






Düzce University Journal of Science & Technology

Research Article

CNN-based Gender Prediction in Uncontrolled Environments

 Kazım YILDIZ ^{a,*},  Engin GÜNEŞ ^b,  Anil BAS ^a

^a Department of Computer Engineering, Faculty of Technology, Marmara University, İstanbul, TURKEY

^b Department of Computer Engineering, Institute of Pure and Applied Sciences, Marmara University, İstanbul, TURKEY

* Corresponding author's e-mail address: kazim.yildiz@marmara.edu.tr

DOI: 10.29130/dubited.763427

ABSTRACT

With the increasing amount of data produced and collected, the use of artificial intelligence technologies has become inevitable. By using deep learning techniques from these technologies, high performance can be achieved in tasks such as classification and face analysis in the fields of image processing and computer vision. In this study, Convolutional Neural Networks (CNN), one of the deep learning algorithms, was used. The model created with this algorithm was trained with facial images and gender prediction was made. As a result of the experiments, 93.71% success rate was achieved on the VGGFace2 data set and 85.52% success rate on the Adience data set. The aim of the study is to classify low-resolution images with high accuracy.

Keywords: CNN, VGGFace2, Adience, Image classification, Gender classification

Kontrolsüz Ortamlarda CNN Tabanlı Cinsiyet Tahmini

ÖZET

Üretilen ve toplanan veri miktarının giderek artması ile birlikte yapay zekâ teknolojilerin kullanılması kaçınılmaz hale gelmiştir. Bu teknolojilerden biri olan derin öğrenme teknikleri kullanılarak bilgisayarlı görü ve görüntü işleme alanlarında yüz analizi ve sınıflandırma gibi görevlerde yüksek performans alınabilmektedir. Bu çalışmada derin öğrenme algoritmalarından Evrişimsel Sinir Ağları (CNN) kullanılmıştır. Bu algoritma ile oluşturulan model, yüz görüntüleriyle eğitilmiş ve cinsiyet tahmini yapılmıştır. Yapılan deneyler sonucunda VGGFace2 veri seti üzerinde 93.71% ve Adience veri seti üzerinde 85.52% oranında başarı sağlanmıştır. Çalışmanın amacı düşük çözünürlükteki resimleri yüksek doğrulukla sınıflandırabilmektir.

Anahtar Kelimeler: CNN, VGGFace2, Adience, Görüntü sınıflama, Cinsiyet sınıflama

I. INTRODUCTION

Gender classification is one of the areas used in many applications. This is a significant task especially in the fields on human-computer interaction, control system, visual surveillance, and yet particularly difficult when only one single image available.

In this study, we propose a deep learning approach to get better results in gender classification. Deep learning is generally used in object recognition [1], speech recognition [2] and image classification [3]. These models have been successful in high-resolution images [4]. In this study, we focus on uncontrolled images. The challenge working with uncontrolled images is to predict the gender accurately while dealing with lighting, angle, facial expression and occlusion [5]. Particularly, we used datasets that have large variations in ethnicity, pose, age and illumination. In-the-wild images from such datasets make the problem somewhat difficult, since it requires modeling intrinsic and extrinsic parameters. However, images taken in real world setting, by surveillance and security systems for example, are even more challenging. We observe that these images are in considerably poor quality and often blurry or pixelated. To our knowledge, there is no low-resolution dataset explicitly for gender classification available in the literature. For this reason, we generate low-resolution images from available datasets by downsampling to establish approximate real-world imagery setting.

The aim of this study to establish a simple convolutional neural network structure for gender classification using low-resolution images. We have following contributions in this study: (1) A powerful CNN architecture is proposed. This is a modified version of VGGNet, the originally proposed in [3], widely used in image classification. (2) We observe that the proposed CNN architecture would maintain its success in low-quality mages as well. The proposed approach yields results that are superior to previous studies, even using reduced image quality. (3) We investigate how image quality affects the gender detection accuracy. We trained the network with various image resolutions and demonstrate the gender detection performance drops as resolution decrease.

