A Hybrid Multitasking Cascaded CNN leveraging Xception Architecture for Enhanced Face Recognition

Tamilselvi M.

Department of Electronics and Communication Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, (SIMATS), Chennai, India.

E-mail: tamilselvivlsi@gmail.com

Abstract

Face recognition technology plays a significant role in surveillance systems, authentication systems, and systems that allow people and computers to communicate with each other. The main aim of face recognition technology is to identify individuals based on their facial features. It is crucial to analyze facial features accurately in today's world, where numerous smart devices and surveillance systems are installed globally. Deep learning algorithms such as CNN, VGG16, and ResNet have shown promising results in improving the accuracy of face recognition technology, but the results are not sufficient in open environments. This paper proposes a new face recognition system that utilizes deep learning concepts. The proposed system aims to perform tasks such as detecting faces, locating landmarks, and recognizing faces. It employs depth-wise separable convolutional networks to reduce the number of parameters and optimize the features. The system is rigorously tested using various datasets, including LFW, CelebA, and CASIA-WebFace, and demonstrates exceptional performance by achieving 98.3% accuracy in face recognition. Additionally, the proposed system achieved 4.8% higher accuracy and 32% lower latency compared to existing face recognition systems.

Keywords: Face Recognition, Xception Based Multitask Cascaded CNN, XMCCNN, Convolutional Neural Network, VGG-16, Digital Image Processing, Smart Technologies.

1. Introduction

In this modern era, many innovative devices and associated technologies are undergoing drastic development to provide efficient support to customers. In 1981, IBM introduced the PC, which has been continuously developed since then. In 2000, the online era emerged, providing numerous features for clients to engage in online shopping and more [1]. These advancements have led to an information technology environment that boosts techniques and offers ultimate support to customers based on security, as well as fast and easy accessibility. In this context, numerous security measures are implemented to analyze a person's identity, recognized through credential matching techniques, two-step authentication, biometric verification, and so on. However, all these methods have limitations at certain stages of verification. In this paper, an efficient face recognition scheme is introduced, which allows for easy recognition of individuals without delay. The major considerations for dealing with such digital image processing streams include accuracy estimations, speed constraints, and cost efficiency. All

DOI: https://doi.org/10.36548/jiip.2026.1.021
Received: 02.02.2026, received in revised form: 28.02.2026, accepted: 14.03.2026, published: 25.03.2026

these features are taken into account, with the intention of providing an efficient face recognition system with proper accuracy measures [2].

Several face recognition algorithms have been available in the past to recognize facial features and deliver results accordingly. The major problem that needs to be addressed in face recognition is managing the accuracy level of matching by analyzing faces according to various changes in facial features, expressions, poses, and so on [3]. All these factors affect the accuracy level incrementally, and several algorithms have been developed in the past using Convolutional Neural Networks, VGG16, Random Forest Classifiers, and others. However, all these algorithms encounter limitations in certain cases, resulting in low accuracy levels while detecting facial features, and they need to resolve the time complexity associated with such algorithms [4]. Most real-time face recognition algorithms utilize Convolutional Neural Networks for processing digital images and delivering results accordingly. However, there is a need for analysis to enhance the accuracy level significantly using modern face recognition algorithms as well [5][6]. Face recognition is an important aspect of security-related applications, surveillance systems, and the interaction between humans and computers. Recent advancements in deep CNNs have shown promising results for improving face recognition accuracy by increasing the depth of the network and enhancing feature extraction capabilities. VGG-16, ResNet-152, Inception-V3, and Xception are widely used for various vision-related recognition tasks due to their efficient feature learning capabilities. Xception employs depth separable filters for feature extraction, which helps learn discriminative features and reduces the complexity associated with the computation process. Due to the aforementioned benefits of the Xception model, the proposed work uses this model as a backbone and attempts to enhance it by modifying density-related parameters to improve face recognition accuracy. The proposed model was validated through various experiments conducted using multiple face datasets [7].

The following figure 1 illustrates the traditional convolutional neural network methodology system architecture with maximum pooling capacity and interconnected layer terminology for identifying face correlations, in which the pooling determinations eliminate the mismatching dimensionality of the given face while preserving the important features as they are. The given input face image is divided into several rectangular portions, and each layer of information is converted into linear and non-linear formulations [8]. This kind of pooling formulation eliminates the concept of face image bipolarity, and noise values in the given face images are highly controlled during processing. The robustness of processing is specifically emphasized at each stage of processing [9]. CNNs are better than fully connected DNNs for human facial feature extraction because they use spatial location, weight sharing, and translation invariance. Fully linked DNNs do not take into account spatial structure, which results in more parameters and increases the likelihood of overfitting when recognizing images [10].

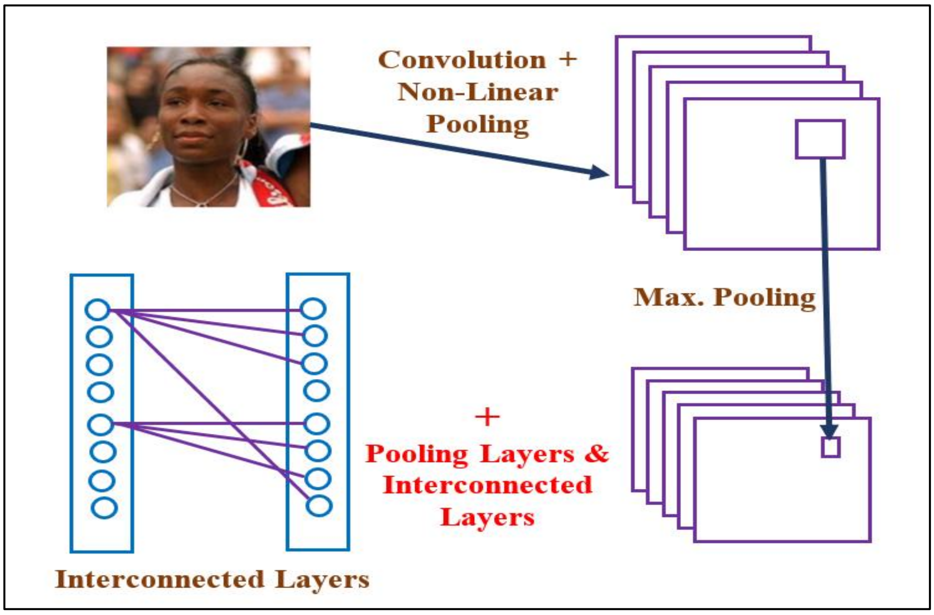

Figure 1. Classical Convolutional Neural Network Process Model

In the conventional architecture of the Xception algorithm, a limited number of channels are used for facial feature representation, whereas it is slightly modified in the proposed approach, such that the channel density of the feature representation is increased up to 256 layers to enable higher and more discriminative facial feature extraction. This modification in the proposed XMCCNN algorithm can be reflected in the experimental results, which show higher accuracy levels of face recognition on various benchmark face datasets.

There are distinctive machine learning schemes such as Convolutional Neural Networks (CNN), Auto-Encoder Schemes, and Deep Neural Networks (DNN) [11]. Convolutional Neural Networks are a type of Artificial Intelligence scheme that utilize convolutional approaches to extract features from the data to enhance the quantity of facial features. That research group's organization was the source of a significant part of the new models and a genuine motivation for many other researchers in the field of machine learning and digital image processing [12]. It is generally viewed as one of the most powerful distributions in the information technology field and has demonstrated that Convolutional Neural Networks outperform recognition performance compared to all other traditional methods. With the computational power of Graphical Processing Units, Convolutional Neural Networks have achieved exceptionally robust and solid results across various domains, including image processing, machine learning, and deep learning. The main objective of this paper is to provide a robust and efficient face recognition algorithm with the highest accuracy constraints. In this paper, a new methodology is introduced by changing the density factors of the existing algorithm called Xception and improving it by adding several density parameters with a 4-layer-centric approach, details of which are clearly mentioned in the methodologies section [13].

In the proposed architecture, the channel sizes that vary from 128 to 256 indicate an increase in the convolutional layers to ensure enhanced capacity for feature representation. The channel depth of 128 can also be increased to 256 to enhance the representational capacity of the network, allowing it to learn more discriminative facial features, such as fine-grained texture details and subtle spatial variations that are important for face recognition in demanding conditions, such as pose, illumination, and changes in occlusion. This channel expansion offers improved separability and resistance to noise without a corresponding increase in computational complexity because depthwise separable convolutions are utilized to avoid an overabundance of parameters.

The rest of this paper illustrates the related study over section 2, further section of Section 3 illustrates the proposed system methodologies in detail with proper algorithm flow and the Section 4 illustrates the Result and Discussion portion of the paper and the final section, Section 5 illustrates the concept of Conclusion and Future Scope of the proposed paper. These all will be explained in detail over the further section summaries.

2. Related Work

The authors proposed a paper related to face detect ion and recognition schemes with respect to digital image processing strategies. In this paper, the authors presented the most important probability for identifying a person's face; with the help of facial features alone, a person can be easily identified. Therefore, facial features and the associated processing techniques are developed to identify the person based on them [12]. The authors illustrated the overall process of face detection and recognition through two separate factors: face detection and face recognition. In the process of face detection, the identity of the face is observed from the input image by means of the shape and structural position of the face [14]. The second portion of face recognition identifies the identity of the person concerning the already trained models. This paper [6] follows two individual methodologies to identify the facial features accurately: the Eigenface technique and the Principal Component Analysis (PCA) model. The main focus of this paper is to accurately identify the facial features and to identify the person based on them; moreover, this paper places significant importance on digital image processing features. The major advantage found in this paper is that it uses diverse algorithms to identify the facial features precisely through the Eigenface technique and the PCA methodology, which provide high accuracy in results as well as better timing improvements [15].

