific datasets for testing, we ensure that:

- **The evaluation is not influenced by dominant data in training:** The separation of datasets preserves the necessary independence between the development (training/validation) and evaluation (testing) phases, avoiding the dominance of characteristics from larger datasets in the model's performance.
- **The model is exposed to varied and challenging scenarios:** Smaller datasets often possess unique characteristics, representing challenging and complementary scenarios. This provides a more realistic and robust metric for its generalization ability.
- **The risk of overfitting is minimized:** As the test datasets do not participate in any stage of the model adjustment process, the results solely reflect the model's generalization capability rather than memorization of specific patterns.

This methodology reflects one of the objectives of the study: to verify whether the model can sustain good performance under adverse and realistic conditions, which is a crucial indicator of its reliability and effectiveness in production environments. We acknowledge that the use of smaller datasets for testing may limit the statistical analysis due to the reduced sample size. However, the selection of these datasets was motivated by their relevance and diversity, providing a robust evaluation of the model's ability to handle challenging scenarios.

2.3 Performance Metrics

To evaluate our models, particularly in the context of medical image classification, we employed a set of performance metrics that captured various aspects of model effectiveness. To facilitate understanding, Figure 6 provides an intuitive summary of these performance metrics. These performance metrics include:

- **Accuracy:** Defined as the ratio of correctly predicted observations to the total observations:

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

- **Precision (Positive Predictive Value):** Measures the proportion of actual positives among predicted positives:

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

- **Recall (Sensitivity or True Positive Rate):** Indicates the model's ability to identify all actual positives:

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

- **F1-Score:** The harmonic mean of precision and recall, useful for cases where an equal balance of precision and recall is crucial:

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Area Under the ROC Curve (AUC-ROC):** Evaluates the model's discrimination capability, with a value between 0.5 (no discriminative power) and 1.0 (perfect discrimination).

2.4 Hyperparameter Tuning

Hyperparameter optimization plays a pivotal role in enhancing the performance of machine learning models. Akiba *et al.* [2019] introduces Optuna, an advanced hyperparameter optimization framework based on a "define-by-run" strategy. This approach adapts dynamically to different computational environments, from large distributed systems to localized experimental setups, and incorporates innovative search and pruning mechanisms to explore the parameter space efficiently. This flexibility and efficiency make Optuna particularly effective, signifying a major shift in how hyperparameter tuning is approached in the field.

Further research by Shekhar *et al.* [2021], examines the efficacy of various hyperparameter optimization libraries, such as Optuna, HyperOpt, Optunity, and SMAC, across six real-world datasets. Their study underscores Optuna's superior performance in selecting algorithms and tuning parameters, although it also notes the need for a clearer rationale behind the choice of tools and more detailed insights into methodological approaches and performance trade-offs.

In a related study, Hamza *et al.* [2022] demonstrates the application of Bayesian optimization with deep convolutional neural networks (DCNNs) for detecting COVID-19 from chest X-ray images. By enhancing image contrast and employing pre-trained models like EfficientNet-B0

and MobileNet-V2, the researchers achieved an impressive 99.4% accuracy across three public datasets. Despite these advancements, the study highlights the necessity of further evaluations regarding model robustness, computational efficiency, and ethical considerations in AI applications.

Our experiments employed models such as ResNet50, ViT, and SwinT, optimized using the Optuna library to significantly improve performance metrics like AUC, accuracy, precision, sensitivity, and F1-score. We conducted experiments using two distinct data partitioning strategies, 70/15/15 and 60/20/20, to determine the optimal balance between training set size and generalization capabilities. This extensive testing, performed on a robust hardware setup featuring high-performance GPUs and CPUs, confirms the critical impact of optimized hyperparameters on the efficacy of DL models in various practical applications.

The analysis of the standard deviation differences between the 70/15/15 and 60/20/20 splits reveals that the 70/15/15 split generally provides more consistency, particularly when using the Adaptive Moment Estimation (Adam) optimizer, as seen in lower standard deviations for metrics like AUC, accuracy, precision, and recall. In contrast, for the Stochastic Gradient Descent (SGD) optimizer, the 60/20/20 split shows greater consistency in precision and F1-Score. These results suggest that while Adam performs more consistently in the 70/15/15 split, SGD benefits from the 60/20/20 split in specific metrics, indicating that the choice of split and optimizer can influence the model's stability.

2.4.1 Justification for Comparing Adam and SGD Optimizers

The decision to compare Adam and SGD optimizers in our study is rooted in the fundamental differences in their optimization approaches, which significantly impact model training, convergence, and generalization capabilities. This comparison is essential for understanding the optimal optimizer choice for various machine learning tasks.

1. **Different Optimization Behaviors:** Adam combines advantages from AdaGrad and RMSProp optimizers, adapting the learning rate for each parameter, which facilitates quicker convergence in scenarios involving large-scale data and sparse gradients. Conversely, SGD maintains a uniform learning rate and is known for its simplicity and reliability in numerous settings.
2. **Impact on Model Training and Generalization:** Research such as by Wilson *et al.* [2018] suggests that while adaptive methods like Adam can accelerate the initial phase of learning, they might underperform in terms of generalization compared to SGD under certain conditions. This makes it crucial to analyze both optimizers to identify the most suitable one based on the specific requirements of the dataset and model architecture.
3. **Hyperparameter Sensitivity:** Both optimizers are sensitive to their hyperparameter settings, which influences their performance significantly. Evaluating both within the same experimental framework allows us to understand this sensitivity better and optimize the hyperparameters effectively.

4. **Robustness Across Various Architectures:** Different neural network architectures may benefit differently from each optimizer. For instance, architectures prone to overfitting might benefit from SGD's inherent noise, helping them escape local minima, while Adam might be better suited for models requiring fast convergence.
5. **Comprehensive Evaluation Strategy:** Including a comparative analysis of Adam and SGD helps in establishing a robust evaluation strategy, informing the selection process for the optimizer that best suits the model's needs. This approach enhances the reliability and performance of the models in diverse operational environments.

In conclusion, our comprehensive comparative analysis of Adam and SGD is aimed at elucidating their respective impacts on the performance across different models and datasets, thereby facilitating an informed choice of the optimizer that enhances model efficacy and robustness in practical applications.

2.5 Experimental Setup

2.5.1 Hardware Configuration

The primary hardware used for the experiments included a Dell Alienware m15 R7 laptop equipped with an Intel Core i7 12700H CPU, 16 GB RAM, an NVIDIA RTX 3070 GPU with 8GB VRAM, and 1 TB SSD. For tasks requiring additional memory capacity, such as Bayesian Optimization, a desktop computer was used, featuring an AMD A10-7850K CPU, 16GB RAM, an NVIDIA RTX 3060 GPU with 12 GB VRAM, and a 500 GB hard drive.

2.5.2 Model Implementation and Optimization

Our experiments utilized three advanced machine learning models: ResNet50, ViT, and Swin Transformer (SwinT), all implemented using PyTorch 1.12.1 PyTorch [2023]. We leveraged TL by utilizing the pre-trained weights from the ImageNet-1K dataset ImageNet1k [2017] to enhance initial model performance. Furthermore, we used Bayesian optimization through the Optuna library Akiba *et al.* [2019] to fine-tune the hyperparameters, resulting in significant improvements across various performance metrics, including AUC, Accuracy, Precision, Sensitivity, and F1-Score.

2.5.3 Data and Methodology

Our research encompassed nine datasets, consisting of four original sets and five derivatives, as detailed in Table 2. The optimized hyperparameters are outlined in Table 9. We experimented with two distinct data partitioning strategies (70/15/15 and 60/20/20) to determine the most effective training and validation balance. Optimization strategies included the Adam and SGD techniques. In total, we conducted 135 unique experimental runs, compiling a comprehensive dataset of 675 metric evaluations.

3 Results

To assess the effectiveness of our models, we utilized a comprehensive set of performance metrics, including weighted average accuracy, precision, recall, F1-score, and AUC. In medical diagnostics, recall becomes a crucial metric because it measures the model's ability to identify all positive instances correctly. High recall is essential in clinical settings as failing to detect conditions like diseases could have dire consequences. Thus, enhancing recall is often prioritized to ensure that few to no positive cases go undetected.

3.1 Performance Analysis with Adam Optimizer using 70/15/15 Split Dataset

3.1.1 Model Evaluation Across Datasets

We evaluated three advanced models, Swin Transformer (SwinT), CNN, and ViT, across multiple datasets, utilizing a 70/15/15 split dataset. The performance metrics, including AUC, accuracy, precision, recall, and F1-score are summarized in Table 10.

Detailed Dataset Performance

- **CDE Hybrid1&2:** SwinT demonstrated superior performance with AUC and accuracy at 0.74 and balanced performance metrics at 0.75. In contrast, CNN and ViT showed lower performance.
- **CDE TC:** CNN achieved the highest AUC at 0.78 but with lower accuracy at 0.62. SwinT and ViT both recorded an AUC of 0.75.
- **CDE X-Ray:** SwinT outperformed others with an AUC of 0.5 and higher precision and recall values.
- **COVID-QU-Ex** and **DSHybrid1:** Both SwinT and CNN achieved nearly perfect scores, with ViT also performing robustly.
- **HCV-UFPR-COVID-19:** SwinT excelled with an AUC of 0.94, significantly higher than CNN and ViT.

Comprehensive Insights Results underscore SwinT's robust performance, establishing it as a potent tool for medical image analysis. CNN shows promise in specific conditions but generally falls short of SwinT and ViT in balanced performance. Figures 7, 8, and 9 display ROC curves and confusion matrices for these models, providing visual support for the performance metrics discussed.

3.2 Performance Analysis with SGD Optimizer Using 70/15/15 Split Dataset

3.2.1 Model Evaluation Across Datasets

Using the SGD optimizer, the models were assessed across similar datasets with the same partition strategy. The performance metrics including AUC, accuracy, precision, recall, and F1-score are summarized in Table 11.

Detailed Dataset Performance

- **CDE Hybrid1&2:** SwinT and ViT both showed consistent performance with an AUC of 0.71. The CNN model lagged slightly with an AUC of 0.67.
- **CDE TC:** CNN led with an AUC of 0.77, although with a lower accuracy. SwinT and ViT achieved competitive but balanced performance metrics.
- **CDE X-Ray:** SwinT and ViT each had an AUC of 0.5, indicating poor predictive performance. However, these models showed slightly better results compared to the other techniques evaluated.
- **COVID-QU-Ex:** All models demonstrated exceptional performance, achieving AUC and other metrics above 0.97.
- **DSHybrid1:** SwinT displayed superior performance with nearly perfect scores across all performance metrics.
- **HCV-UFPR-COVID-19:** SwinT and ViT performed well, with SwinT slightly outperforming ViT.

Comprehensive Insights These findings confirm the efficacy of the SGD optimizer, with SwinT consistently showing high performance across most performance metrics and datasets, suggesting its effectiveness for complex medical imaging tasks. Figures 10, 11, and 12 show the performance metrics obtained using SGD optimization, Adam optimization, and without optimization.

3.3 Summary of Findings with 70/15/15 Split Dataset

Across both Adam and SGD optimizers with a 70/15/15 split, SwinT frequently emerged as the leading model, particularly in challenging datasets. This consistency highlights its suitability for high-stakes applications such as medical diagnostics. CNN and ViT also showed commendable performances but were generally outpaced by SwinT regarding overall recall and F1-Score.

3.4 Performance Analysis with Adam Optimizer Using 60/20/20 Split Dataset

3.4.1 Model Evaluation Across Datasets

SwinT, CNN, and ViT's performance was evaluated across multiple datasets with a partitioning strategy of 60/20/20 for training, validation, and testing respectively. The performance metrics, including AUC, accuracy, precision, recall, and F1-score, are detailed in Table 12.

Detailed Dataset Performance

- **CDE Hybrid1&2:** SwinT maintained superior performance with an AUC of 0.74, similarly demonstrating higher accuracy and balanced performance metrics. CNN and ViT presented lower performance metrics, aligning with their results under the 70/15/15 split.
- **CDE TC:** The CNN model excelled with the highest AUC of 0.78, surpassing SwinT and ViT, which both

reported an AUC of 0.75. This underscores CNN's strength in this particular dataset.

- **CDE X-Ray:** CDE X-Ray: SwinT demonstrated poor predictive performance with an AUC of 0.5 but achieved slightly better precision and recall metrics compared to the other models.
- **COVID-QU-Ex and DSHybrid1:** Both SwinT and CNN almost reached perfect performance metrics, with ViT also showing robust outcomes, especially in precision and recall.
- **HCV-UFPR-COVID-19:** SwinT was notably effective with an AUC of 0.94, proving its applicability in highly specialized medical datasets.

Comprehensive Insights The extended validation set in the 60/20/20 split allowed for more nuanced tuning and testing, reinforcing the robustness of SwinT across datasets. This setup confirmed SwinT's consistent performance, with CNN and ViT showing varying degrees of efficacy depending on the dataset. Figures 10, 11, and 12 show the performance metrics obtained using SGD optimization, Adam optimization, and without optimization.

3.5 Performance Analysis with SGD Optimizer Using 60/20/20 Split Dataset

3.5.1 Model Evaluation Across Datasets

Using the SGD optimizer, we further assessed the models across the same datasets but with a 60/20/20 split dataset, focusing on how the different partitioning affected performance metrics. The performance metrics, including AUC, accuracy, precision, recall, and F1-score, are detailed in Table 13.

Detailed Dataset Performance

- **CDE Hybrid1&2:** Both SwinT and ViT achieved consistent results with AUCs of 0.71, while CNN slightly lagged with an AUC of 0.67.
- **CDE TC:** The CNN model showcased its capability with the highest AUC of 0.77, displaying strong precision and F1-score performance metrics, although SwinT and ViT were not far behind.
- **CDE X-Ray:** SwinT and ViT matched each other with an AUC of 0.5, reflecting poor predictive performance across this challenging dataset. Nonetheless, they performed slightly better than the other techniques.
- **COVID-QU-Ex:** All models excelled, demonstrating their effectiveness in diagnosing COVID-related anomalies with AUCs nearing perfection.
- **DSHybrid1:** SwinT continued its dominance with nearly perfect scores, while CNN and ViT also performed admirably, showing high AUC and balanced performance metrics.
- **HCV-UFPR-COVID-19:** SwinT outperformed other models with a higher AUC, validating its superior image analysis capabilities.