Face recognition technologies can offer different application areas. Our study directly contributes to related topics such as face analysis and similarity analysis by examining images taken at different times and places. The proposed study has the potential to be implemented in healthcare, entertainment, access control and security systems.

The rest of this paper is organised as follows: In Section 2, we review the relevant studies in the literature. Section 3 describes our network model and introduces the datasets used in training and evaluation in detail. In Section 4 experimental results in two different datasets are listed. Finally, we draw a conclusion in Section 5.

II. LITERATURE REVIEW

We start by examining some of the examples in the literature. There are many ways to determine gender. For example, mouse movements can be analysed for gender classification [6]. Another approach is to use the properties of iris images [7]. Profile face image can be used together with the ear images [8]. It is possible to estimate gender by analysing speech signals [9]. Another noteworthy example is to use neural signals from EEG sensors [10]. Gender classification can be established successfully from word choices [11], human gait [12], the smartphone usage habits [13] as well as using hand pictures [14] and three-dimensional face images [15]. While skin colour can affect performance in gender classification [16], facial texture can be used as identifiers [17].

CNN have respectably improved the performance in visual classification [18-22]. Milki et al. suggested a deep learning-based method for face detection in an uncontrolled environment. They combined at multiple scales both local and global features. The experimental results show that the suggested method works well on multi-scale face detection problems [23]. Masud et al. suggested a

tree-based model which is using deep model for automatic face recognition in a cloud environment which is computationally less expensive. The model is evaluated on various public datasets. The proposed model accuracy is more than 95% [24]. Lozoya et al. suggested a method based on features which are extracted CNN. The facial expression recognition model learns from different databases with mixed instances to increase generalization. The accuracy of the model is above 92% [25]. Chaudri proposed a comparative analysis on face recognition with deep learning models [26].

III. MATERIALS AND METHODS

We implement CNN architecture using Keras library in TensorFlow platform in this study. Here, we will describe our network and used datasets in detail. In the next subsection, we start by explaining how convolutional layers structured, then, present our own proposed model. The following subsection reports the image preprocessing stage and the datasets we used. The block diagram of the study is shown in Figure 1.

A. PROPOSED MODEL

For tasks such as image segmentation and classification, traditional networks were in use before the CNN. While these networks are often limited in their capacity, CNN have emerged as data grow. The most important feature of the convolutional neural networks is that it has a layered structure. In this architecture, the algorithms are modelled similar to the structure of human neurons. Convolutional neural networks in many cases give better results than traditional networks [27-28]. However, we see that the performance is significantly reduced when these networks work with low-resolution images [29]. We briefly explain how CNN layer structure works. The convolution layer calculates the output of neurons. It provides a feature map from each input image. To give an example, there is an image matrix and this matrix is multiplied by a 3x3 matrix then this 3x3 matrix is shifted, and the feature map is created in this way. The pooling layer also emerges as a sub-sampling layer. The purpose of the pooling layer is to reduce the size and is to get an input notation. Max-pooling [30] is most commonly used in pooling. For example, when using a 2x2 maximum pooling in an image matrix, taking the highest value in the field on which it comes, and it writes to the new matrix. Then the matrix is shifted, and the process continues with the highest value there. A new matrix is created by applying this process to each element of the image matrix. The fully connected layer is the layer that combines and flattens all high-level features. Classification is performed using SoftMax [31] or sigmoid [32] classifier using the given data in this layer.

The structure of our model is shown in Figure 1. We build a smaller and compact version of the VGGNet architecture [3], which is widely used for image classification and face recognition [33]. Our proposed network consists of five convolution layers, three pooling layers and one fully connected layer. The most prominent features of the images are extracted with the help of the convolution layers in CNN. These extracted features are sent to the output layer and classification is performed on these features.