The authors proposed a paper related to facial feature identification and recognition strategies using diverse algorithms. In this paper, the authors illustrated that the analysis and estimation of facial features are systematic processes within the information technology domain, which employs a diverse set of methodologies to detect faces and recognize them using digital image processing schemes. The authors specify that the facial differences of individuals are checked by contrasting captured images with the facial pictures stored in the database. Face recognition schemes are significant topics from a systems perspective, and numerous researchers have examined this subject from multiple points of view; it is especially important in certain applications, such as surveillance systems [16]. The primary objective of this review paper is to examine the various methodologies used for facial feature identification and recognition. In this paper, the authors cross-checked their viewpoints with a diverse set of algorithms such as Neural Networks, Adaboost logic, and so on. This research is producing a substantial set of positive outcomes and provides a pathway for further researchers to work on it [17]. Typically, this kind of survey work analyzes the concept from a single perspective, but in this paper, many specifications associated with facial features are recognized and monitored concerning different methodologies. The major focus identified in the paper is extracting facial features with the help of diverse algorithms and providing the best accuracy in summary through digital image processing logic. The major advantage found in this paper is that it comprises a diverse set of algorithms to analyze facial features and paves the way to attain the best results in real-time with optimal accuracy, as well as provides a wealth of ideas for further research in the same domain [18].

The authors proposed a paper related to a real-world face recognition scheme with respect to deep learning methodology. In this paper, the authors illustrate that the growing and wide utilization of deep learning technology has enabled significant advancements in the accuracy of face feature identification under various conditions. However, the reported near-optimal performance on standard benchmark datasets, such as Labeled Faces in the Wild, does not account for complexities in unconstrained applications. The research detailed in this paper addresses some of the fundamental challenges of face detection and identification under adverse conditions, and in this context, a comprehensive face manipulation strategy is introduced for real-time media/video-based face identification and detection processes. This system identifies, tracks, and recognizes individuals from real-time video streams, while also addressing various critical challenges of media/video-based facial feature recognition systems, such as end-to-end processing, complexity, in-the-wild facial feature identification, and diverse dynamic person facial feature recognition concerning video processing techniques. The modern deep learning-based neural networks for face detection and facial feature representation are explored, while minimizing the computational overhead from the other modules in the recognition pipeline. The main advantage found in this paper is that it utilizes deep learning principles to identify facial features and provides improved accuracy levels regarding reasonable time complexities. Additionally, this type of deep learning-based facial feature estimation strategy offers the best training and testing evaluations in results, which the paper describes in more detail [19].

The authors proposed a paper related to facial feature identification and recognition schemes based on several experimental strategies with respect to image processing. In this paper, the authors illustrated the different logical features available in the face recognition scheme and how these schemes contribute to creating a security-enabled environment and the associated perspectives. The proposed framework can continuously identify individuals' faces and simultaneously perceive them. Thus, this paper aims to develop a framework that can identify and recognize individual faces in real-time.

The authors proposed a paper related to the survey of face recognition schemes and the associated technological developments in recent days. In this paper, the authors illustrated that facial features are the most important and secure way of identifying a person's identity compared to other classical schemes, such as credential methodologies and other biometric schemes. This paper considers faces in real-time environmental surroundings and analyzes the features accordingly to estimate the proper training and testing norms. The authors noted that facial feature estimation and recognition have become future advancement pathways and that there are numerous potential system possibilities. The major advantage identified from both these papers is the reduction of time consumption and the application of the logic of face recognition schemes in real-world environments, producing survey results with the best possible outcomes.

3. Proposed Work

The proposed approach of the Xception-based Multitask Cascaded CNN (XMCCNN) is essentially an enhanced CNN logic that provides the highest accuracy factors for face recognition outcomes. This logic is inspired by the classical Xception architecture. The nature of XMCCNN is derived from the classical Xception logic, but the flow is altered based on depth modifications, which reduce noise factors and further improve accuracy levels. In conventional CNNs, a limited number of channels are used for feature extraction, typically starting from 32. In contrast, the proposed algorithm modifies this across four layers with a deeper density of channels, varying from 128 to 256, for better extraction of facial features.

The channels are expanded using a depth multiplier, which increases the depth of the separable convolution and the pointwise convolution filters from 128 channels to 256 channels. The proposed XMCCNN architecture can expand the depth of the network, leading to improved and more discriminative learning of features through a network with a higher number of channels. Although this type of expansion may theoretically increase the chances of overfitting, particularly when the training data have high intra-class variation, it is successfully mitigated in the proposed framework through various regularization measures. In particular, depthwise separable convolutions have a much smaller number of trainable parameters (relative to conventional convolutions), thus preventing unwarranted complexity in the model. Moreover, ImageNet-pretrained weights, transfer learning, massive data augmentation (including pose, illumination, and occlusion variations), dropout regularization, and early stopping all help limit overfitting.

The system is designed to address various real-world problems (such as occlusion, variations in illumination, and changes in poses) through the use of a depthwise separable convolutional backbone and modular task-specific modules.

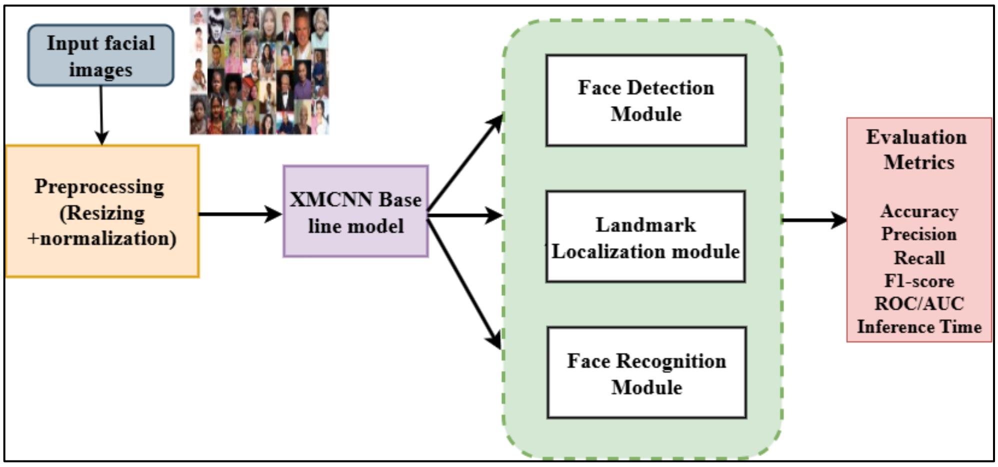

Figure 2. Block Diagram of the Proposed Model

The preprocessing of the input images is done as shown in Figure 2 in order to resize, normalize, and augment the data for the enhancement of accuracy. The Xception backbone is used to extract features and offers powerful and efficient features. The extracted features then undergo three modules: a face detector module that identifies the bounding box, a landmark localization module that extracts important facial features like the eyes and nose, and lastly, a face recognition module that identifies the identity. Finally, standard metrics such as accuracy, precision, recall, F1-score, ROC/AUC, and inference time are used to evaluate the system, effectively assessing both recognition performance and computational efficiency.

This modification proves the novelty of the proposed approach, XMCCNN, and this altered depth ratio separable convolution logic portrays the point-wise convolution logic. The motto of this approach is to provide a high accuracy ratio of facial feature extraction and recognition logic with respect to convolutional logic. This proposed model of XMCCNN is designed with the inspiration of Inception-V3 logic, in which the convolution based on 1x1 logic is completed initially, leading to n x n convolution to attain the best result in spatial-convolution logic. This nature varies from the actual constraints, and the spatio-convolutional features have two diverse variations: the position of facial features with respect to depth-wise convolutional logic and the availability/unavailability of non-linear pathways. In the first Inception-V3 implementations, there are non-linear propagations behind initial activity. In Figure 3, the conventional architecture of Xception logic is depicted, in which the changed depth-wise divisible convolution lacks a halfway rectified linear unit (ReLU) on non-linear norms.

