IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 September 2023, accepted 17 October 2023, date of publication 6 November 2023,
date of current version 17 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3330242

RESEARCH ARTICLE

A Transfer Learning Approach for Facial Paralysis Severity Detection

WASIF ALI[1], MUHAMMAD IMRAN[1], MUHAMMAD USMAN YASEEN[1], KHURSHEED AURANGZEB[2], (Senior Member, IEEE), NOUMAN ASHRAF[3], AND SHERAZ ASLAM[4,5], (Member, IEEE)

[1]Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad 45550, Pakistan
[2]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[3]School of Electrical and Electronic Engineering, Technological University Dublin, Dublin, D07 EWV4 Ireland
[4]Department of Electrical Engineering, Computer Engineering and Informatics, Cyprus University of Technology, 3036 Limassol, Cyprus
[5]Department of Computer Science, CTL Eurocollege, 3077 Limassol, Cyprus

Corresponding authors: Wasif Ali (aliwasif072@gmail.com) and Nouman Ashraf (nouman.ashraf@tudublin.ie)

This Research is funded by Researchers Supporting Project Number (RSPD2023R947), King Saud University, Riyadh, Saudi Arabia.

ABSTRACT Facial paralysis is a debilitating condition that weakens or damages facial muscles resulting in asymmetric or abnormal facial movements. To aid in the diagnosis and rehabilitation of facial paralysis, researchers have developed machine learning and deep learning computer-aided diagnosis systems. However, machine learning models have limitations as they rely on facial landmark techniques and manual face palsy region extraction methods to obtain spatial information. Moreover, deep learning models need large, labelled datasets for training whereas existing available facial paralysis datasets are small and restricted. This presents significant challenges, including difficulties in data acquisition, insufficient patient numbers, and inadequate diversity within the datasets. These limitations can potentially restrict the generalizability of these models and introduce biases in the resulting outcomes. In this study, we propose an approach for the diagnosis and grading of facial paralysis comprised of two datasets, one from MEEI (Massachusetts Eye and Ear Infirmary) videos of patients and the other from the YFP (YouTube Face Palsy) dataset. The model uses a transfer learning approach to fine-tune the VGGFace model, which is pre-trained on facial images, on the prepared datasets for facial paralysis. The resultant model was subsequently renamed as FP-VGGFace for the purpose of this research. Additionally, two more pre-trained models on facial images, ResNet50 and VGG16, are also fine-tuned for the facial paralysis task. This was undertaken to conduct a performance comparison of multiple models on the prepared dataset. The findings indicate that the models exhibit high accuracy, benefiting from pre-training on a diverse dataset that enables the capture of spatial information from facial images. The FP-VGGFace model achieves the best accuracy (99.3%) and F1-score (99.3%) surpassing all benchmark models. This study underscores the potential of utilizing pre-trained deep learning models for the diagnosis and rehabilitation of facial paralysis.

INDEX TERMS Bell's palsy, deep learning, facial palsy, facial paralysis, transfer learning, VGGFace.

I. INTRODUCTION

Face paralysis is the abnormal movement of the facial muscles which are responsible for several functions in a face [1]. Paralyzed patients may exhibit asymmetric facial expression, eye blinking, speaking, and movement of mouth muscles [2]. Face paralysis is known as idiopathic facial palsy, bell's palsy, and facial nerve paralysis. Facial paralysis is normally caused by inflammation or infection of the head trauma, facial nerve, neck or head tumour, and stroke [3]. It causes sudden weakness and damage to one or both sides of facial muscles. There are 150,000 people affected by facial palsy in the United States every year [4]. Although, face palsy diagnosis and grading of the severity level of diseases is a crucial task. In the whole process, rapid and objective assessment [5] helps to choose the optimal rehabilitation

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

treatment by physicians. The Electromyography test is widely used to detect facial palsy and determine the severity of the nerve damage [1]. eFACE, House-Brackmann, and Sunnybrook scales assist the clinical method to categorize the severity level of face paralysis [5]. Mostly these grades are classified into *normal*, *near normal*, *mild*, *moderate*, *severe*, and *completely* paralyzed.

In existing studies, both machine learning (ML) and deep learning (DL) models are proposed for facial palsy detection and severity prediction. Most studies in the literature used ML algorithms to detect face paralysis with the help of facial landmarks [6], [7], [8]. These models first perform landmark detection using other models like the MEE shape predictor [4] and then pass it to the classifier for prediction. However, landmarks systems are only capable of detecting the shape of the paralyzed face and are unable to capture texture information [9]. Consequently, the accuracy of landmark detection impacts the performance of the models. Deep learning models typically use convolutional neural networks (CNN) to extract deep features from the facial images [10], [11]. They detect subtle changes and identify patterns from facial features which help in better detection of facial palsy. However, DL techniques are computationally expensive to train models from scratch. Moreover, these models are data-hungry, they need a large labelled dataset to train resourcefully [11], [12]. Those employing pre-trained models are based on natural images, which causes inadequate feature learning in the case of facial paralysis classification. Hence, they fail to produce high-quality results for facial paralysis tasks [10], [13]. Moreover, numerous computer-aided approaches employ specialized optical equipment and multidimensional imaging techniques to quantify facial paralysis [3], [14]. Although these methods have high accuracy they are expensive and complex for common use [14]. The aforementioned issues need a robust approach to diagnose and grade the severity level of facial paralysis.

The existing facial paralysis diagnosis studies are limited to small and private datasets. A limited dataset causes the issues of class imbalance, fewer subjects, and the lowest diversity in data. Consequently, it produces a less generalized model having biased results towards the minority class instances. In this study, we propose a model for the diagnosis of the severity level of paralysis in patients. We intend to solve the matter with a transfer learning approach. The VGGFace model is pre-trained on the VGGFace dataset, which has more than 2.6 million face images of 2622 individual identities. Additionally, the merger of two face palsy datasets, Massachusetts Eye and Ear Infirmary (MEEI) [5] and YouTube Facial Palsy (YFP) [15] datasets is the first time used to fine-tune the model. The MEEI face palsy images are extracted at the rate of three frames per second from MEEI videos of patients. The YFP dataset is extracted with six frames per second from videos of facial palsy patients. The fusion of both datasets achieves balanced instances, more subjects for each class, and diverseness in the dataset.

The proposed novelties in this article make our approach more robust, efficient, and reliable in classifying face palsy in patients. This work presents several key contributions in the field of facial paralysis diagnosis and grading.

- To obtain more accurate and detailed feature extraction from facial paralysis images, we use a transfer learning approach to leverage the extracted features from the VGGFace model, allowing us to capture spatial texture information more effectively.
- A novel dataset is created by combining two existing datasets, namely MEEI and YFP. This approach increases the number of subjects and introduces diversity to the dataset, which improves the robustness and accuracy of the model.
- A comparative analysis is conducted on various pre-trained deep learning models trained on facial images to determine the optimal approach for facial paralysis classification.

These contributions offer a promising solution for addressing the limitations of existing methods and improving the accuracy and robustness of facial paralysis diagnosis and grading. The remaining paper is organised as follows. Section II explores the literature on automatic facial palsy detection and severity classification. Section III provides the dataset and methodology details. Section IV presents experimental settings. Results are analysed in Section V. Section VI discusses the findings of this work. Finally, Section VII presents the conclusion of the paper and provides future work directions.

II. RELATED WORK

Computer-aided diagnosis methods for facial paralysis detection have been created using both ML and DL techniques, according to the literature. These techniques are found to be especially useful in strengthening the efficacy of rehabilitation treatments as well as accelerating the process of diagnosis. The literature on the diagnosis and rehabilitation of facial paralysis can be broadly categorised into two dimensions: machine learning and deep learning. Fig. 1 shows the taxonomy of the ML and DL techniques used for facial palsy detection in the literature.

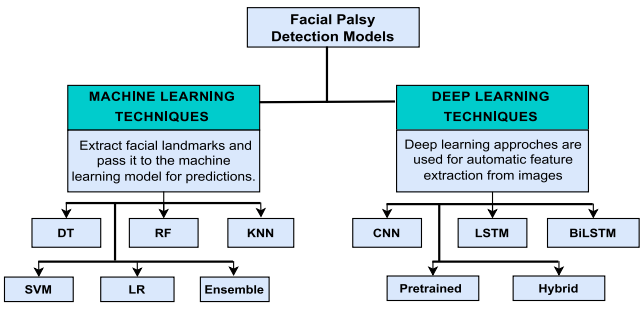


FIGURE 1. Dimensions of the literature review for facial paralysis detection. DT: decision tree, RF: random forest, KNN: K-nearest neighbour, SVM: support vector machine, LR: logistic regression, CNN: convolutional nerural network, LSTM: long short-term memory, BiLSTM: bidirectional long short-term memory.

A. MACHINE LEARNING MODELS

Machine learning approaches rely on facial landmark techniques for diagnosing facial paralysis in the affected regions of the face. The key points around the eyes and mouth are particularly useful for measuring the severity level of paralysis. Several algorithms have been designed to automatically identify the crucial key points on the face that are necessary for diagnosing facial paralysis.

For example, authors in [8] proposed ML-based approaches for paralysis detection using facial landmarks techniques. The facial landmark technique is used for feature extraction and the model selects ten significant features using the Gini index. These features are then fed to logistic regression (LR) and support vector machine (SVM) models. The SVM performed better than LR and achieved an average accuracy of 76.87%. However, the model accuracy is significantly low. Furthermore, the dataset used for training is limited and has only 757 healthy faces and 717 paralyzed faces.

In this study [16], an ensemble approach with five SVMs and a rule-based classifier is proposed to predict the severity level of facial palsy patients. Kinect V2 library is used to extract 32 landmark points on the face which are then passed to the symmetry analysis unit for feature extraction. The model was tested on the augmented dataset of originally 13 facial palsy patients with 375 images and 50 healthy participants with 1650 records. The proposed model achieved 96.8% accuracy in detecting the correct face palsy class.

The work in [3] used three ML models: decision tree (DT), SVM and an ensemble approach to grade the severity of the facial paralysis. A dynamic three-dimensional stereo photogrammetry imaging system is used to capture asymmetry in facial images. From these images facial landmarks are extracted that help to measure facial expression on images dataset. The model was simulated on 16 facial palsy patients and it achieved the best accuracy of 91.1% for the SVM classifier. However, the accuracy and number of subjects represent inadequate model performance.

The study in [17] detects bell's palsy by identifying dissimilarity between the blinking patterns of eyes. The proposed method calculates blink similarity between two eyes and passes the extracted features to a range of ML classifiers. The stochastic gradient descent (SGD) displayed the best accuracy of 94.7% among other ML models. Although this model achieves better accuracy with the limited dataset, it only considers the eye portion for palsy detection. It may fail if the video inadequately captures the eye part of the face.

