IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 August 2024, accepted 20 August 2024, date of publication 26 August 2024, date of current version 3 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3449802

APPLIED RESEARCH

Exploring Transformer for Face Mask Detection

YONGHUA MAO[1], YUHANG LV[1], GUANGXIN ZHANG[1], AND XIAOLIN GUI[2], (Member, IEEE)

[1]School of Computer Science/Shaanxi Key Laboratory of Clothing Intelligence, Xi'an Polytechnic University, Xi'an 710048, China
[2]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Corresponding author: Xiaolin Gui (xlgui@mail.xjtu.edu.cn)

This work was supported by the Natural Science Basic Research Program of Shaanxi Province (2023-JC- ZD-38), the National Key Research and Development Project (2022YFB3305503), the Key Science and Technology Project of Henan Province (201300210400), and the Key Research and Development Program of Shaanxi Province (2023-YBGY-403).

ABSTRACT The COVID-19 pandemic has underscored the importance of face masks in curbing viral transmission, prompting governments worldwide to enforce stringent public health mandates requiring mask usage in public areas. Consequently, there is a growing focus on developing automated mask detection technologies to augment these measures and minimize viral spread. In this study, we explore the potential of the Swin Transformer architecture for accurately identifying face mask usage, aiming to surpass the current performance limitations of existing face mask detection models. We evaluate the performance of our proposed model and comparison models using comprehensive evaluation metrics, including accuracy, precision, recall, specificity, F1-score, Kappa coefficient, and MCC. Our experiments yield several notable findings. Firstly, MobileNetV2 demonstrates superior performance compared to the baseline CNN model across all seven evaluation metrics within the face mask datasets. Secondly, within the category of convolutional neural networks (CNNs), EfficientNetV2 outperforms MobileNetV2, a classic lightweight network, across all metrics. DenseNet exhibits better performance than ResNet-50 across all seven evaluation metrics. Most significantly, the Swin Transformer architecture emerges as the most effective model, surpassing not only MobileNetV2 but also EfficientNetV2. The empirical results confirm that our Swin Transformer achieves statistically significant improvements in accuracy, precision, recall, specificity, F1-score, Kappa coefficient, and MCC compared to the other models.

INDEX TERMS Face mask detection, swin transformer, EfficientNet, MobileNet.

I. INTRODUCTION

The COVID-19 virus, a type of coronavirus, spread rapidly worldwide within a few months and continued to evolve through mutations. The World Health Organization classified it as a global pandemic in 2020 [1], [2], [3]. As of June 14, 2023, there have been 767,984,989 reported cases of COVID-19 and 6,943,390 deaths globally.

In response to the escalating situation, it has been found that the virus is primarily transmitted via droplets and airborne particles. When an infected individual coughs, sneezes, talks, or breathes, they release respiratory droplets carrying the virus. These droplets can be directly inhaled by nearby individuals or land on the mucous membranes of the mouth, nose, or eyes, causing infection.

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

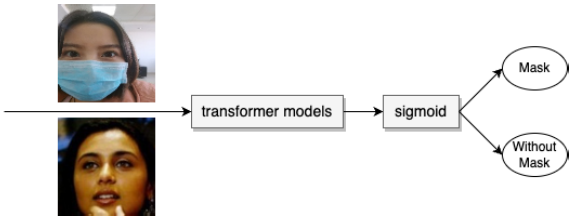

FIGURE 1. The face mask detection transformer model.

Besides droplet transmission, the COVID-19 virus can also exist in the air as small suspended particles called aerosols. If someone inhales aerosols containing the virus exhaled by an infected person, they may become infected [4], [5], [6]. Therefore, wearing masks is an effective way to prevent cross-infection and control the virus's spread. Masks protect the wearer from exposure to respiratory droplets and aerosols carrying the virus, reducing the risk of infection

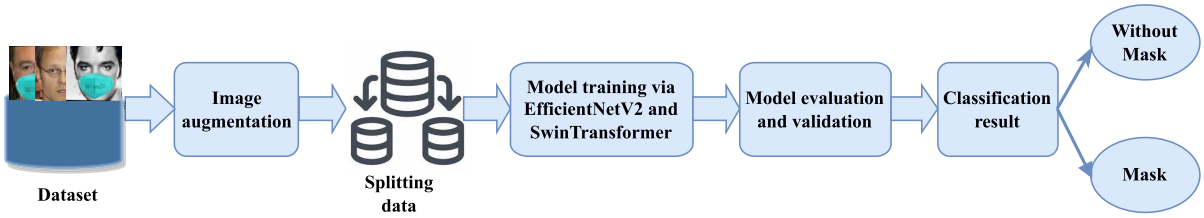

FIGURE 2. The schematic diagram of the proposed system.

from respiratory diseases like COVID-19, especially when in contact with patients or individuals who may carry the virus, as well as in enclosed and crowded environments [7], [8], [9].

Most governments are enforcing strict guidelines to wear masks in public places. It is not feasible to manually check if people are wearing masks [10]. Developing tools and techniques to detect mask usage can significantly reduce the spread of infections like COVID-19.

The automatic detection of masks on faces has become an increasingly crucial research topic, especially in the context of the COVID-19 pandemic. With the mandatory use of masks in many countries to curb the spread of the virus, the ability to automatically detect whether a person is wearing a mask has significant implications for public health and safety.

Moreover, the development of efficient and accurate mask detection algorithms can have wide-ranging applications beyond health and safety. For instance, they can be used in facial recognition systems to ensure that only masked individuals are granted access to certain areas, such as hospitals or public transportation. Additionally, these algorithms can be integrated into smart home systems to control the entry of guests or visitors and in retail spaces to monitor compliance with mask-wearing regulations.

As the use of masks becomes more prevalent in daily life, the ability to automatically detect masked faces will become increasingly important in various fields, including healthcare, public transportation, and security. Therefore, research into this area is not only essential but also timely and will undoubtedly continue to be a key focus of study in the coming years.

Due to the importance of face mask detection, this paper explores the use of transformer architectures, specifically the Swin transformer (see Figure 1), to detect whether an individual is wearing a mask. We focus on pushing the performance limits of face mask detection. Additionally, we propose four comparative network architectures: ResNet-50, DenseNet, MobileNet V2, and EfficientNet V2. We evaluate the performance of these face mask detection models using seven metrics.

Our study yields several noteworthy observations. First, we confirm that MobileNet V2 outperforms CNN in all seven evaluation metrics on our face mask datasets. Between the two comparative CNN models, DenseNet performs better than ResNet-50 [11] in all metrics. Second, EfficientNet V2 surpasses MobileNet V2 [12] in all evaluation metrics for face mask detection. Third, the Swin transformer outperforms both MobileNet V2 and EfficientNet V2. Our experimental results indicate that the Swin transformer exceeds the performance of the other four models across all seven evaluation metrics.

The research objectives of this study are twofold: first, to develop and evaluate Swin Transformer architectures for accurate face mask detection, and second, to assess their performance compared to existing convolution neural network (CNN) face mask detection models, which are MobileNet V2 and ResNet-50. The problem statement revolves around the need for effective automated face mask detection technologies in the context of the COVID-19 pandemic. By leveraging Swin Transformer architectures, we aim to overcome the current limitations of CNN-based models and achieve superior performance in face mask detection.

Addressing these goals contributes significantly to both computer vision and public health domains. In computer vision, the study advances the state-of-the-art by exploring novel architectures specifically tailored for face mask detection tasks, potentially leading to more efficient and accurate detection systems. From a public health perspective, the development of robust automated mask detection technologies can aid in enforcing mask mandates, thereby mitigating the spread of infectious diseases such as COVID-19 in public settings. Ultimately, this research aims to bridge the gap between computer vision advancements and real-world applications in public health, with the overarching aim of enhancing disease prevention strategies.