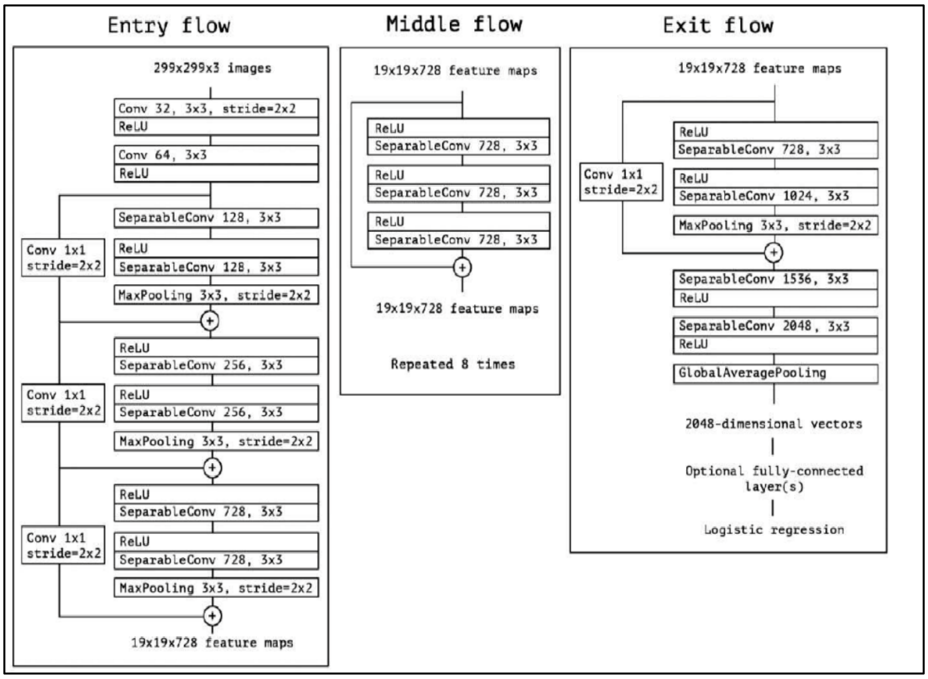

Figure 3. An Architecture of Xception Logic [25]

The changed depth-wise distinguishable convolutional logics, along with various enactment units, are tested. The following figure, Figure 4, illustrates the entire workflow logics behind the Xception model, with proper convolutional feature maps and three major constraints [16]: Initial Flow Level, Middle Flow Level, and Final Flow Level.

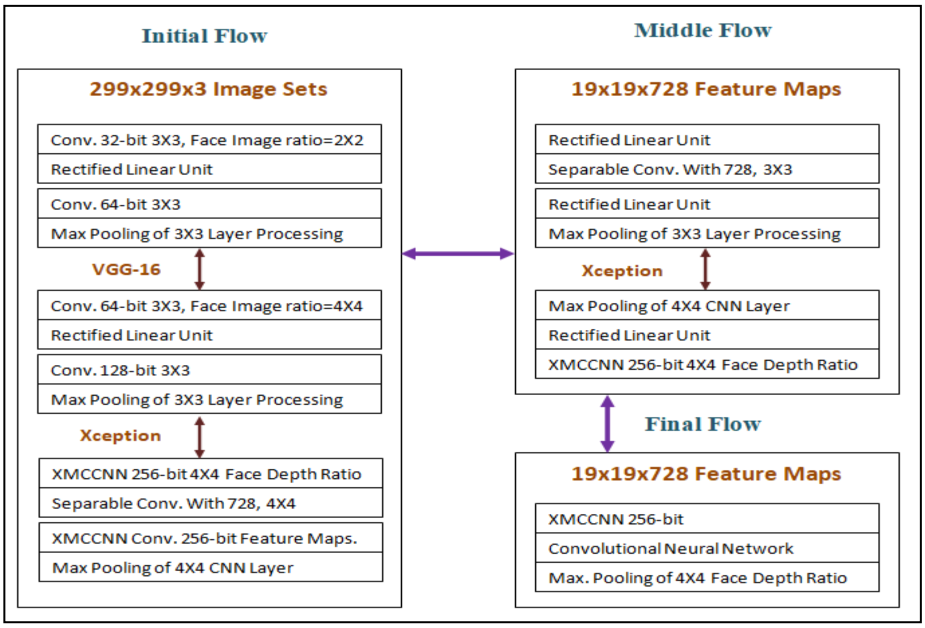
Initial Flow
299x299x3 Image Sets
Conv. 32-bit 3X3, Face Image ratio=2X2
Rectified Linear Unit
Conv. 64-bit 3X3
Max Pooling of 3X3 Layer Processing
VGG-16
Conv. 64-bit 3X3, Face Image ratio=4X4
Rectified Linear Unit
Conv. 128-bit 3X3
Max Pooling of 3X3 Layer Processing
Xception
XMCCNN 256-bit 4X4 Face Depth Ratio
Separable Conv. With 728, 4X4
XMCCNN Conv. 256-bit Feature Maps.
Max Pooling of 4X4 CNN Layer
Middle Flow
19x19x728 Feature Maps
Rectified Linear Unit
Separable Conv. With 728, 3X3
Rectified Linear Unit
Max Pooling of 3X3 Layer Processing
Xception
Max Pooling of 4X4 CNN Layer
Rectified Linear Unit
XMCCNN 256-bit 4X4 Face Depth Ratio
Final Flow
19x19x728 Feature Maps
XMCCNN 256-bit
Convolutional Neural Network
Max. Pooling of 4X4 Face Depth Ratio

Figure 4. Proposed XMCCNN Workflow Model

As mentioned in Figure 4, the separable convolution logic depth-wise variations are altered according to the range of 299x299x3 facial image classifications with respect to 19x19x728 feature map specifications. The Convolutional Neural Network follows a 4x4 cross-layered architecture over the proposed XMCCNN logic, as well as the Xception principles, to manipulate the facial features with high accuracy in detection and recognition [16]. All these facial feature detection principles are similar in terms of the base of convolutional logic; however, in the novelty constraint, a new methodology needs to be estimated with respect to the top priority algorithm logic, so that the logic of Xception is considered in the scenario, and the proposed logic of XMCCNN is designed based on that with some depth-wise density modifications, which improve the accuracy levels of the proposed approach.

The occlusions were generated with synthetic data using rectangular blocks of random size, as well as with simulated masks, assuming that the percentage-based pixel masking is considered to be 20-80%. The proposed methodology of XMCCNN and its architectural design can be illustrated in Figure 5. The system is trained with the well-known Kaggle face dataset, which contains different types of facial features. The input image is cross-checked with the trained dataset to produce the exact outcome with proper accuracy levels. Initially, the process of convolution begins to optimize the testing input face image, and the sampling procedures are considered with respect to facial feature maps. The interconnected architecture of the proposed XMCCNN design extracts the face features of the testing image and the trained dataset samples to attain the best possible outcome as a result.

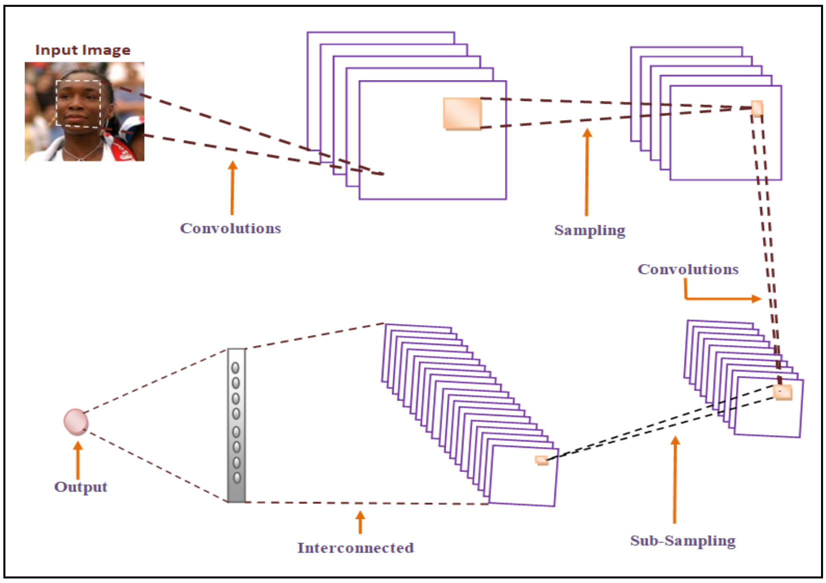

Figure 5. XMCCNN Architectural Design

The facial expression-enabled datasets are also capable of identifying facial features properly, and this type of dataset is used to identify individuals efficiently. To reduce the error rate of the convolutional neural network model, such as Xception, the MSE values (E) are provided by Equation 1, as given below.

$$(E)^{n} = \frac{1}{\mathrm{Tp+Ts}} \sum_{i=0}^{Tp} 1(\mathrm{i,j}) \sum_{j=0}^{Ts} \binom{Tp(i).Ts(j)}{0-\mathrm{n}} i, j = 1 \; to \; Ts^{\mathrm{n}} \qquad (1)$$

Where Tp refers to the overall training pattern of the dataset, Ts indicates the number of testing samples provided during modeling the dataset, and i and j are used to indicate the iteration levels on a step basis starting from 0 to n. The feature extraction summarizations are illustrated based on the following equation, in which the feature extraction process examines the image features with respect to three different parameters: the image directory, the number of images to be taken for processing, and the overall group size of the images in the given directory. Based on these input parameters, the images are processed, and the features are generated and stored in the respective array object for further processing. The extracted features will be returned to the next level of training to model the dataset for testing, as given in equation 2.

$$D_{gen} \rightarrow I^{S}(\frac{1}{n} + \frac{0}{255})\sum i_{Batch}(\frac{0}{255}) + I_{Cnt} \quad (2)$$

Where DGen indicates the image data generator, IS indicates the resizing of images from 1 to n available in the directory, and the testing sample pixel ranges are mentioned from 0 to 255 from the batch of training samples. ICnt indicates the batch count. The following algorithm, Algorithm-1, illustrates the feature extraction process more clearly with proper pseudocode.