Comprehensive Insights This analysis confirms the benefits of the 60/20/20 split in providing a larger validation set that enhances model tuning and validation, particularly benefiting the CNN in certain datasets. SwinT remains consistently strong, affirming its robustness and reliability. Figures 10, 11, and 12 show the performance metrics obtained using SGD optimization, Adam optimization, and without optimization.

3.6 Summary of Findings with 60/20/20 Split Dataset

The 60/20/20 split facilitated rigorous model validation, allowing SwinT to consistently demonstrate superior performance, while CNN showed improvements in specific datasets when compared to the 70/15/15 split. ViT's performance remained strong but less consistent, highlighting the impact of data partitioning strategies on model efficacy. Future research should explore the scalability of these findings across larger, more diverse datasets and consider integrating multimodal data to further enhance diagnostic accuracy.

3.7 Comparative Analysis of Datasets, Optimizers, and Split Datasets

3.7.1 Comparative Overview

This analysis synthesizes findings across various datasets, optimizers (Adam and SGD), and data partitioning strategies (70/15/15 and 60/20/20). The aim is to delineate the influence of these variables on model performance and identify optimal configurations for specific types of data.

Impact of Datasets The performance variance across different datasets highlights the adaptability of models like SwinT, which consistently showed superior performance metrics, particularly in medically oriented datasets such as HCV-UFPR-COVID-19 and COVID-QU-Ex. CNN tended to excel in more structured datasets like CDE TC, while ViT showed fluctuating but generally robust performance across broader datasets.

It is worth noting that regarding the CDE X-ray dataset, the models were trained and validated on the COVID-QU-Ex dataset and tested on the HCV-UFPR-COVID-19. The possible reasons for the poor results (AUC around 0.5) obtained in the CDE X-ray dataset by all models may be:

- **Differences in Images:** It is important to note that the two datasets may present significant differences in the images. These variations may be related to the quality of imaging equipment used to capture the radiological exams. For example, different X-ray machines can produce images with distinct characteristics, such as resolution, contrast, and artifacts. These differences can affect the models' ability to generalize correctly.
- **Variation in Patients:** Moreover, the variation in patients examined in the images is crucial. Radiological characteristics can vary significantly with disease progression. For instance, a patient in the early stage of

COVID-19 may present different opacity patterns compared to a patient in an advanced stage. This diversity in patients can introduce noise into the data and complicate the classification task.

- **Distinct Features:** The combination of differences in images and patient characteristics can lead to very distinct features in the datasets. Models trained on COVID-QU-Ex may have learned to recognize specific patterns in that set, but these patterns may not directly apply to CDE X-ray. As a result, the models may not be able to generalize well to the new dataset.

Optimizer Influence Comparing the Adam and SGD optimizers, Adam consistently facilitated better performance in terms of AUC and balanced performance metrics, particularly with the SwinT model. The usage of SGD showed slightly less consistency but provided valuable insights into model robustness and efficiency under different optimization pressures.

Data Set Split Configurations The 70/15/15 split proved effective for a general assessment of model capabilities, particularly benefiting the SwinT and ViT models with higher overall performance metrics. Offering a larger validation set, the 60/20/20 split proved advantageous for nuanced model tuning and showed particular benefits for the CNN model. This suggests that more extensive validation phases might be conducive to optimizing models that initially display lower performance metrics. We synthesized the results using tables from performance metrics (AUC, Accuracy, Precision, Recall, F1-Score) compiled in five tables (Tables 10, 11, 12, 13) representing 675 aggregated metric values.

3.8 Statistical Analysis of the Results

3.8.1 Exploratory Analysis of the Aggregated Performance of Models

Figures 14 to 17 present radar and bar charts comparing performance metrics, including AUC, accuracy, precision, recall, and F1-Score, across different models. These visualizations provide valuable insights into the aggregated performance of the analyzed architectures.

The radar charts (Figures 14 and 15) show that SwinT models consistently achieved high median values across most metrics, particularly in precision and recall, standing out as the most balanced architecture among those evaluated. The bar chart (Figure 16) corroborates these findings, highlighting the overall superior performance of SwinT compared to CNN and ViT models. However, in terms of AUC, CNN models demonstrated a slight advantage, as shown.

The percentage difference chart (Figure 17) highlights the significant advantage of SwinT models over ViT models in metrics such as precision and recall, surpassing 80% in some cases. CNN models also outperformed ViT models in most metrics, albeit with smaller margins. These results emphasize SwinT's robustness for tasks requiring high precision and recall, while CNN models proved to be particularly effective in tasks prioritizing AUC.

Our statistical analysis aimed to compare model performances across different split datasets, optimization strategies, and optimizer types.

Initially, the Shapiro-Wilk Test indicated a non-normal distribution for most data (Figure 18), prompting the use of non-parametric tests, specifically the Friedman Test, alongside parametric Paired Samples T-Tests where appropriate. The Friedman Test results (Table 14) revealed statistically significant differences across models and strategies, confirmed by pairwise comparisons using the Durbin-Conover method.

The analyses, supported by heatmaps (Figures 10 to 13), indicated that the SwinT model consistently outperformed the other models across most datasets and performance metrics, particularly excelling in settings where traditional CNN showed limitations. Detailed results from each dataset highlight the Swin Transformer's robustness across varying conditions and its effectiveness in handling complex image classifications like those required in medical diagnostics.

3.8.2 Descriptive Analysis of Model Performance

The analyzed models were evaluated on five performance metrics: Accuracy, AUC, F1-Score, Precision, and Recall, using descriptive statistics from Figures 18 to 23. Below, we present a concise summary of the main results:

- The SwinT models demonstrated the most consistent and superior performance across the majority of analyzed metrics. Models configured with the Adam and SGD optimizers, such as SwinT-DS701515-Adam (I) and SwinT-DS602020-SGD (L), achieved the highest means and medians in Accuracy, F1-Score, Precision, and Recall. Additionally, the SwinT architecture showed robustness even in the absence of optimizers, as evidenced by the SwinT-DS701515-No Optimizer (H) model. However, the improvements provided by optimization strategies were significant. It is worth noting that in terms of AUC, SwinT models were not the best, with a mean of 0.8516 compared to 0.852 for CNN models, but they were well ahead of ViT models, which achieved 0.809.
- The CNN models, especially those optimized with Adam (CNN-DS701515-Adam (D) and CNN-DS602020-Adam (F)), also delivered competitive performance, with consistent results in AUC, Accuracy, and Precision. Nevertheless, these models consistently lagged behind SwinT models in terms of stability and overall performance. Conversely, CNN models configured with SGD or without optimizers exhibited greater variability and lower performance.
- The ViT architecture exhibited greater variability in its results, with mixed performance across metrics. Models optimized with SGD (ViT-DS701515-SGD (O) and ViT-DS602020-SGD (Q)) outperformed those optimized with Adam, excelling in metrics such as F1-Score, Precision, and Recall, yet still falling short of the optimized SwinT and CNN models. The absence of optimization in ViT models (ViT-DS701515-No Optimizer (M)) resulted in significantly lower performance, underscoring the importance of optimization strategies for this architecture.

- Overall, the Adam and SGD optimizers were pivotal in achieving high model performance across all metrics, with Adam being particularly effective for CNN and SwinT architectures, while SGD demonstrated greater impact on ViT and SwinT models. The absence of optimization proved detrimental in almost all architectures, except for SwinT, which displayed a degree of robustness even in scenarios without advanced optimization.

4 Discussion

This study has effectively evaluated the performance of advanced DL models: CNN, SwinT, and ViT, across various configurations. Through extensive statistical analysis and visual data representation via heatmaps (see Figures 10 to 13), the SwinT has demonstrated consistently superior performance across key performance metrics such as AUC, accuracy, precision, recall, and F1-score.

Theoretical and Practical Implications Results affirm SwinT's theoretical capacity to integrate global contextual information, which is crucial in medical image analysis for detecting subtle and nuanced patterns. The practical application of such advanced models in medical diagnostics could significantly streamline workflows, enhance diagnostic accuracy, and expedite patient management, particularly in high-stakes environments like infectious disease diagnosis.

Challenges and Ethical Considerations Despite promising outcomes, the integration of these models into clinical practice raises significant ethical and operational concerns. These models should complement, not replace, the expertise of medical professionals, ensuring that all automated diagnostics are verified by experienced radiologists to manage ambiguities or anomalies not represented in the training data.

Limitations and Future Research The potential data leakage in datasets like HUST-19 could affect the reliability of model evaluations. Future studies should focus on robust dataset partitioning and validation through multicenter studies to verify the effectiveness of these models in varied clinical environments. Additionally, exploring next-generation models and incorporating Explainable AI (XAI) will enhance model transparency and facilitate broader clinical adoption.

Literature Context and Benchmarking Comparatively, the SwinT not only meets but often exceeds the performance of other models reported in the literature (see Table 15, Luz *et al.* [2021] and Mehboob *et al.* [2022]), reinforcing its potential as a significant advancement in medical imaging technologies. However, the performance variability across datasets highlights the need for customized training and optimization strategies to maximize each model's efficacy.

5 Conclusions

The study conclusively demonstrates SwinT's effectiveness in classifying COVID-19 from radiographic images, empha-

sizing its superior generalization capabilities across diverse datasets, a critical factor for real-world clinical applications.

Integration into Clinical Practice The integration of the SwinT into clinical settings should be judicious, with the model serving as an augmentation to, rather than a replacement for, human expertise. This ensures that diagnostic processes benefit from both advanced AI capabilities and professional medical judgment.

Challenges and Future Directions Addressing the challenges of data leakage and dataset diversity is critical for advancing the practical application of these models. Future research should aim to broaden the Swin Transformer's application scope to include other medical conditions and imaging modalities and integrate XAI techniques to enhance the transparency of its diagnostic processes.

Advancing Medical AI Research Building partnerships with medical institutions for access to comprehensive and varied datasets will be essential. These collaborations, along with ongoing model refinement based on real-world clinical feedback, will be crucial for the successful implementation of AI technologies in healthcare settings.

Concluding Reflections The Swin Transformer stands out as a transformative tool in medical diagnostics, capable of significantly enhancing the accuracy and efficiency of radiological assessments. As the medical AI field evolves, the thoughtful integration of such technologies into clinical practice is imperative, ensuring they align with ethical standards and contribute positively to patient care.

Declaration

Acknowledgements

We would like to thank professors David Menotti (Federal University of Paraná – UFPR), João Guimarães, and Gustavo Miozzo (Hospital Cruz Vermelha—HCV) for their indispensable support in providing the images of the HCV-UFPR-COVID-19 dataset, without which the work would have been much more difficult.

Funding

This research was not funded.

Authors' Contributions

All authors were involved in every stage of the work and contributed equally to its completion.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Three datasets used in the work are public: the HUST-19 Dataset [CNBC, 2022], the COVID-QU-Ex Dataset [Tahir, 2022], and the SARS-COV2 CT dataset [Soares and Angelov, 2020]. The fourth dataset, HCV-UFPR-COVID-19 X-ray, is made available to researchers upon request on a case-by-case basis.

References

- Abiyev, R. and Ismail, A. (2021). COVID-19 and Pneumonia Diagnosis in X-Ray Images Using Convolutional Neural Networks. *Mathematical Problems in Engineering*, 2021. DOI: 10.1155/2021/3281135.
- Ahamed, K., Islam, M., Uddin, A., Akhter, A., Paul, B., Yousuf, M., Uddin, S., Quinn, J., and Moni, M. (2021). A deep learning approach using effective preprocessing techniques to detect COVID-19 from chest CT-scan and X-ray images. *Computers in Biology and Medicine*, 139. DOI: 10.1016/j.compbiomed.2021.105014.
- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631. DOI: 10.1145/3292500.3330701.
- Al Rahhal, M., Bazi, Y., Jomaa, R., Alshibli, A., Alajlan, N., Mekhalfi, M., and Melgani, F. (2022). COVID-19 detection in CT/x-ray imagery using vision transformers. *Journal of Personalized Medicine*, 12(2). DOI: 10.3390/jpm12020310.
- Asif, S., Zhao, M., Tang, F., and Zhu, Y. (2022). A deep learning-based framework for detecting COVID-19 patients using chest X-rays. *Multimedia Systems*, 28(4):1495–1513. DOI: 10.1007/s00530-022-00917-7.
- Awan, M., Bilal, M., Yasin, A., Nobanee, H., Khan, N., and Zain, A. (2021). Detection of covid-19 in chest x-ray images: A big data enabled deep learning approach. *International Journal of Environmental Research and Public Health*, 18(19). DOI: 10.3390/ijerph181910147.
- Aytekin, I., Dalmaz, O., Ankishan, H., Saritas, E., Bagci, U., Cukur, T., and Celik, H. (2022). Detecting COVID-19 from respiratory sound recordings with transformers. In Drukker K., I. K., editor, *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, volume 12033. SPIE. ISSN: 16057422. DOI: 10.1117/12.2611490.
- Bakator, M. and Radosav, D. (2018). Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3). DOI: 10.3390/mti2030047.
- Balderas, L., Lastra, M., Láinez-Ramos-Bossini, A. J., and Benítez, J. M. (2023). Covid-vit: Covid-19 detection method based on vision transformers. In *Intelligent Systems Design and Applications*, volume 716. Springer International Publishing. DOI: 10.1007/978-3-031-35501-1_8.
- Banerjee, A., Bhattacharya, R., Bhateja, V., Singh, P., Lay-Ekuakille, A., and Sarkar, R. (2022). COFE-Net: An ensemble strategy for Computer-Aided Detection for COVID-19. *Measurement: Journal of the International Measurement Confederation*, 187. DOI: 10.1016/j.measurement.2021.110289.
- Cao, K., Deng, T., Zhang, C., Lu, L., and Li, L. (2022). A CNN-transformer fusion network for COVID-19 CXR image classification. *PLoS ONE*, 17(10 October). DOI: 10.1371/journal.pone.0276758.
- Castiglione, A., Vijayakumar, P., Nappi, M., Sadiq, S., and Umer, M. (2021). COVID-19: Automatic detection of the novel coronavirus disease from CT images using an optimized convolutional neural network. *IEEE Transactions on Industrial Informatics*, 17(9):6480–6488. DOI: 10.1109/TII.2021.3057524.
- Castro, R., Luz, P. M., Wakimoto, M. D., Veloso, V. G., Grinsztejn, B., and Perazzo, H. (2020). COVID-19: a meta-analysis of diagnostic test accuracy of commercial assays registered in brazil. *The Brazilian Journal of Infectious Diseases*, 24(2):180–187. DOI: 10.1016/j.bjid.2020.04.003.
- Cha, S.-M., Lee, S.-S., and Ko, B. (2021). Attention-based transfer learning for efficient pneumonia detection in chest x-ray images. *Applied Sciences*, 11(3). DOI: 10.3390/app11031242.
- Chen, H., Zhang, T., Chen, R., Zhu, Z., and Wang, X. (2023). A Novel COVID-19 Image Classification Method Based on the Improved Residual Network. *Electronics (Switzerland)*, 12(1). DOI: 10.3390/electronics12010080.
- Chetoui, M. and Akhloufi, M. (2022). Explainable vision transformers and radiomics for COVID-19 detection in chest x-rays. *Journal of Clinical Medicine*, 11(11). Publisher: MDPI. DOI: 10.3390/jcm11113013.
- Chollet, F. (2017). Xception: Deep learning with depth-wise separable convolutions. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 1800–1807. Institute of Electrical and Electronics Engineers Inc.. DOI: 10.1109/CVPR.2017.195.
- Chouhan, V., Singh, S., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R., and de Albuquerque, V. (2020). A novel transfer learning based approach for pneumonia detection in chest x-ray images. *Applied Sciences (Switzerland)*, 10(2). DOI: 10.3390/app10020559.
- Cireşan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. DOI: 10.1109/CVPR.2012.6248110.
- CNBC, N. G. D. C. (2022). Hust-19 database. Available at: <https://ngdc.cncb.ac.cn/ictcf/HUST-19.php>. Accessed: 05/05/2023.
- Dehkordi, H., Kashiani, H., Hamidi Imani, A., and Shokouhi, S. (2021). Lightweight local transformer for COVID-19 detection using chest CT scans. In *ICCKE 2021 - 11th International Conference on Computer Engineering and Knowledge*, pages 328–333. Institute of Electrical and Electronics Engineers Inc.. DOI: 10.1109/ICCKE54056.2021.9721517.
- Dinh, T., Lee, S.-H., Kwon, S.-G., and Kwon, K.-R. (2022). COVID-19 Chest X-ray Classification and Severity As-