The authors in [1] performed facial paralysis classification using images captured with laser speckle contrast imaging (LSCI). First, LSCI is employed to generate RGB colour images and blood flow images of the paralyzed faces. Then, the patient's face is divided into areas of concern using an improved segmentation approach to assess the facial blood flow. The three HB score classifiers are used to determine the facial paralysis severity. The model achieved 97.14% accuracy which is relatively better as compared to other ML models.

In the article [18], multiple ML models are employed for facial palsy classification in three severity levels. The image dataset is preprocessed using facial landmarks and it measures the symmetry and asymmetry between face sides in images. The images are then fed to the SVM model for severity prediction which generates the best accuracy of 95.58%. However, the accuracy of the approach relies on the careful collection of data, which involves ensuring consistency in the angle of the face and the camera used to capture the images.

The study in [7] refers to machine learning models to extract images, acquire facial landmarks, and compute the feature values to train the model and predict outcomes for new patients. The study compares RF and SVM classifiers to evaluate their effectiveness in the proposed method. The RF performed better than SVM and achieved an overall accuracy of 88.9%. However, the study uses a limited number of subjects and excludes texture features while feature learning.

The overall findings indicate that ML approaches are not suitable for facial palsy detection as their dependence on facial landmark techniques for feature extraction generates poor facial palsy detection results.

B. DEEP LEARNING MODELS

Deep learning approaches are particularly effective for automatic feature extraction from images using convolutional operations. Multiple deep learning models have been developed to automatically extract features from facial images. These models are capable of identifying patterns and subtle changes in facial expressions that may indicate the presence of facial paralysis.

For example, the article [10] proposed a CNN model for the classification of facial palsy images in five different severity levels. The authors used the generative adversarial network (GAN) model to address the class imbalance issue in the dataset. However, the synthetic images generated by GANs do not retain the same level of quality as the real images which negatively impacts the performance of the model. As a result, the model produces a low accuracy of 92.60%.

The work in [19] proposed 3DPalsyNet for grading facial palsy and detecting facial emotion using a 3D CNN architecture. Dynamic actions were captured by ResNet and served as the foundation for recording the video data's dynamic behaviours. Two datasets were used for training, one with face palsy images and the other for facial expressions and the model achieved significantly low accuracies of 82% and 86% respectively.

Authors in [20] proposed a hierarchical framework comprising three components to detect the local palsy region of facial palsy patients. The first one detects faces, the second component performs landmark detection and the final component detects facial palsy regions. To train and evaluate the model YouTube faces facial palsy dataset of 21 patients

was collected and labeled by the neurologists. The CK+ dataset was used to enhance the dataset size for normal images. The model achieved precision and recall of 93% at recall 88% respectively.

In [21], the authors proposed a transfer learning approach for predicting the stress state of facial nerve paralysis (FNP) patients. An FNP dataset is created with the facial emotion expressions of paralyzed patients. The prepared dataset is used to fine-tune the VGGNet model, which is pertained on the facial emotion recognition (FER2013) data set. The results generate an accuracy of 66.58% for the VGGNet model, which is significantly less mainly due to the low-quality dataset.

The work in [12] proposed an automatic facial paralysis detection model using a cascaded encoder architecture. The first encoder generates spatial information about facial attributes depicted through semantically segmented facial regions in input images. The second encoder performs an assessment of facial paralysis based on the extracted spatial information. The model produced a facial palsy detection accuracy of 95.60% which can be improved by employing better feature extraction techniques.

To categorize the facial nerve function in facial areas, a region-based parallel convolutional network model, named TPCNN, is presented in [11]. The dataset used in the study is small and unbalanced, hence the model utilized strategies such as pre-padding, data augmentation, and optimal weights for each class during the training phase. The model achieved a classification accuracy of 69% to differentiate between normal faces and faces with facial nerve palsy.

The study in [13] used deep convolutional networks for feature extraction for the assessment process of unilateral peripheral facial paralysis (UPFP). The GoogLeNet, pre-trained on non-biological images, was used for fine-tuning on the UPFP dataset. After fine-tuning the model accurately differentiated House-Brackmann (HB) degrees on the provided image dataset with 91.25% accuracy.

In [14], authors proposed a hybrid approach by combining CNN and LSTM models to quantitatively assess the grade for face palsy on the YFP dataset. First, the model captures the image sequence from the video and performs face detection and image segmentation on the paralysis area. The local and global CNN models perform feature extraction, which after concatenation are passed to LSTM. The LSTM output feature vector is passed to the dense layer for classification using the softmax activation function. The model achieves an accuracy of 94.8% and beats other state-of-the-art models.

In this study [9], authors focus on asymmetric facial muscle movement during the facial paralysis diagnosis task. For that, it uses double-path LSTM which employs deep differentiated networks to extract global and local feature vectors. The resultant feature vectors and then concatenated and fed to a dense layer for classification. However, the proposed model achieves low accuracy of 73.4% which can be easily improved using the transfer learning approach.

Authors in [22] proposed a Bi-LSTM-based methodology for identifying facial weakness in videos. The model was tested on an ''in-the-wild'' video dataset validated by neurologists and emergency medical service personnel. The results beat various facial weakness detection options already in use with an accuracy of 94.3%. However, the achieved accuracy is quite low for medically critical applications and needs to be improved.

The triple-stream LSTM was proposed by authors in [23] to automatically assess the severity of facial paralysis. For each facial action, the model first performs preprocessing on the video and splits it into three overlapping videos. Then, triple-stream LSTM extracts face features from the video segments of several facial regions simultaneously. These features are subsequently concatenated through parameters automatically learned by the network. The average accuracy of all facial actions is 86.37%. When working with large datasets, using a three-stream LSTM model is computationally expensive.

The study in [24] proposed a hybrid deep learning model for facial nerve paralysis detection using DeepID and Inception-v3 pre-trained model. The hybrid approach helped to achieve an accuracy of 97.5% even with a small dataset of 1049 facial images. However, retraining the model with a bigger dataset may improve the results even further.

Based on the literature review, the following research directions are set for this study. The existing studies mainly use limited datasets; the proposed model should use a relatively bigger dataset. The deep learning model learns spatial information from facial images and generates better results. If coupled with transfer learning, deep learning models can demonstrate better performance even with the limited data. The proposed model attempts to improve the severity prediction accuracy using the transfer learning approach.

III. PROPOSED METHODOLOGY

The proposed methodology for facial palsy severity grading is depicted in Fig. 2. The model takes two facial palsy datasets: MEEI and YFP and merges them to create a more diverse and robust dataset. To balance the merged dataset for normal faces, CK+ dataset is combined with the MEEI & YFP datasets to create a balanced dataset. In the third step, data preprocessing operations are performed to improve model accuracy and reduce computation overhead. The fourth step describes the transfer learning approach of using the pre-trained VGGFace model and fine-tuning it for the facial paralysis task. Finally, the feature vector is fed to the classifier for detecting the severity of facial paralysis.

A. DATA ACQUISITION

The development of effective diagnosis techniques heavily relies on the availability of suitable datasets. However, the datasets used in previous studies are mainly private and non-shareable, which negatively impacts the progress of research in this field. Obtaining appropriate facial paralysis datasets is a challenging task that often requires requests to multiple

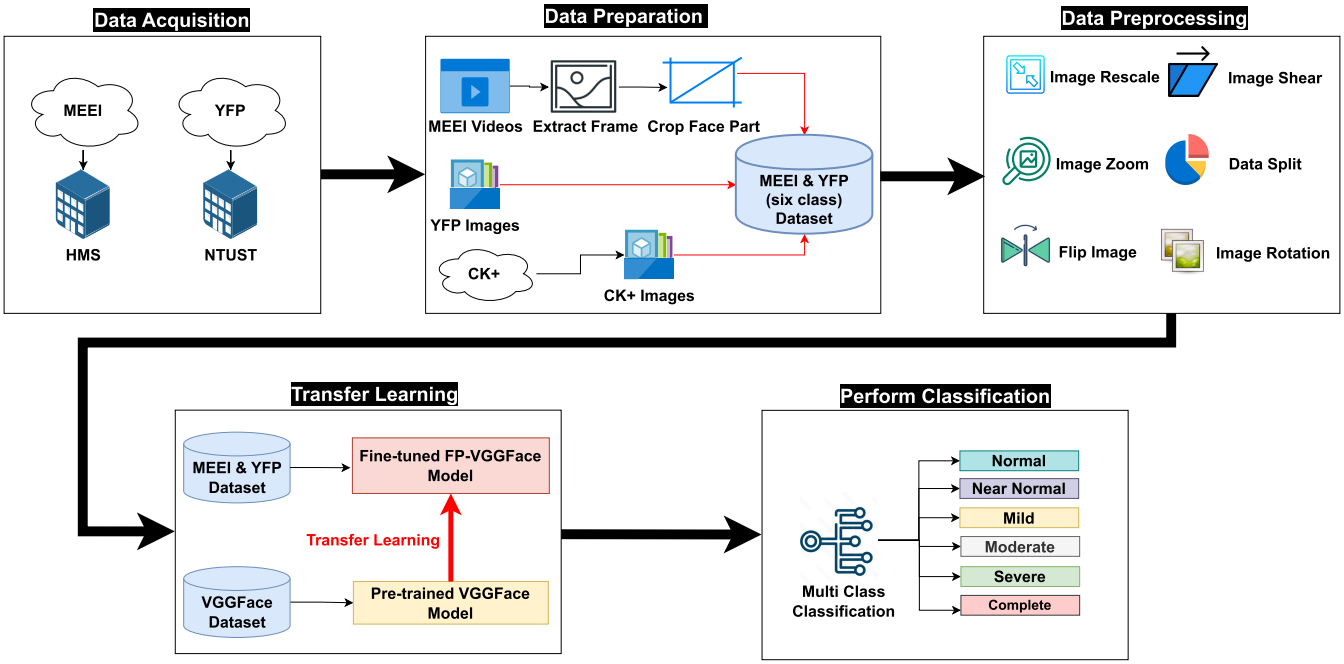


FIGURE 2. Illustration of the proposed model for facial paralysis severity detection. HMS: Harvard Medical School, NTUST: National Taiwan University of Science and Technology.

sources and publications. In our effort for suitable datasets, we are fortunate enough to acquire two datasets.

We obtained two facial paralysis datasets, namely Massachusetts Eye and Ear Infirmary (MEEI) [5] and YouTube Facial Palsy (YFP) [15] datasets. The MEEI is a private dataset and can be requested by sending an email to d.guarinlopez@ufl.edu. On the other hand, the YFP dataset can be obtained from the following GitHub link: https://github.com/AvLab-CV/YouTube-Facial-Palsy-Database. The first dataset, MEEI, was gathered by Jacqueline J. Greene (MD) and his team from the MEEI, Harvard Medical School, Boston, USA and is clinically approved by the Facial Nerve Center at MEEI. The second dataset, YFP, was acquired by Gee-Sern Jison Hsu (Professor) and his colleagues from the National Taiwan University of Science and Technology. Access to these datasets enables the development of more accurate and efficient techniques for the diagnosis of facial paralysis and benefiting patients. Please note that these datasets are not publicly available and require specific permissions or requests for access.