The rest of this paper is organized as follows: Section II reviews related work on face mask detection. Section III describes our experimental methodologies. In Section IV, we present the datasets, evaluation metrics, comparison results, and analysis of transformers and other related models. Section V concludes our work, and Section VI discusses future directions.

II. RELATED WORK

Our research focuses on detecting individuals not wearing face masks to help interrupt the transmission and spread of serious infectious diseases like COVID-19. Researchers have proposed various deep learning architectures to enhance the accuracy and efficiency of face mask recognition [10], [13], [14], [15], [16], [17].

Traditional face mask recognition methods face challenges in accuracy and efficiency. Venkateswarlu et al. [18] addressed these issues by employing a MobileNet model that can detect mask parts of faces. Sanjaya and

Rakhmawan [19] improved upon MobileNet by developing MobileNetV2, increasing the accuracy of face mask detection to 96.85%. Hussain and his team [20] presented a deep convolutional neural network (CNN) and MobileNetV2 based on transfer learning for mask detection. They tested their approach on two datasets, one with 2,500 images and the other with 4,436 images. Their comparative experiments showed that MobileNetV2 outperforms the deep CNN, achieving accuracy rates of 99% and 98% on these datasets, respectively. Elnady and Almghraby [12] utilized the deep learning architecture MobileNetV2, employing the optimization method 'stochastic gradient descent' with a learning rate set to 0.001 and a momentum of 0.85. The accuracy peaked after 12 epochs, reaching 99% for training and 98% for validation, with training and validation losses minimized at 0.05% and 0.025%, respectively.

On the other hand, Golwalkar and Mehendale [21] applied deep metric learning and their improved FaceMaskNet-21 to produce 128-dimensional encoding, facilitating face recognition from static images, real-time video streams, and static video files. They achieved a testing accuracy of 88.92% while maintaining an execution time below 10 ms.

Bishwas et al. [11] have made significant contributions in mask detection using a ResNet-50 model. Their work includes fine-tuning a pre-trained ResNet-50 model on their datasets, achieving an accuracy of 89%. This demonstrates the effectiveness of transfer learning for this specific task. They also developed and fine-tuned the hyperparameters of a ResNet-50 based architecture, resulting in approximately 47% accuracy in identifying faces within the masked face datasets. This indicates the challenges and complexities involved in distinguishing masked faces. They provided a detailed description of hyperparameter tuning for the model, offering insights into the optimization process and contributing valuable knowledge to the field.

Tan and Le [22] introduced EfficientNetV2, which surpasses its predecessor EfficientNet in training acceleration and parameter efficiency. EfficientNetV2 exhibits exceptional performance across various datasets, including ImageNet, CIFAR-10, and CIFAR-100. It attains 87.3% top-1 accuracy on the ImageNet dataset, with a 3-9x improvement in training speed and a 6.8x reduction in model size compared to previous models.

Liu et al. [23] proposed the network model Swin Transformer, which achieves superior results in COCO object detection and ADE20K semantic segmentation, significantly outperforming previous methods.

III. MATERIALS AND METHODS

In our research methodology, we selected the Swin Transformer architecture for face mask detection. The process flow is illustrated in Fig. 2. The first step involves image augmentation, including rotations, scaling, and other transformations. The second step is data splitting, dividing the datasets into training, validation, and testing sets in a 7:2:1 ratio. The third step involves constructing and training the face mask detection model using the training set. Next, model evaluation and validation are conducted to ensure proper training and prevent overfitting. Training is terminated when the validation accuracy stops improving. Finally, classification is performed on the testing set to determine whether faces are masked or not.

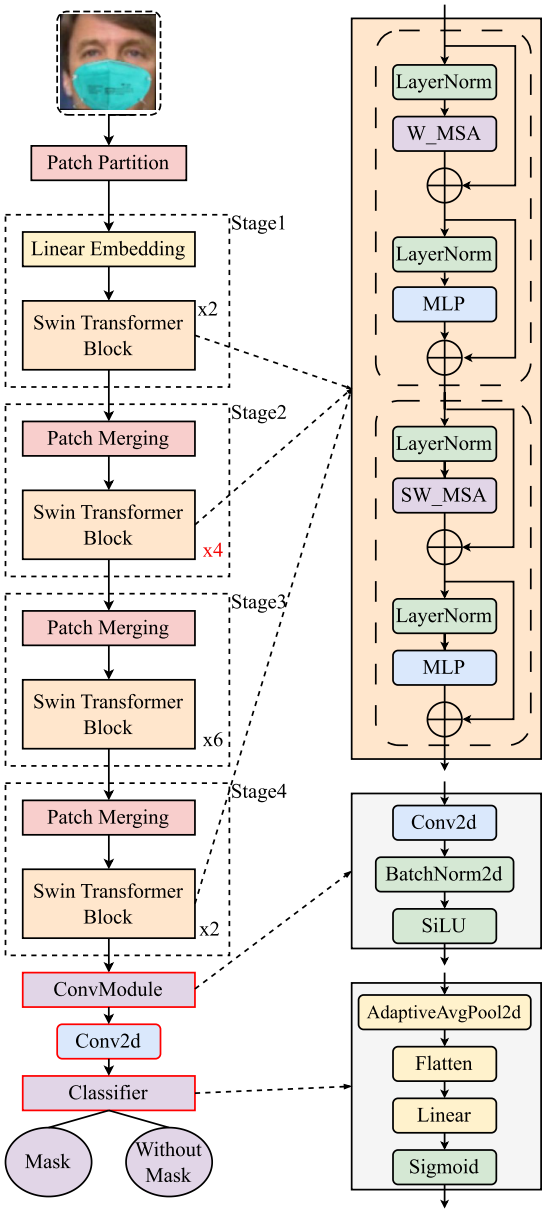

FIGURE 3. The Swin Transformer face mask detection model.

Additionally, we conducted an exhaustive evaluation of the performance of EfficientNetV2 and the Swin Transformer on this dataset and compared their performance with MobileNetV2 [12] and ResNet-50 [11], which were previously used for face mask detection in related works.

A. SWIN TRANSFORMER

Swin Transformer [23] is a novel vision transformer capable of functioning as a general-purpose backbone for

TABLE 1. Detailed architecture specifications.

	output size	Swin-T	Swin-st	Swin-h	Swin-d
stage 1	4×	concat 4×4, 96d, LN	concat 4×4, 96d, LN	concat 4×4, 96d, LN	concat 4×4, 128d, LN
	[56×56]	$[win.sz.7 \times 7, 96d, 3h] \times 2$	$[win.sz.7 \times 7, 96d, 3h] \times 2$	$[win.sz.7 \times 7, 96d, 3h] \times 2$	$[win.sz.7 \times 7, 128d, 3h] \times 2$
stage 2	8×	concat 2×2, 192d, LN	concat 2×2, 192d, LN	concat 2×2, 192d, LN	concat 2×2, 256d, LN
	[28×28]	$[win.sz.7 \times 7, 192d, 6h] \times 2$	$[win.sz.7 \times 7, 192d, 6h] \times 4$	$[win.sz.7 \times 7, 192d, 8h] \times 2$	$[win.sz.7 \times 7, 256d, 6h] \times 2$
stage 3	16×	concat 2×2, 384d, LN	concat 2×2, 384d, LN	concat 2×2, 384d, LN	concat 2×2, 512d, LN
	[14×14]	$[win.sz.7 \times 7, 384d, 12h] \times 6$	$[win.sz.7 \times 7, 384d, 12h] \times 6$	$[win.sz.7 \times 7, 384d, 16h] \times 6$	$[win.sz.7 \times 7, 512d, 12h] \times 6$
stage 4	32×	concat 2×2, 768d, LN	concat 2×2, 768d, LN	concat 2×2, 768d, LN	concat 2×2, 1024d, LN
	[7×7]	$[win.sz.7 \times 7, 768d, 24h] \times 2$	$[win.sz.7 \times 7, 768d, 24h] \times 2$	$[win.sz.7 \times 7, 768d, 32h] \times 2$	$[win.sz.7 \times 7, 1024d, 24h] \times 2$

computer vision. It builds hierarchical feature maps to capture different-scale information within an image. Swin Transformer models at diverse scales and has linear computational complexity relative to image size.

