An Evaluation of the performance of Convolution Neural Network and Transfer Learning on Face gender Recognition*

Ronke Seyi Babatunde[1], Akinbowale Nathaniel Babatunde[1], Roseline Oluwaseun Ogundokun[2,3*[0000−0002−2592−2824]], AbdulRahman Tosho Abdulahi[4,[0000−0003−2899−9883]], and Alfonso González-Briones[5[0000−0002−3444−4393]]

[1] Department of Computer Science, Kwara State University, Malete, Kwara State, Nigeria. {ronke.babatunde, akinbowale.babatunde}@kwasu.edu.ng
[2] Department of Multimedia Engineering, Kaunas University of Technology, Kaunas, Lithuania. rosogu@ktu.lt
[3] Department of Computer Science, Landmark University Omu Aran, Nigeria. ogundokun.roseline@lmu.edu.ng
[4] Department of Computer Science, Institute of Information and Communication Technology, Kwara State Polytechnic, Nigeria. AbdulRahman.t@kwarastatepolytechnic.edu.ng
[5] Grupo de Investigación BISITE, Departamento de Informática y Automática, Facultad de Ciencias,University of Salamanca, Instituto de Investigación Biomédica de Salamanca, Calle Espejo 2, 24.2, 37007, Salamanca, Spain alfonsogb@usal.es

Abstract. Gender detection has become an important component of the biometric system and protection. It is used to recommend age-appropriate content and to broaden the reach of multilayer advertising and promotions, among others. Many challenges emanate from using contemporary machine learning algorithms for face gender recognition, ranging from low detection accuracy and overfitting. Most recognition tasks, including object identification, segmentation, and classification, are based on convolutional neural networks. The human Face is a nonlinear manifold in which slight input changes can result in the invariance of the output. This prompts the need for continuous research to achieve more accurate recognition, especially with experimenting on ground truth data. This research compares the performance of CNN and TL on faces acquired in real-time scenarios to detect gender. The empirical result based on the study demonstrates that TL outperforms CNN in terms of prediction accuracy, with TL having 95.83% and CNN having 83.11%, and also in terms of recognition time, with TL having 0.23secs and CNN having 0.98 secs, amongst other evaluation metrics. The novelty of the research is in the recognition time, which is not a metric commonly reported in the literature. The work demonstrates the effectiveness of using pre-trained

* This research has been supported by the project "COordinated intelligent Services for Adaptive Smart areaS (COSASS), Reference: PID2021-123673OB-C33, financed by MCIN /AEI /10.13039/501100011033 / FEDER, UE.

models based on the dataset used and evaluation metrics for the experiment.

Keywords: Gender detection· CNN· Transfer learning· Face gender recognition

1 Introduction

Gender recognition was first sighted as a challenge in psychological research. Several academics are now focusing on human detection, gender identification, age prediction, sentiment evaluation, and race identification. Although this gender identification is a binary classification issue, it might be considered the easiest face estimation challenge. In contemporary technology, the ability to determine a person's gender by face analysis has become very important. Face recognition assessment has evolved into one of the greatest significant subfields in computer vision, as it plays a crucial role in social contact. Automated gender categorisation has major applications in smart user interface design, biometric detection, interpersonal interactions, visual monitoring, collecting demographic information for advertisement, behaviour recognition, and numerous other fields. These responsibilities include characterising age, gender, facial characteristics, gestures, and personality. As a result, much research has gone into creating automatic systems that can classify genders. The primary goal of gender recognition is to distinguish between males and females founded on various human characteristics. The two eyes, nose, and mouth are basic facial traits used to distinguish between genders González-Briones et al. (2018) and Lawal et al. (2019). Although this task has been extensively studied in the past, the reported performances are far from ideal, particularly in unrestricted settings. Furthermore, the application context and training procedures significantly impact how difficult this task is Sumi et al. (2021). In-dataset inference, also known as an inference from separate datasets, can be used to train and test gender classification models. Additionally, these datasets allow for capturing facial images in both controlled and uncontrolled/unconstrained contexts, which adds to the difficulty of the endeavour. Jia et al. (2016) achieved one of the most modern methods for gender classification by employing 4 million photos to train the approach and testing it on the unrestricted Labelled Face in the wide (LFW) dataset. Convolutional neural networks, a kind of deep neural network, have recently emerged as the industry standard for object recognition. A CNN is a deep learning system that distinguishes one image from another while allowing an input image to take on several objects. One of the reasons why convolutional neural networks have emerged as one of the most popular algorithms with a strong reputation in picture feature extraction is that they require significantly less pre-processing than other classification techniques. Due to their outstanding performance in facial analysis, the design models of this deep-learning neural network have been used in many detection tasks. Nearly every aspect of computer vision has improved because of it, including the ability to recognise facial expressions, handwritten