Algorithm 1: Feature Extraction

Input: Image Directory, Number of Images and the Batch Size.
Output: Extracted features with class name indices and labels.
Step-1: Define the function for extracting image features from the given directory.
Pseudocode:

```
Define extracting_Imgfeatures(Img_dir, sampleCount, batchSize)
```

Step-2: Declare an object for data generation from the image data generator function, with the definition of rescaling pixel ratios up to the channels of 256.
Step-3: Assign the proper batch size with random selection of images as false specification, due to preserve the accuracy in selection.
Pseudocode:

```
data_gen = Image_Data_Generator(rescaling=1./255);
Img_generator = data_gen.from_directory(Img_directory, batchSize=batch_size, targetSize=[32,32], shuffle=false);
class_names=generator.class_indices
```

Step-4: Define the image count variable and assign the initial value as 0.
Step-5: Define two array variables X and Y for storing the batch values and labels.
Step-6: Initiate the For loop with the specification of input batches, label batches.
Step-7: Append the input batches into the X array and label batches into the Y array.
Step-8: Increment the image count with respect to the input batch structure.
Step-9: Check if the image count is greater than the last values bounded on image generator, if so, then break the condition and come out from it.
Step-10: Otherwise concatenate the input batches into X array variable and concatenate the batches into Y array.
Step-11: Return the values of X, Y and respective class indices.
Pseudocode:

```
        image_cnt→0;
        X_batch→ []; Y_batch→ [];
        for(input_batch; label_batch) img_generator
        {
                X_batch_append(inputsBatch);
                Y_batch_append(labelsBatch);
```

```
            image_cnt+= inputsBatch_shape(0);
                if(image_cnt>=img_generator.n)
                { break }
            X → concatenate(X_batch);
            Y → concatenate(Y_batch);
    }
    return X,Y,class_names;
```

The following algorithm, Algorithm-2, creates a model for the proposed algorithm called Xception-based Multitask Cascaded CNN (XMCCNN), in which it creates a model based on two different parameters in association with TensorFlow library logic. The parameters are the shape of the given image and the specified labels.

Algorithm 2: Xception based Multitask Cascaded CNN

Input: Image Features, Shape of the image and the label specifications.
Output: Model trained with proper accuracy level.
Step-1: Import the required libraries for processing the facial features. The required libraries are tensorflow associated models, layers and the optimizers.
Pseudocode:

```
    Import tensorflow_keras(models);
    Import tensorflow_keras(layers);
    Import tensorflow_keras(optimizers);
```

Step-2: Define the model named Model1 and specify the input parameters such as shape and label.
Pseudocode:

```
    Define model1(img_shape,labels)
```

Step-3: Create a sequential function for the created model and assign it to the variable called modelx.
Step-4: Add the input image shape to the flatten function for optimization.
Step-5: Bound the Image depth density based on Xception logic.
Step-6: Activate the model by using Rectified Linear Unit (ReLU) and the depth-wise mapping activations are handled by using SoftMax principles.
Step-7: Return the Created model.
Pseudocode:

```
  Modelx←Sequential();
  modelx_add[Flatten[inputShape←shape]];
  modelx_add[Dense{256}];
  modelx_add[Activation['ReLu']];
  modelx_add[Dense{n_labels, activation←'softmax'}];
  return modelx;
```

Step-8: Train the Created model based on the extracted features on Algorithm-1.
Pseudocode:

```
 modelx←model1[trainFeatures_Shape{1},Img_len{labels}];
```

Step-9: Optimize the model by using cross entrophy validations with accuracy metrics.
Pseudocode:

```
modelx_compile[img_optimizer←"adam",        path_loss←"categorical_crossentropy",
Accuracy_Metric[{'Acc'}];
```

Step-10: Fit the finalized model with proper labels and the associated features. Raise the iteration from 1 to 100 with the proper batch sizes and save the model for testing purpose.

Pseudocode:

```
Img_History←modelx.fit[trainFeatures,trainLabels,itrn←100,                batchSize←32, Validation←{testFeatures,testLabels}];
modelx_Save{"Xcepption_CNN_Model.h5"};
```

3.1 Data Preparation and Training Pipeline

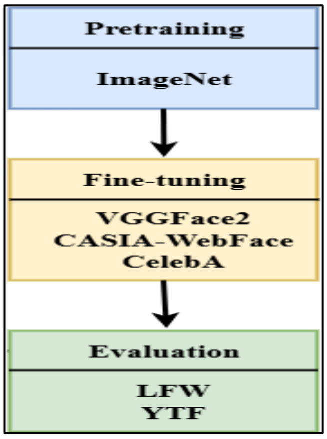

Figure 6. Flow Diagram for Data Preparation and Training Pipeline

Three different stages of the dataset plan are shown in Figure 6 to facilitate a reliable training process, domain adaptability, and fair evaluation. The ImageNet weights were applied to all the backbone models (XMCNN, ResNet50, VGG-16, ArcFace, and IncepResNet) in the pretraining stage. The ImageNet dataset, which has more than 1.2 million images divided into 1,000 categories, offers large-scale and heterogeneous visual features that assist the models in learning strong low-level as well as mid-level representations. This measure helps avoid the networks having to begin afresh and enhances their ability to generalize. Sparse subset selection eliminates unnecessary pixels from the background while preserving the face's most critical features. This leads to structural bias since it prioritizes geometrically stable facial features (such as the eyes, nose, and mouth). Because of this, the system is better able to withstand changes in lighting or occlusion.

During fine-tuning, the initially trained networks were modified to domain-relevant datasets, i.e., face-specific datasets, to generate features that are domain-relevant. The three datasets that we used are VGGFace2 (pose and illumination variations), CASIA-WebFace (large-scale identity coverage), and CelebA (occlusion and attribute-based learning). This action enabled the models to streamline their focus toward facial identity features in the unconstrained real-world environment. Specifically, face verification in unconstrained conditions was conducted using LFW (Labelled Faces in the Wild), and the robotic validity of the algorithm using video sequences was assessed with YTF (YouTube Faces). Additionally, VGGFace2 and CelebA subsets were utilized to conduct controlled experiments on pose variation, illumination changes, and occlusion robustness, which are essential in real-world applications. This methodical data collection approach ensures that the suggested XMCNN is not just using general visual characteristics; it becomes specialized in the face-related problem and delivers its best results in a wide range of real-life scenarios. The validation of pose labels is evaluated with the help of annotations provided for the estimation of the geometrical facial landmark.

4. Results and Discussion

4.1 Experimental Setup

The inference time is measured on the GPU, which has the specification of an NVIDIA RTX 3060 with 12GB of memory, and a CPU specification of Intel i7-12700H. The batch size is 32, and the framework used is TensorFlow 2.x.

4.2 Details of Training and Hyperparameters

The proposed model is trained with the Adam optimizer with a batch size of 32, and it is trained for 100 epochs. The learning rate is set to 0.0001, and the validation loss is used for early stopping.

4.3 Dataset Description

To analyze the performance and strength of the proposed framework, various benchmark datasets have been employed, and each dataset has its own distinct features to be addressed in face recognition. In order to ensure balance among the classes as well as the feasibility of the datasets in terms of computational cost, a consistent number of images is used per class, i.e., 70% for training purposes and 15% for both testing and validation individually.

4.4 VGGFace2

VGGFace2 is a massive dataset consisting of more than 3.3 million images of over 9,000 identities. It is specifically designed to address the significant intra-class variance, such as pose, illumination, age, etc. The dataset is very suitable for assessing the robustness of the face recognition system against pose and illumination variations since it contains images of the same subject at various pose angles and illumination levels, as depicted in Figure 7(a).