We propose an enhanced face mask detection framework based on the Swin Transformer (Swin-T), as illustrated in Figure 3. The framework comprises four stages, each progressively reducing the resolution of the input feature image. Initially, an input face image tensor is divided into non-overlapping patches by a patch splitting module. Each patch, treated as a "token," features a concatenation of raw pixel values and relative positional information.

In Stage 1, the raw-valued face features are projected to an arbitrary dimension using a linear embedding layer. In Stages 2 through 4, the feature tensor is subsampled by a patch merging layer. Each of these stages contains duplicated stacks of Swin Transformer blocks, with the number of blocks being 2, 4, 6, and 2, respectively, as optimized through extensive ablation experiments.

Subsequently, the features extracted by the Swin Transformer block are processed through a convolutional block. This block consists of a 3 × 3 Conv2d layer followed by BatchNorm2d and a SiLU activation function. The processed features are further refined using a 1 × 1 Conv2d layer. The output is then fed into a classifier layer for categorization. The classifier layer is composed of an AdaptiveAvgPool2d layer, a Flatten layer, a Linear layer, and a Sigmoid activation function. This classifier is designed to detect the presence or absence of a mask.

The red components in Figure 3 represent our improved module.

1) SWIN TRANSFORMER BLOCK

The Swin Transformer is designed by replacing the standard multi-head self-attention (MSA) module in a Transformer block with a module based on shifted windows. Other components, such as the LayerNorm layer, two-layer multi-layer perceptron (MLP), and residual connection, remain unchanged. Each Swin Transformer block primarily consists of window-based multi-head self-attention (W-MSA) module and a multi-layer perceptron module. The LayerNorm layer and the residual connection are applied to each module.

As shown in the right of Figure 3, two successive Swin Transformer blocks, which form a basic computational module of the Swin Transformer, differ in that one uses Window-based Multi-head Self-Attention (W-MSA) and the other uses Shifted Window-based Multi-head Self-Attention (SW-MSA)." W-MSA restricts attention to the interior of each blocklet, whereas SW-MSA introduces translational operations that allow the blocklet to interact with its surrounding blocklets. This combination of mechanisms enables the Swin Transformer to capture correlation information of an image at different granularities, leading to more powerful feature representation and learning capabilities. This also explains why the number of Swin Transformer blocks is even.

2) HYPER PARAMETER SPECIFICATION

First, we investigate four original Swin Transformer architectures, including Swin-T, Swin-S, Swin-B, and Swin-L [23], for face mask detection.

Using Swin-T as a baseline, we then examine the impact of different hyperparameter settings on the performance of the Swin Transformer model. Our experiments involve adjusting the depths (Swin-st), the number of attention heads (Swin-h), and the embedding dimensions (Swin-d) of the model, followed by an evaluation of the metrics on a test set.

As shown in Table 1, an input face image size of 224 × 224 is assumed for all architectures. "Concat n×n" indicates a concatenation of n×n neighboring features in a patch, resulting in a downsampling of the feature map by a rate of n. "96d" denotes a linear layer with an output dimension of 96. In Swin-d, we focus on different embedding dimensions. "win. sz. 7 × 7" indicates a multi-head self-attention module with a window size of 7 × 7. "3h" indicates three heads in the multi-head attention module. In Swin-h, we focus on varying the number of heads in different stages. "×2" means that the stage contains two Swin Transformer blocks. In Swin-st, we adjust the number of Swin Transformer blocks in stage 2. LN stands for LayerNorm.

B. COMPARISON METHODOLOGY

Several other face detection models are used in our experiments for comparison, namely MobileNetV2, DenseNet, and ResNet-50 [11]. MobileNetV2 is a significant model for face mask detection [12], [17], [19], [20], [24] and serves as our primary comparison methodology. ResNet-50 and DenseNet, both convolutional neural networks (CNNs) [10], [17], are key comparison models for MobileNetV2. Additionally, we develop an EfficientNet-based face mask detection model for comparative analysis.

1) EfficientNetV2

There are some challenges associated with the EfficientNet [25] family of models. EfficientNetB0 to EfficientNetB1 are suitable for scenarios with limited computational resources, while EfficientNetB2 to EfficientNetB3 build on these models and are suitable for a wider range of tasks and datasets. EfficientNetB4 to EfficientNetB7, though more powerful, have significantly more parameters and computational complexity, resulting in slow training speeds when dealing with large image sizes.

To address these limitations, Tan and Le introduced a novel convolutional neural network called EfficientNetV2. Unlike EfficientNet, which employs a composite scaling method, EfficientNetV2 adopts a non-uniform scaling strategy by dividing model scaling into two distinct branches: EfficientNetV2-S and EfficientNetV2-M. This non-uniform scaling strategy makes EfficientNetV2 more adaptable in design compared to its predecessor. Additionally, EfficientNetV2 incorporates a 'breadth boosting' strategy, which enhances the model's capacity by increasing the channel dimensions, allowing it to better capture data features and potentially improve performance across various tasks.

Tan and Le conducted experiments with EfficientNet-B4, replacing the MBConv with the Fused-Conv structure in the shallow layers, and observed a significant improvement in speed. After a series of experiments, they identified the optimal configuration by replacing the Fused-Conv structure from Stage 1 to Stage 3. This is illustrated in Figure 4.

Table 2 presents the architecture of our face mask detection model. The input layer processes a face image with 3 channels, and the output is a logistic regression layer. Stage 1 uses a 3×3 small convolution kernel to expand the number of channels to 24. Stages 2 to 4 consist of Fused-MB blocks, whose architecture is shown in Figure 4. Stages 5 to 7 are composed of MB Conv blocks. Stage 8 comprises a 1×1 convolution, pooling, and fully-connected (FC) layers.

2) MobileNetV2

MobileNetV2 [26], proposed by Google, is a lightweight network model that offers improved performance compared to its predecessor.

As described in related work, Elnady's model [12] achieves the highest performance among several MobileNetV2-based models [12], [17], [19], [20], [24] for face mask detection. Therefore, Elnady's model is selected as a key previous work for comparison with our model.

MobileNetV2 introduces an inverted residual structure. The traditional residual structure uses a $1\times$ 1 convolution to reduce the number of channels, followed by a 3×3 convolution, and finally another 1×1 convolution to restore the number of channels. In contrast, the inverted residual block in MobileNetV2 replaces the 3×3 convolution with a depthwise separable convolution. It first uses a 1×1 convolution to increase the number of channels, then applies a depthwise 3×3 convolution, and finally uses another 1×1 convolution to reduce the number of channels.