B. DATASET PREPARATION

The facial paralysis dataset consists of MEEI, YFP, and CK+ datasets. Fig. 3 shows the complete process of the dataset preparation. From the MEEI videos, it captures frames from the images. The face part of the images is cropped to extract the relevant portion of the image. YFP and CK+ image datasets are merged with the MEEI dataset. Data preprocessing and data augmentation operations are applied on the merged dataset to add diversity to the dataset. Finally, the dataset is arranged into six classes based on the severity levels of facial palsy. The subsequent subsections provide details about the individual operations performed during the data preparation phase.

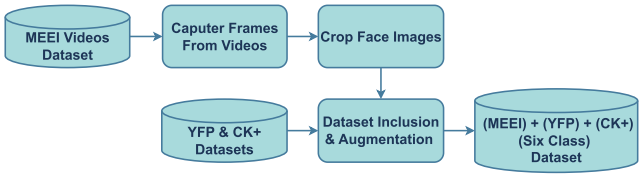


FIGURE 3. Dataset preparation steps used to generate a target facial paralysis dataset.

1) MEEI DATASET

The MEEI data repository contains videos of 60 subjects with six severity levels of facial paralysis measured by the eFACE scale [5]. In each video, the patients pretend to perform some expressions and repeat dialogues that help illustrate the palsy region. We extracted frames from the videos at a rate of three frames/second and cropped the face parts from images using the Dlib library, a popular software tool for machine learning and computer vision tasks. Specifically, we captured frames from videos and performed face cropping on each frame to remove other identifying information from the image, such as the background. By focusing on the face part of each image, we can simplify the task of facial paralysis detection and classification by removing extraneous information and focusing on the most relevant part of the image. Fig. 4 shows the samples before and after the face cropping.

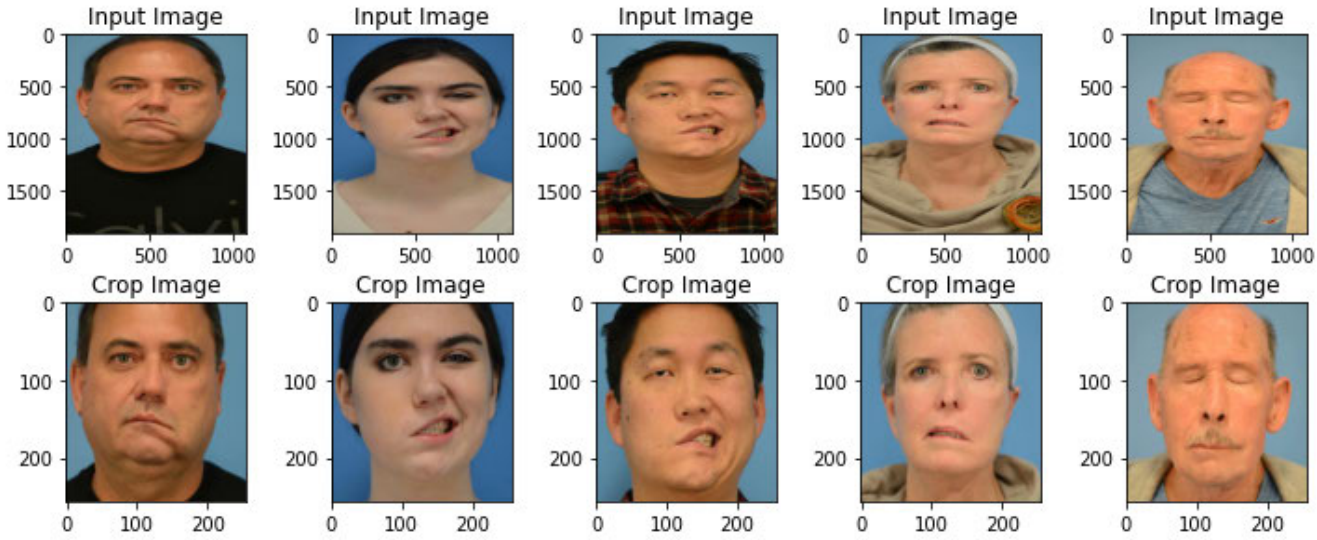


FIGURE 4. Face cropping operation depicted over MEEI (Massachusetts Eye and Ear Infirmary) dataset [5].

Face cropping and image resizing can improve the accuracy of the task of facial paralysis detection and severity classification. The severity levels of both flaccid and non-flaccid categories of facial palsy are merged into *normal*, *near normal*, *mild*, *moderate*, *severe*, and *complete* classes.

2) YFP DATASET

To increase the number of subjects, we merged the YFP dataset with the MEEI dataset. The YFP dataset consists of 32 videos featuring 21 patients diagnosed with facial palsy. The patients were instructed to speak and perform various facial expressions while being captured by the camera. Each video was transformed into a sequential collection of images at a sampling rate of six frames per second. Some sample images of the YFP dataset are provided in Fig. 5.

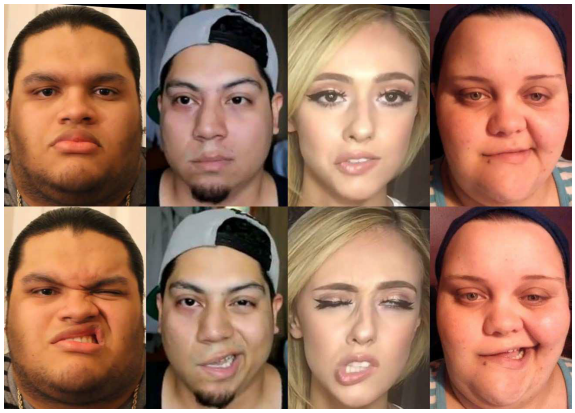


FIGURE 5. YFP (YouTube Face Palsy) dataset sample images [15], [20].

Subsequently, the resulting frames are labeled using the House-Brackmann scale [13], a widely employed grading system for evaluating the severity of facial paralysis. The scale employs grades I, II, III, IV, V, and VI to signify different levels of facial nerve dysfunction or paralysis. Grade I represents *normal*, grade II represents *near normal*, grade III signifies *mild* paralysis, grade IV is for *moderate*, grade V indicates *severe* and grade VI represents *complete* paralysis. These grades enable us to represent distinct levels of impairment experienced by the patients. The scores on this scale provide valuable insights into the extent of impairment faced by the individuals. The YFP dataset was highly imbalanced, with a huge variance in image count per subject. Therefore, we used limited images from each subject to create a well-balanced and diverse dataset.

3) CK+ DATASET

To address the issue of class imbalance between normal and paralyzed images, we used the CK+ dataset. The CK+ dataset is a normal images dataset which contains facial images of healthy human beings. It was originally extracted from 593 videos of 123 different people in the age range of 18 to 50 years using 30 frames per second. The CK+ dataset is a widely used dataset for facial expression classification experiments. In this study, we have used 350 images from the original dataset to balance normal facial images in our prepared dataset.

4) DATA PREPROCESSING

Data preprocessing is required in the merged dataset to improve its quality by performing image preprocessing and augmentation operations. The cropped images in the prepared dataset are resized to a standard size of 224 $\times$ 224 for input to the transfer learning model. Further, they are rescaled in the range 0 to 1 by dividing all pixel values by 225. Rescaling is used to normalize the image data in the standard range of 0 to 1. Some classes have more images than

others whereas certain images in the dataset are noisy and represent inadequate information about facial palsy. First, noisy images are removed from the dataset. Then, to achieve the balance among all the classes additional images were generated using data augmentation techniques. Traditional image augmentation techniques such as shearing, flipping, rotation, and zooming are used for data augmentation. The aim is to achieve the same number of images in each facial palsy class to increase the diversity of the training data and improve the generalization ability of the model. Fig. 6 illustrates the augmentation steps by taking a sample original image as an input. These augmentation steps are crucial in preparing a high-quality dataset for training an accurate facial paralysis classification model. Both image preprocessing and augmentation operations are performed using the Keras library *ImageDataGenerator* class. Table 1 describes the specifics of the operations performed during the augmentation and preprocessing phases.

TABLE 1. Description of the data preprocessing and data augmentation steps perfomed in the study.

Preprocessing & Augmentation steps	Description
Resizing	Resize images of the merged dataset to 224x224.
Rescaling	Rescale images by dividing all pixel values by 225.
Shearing	Apply shear transformation with a shear range of up to 0.2.
Flipping	Apply horizontal flip to augment the dataset.
Rotation	Apply rotation transformation with a rotation range of up to 20%.
Zooming	Apply zoom transformation with a zoom range of up to 0.2.

5) DATASET DISTRIBUTION

We created a balanced facial palsy dataset with six classes representing the six severity levels of facial paralysis. The prepared dataset consists of 6,600 images divided into six different classes. Each class contains 1100 images, with 770 images allocated for training, 165 images for validation, and 165 images for testing. Consequently, we had a total of 4620 images dedicated to training and 1980 images equally divided for validation and testing, resulting in a distribution ratio of 75% for training and 15% for both validation and testing. Table 2 provides specifics of the distribution.

TABLE 2. Facial paralysis dataset organization into training, validation, and testing sets.

Severity classes	Training	Validation	Testing	Total
Normal	770	165	165	1100
Near Normal	770	165	165	1100
Mild	770	165	165	1100
Moderate	770	165	165	1100
Severe	770	165	165	1100
Complete	770	165	165	1100
Total	**4620**	**990**	**990**	**6600**

C. TRANSFER LEARNING

To automatically extract features from facial paralysis images, we employed a transfer learning approach that addresses the issue of data scarcity. This study possesses a unique aspect, focusing on the utilization of pre-trained models such as VGGFace [25], ResNet50 [25], and VGG16 [25] which are originally trained on facial images, as opposed to natural images. The reason for selecting these models is their availability as facial images pre-trained models among all best-performing models available on Keras applications. This approach enables us to learn discriminative features that can aid in extracting pertinent information from paralyzed facial images. VGGFace, being a well-generalized model for face recognition and detection, serves as our primary choice.

Inspired from [26] and [27], transfer learning for this work is formally defined as:

Definition 1: "*Given any source domain D_S trained on a learning task T_S, and a target domain D_T for a learning task T_T, transfer learning aims to improve the learning of a target function f_T in D_T by using learned knowledge from D_S, provided that $D_S \neq D_T$ or $T_S \neq T_T$.*"

In our facial palsy severity prediction work, VGGFace is the source model D_S and the FP-VGGFace is the target model D_T. Since both domains are different but have overlapping features, hence, transfer learning logically applies to our target learning task T_T. Fine-tuning is performed to adjust the parameters of pre-trained models for our facial paralysis task. Fine-tuning involves adapting the parameters of the pre-trained model to extract relevant features for the new task. In the context of feature extraction from images, we used the pre-trained model as a feature extractor by passing images through the network and extracting features from one of its intermediate layers. These extracted features are then used as input to a model that is for training on the new task of classifying facial paralysis. Fine-tuning is done by further training the entire pre-trained model on the new task or by training only the last few layers of the model while keeping the earlier layers frozen.