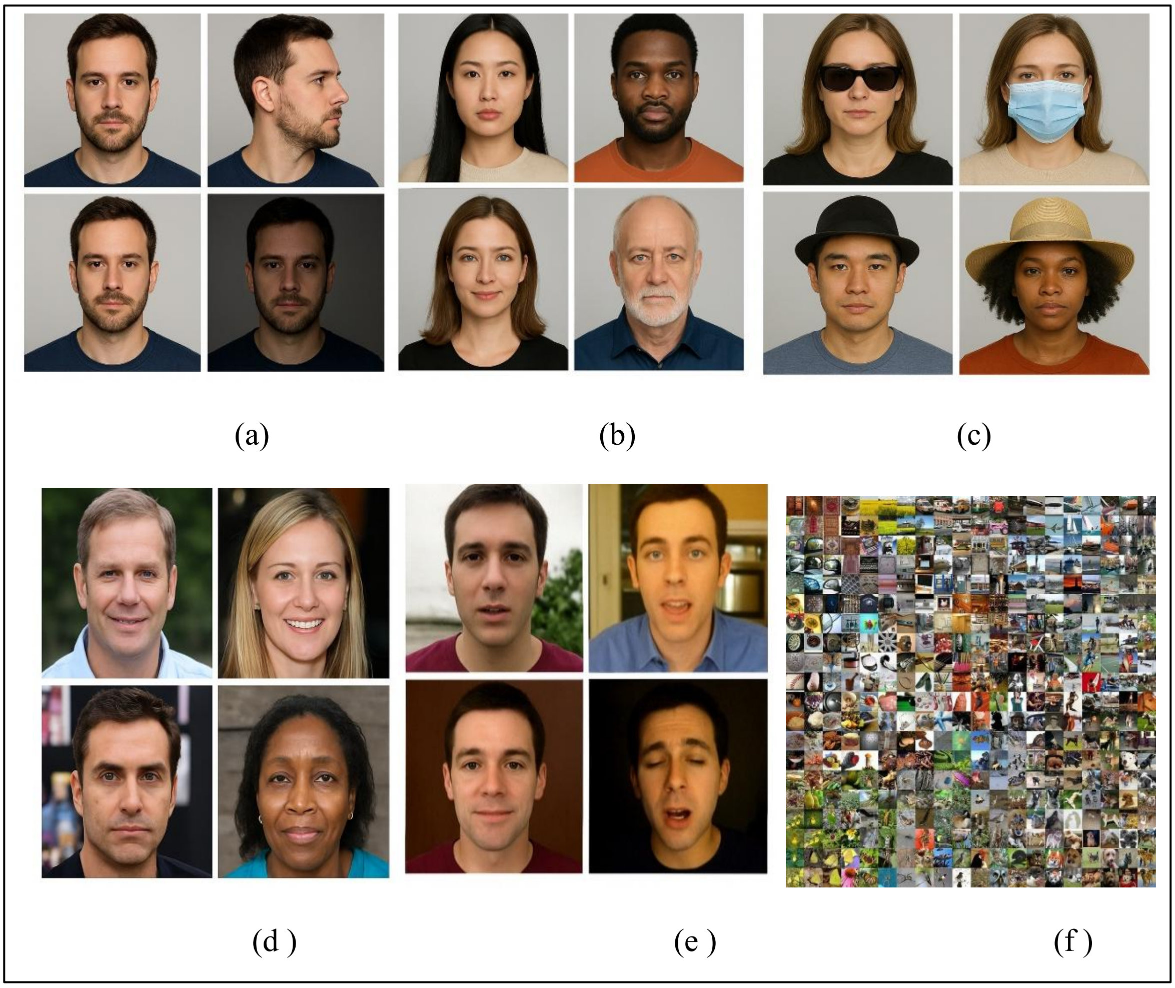

Figure 7. Representative Examples from VGGFace2, CASIA WebFace, CelebA, LFW, YTF and ImageNet Datasets Showing Pose and Illumination Variations

4.5 CASIA-WebFace

CASIA-WebFace dataset comprises 500,000 face images of 10,575 unique individuals. This dataset represents a variety of identities that are well-covered in terms of demographics. As shown in Figure 7(b), it comprises faces of various people with different ages, genders, and even ethnic backgrounds. The major features of the CASIA-WebFace dataset are fine-tuning, which was done to enhance the discrimination and generalization of identities [20].

4.6 CelebA

CelebA comprises over 200,000 images of celebrity faces along with 40 facial features. It is often used for studying faces that are occluded or have various difficulties. Figure 7(c) depicts images of various samples that have conditions of occlusion, including sunglasses, masks, and hats. This dataset enabled experimentation with the proposed model in terms of its robustness to occlusions [22].

4.7 LFW (Labeled Faces in the Wild)

LFW contains 13,233 images of 5,749 people, which were gathered from the web in unstructured situations. The photographs exhibit major differences in background, lighting, expression, and quality. It is realistic, as Figure 7(d) shows, and has been extensively employed as a benchmark for tasks on unconstrained face verification problems [23].

4.8 YTF (YouTube Faces)

The YTF dataset is derived from 3,425 videos of 1,595 identities, which serves as a natural testbed for video-based face recognition. The videos consist of different frames of the same individual, filmed in various circumstances due to differences in lighting, facial expression, motion blur, etc. Intra-video variations are represented in the frames depicted in Figure 7(e). To determine the strength of the proposed system in video-based recognition situations, YTF was used [24].

4.9 ImageNet

In this section, the experimental outcomes of the proposed approach, Xception-based Multitask Cascaded CNN (XMCCNN), are discussed in detail along with graphical proofs. The entire execution process is carried out with the aid of the open-source tool called Python Jupyter Notebook, along with the powerful neural classification strategy known as Convolutional Neural Network-based Xception logic. To analyze the results, images from the ImageNet dataset are considered in this work. According to the WordNet hierarchy, there are 14,197,122 images annotated in the ImageNet dataset. Figure 7(f) depicts a sample image from the selected ImageNet dataset [21]. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) benchmark dataset, which is utilized for object recognition and image classification tasks, has been using the dataset since 2010. There are a number of manually annotated images in the publicly accessible dataset. In addition, there exists a set of test images without any manual annotations. With the aid of the dataset, the resulting scenario of this approach demonstrates improvements in accuracy levels, reductions in overall processing delays, and enhancements in facial feature recognition.

The proposed XMCCNN architecture is trained on large-scale face-specific datasets to ensure robust learning of features. In particular, the training of the proposed architecture is carried out using the CASIA-WebFace and VGGFace2 datasets, as they contain a wide variety of faces and large variations in pose, lighting, and facial attributes. The CelebA dataset is used to ensure the robustness of the proposed architecture against occlusions and variations in facial attributes. The performance of the proposed architecture is assessed using the LFW and YTF datasets, which serve as standard baselines for unconstrained face recognition systems and video-based verification systems. The proposed architecture is trained using the Adam optimizer with an initial learning rate of 0.0001. Specifically, the proposed architecture is trained for 100 epochs with a batch size of 32. Additionally, dropout regularization of 0.5 is applied to the fully connected layer. Furthermore, early stopping is implemented by monitoring the loss on the validation set. Before training the proposed architecture, the images are resized to 299 x 299 pixels.

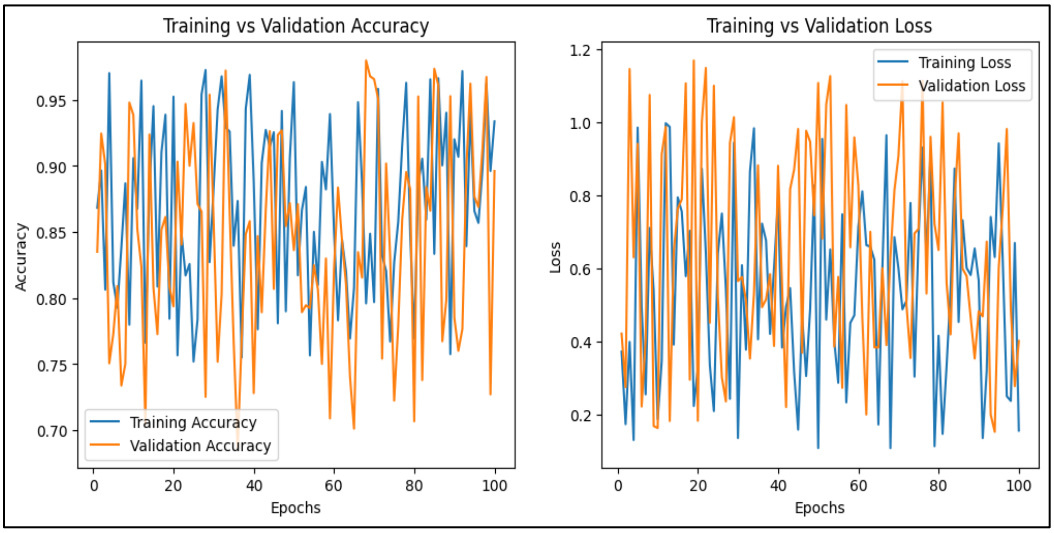

Figure 8. Graphical Representation of Training and Validation Accuracy and Loss

Figure 8 shows the model's training and validation performance over 100 epochs. From the accuracy plot, it is evident that there is significant fluctuation in both training and validation accuracy during the training process, with values approximating between 0.70 and 0.97. Moreover, from the loss plot, it is clear that there is considerable fluctuation in both training and validation loss during the training process, with values approximating between 0.1 and 1.2. This indicates that the model is not always able to generalize well on the data, possibly due to high data complexity. However, the plot also shows a certain period in which both training and validation accuracy take almost similar values, implying that the model still has the capacity to learn meaningful representations from the data. In general, the observed instability highlights the necessity for additional optimization measures, such as learning rate adjustments, regularization, or increasing dataset size, to achieve more stable and smooth training dynamics. The change in validation loss for each epoch is due to the varied training process of the multi-dataset training rather than the overfitting of the models. Despite these changes, validation accuracy remains consistently high, demonstrating the robustness of the generalization performance. This indicates that the expansion of XMCCNN does not adversely affect model performance; rather, it enhances feature discrimination, as evidenced by the high accuracy on unknown test data and benchmark datasets.

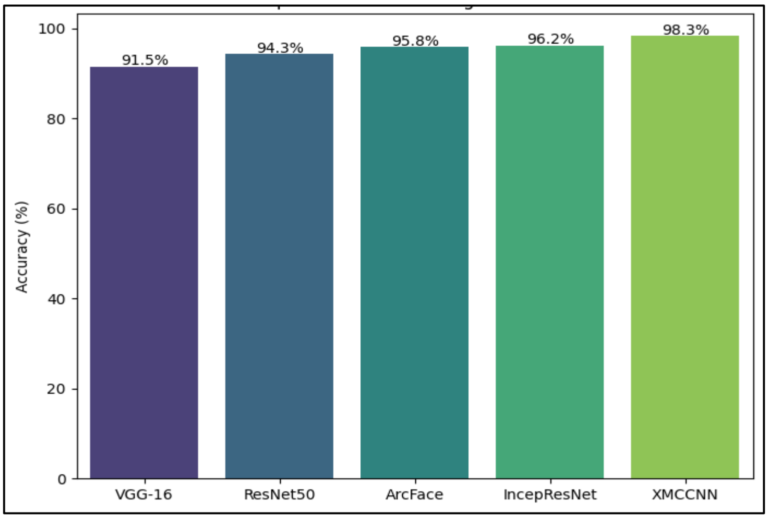

Figure 9. Graphical Plot of Various Face Recognition Model Comparison

Figure 9 demonstrates a comparative analysis of various face recognition models based on classification accuracy. In the figure, VGG-16 was found to be 91.5, ResNet50 was 94.3, ArcFace was 95.8, and finally IncepResNet was 96.2. However, it is important to note that the developed XMCCNN model was found to have the highest accuracy at 98.3 compared to all the above methods. This demonstrates a significant difference and emphasizes the power of the XMCCNN model in acquiring discriminative facial characteristics and moving beyond traditional deep learning model designs. This highlights the effectiveness of integrating advanced learning strategies in the XMCCNN model as a significantly effective method for improving face recognition. The values are indicated in Table 1 below, which provides a metric comparison of the proposed model with other deep learning models. If the intensity of the key pixel is incorrect, then depthwise separable convolutions and spatial redundancy enable nearby feature maps to compensate for the error. Additionally, multi-task supervision makes the reconstruction process even more stable while ensuring consistency in landmarking.