- essment Using Convolutional and Transformer Neural Networks. *Applied Sciences (Switzerland)*, 12(10). DOI: 10.3390/app12104861.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929. DOI: 10.48550/arXiv.2010.11929.
- Geng, L., Zhang, S., Tong, J., and Xiao, Z. (2019). Lung segmentation method with dilated convolution based on VGG-16 network. *Computer Assisted Surgery*, 24:27–33. Publisher: Taylor & Francis. DOI: 10.1080/24699322.2019.1649071.
- Hamza, A., Attique Khan, M., Wang, S., Alhaisoni, M., Alharbi, M., Hussein, H., *et al.* (2022). Covid-19 classification using chest x-ray images based on fusion-assisted deep bayesian optimization and grad-cam visualization. *Frontiers in Public Health*, 10. Available at: <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.1046296/full>.
- Hashmi, M., Katiyar, S., Keskar, A., Bokde, N., and Geem, Z. (2020). Efficient pneumonia detection in chest xray images using deep transfer learning. *Diagnostics*, 10(6). DOI: 10.3390/diagnostics10060417.
- Hassan, H., Ren, Z., Zhao, H., Huang, S., Li, D., Xiang, S., Kang, Y., Chen, S., and Huang, B. (2022). Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision tasks. *Computers in Biology and Medicine*, 141. Publisher: Elsevier Ltd. DOI: 10.1016/j.combiomed.2021.105123.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. DOI: 10.1109/CVPR.2016.90.
- Hussain, A., Hussain, T., Ullah, W., and Baik, S. (2022). Vision transformer and deep sequence learning for human activity recognition in surveillance videos. *Computational Intelligence and Neuroscience*, 2022:3454167. DOI: 10.1155/2022/3454167.
- ImageNet1k (2017). Imagenet. Available at: <https://www.image-net.org/download.php>.
- Jain, R., Nagrath, P., Kataria, G., Sirish Kaushik, V., and Jude Hemanth, D. (2020). Pneumonia detection in chest x-ray images using convolutional neural networks and transfer learning. *Measurement: Journal of the International Measurement Confederation*, 165. DOI: 10.1016/j.measurement.2020.108046.
- Jiang, X., Zhu, Y., Cai, G., Zheng, B., and Yang, D. (2022). MXT: A new variant of pyramid vision transformer for multi-label chest x-ray image classification. *Cognitive Computation*. Publisher: Springer. DOI: 10.1007/s12559-022-10032-4.
- Jung, M. and Chi, S. (2020). Human activity classification based on sound recognition and residual convolutional neural network. *Automation in Construction*, 114:103177. DOI: <https://doi.org/10.1016/j.autcon.2020.103177>.
- Kapoor, N. (2021). Recall, specificity, precision, f1-score and accuracy. Available at: <https://tinyurl.com/287xse3x>.
- Kathamuthu, N., Subramaniam, S., Le, Q., Muthusamy, S., Panchal, H., Sundararajan, S., Alrubai, A., and Maher Abdul Zahra, M. (2023). A deep transfer learning-based convolution neural network model for COVID-19 detection using computed tomography scan images for medical applications. *Advances in Engineering Software*, 175. DOI: 10.1016/j.advengsoft.2022.103317.
- Khobragade, P. P. and Manthalkar, R. (2024). Thoracic computed tomography (ct) image-based identification and severity classification of covid-19 cases using vision transformer (vit). *SN Applied Sciences*, 6:48. DOI: 10.1007/s42452-024-06048-0.
- Kibriya, H. and Amin, R. (2023). A residual network-based framework for COVID-19 detection from CXR images. *Neural Computing and Applications*, 35(11):8505–8516. DOI: 10.1007/s00521-022-08127-y.
- Konwer, A. and Prasanna, P. (2022). Clinical outcome prediction in COVID-19 using self-supervised vision transformer representations. In Drukker K., I. K., editor, *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, volume 12033. SPIE. ISSN: 16057422. DOI: 10.1117/12.2612957.
- Krishnan, K. and Krishnan, K. (2021). Vision transformer based COVID-19 detection using chest x-rays. In Kumar R., Jain S., S. H., editor, *Proceedings of IEEE International Conference on Signal Processing, Computing and Control*, volume 2021-October, pages 644–648. Institute of Electrical and Electronics Engineers Inc. ISSN: 26438615. DOI: 10.1109/ISPC53510.2021.9609375.
- Lanjewar, M., Shaikh, A., and Parab, J. (2022). Cloud-based COVID-19 disease prediction system from X-Ray images using convolutional neural network on smartphone. *Multi-media Tools and Applications*. DOI: 10.1007/s11042-022-14232-w.
- Li, J., Yang, Z., and Yu, Y. (2021). A medical AI diagnosis platform based on vision transformer for coronavirus. In *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology, CEI 2021*, pages 246–252. Institute of Electrical and Electronics Engineers Inc.. DOI: 10.1109/CEI52496.2021.9574576.
- Liang, G. and Zheng, L. (2020). A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer Methods and Programs in Biomedicine*, 187. DOI: 10.1016/j.cmpb.2019.06.023.
- Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26. DOI: 10.1016/j.neucom.2016.12.038.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002. DOI: 10.1109/ICCV48922.2021.00986.
- Luján-García, J. E., Yáñez-Márquez, C., Villuendas-Rey, Y., and Camacho-Nieto, O. (2020). A transfer learning method for pneumonia classification and visualization.

- Applied Sciences*, 10(8). DOI: 10.3390/app10082908.
- Luz, E., Silva, P., Silva, R., Silva, L., Guimarães, J., Miozzo, G., Moreira, G., and Menotti, D. (2021). Towards an effective and efficient deep learning model for COVID-19 patterns detection in x-ray images. *Research on Biomedical Engineering*, 38(1):149–162. Publisher: Springer Science and Business Media Deutschland GmbH. DOI: 10.1007/s42600-021-00151-6.
- Ma, Y. and Lv, W. (2022). Identification of pneumonia in chest x-ray image based on transformer. *International Journal of Antennas and Propagation*, 2022. DOI: 10.1155/2022/5072666.
- Marefat, A., Marefat, M., Hassannataj Joloudari, J., Nematollahi, M. A., and Lashgari, R. (2023). Cctcovid: Covid-19 detection from chest x-ray images using compact convolutional transformers. *Frontiers in Public Health*, 11. DOI: 10.3389/fpubh.2023.1025746.
- Meedeniya, D., Kumarasinghe, H., Kolonne, S., Fernando, C., Díez, I., and Marques, G. (2022). Chest x-ray analysis empowered with deep learning: A systematic review. *Applied Soft Computing*, 126. DOI: 10.1016/j.asoc.2022.109319.
- Mehboob, F., Rauf, A., Jiang, R., Saudagar, A. K. J., Malik, K. M., Khan, M. B., Hasnat, M. H. A., Altameem, A., and AlKhathami, M. (2022). Towards robust diagnosis of COVID-19 using vision self-attention transformer. *SCIENTIFIC REPORTS*, 12(1). DOI: 10.1038/s41598-022-13039-x.
- Mondal, A., Bhattacharjee, A., Singla, P., and Prathosh, A. (2022). XViTCOS: Explainable vision transformer based COVID-19 screening using radiography. *IEEE Journal of Translational Engineering in Health and Medicine*, 10. Publisher: Institute of Electrical and Electronics Engineers Inc.. DOI: 10.1109/JTEHM.2021.3134096.
- Murphy, Z., Venkatesh, K., Sulam, J., and Yi, P. (2022). Visual Transformers and Convolutional Neural Networks for Disease Classification on Radiographs: A Comparison of Performance, Sample Efficiency, and Hidden Stratification. *Radiology: Artificial Intelligence*, 4(6). DOI: 10.1148/ryai.220012.
- Ning, W., Lei, S., Yang, J., Cao, Y., Jiang, P., Yang, Q., and et al. (2020). Open resource of clinical data from patients with pneumonia for the prediction of covid-19 outcomes via deep learning. *Nature Biomedical Engineering*, 4:1197–1207. DOI: 10.1038/s41551-020-00633-5.
- Nishio, M., Kobayashi, D., Nishioka, E., Matsuo, H., Urase, Y., Onoue, K., Ishikura, R., Kitamura, Y., Sakai, E., Tomita, M., Hamanaka, A., and Murakami, T. (2022). Deep learning model for the automatic classification of COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy: a multi-center retrospective study. *Scientific Reports*, 12(1). DOI: 10.1038/s41598-022-11990-3.
- Pan, S., Wang, T., Qiu, R., Axente, M., Chang, C.-W., Peng, J., Patel, A., Shelton, J., Patel, S., Roper, J., and Yang, X. (2023). 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine and Biology*, 68(10). DOI: 10.1088/1361-6560/acca5c.
- Park, S., Kim, G., Oh, Y., Seo, J., Lee, S., Kim, J., Moon, S., Lim, J.-K., Park, C., and Ye, J. (2022a). Self-evolving vision transformer for chest x-ray diagnosis through knowledge distillation. *Nature Communications*, 13(1). DOI: 10.1038/s41467-022-31514-x.
- Park, S., Kim, G., Oh, Y., Seo, J., Lee, S., Kim, J., Moon, S., Lim, J.-K., and Ye, J. (2022b). Multi-task vision transformer using low-level chest x-ray feature corpus for COVID-19 diagnosis and severity quantification. *Medical Image Analysis*, 75. DOI: 10.1016/j.media.2021.102299.
- Peng, L., Wang, C., Tian, G., Liu, G., Li, G., Lu, Y., Yang, J., Chen, M., and Li, Z. (2022). Analysis of ct scan images for covid-19 pneumonia based on a deep ensemble framework with densenet, swin transformer, and regnet. *Frontiers In Microbiology*, 13:1. DOI: 10.3389/fmicb.2022.995323.
- PyTorch (2023). Previous pytorch versions | pytorch. Available at: <https://pytorch.org/get-started/previous-versions/>.
- Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *CoRR*, abs/2108.08810.
- Rahman, T., Chowdhury, M., Khandakar, A., Islam, K., Islam, K., Mahbub, Z., Kadir, M., and Kashem, S. (2020). Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest x-ray. *Applied Sciences (Switzerland)*, 10(9). DOI: 10.3390/app10093233.
- Raman, B., Cassar, M. P., Tunnicliffe, E. M., Filippini, N., Griffanti, L., Alfaro-Almagro, F., Okell, T., Sheerin, F., Xie, C., Mahmood, M., Mózes, F. E., Lewandowski, A. J., Ohuma, E. O., Holdsworth, D., Lamlum, H., Woodman, M. J., Krasopoulos, G., Mills, R., McConnell, F. A. K., Wang, C., Arthofer, C., Lange, F. J., Andersson, J., Jenkinson, M., Antoniadis, C., Channon, K., Shanmuganathan, M., Ferreira, V. M., Piechnik, S. K., Klennerman, P., Brightling, C., Talbot, N. P., Petousi, N., Rahman, N. M., Ho, L. P., Saunders, K., Geddes, J. R., Harrison, P. J., Pattinson, K., Rowland, M. J., Angus, B. J., Gleeson, F., Pavlides, M., Koychev, I., Miller, K. L., Mackay, C., Jezard, P., Smith, S. M., and Neubauer, S. (2021). Medium-term effects of sars-cov-2 infection on multiple vital organs, exercise capacity, cognition, quality of life and mental health, post-hospital discharge. *EClinicalMedicine*, 31:100683. DOI: 10.1016/j.eclim.2020.100683.
- Shekhar, S., Bansode, A., and Salim, A. (2021). A comparative study of hyper-parameter optimization tools. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2021*.
- Shome, D., Kar, T., Mohanty, S., Tiwari, P., Muhammad, K., Altameem, A., Zhang, Y., and Saudagar, A. (2021). Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health*, 18(21). Publisher: MDPI. DOI: 10.3390/ijerph182111086.
- Soares, E. and Angelov, P. (2020). Sars-cov-2 ct-scan dataset. DOI: 10.34740/KAGGLE/DSV/1199870.
- Srivastava, G., Pradhan, N., and Saini, Y. (2022). Ensemble of Deep Neural Networks based on Condorcet's Jury Theorem for screening Covid-19 and Pneumonia from radiograph images. *Computers in Biology and Medicine*, 149.

