KSÜ Mühendislik Bilimleri Dergisi, 27(3), 2024
Araştırma Makalesi

KSU J Eng Sci, 27(3), 2024
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 24.11.2023
Kabul Tarihi :30.11.2023

Received Date :24.11.2023
Accepted Date : 30.11.2023

TRANSFORMER BASED COVID-19 DETECTION USING CHEST X-RAYS

AKCİĞER GÖRÜNTÜLERİNDEN TRANSFORMER TABANLI COVID-19 TESPİTİ

Hacı Ömer DOKUMACI (ORCID: 0000-0003-4032-0669)

Kahramanmaraş Sütçü İmam Üniversitesi, Elektrik Elektronik Mühendisliği Bölümü, Kahramanmaraş, Türkiye

*Sorumlu Yazar / Corresponding Author: Hacı Ömer DOKUMACI, omer_dokumacı@ksu.edu.tr

ABSTRACT

Covid-19 has affected millions globally, leading to substantial illness and mortality. Chest X-rays serve as a rapid and effective means of tracking the progression of Covid-19. However, diagnosing Covid-19 from a chest X-ray can be complex, and even skilled radiologists may not always provide a conclusive diagnosis. In our research, we utilized a dataset comprising X-ray images of Covid-19, lung opacity, viral pneumonia, and healthy patients to assess the efficacy of various vision transformer-based models. A modified version of the Swin Transformer achieved an accuracy of 98.9% and a precision of 99.2% on Covid-19 images in a four-way classification task. Our findings are competitive with cutting-edge techniques for diagnosing Covid-19. This method could aid healthcare professionals in screening patients for Covid-19, thereby enabling quicker treatment and improved health outcomes for those affected by the virus.

Keywords: Covid-19, deep learning, transformer.

ÖZET

Covid-19, küresel olarak milyonlarca kişiyi etkileyerek önemli hastalıklara ve ölümlere yol açmıştır. Akciğer röntgenleri, Covid-19'un ilerlemesini izlemek için hızlı ve etkili bir yöntem olarak hizmet vermektedir. Ancak, bir akciğer röntgeninden Covid-19'u teşhis etmek karmaşık olabilir ve hatta deneyimli radyologlar bile her zaman kesin bir teşhis koyamayabilir. Bu araştırmada, Covid-19, akciğer opasitesi, viral pnömoni ve sağlıklı hastaların X-ray görüntülerinden oluşan bir veri setini kullanarak çeşitli vision transformer tabanlı modellerin etkinliği değerlendirildi. Swin Transformer'ın modifiye edilmiş bir versiyonu, Covid-19 görüntülerinde dört yönlü sınıflandırmada %98.9 doğruluk ve %99.2 hassasiyet elde etti. Bulgularımız, Covid-19 teşhisi için, son teknoloji tekniklerle rekabet edebilecek düzeydedir. Bu yöntem, sağlık profesyonellerinin Covid-19 için hastaları taramasına yardımcı olabilir, böylece daha hızlı tedavi sağlanabilir ve Covid-19 hastaları için daha iyi sağlık sonuçları elde edilebilir.

Anahtar Kelimeler: Covid-19, derin öğrenme, transformer.

ToCite: DOKUMACI, H. Ö., (2024). TRANSFORMER BASED COVID-19 DETECTION USING CHEST X-RAYS. *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, 27(3), 679-687.

INTRODUCTION

COVID-19, an infectious disease caused by the SARS-CoV-2 virus, has had a profound impact on the world (Wang et al., 2020; Gorbalenya et al., 2020; Phelan et al., 2020). Most of those infected will experience mild to moderate respiratory symptoms and recover without needing special care. However, some individuals will become severely ill and require medical intervention. The virus primarily spreads through droplets expelled from the mouth or nose of an infected person. It can also be transmitted by touching surfaces or objects contaminated with the virus. The pandemic has caused widespread sickness, death, and economic disruption. In response, governments and organizations globally have taken steps to control the virus's spread, such as implementing lockdowns, travel restrictions, and social distancing guidelines. The crisis has also spurred extensive research into the virus and its impact on the human body.

The RT-PCR (Reverse Transcription Polymerase Chain Reaction) test is a type of molecular diagnostic test that identifies the genetic material of the SARS-CoV-2 virus, the causative agent of COVID-19 (Kucirka et al., 2020; Khan et al., 2020). It's frequently used to ascertain if a person is presently infected with the virus. The RT-PCR test is highly valued for its precision. When conducted correctly by a healthcare professional, it provides very accurate results and is considered the benchmark for identifying the presence of the infectious virus. Another benefit of the RT-PCR test is its ability to detect the virus in asymptomatic individuals. This is vital for pinpointing and isolating infected individuals to curb the virus's spread. Despite its advantages, the RT-PCR test has some shortcomings. A significant drawback is the extended time it takes to get results. While results can sometimes be obtained within a day or two, there have been instances during the pandemic where it took up to one to two weeks. Another concern with RT-PCR testing is the potential for a high number of false negatives, which can occur if the viral load is low, causing the test to fail to detect the virus.