numbers, and human actions. Zaman (2020). The development and maturation of facial recognition over time have been made possible by technological progress and the quick rise in the image and video quality. Finding missing children, combating human trafficking, improving healthcare, and other areas have benefited greatly. Healthcare facilities have employed facial recognition in many ways to safeguard the health and safety of patients and staff. These include protecting priceless lab equipment and enabling touchless access so that solitary skilled and approved staff can enter private spaces in a facility, lowering the danger of contamination. However, a user-friendly mechanism and timely and easy-to-interpret interface for gender recognition is still lagging. Hence, this research aims to compare the effectiveness of the face gender recognition scheme using CNN and Transfer learning.

2 Related work

Babatunde et al. (2016) developed a real-time automatic system recognising face gender. The face images were captured in real-time via webcam, whereby an individual is made to position themselves in front of the system to acquire a frame. A green outline on the face region indicated a tracked face in the frame. Gender recognition takes effect in nanoseconds once the Face is detected. The work observed that the image quality degraded the system's resolution and varying illumination performance. Afifi and Abdelhamed (2019) conducted comparable cross-dataset tests on face images. According to the work, the face images were collected and assembled into individual datasets in various ways, with occlusions, changes in illumination, backdrops, etc. The paper recommended that a foggy face that combines discrete facial elements with a holistic attribute will be considered instead of treating the face image as a separate feature. This will lead to better classification performance. Hence, the work trained CNN on a cloudy face and combined it with AdaBoost to determine the gender category. For gender categorisation, a combination of local traits and contextual features (the clouded Face) were derived from the images. To extract the necessary attributes for the experiment, four CNNs were used, along with an Ada-boost-based classification. To build the system, the experiment employed five (5) popular datasets: LFW, FERET, Adience dataset, and image of groups dataset. The network was tested and assessed using Specs on Faces (SoF). Their approach produced good validation accuracy. Jewel et al. (2019) applied the CNN structured method and TL to recognise the ethnicity of Bengali public and did a systematic gender categorisation. To determine which transfer learning model is more practical for achieving the requisite accuracy, the work applied numerous models, including VGG16, Mobile net, and Resnet50. Since there are many Indians that dress and appear like Bengalis, the investigation identified various issues. As a result, the model was created using the custom of their clothes and facial images. A complex model was developed utilising CNN and transfer learning, showing that transfer learning led to better results. Moeini and Mozaffari (2017) suggested a technique for gender categorisation based on vocabulary acquisition. The three