- DOI: 10.1016/j.compbmed.2022.105979.
- Sun, W., Qin, Z., Deng, H., Wang, J., Zhang, Y., Zhang, K., Barnes, N., Birchfield, S., Kong, L., and Zhong, Y. (2022). Vicinity vision transformer. DOI: 10.48550/arXiv.2206.10552.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. DOI: 10.1109/CVPR.2016.308.
- Tahir, A. M. (2022). Covid-qu dataset. Available at: <https://www.kaggle.com/datasets/anasmohammedtahir/covidqu>.
- Tahir, A. M., Chowdhury, M. E. H., Qiblawey, Y., Khandakar, A., Rahman, T., Kiranyaz, S., Khurshid, U., Ibtehaz, N., Mahmud, S., and Ezeddin, M. (2021). Covid-qu. DOI: 10.34740/KAGGLE/DSV/2759090.
- Than, J., Thon, P., Rijal, O., Kassim, R., Yunus, A., Noor, N., and Then, P. (2021). Preliminary study on patch sizes in vision transformers (ViT) for COVID-19 and diseased lungs classification. In *1st National Biomedical Engineering Conference, NBEC 2021*, pages 146–150. Institute of Electrical and Electronics Engineers Inc.. DOI: 10.1109/NBEC53282.2021.9618751.
- Tian, G., Wang, Z., Wang, C., Chen, J., Liu, G., Xu, H., Lu, Y., Han, Z., Zhao, Y., and Li, Z. (2022). A deep ensemble learning-based automated detection of covid-19 using lung ct images and vision transformer and convnext. *Frontiers In Microbiology*, 13:1. DOI: 10.3389/fmicb.2022.1024104.
- Tuli, S., Dasgupta, I., Grant, E., and Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? *CoRR*, abs/2105.07197. DOI: 10.48550/arXiv.2105.07197.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Fergus, R., Wallach, H., Wallach, H., Guyon, I., Vishwanathan, S., Luxburg, U. v., Garnett, R., Vishwanathan, S., Bengio, S., and Fergus, R., editors, *Advances in Neural Information Processing Systems*, volume 2017-December, pages 5999 – 6009. Neural information processing systems foundation. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Wang, T., Nie, Z., Wang, R., Xu, Q., Huang, H., Xu, H., Xie, F., and Liu, X.-J. (2023). PneuNet: deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using Vision Transformer. *Medical and Biological Engineering and Computing*, 61(6):1395–1408. DOI: 10.1007/s11517-022-02746-2.
- Wei, Z., Pan, H., Li, L., Lu, M., Niu, X., Dong, P., and Li, D. (2022). Dmformer: Closing the gap between cnn and vision transformers. *arXiv preprint arXiv:2209.07738*. DOI: 10.48550/ARXIV.2209.07738.
- WHO (2024). Who coronavirus (covid-19) dashboard | who coronavirus (covid-19) dashboard with vaccination data. <https://covid19.who.int/>. Accessed on May 23, 2024.
- Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. (2018). The marginal value of adaptive gradient methods in machine learning. DOI: <https://doi.org/10.48550/arXiv.1705.08292>.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-demos.6.
- Yao, Z., Li, J., Guan, Z., Ye, Y., and Chen, Y. (2020). Liver disease screening based on densely connected deep neural networks. *Neural networks : the official journal of the International Neural Network Society*, 123:299–304. Place: United States. DOI: 10.1016/j.neunet.2019.11.005.
- Yuan, J., Wu, F., Li, Y., Li, J., Huang, G., and Huang, Q. (2023). DPDH-CapNet: A Novel Lightweight Capsule Network with Non-routing for COVID-19 Diagnosis Using X-ray Images. *Journal of Digital Imaging*. DOI: 10.1007/s10278-023-00791-3.
- Zhang, L. and Wen, Y. (2021). A transformer-based framework for automatic COVID19 diagnosis in chest CTs. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2021-October, pages 513–518. Institute of Electrical and Electronics Engineers Inc. ISSN: 15505499. DOI: 10.1109/ICCVW54120.2021.00063.
- Zhao, J., Zhang, Y., He, X., and Xie, P. (2020). Covid-ct-dataset: A CT scan dataset about COVID-19. *CoRR*, abs/2003.13865. DOI: 10.48550/arXiv.2003.13865.

Appendix: Extra Figures and Tables

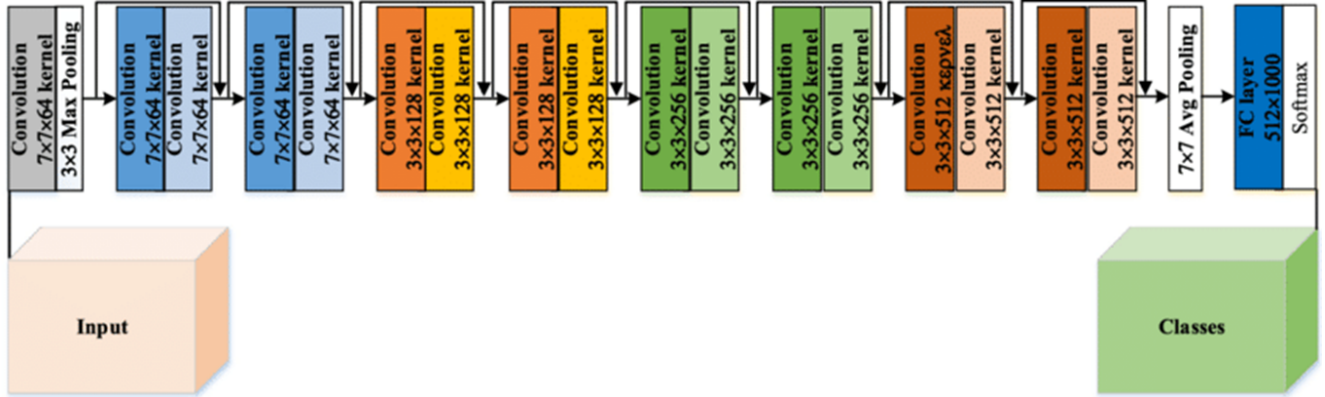


Figure 1. ResNet Architecture. [He *et al.*, 2016]

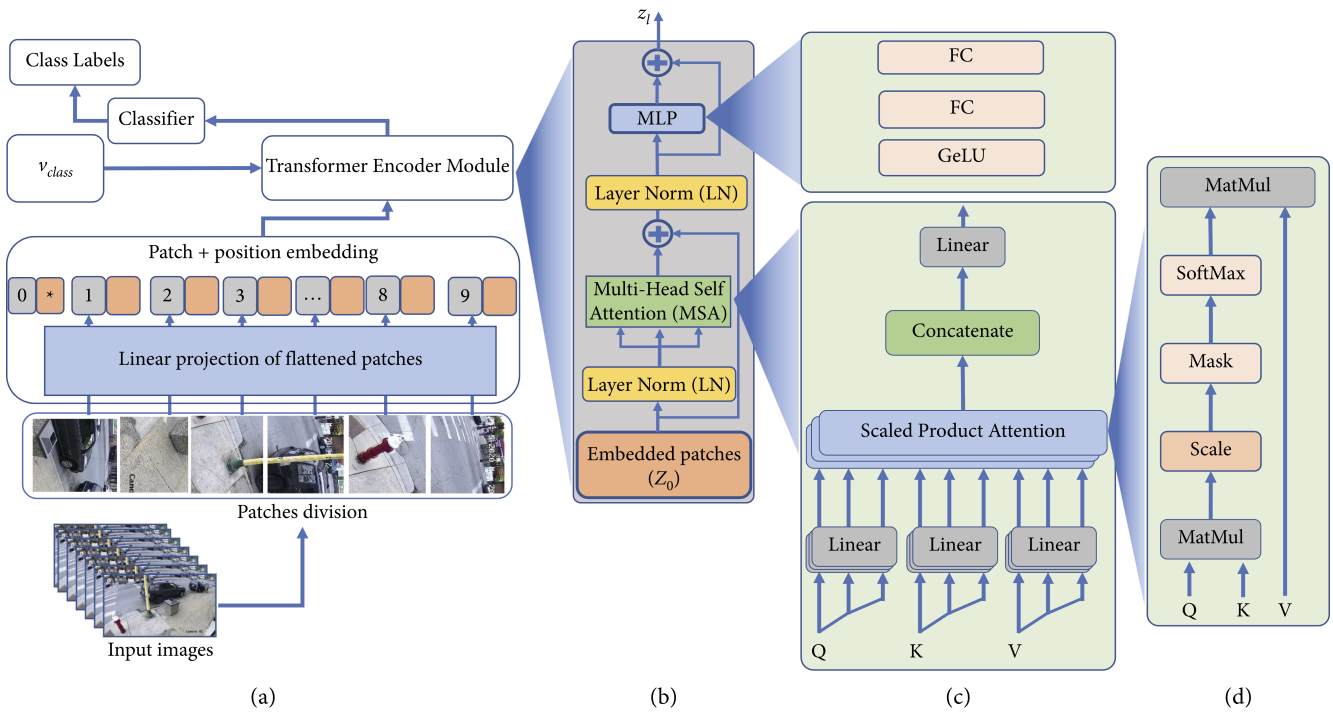


Figure 2. The Vision Transformer architecture: (a) the main architecture of the model, (b) the transformer encoder module and (c) multiscale self-attention (MSA) head and (d) the self-attention (SA) head. [Hussain *et al.*, 2022]

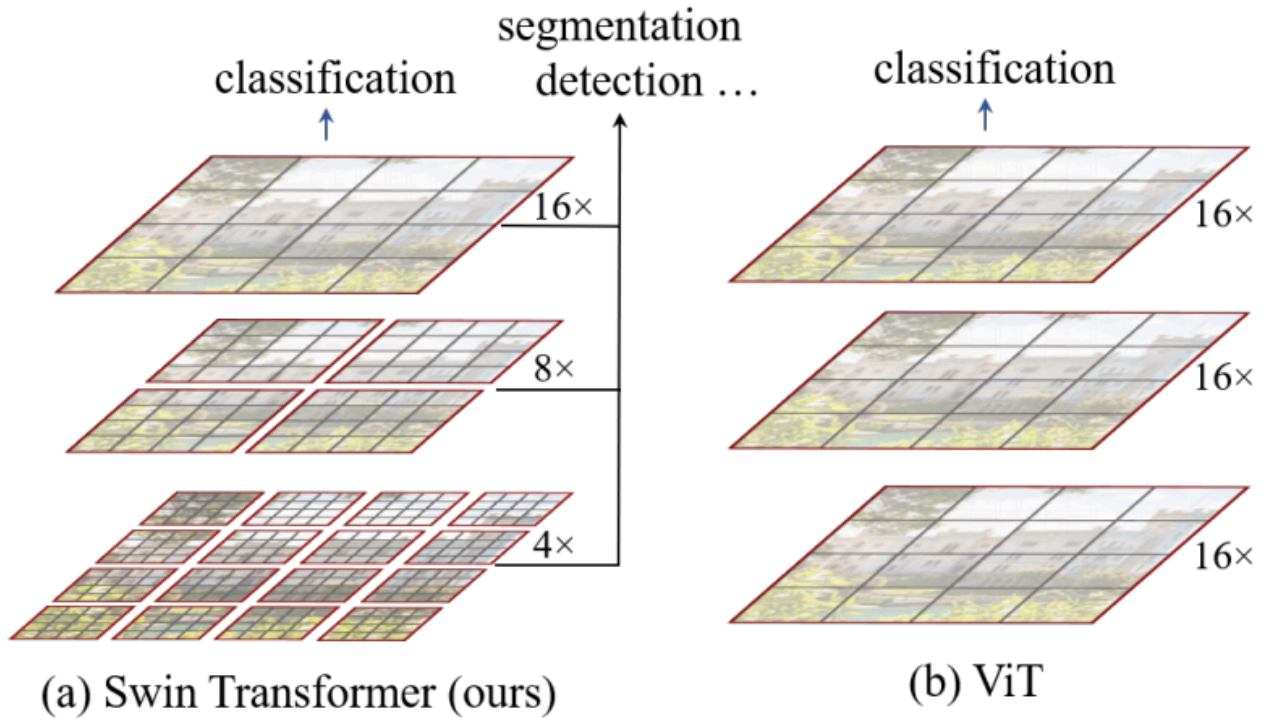


Figure 3. (a) The Swin Transformer [Liu *et al.*, 2021] builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous ViT [Dosovitskiy *et al.*, 2020] produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.