FP-VGGFace CNN is used in this study to extract features from facial paralysis images. FP-VGGFace is a modified version of VGGFace [28] that was trained specifically for extracting features from faces. VGGFace consists of 11 blocks, each of which starts with a linear operator and then includes one or more non-linearity, such as max pooling and ReLU. Since the linear operator in the first eight of these blocks is a bank of linear convolution filters, they are referred to as convolutional blocks. The last three blocks are fully connected layers and they resemble the convolutional layers. These layers take input feature maps from the preceding convolutional blocks and capture high-level relationships between the features. A ReLU rectification layer is placed after every convolution layer. The first two fully connected layers' output has 4,096 dimensions, and the last fully connected layer generates an output of 2,622 dimensions. Hence, the VGGFace model is trained to identify 2,622

FIGURE 6. Illustration of the dataset augmentation techniques over a sample image [5].

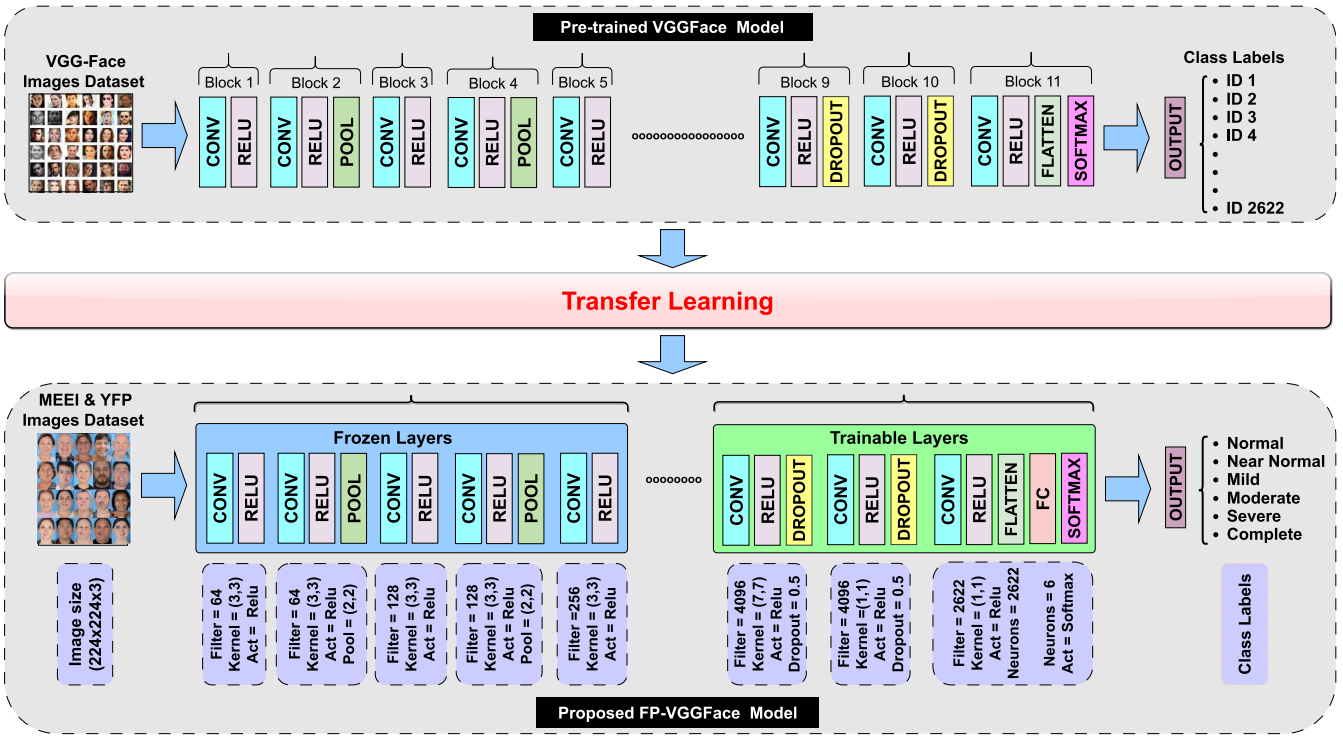


FIGURE 7. Illustration of transfer learning approach for proposed FP-VGGFace model.

unique individual faces. A softmax layer is used to evaluate the class probability for the result.

FP-VGGFace is the fine-tuned version of VGGFace on the prepared dataset using a transfer learning approach. The fine-tuning process uses the frozen convolution base with a trainable head. The frozen base refers to the concept of freezing the initial layers of the model, which means they are not further trained. This approach leverages the prior knowledge gained from a previous task. On the other hand, the trainable head refers to the last layers of the model that are trained specifically for the new task of multiclass classification of facial paralysis. The final layer of the model is customized based on the number of classes. Furthermore, the classification heads of the models are also adjusted accordingly. This architecture allows for the reuse of spatial information extracted from face images to accurately classify facial paralysis in patients. The architecture of the proposed methodology for facial palsy classification is shown in Fig. 7.

Algorithm (1) defines the transfer learning approach on the prepared dataset for facial paralysis severity detection. The input N represents the merged facial paralysis dataset, and W represents the pre-trained weights of the VGGFace model. The goal is to utilize this input to build a model using transfer learning and obtain evaluation matrices as output. The algorithm's output includes the model itself and the evaluation matrices, which provide valuable insights into the model's performance. In step 3, the facial paralysis dataset N is split into training and testing sets, denoted as X_{train}, Y_{train}, X_{test}, and Y_{test}, respectively. This division allows the evaluation of the model during the training and testing phases.

The pre-trained VGGFace model is denoted as M in the algorithm (step 4). The model includes the pre-trained weights W obtained from the training process on the

Algorithm 1 FP-VGGFace Transfer Learning

1: **Input** ← N represents the prepared facial paralysis dataset. W represent the training weights of VGGFace model pre-trained on VGGFace dataset.
2: **Output** ← Model and evaluation matrices.
3: $(X_{\text{train}}, Y_{\text{train}}, X_{\text{test}}, Y_{\text{test}})$ ← splitDataset(N)
4: M ← VGGFaceModel(W)
5: $\hat{M}$ ← M+ adjustLastLayers()
6: FP-VGGFace ← freezingModelBase($\hat{M}$)
7: H ←setHyperparameters()
8: **while** (model not converged) **do**
9: $train_{\text{acc}}$ ← [] ▷ List for training accuracy
10: $train_{\text{loss}}$ ← [] ▷ List for training loss
11: val_{acc} ← [] ▷ List for validation accuracy
12: val_{loss} ← [] ▷ List for validation loss
13: $train_{\text{acc}}, train_{\text{loss}}, val_{\text{acc}}, val_{\text{loss}}$ ← FP-VGGFace.Fit($H, X_{\text{train}}, Y_{\text{train}}, X_{\text{val}}, Y_{\text{val}}$)
14: $Pred_{\text{label}}$ ← modelPredict($X_{\text{test}}, Y_{\text{test}}$)
15: plotAccuarcy($train_{\text{acc}}, val_{\text{acc}}$)
16: plotLoss($train_{\text{loss}}, val_{\text{loss}}$)
17: $True_{\text{label}} = Y_{\text{test}}$
18: plotConfussionMaterics($True_{\text{label}}, Pred_{\text{label}}$)
19: plotROC($True_{\text{label}}, Pred_{\text{label}}$)
20: **end while**

VGGFace dataset. This initialization serves as a starting point for fine-tuning. The last layers of the model M are adjusted to align with the number of classes in the new task, denoted as $\hat{M}$ (step 5). This adjustment is performed using the adjustLastLayers() function, ensuring compatibility between the model and the facial paralysis classification problem. The freezingModelBase() function is used to transfer the weights from the previous task and prevents the model's base layers from being trained during the training process. This step helps retain the valuable pre-trained features while allowing the head, which includes the last several layers and the final dense layers, to be trained on the facial paralysis dataset. The fine-tuned model, named as FP-VGGFace, is saved in step 6, for training and testing in subsequent steps of the algorithm. Hyperparameters H are set to specify various aspects of the model training process, such as learning rate, optimizer, and batch size. These hyperparameters are fine-tuned to optimize the model performance.

The algorithm calculates various metrics to assess the model's performance. These include training accuracy ($train_{acc}$), training loss ($train_{loss}$), validation accuracy (val_{acc}), and validation loss (val_{loss}). These metrics provide insights into the model's ability to learn and generalize from the data. Equation (1) provides mathematical formula for calculating categorical cross entropy loss.

$$train_{loss}, val_{loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1_{y_i \in C_c} log p_{model}[y_i \in C_c] \tag{1}$$

where N represents the number of observations and C shows the number of classes. $1_{y_i \in C_c}$ indicates membership of the sample i for a category c. $p_{model}[y_i \in C_c]$ is the probability predicted for the sample i belonging to the class c. The model FP-VGGFace is trained using the hyperparameters H for 25 epochs or is stopped if it fails to improve. The FP-VGGFace.Fit() function is used to update the model's performance metrics, including $train_{acc}$, $train_{loss}$, val_{acc}, and val_{loss} (step 13). After training, the proposed model FP-VGGFace is utilized to predict the labels ($Pred_{label}$) for the test set (X_{test}, Y_{test}) (step 14). These predictions serve as the basis for evaluating the model's performance on unseen data. The algorithm visualizes the training and validation accuracy ($train_{acc}$ and val_{acc}) as well as the training and validation loss, ($train_{loss}$ and val_{loss}). These visualizations help assess the model's learning progress and identify any signs of overfitting or underfitting. The true labels ($True_{label}$) for the test set (Y_{test}) are set, and a confusion matrix is plotted using the plotConfusionMatrix() function. This matrix provides a comprehensive overview of the model's predictions, enabling detailed analysis of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Finally, plotROC() performs receiver operating characteristic (ROC) curve analysis using area under the curve (AUC) metric.

In order to handle a multi-class classifier, we modified the model's last layer and employed a softmax activation function to categorize between six classes. This categorization represents the severity levels of paralysis, including *normal*, *near normal*, *mild*, *moderate*, *severe*, and *completely* paralyzed. The softmax function is provided in Equation (2) which normalizes a vector of output from the previous layer's perceptrons and converts it into a probability distribution with values between 0 and 1. It calculates the likelihood that an input belongs to each of the potential classes. The target class is determined based on the highest probability.

$$\text{Softmax}(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} \tag{2}$$

where x is the input vector, e^{x_i} is the exponential of the input vector and C is the number of classes in the multi-class classifier.