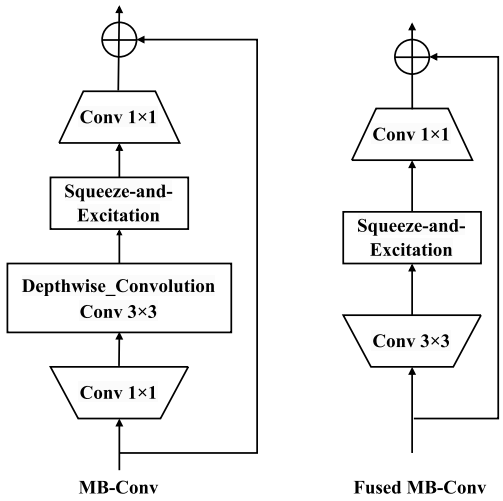

FIGURE 4. The Architecture of MBConv and Fused-MBConvt.

TABLE 2. The architecture of EfficientNetV2-S face mask detection model.

Stage	Operator	Stride	Channels	Layers
Input	Face Image	-	3	-
1	Conv 3×3	2	24	1
2	Fused-MB Conv1, k 3×3	1	24	2
3	Fused-MB Conv4, k 3×3	2	48	4
4	Fused-MB Conv4, k 3×3	2	64	4
5	MB Conv4, k 3×3,SE0.25	2	128	6
6	MB Conv4, k 3×3, SE0.25	1	160	9
7	MB Conv4, k 3×3, SE0.25	2	256	15
8	Conv 1×1 & Pooling & FC	-	1280	1
Output	sigmoid	-	1	1

3) ResNet-50

ResNet-50 [11] is another previous work used for comparison with our model.

ResNet-50 [27] serves as a feature extractor, coupled with a fully connected layer and a sigmoid output function for our face mask classification. Residual connections are introduced to address issues such as gradient vanishing and gradient explosion in deep neural network training. These connections create direct, skip paths between different layers, allowing residual blocks to learn how to add input features to output features effectively. This enables the construction of deeper neural networks.

4) DenseNet

We use DenseNet-169 [28] as a feature extractor, coupled with a fully connected layer and a sigmoid output function as a classifier. The building blocks of DenseNet-169 primarily consist of DenseBlocks and Transition layers. The structure of a DenseBlock is BN+ReLU+1 $\times$ 1Conv+BN+ReLU+3 $\times$ 3Conv. This internal structure is repeated multiple times within the DenseBlock, allowing the feature maps to be densely connected, which enhances feature representation and reduces the number of parameters.

The Transition layer consists of BN+ReLU+1 $\times$ 1Conv+2$\times$2AvgPooling, and it connects each DenseBlock. When the input to the Transition layer has a feature mapping with C channels, the output feature mapping has C times the compression coefficient channels (the compression

FIGURE 5. Sample images in the dataset.

coefficient is usually between 0 and 1). This effectively reduces the computational complexity and the number of parameters of the model.

IV. EXPERIMENTAL RESULT AND DISCUSSION

All models were trained using an NVIDIA GeForce RTX4060 laptop GPU. The algorithms were implemented using the Pytorch framework, and several experiments were conducted to evaluate the performance of the models.

A. DATASETS

The datasets used in this study were obtained from the National Multimedia Software Engineering Technology Research Centre of Wuhan University. The primary sources of data were the Real World Masked Face Dataset (RMFRD) [29] and the Simulated Masked Face Dataset (SMFRD) [30], with supplementary data from the Moxa3K dataset [31]. To ensure diversity and balance, a combined total of 24,594 images were selected, maintaining an almost equal distribution of masked and unmasked face images, as detailed in Table 3. Sample images from these datasets are displayed in Figure 5.

TABLE 3. List of datasets.

Datasets	Mask	No-Masks
RMFRD	5000	90000
SMFRD	6559	6558
Moxa3K	5380	14
Ours(no synth.)	7695	8868
Ours	12019	12575

The RMFRD dataset contributed 5,000 masked images and 90,000 non-masked images, making it the largest source of non-masked data. SMFRD provided 6,559 masked and 6,558 non-masked images, offering a balanced dataset for both categories. The Moxa3K dataset, while smaller in size, included 5,380 masked images but only 14 non-masked images, highlighting a significant imbalance that necessitated careful handling during the augmentation process.

From these datasets, we selected a subset of 3,680 masked images from RMFRD, 3,946 from SMFRD, and 69 from Moxa3K, amounting to a total of 7,695 masked images. To further improve the robustness and generalization capabilities of our model, we applied data augmentation techniques to these images, involving random rotations between -15 and 15 degrees. This choice avoided large-angle rotations that could distort natural facial poses, which is critical for maintaining accuracy in mask detection. The augmentation process resulted in an additional 4,324 images, which were incorporated into the training set.

For non-masked images, we selected 4,809 from RMFRD, 4,059 from SMFRD, and those from Moxa3K, and generated 3,707 additional augmented images. These images were also added to the training set. After augmentation, the final dataset was organized into training, validation, and test sets in a 7:2:1 ratio, comprising 17,216 images for training, 4,919 for validation, and 2,459 for testing.

Overall, our dataset preparation process ensured a well-balanced and diverse dataset, critical for training a robust and generalizable mask detection model.

B. EVALUATION METRICS

To compare and evaluate the performance of different face mask detection models, several evaluation metrics [32], including accuracy, precision, recall, specificity, F1-score, and kappa coefficient, have been investigated in this paper. These metrics are calculated using the confusion matrix shown in Table 4, which includes True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values.

TABLE 4. Confusion matrix.

Confuse / Predict / Actual	Positive	Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP(False Positive)	TN(True Negative)

The confusion matrix components are defined as follows:

- True Positive (TP): The number of masked images correctly identified as masked.
- True Negative (TN): The number of unmasked images correctly identified as unmasked.
- False Positive (FP):The number of unmasked images mistakenly labeled as masked.
- False Negative (FN):The number of masked images mistakenly classified as unmasked.

1) ACCURACY

Accuracy is defined as shown in (1). It measures the ratio of correctly classified samples to the total number of samples in the testing data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

2) PRECISION

Precision calculates the proportion of true positive predictions out of all positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

3) RECALL

Recall represents the ratio of true positive predictions to the actual number of positive samples in the

testing data.

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

4) SPECIFICITY

Specificity indicates the proportion of true negative predictions out of all negative predictions made by the model.

$$Specificity = \frac{TN}{TN + FP} \times 100\% \tag{4}$$

5) F1-SCORE

The F1-score is the harmonic mean of recall and precision, offering a balance between these two metrics.

$$\begin{aligned} F1 &= \frac{2 \times precision \times recall}{precision + recall} \times 100\% \\ &= \frac{2TP}{2TP + FP + FN} \times 100\% \end{aligned} \tag{5}$$

6) KAPPA COEFFICIENT

The kappa coefficient measures the agreement between the model's predictions and the actual class labels, taking into account the possibility of agreements due to chance.

$$Kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy} \times 100\% \tag{6}$$

7) MATTHEWS COEFFICIENT

The Matthews correlation coefficient (MCC) is a comprehensive performance indicator used to evaluate binary classification models, as shown in (7). A higher MCC value indicates better performance, a value of approximately 0 indicates random performance, and a negative value implies poorer performance, suggesting the model performs worse than random guessing. The MCC is particularly advantageous in cases of imbalanced categories and small sample sizes because it considers all four classification outcomes when evaluating performance, thereby reducing the randomness of the results.

$$\begin{aligned} M = &\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \\ &\times 100\% \end{aligned} \tag{7}$$

In summary, the evaluation of classification models involved comprehensive analysis using the mentioned metrics, providing a well-rounded assessment of their performance on the testing data.