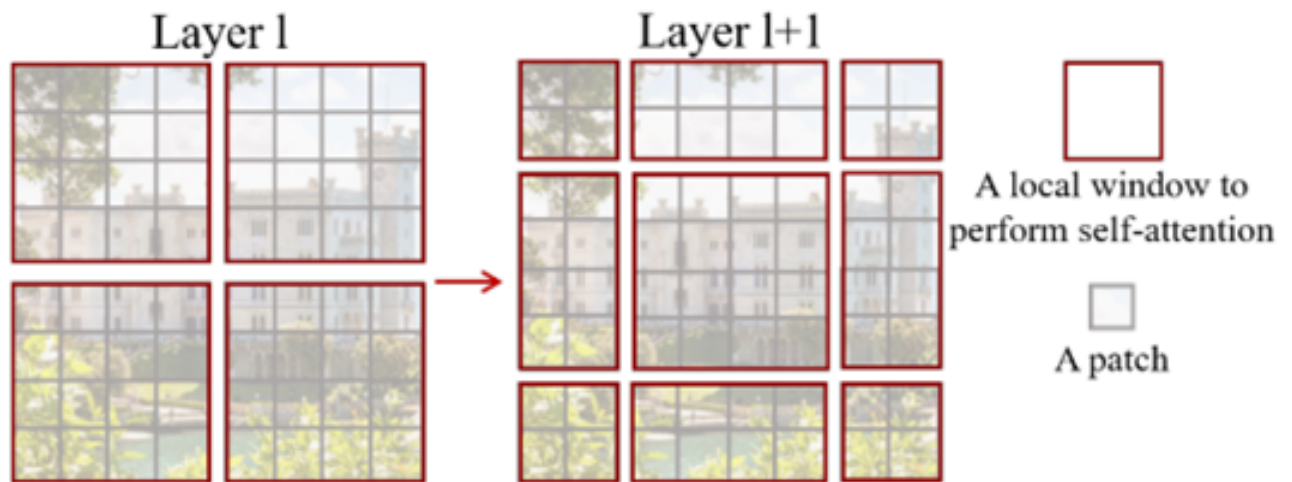


Figure 4. Swin Transformer Shifted Window. [Liu *et al.*, 2021]

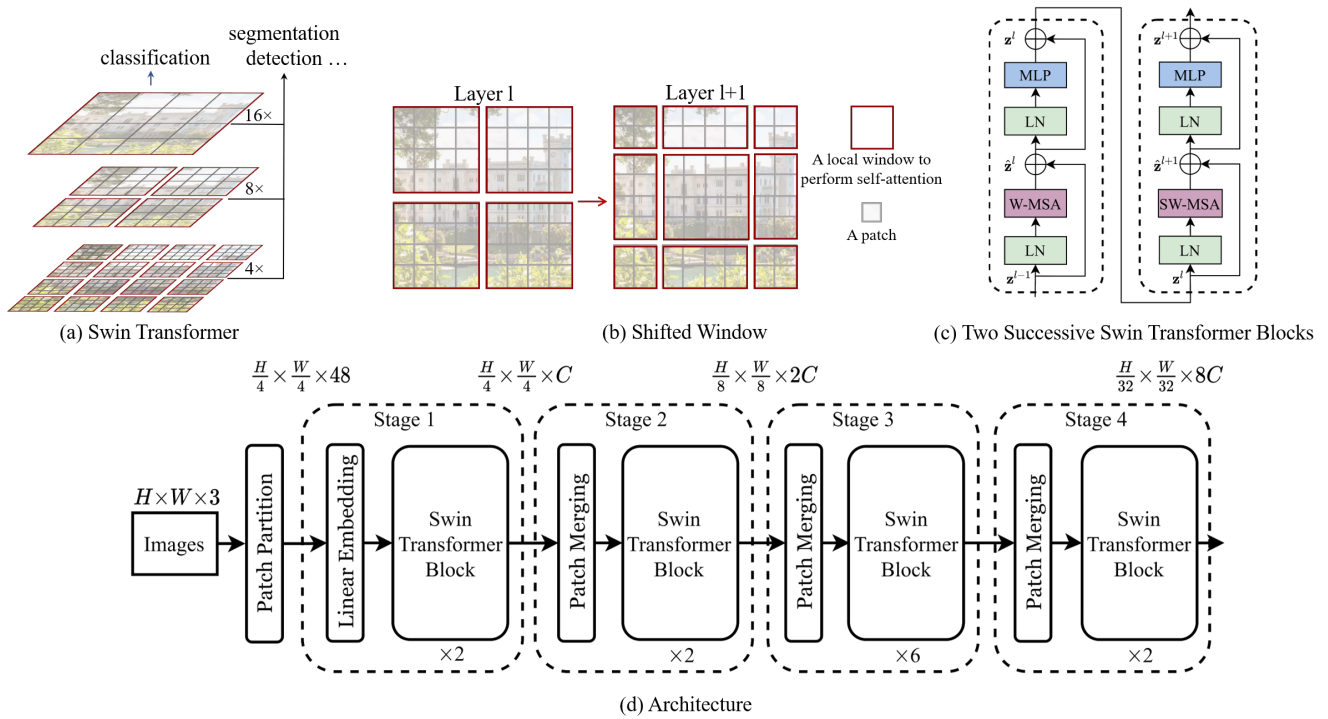


Figure 5. (a) Swin Transformer patches splitting strategy, (b) Shifted Window approach, (c) Swin Transformer Blocks and (d) Swin Transformer Architecture overview. [Liu *et al.*, 2021]

		Predicted		
		Positive	Negative	
Actual	Positive	True positive(TP)	False Negative(FN)	Sensitivity or Recall or True Positive Rate=TP/(TP+FN)
	Negative	False Positive (FP)	True Negative(TN)	Specificity or True Negative Rate=TN/(TN+FP)
		Precision or Positive Predictive Value=TP/(TP+FP)	Negative Predictive Value=FN/(FN+TN)	Accuracy=TP+TN/TP+TN+FP+FN

Figure 6. Performance Metrics used in Deep Learning. [Kapoor, 2021]

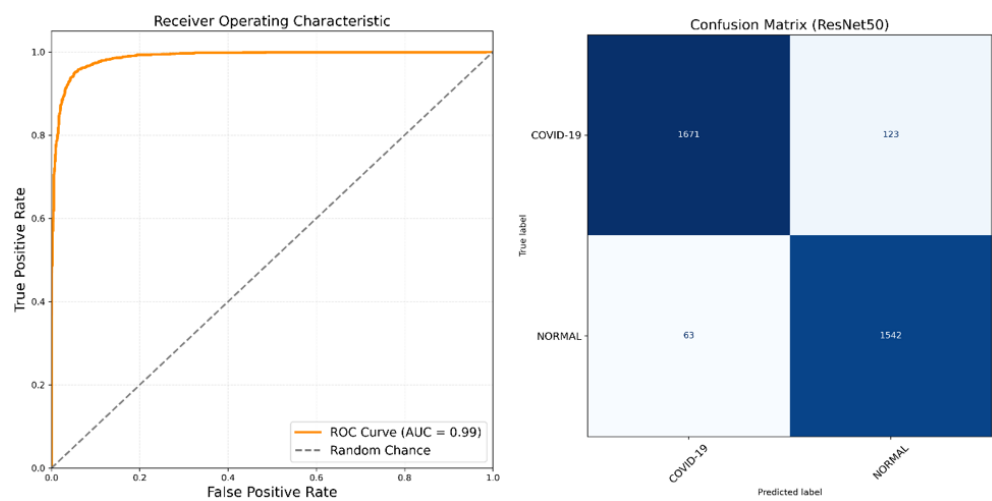


Figure 7. Performance Metrics obtained in dataset COVID-QU-Ex by the ResNet50. (Created by the author)

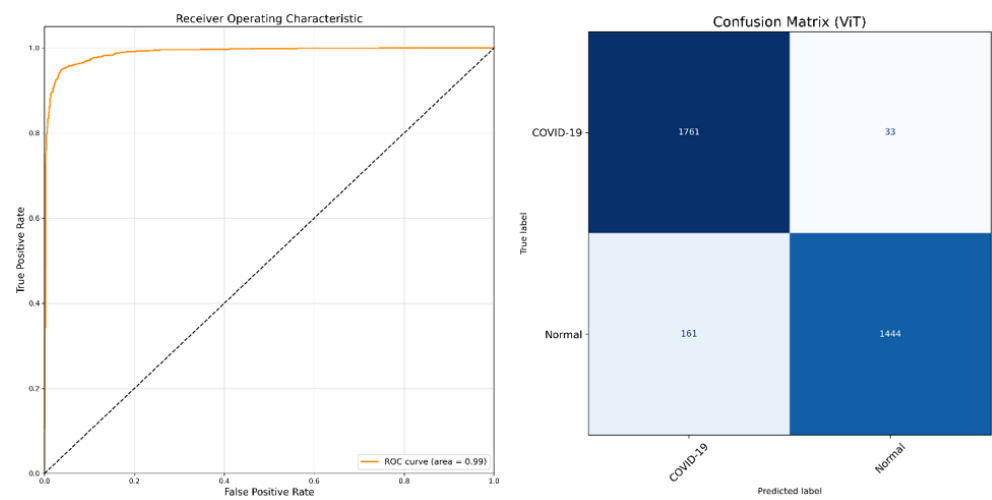


Figure 8. Performance Metrics obtained in dataset COVID-QU-Ex by the Vision Transformer. (Created by the author)

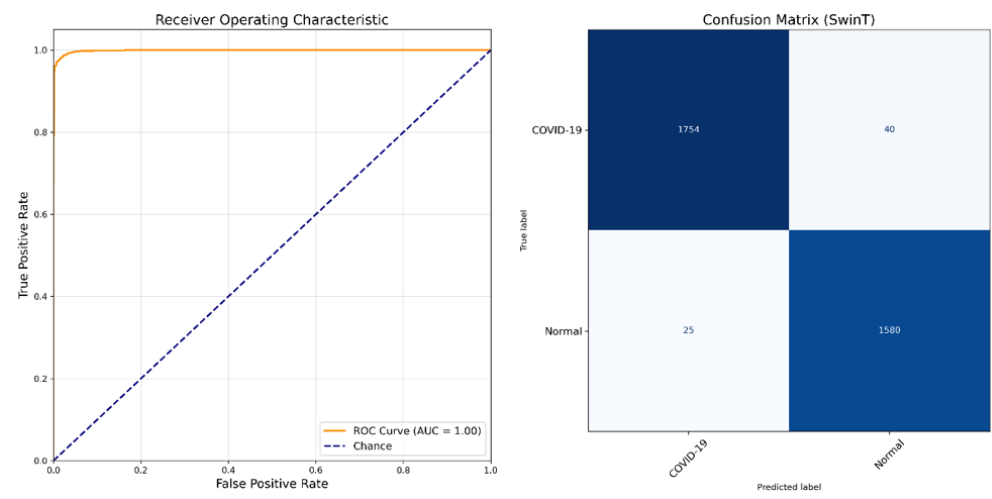


Figure 9. Performance Metrics obtained in dataset COVID-QU-Ex by the Swin Transformer. (Created by the author)

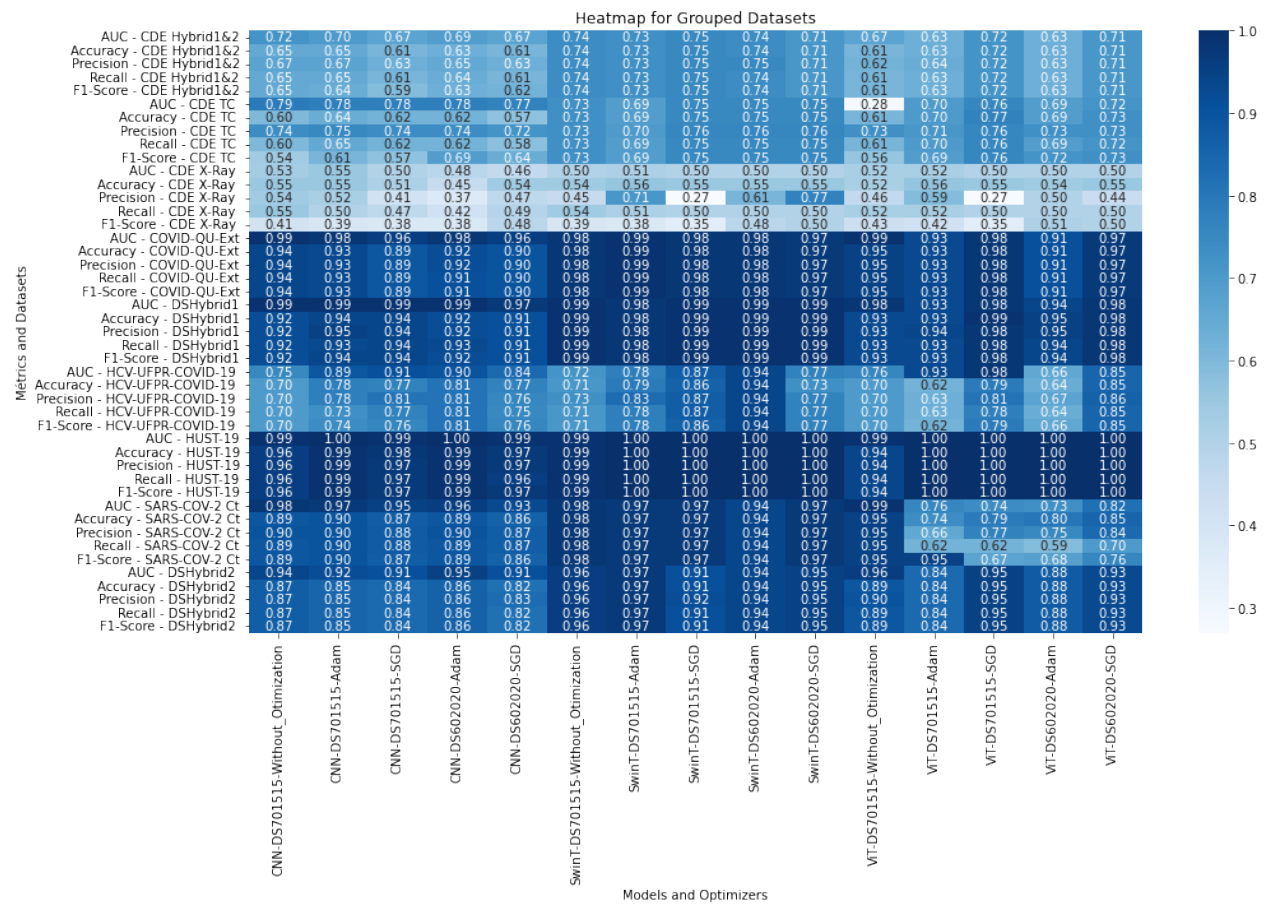
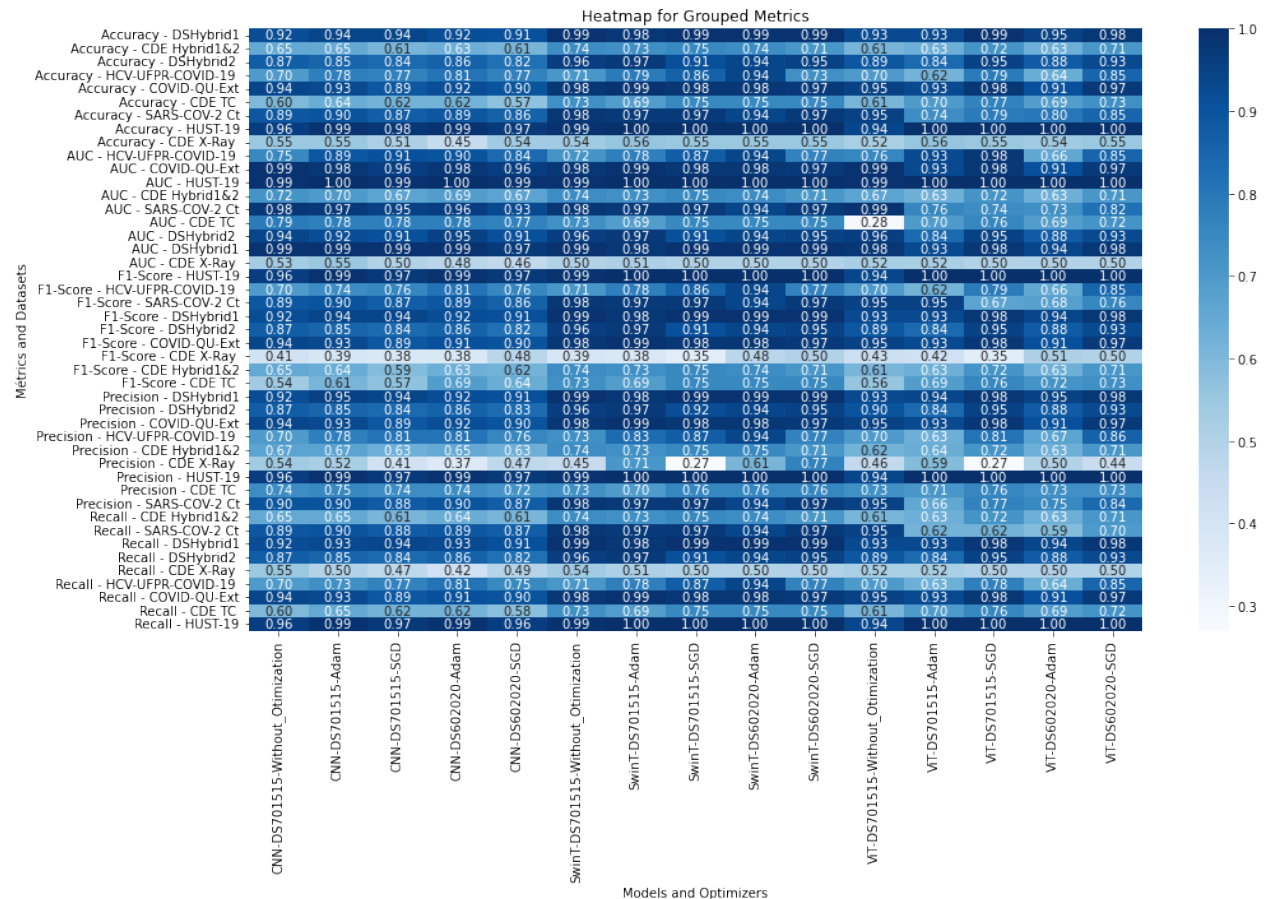