IV. EXPERIMENTAL SETTINGS

For training deep learning models, we utilize Python 3.6 and Google Colab GPUs. TensorFlow and Keras, popular open-source libraries for building deep neural networks, are used throughout the study. The implementation of the baseline models and the overall code is done using Keras library.

A. EVALUATION METRICS

The evaluation metrics used in this study are the standard statistical measures to evaluate the performance of deep learning models. These include accuracy, precision, recall and F1-score. The mathematical expressions to evaluate their

values are given in Equations (3), (4), (5), (6).

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

$$Precision = \frac{TP}{(TP + FP)} \tag{4}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{5}$$

$$F1 - score = \frac{(2 * Precision * Recall)}{(Precision + Recall)}, \tag{6}$$

These measures are evaluated from the confusion matrix and they represent correct and incorrect classifications of the model, thus depicting the overall model performance. TP represent the correct severity class prediction, and TN shows the correct healthy classification. FP is incorrectly classified facial palsy patient and FN is incorrectly classified as a healthy person.

B. HYPERPARAMETERS

The hyperparameter value selection uses a grid search mechanism where we systematically try different combinations of values to judge the model performance. As deep learning models are non-deterministic and exact outcomes cannot be predicted, a series of experiments were run with different hyperparameter values. These hyperparameters include learning rate, epoch, optimizer, loss function, dropout and trainable layers. The loss function is always categorical cross entropy for multiclass classification experiments. For FP-VGGFace and VGG16 models, the learning rate is chosen as 0.001 and the optimizer is set as Adam. For the ResNet50 model, the learning rate is 0.0001 and the optimizer is stochastic gradient descent (SGD). The remaining details about the hyperparameters are provided in Table 3.

TABLE 3. Hyperparameter settings for the experimentation.

Hyperparameter	FP-VGGFace	VGG16	ResNet50
Learning Rate	0.001	0.001	0.0001
Epoch	25	25	25
Optimizer	Adam	Adam	SGD
Loss Function	Categorical Cross Entropy	Categorical Cross Entropy	Categorical Cross Entropy
Dropout	-	0.6	0.5
Trainable Layers	10	8	150

V. RESULTS AND ANALYSIS

In this section, we present graphical results that represent the model accuracy and loss for both the training and validation phases. It also presents confusion matrices that show the accuracy achieved by each model on the test set. The receiver operating characteristic (ROC) curve demonstrates the classification performance of the models using area under the curve (AUC) metric.

A. FP-VGGFACE MODEL

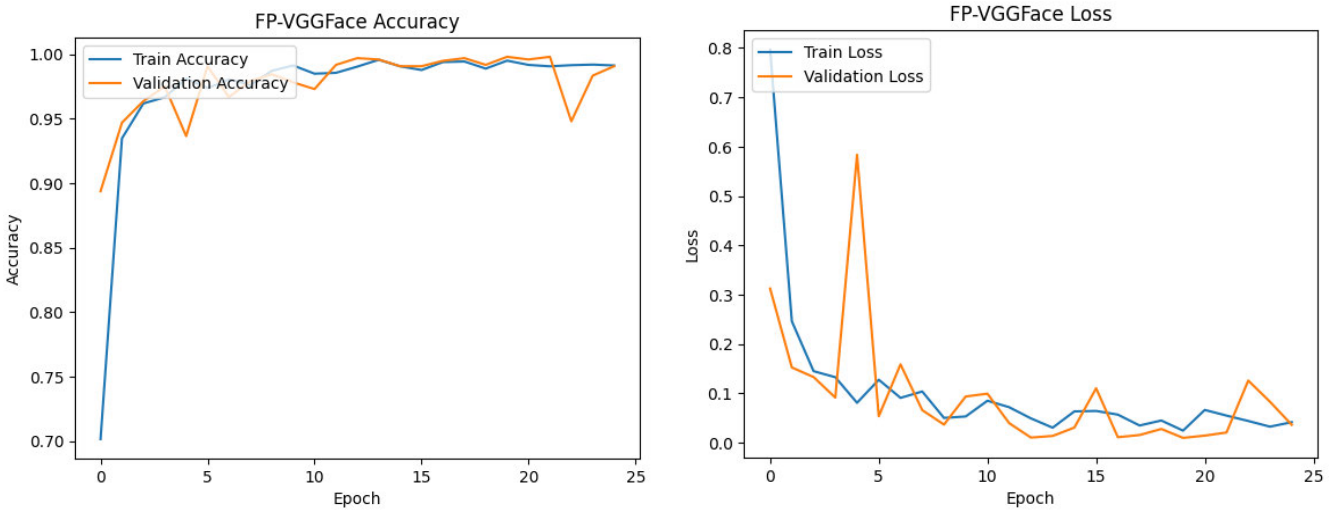
Fig. 8 displays the accuracy and loss curves for training and validation, demonstrating the performance of the proposed FP-VGGFace model on the prepared dataset. The blue curve represents training accuracy, and the orange curve shows validation accuracy. Training accuracy demonstrates a gradual learning trend of the model by progressively improving the accuracy with the increase in the number of epochs. The model starts from 0.71 accuracy and attains a maximum value of 0.98 after 20 epochs. Validation accuracy closely follows the training accuracy and achieves an overlapping value of 0.98 after 25 epochs except for some spikes. At certain epochs (e.g., 20) validation accuracy also becomes superior to training accuracy. The curves did not change afterwards and hence training was stopped after 25 epochs.

The loss curves for both training and validation also follow a similar pattern and gradually decrease as the number of epochs increases. The two curves illustrate the loss of the model on the training and validation datasets, respectively. The training loss is the value of the loss function on the training set at each epoch of training. The validation loss is the loss function's value on the validation set at each epoch of training. Furthermore, it indicates the model's capacity to generalize on unseen data. The validation loss beats training loss at intermediate epochs, which demonstrates the model's capability to perform well for unseen data. The curves show the model's tendency to learn the patterns from the facial palsy dataset by gradual improvement of the training and validation accuracies and gradual reduction of the training and validation loss.

To assess the performance of the FP-VGGFace model on unseen data, we simulated it on the test dataset. Fig. 9 shows the model results using the confusion matrix and AUC-ROC curve. These results are achieved after retraining 10 layers of the VGGFace model and retaining the remaining weights from the pre-trained model. The model performed on the optimized hyperparameter values on the prepared dataset. The confusion matrix provides the summary of both correct and incorrect predictions while the AUC-ROC curve demonstrates the classification accuracy using the AUC metric. The confusion matrix displays superior model performance as the majority of datapoints are TP and only a few are FN or FP. Hence, the misclassification number is negligible and a maximum of only three samples for *severe* class are misclassified while the majority are correctly classified. The ROC curve validates the results of the confusion matrix and remains closer to the top left corner for all classes of the model. The ideal value of AUC is 1 and the FP-VGGFace model achieves it except for *severe* and *near normal* classes, which shows the best performance of the proposed model.

B. RESNET50 MODEL

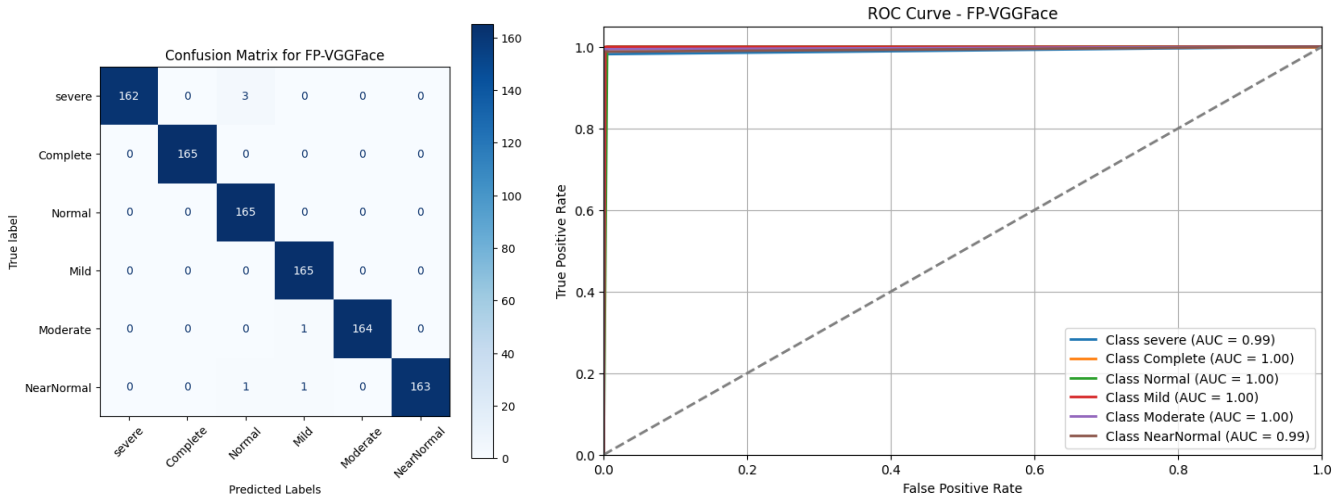
The training and validation performance of the fine-tuned Resnet50 model on our prepared dataset is depicted in Fig. 10. The graphs consist of training and validation curves for accuracy and loss. These curves depict progressive learning but irregular spikes demonstrate model challenges in



(a) Training and validation accuracy.

(b) Training and validation loss.

FIGURE 8. **Training and validation performance analysis of the proposed FP-VGGFace model using accuracy and loss curves.**



(a) Confusion matrix of FP-VGGFace model.

(b) AUC-ROC curve analysis for FP-VGGFace model.

FIGURE 9. **Proposed FP-VGGFace model performance analysis on test data using confusion matrix and AUC-ROC curve.**

acquainting with the target facial paralysis dataset. The high value of the training accuracy curve means that the model fits on the training data with an increase in epochs number. The training accuracy progresses gradually; however, validation accuracy struggles to keep pace and hence drops below the training accuracy for epochs number 15 to 20. Although it matches with the training accuracy subsequently at 25 epochs and remains same afterwards but the results display model difficulty to learn facial paralysis features.