C. HYPER PARAMETER OPTIMIZATION

Through ablation experiments, we determined that tuning the model depth had the most significant effect on performance. As shown in Table 5, Swin-T outperforms the other architectures (Swin-B, Swin-S, and Swin-L) in most evaluation metrics. Therefore, we chose Swin-T as the baseline for further optimization.

The optimal configuration was found to be Swin-st, which optimally adjusts the depth of the Swin Transformer blocks. This suggests that, with the same embedding dimension and number of attention heads, appropriately increasing the depth of the middle layers of the model can substantially improve its representation ability and classification effectiveness for face mask detection.

Furthermore, we observed that adjusting the number of attention heads (Swin-h) and embedding dimensions (Swin-d) also influences model performance, albeit to a lesser extent. Specifically, increasing the number of attention heads slightly improved some metrics, while increasing the embedding dimension had a slightly lesser impact.

Thus, Swin-st was selected as our base architecture for face mask detection. To enhance the performance of the transformer network, we added a classification module at the end of the Swin Transformer architecture, resulting in our final model.

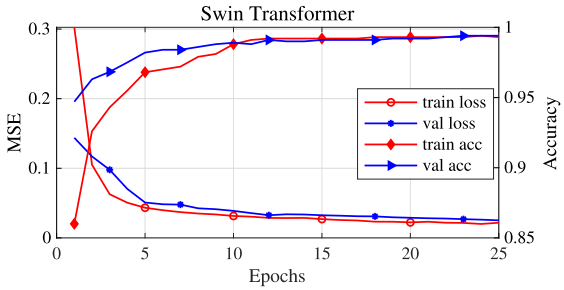

FIGURE 6. The training result of face mask Swin Transformer model.

D. RESULTS OF TRAINING AND VALIDATION DATASET

Figure 6 illustrates the training results of the Swin Transformer model for face mask detection. 'train loss' and 'val loss' denote the mean square error between the labels and detection results on the training and validation sets, respectively. 'train acc' and 'val acc' represent the accuracy of face mask detection on the training and validation sets, respectively. As shown in Figure 6, the training accuracy mirrors the validation accuracy after epoch 15, with both metrics plateauing. Similarly, the loss function trends for both the training and validation sets converge after epoch 20.

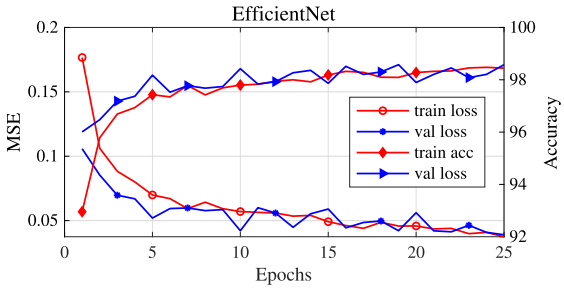

FIGURE 7. The training result of the face mask EfficientNet model.

Figures 7, 8, 9, and 10 depict the training results of our face mask detection models using different architectures—EfficientNet V2, MobileNet V2, ResNet-50, and DenseNet, respectively. These models were trained over multiple epochs to enhance their accuracy and performance.

TABLE 5. Results obtained for swin transformer architecture.

Results / Metric / Model	TP	FP	TN	FN	Accuracy	Precision	Recall	Specificity	F1-score	Kappa coefficient	Mcc	Parameters
Swin-T	1239	4	1214	2	99.8%	99.7%	99.8%	99.7%	99.8%	99.5%	99.5%	27.5M
Swin-B	1240	5	1213	1	99.8%	99.6%	99.9%	99.6%	99.8%	99.5%	99.5%	86.7M
Swin-S	1230	6	1212	1	99.7%	99.5%	99.9%	99.4%	99.7%	99.4%	99.4%	48.8M
Swin-L	1240	5	1213	1	99.8%	99.6%	99.9%	99.6%	99.8%	99.5%	99.5%	195.0M
Swin-st	1240	3	1215	1	99.8%	99.8%	99.9%	99.8%	99.8%	99.7%	99.7%	28.4M
Swin-h	1238	3	1215	3	99.8%	99.8%	99.8%	99.8%	99.8%	99.5%	99.5%	27.5M
Swin-d	1237	3	1216	4	99.7%	99.8%	99.7%	99.8%	99.7%	99.4%	99.4%	48.9M

Figure 7 shows the training progress of the EfficientNet V2-based model, demonstrating a gradual improvement in performance. The loss function steadily decreases, while the accuracy metric consistently increases.

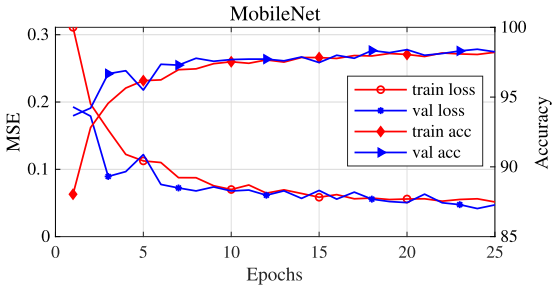

FIGURE 8. The training result of face mask MobileNet model.

Figure 8 presents the training outcomes of the face mask detection model using MobileNet V2 architecture. Similar to the previous figure, there is an upward trend in accuracy and a downward trend in loss after each epoch, indicating effective learning from the dataset.

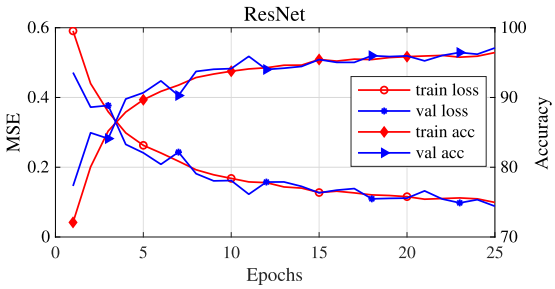

FIGURE 9. The training result of face mask ResNet-50 model.

Figure 9 displays the training results of the face mask detection model based on ResNet-50 architecture, showing promising progress with decreasing loss values and increasing accuracy scores over time.

Figure 10 illustrates the training results of the face mask detection model built on DenseNet architecture. It exhibits similar trends, with diminishing loss values and increasing accuracy rates throughout successive epochs.

After 15 epochs of training, all models, including Swin Transformer, EfficientNet V2, MobileNet V2, ResNet-50, and DenseNet, converge towards stability. Further improvements become marginal beyond this point, suggesting that 25 epochs of training are sufficient for these models.

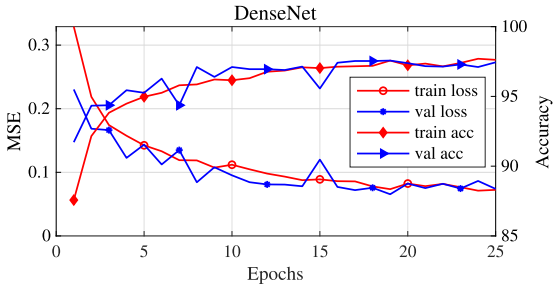

FIGURE 10. The training result of face mask DenseNet model.

Additional iterations may not significantly enhance their performance.

In conclusion, through 25 epochs of training, all face mask detection models—including our Swin Transformer, EfficientNet V2, MobileNet V2, ResNet-50, and DenseNet—were trained effectively.

E. COMPARISON BETWEEN OUR PROPOSED MODELS AND RELATED WORKS

1) RESULT OF OUR DATASETS

Table 6 provides a comprehensive comparative analysis of the evaluation metrics—Accuracy, Precision, Recall, F1 Score, Specificity, Kappa Coefficient, and MCC—for five different model architectures. Our improved Swin Transformer model, referred to as Swin-st in Table 5, is our proposed model. MobileNetV2, which is one of the state-of-the-art models in face mask detection, is highlighted in related works [12], [17], [19], [20], [24]. ResNet-50 [11] is another significant model in face mask detection. All results are derived from our test set, including those for MobileNetV2, which were obtained after training and validation on our datasets.