Figure 10. Heatmap for Grouped Datasets. (Created by the author)



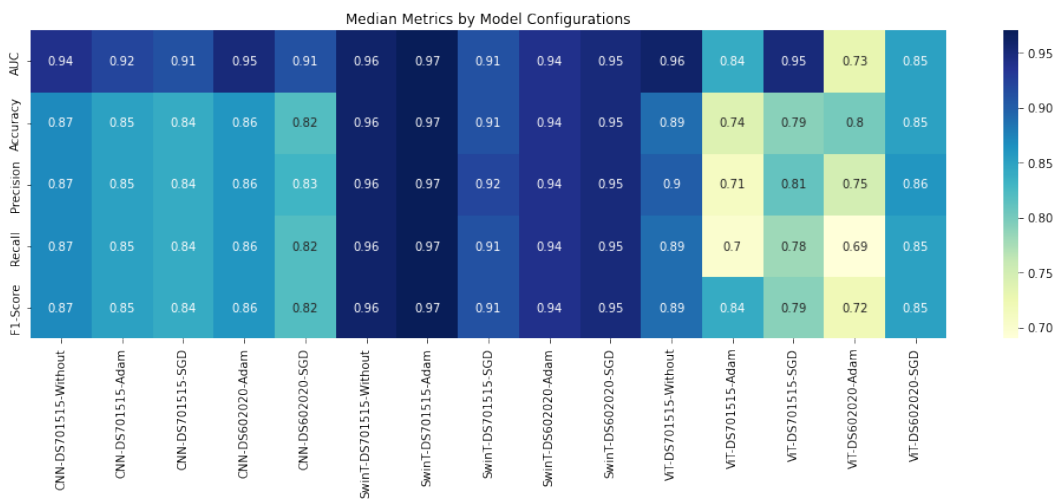


Figure 12. Heatmap Medians Performance Metrics by Model. (Created by the author)

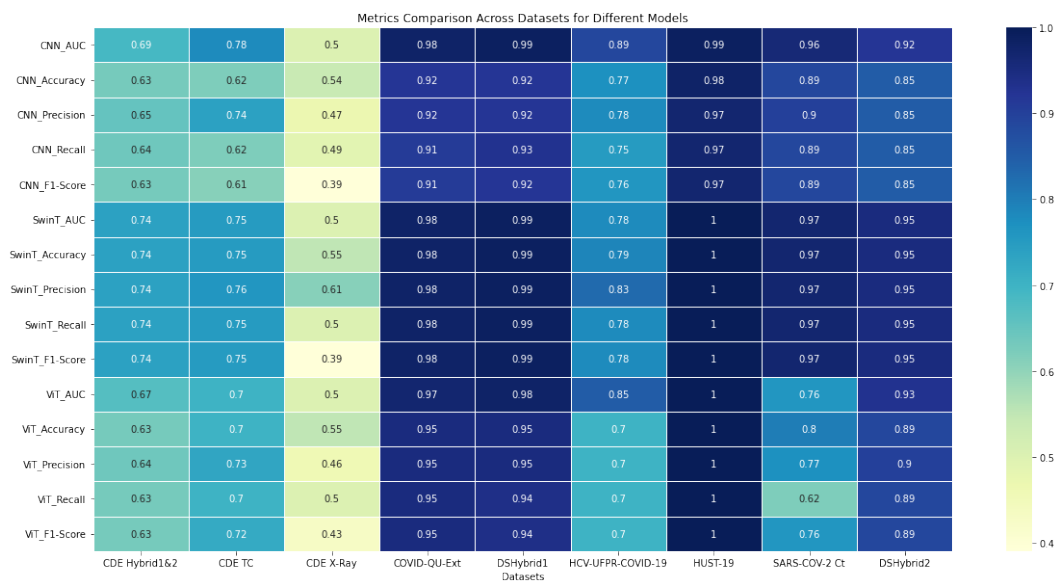


Figure 13. Comparison of Performance Metrics by Models. (Created by the author)

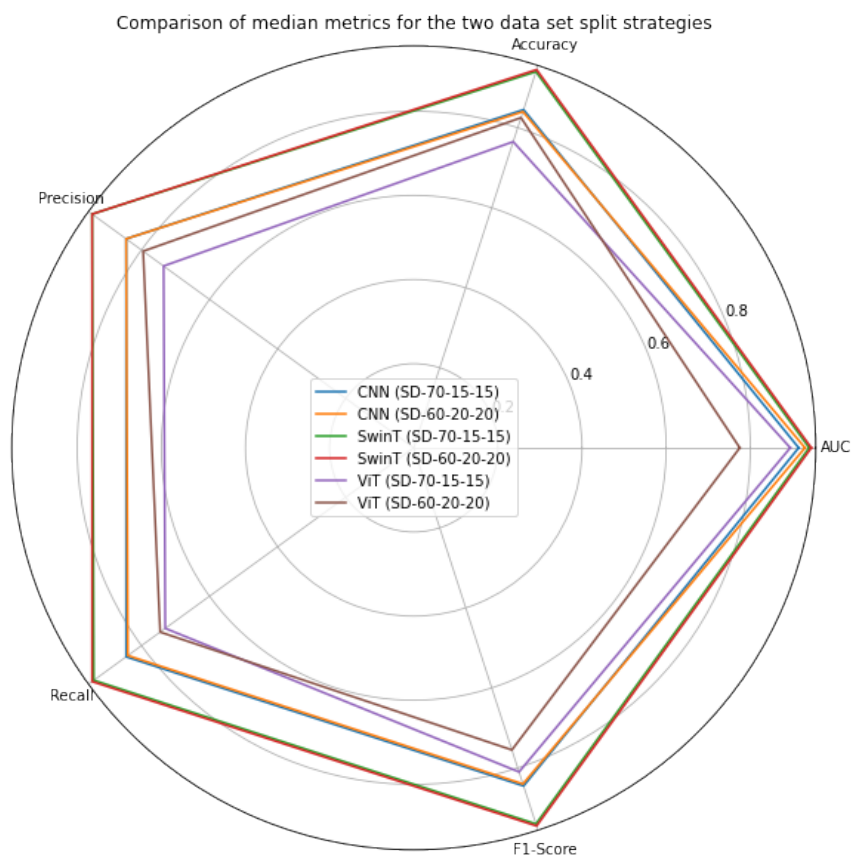


Figure 14. Medians of 5 performance metrics by strategy. (Created by the author)

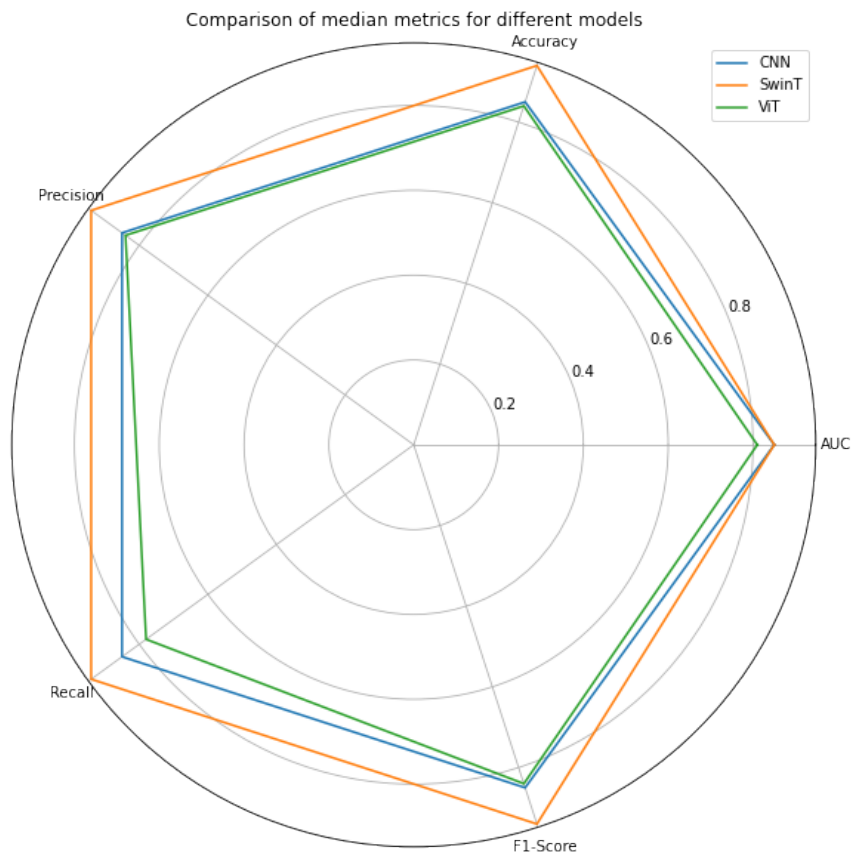


Figure 15. Medians of 5 Performance Metrics by 3 Models. (Created by the author)

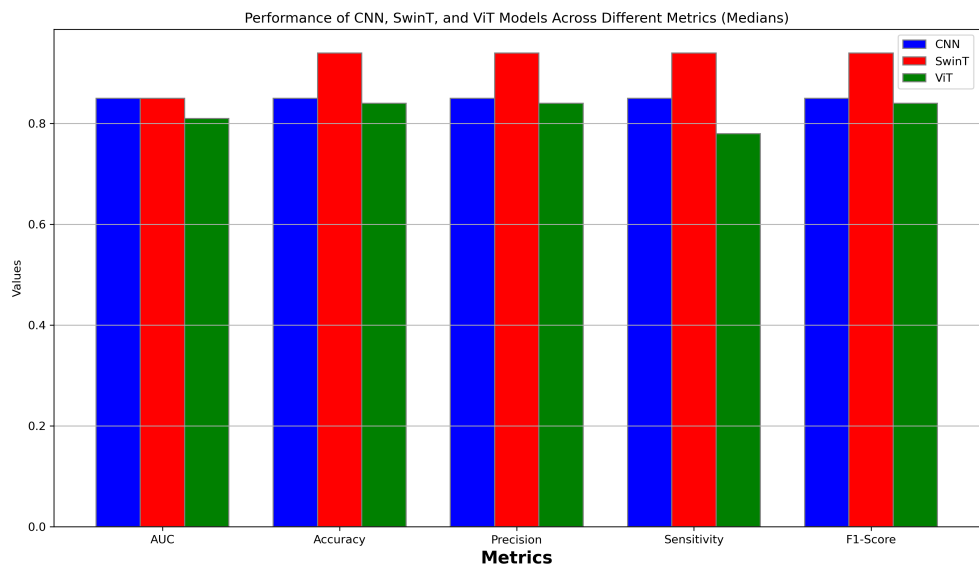


Figure 16. Performance Metrics vs Models. (Created by the author)

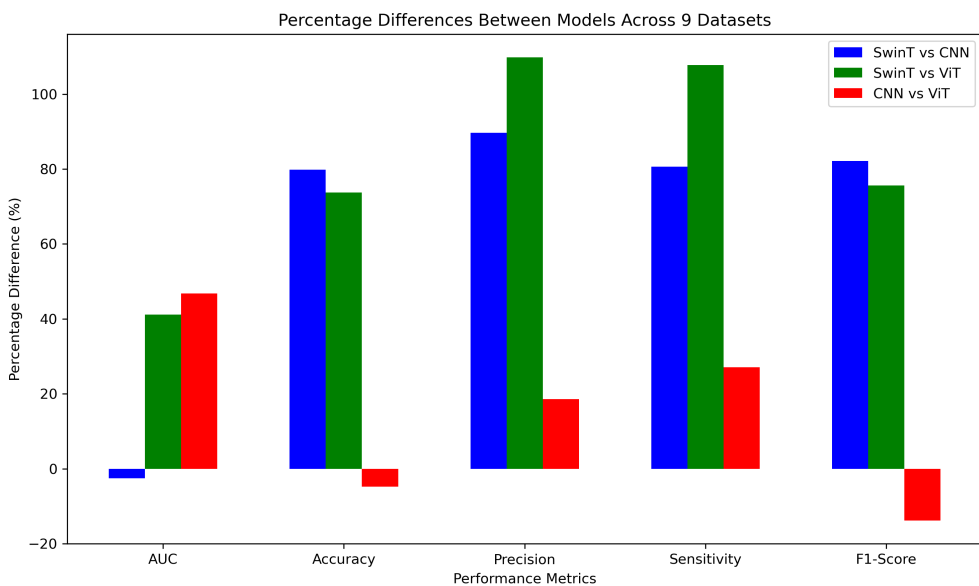


Figure 17. Percentage difference between models. (Created by the author)