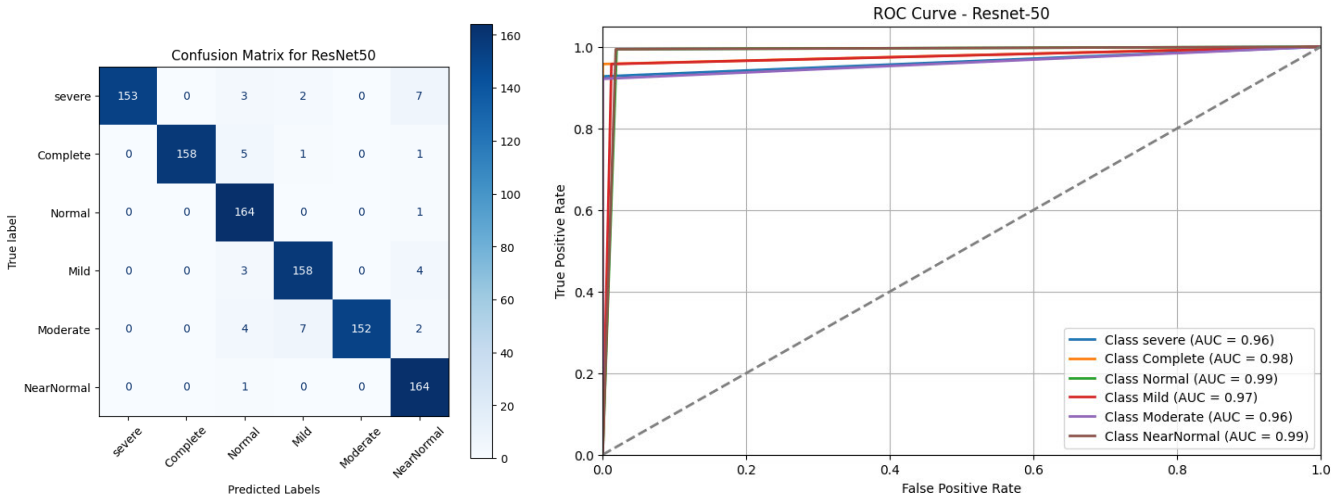
The model loss demonstrates a similar pattern to the accuracy curves. For both training and validation, it starts from a high initial value of 2.5 but decreases with the model training. The low value of training loss indicates that the model fits the training data well, while the high loss means that the model is making many incorrect predictions. The validation loss measures the model's ability to generalize to new data. The loss curve for training and validation ultimately drops to 0.2 at which point training is stopped as no further improvement is observed.

The results of the ResNet50 model using the test dataset are competitive with the FP-VGGFace model (Fig. 11). With the 150 trainable layers on the prepared dataset in comparison with only 10 layers of the FP-VGGFace model, the model demonstrates its validity in the facial paralysis prediction domain. However, the performance still lags behind the FP-VGGFace model. The confusion matrix demonstrates few FP and FN cases with the majority of the datapoints being either TP or TN. The highest number of misclassifications are with

(a) Training and validation accuracy.

(b) Training and validation loss.

FIGURE 10. **Training and validation performance analysis of the ResNet50 model using accuracy and loss curves.**



(a) Confusion Matrix for ResNet50.

(b) AUC-ROC curve analysis for ResNet50 model.

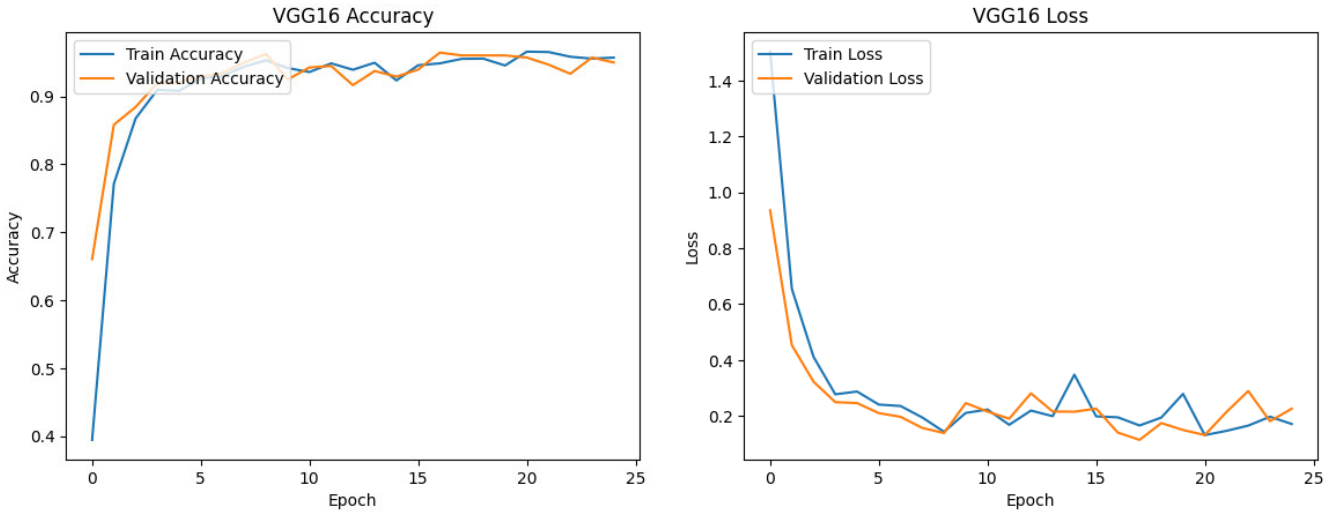
FIGURE 11. **ResNet50 model performance analysis on test data using confusion matrix and AUC-ROC curve.**

the *moderate* facial palsy class where 13 test datapoints are misclassified. For the remaining classes, misclassifications are negligible. The ROC curve also displays the competitive performance of the model with a maximum AUC of 0.99 for *normal* and *near normal* classes. The remaining AUC values are also greater than 0.96, hence it can be classified as a better model performance.

C. VGG16 MODEL

Fig. 12 displays the training and validation accuracy graphs, which illustrate the performance of the fine-tuned VGG16 model on our prepared dataset of facial paralysis detection. The graphs include two curves: the blue curve represents training accuracy, and the orange curve shows validation accuracy. The accuracy curve exhibits progressive learning of the model during the training and validation phases and is superior to the ResNet50 model. Both curves depict learning tendency during the training process and validation results occasionally surpass training results. Finally, both curves overlap at 0.97 accuracy with no further improvement. Hence, the training was stopped at 25 epochs.

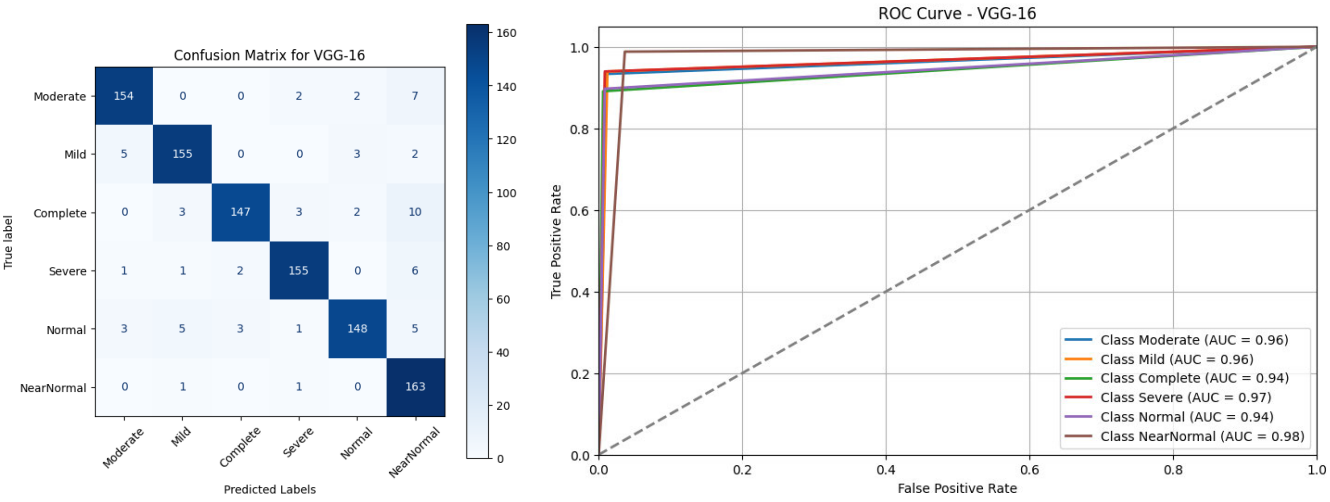
The loss curves also demonstrate the learning performance of the model on training and validation data. The training loss measures the accuracy of the model with respect to the training data, while the validation loss indicates the model's ability to generalize on unseen data. The training and validation loss curves represent steady learning of the model due to a gradual decrease in the loss. The low



(a) Training and validation accuracy.

(b) Training and validation loss.

FIGURE 12. **Training and validation performance analysis of the VGG16 model using accuracy and loss curves.**



(a) Confusion matrix of VGG16 model.

(b) AUC-ROC curve analysis for VGG16 model.

FIGURE 13. **VGG16 model performance analysis on test data using confusion matrix and AUC-ROC curve.**

validation loss from 15 epochs to 20 epochs also demonstrates better performance of the validation curve than the training curve.

Fig. 13 illustrates the model performance on test data using the confusion matrix and AUC-ROC curve. The confusion matrix evaluates the model's performance by summarizing the number of correct and incorrect predictions made on the test dataset. The results show an overall better performance of the VGG16 model on the test data except few exceptions. The *complete* and *normal* classes have more FP and FN instances as compared to other subtypes. These results are achieved with 8 trainable layers of the pre-trained VGG16 model. This may be improved by raising the count of trainable layers but then the model freezes already learned weights and starts to unlearn basic facial features. We also tried tuning the learning rate and dropout rate but that also did not improve model performance. An increase in the number of epochs also failed to improve the learning rather it stayed the same. So, the optimal results achieved for facial paralysis severity classification are AUC = 0.98 for *near normal*, AUC = 0.97 for *severe*, AUC = 0.96 for *mild* and *moderate* while least AUC was recorded for *complete* and *normal* classes having AUC = 0.94. Although the overall VGG16 model performance is better from the confusion matrix and ROC curve analysis but stands worse than both benchmark models, ResNet50 and FP-VGGFace.

D. COMPARATIVE ANALYSIS

Table 4 shows the overall accuracy, precision, recall and F1-score of FP-VGGFace, ResNet50, and VGG16 models. These metrics provide a holistic view of the model's performance in accurately classifying facial paralysis. FP-VGGFace displays superior performance for all metrics as compared to the other two benchmark models. It achieved an accuracy of 99.3% in facial palsy severity detection. The second-best performing model is ResNet50, which achieved an accuracy of 95.8%. Lastly, VGG16 achieved an accuracy of 93.1% on our prepared dataset. The precision and recall for FP-VGGFace are estimated as 99.4% and 99.3% respectively. However, for ResNet50 and VGG16 benchmark models, they remain in the range of 93.1% to 96.1%, which is quite low in comparison with FP-VGGFace model. The F1-score which demonstrates the cumulative performance of the model in terms of both precision and recall also declares FP-VGGFace as the best-performing model.

TABLE 4. Performance comparison of proposed FP-VGGFace model with two benchmark models ResNet50 and VGG16.