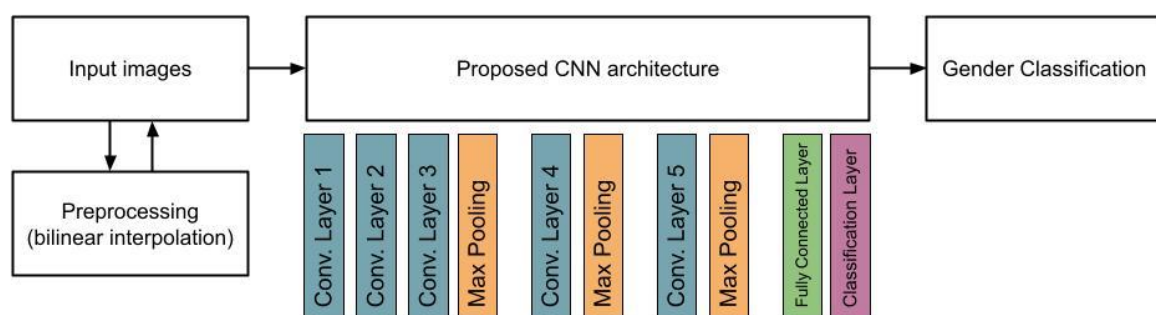


Figure 1. Overview of the study with proposed CNN architecture.

B. DATASET

In this study, we first used VGGFace2, a dataset consisting of more than 9000 different people and 3.3 million images [34]. On average, there are 362 pictures for each person. The dataset consists of different ethnicities, different age groups and different professions and images have been downloaded from google image search. The dataset is organized as 3.1M train data and 200K test data.

The images in the dataset have various resolutions and differ from each other. Since our network takes fixed-size images as input, we use bilinear interpolation to downsample face images to $32 \times 32 \times 3$. This also allows us to experiment with low-resolution images. There is no cropping process, however, resizing some images caused aspect ratio change.

We also used Adience dataset [35] as a benchmark to evaluate our results. Similar preprocessing steps were applied to images in this dataset as well. Figure 2 shows the original and downsampled images from VGGFace2 and Adience datasets.