Considering the limitations of RT-PCR testing, it's essential to consider other diagnostic methods. Medical imaging techniques such as computed tomography (CT) scans and chest X-rays (CXR) can be effective in diagnosing COVID-19. These methods are particularly useful when a patient's RT-PCR test is negative, but there is still a suspicion of infection. CT scans have proven to be more precise than CXR in detecting COVID-19, but they come with their own set of challenges, including higher costs, longer scanning times, increased radiation exposure, and the inability to perform at the patient's bedside. On the other hand, CXR scans are less costly, quicker, and expose patients to less radiation compared to CT scans. However, diagnosing COVID-19 from a CXR scan can be difficult, and even skilled radiologists may not always be able to provide a definitive diagnosis.

Deep learning (DL) has had a significant impact on the diagnosis of COVID-19 using chest X-rays. DL methods have been widely applied to analyze radiographic images such as Chest X-Rays (CXR) and CT scan images (Apostolopoulos et al., 2020; Wang et al., 2020; Umer et al., 2022). The use of deep learning to analyze X-rays could greatly reduce the time taken to diagnose patients. An AI model can process hundreds of images in the average time taken for a radiologist to analyze one. Furthermore, deep learning technology is a practical and affordable modality that can be deemed a reliable technique for adequately diagnosing the COVID-19 disease.

PREVIOUS WORK

Vision transformers (ViTs) are a recent type of neural network architecture that has been shown to achieve state-of-the-art results on a variety of computer vision tasks (Vaswani et al., 2017; Dosovitskiy et al., 2020; Wu et al., 2020). ViTs are inspired by the transformer architecture which was originally developed for natural language processing (NLP). One of the key differences between ViTs and convolutional neural networks (CNNs) is that ViTs use self-attention, which allows them to attend to any part of the input image regardless of its spatial location. This makes ViTs more powerful than CNNs at learning long-range dependencies in images. ViTs have shown promising results on a variety of computer vision tasks, including image classification, object detection, and semantic segmentation.

Rahhal et al. utilized a model grounded on the Data-Efficient Image Transformer (DeiT) architecture, an enhanced version of the Vision Transformer (ViT). The model subdivided both original and augmented images into non-overlapping patches, which were then processed through the embedding layer and a Siamese encoder, achieving an accuracy of 94.62% on CXR data. Chetoui et al. conducted a comparative study between several ViT models and CNN models for multi-class classification problems. Their findings revealed that ViT-B32 outperformed CNN architectures in detecting Covid-19 on CXR images and was effective in pinpointing the most significant pathology regions, with an accuracy rate of 96%.

Yang et al. adapted the vision transformer to leverage all outputs from the encoder, yielding superior results than CNNs for Covid-19 diagnosis, with an accuracy of 98.2%. Okolo et al. introduced an architecture based on vision transformer named Input Enhanced ViT (IEViT). Inspired by ResNet skip connections, this approach iteratively added a representation of the original input image to the output of each transformer encoder layer, proving to be more effective than standard ViTs. These studies collectively suggest that transformers hold an edge over CNNs in evaluating Covid-19 Chest X-rays.