Strategy	Variable Name	AUC	Accuracy	Precision	Sensitivity	F1-Score
CNN-DS701515-No_Optimizer	C	No	Yes	Yes	Yes	Yes
CNN-DS701515-Adam	D	Yes	Yes	Yes	Yes	Yes
CNN-DS701515-SGD	E	No	Yes	Yes	Yes	Yes
CNN-DS602020-Adam	F	No	Yes	Yes	Yes	Yes
CNN-DS602020-SGD	G	Yes	Yes	Yes	Yes	Yes
SwinT-DS701515-No_Optimizer	H	No	No	No	No	No
SwinT-DS701515-Adam	I	No	No	No	No	No
SwinT-DS701515-SGD	J	No	Yes	No	No	No
SwinT-DS602020-Adam	K	No	No	No	No	No
SwinT-DS602020-SGD	L	Yes	Yes	No	Yes	Yes
ViT-DS701515-No_Optimizer	M	No	No	Yes	No	Yes
ViT-DS701515-Adam	N	Yes	Yes	Yes	Yes	Yes
ViT-DS701515-SGD	O	No	Yes	No	Yes	Yes
ViT-DS602020-Adam	P	Yes	Yes	Yes	Yes	Yes
ViT-DS602020-SGD	Q	Yes	Yes	Yes	Yes	Yes

Figure 18. Shapiro-Wilk for performance metrics. (Created by the author)

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.787	0.803	0.781	0.788	0.772	0.847	0.853	0.862	0.870	0.847	0.789	0.772	0.838	0.782	0.841
Median	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.740	0.790	0.800	0.850
Standard deviation	0.160	0.157	0.165	0.180	0.160	0.169	0.164	0.152	0.155	0.164	0.176	0.158	0.153	0.163	0.151
Variance	0.026	0.024	0.027	0.032	0.026	0.028	0.027	0.023	0.024	0.027	0.031	0.025	0.024	0.027	0.023
Minimum	0.550	0.550	0.510	0.450	0.540	0.540	0.560	0.550	0.550	0.550	0.520	0.560	0.550	0.540	0.550
Maximum	0.960	0.990	0.980	0.990	0.970	0.990	1.000	1.000	1.000	1.000	0.950	1.000	1.000	1.000	1.000
25th percentile	0.650	0.650	0.620	0.630	0.610	0.730	0.730	0.750	0.750	0.730	0.610	0.630	0.770	0.640	0.730
50th percentile	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.740	0.790	0.800	0.850
75th percentile	0.920	0.930	0.890	0.920	0.900	0.980	0.980	0.980	0.980	0.970	0.940	0.930	0.980	0.910	0.970

Figure 19. Descriptive Statistics of Accuracy. (Created by the author)

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.853	0.864	0.851	0.859	0.833	0.843	0.847	0.858	0.864	0.846	0.793	0.804	0.846	0.771	0.831
Median	0.940	0.920	0.910	0.950	0.910	0.960	0.970	0.910	0.940	0.950	0.960	0.840	0.950	0.730	0.850
Standard deviation	0.165	0.156	0.169	0.177	0.175	0.177	0.176	0.166	0.168	0.173	0.258	0.162	0.174	0.168	0.165
Variance	0.027	0.024	0.029	0.031	0.031	0.031	0.031	0.027	0.028	0.030	0.067	0.026	0.030	0.028	0.027
Minimum	0.530	0.550	0.500	0.480	0.460	0.500	0.510	0.500	0.500	0.500	0.280	0.520	0.500	0.500	0.500
Maximum	0.990	1.000	0.990	1.000	0.990	0.990	1.000	1.000	1.000	1.000	0.990	1.000	1.000	1.000	1.000
25th percentile	0.750	0.780	0.780	0.780	0.770	0.730	0.730	0.750	0.750	0.750	0.670	0.700	0.740	0.660	0.720
50th percentile	0.940	0.920	0.910	0.950	0.910	0.960	0.970	0.910	0.940	0.950	0.960	0.840	0.950	0.730	0.850
75th percentile	0.990	0.980	0.960	0.980	0.960	0.980	0.980	0.980	0.980	0.970	0.990	0.930	0.980	0.910	0.970

Figure 20. Descriptive Statistics of AUC. (Created by the author)

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.764	0.777	0.757	0.787	0.773	0.830	0.832	0.840	0.862	0.846	0.773	0.779	0.800	0.770	0.826
Median	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.840	0.790	0.720	0.850
Standard deviation	0.198	0.198	0.200	0.190	0.162	0.206	0.209	0.208	0.174	0.173	0.201	0.197	0.211	0.167	0.166
Variance	0.039	0.039	0.040	0.036	0.026	0.042	0.044	0.043	0.030	0.030	0.040	0.039	0.044	0.028	0.027
Minimum	0.410	0.390	0.380	0.380	0.480	0.390	0.380	0.350	0.480	0.500	0.430	0.420	0.350	0.510	0.500
Maximum	0.960	0.990	0.970	0.990	0.970	0.990	1.000	1.000	1.000	1.000	0.950	1.000	1.000	1.000	1.000
25th percentile	0.650	0.640	0.590	0.690	0.640	0.730	0.730	0.750	0.750	0.750	0.610	0.630	0.720	0.660	0.730
50th percentile	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.840	0.790	0.720	0.850
75th percentile	0.920	0.930	0.890	0.910	0.900	0.980	0.980	0.980	0.980	0.970	0.940	0.930	0.980	0.910	0.970

Figure 21. Descriptive Statistics of F1-Score. (Created by the author)

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.804	0.816	0.790	0.796	0.784	0.839	0.876	0.834	0.879	0.877	0.798	0.771	0.804	0.780	0.829
Median	0.870	0.850	0.840	0.860	0.830	0.960	0.970	0.920	0.940	0.950	0.900	0.710	0.810	0.750	0.860
Standard deviation	0.147	0.152	0.176	0.190	0.158	0.189	0.132	0.232	0.138	0.120	0.178	0.157	0.228	0.166	0.180
Variance	0.022	0.023	0.031	0.036	0.025	0.036	0.017	0.054	0.019	0.014	0.032	0.025	0.052	0.028	0.032
Minimum	0.540	0.520	0.410	0.370	0.470	0.450	0.700	0.270	0.610	0.710	0.460	0.590	0.270	0.500	0.440
Maximum	0.960	0.990	0.970	0.990	0.970	0.990	1.000	1.000	1.000	1.000	0.950	1.000	1.000	1.000	1.000
25th percentile	0.700	0.750	0.740	0.740	0.720	0.730	0.730	0.760	0.760	0.770	0.700	0.640	0.760	0.670	0.730
50th percentile	0.870	0.850	0.840	0.860	0.830	0.960	0.970	0.920	0.940	0.950	0.900	0.710	0.810	0.750	0.860
75th percentile	0.920	0.930	0.890	0.920	0.900	0.980	0.980	0.980	0.980	0.970	0.940	0.930	0.980	0.910	0.970

Figure 22. Descriptive Statistics of Precision. (Created by the author)

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mean	0.787	0.792	0.777	0.786	0.766	0.847	0.847	0.858	0.864	0.846	0.789	0.756	0.810	0.753	0.818
Median	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.700	0.780	0.690	0.850
Standard deviation	0.160	0.166	0.172	0.186	0.168	0.169	0.176	0.166	0.168	0.173	0.176	0.172	0.179	0.180	0.170
Variance	0.026	0.028	0.030	0.035	0.028	0.028	0.031	0.027	0.028	0.030	0.031	0.030	0.032	0.032	0.029
Minimum	0.550	0.500	0.470	0.420	0.490	0.540	0.510	0.500	0.500	0.500	0.520	0.520	0.500	0.500	0.500
Maximum	0.960	0.990	0.970	0.990	0.960	0.990	1.000	1.000	1.000	1.000	0.950	1.000	1.000	1.000	1.000
25th percentile	0.650	0.650	0.620	0.640	0.610	0.730	0.730	0.750	0.750	0.750	0.610	0.630	0.720	0.630	0.710
50th percentile	0.870	0.850	0.840	0.860	0.820	0.960	0.970	0.910	0.940	0.950	0.890	0.700	0.780	0.690	0.850
75th percentile	0.920	0.930	0.890	0.910	0.900	0.980	0.980	0.980	0.980	0.970	0.940	0.930	0.980	0.910	0.970

Figure 23. Descriptive Statistics of Recall (Created by the author)

Table 1. Comparison of Performance Metrics Between Models in Different Papers.

Model	Author	Dataset	AUC	Acc	Prec	Recall	F1-Score
CNN	[Castiglione <i>et al.</i> , 2021]	SARS-COV-2 CT-Scan	0.98	0.9999	0.9992	0.9996	-
	[Awan <i>et al.</i> , 2021]	Coronavirus chest x-ray images and Chest X-Ray images (Pneumonia)	-	0.97	-	-	-
	[Banerjee <i>et al.</i> , 2022]	SARS-COV-2 CT Scan Dataset and Montgomery Dataset (CXRs)	-	0.9949	-	-	-
	[Hamza <i>et al.</i> , 2022]	COVID-GAN, COVID-Net small chest x-ray and CXR	-	0.994	-	-	-
	[Kathamuthu <i>et al.</i> , 2023]	Several	-	0.98	-	-	-
	[Nishio <i>et al.</i> , 2022]	COVID-BIMCV and COVID-PRIVATE	-	0.9912	-	-	-
	[Srivastava <i>et al.</i> , 2022]	Chest X-rays and CT Scan Dataset	-	0.9978	-	-	-
	[Abiyev and Ismail, 2021]	CXR images and COVID-19 Radiography database	-	0.983	0.983	0.979	0.98
	[Ahamed <i>et al.</i> , 2021]	COVID-19 Radiography Database	-	0.9901	-	-	-
	[Asif <i>et al.</i> , 2022]	COVID-19 radiography database	0.9998	0.9968	-	0.9966	-
	[Kibriya and Amin, 2023]	Chest X-ray	-	0.973	-	-	-
	[Lanjewar <i>et al.</i> , 2022]	COVID-19 Radiography Database and Chest X-Ray Images	-	0.9881	-	-	-
	[Yuan <i>et al.</i> , 2023]	Covid chest x-ray and Mixed Dataset	-	0.9799	0.9805	0.9802	0.9803
ViT	[Shome <i>et al.</i> , 2021]	Mixed Dataset	0.99	0.98	-	-	-
	[Cao <i>et al.</i> , 2022]	Guangzhou Women and Children Medical Centre dataset, MIDRC-RICORD and COVIDx CXR dataset	-	0.9709	0.9716	0.9693	0.9704
	[Chetoui and Akhloufi, 2022]	ChestX-ray8 dataset, NIH dataset and RSNA dataset	0.99	-	-	0.99	-
	[Park <i>et al.</i> , 2022a]	CheXpert, BIMCV, PADChest and Montgomery and Shenzen dataset	0.99	-	-	0.99	-
	[Park <i>et al.</i> , 2022b]	CheXpert, BIMCV, Brixia, NIH, AMC, CNUH, YNU, KNUH datasets	0.998	-	-	-	-
	[Murphy <i>et al.</i> , 2022]	Chest X-ray 14 dataset and MURA dataset	0.78	-	-	-	-
	[Jiang <i>et al.</i> , 2022]	MXT dataset and Chest X-ray14 dataset	0.83	-	-	-	-
	[Mehboob <i>et al.</i> , 2022]	HUST-19 dataset	-	-	0.997	-	-
	[Mondal <i>et al.</i> , 2022]	COVIDx CT-2A dataset and CHEST X-RAY DATASET	-	-	0.981	-	-
	[Konwer and Prasanna, 2022]	COVID-19 Radiography Database, SIIM-FISABIO-RSNA COVID-19 Detection	-	-	-	-	-
	[Wang <i>et al.</i> , 2023]	Mixed 7 datasets	-	-	0.9513	-	-
	[Chen <i>et al.</i> , 2023]	COVID-19 Radiography database	-	0.9827	0.9891	0.978	0.9913
	[Marefat <i>et al.</i> , 2023]	COVIDx CXR-3	-	0.9922	-	-	-

SwinT	[Dinh <i>et al.</i> , 2022]	COVID CXR, Chest X-ray, RICORD dataset and RALO dataset	-	-	0.99	-	-
	[Peng <i>et al.</i> , 2022]	COVID multiclass dataset of CT scans, SARS-COV-2 Ct-Scan Dataset and COVID-CT	0.9991	0.9894	0.9833	0.9895	0.9864
	[Tian <i>et al.</i> , 2022](Binary)	COVID multiclass dataset of CT scans, SARS-COV-2 Ct-Scan Dataset and COVID-CT	0.9985	0.9821	-	0.9907	0.9855
	[Tian <i>et al.</i> , 2022](multiclass)	COVID multiclass dataset of CT scans, SARS-COV-2 Ct-Scan Dataset and COVID-CT	-	0.9696	0.9668	-	0.9631
	[Ma and Lv, 2022] (Dataset 1)	NIH dataset and Chest X-Ray dataset	-	0.873	-	-	-
	[Ma and Lv, 2022] (Dataset 2)	NIH dataset and Chest X-Ray dataset	-	0.972	-	-	-
	[Pan <i>et al.</i> , 2023]	NIH Chest x-rays dataset, MRI dataset and CT dataset	-	0.93	-	-	-