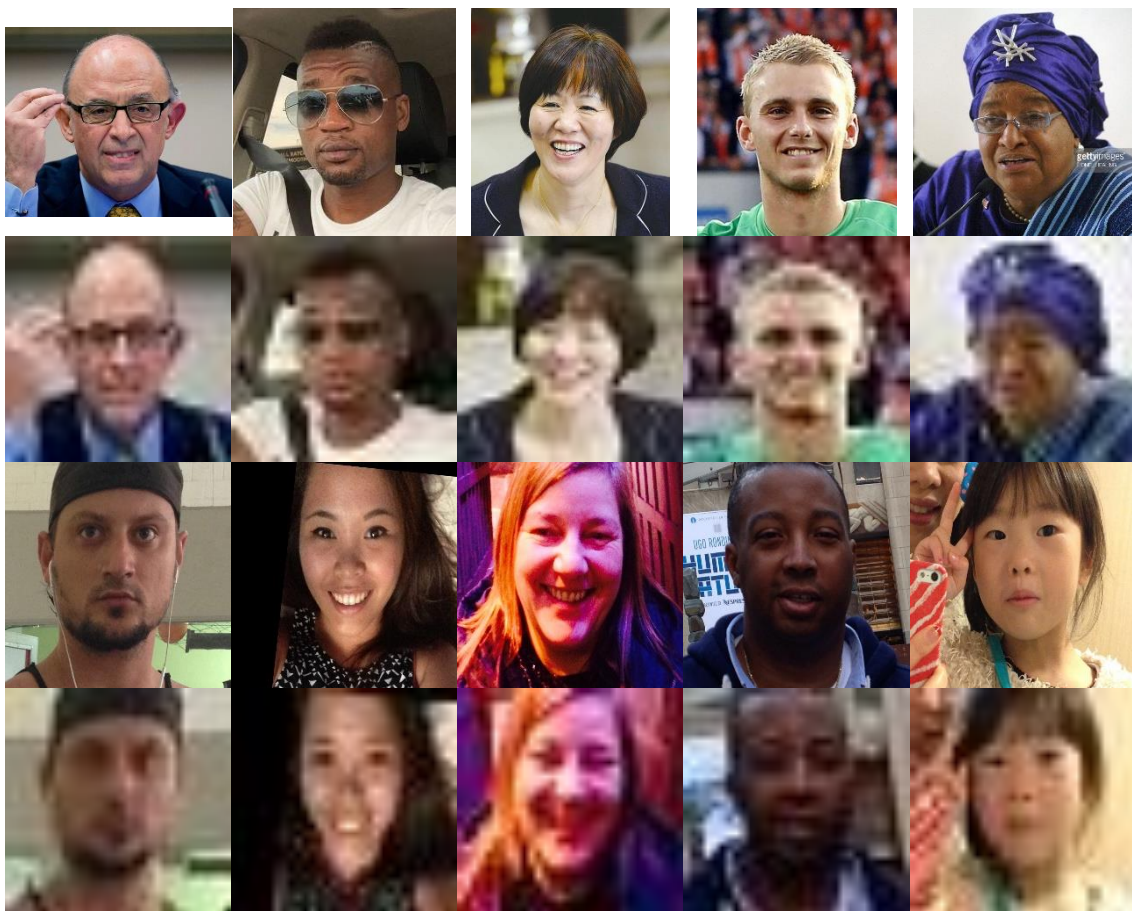


Figure 2. Original and downsampled image examples from VGGFace2 (top two rows) and Adience (bottom two rows) datasets.

IV. RESULTS

In this part, we propose two sets of experiments on two datasets. First, we trained our network on downsampled images from VGGFace2 dataset. We used learning rate = 0.000001, batch size = 64 and trained the network for 300 epochs. Our network achieved 93.71% success rate (96.29% test accuracy,

93.71% validation accuracy, 10.80% train loss and 17.18% validation loss). Model's accuracy and loss curves over epochs are plotted in Figure 3.

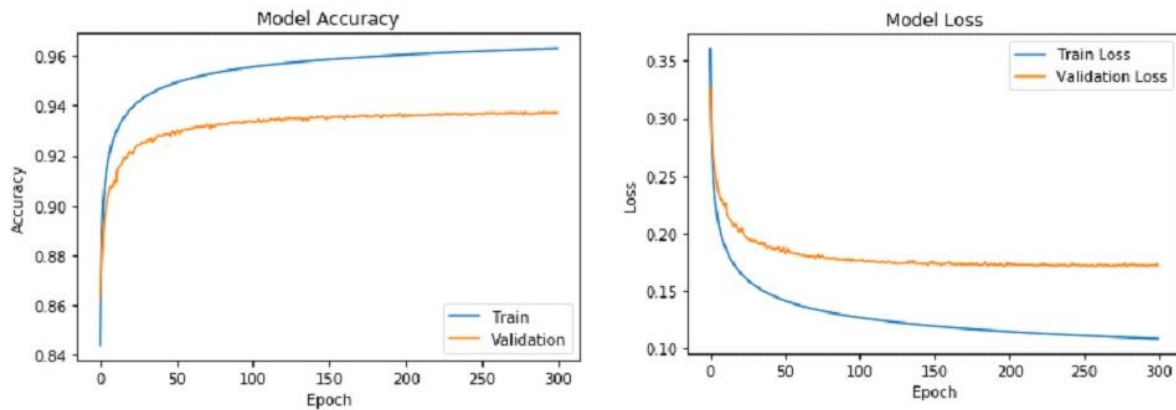


Figure 3. Model Accuracy and Loss on VGGFace2 dataset.