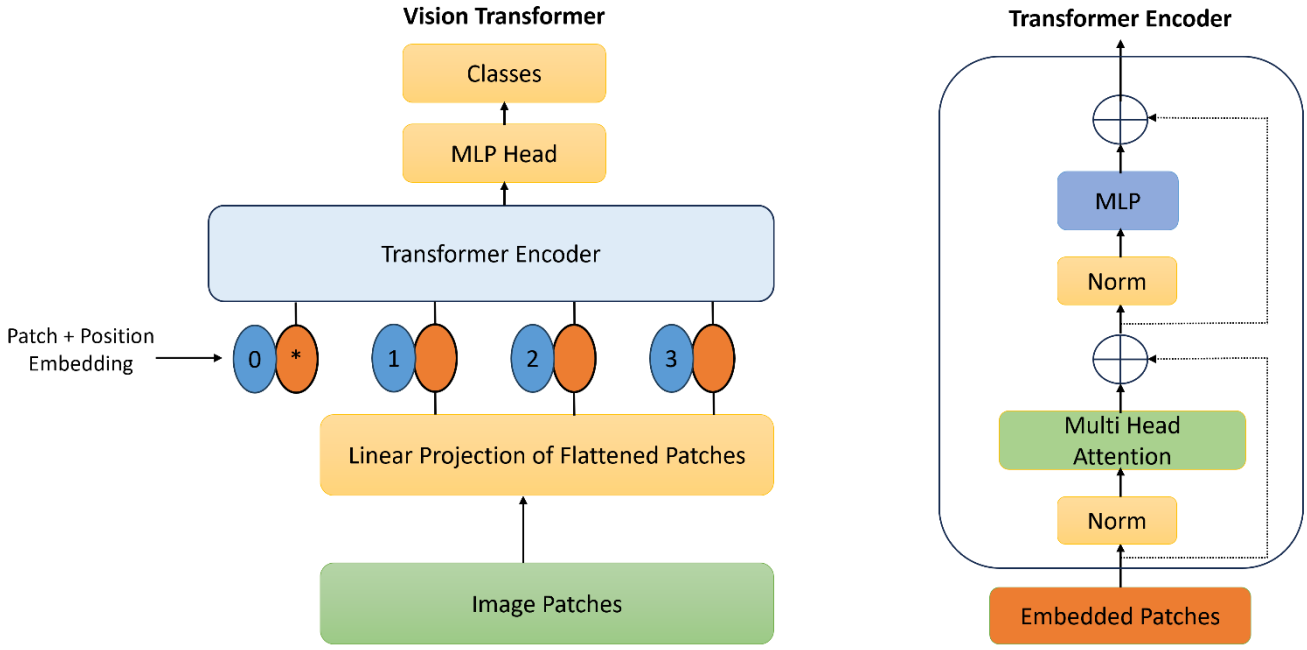

Figure 1. Structure of the Vision Transformer

MATERIALS AND METHODS

In this study, the performance of the latest vision transformer based neural net architectures are evaluated and compared for the detection of Covid-19 on one of the most comprehensive Covid-19 public datasets. This work provides a baseline for these models without modification of the underlying network so that a fair assessment can be made on the vanilla models. Only the classification head has been changed for all networks in this study. In a significant part of the literature, parts of the networks have been changed, but these changes and corresponding results are hard to replicate since accompanying source code or hyperparameter values are rarely given. Since the performances of standard open-source networks are evaluated in this study, the results can be replicated by experts in the field. The networks that are benchmarked in this study are: ViT, MaxViT, and Swin Transformer.

Transformer Based Models

Vision Transformer (ViT) is a model that applies the transformer to the image classification task (Dosovitskiy et al., 2020). It emerged as a competitive alternative to CNNs that are currently state-of-the-art in computer vision and widely used for different image recognition tasks. The structure of ViT can be seen in Fig. 1. ViT is a structure that employs self-attention mechanisms for image processing. It is composed of a sequence of transformer blocks. Each block has two layers: a multi-head self-attention layer and a feed-forward layer. The self-attention layer computes attention weights for each pixel in the image. The feed-forward layer applies a non-linear transformation to the self-attention layer's output. The multi-head attention allows the model to focus on different parts of the input sequence at the same time. Additionally, ViT includes a patch embedding layer that splits the image into fixed-size patches and converts each patch into a high-dimensional vector representation. These patch embeddings are then processed further in the transformer blocks. The output of the ViT structure is a class prediction, which is achieved by passing the last transformer block's output through a classification head, typically a single fully connected layer.

MaxViT is a novel vision transformer module (Tu et al., 2020). This paper introduces a multi-axis attention model that is both efficient and scalable. It comprises two components: blocked local attention and dilated global attention. These components enable interactions between global and local spatial aspects on any input resolution, maintaining only linear complexity. Furthermore, the authors introduce a new architectural component that integrates their proposed attention model with convolutions. This leads to the proposal of a straightforward hierarchical vision backbone, achieved by repeating the basic building block across multiple stages.

The Swin Transformer is a vision transformer that functions as a universal backbone for computer vision (Liu et al., 2021). It was developed to overcome the difficulties of applying transformers from language to vision, such as the large-scale variations in visual objects and the high resolution of image pixels compared to text words. To tackle these large scales, the Swin Transformer introduces a hierarchical structure that calculates representations with shifted windows. This shifted windowing approach enhances efficiency by confining self-attention computation to non-overlapping local windows and enables cross-window connection. The hierarchical structure provides the flexibility to model at different scales and maintains linear computational complexity relative to image size. These characteristics make the Swin Transformer suitable for a wide variety of vision tasks, including image classification, object detection, and semantic segmentation.