Model	Accuracy	Precision	Recall	F1-score
FP-VGGFace	0.993	0.994	0.993	0.993
ResNet50	0.958	0.961	0.958	0.958
VGG16	0.931	0.934	0.931	0.931

For the experimentation, both the FP-VGGFace and VGG16 models converged by retaining some of the last layers. However, for the ResNet50 model, a large portion of its layers required further training to converge. This was necessary to accurately diagnose the severity level of facial paralysis. FP-VGGFace performed exceptionally well for classifying the grading of facial paralysis because it is based on the pre-trained VGGFace model, which was originally designed for face recognition and face detection tasks. The model's weights, being generalized for facial tasks, contribute to its outstanding performance in facial paralysis classification. On the other hand, the ResNet50 model has a significant number of layers and parameters compared to other models. During the fine-tuning process, when some of its layers are retrained, it tends to suffer from overfitting. To achieve better results in facial paralysis classification with the ResNet50 model, it is necessary to further fine-tune a substantial portion of the model. Hence, to achieve competitive performance, we retrained its 150 layers. Lastly, VGG16 does not learn outstanding weights for the facial paralysis task, as the FP-VGGFace model has strong generalizability for facial tasks, leading to its superior performance in this specific classification task. Fig. 14 illustrates the comparison of FP-VGGFace with benchmark models using accuracy metric.

In the final phase of our study, we conducted extensive comparisons between our proposed FP-VGGFace model and several existing models commonly used for face paralysis classification. The models we evaluated include PHCNN-LSTM [14], VGG16 Net [10], Cascade Training [12], TPCNN [11] and FaceDisNet [29]. To ensure a comprehensive evaluation, we considered multiple performance metrics, including accuracy, precision, recall, and F1-score. The results, as shown in Fig. 15, highlight the superiority of our proposed FP-VGGFace model. It achieved an impressive accuracy of 99.3%, outperforming all the compared models. Specifically, PHCNN-LSTM achieved an accuracy of 94.8%, VGG16 Net achieved 92.6%, Cascade Training achieved 95.6%, TPCNN achieved 89.0%, and FaceDisNet achieved 98.0% for multi-class disease diagnosis. The results of our proposed model for other evaluation metrics are also better than existing models by a wide margin. These results showcase the effectiveness of our proposed FP-VGGFace model in face paralysis classification. It demonstrates the potential for diagnosing the severity level of facial paralysis.

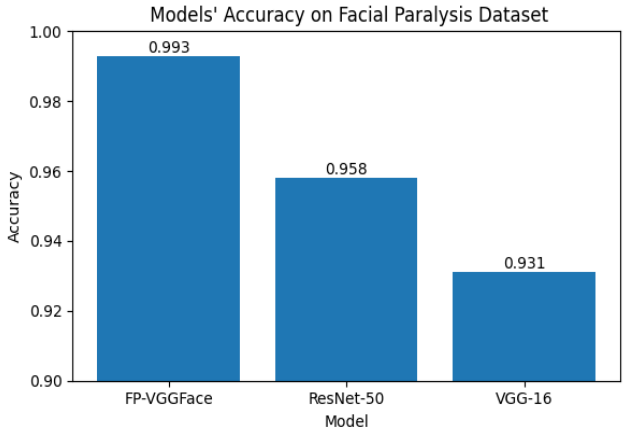


FIGURE 14. Accuracy metric comparison of the proposed FP-VGGFace model with benchmark models.

In each of the existing models, certain limitations were observed. The PHCNN-LSTM [14] model was trained on a class-imbalanced dataset, leading to a significant variance in the inter-class counts, affecting its overall accuracy. The VGG16 Net [10] model, being pre-trained on natural images without facial data, lacks specific facial features and may not perform optimally in facial-related tasks. The triple-path convolutional neural network (TPCNN) [11] has the lowest accuracy among all models. In addition, the three-model approach and the parallel convolutional neural network structure require substantial computational resources and time for training. The FaceDisNet [29] computer-aided diagnosis system for facial diseases uses integrated spatial information from several CNNs of various architectures, which is computationally expensive. It requires a large amount of computational resources. Lastly, the Cascaded Training [12] showed a strong emphasis on segmented facial regions like the mouth, nose, and eyes, but it appeared to be less attentive to other crucial parts of the patients' faces, potentially affecting its comprehensive facial analysis.

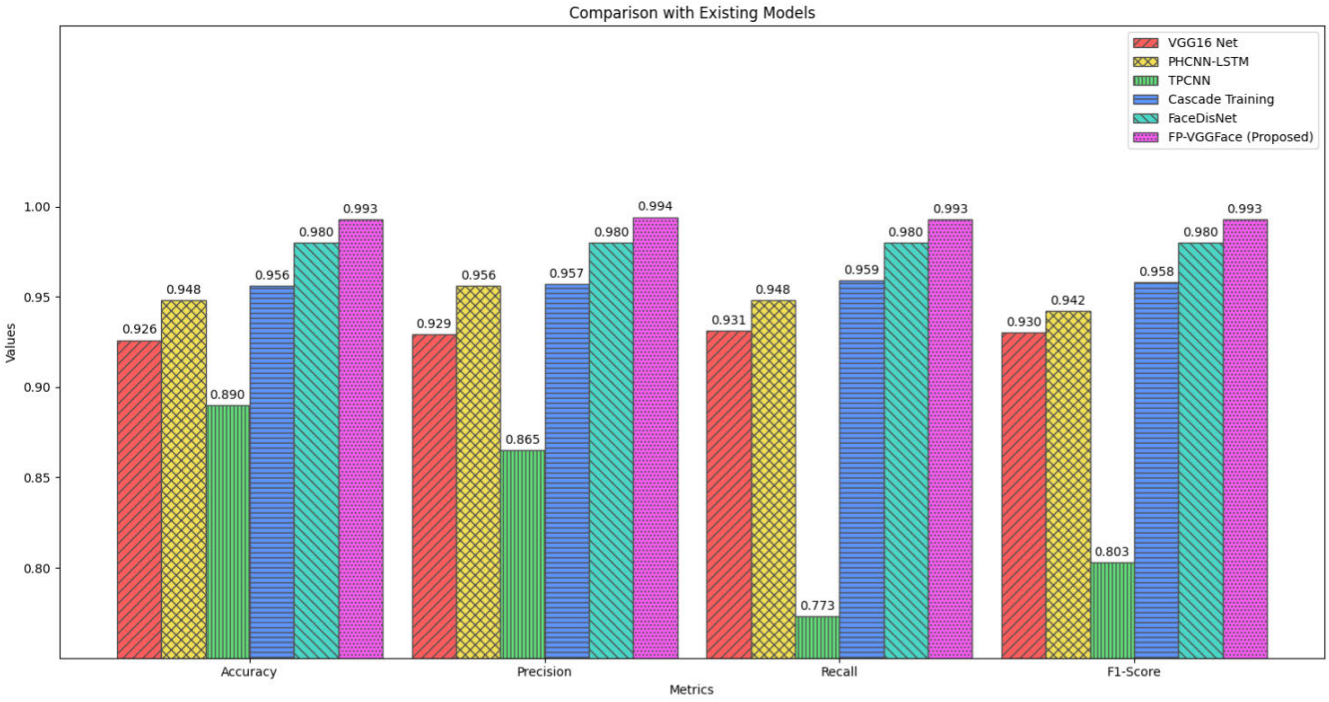


FIGURE 15. Performance comparison of the proposed FP-VGGFace model with existing models using standard evaluation matrices.

VI. DISCUSSION

The existing models for facial paralysis detection and classification have utilized both machine learning and deep learning techniques. However, these models have been limited by the use of private and small datasets, resulting in a small number of subjects and limited diversity in the data. The recent machine learning approaches in facial paralysis have focused on facial landmark detection, which may overlook texture information and rely on external methods for landmark detection. On the other hand, deep learning approaches require large labelled datasets, but the field of facial paralysis suffers from limited data availability. Transfer learning, commonly used with natural images, lacks knowledge and features specific to facial images, leading to insufficient feature learning for facial paralysis images.

In this study, we utilized a combination of two datasets: the MEEI dataset, which is clinically approved, and the YouTube Face Palsy dataset, widely used in facial paralysis research. This merger allowed us to increase the number of images, subjects, and diversity within the dataset. Both datasets primarily consist of paralyzed facial images. To address the issue of class imbalance between normal and paralyzed images, we incorporated images from the CK+ dataset, which contains normal facial expressions. Our goal was to detect facial paralysis and classify its severity level, aiding in selecting appropriate treatment options and guiding the rehabilitation process.

After merging the datasets for face paralysis detection and classification, we still faced the challenge of insufficient data to train a deep learning model from scratch. Recognizing this limitation, we turned to transfer learning as a solution. This approach allowed us to capitalize on the knowledge gained from pre-trained models in the domain of facial images. Hence, we utilized pre-trained models such as VGGFace, VGG16, and ResNet50, which are trained on larger facial image datasets like VGGFace and VGGFace2, to effectively transfer the learned knowledge to our face paralysis analysis task.

Considering the aforementioned challenges, we conducted several experiments to address face paralysis detection and classification. These experiments yielded promising results, showcasing the effectiveness of our proposed model. Furthermore, we conducted a comparative analysis to evaluate the accuracy of the proposed model in comparison to existing approaches. This analysis provides valuable insights into the performance and potential of our proposed method in accurately detecting and classifying facial paralysis severity.

VII. CONCLUSION

This paper presents an approach for the severity classification of face paralysis, which occurs due to facial muscle weakness and nerve damage, resulting in impaired facial function. Our transfer learning approach addresses the limitations of existing models and proposes a facial palsy classification model for grading the severity level of patients. The results reveal the superior performance of our model by achieving a remarkable accuracy of 99.3%. Overall, our deep learning approach, incorporating transfer learning and fine-tuning, enables us to leverage pre-existing knowledge and features extracted from facial images for accurate and efficient face paralysis detection and classification. An automatic facial paralysis detection and classification system has significant

benefits for both physicians and patients involved in the rehabilitation process. It can assist physicians in selecting the most suitable treatment plan based on accurate and objective assessments of the patient's condition. Additionally, it enables patients to track and evaluate their recovery progress throughout the treatment process.

Future research in the field of facial paralysis holds significant potential for advancements and improvements in healthcare facilities. The domain of facial paralysis lacks generalized models for effective detection and classification. Key areas where advancements can be made include the preparation of datasets related to facial paralysis. This involves the collection of diverse and representative datasets featuring cases of facial paralysis. In the future, an incremental learning approach could be employed. The model can be designed to learn incrementally as new data becomes available over time.

ACKNOWLEDGMENT

This Research is funded by Researchers Supporting Project Number (RSPD2023R947), King Saud University, Riyadh, Saudi Arabia.

REFERENCES

[1] C. Jiang, J. Wu, W. Zhong, M. Wei, J. Tong, H. Yu, and L. Wang, "Automatic facial paralysis assessment via computational image analysis," *J. Healthcare Eng.*, vol. 2020, pp. 1–10, Feb. 2020.

[2] G. S. Parra-Dominguez, R. E. Sanchez-Yanez, and C. H. Garcia-Capulin, "Facial paralysis detection on images using key point analysis," *Appl. Sci.*, vol. 11, no. 5, p. 2435, Mar. 2021.