vocabularies they utilised were a male glossary, a female glossary, and a feature glossary. The three vocabularies were learned using two vocabulary learning methods: dictionary learning gender classification (DL-GC) and separate dictionary learning gender classification (SDL-GC). Furthermore, they used three (3) publicly available datasets: FERET, LFW, and Groups Dataset. Only two features were employed to extract the essential characteristics and were removed automatically. Sparse Representation Classification was then used to categorise the gender from the estimated values. Convincing results were obtained for gender classification when evaluated on the datasets mentioned above. Comparing other cutting-edge technologies revealed an improvement over prior gender classification accuracy. Sumi et al. (2021) developed a human gender recognition system from facial images utilising CNN. Two different datasets, each made of 100 males and females, were obtained from other sources for comparison. The first source contained up to 100 females and the same for males. The image size used was 96 X 96. There are 32 filters in the convolution layer, each having a 3 X 3 kernel and the pooling layer with a 3 X 3 kernel, thereby reducing the spatial dimension from 96 X 96 to 32 X 32. The convolution and RELU were repeated to allow a richer set of features. The work used Adam optimiser, and the result gave a high accuracy compared with similar systems. The work also discovered that a proper dataset and optimisers could improve recognition accuracy. Sharma et al. (2020) reviewed real-time sentiment, gender, and age recognition utilising CNN. The work used the Viola-Jones algorithm, CNN, and HOG (Histogram of Oriented Gradient). The pre-processing stage enhances the quality of the input image and generated data by removing noise and minimising redundancy. The result shows that the HOG+viola jones algorithm yields a good accuracy of 70% on FER13 datasets. At the same time, the CNN model achieved 97%. Greco et al. (2021) carried out Gender recognition in the wild using a robustness assessment over tainted imageries. Cropped images were used as the network's input, and a face detector based on SSD and ResNet-10 was used as the backbone technique. Then the images were rescaled to adapt to the native input size. A variety of data augmentation techniques were used to replicate changes that are likely to happen in real circumstances. The dataset consists of 9131 subjects and 3.31 million imageries with variations. Different convolutional neural networks were tested, and cross-dataset evaluation was performed. The SGD optimiser was used, and the result showed that VGG attained an accuracy of 96.6%, which is more advanced than the other types of CNN. The work discovered that a large dataset would increase processing time, which may necessitate using high-end systems for real-time implementation. YILDIZ et al. (2021) developed a CNN-founded gender recognition in unrestrained settings. 3.3 million images of more than 9000 unique individuals make up the VGGFace2 dataset, which uses 3.1 million images as train data and 200 000 images as test data. An average of 362 images represents each individual. The images in the dataset differ from each other. Bilinear interpolation was used to resize face images to 32x32x3 before pre-processing. The Adience dataset was used as a benchmark to evaluate the results. The classification was performed using the sigmoid classifier. The VG-

GNet architecture was scaled down and made more manageable. The suggested network has five convolution layers, three pooling levels, and one fully linked layer. The work achieved 96.29% test accuracy and 93.71% validation accuracy. However, comparison between two different datasets may lead to bias in some instances. The deductions from the literature indicate that CNN is widely used in the recognition task. However, a face image's pose, illumination and expression challenges call for continual research to achieve an improved result.

3 Methodology

3.1 Dataset collection and pre-processing

The dataset used in this research consists of two hundred and forty (240) faces of forty (40) students taken from six distinct camera angles, poses and illumination to ensure adequate sample variations. Additionally, 4 of each face image was used to populate the training set to avoid overfitting, while the other two were used to aggregate the testing set. The photos were labelled, resized and converted to a greyscale to provide a suitable template for the next stage. The images were first scaled down to an arbitrary size of 32×32. Then the images were transformed to grayscale to reduce the channel of the idea from a 3-channel (RGB/YCM) image to a 1-channel image to prevent creating excessively large images. This process also removed noise and other undesired distortion. The normalised images were then passed on to the machine learning models. The CNN and TL approaches were trained on the human faces dataset. The images were split into a train set of 70% and a test set of 30%. Hence the training set consists of 168 pictures, and the testing set consists (of 72) images. The images were trained using a 15-piece batch size and 20-iteration training epochs. Following model implementation, the accuracy and loss for each period were calculated.

3.2 Classification using CNN

The resultant face images are in the form of matrices passed onto the CNN model as input into the input layer. Most of the framework's computational effort is carried out by the convolution layer, which is the essential element of CNN. The restricted province of the approachable field and the kernel, a collection of learnable parameters, merge two matrices in this layer to create a dot product. The kernel is smaller in space but deeper when using a face image. The output volume was computed utilising the formula below for an input of size W x W x D, D is the number of kernels with a spatial dimension of F, stride S, and amount of padding P Ogundokun et al. (2022).

$$W^{out} = \left(\frac{W - F + 2P}{S}\right) + 1 \tag{1}$$

The second layer, known as the pooling layer, substitutes a summary statistic produced from nearby outputs for the network's creation in certain areas. This

helps reduce the spatial size of the representation, which minimises the amount of computation and weights required. Each slice of the model is subjected to the pooling operation separately. The output volume was determined using the formula below, with a pooling kernel of spatial size F, a pooling map of size W x W x D, and stride S.

$$W^{out} = \left(\frac{W - F}{S}\right) + 1 \tag{2}$$

Since convolution is a linear operation and Face images are nonlinear, the third layer, known as the non-linearity layer, is typically added straight after the convolutional layer to add non-linearity to the activation map. ReLU is more reliable than all the activation functions, as it accelerates convergence up to six times. As in a standard FCNN, all of the neurons in the fourth layer are fully connected to all the neurons in the layers above and below. This layer is known as the fully connected layer. Therefore, it can be determined by matrix multiplication, adding a bias effect. As seen in Figure 1, the completely connected layer makes it easier to map the representation between the input and output.