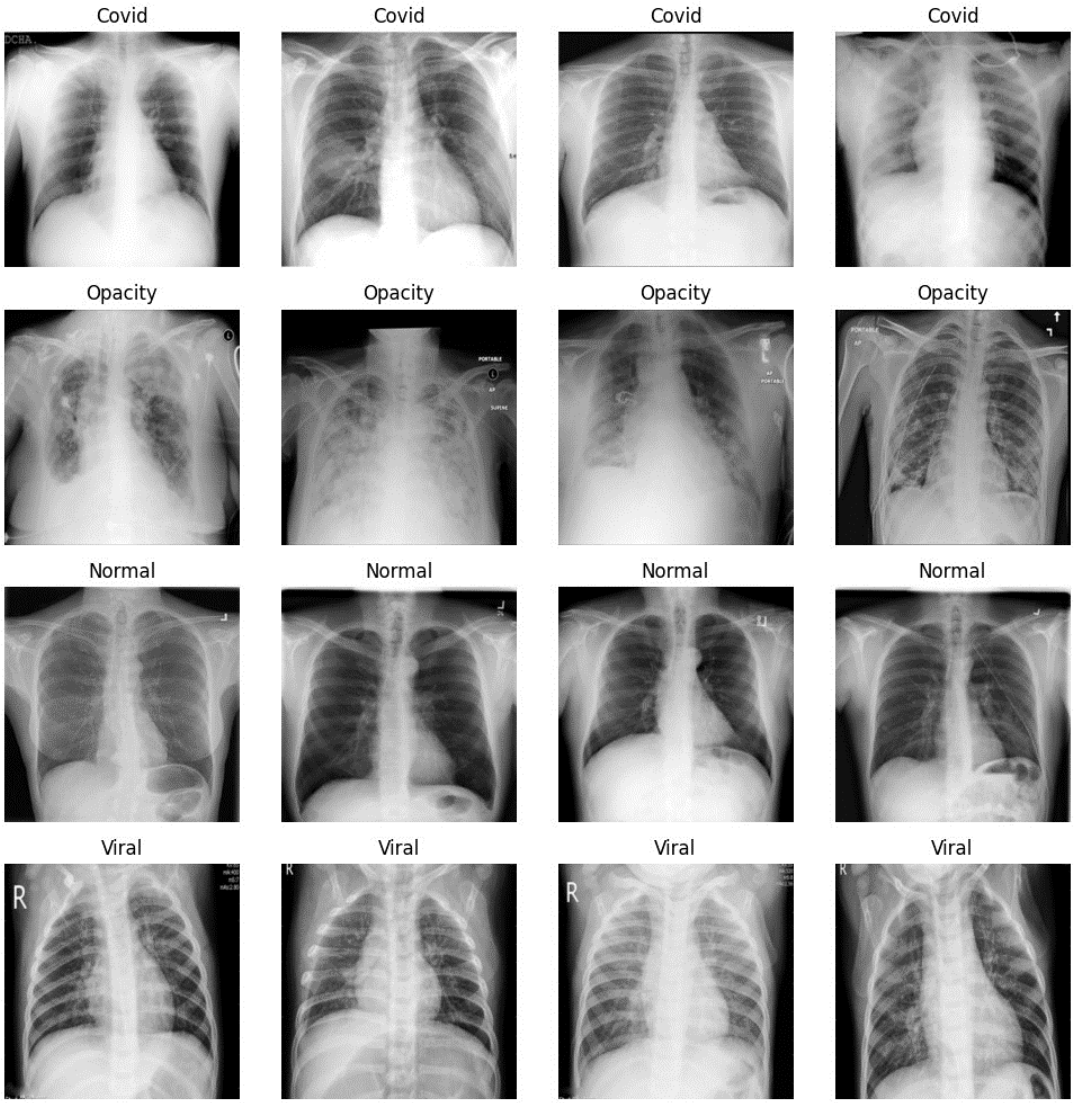

Figure 2. Sample Images in the Dataset with Their Labels

Dataset

The Covid-19 dataset utilized in this study is composed of Chest X-Ray (CXR) images representing four conditions: Normal, Covid-19, Lung Opacity, and Viral Pneumonia (Chowdhury et al., 2020; Rahman et al., 2021). Lung Opacity refers to non-Covid lung infections. The dataset includes 3616 Covid-19 positive cases, 10,192 Normal cases, 6012 Lung Opacity cases, and 1345 Viral Pneumonia cases, along with corresponding lung masks, totaling 21,165 CXR images. All images have been standardized to a resolution of 299×299. These images were sourced from various medical institutions. Representative images along with their labels are displayed in Fig.2.The dataset was randomly divided into three categories: training, validation, and test. The test dataset consisted of 10% of the dataset while 80% of the rest was assigned to the training dataset and 20% of the rest was held out as the validation dataset.