[3] M. A. Alagha, A. Ayoub, S. Morley, and X. Ju, "Objective grading facial paralysis severity using a dynamic 3D stereo photogrammetry imaging system," *Opt. Lasers Eng.*, vol. 150, Mar. 2022, Art. no. 106876.

[4] D. L. Guarin, Y. Yunusova, B. Taati, J. R. Dusseldorp, S. Mohan, J. Tavares, M. M. van Veen, E. Fortier, T. A. Hadlock, and N. Jowett, "Toward an automatic system for computer-aided assessment in facial palsy," *Facial Plastic Surg. Aesthetic Med.*, vol. 22, no. 1, pp. 42–49, Feb. 2020.

[5] J. J. Greene, D. L. Guarin, J. Tavares, E. Fortier, M. Robinson, J. Dusseldorp, O. Quatela, N. Jowett, and T. Hadlock, "The spectrum of facial palsy: The MEEI facial palsy photo and video standard set," *Laryngoscope*, vol. 130, no. 1, pp. 32–37, Jan. 2020.

[6] Z. Guo, W. Li, J. Dai, J. Xiang, and G. Dan, "Facial imaging and landmark detection technique for objective assessment of unilateral peripheral facial paralysis," *Enterprise Inf. Syst.*, vol. 16, nos. 10–11, pp. 1556–1572, Oct. 2022.

[7] Y. Liu, Z. Xu, L. Ding, J. Jia, and X. Wu, "Automatic assessment of facial paralysis based on facial landmarks," in *Proc. IEEE 2nd Int. Conf. Pattern Recognit. Mach. Learn. (PRML)*, Chengdu, China, Jul. 2021, pp. 162–167.

[8] A. Arora, A. Sinha, K. Bhansali, R. Goel, I. Sharma, and A. Jayal, "SVM and logistic regression for facial palsy detection utilizing facial landmark features," in *Proc. 14th Int. Conf. Contemp. Comput.*, New York, NY, USA, Aug. 2022, pp. 43–48.

[9] P. Xu, F. Xie, T. Su, Z. Wan, Z. Zhou, X. Xin, and Z. Guan, "Automatic evaluation of facial nerve paralysis by dual-path LSTM with deep differentiated network," *Neurocomputing*, vol. 388, pp. 70–77, May 2020.

[10] M. Sajid, T. Shafique, M. Baig, I. Riaz, S. Amin, and S. Manzoor, "Automatic grading of palsy using asymmetrical facial features: A study complemented by new solutions," *Symmetry*, vol. 10, no. 7, p. 242, Jun. 2018.

[11] M. Shayestegan, J. Kohout, K. Štícha, and J. Mareš, "Advanced analysis of 3D kinect data: Supervised classification of facial nerve function via parallel convolutional neural networks," *Appl. Sci.*, vol. 12, no. 12, p. 5902, Jun. 2022.

[12] T. Wang, S. Zhang, L. Liu, G. Wu, and J. Dong, "Automatic facial paralysis evaluation augmented by a cascaded encoder network structure," *IEEE Access*, vol. 7, pp. 135621–135631, 2019.

[13] Z. Guo, M. Shen, L. Duan, Y. Zhou, J. Xiang, H. Ding, S. Chen, O. Deussen, and G. Dan, "Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Melbourne, VIC, Australia, Apr. 2017, pp. 135–138.

[14] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, and T. D. Pham, "Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 10, pp. 2325–2332, Oct. 2020.

[15] G. J. Hsu, J.-H. Kang, and W.-F. Huang, "Deep hierarchical network with line segment learning for quantitative analysis of facial palsy," *IEEE Access*, vol. 7, pp. 4833–4842, 2019.

[16] A. Gaber, M. F. Taher, M. A. Wahed, N. M. Shalaby, and S. Gaber, "Classification of facial paralysis based on machine learning techniques," *Biomed. Eng. OnLine*, vol. 21, no. 1, pp. 1–10, Sep. 2022.

[17] S. A. Ansari, K. R. Jerripothula, P. Nagpal, and A. Mittal, "Eye-focused detection of Bell's Palsy in videos," 2022, *arXiv:2201.11479*.

[18] G. S. Parra-Dominguez, C. H. Garcia-Capulin, and R. E. Sanchez-Yanez, "Automatic facial palsy diagnosis as a classification problem using regional information extracted from a photograph," *Diagnostics*, vol. 12, no. 7, p. 1528, Jun. 2022.

[19] G. Storey, R. Jiang, S. Keogh, A. Bouridane, and C.-T. Li, "3DPalsyNet: A facial palsy grading and motion recognition framework using fully 3D convolutional neural networks," *IEEE Access*, vol. 7, pp. 121655–121664, 2019.

[20] G. J. Hsu, W.-F. Huang, and J.-H. Kang, "Hierarchical network for facial palsy detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 580–586.

[21] C. Xu, C. Yan, M. Jiang, F. Alenezi, A. Alhudhaif, N. Alnaim, K. Polat, and W. Wu, "A novel facial emotion recognition method for stress inference of facial nerve paralysis patients," *Expert Syst. Appl.*, vol. 197, Jul. 2022, Art. no. 116705.

[22] Y. Zhuang, M. M. McDonald, C. M. Aldridge, M. A. Hassan, O. Uribe, D. Arteaga, A. M. Southerland, and G. K. Rohde, "Video-based facial weakness analysis," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 9, pp. 2698–2705, Sep. 2021.

[23] C. Yang, J. Kang, X. Xue, Y. Zhou, H. Wang, Z. Wan, T. Su, F. Xie, and P. Xu, "Automatic degree evaluation of facial nerve paralysis based on triple-stream long short term memory," in *Proc. 3rd Int. Symp. Image Comput. Digit. Med.*, Xi'an, China, Aug. 2019, pp. 7–11.

[24] A. Song, Z. Wu, X. Ding, Q. Hu, and X. Di, "Neurologist standard classification of facial nerve paralysis with deep neural networks," *Future Internet*, vol. 10, no. 11, p. 111, Nov. 2018.

[25] R. Malli. *VGGFace Implementation With Keras Framework*. Accessed: Oct. 5, 2023. [Online]. Available: https://github.com/rcmalli/keras-vggface

[26] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[27] E. Rasool, M. J. Anwar, B. Shaker, M. H. Hashmi, K. U. Rehman, and Y. Seed, "Breast microcalcification detection in digital mammograms using deep transfer learning approaches," in *Proc. 9th Int. Conf. Comput. Data Eng.*, Jan. 2023, pp. 58–65.

[28] O. Parkhi, V. Andrea, and Z. Andrew, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Swansea, U.K., Sep. 2015, pp. 1–12.

[29] O. Attallah, "A deep learning-based diagnostic tool for identifying various diseases via facial images," *Digit. Health*, vol. 8, Jan. 2022, Art. no. 205520762211244.

WASIF ALI received the Bachelor of Science degree in computer science from International Islamic University Islamabad, Pakistan, and the Master of Science degree in computer science from COMSATS University Islamabad, Islamabad Campus, Pakistan, in June 2023. He is currently a Dedicated Research Scholar. His research interests include artificial intelligence, deep learning, computer vision, and data mining. With a passion for research and a strong academic foundation, he has honed a diverse skill set, including research expertise, technical writing proficiency, critical thinking abilities, data analysis, problem-solving skills, and adaptability.

MUHAMMAD IMRAN received the Graduate degree in software engineering from the University of Engineering and Technology Taxila, Pakistan, in 2006, and the master's degree in software engineering and the Ph.D. degree in computer science from the University of Southampton, U.K., in 2009 and 2015, respectively. He was a Lecturer with COMSATS University Islamabad (CUI), Islamabad, Pakistan, from 2007 to 2008, where he is currently an Assistant Professor with the Department of Computer Science. His research interests include artificial intelligence, machine learning, natural language processing, and semantic web.

MUHAMMAD USMAN YASEEN is currently an Assistant Professor with COMSATS University Islamabad (CUI), Islamabad, Pakistan. His current research interests include video analytics, big data analysis, machine learning, and distributed systems.

KHURSHEED AURANGZEB (Senior Member, IEEE) received the B.S. degree in computer engineering from the COMSATS Institute of Information Technology Abbottabad, Pakistan, in 2006, the M.S. degree in electrical engineering (system on chip design) from Linköping University, Sweden, in 2009, and the Ph.D. degree in electronics design from Mid Sweden University, Sweden, in June 2013. He is currently an Associate Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), Riyadh, Saudi Arabia. He has obtained more than 15 years of excellent experience as an Instructor and a Researcher in data analytics, machine/deep learning, signal processing, electronics circuits/systems, and embedded systems. He has been involved in many research projects as a principal investigator and a co-principal investigator. He has authored or coauthored more than 90 publications, including IEEE/ACM/Springer/Hindawi/MDPI journals and flagship conference papers. His research interests include embedded systems, computer architecture, signal processing, wireless sensor networks, communication, and camera-based sensor networks, with an emphasis on big data and machine/deep learning with applications in smart grids, precision agriculture, and healthcare.

NOUMAN ASHRAF received the B.S. and M.S. degrees in electrical engineering from COMSATS University Islamabad, Pakistan, and the Ph.D. degree in electrical engineering from Frederick University, Cyprus, under the Erasmus Mundus Scholarship Program. He was with the Turku University of Applied Sciences, Finland, the TSSG, Waterford Institute of Technology, Ireland, and the University of Cyprus. Currently, he is with Technological University Dublin, Ireland. His research interests include the application of control theory for management of emerging networks with applications in the Internet of Things, 5G and beyond communication networks, electric vehicles, metamaterial-based beam steering, wireless sensor networks, smart grids, networks resilience, and power load modeling.

SHERAZ ASLAM (Member, IEEE) received the B.S. degree in computer science from Bahauddin Zakariya University (BZU), Multan, Pakistan, in 2015, the M.S. degree in computer science with a specialization in energy optimization in the smart grid from COMSATS University Islamabad (CUI), Islamabad, Pakistan, in 2018, and the Ph.D. degree in computer engineering and informatics from the Cyprus University of Technology (CUT), Limassol, Cyprus, under the supervision of Dr. Herodotos Herodotou. He was a Research Associate during the M.S. period with CUI under the supervision of Dr. Nadeem Javaid. Currently, he is a Postdoctoral Researcher with the DICL Research Laboratory, CUT, where he is also a part of an European Union funded research project named as aerOS. Previously, he worked on several EU funded research projects, i.e., STEAM and MARI-Sense. He has authored more than 80 research publications in ISI-indexed international journals and conferences, including IEEE Internet of Things Journal, *Renewable and Sustainable Energy Reviews*, and *Electric Power System Research*. His research interests include data analytics, generative adversarial networks, wireless networks, smart grids, and cloud computing. He is also constantly looking for collaboration opportunities with professors and students from different universities around the globe. He also served as a TPC member and an invited reviewer for international journals and conferences.