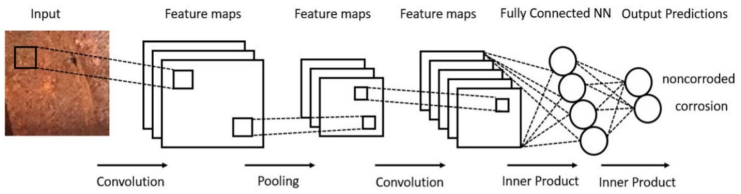


Fig. 1. The CNN architecture (Atha and Jahanshahi, 2018)

3.3 Classification using VGG 16 TL

The facial images in this study were also exposed to the VGG16 algorithm, which uses a max pool layer of 2x2 filters with a stride of 2 and convolution layers of 3x3 filters with a stride of 1. Max pool and convolution layers are evenly spaced throughout the whole architecture. Two fully connected layers (FC) and a softmax for output are included as a conclusion. The 16 in VGG16 refers to the network's large size and the fact that it has 16 layers with weights. Figure **??** displays the architecture of the VGG16.

3.4 Testing and Validation

To measure the effectiveness of the models, after the training, the models were tested and validated by obtaining the accuracy of classification in terms of correctly and incorrectly classified images. Additional metrics of evaluation of the efficiency of the models are precision, recall, and F-measure values for CNN and TL.

3.5 Results and discussion

The validation accuracy and loss of the CNN architecture implemented in this research work, after training the model with the acquired dataset, with an epoch of 20 and validation step of 20, based on the accuracy of the last epoch, is presented in Figure 3.

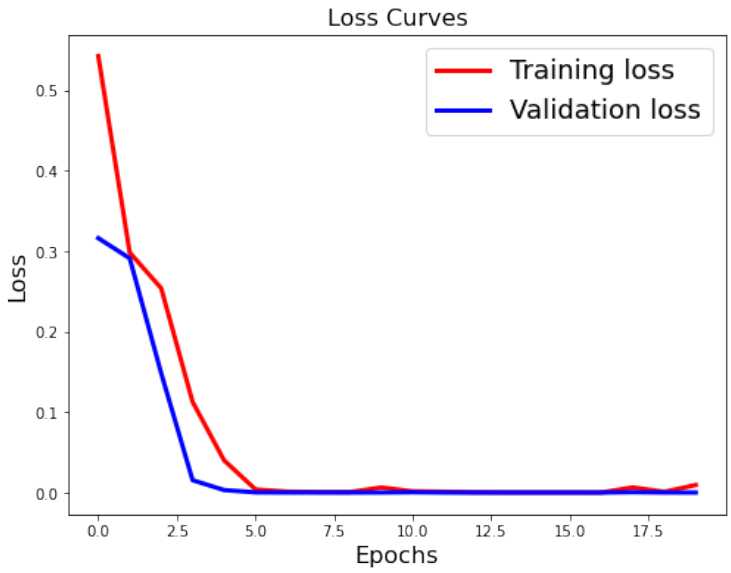


Fig. 2. The loss in training and validation

At the start of the training, there was a high training loss of over 0.5 and a validation loss of slightly over 0.3, as evident. As the epochs are further increased, the network can learn more, and the loss of the training and validation experiences declines as it gets closer to 0.

The accuracy rises to a good range near 0.75 at the beginning of training, and the validation accuracy rises to well above 0.80. As the number of epochs increases, the network experiences an increase and gets close to the range of 0.97 and above. Similarly, the TL VGG16 architecture shows the loss and accuracy curves in Figure **??** and Figure **??**.