Data augmentation is a strategy that significantly increases the variety of data available for training models without the need to collect more data. Widely used data augmentation techniques, such as rotation and horizontal flipping, are often employed when training large neural networks. These techniques can be applied individually or in combination. The goal is to create a more robust model by including a wider range of samples in the training data. In this study, the images underwent random rotation and horizontal flipping during the training phase. The rotation angle was limited to a maximum of 10 degrees. Before augmentation, the images were also scaled to match the input resolution of the networks.

Transfer Learning

Transfer learning is a technique in machine learning where a model, initially developed for one task, is repurposed as the foundation for a model on a different task. This approach is widely used in deep learning where models that have been pre-trained are employed as the starting point for tasks related to computer vision and natural language processing because of the extensive computational and temporal resources needed to develop neural network models from scratch for these problems.

In this study, the output layers of the pre-trained models are removed, and a new classification head, corresponding to the number of classes in this study, is added. The model is then retrained on the dataset, leveraging the pre-trained weights. The classification head comprises a fully connected (FC) layer with an output size of 512, a dropout layer, and a final classification layer for the four classes. The model training was conducted with an initial learning rate of 1e-5, a batch size of 32 images, and over 25 epochs. The optimizer used was AdamW, with a step learning rate decay of 0.1 for every 10 epochs.

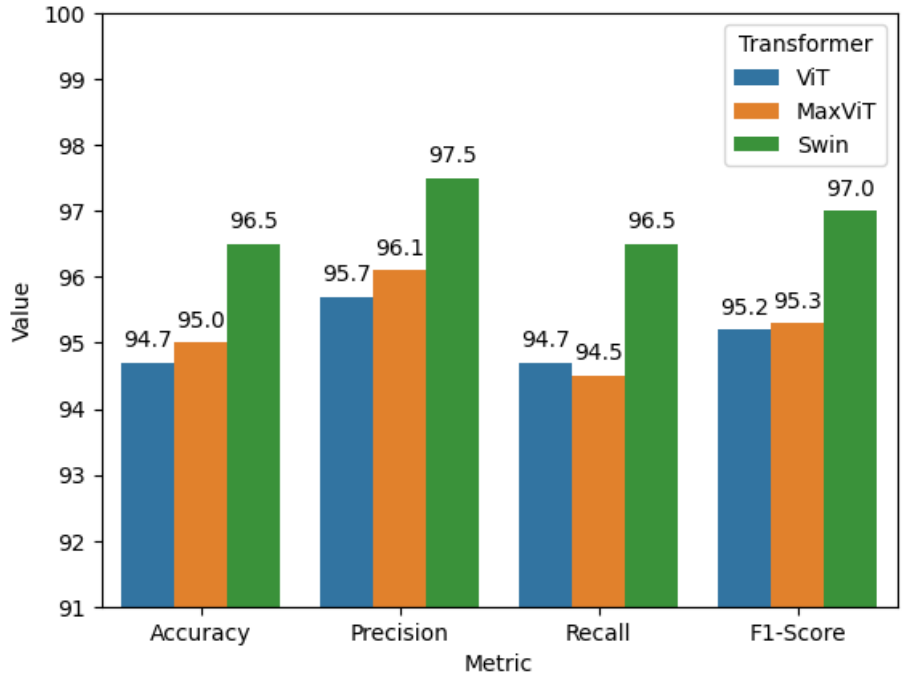

Figure 3. Results for Macro Average Metrics (%)

RESULTS AND DISCUSSION

In this study, four indicators are used to evaluate the performance of the models. Precision can be represented by Eq. 1.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Recall is defined by Eq. 2.

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

Accuracy is represented by Eq. 3.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

F_1 score is the harmonic average between precision and recall. If we need to find a balance between precision and recall, or if there is an uneven class distribution, then F_1 score is important (Eq. 4).

$$F_1 = \frac{2\,.\,Recall\,.\,Precision}{Recall+Precision} \tag{4}$$

The results obtained for macro average metrics are shown in Fig. 3. Macro average is the averaged metric over all 4 classes. All models have metrics between %94 and %98, indicating very good performance for overall detection. Swin Transformer is the best model, making it the obvious choice for diagnosis. The overall recall and precision of the Swin Transformer is above 96.5%, therefore it is good not only for catching true positives but also for minimizing false positives. MaxViT and ViT have very similar metrics with MaxViT being a little better.