Second, we trained our network separately on the subset (%10 of the dataset) of downsampled images from VGGFace2 dataset. One would assume that some features that are available in high-resolution would be lost when the image resolution is reduced. In other words, recognition and classification task would be much harder in a low-resolution setting. To prove that this is a correct assumption, we follow the approach of Mynepalli et al. [36]. We trained our proposed model using images with various resolutions (particularly, 32x32x3, 64x64x3 and 128x128x3). In Figure 4, ROC plots and AUC scores illustrate that the detection performance drops as resolution decreases.

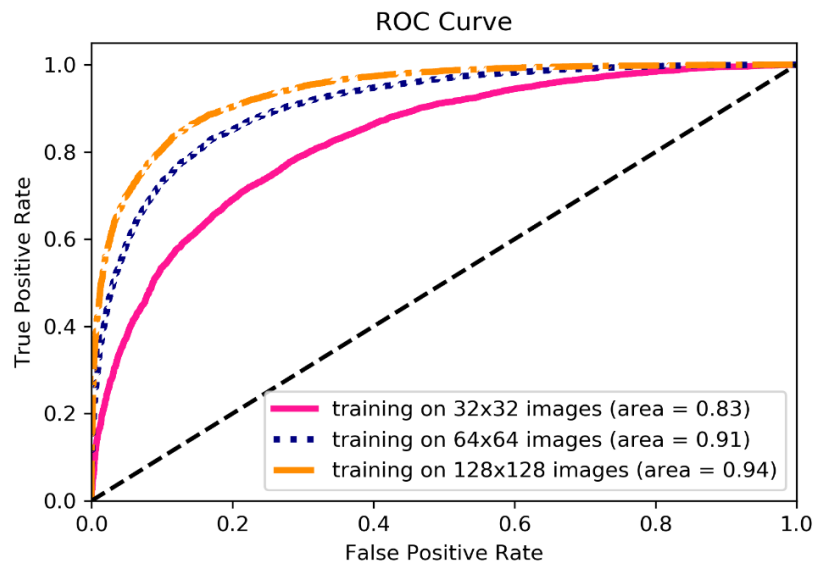


Figure 4. Training results on various image resolutions. The gender detection performance drops as resolution decreases.

For quantitative comparison, we test our network's performance on Adience dataset. The experimental results are listed in Table 1. For training phase, we used learning rate = 0.00001, decay rate = $1e-6$ and batch size = 24. For training set we used 80% of the dataset and rest for validation set. Our proposed network is trained for 400 epochs in a low-resolution setting. In Table 2, we show a quantitative comparison with previous studies. Please note that although our network trained on low-resolution images, it offers equal or better performance. The proposed approach surpasses prominent studies in the field and very close to the performance of [39].

Table 1. Training/Validation accuracy and loss results on Adience dataset.

Epoch	Accuracy	Loss	Validation Accuracy	Validation Loss
200	0.9172	0.2013	0.8521	0.3518
400	0.9649	0.0875	0.8552	0.4502

Table 2. Gender classification accuracy comparison on Adience Dataset. Please note that although our network trained on low-resolution images, it offers equal or better performance.

Method	Accuracy (%)
Eidinger et al. [35]	77.8
Liao et al. [37]	78.63
Hassner et al. [38]	79.3
Ours	85.52
Levi and Hassner [39]	86.80

As a final experiment, we provide example results in Figure 5. Prediction rates are shown under each image. This experiment clearly illustrates the difficulty of low-resolution data. The pixelation and the loss of appearance-based information is hugely visible. Nevertheless, our method exhibits very high success rate.