The ResNet-50 and DenseNet architectures, widely adopted in computer vision tasks, both achieve approximately 97% in accuracy, precision, recall, specificity, and F1 score. These architectures provide strong baseline performance in our face mask detection task, with parameters in the tens of millions. They exhibit good generalization ability and are relatively easy to train, making them suitable for tasks with limited computational resources. However, in terms of the Kappa coefficient and Matthews correlation coefficient (MCC), ResNet-50 and DenseNet perform much lower than other models, as shown in Figure 11. Compared to newer architectures like EfficientNetV2 and the Swin Transformer, ResNet-50 and DenseNet may struggle to

TABLE 6. Results obtained for the classification models on our test set.

Results / Metric / Model	Accuracy	Precision	Recall	Specificity	F1-score	Kappa coefficient	MCC	Parameters
ResNet-50 [11]	97.2%	97.3%	97.2%	97.3%	97.3%	94.5%	94.5%	24.5M
DenseNet	98.3%	98.1%	98.5%	98.0%	98.3%	96.6%	96.6%	12.5M
MobileNetV2 [12]	99.0 %	99.1%	99.0%	99.1%	98.6%	98.0%	98.0%	3.2M
EfficientNetV2	99.3 %	99.2%	99.4%	99.2%	99.3%	98.6%	98.6%	20.2M
Our Swin transformer	99.8%	99.9%	99.8%	99.9%	99.8%	99.7%	99.7%	28.4M

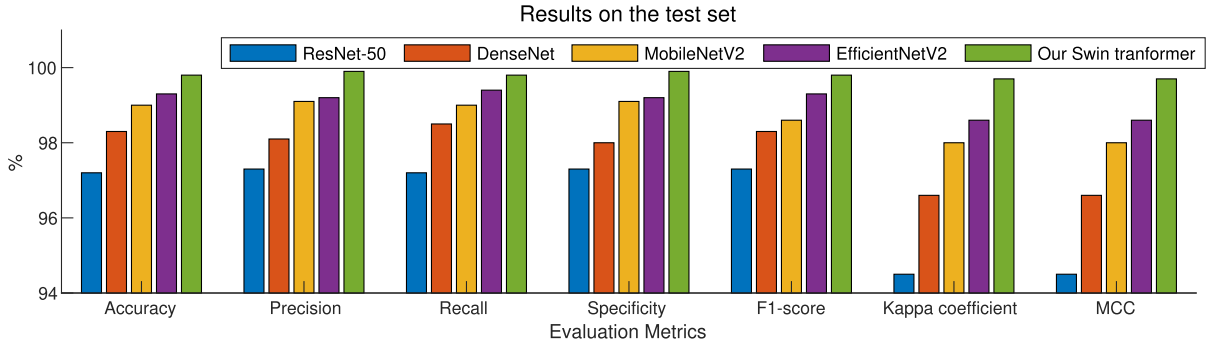

FIGURE 11. Accuracy, precision, recall, specificity, F1-score, and kappa coefficient of different models on the test set.

capture fine-grained details and long-range dependencies in images, potentially limiting their performance in complex scenarios.

MobileNetV2, known for its lightweight design and efficient inference, has only 3 million parameters and achieves over 99% in accuracy, precision, recall, and specificity in our face mask detection task, outperforming ResNet-50 and DenseNet. These characteristics make it suitable for deployment on resource-constrained devices. MobileNetV2 balances model size and accuracy well, making it a practical choice for real-time applications. However, in terms of F1-score, Kappa coefficient, and MCC, MobileNetV2 performs lower than more complex architectures like Swin Transformer and EfficientNet V2, as shown in Figure 11. This suggests that MobileNetV2 may sacrifice some accuracy and representational capacity, especially in tasks requiring fine-grained feature extraction.

EfficientNetV2, with 20.3 million parameters, surpasses MobileNetV2 in all seven evaluation metrics, especially in recall, F1-score, Kappa coefficient, and MCC. EfficientNetV2 shows competitive performance in face mask detection tasks, effectively balancing model complexity and accuracy. It uses compound scaling to efficiently balance model depth, width, and resolution, leading to improved performance. However, compared to the Swin Transformer, EfficientNetV2 scores lower in every evaluation metric, as shown in Figure 11. Despite its efficiency, EfficientNetV2 may not capture intricate spatial relationships in images as effectively as more complex architectures like the Swin Transformer. The fixed scaling coefficients may limit its adaptability to diverse datasets and scenarios compared to architectures with more flexible hyperparameters.

As shown in Figure 11, our Swin Transformer model surpasses the other four models across all evaluation metrics. Specifically, in terms of Kappa coefficient and MCC, ResNet-50 [11] achieves approximately 94%, while our Swin Transformer reaches an impressive 99.7%, a 5.7% improvement. From Table 6, it is clear that our Swin Transformer model excels in mask detection across all evaluation metrics, outperforming the other four neural network models. It is notable that MobileNetV2 [12], a state-of-the-art model for face mask detection, is outperformed by our model by 1.7% in terms of Kappa coefficient and MCC. The Swin Transformer shows significant gains in face mask detection, particularly in capturing long-range dependencies and contextual information within images. Its self-attention mechanism effectively models global relationships among image pixels, leading to superior feature representation. However, the Swin Transformer, with 28.4 million parameters, has higher computational complexity than simpler architectures like MobileNetV2, which has only 3 million parameters. This results in increased resource requirements during training and inference. Despite its effectiveness, the Swin Transformer may require larger datasets and longer training times to fully utilize its capabilities compared to lightweight architectures like MobileNetV2.

In our comparative analysis of different architectures for face mask detection, including EfficientNet V2, Swin Transformer, ResNet-50, DenseNet, and MobileNetV2, we observed distinct strengths and weaknesses in each model. The Swin Transformer shows significant performance gains in face mask detection, especially in capturing long-range dependencies and contextual information within images. Its self-attention mechanism effectively models global relationships among image pixels, leading to superior feature representation. However, our Swin Transformer's computational complexity is higher than that of simpler architectures like MobileNetV2, resulting in increased resource requirements

TABLE 7. Results obtained by the classification model on the SMFRD test set.

Results \ Metric / Model	Accuracy	Precision	Recall	Specificity	F1-score	Kappa coefficient	MCC	Parameters
ResNet-50 [11]	97.2%	97.2%	97.3%	97.1%	97.2%	94.5%	94.5%	24.5M
DenseNet	98.5%	98.3%	98.8%	98.3%	98.5%	97.0%	97.0%	12.5M
MobileNetV2 [12]	98.9%	99.1%	98.8%	99.1%	98.9%	97.9%	97.9%	3.2M
EfficientNetV2	99.4 %	99.3%	99.5%	99.3%	99.4%	98.8%	98.8%	20.2M
Our Swin transformer	99.7%	99.8%	99.6%	99.8%	99.7%	99.5%	99.5%	28.4M