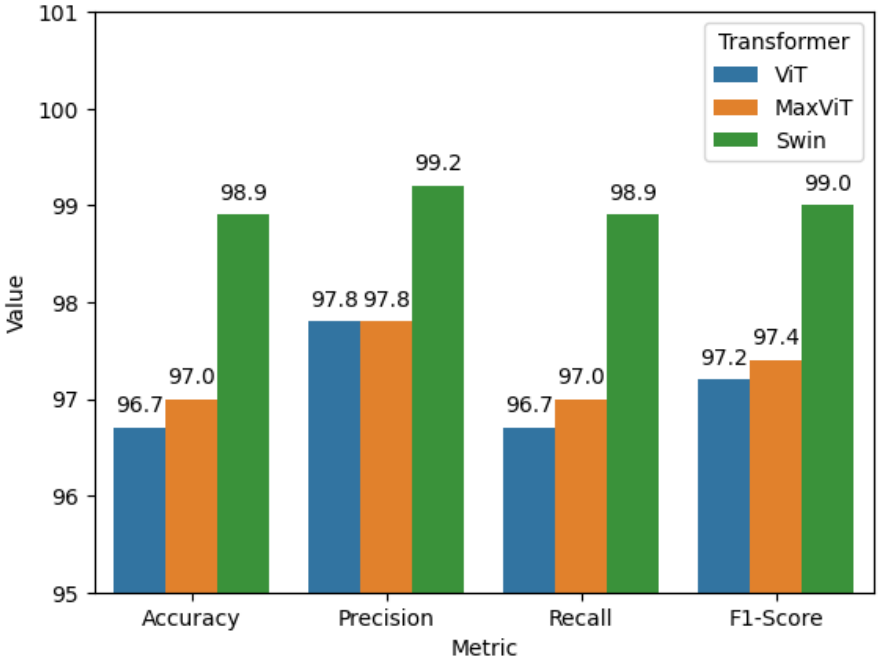

Figure 4. Results for Covid-19 Detection Metrics (%)

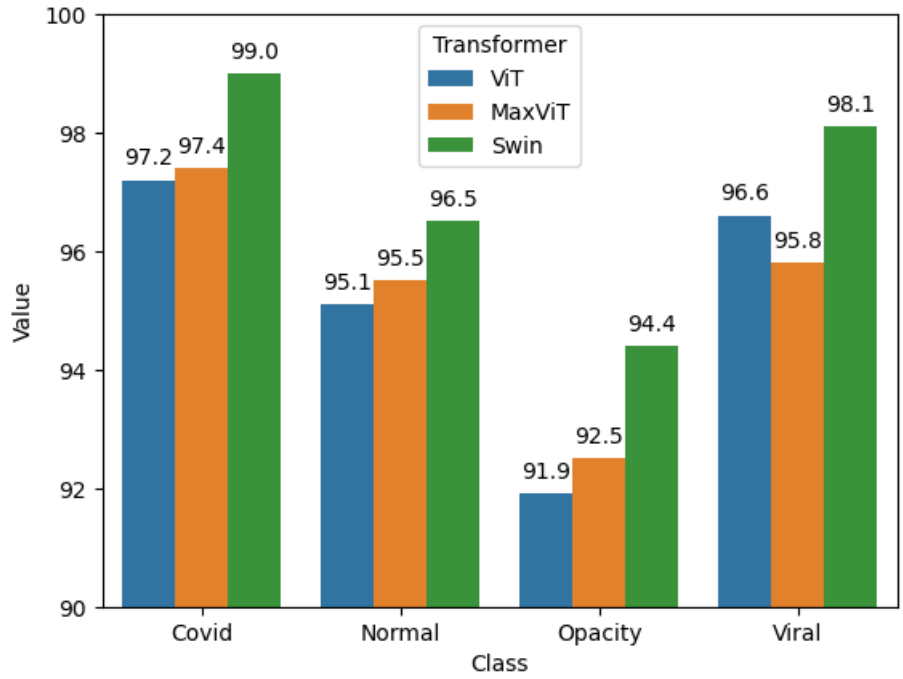

Figure 5. F_1 Scores for Each Class and Model (%)

Fig.4 shows the results of the Covid-19 detection metrics. These metrics have been calculated only on the Covid-19 classification performance. Again, the models show very high metrics. Swin Transformer is again the best model out of these three transformer variants, making it the best candidate for Covid-19 diagnosis. The overall scores for the Swin Transformer are around 99%, therefore it is good for detecting both Covid-19 positives and negatives. ViT and MaxViT have similar performances.