The training starts with a training loss of up to 0.175, but there was no consistent decrease as the epoch increased further. It ended up at a value slightly above 0.050. The validation loss started well and ended at the same weight as the training loss. This means that the TL model's performance will keep increasing with an increase in the number of epochs.

The validation accuracy was steady until the last epoch when the model's performance dropped as a result of the early stopping callback included to avoid

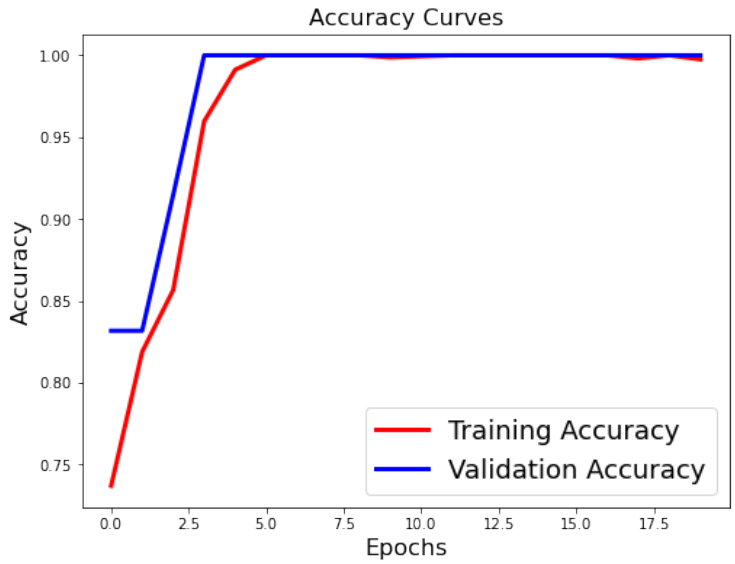


Fig. 3. The accuracy in training and validation

overfitting. The result of implementing using the CNN, as well as TL models for the face gender recognition, is further presented in the below tables. Table 1 reveals the result of the time taken to recognise the gender of the queried faces and the accuracy of recognition.

Table 1. Performance result of recognition.

Model	Recognition time (sec)	No. of tested samples	Correct Classification	Incorrect Classification	Accuracy (%)
CNN	0.98	72	62	10	86.11
TL	0.23	72	69	3	95.86

The result in Table 1 reveals that the TL model outperforms the CNN model in terms of faster recognition time, higher accuracy and a limited number of incorrect classifications. This result could be because only the early and central layers are employed in transfer learning, while the latter layers are only retrained. The overall effect is a shorter training time, and as can also be observed, TL takes 0.23 secs to recognise the test image, unlike CNN, which takes 0.98 secs. This indicates that the TL that involves pre-trained models to carry out the new task is faster than the existing CNN.

Furthermore, the TL has a higher accuracy in recognising the queried image. A possible reason is that CNN loses information in the pooling layers due to changes in pose and orientation, which makes it difficult to handle ambiguity in terms of a slight shift in face orientation.

The contingency table for the CNN and TL models are presented in Table 2 and Table 3, respectively. The letters TP, FP, FN, and TN denote the true positives, false positives, and false negatives, respectively.

Table 2. Contingency table for CNN.

Model	Accuracy (%)
TP (36)	FN (6)
FP(4)	TN (26)

Table 3. Contingency table for TL.

Model	Accuracy (%)
TP (48)	FN (1)
FP(2)	TN (21)

These values were used to obtain the precision, recall and f-measure based on the widely adopted relations indicated in Equations 3, 4 and 5 Ogundokun et al. (2022).

$$Precision = \left(\frac{TP}{TP + FP}\right) \tag{3}$$

$$Recall = \left(\frac{TP}{TP + FN}\right) \tag{4}$$

$$F - measure = 2 * \left(\frac{ecall * precision}{recall + precision}\right) \tag{5}$$

The measure of the efficiency of the models based on the experiment carried out in this research is presented in Table 4 in terms of recall, precision and f-measure, which are computed using the values from the contingency tables above for CNN and TL.

Table 4. Result of precision, recall and f-measure.