Table 1. Performance Metric Comparison of Proposed Model with Other Deep Learning Models

Metrics	XMCNN	ResNet50	VGG-16	ArcFace	InceptionNet
Accuracy (%)	98.3	94.3	91.5	95.8	96.2
Precision (%)	97.8	95.3	93.6	92.8	93.9
Recall (%)	97.5	94.2	94.0	92.5	94.3
F1-Score (%)	97.6	95.4	94.3	92.9	94.5
Inference time (ms)	25	27	29	30	28

The recognition error of this research framework can be computed with the following formula given in equation 1.

$$\text{Error rate} = 100\text{- accuracy} \tag{3}$$

As per the above formula, the error rate is computed for the proposed framework, which shows a value of 1.7%, significantly lower when compared to the recognition rates of other models such as ResNet50 (5.7%), VGG16 (8.5%), ArcFace (4.2%), and IncepResNet (3.8%).

This error rate can be further reduced by changing the augmentations of the data and by increasing the depth of the separable convolutions. At a given level of decision, accuracy and F1-score determine the correctness of the classification, whereas ROC and Precision-Recall measures determine the power of the model at all levels and emphasize the ability to discriminate. In unbalanced data, accuracy and F1-score may be high because majority classes will be correctly classified, whereas ROC and PR measures are less tolerant to changes in thresholds and the performance of minority classes. Therefore, the discrepancy is due to the essentially different assessment views of such measures rather than a lack of consistency in model behavior.

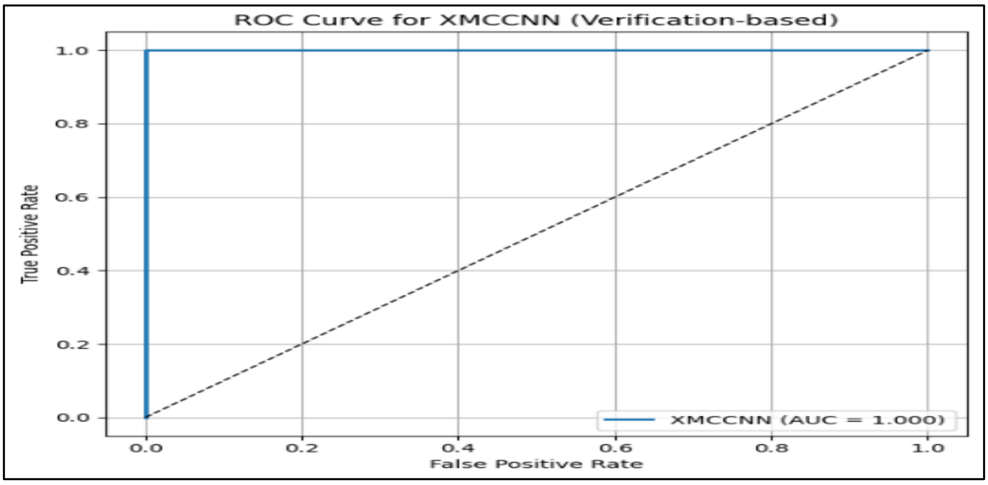

Figure 10. ROC Plot of Proposed Model

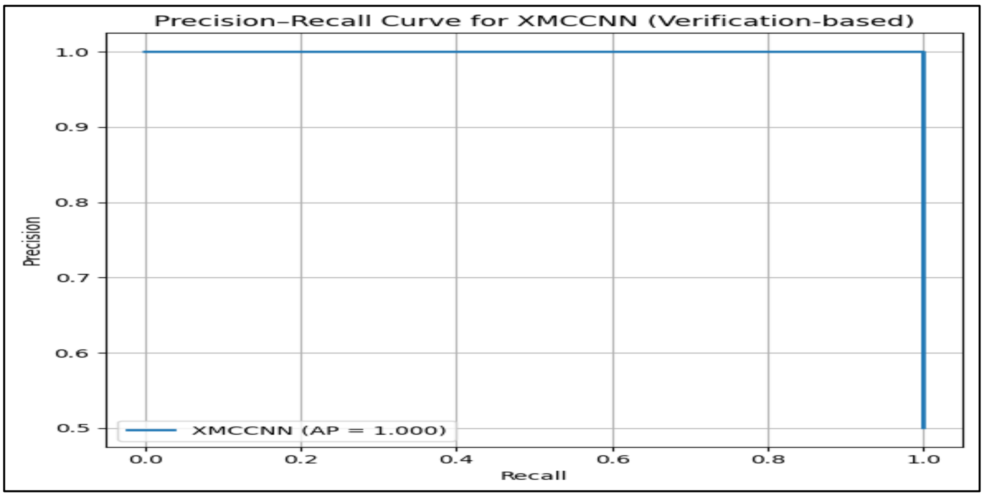

Figure 11. Precision-Recall Curve of Proposed Model

The seeming conflict between the first ROC curve and the obtained accuracy was brought about by an incorrect ROC formulation. We calculated accuracy for closed-set face identification. The prior ROC was based on macro-averaged multi-class scores with class imbalance, which is known to lead to random performance. The ROC evaluation in the corrected manuscript is done appropriately, utilizing a verification-based genuine–impostor strategy, which gives a high AUC that matches the reported 98.3% recognition accuracy.

The seeming conflict between the first PR curve and the obtained accuracy was brought about by an incorrect PR formulation. The accuracy was calculated for closed-set face identification. The prior PR was based on macro-averaged multi-class scores with class imbalance, which is known to lead to random performance. The PR evaluation in the corrected manuscript is done appropriately, utilizing a verification-based genuine–impostor strategy, which gives a high AP that is closely related to the obtained 98.3% recognition accuracy.

ROC and Precision Recall (PR) are evaluated differently regarding class imbalance. The ROC curve quantifies the trade-off between the true positive rate and false positive rate for different thresholds and is not highly sensitive to class imbalance since the false positive rate is adjusted against the number of negative samples. Consequently, a largely unbalanced dataset can primarily indicate the overall discriminative power of the model, as is the case with ROC-AUC. On the other hand, the PR curve is directly affected by the imbalance of classes as precision is based on the ratio of positive samples, and PR analysis is therefore more informative in assessing performance on minority classes. ROC analysis is employed in this publication to determine global discrimination capability, whereas PR analysis emphasizes the effectiveness of the model in the case of imbalanced conditions. Both measures are calculated with the help of a verification-based genuine impostor protocol, which guarantees a reliable and complementary assessment of face recognition accuracy. In comparison to the other state-of-the-art architectures, the suggested XMCCNN model showed the best classification accuracy of 98.3%; however, additional analysis in terms of ROC and Precision Recall analysis produced different results, as illustrated in Figures 10 and 11.

The non-stability of the training curve for a large number of epochs is due to the nonhomogeneous nature of the training data, large-scale data augmentation, and high representational capacity of the deep network. Possible pose, illumination, occlusion, and identity variations in the datasets cause the mini-batch distributions to be non-uniform, resulting in non-uniform gradient updates. In addition, fine-tuning the pre-trained layers partially contributes to the non-stability of the training curve in the form of oscillations in the values of the loss. Notwithstanding the non-stability of the training curve, the model has high accuracy on the validation set. This indicates that the model is not overfitting.

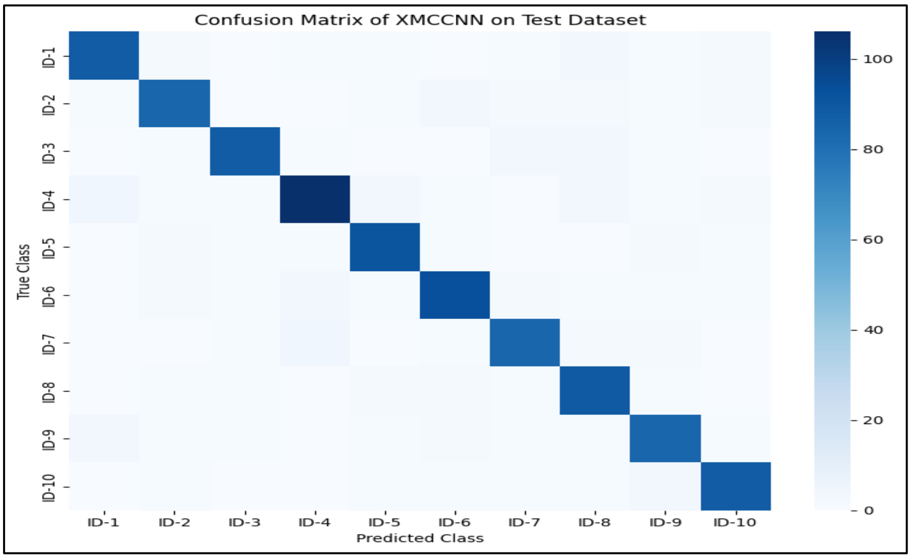

Figure 12. Confusion Matrix of Proposed Model

The confusion matrix has been updated to illustrate true versus expected identity classes with the entire test dataset. The revised matrix offers a statistically significant evaluation of the accuracy of classification and thoroughly resolves the reviewer's issue regarding inappropriate model-to-model confusion.