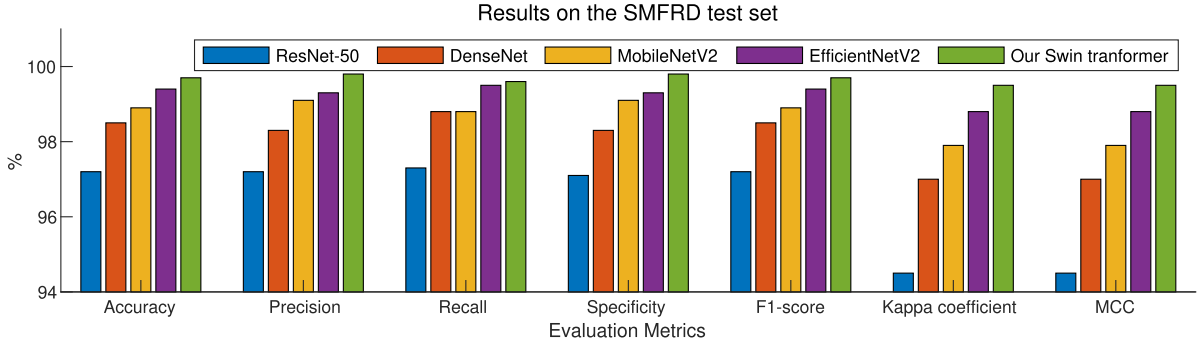

FIGURE 12. Accuracy, precision, recall, specificity, F1-score, and kappa coefficient of different models on the SMFRD test set.

during training and inference. Despite its effectiveness, our Swin Transformer may require larger datasets and longer training times to fully utilize its capabilities compared to lightweight architectures.

This finding underscores the importance of leveraging advanced architectural components, such as self-attention mechanisms, to capture long-range dependencies and contextual information within images.

2) RESULT OF THE SMFRD DATASET

In this subsection, we compare the performance of our proposed model on the SMFRD dataset with four other models to enhance the quality and robustness of the experiments. The results for ResNet-50 and MobileNetV2 were tested on the same dataset and did not come directly from the original papers. The experimental results on the SMFRD dataset, summarized in Table 7, present a comparative evaluation of several models, including ResNet-50, MobileNetV2, and our proposed Swin Transformer. This subsection provides a detailed analysis of these results, focusing on the comparison between our Swin Transformer and the baseline models ResNet-50 [11] and MobileNetV2 [12].

As shown in Figure 12, our Swin Transformer achieves a superior accuracy of 99.7%, with minor variations between 99.6% and 99.8%. In comparison, ResNet-50 achieves around 97.2%, and MobileNetV2 records approximately 98.9%. This significant improvement demonstrates the Swin Transformer's effectiveness in capturing intricate patterns and generalizing well across the dataset. The accuracy margin of approximately 2.5% over ResNet-50 and 0.8% over MobileNetV2 underscores its robustness.

The parameter count is crucial for evaluating model efficiency and scalability. Our Swin Transformer has 28.4 million parameters, compared to ResNet-50's 24.5 million and MobileNetV2's 3.2 million. While the Swin Transformer's larger parameter count indicates higher computational and memory requirements, it also enhances learning capacity and accuracy. Conversely, MobileNetV2's lower parameter count balances efficiency and performance, making it suitable for resource-constrained environments.

Beyond accuracy, our Swin Transformer performs superiorly in other metrics, consistently scoring 99.5%. In contrast, ResNet-50 scores 94.5%, and MobileNetV2 scores 97.9%. These results highlight the comprehensive performance advantages of the Swin Transformer, indicating its effectiveness in accuracy and other critical evaluation parameters.

The comparative analysis reveals that our Swin Transformer outperforms both ResNet-50 and MobileNetV2 across all primary evaluation metrics. The increase in parameter count is justified by substantial gains in accuracy and performance. MobileNetV2, while less accurate than the Swin Transformer, offers a favorable trade-off with its minimal parameter count, making it attractive for resource-limited applications. ResNet-50, although reliable, falls short in accuracy and additional performance metrics compared to the other two models.

V. CONCLUSION

The ongoing COVID-19 pandemic, which has significantly impacted global health since its emergence in 2019, has underscored the effectiveness of face masks in curbing viral transmission. As a result, many governments worldwide have enforced stringent public health measures mandating mask usage in public settings. In response, there has been a growing emphasis on the development of automated mask detection technologies to reinforce these efforts and reduce viral transmission rates.

In this study, we proposed and evaluated the Swin Transformer model for the task of face mask detection. Our extensive experiments on the SMFRD and our own datasets demonstrated that the Swin Transformer significantly outperforms existing models, including ResNet-50 [11] and MobileNetV2 [12], [17], [19], [20], [24], across all key evaluation metrics such as accuracy, precision, recall, specificity, F1-score, Kappa coefficient, and MCC.

Our experiments yield several noteworthy findings. Firstly, MobileNetV2 demonstrates superior performance compared to the baseline CNN model across all seven evaluation metrics within the face mask datasets. Secondly, within the category of convolutional neural networks (CNNs), EfficientNetV2 outperforms MobileNetV2, a state-of-the-art model for face mask detection, across all metrics. Additionally, DenseNet exhibits better performance than ResNet-50, another related work for face mask detection, across all metrics. Most significantly, our Swin Transformer architecture emerges as the most effective model, surpassing not only MobileNetV2 but also EfficientNetV2. The empirical results confirm that our Swin Transformer achieves statistically significant improvements in accuracy, precision, recall, specificity, F1-score, Kappa coefficient, and MCC compared to the other models.

In conclusion, our experimental results offer valuable insights into the field of face mask detection, highlighting the potential of advanced architectures like the Swin Transformer for real-world applications. By leveraging these findings, researchers and practitioners can continue to innovate and develop cutting-edge solutions to address pressing challenges in public health and safety. Our Swin Transformer's remarkable robustness and generalization capabilities make it a promising candidate for practical applications involving diverse datasets, contributing significantly to the enhancement of automated mask detection technologies.

VI. LIMITATIONS AND FUTURE WORK

While our study has provided valuable insights into face mask detection using state-of-the-art architectures, it is important to acknowledge several limitations and potential areas for future research:

1) DATASET LIMITATIONS

The performance of our models is contingent upon the quality and diversity of the dataset used for training and evaluation. While we made efforts to curate a comprehensive dataset, there may still be inherent biases or limitations in the data, such as imbalanced class distributions or variations in image quality.

Future research could explore the use of larger and more diverse datasets to improve model generalization and robustness across different demographic groups, environmental conditions, and mask types. Additionally, incorporating domain-specific data augmentation techniques tailored to the characteristics of face mask images could further enhance model performance.

2) MODEL INTERPRETABILITY AND EXPLAINABILITY

While our models have demonstrated high accuracy and performance, their internal mechanisms may lack interpretability, making it challenging to understand the reasoning behind model predictions. This could pose challenges in gaining trust and acceptance from end-users and stakeholders.

Future research could focus on enhancing the interpretability and explainability of deep learning models for face mask detection, employing techniques such as attention mechanisms, saliency maps, and model visualization methods. By providing insights into model decision-making processes, these approaches can enhance transparency and facilitate model adoption in real-world settings.

3) REAL-WORLD DEPLOYMENT CONSIDERATIONS

The practical deployment of face mask detection systems involves various logistical, ethical, and legal considerations that must be addressed. These include privacy concerns, data security, regulatory compliance, and societal acceptance.

Future research should engage stakeholders from diverse backgrounds, including policymakers, privacy advocates, and end-users, to develop ethically sound and socially responsible deployment strategies. Collaboration with domain experts and interdisciplinary teams can help navigate these complex challenges and ensure the responsible adoption of face mask detection technology.

In summary, while our study has made significant contributions to the field of face mask detection, there remain important avenues for future research to address limitations and advance the state-of-the-art. By addressing these challenges and embracing interdisciplinary collaboration, we can continue to innovate and develop robust, effective, and ethically responsible solutions for public health and safety.