Classifier	Recall	precision	F-measure
CNN	0.86	0.90	0.88
TL	0.98	0.96	0.97

It can be observed from Table 4 that the TL outperforms the CNN in terms of the F-measure, which is a result of the very low false negative (FN) outcome.

This implies that the model will not deny authentic users (as obtained from the recall). Additionally, the low false positive (FP) outcome indicates that the model will not wrongly accept unauthorised users.

It is worthy of note that much of the reviewed research related have not satisfactorily reported information on the evaluation metrics such as recall, precision, f-measure and recognition time, which are important determinants of the efficiency of a machine learning model. Therefore, this work recommends that recognition accuracy alone is insufficient to measure a model's performance.

4 Conclusion

In this study, a performance comparison between CNN and TL on classifying human face gender was carried out. The experimental results indicate that both algorithms performed optimally. However, TL outperforms CNN regarding recognition accuracy, recognition time, precision, recall and f-measure. The TL approach proves effective and efficient based on the performance metrics employed and the dataset used. Furthermore, the dataset used for training is relatively small. This work intends to extend the experiment to a larger and augmented dataset. This technique is important to address the problem of distributed coordination and decision making in hostile environments with heterogeneous physical IoT (Internet of Things) devices operating with images in Edge-Computing, to achieve their goals more efficiently. This enables solutions to flexible, adaptive and seamless decision-making problems at the edge of the cloud, where complex predictions are processed in the cloud in an optimised way and operational decisions are made locally at the edge. in the cloud and operational decisions are made locally at the edge.

Bibliography

M. Afifi and A. Abdelhamed. Afif4: Deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces. *Journal of Visual Communication and Image Representation*, 62:77–86, 2019.

D. J. Atha and M. R. Jahanshahi. Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. *Structural Health Monitoring*, 17(5):1110–1128, 2018.

R. Babatunde, S. Abdulsalam, S. Yusuff, and A. Babatunde. Gender recognition using local binary pattern and naive bayes classifier. *Journal of Computer Science and Its Application*, 23(1):65–74, 2016.

A. González-Briones, G. Villarrubia, J. F. De Paz, and J. M. Corchado. A multi-agent system for the classification of gender and age from images. *Computer Vision and Image Understanding*, 172:98–106, 2018.

A. Greco, A. Saggese, M. Vento, and V. Vigilante. Gender recognition in the wild: a robustness evaluation over corrupted images. *Journal of Ambient Intelligence and Humanized Computing*, 12:10461–10472, 2021.

M. Jewel, M. I. Hossain, and T. H. Tonni. Bengali ethnicity recognition and gender classification using cnn & transfer learning. In *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 390–396. IEEE, 2019.

S. Jia, T. Lansdall-Welfare, and N. Cristianini. Gender classification by deep learning on millions of weakly labelled images. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 462–467. IEEE, 2016.

C. Lawal, A. Akinrinmade, and J. Badejo. Face-based gender recognition analysis for nigerians using cnn. In *Journal of Physics: Conference Series*, volume 1378, page 032014. IOP Publishing, 2019.

H. Moeini and S. Mozaffari. Gender dictionary learning for gender classification. *Journal of Visual Communication and Image Representation*, 42:1–13, 2017.

R. O. Ogundokun, R. Maskeliūnas, and R. Damaševičius. Human posture detection using image augmentation and hyperparameter-optimized transfer learning algorithms. *Applied Sciences*, 12(19):10156, 2022.

D. J. Sharma, S. Dutta, and D. J. Bora. Rega: Real-time emotion, gender, age detection using cnn-a review. *Annals of Computer Science and Information Systems*, 24, 2020.

T. A. Sumi, M. S. Hossain, R. U. Islam, and K. Andersson. Human gender detection from facial images using convolution neural network. In *Applied Intelligence and Informatics: First International Conference, AII 2021, Nottingham, UK, July 30–31, 2021, Proceedings 1*, pages 188–203. Springer, 2021.

K. YILDIZ, E. GÜNEŞ, and B. Anil. Cnn-based gender prediction in uncontrolled environments. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 9(2): 890–898, 2021.

F. Zaman. Gender classification using custom convolutional neural networks architecture. *International Journal of Electrical and Computer Engineering (IJECE)*, 10(6):5758–5771, 2020.