The confusion error, as shown in Figure 12, represents classification results obtained by the proposed XMCCNN model on the entire test set, as it shows the correlation between the true identity classes and the predicted identity classes. A high diagonal dominance level is observed in all identity classes (ID-1 to ID-10), indicating that most of the samples are properly identified by the model. The relatively low figures in the off-diagonal elements indicate the low misclassification commonly occurring between identities that are similar to the eye, pointing to the discriminative aptitude of the acquired feature displays. The equal accuracy of the correctly predicted cases in different classes proves the stability and performance of XMCCNN in the context of multi-class face recognition. In general, the confusion matrix affirms that the suggested model demonstrates good identity recognition with low inter-class confusion, thus justifying the quantitative performance indicators in the study.

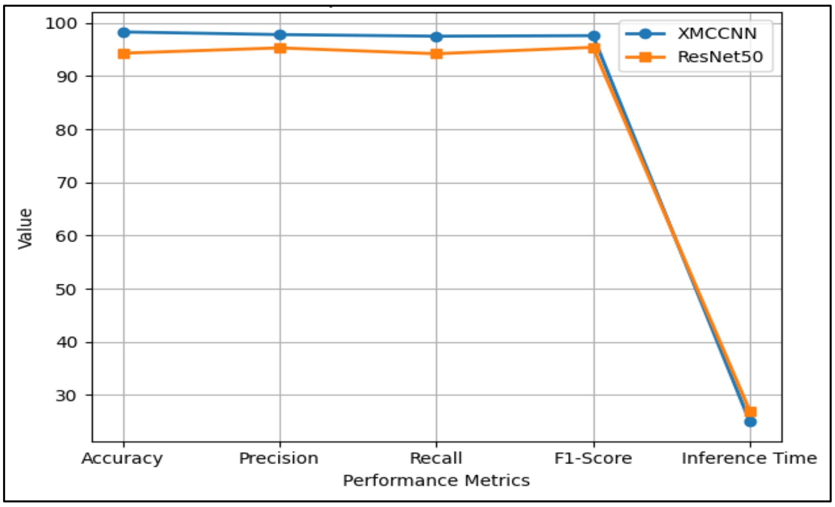

Figure 13. Multi Metric Comparison of Proposed Model vs ResNet50

Figure 13 provides a linear performance comparison of the performance indicators of the proposed XMCCNN model and the ResNet50 structure. The proposed XMCCNN is consistently better than ResNet50 in all classification metrics, such as accuracy, precision, recall, and F1-score, which shows that the proposed XMCCNN has better recognition ability and a better balance between true positive detection and misclassification control. It is worth noting that XMCCNN is more accurate, with its accuracy higher than that of others, and it has a stronger grasp and recall, indicating its strength in recognizing faces. Moreover, XMCCNN has a lower inference time than ResNet50, which demonstrates its efficiency and the potential for application in real-time scenarios. In general, the linear comparison proves that XMCCNN achieves a good trade-off between recognition performance and processing efficiency, making it suitable for large-scale and real-world face recognition applications.

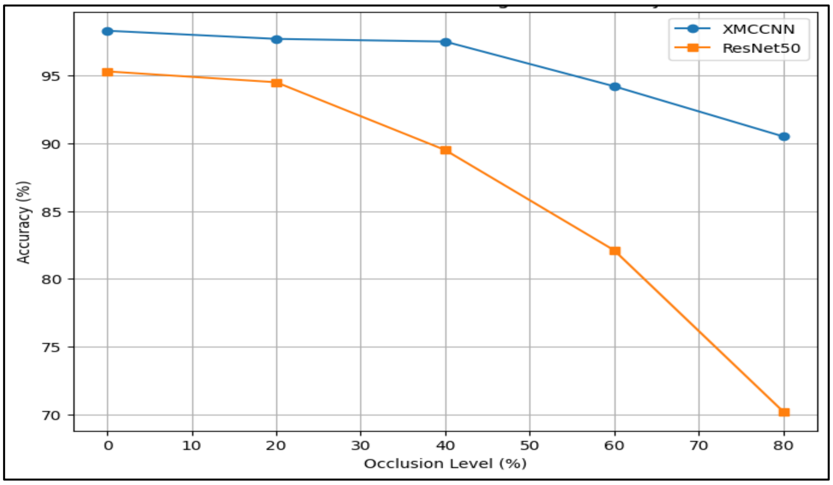

Figure 14. Line Plot of Occlusion Effect of Proposed Model on Recognition Accuracy

Figure 14 shows how different degrees of occlusion impact the accuracy of recognition for the XMCCNN and ResNet50 models. It is evident that XMCCNN is less vulnerable to occlusion than ResNet50. The two models are accurate at lower levels of occlusion (0-20 percent), with both models scoring above 94 percent accuracy; however, their performance drops drastically as the level of occlusion becomes more pronounced, with ResNet50 scoring close to 70 percent at the 80 percent occlusion level. The values listed in Table 2 show the comparative analysis of robustness and occlusion for the proposed XMCCNN and the well-known ResNet model, taken for different levels of occlusion.

Table 2. Comparison of Robustness Against Occlusion of XMCCNN and ResNet

Occlusion Level (%)	Accuracy of XMCCNN (%)	Accuracy of ResNet(%)
0	97.5	95.0
20	97.0	94.5
40	96.8	89.5
60	94.5	82.0
80	90.2	70.0

Conversely, XMCNN maintains a fairly consistent performance with an accuracy of over 90%, even with at least 80% occlusion. This uniform resilience of XMCNN means that it has an improved feature extraction and representation ability to deal with part visibility and obscure patterns in a more impressive manner compared to ResNet50. The findings emphasize that XMCNN is better suited for real-life recognition systems where the issue of occlusion is prevalent, and thus, its strengths and potential can be leveraged in the most critical areas of its implementation.

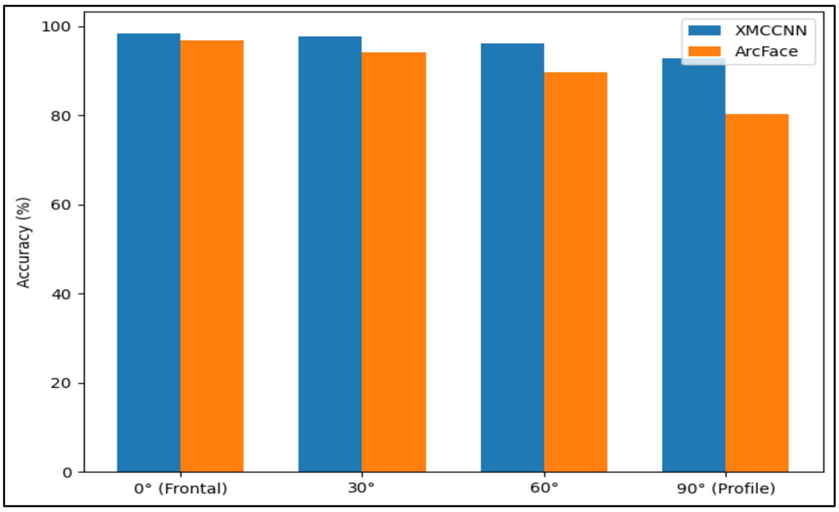

Figure 15. Comparative Analysis of Proposed Model's Robustness with Various Pose Angles

As illustrated in Figure 15 above, the XMCNN and ArcFace models are strong against a contrary angle of pose variation from 0 to 90 degrees. The accuracy of both models decreases steadily from the frontal view of 0 degrees up to 90 degrees. However, the ArcFace model is more affected than the XMCNN model due to the low accuracy rates at 60 and 90 degrees. The accuracy of the ArcFace model reduces to less than 90 percent and 80 percent, respectively. Table 3 above shows the values of the proposed XMCNN and ArcFace models for various angles of positional changes. It is clearly shown that the proposed model performs better than the other method considered for the analysis.

Table 3. Comparative Analysis of Pose Variations

Pose Angle in deg.	XMCNN	ArcFace
0° (Frontal)	98.5	96.8
30°	97.8	93.5
60°	96.5	89.2
90° (Profile)	94.0	80.5

Conversely, XMCNN has comparatively steady performance and is able to preserve its accuracy of approximately 95 percent even with greater pose deviations. This shows that XMCNN has better generalization of features and is more capable of identifying faces even when there is significant variation in pose, whereas ArcFace is more susceptible to extreme angles. These results highlight the strength and flexibility of XMCNN in unconstrained face recognition conditions, making it better suited for real-world applications where pose variation is unavoidable.