Table 2. Datasets used in the study

#	Source	Dataset	Type	Images	COVID-19	Normal
1	[Tahir <i>et al.</i> , 2021]	COVID-QU-Ex	X-ray	22,657	11,956	10,701
2	[Luz <i>et al.</i> , 2021]	HCV-UFPR-COVID-19	X-ray	513	281	232
3	[Mehboob <i>et al.</i> , 2022; Soares and Angelov, 2020]	SARS-COV2 CT	CT	2,481	1,252	1,229
4	Ning <i>et al.</i> [2020]	HUST-19	CT	13,980	4,001	9,979
5	The Author	DSHybrid1 (COVID-QU-Ex + HUST-19)	X-ray + CT	36,637	15,957	20,680
6	The Author	DSHybrid2 (HCV-UFPR + SARS-COV2 CT)	X-ray + CT	2,994	1,533	1,461
7	The Author	CDE Hybrid1&2	X-ray + CT	39,631	17,490	22,141
8	The Author	CDE Raio X (COVID-QU-Ex& HCV-UFPR)	X-ray	23,170	12,237	10,933
9	The Author	CDE CT (HUST-19& SARS-COV2 CT)	TC	16,461	5,253	11,208

Table 3. CDE X-Ray Dataset Split COVID-19 Class

Dataset Name	Training (70%)	Validation (30%)	Test (100%)
COVID-QU-Ex Dataset	8,369	3,587	0
HCV-UFPR-COVID-19	0	0	281

Table 4. CDE X-Ray Dataset Split Normal Class

Dataset Name	Training (70%)	Validation (30%)	Test (100%)
COVID-QU-Ex Dataset	7,491	3,210	0
HCV-UFPR-COVID-19	0	0	232

Table 5. CDE CT Dataset Split COVID-19 Class

Dataset Name	Training (70%)	Validation (30%)	Test (100%)
HUST-19	2,801	1,200	0
SARS-COV2 CT dataset	0	0	1,252

Table 6. CDE CT Dataset Split Normal Class

Dataset Name	Training (70%)	Validation (30%)	Test (100%)
HUST-19	6,985	2,994	0
SARS-COV2 CT dataset	0	0	1,229

Table 7. CDE Hybrid1&2 Dataset Split COVID-19 Class

Dataset Name	Training (70%)	Validation (30%)	Test (100%)
DSHybrid1	11,169	4,788	0
DSHybrid2	0	0	1,533

Table 8. CDE Hybrid1&2 Dataset Split Normal Class

Dataset Name	Training (70%)	Validation (30%)	Test (100%)
DSHybrid1	14,476	6,204	0
DSHybrid2	0	0	1,461

Table 9. Hyperparameters tuning by Optuna library

Optimizer	Learning Rate	Weight Decay	Batch Size	Mom.	Beta 1	Beta 2	Epsilon	Epochs
ViT								
SGD	0.000847	0.000248	8	0.866	N/A	N/A	N/A	25
AdamW	0.000015	0.000029	16	N/A	0.851	0.995	0.0	25
SwinT								
SGD	0.000847	0.000248	8	0.866	N/A	N/A	N/A	25
AdamW	0.000015	0.000029	16	N/A	0.851	0.995	0.0	25
SwinS								
SGD	0.000847	0.000248	8	0.866	N/A	N/A	N/A	25
CNN								
SGD	0.000767	0.000004	32	0.173	N/A	N/A	N/A	25
Adam	0.002931	N/A	N/A	N/A	N/A	N/A	N/A	25

Table 10. Performance Metrics for Models DS-70-15-15-Dataset-Adam

Dataset	Model	AUC	Accuracy	Precision	Recall	F1-Score
CDE Hybrid1&2	CNN	0.69	0.63	0.65	0.64	0.63
CDE Hybrid1&2	SwinT	0.74	0.74	0.75	0.74	0.74
CDE Hybrid1&2	ViT	0.63	0.63	0.63	0.63	0.63
CDE TC	CNN	0.78	0.62	0.74	0.62	0.69
CDE TC	SwinT	0.75	0.75	0.76	0.75	0.75
CDE TC	ViT	0.69	0.69	0.73	0.69	0.72
CDE X-Ray	CNN	0.48	0.45	0.37	0.42	0.38
CDE X-Ray	SwinT	0.5	0.55	0.61	0.5	0.48
CDE X-Ray	ViT	0.5	0.54	0.5	0.5	0.51
COVID-QU-Ex	CNN	0.98	0.92	0.92	0.91	0.91
COVID-QU-Ex	SwinT	0.98	0.98	0.98	0.98	0.98
COVID-QU-Ex	ViT	0.91	0.91	0.91	0.91	0.91
DSHybrid1	CNN	0.99	0.92	0.92	0.93	0.92
DSHybrid1	SwinT	0.99	0.99	0.99	0.99	0.99
DSHybrid1	ViT	0.94	0.95	0.95	0.94	0.94
DSHybrid2	CNN	0.95	0.86	0.86	0.86	0.86
DSHybrid2	SwinT	0.94	0.94	0.94	0.94	0.94
DSHybrid2	ViT	0.88	0.88	0.88	0.88	0.88
HCV-UFPR-COVID-19	CNN	0.9	0.81	0.81	0.81	0.81
HCV-UFPR-COVID-19	SwinT	0.94	0.94	0.94	0.94	0.94
HCV-UFPR-COVID-19	ViT	0.66	0.64	0.67	0.64	0.66
HUST-19	CNN	1.00	0.99	0.99	0.99	0.99
HUST-19	SwinT	1.00	1.00	1.00	1.00	1.00
HUST-19	ViT	1.00	1.00	1.00	1.00	1.00
SARS-COV-2 Ct	CNN	0.96	0.89	0.9	0.89	0.89
SARS-COV-2 Ct	SwinT	0.94	0.94	0.94	0.94	0.94
SARS-COV-2 Ct	ViT	0.73	0.8	0.75	0.59	0.68

Table 11. Performance Metrics for Models DS-70-15-15-Dataset-SGD

Dataset	Model	AUC	Accuracy	Precision	Recall	F1-Score
CDE Hybrid1&2	CNN	0.67	0.61	0.63	0.61	0.62
CDE Hybrid1&2	SwinT	0.71	0.71	0.71	0.71	0.71
CDE Hybrid1&2	ViT	0.71	0.71	0.71	0.71	0.71
CDE TC	CNN	0.77	0.57	0.72	0.58	0.64
CDE TC	SwinT	0.75	0.75	0.76	0.75	0.75
CDE TC	ViT	0.72	0.73	0.73	0.72	0.73
CDE X-Ray	CNN	0.46	0.54	0.47	0.49	0.48
CDE X-Ray	SwinT	0.5	0.55	0.77	0.5	0.5
CDE X-Ray	ViT	0.5	0.55	0.44	0.5	0.5
COVID-QU-Ex	CNN	0.96	0.9	0.9	0.9	0.9
COVID-QU-Ex	SwinT	0.97	0.97	0.97	0.97	0.97
COVID-QU-Ex	ViT	0.97	0.97	0.97	0.97	0.97
DSHybrid1	CNN	0.97	0.91	0.91	0.91	0.91
DSHybrid1	SwinT	0.99	0.99	0.99	0.99	0.99
DSHybrid1	ViT	0.98	0.98	0.98	0.98	0.98
DSHybrid2	CNN	0.91	0.82	0.83	0.82	0.82
DSHybrid2	SwinT	0.95	0.95	0.95	0.95	0.95
DSHybrid2	ViT	0.93	0.93	0.93	0.93	0.93
HCV-UFPR-COVID-19	CNN	0.84	0.77	0.76	0.75	0.76
HCV-UFPR-COVID-19	SwinT	0.77	0.73	0.77	0.77	0.77
HCV-UFPR-COVID-19	ViT	0.85	0.85	0.86	0.85	0.85
HUST-19	CNN	0.99	0.97	0.97	0.96	0.97
HUST-19	SwinT	1.00	1.00	1.00	1.00	1.00
HUST-19	ViT	1.00	1.00	1.00	1.00	1.00
SARS-COV-2 Ct	CNN	0.93	0.86	0.87	0.87	0.86
SARS-COV-2 Ct	SwinT	0.97	0.97	0.97	0.97	0.97
SARS-COV-2 Ct	ViT	0.82	0.85	0.84	0.7	0.76

Table 12. Performance Metrics for Models DS-60-20-20-Dataset-Adam

Dataset	Model	AUC	Accuracy	Precision	Recall	F1-Score
CDE Hybrid1&2	CNN	0.67	0.61	0.63	0.61	0.62
CDE Hybrid1&2	SwinT	0.71	0.71	0.71	0.71	0.71
CDE Hybrid1&2	ViT	0.71	0.71	0.71	0.71	0.71
CDE TC	CNN	0.77	0.57	0.72	0.58	0.64
CDE TC	SwinT	0.75	0.75	0.76	0.75	0.75
CDE TC	ViT	0.72	0.73	0.73	0.72	0.73
CDE X-Ray	CNN	0.46	0.54	0.47	0.49	0.48
CDE X-Ray	SwinT	0.5	0.55	0.77	0.5	0.5
CDE X-Ray	ViT	0.5	0.55	0.44	0.5	0.5
COVID-QU-Ex	CNN	0.96	0.9	0.9	0.9	0.9
COVID-QU-Ex	SwinT	0.97	0.97	0.97	0.97	0.97
COVID-QU-Ex	ViT	0.97	0.97	0.97	0.97	0.97
DSHybrid1	CNN	0.97	0.91	0.91	0.91	0.91
DSHybrid1	SwinT	0.99	0.99	0.99	0.99	0.99
DSHybrid1	ViT	0.98	0.98	0.98	0.98	0.98
DSHybrid2	CNN	0.91	0.82	0.83	0.82	0.82
DSHybrid2	SwinT	0.95	0.95	0.95	0.95	0.95
DSHybrid2	ViT	0.93	0.93	0.93	0.93	0.93
HCV-UFPR-COVID-19	CNN	0.84	0.77	0.76	0.75	0.76
HCV-UFPR-COVID-19	SwinT	0.77	0.73	0.77	0.77	0.77
HCV-UFPR-COVID-19	ViT	0.85	0.85	0.86	0.85	0.85
HUST-19	CNN	0.99	0.97	0.97	0.96	0.97
HUST-19	SwinT	1.00	1.00	1.00	1.00	1.00
HUST-19	ViT	1.00	1.00	1.00	1.00	1.00
SARS-COV-2 Ct	CNN	0.93	0.86	0.87	0.87	0.86
SARS-COV-2 Ct	SwinT	0.97	0.97	0.97	0.97	0.97
SARS-COV-2 Ct	ViT	0.82	0.85	0.84	0.7	0.76

Table 13. Performance Metrics for Models DS-60-20-20-Dataset-SGD

Dataset	Model	AUC	Accuracy	Precision	Recall	F1-Score
CDE Hybrid1&2	CNN	0.67	0.61	0.63	0.61	0.62
CDE Hybrid1&2	SwinT	0.71	0.71	0.71	0.71	0.71
CDE Hybrid1&2	ViT	0.71	0.71	0.71	0.71	0.71
CDE TC	CNN	0.77	0.57	0.72	0.58	0.64
CDE TC	SwinT	0.75	0.75	0.76	0.75	0.75
CDE TC	ViT	0.72	0.73	0.73	0.72	0.73
CDE X-Ray	CNN	0.46	0.54	0.47	0.49	0.48
CDE X-Ray	SwinT	0.5	0.55	0.77	0.5	0.5
CDE X-Ray	ViT	0.5	0.55	0.44	0.5	0.5
COVID-QU-Ex	CNN	0.96	0.9	0.9	0.9	0.9
COVID-QU-Ex	SwinT	0.97	0.97	0.97	0.97	0.97
COVID-QU-Ex	ViT	0.97	0.97	0.97	0.97	0.97
DSHybrid1	CNN	0.97	0.91	0.91	0.91	0.91
DSHybrid1	SwinT	0.99	0.99	0.99	0.99	0.99
DSHybrid1	ViT	0.98	0.98	0.98	0.98	0.98
DSHybrid2	CNN	0.91	0.82	0.83	0.82	0.82
DSHybrid2	SwinT	0.95	0.95	0.95	0.95	0.95
DSHybrid2	ViT	0.93	0.93	0.93	0.93	0.93
HCV-UFPR-COVID-19	CNN	0.84	0.77	0.76	0.75	0.76
HCV-UFPR-COVID-19	SwinT	0.77	0.73	0.77	0.77	0.77
HCV-UFPR-COVID-19	ViT	0.85	0.85	0.86	0.85	0.85
HUST-19	CNN	0.99	0.97	0.97	0.96	0.97
HUST-19	SwinT	1.00	1.00	1.00	1.00	1.00
HUST-19	ViT	1.00	1.00	1.00	1.00	1.00
SARS-COV-2 Ct	CNN	0.93	0.86	0.87	0.87	0.86
SARS-COV-2 Ct	SwinT	0.97	0.97	0.97	0.97	0.97
SARS-COV-2 Ct	ViT	0.82	0.85	0.84	0.7	0.76

Table 14. Friedman Test for Model Performance Metrics

Metric	χ^2	df	p
AUC	32.8	14	0.003
Accuracy	71.8	14	< 0.001
Precision	57.4	14	< 0.001
Sensitivity	67.4	14	< 0.001
F1-Score	58.4	14	< 0.001
Medians (Models)	9.58	2	0.008

Table 15. Comparative Analysis with Benchmark Models. The symbol “-” indicates that the corresponding metric is not available in the referenced study. “Our implementation” refers to the results obtained using the models developed and evaluated in this work.

Model	Dataset	Author	AUC	Accuracy	Precision	Recall	F1-Score
CNN	SARS-COV-2 Ct	[Castiglione <i>et al.</i> , 2021]	0.98	0.99	0.99	0.99	-
	SARS-COV-2 Ct	Our implementation	0.98	0.89	0.90	0.89	0.89
ViT	HUST-19	[Mehboob <i>et al.</i> , 2022]	-	-	0.99	-	-
	HUST-19	Our implementation	1.00	1.00	1.00	1.00	1.00
SwinT	SARS-COV-2 Ct	[Peng <i>et al.</i> , 2022]	0.99	0.98	0.98	0.98	0.98
	SARS-COV-2 Ct	[Tian <i>et al.</i> , 2022] (Binary)	0.99	0.98	-	0.99	0.98
	SARS-COV-2 Ct	Our implementation	0.98	0.98	0.98	0.98	0.98