REFERENCES

[1] M. Ciotti, M. Ciccozzi, A. Terrinoni, W. C. Jiang, and S. Bernardini, "The COVID-19 pandemic," *Crit. Rev. Clin. Lab. Sci.*, vol. 57, no. 1, pp. 365–388, 2020.

[2] J. Watkins, "Preventing a COVID-19 pandemic," *BMJ*, vol. 368, p. 810, Feb. 2020.

[3] A. Haleem, M. Javaid, and R. Vaishya, "Effects of COVID-19 pandemic in daily life," *Current Med. Res. Pract.*, vol. 10, no. 2, pp. 78–79, Mar. 2020.

[4] L. Li, Z. Yang, Z. Dang, C. Meng, J. Huang, H. Meng, D. Wang, G. Chen, J. Zhang, H. Peng, and Y. Shao, "Propagation analysis and prediction of the COVID-19," *Infectious Disease Model.*, vol. 5, pp. 282–292, Jan. 2020.

[5] A. Masrur, M. Yu, W. Luo, and A. Dewan, "Space-time patterns, change, and propagation of COVID-19 risk relative to the intervention scenarios in Bangladesh," *Int. J. Environ. Res. Public Health*, vol. 17, no. 16, p. 5911, Aug. 2020.

[6] T. Galbadage, B. M. Peterson, and R. S. Gunasekera, "Does COVID-19 spread through droplets alone?" *Frontiers Public Health*, vol. 8, Apr. 2020, Art. no. 163.

[7] S. H. Park, "Personal protective equipment for healthcare workers during the COVID-19 pandemic," *Infection Chemotherapy*, vol. 52, no. 2, p. 165, 2020.

[8] S. E. Eikenberry, M. Mancuso, E. Iboi, T. Phan, K. Eikenberry, Y. Kuang, E. Kostelich, and A. B. Gumel, "To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic," *Infectious Disease Model.*, vol. 5, pp. 293–308, Jan. 2020.

[9] R. Tirupathi, K. Bharathidasan, V. Palabindala, S. A. Salim, and J. A. Al-Tawfiq, "Comprehensive review of mask utility and challenges during the COVID-19 pandemic," *Infezioni Medicina*, vol. 28, no. 1, pp. 57–63, 2020.
[10] A. M. Lad, A. Mishra, and A. Rajagopalan, "Comparative analysis of convolutional neural network architectures for real time COVID-19 facial mask detection," *J. Phys., Conf. Ser.*, vol. 1969, no. 1, Jul. 2021, Art. no. 012037.
[11] B. Mandal, A. Okeukwu, and Y. Theis, "Masked face recognition using ResNet-50," 2021, *arXiv:2104.08997*.
[12] M. Almghraby and A. O. Elnady, "Face mask detection in real-time using MobileNetv2," *Int. J. Eng. Adv. Technol.*, vol. 10, no. 6, pp. 104–108, Aug. 2021.
[13] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (AI) applications for COVID-19 pandemic," *Diabetes Metabolic Syndrome Clin. Res. Rev.*, vol. 14, no. 4, pp. 337–339, 2020.
[14] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement*, vol. 167, Jan. 2021, Art. no. 108288. [Online]. Available: https://api.semanticscholar.org/CorpusID:220835781
[15] A. Chavda, J. Dsouza, S. Badgujar, and A. Damani, "Multi-stage CNN architecture for face mask detection," in *Proc. 6th Int. Conf. for Converg. Technol. (I2CT)*, Apr. 2021, pp. 1–8. [Online]. Available: https://api.semanticscholar.org/CorpusID:221738826
[16] M. A. Chyad, H. A. Alsattar, B. B. Zaidan, A. A. Zaidan, and G. A. Al Shafeey, "The landscape of research on skin detectors: Coherent taxonomy, open challenges, motivations, recommendations and statistical analysis, future directions," *IEEE Access*, vol. 7, pp. 106536–106575, 2019.
[17] U. Kumar, D. Arora, and P. Sharma, "Face mask detection using MobileNetV2 and VGG16," in *Proc. Int. Conf. Recent Trends Comput.*, Mar. 2023, pp. 669–677.
[18] I. B. Venkateswarlu, J. Kakarla, and S. Prakash, "Face mask detection using MobileNet and global pooling block," in *Proc. IEEE 4th Conf. Inf. Commun. Technol. (CICT)*, Dec. 2020, pp. 1–5.
[19] S. A. Sanjaya and S. Adi Rakhmawan, "Face mask detection using MobileNetV2 in the era of COVID-19 pandemic," in *Proc. Int. Conf. Data Anal. Bus. Ind., Way Towards Sustain. Economy (ICDABI)*, Oct. 2020, pp. 1–5.
[20] D. Hussain, M. Ismail, I. Hussain, R. Alroobaea, S. Hussain, and S. S. Ullah, "Face mask detection using deep convolutional neural network and MobileNetV2-based transfer learning," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–10, May 2022.
[21] R. Golwalkar and N. Mehendale, "Masked face recognition using deep metric learning and FaceMaskNet-21," *Appl. Intell.*, vol. 52, no. 1, pp. 13268–13279, 2022.
[22] M. Tan and Q. V. Le, "EfficientNetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[24] S. N. Karthik and S. Raj, "Face mask detection using MobileNetV2 and OpenCV," *Int. J. Eng. Technol. Manage. Sci.*, vol. 7, no. 4, pp. 376–382, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261501041
[25] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[29] Z. Wang, B. Huang, G. Wang, P. Yi, and K. Jiang, "Masked face recognition dataset and application," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 5, no. 2, pp. 298–304, Apr. 2023.
[30] Y. Suryawanshi, V. Meshram, V. Meshram, K. Patil, and P. Chumchu, "Face mask wearing image dataset: A comprehensive benchmark for image-based face mask detection models," *Data Brief*, vol. 51, Dec. 2023, Art. no. 109755.
[31] B. Roy, S. Nandy, D. Ghosh, D. Dutta, P. Biswas, and T. Das, "MOXA: A deep learning based unmanned approach for real-time monitoring of people wearing medical masks," *Trans. Indian Nat. Acad. Eng.*, vol. 5, no. 3, pp. 509–518, Sep. 2020.
[32] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270–64277, 2018.

YONGHUA MAO received the Ph.D. degree in computer science from the School of Computer Science, Xi'an Jiaotong University, in 2019. He was a Visiting Scholar for Dr. H. Zhou in computer engineering with North Carolina State University. He is currently a Master's Tutor with Xi'an Polytechnic University. His research interests include branch prediction, deep learning, and machine learning.

YUHANG LV is currently pursuing the master's degree with the School of Computer Science, Xi'an Polytechnic University. He has a computer science background in deep learning and artificial intelligence. His research interest includes virtual try-on.

GUANGXIN ZHANG is currently pursuing the master's degree with the School of Computer Science, Xi'an Polytechnic University. He has a computer science background in deep learning and artificial intelligence. His research interest includes virtual try-on.

XIAOLIN GUI (Member, IEEE) received the Ph.D. degree in computer science from Xi'an Jiaotong University, China, in 2001. Since 2008, he has been the Director of the Key Laboratory of Computer Network of Shaanxi Province, China. From 2009 to 2012, he was the Vice Head of the Department of Computer Science and Technology, Xi'an Jiaotong University. He is currently a Professor and serviced as the Deputy Dean of the School of Electronic and Information, Xi'an Jiaotong University. He also leads the Center for Grid and Trusted Computing (CGTC). His recent research interests include high performance computing, secure computation of open network systems, dynamic trust management theory, and development on community networks. He was a recipient of the New Century Excellent Talents in Universities of China. He was a recipient of the New Century Excellent Talents in University of China, in 2005.