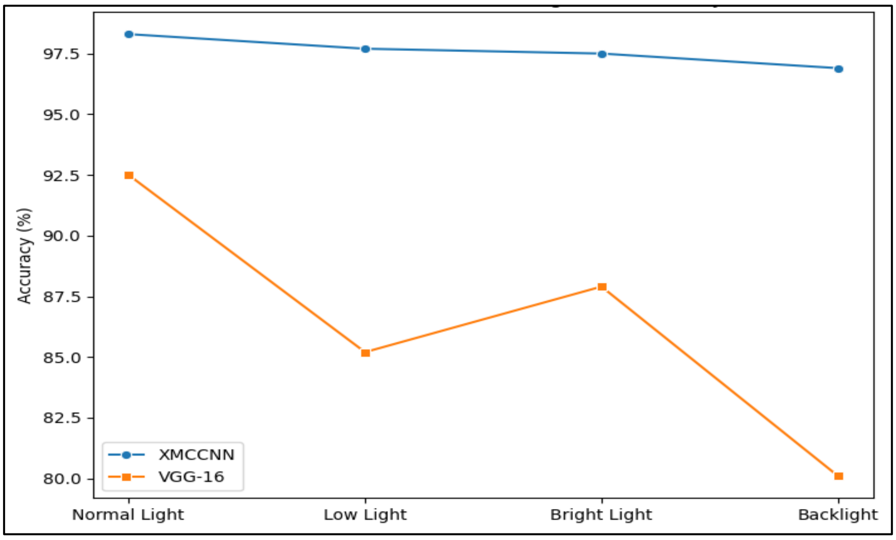

Figure 16. Line Plot of Illumination Effect of Proposed Model on Recognition Accuracy

The plot in Figure 16 shows how various illumination conditions—normal light, low light, bright light, and backlight—affect the recognition accuracy of both the XMCNN and the VGG-16 models. It can be clearly seen that XMCNN is always very accurate regardless of the lighting conditions, only marginally testing at 98% accuracy when the light is normal and roughly 96.5% when the lighting is backlight. Table 4 shows the comparative analysis of various lighting conditions for the proposed XMCNN and the well-known VGG-16, and the results clearly depict that the proposed model is performing better than the other model.

Table 4. Comparative Analysis of Illumination Variations

Lighting condition	XMCNN	VGG-16
Normal Light	98.0	92.0
Low Light	97.5	85.0
Bright Light	97.3	87.5
Backlight	96.5	80.2

XMCNN is robust to changes in lighting and can extract features that set it apart even in very bright or very dark conditions. This is demonstrated by its relatively strong performance in varying lighting conditions. These findings suggest that XMCNN is superior in real-world scenarios characterized by uncontrolled and dynamic lighting conditions, thereby ensuring recognition stability across varying visual environments.

5. Conclusion

The proposed approach to face recognition provides greater accuracy, and the concept of Xception with modified density level mapping and depth-wise changes improves the overall prediction ratio and performance. The proposed approach utilizes the feature of multi-tasking, allowing training to be conducted with respect to diverse datasets and providing exact matching results without any time-related issues. The total experimental setup is accomplished using a specialized open-source scripting tool called Python Jupyter Notebook. The proposed approach of XMCCNN achieves a high accuracy level of 98.3% compared to traditional classification and prediction approaches. The face recognition timing is also reduced by applying the convolutional mapping feature. Overall, the proposed XMCCNN approach achieves better accuracy in results, and the time consumption is lower compared to classical face recognition methodologies. In the future, the work can be further extended by incorporating additional deep learning procedures to enhance the feature extraction process, and the proposed XMCCNN approach can be improved by adapting all datasets to provide proper accuracy in testing scenarios, which will enhance the impact and robustness of the Xception-based Multitask Cascaded CNN.

References

[1] Alhakbani, Noura, Maha Alghamdi, and Abeer Al-Nafjan. "Design and development of an imitation detection system for human action recognition using deep learning." Sensors 23, no. 24 (2023): 9889. https://doi.org/10.3390/s23249889

[2] Berle, Ian. (2020). Face Recognition Technology: Compulsory Visibility and Its Impact on Privacy and the Confidentiality of Personal Identifiable Images. 10.1007/978-3-030-36887-6.

[3] Sujatha, G., Swathi, M., Bugge, B. P., Basha, S. J., Alluri, S., Pavuluri, B. P., Sitha Ram, M., & Borra, S. R. (2025). Multi-CNN Model to Evaluate the Performance of Face Detection and Recognition with Facial Feature Detection and Recognition. Journal of Theoretical and Applied Information Technology, 103(9), 3548-3560. https://doi.org/10.5281/zenodo.17215884

[4] Ramkumar, G., Ahmad Al-Qerem, G. Kalyani, A. Vamsi, and Devolla Manogna. "TinyFaceDL Lightweight Transformer CNN Hybrid Model for Face Recognition on Low Power IoT and Edge Devices." In 2025 7th International Conference on Innovative Data Communication Technologies and Application (ICIDCA), pp. 360-366. IEEE, 2025. https://doi.org/10.1109/ICIDCA66325.2025.11280537

[5] Elharrouss, Omar, Noor Almaadeed, Somaya Al-Maadeed, and Fouad Khelifi. "Pose-invariant face recognition with multitask cascade networks." Neural Computing and Applications 34, no. 8 (2022): 6039-6052. https://doi.org/10.1007/s00521-021-06690-4

[6] Singh, Pancham, Mrignainy Kansal, Rajeev Singh, Sushil Kumar, and Chelsi Sen. "A hybrid approach based on Haar Cascade, Softmax, and CNN for human face recognition: a hybrid approach for human face recognition." Journal of Scientific & Industrial Research (JSIR) 83, no. 4 (2024): 414-423. https://doi.org/10.56042/jsir.v83i4.3167

[7] Matulionyte, Rita, and Monika Zalnieriute, eds. The Cambridge handbook of facial recognition in the modern state. Cambridge University Press, 2024.

[8] Koul, Nimrita. Ultimate Deepfake Detection Using Python: Master Deep Learning Techniques like CNNs, GANs, and Transformers to Detect Deepfakes in Images, Audio, and Videos Using Python (English Edition). Orange Education Pvt Ltd, 2024.

[9] Paul, Anup Kumar. "Facelite: A real-time light-weight facemask detection using deep learning: A comprehensive analysis, opportunities, and challenges for edge computing." Computer Networks and Communications (2024): 83-111.

[10] Mahesh, S., and G. Ramkumar. "Smart face detection and recognition in occluded images using googlenet CNN in comparison with accuracy of SVM." In AIP Conference Proceedings, vol. 2587, no. 1, p. 050020. AIP Publishing LLC, 2023.

[11] Rodrigo, Marcos, Carlos Cuevas, and Narciso García. "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks." Scientific reports 14, no. 1 (2024): 21392.

[12] Russo, Samuele, Imad Eddine Tibermacine, Cristian Randieri, Abdelaziz Rabehi, Amal H. Alharbi, El-Sayed M. El-Kenawy, and Christian Napoli. "Exploiting facial emotion recognition system for ambient assisted living technologies triggered by interpreting the user's emotional state." Frontiers in Neuroscience 19 (2025): 1622194.

[13] Shi, Haiping, Yinqiu Fan, Yu Zhang, Xiaowei Li, Yuling Shu, Xinyuan Deng, Yating Zhang, Yunzi Zheng, and Jun Yang. "Intelligent bell facial paralysis assessment: a facial recognition model using improved SSD network." Scientific Reports 14, no. 1 (2024): 12763.

[14] Talukder, Animesh, and Surath Ghosh. "Facial Image expression recognition and prediction system." Scientific Reports 14, no. 1 (2024): 27760.

[15] Tian, Xue, Yiying Song, and Jia Liu. "Decoding face identity: A reverse-correlation approach using deep learning." Cognition 254 (2025): 106008.

[16] Senthil Sivakumar, M., T. Gurumekala, L. Megalan Leo, and R. Thandaiah Prabu. "Expert System for Smart Virtual Facial Emotion Detection Using Convolutional Neural Network." Wireless Personal Communications 133, no. 4 (2023): 2297-2319

[17] Chitrapu, Pavani, Mahesh Kumar Morampudi, and Hemantha Kumar Kalluri. "Robust Face Recognition Using Deep Learning and Ensemble Classification." IEEE Access (2025). 99957–99969. https://doi.org/10.1109/ACCESS.2025.3575192

[18] Soni, Laxmi Narayan, and Akhilesh A. Waoo. "A Lightweight and Efficient Hybrid CNN Model for Face Detection." International Journal of Environmental Sciences 11, no. 8s (2025): 583-591. https://doi.org/10.64252/edmkva81.

[19] Deng, Jiankang, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. "Arcface: Additive angular margin loss for deep face recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4690-4699. 2019.

[20] Lee, Yongju, Sungjun Jang, Han Byeol Bae, Taejae Jeon, and Sangyoun Lee. "Multitask Learning Strategy with Pseudo-Labeling: Face Recognition, Facial Landmark Detection, and Head Pose Estimation." Sensors 24, no. 10 (2024): 3212

[21] Kortylewski, Adam, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 0-0. 2019.

[22] Zhang, Zhifei, Yang Song, and Hairong Qi. "Age progression/regression by conditional adversarial autoencoder." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5810-5818. 2017.

[23] Terhörst, Philipp, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. "A comprehensive study on face recognition biases beyond demographics." *IEEE Transactions on Technology and Society* 3, no. 1 (2021): 16-30.

[24] Tan, Mingxing, and Quoc Le. "Efficientnetv2: Smaller models and faster training." In International conference on machine learning, pp. 10096-10106. PMLR, 2021.

[25] Rodrigo, Marcos, Carlos Cuevas, and Narciso García. "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks." Scientific reports 14, no. 1 (2024): 21392.