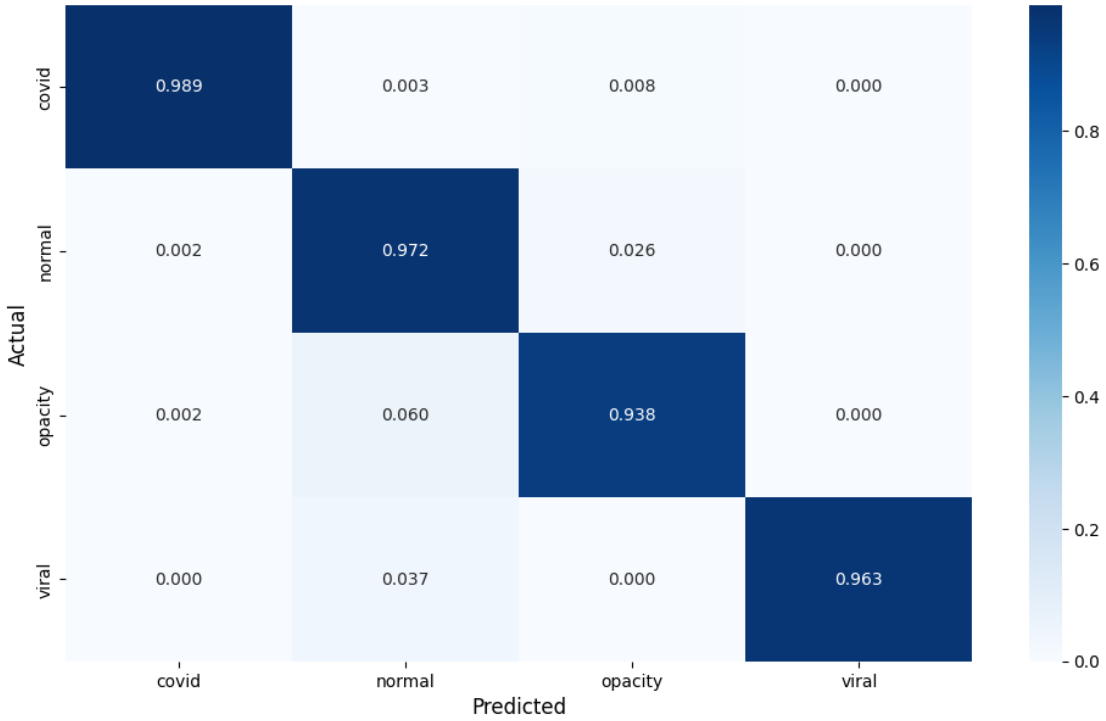

Figure 6. Swin Transformer Confusion Matrix for 4-Way Classification.

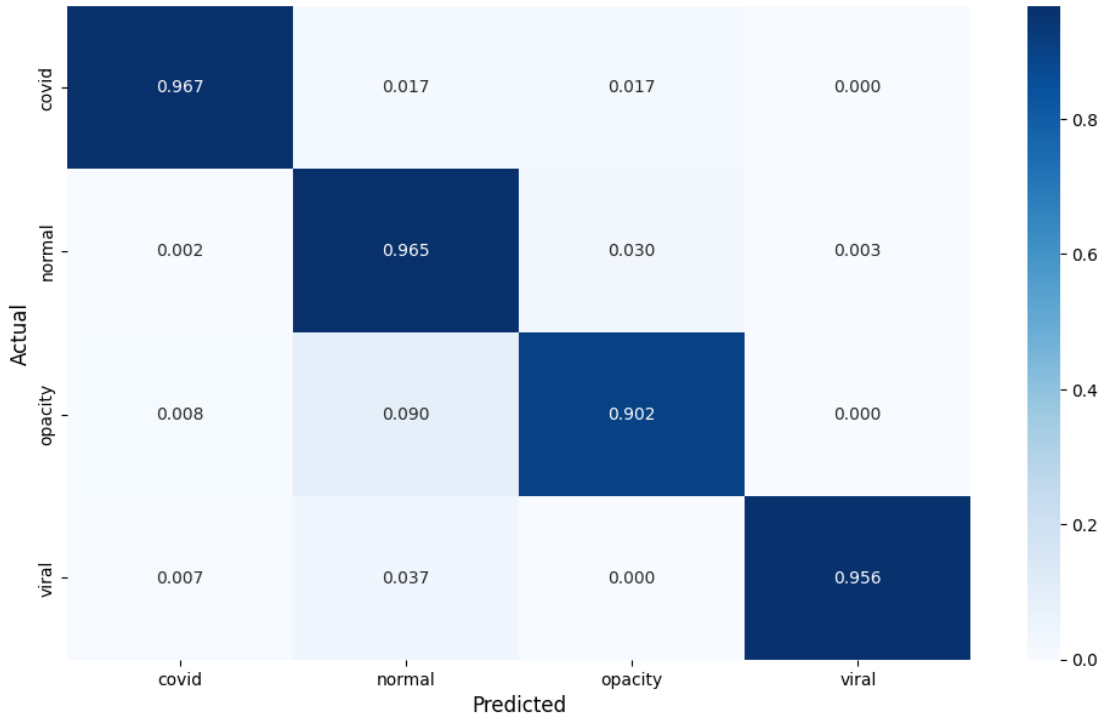

Figure 7. ViT Confusion Matrix for 4-Way Classification.

Fig. 5 displays the F_1 scores for each class and each model. Swin Transformer has the highest score for Covid-19 images at 99%. ViT and MaxViT have similar F_1 scores for Covid-19. All models have the lowest score for the Opacity class which makes this class the hardest to detect among these four.

Fig.6 shows the confusion matrix for the Swin Transformer which is the best transformer model. Opacity class has the smallest recall out of the four classes. Most of the confusion is between normal and opacity classes. There is also some confusion between normal and viral classes. Apart from those, there are not significant false negatives or positives between the classes.

Fig. 7 shows the confusion matrix for ViT. There is significant confusion between normal and opacity classes. Viral-normal confusion can also be seen in the figure. False negatives for Covid-19 come mostly from opacity and normal classes whereas false positives are mostly from viral and opacity classes.

CONCLUSIONS

This study assesses and compares the performance of the latest vision transformer-based neural network architectures for detecting COVID-19 using one of the most comprehensive public COVID-19 datasets. The networks benchmarked include ViT, MaxViT, and Swin Transformer. All models exhibit macro-averaged metrics ranging from 94% to 98%, indicating excellent overall detection performance. Among these three vision transformer variants, the Swin Transformer emerges as the superior model, making it the preferred choice for diagnosis. The Swin Transformer's overall recall and precision exceed 96.5%, making it effective not only in identifying true positives but also in minimizing false positives. With an accuracy of 98.9% for Covid-19, the Swin Transformer achieves one of the highest scores obtained with state-of-the-art techniques for COVID-19 diagnosis.

REFERENCES

Apostolopoulos, I. D. & Mpesiana, T. A. (2020). COVID-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine, 43(2), 635-640.

Chetoui, M., & Akhlouf, M. A. (2022). Explainable vision transformers and radiomics for covid-19 detection in chest x-rays. J. Clin. Med. 11, 3013.

Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Emadi, N.A., et al. (2020). Can AI help in screening viral and covid-19 pneumonia? https://arxiv.org/abs/2003.13145.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929.

Gorbalenya, A. E., Baker, S. C., Baric, R. S., De Groot, R. J., Drosten, C., Gulyaeva, A. A & Ziebuhr, J. (2020). The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol; 5, 536–44.

Khan, A. I., Shah, J. L. & Bhat, M. M. (2020). CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Computer Methods & Programs in Biomedicine, 196(26), 105581.

Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D. & Lessler, J. (2020). Variation in false-negative rate of reverse transcriptase polymerase chain reaction-based SARS-CoV-2 tests by time Since exposure. Annals of Internal Medicine, 173(4), 262–267.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. https://arxiv.org/abs/2103.14030.

Okolo, G. I., Katsigiannis, S. & Ramzan, N. (2022). Ievit: An enhanced vision transformer architecture for chest x-ray image classification. Comput. Methods Programs Biomed. 226, 107141.

Phelan, A. L., Katz, R. & Gostin, L. O. (2020). The novel coronavirus originating in Wuhan, China: Challenges for global health governance. JAMA, 323(8), 709– 710.

Rahhal, A.M.M. et al. (2022). Covid-19 detection in ct/x-ray imagery using vision transformers. J. Personal. Med. 12(2), 310.

Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, S.T., Maadeed, S.A., Zughaier, S.M., Khan, M.S., et al. (2021). Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection Using Chest X-ray Images. Computers in Biology and Medicine, 132, 104319.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022). MaxViT: Multi-Axis Vision Transformer. https://arxiv.org/abs/2204.01697.

Umer, M., Ashraf, I., Ullah, S., Mehmood, A., & Choi, G.S. (2022). Covinet: A convolutional neural network approach for predicting covid-19 from chest x-ray images. Journal of Ambient Intelligence and Humanized Computing, 13(1), 535–547.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. https://arxiv.org/abs/1706.03762.

Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. (2020) A novel coronavirus outbreak of global health concern. Lancet (London, England), 395(10223), 470– 473.

Wang, L., Lin Z. Q., & Wong, A. (2020). Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports, 10(1), 1–12.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). Visual Transformers: Token-based Image Representation and Processing for Computer Vision. https://arxiv.org/abs/2006.03677.

Yang, H., Wang, L., Xu, Y., & Liu, X. (2023). Covidvit: A novel neural network with self-attention mechanism to detect covid-19 through x-ray images. International Journal of Machine Learning and Cybernetics, 14, 973–987.