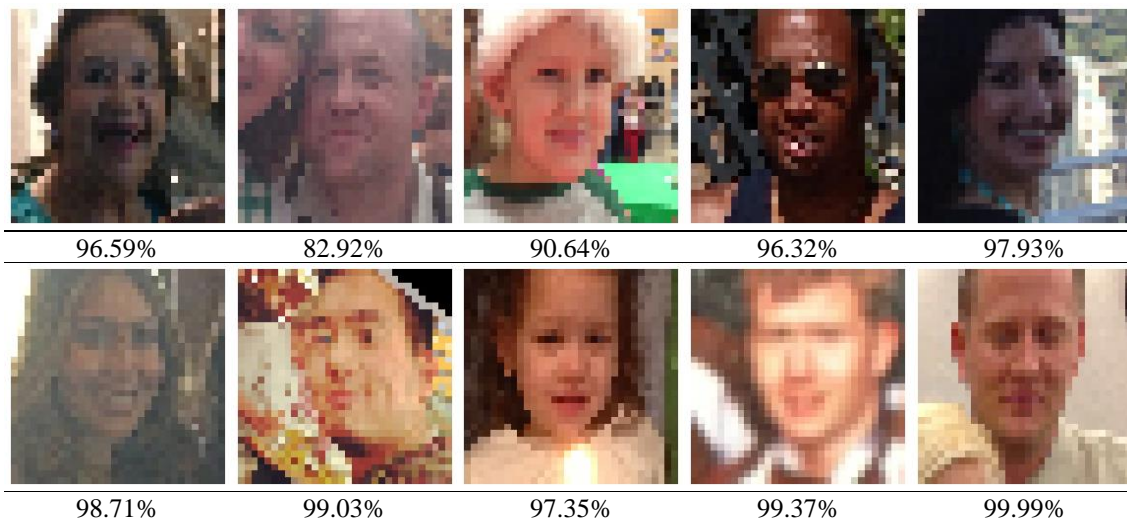


Figure 5. Example results on the Adience dataset. Prediction rates are shown under each image.

In conclusion, these experiments conducted on Adience and VGGFace2 datasets illustrate that the proposed method is capable in correctly predicting gender. Moreover, quantitative comparison in Table 2 shows the performance of our networks is superior even lack of observations and meaningful information from the original data (due to downsampling).

V. CONCLUSION

In this paper, we investigate a CNN-based gender prediction network in uncontrolled environments and tested on two networks. We figure out the problem of gender detection by training a relatively simple deep network on recent large-scale face datasets. To establish approximate real-world imagery setting, we generate low-resolution images by using bilinear interpolation.

We evaluate our network on two datasets. The results show that our network surpasses previously reported results for gender detection on Adience Dataset [35]. As part of the novelty, the experiments on VGGFace2 [34] is important as well since there is almost no gender detection study on this dataset available in the literature. We provide qualitative and quantitative classification results for both datasets for further comparison.

To summarise, we present a robust gender prediction architecture that can be used in low-resolution data. We also provide an analysis on VGGFace2 that shows model performance on various image resolutions. Working with low-resolution uncontrolled images, the proposed network achieves competitive results in gender classification on two datasets.

The study could be used in many application areas. As an example, the proposed technique could be implemented in commercial behaviour analysis where gender prediction could be useful. In the commercial domain, applications could make suggestions considering gender differences. Other examples are, especially mainly working with low-resolution images, security and surveillance systems. This includes CCTV camera systems as well as traffic cameras. Statistical information (e.g. gender-based violation of speed rules and the usage of public transportation) could be extracted from these low-resolution images.

In future studies, instead of the low-resolution pictures we use in this study, we plan to train and develop our model using even lower resolution pictures, such as [36]. In addition, we aim to increase the accuracy rates and further reduce the loss rates by optimizing our model.

VI. REFERENCES

- [1] J. M. Gandarias, A. J. García-Cerezo and J. M. Gómez-de-Gabriel, “CNN-based methods for object recognition with high-resolution tactile sensors,” *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6872–6882, 2019.
- [2] W. Chan, N. Jaitly, Q. Le, O. Vinyals and N. Shazeer, “Speech recognition with attention-based recurrent neural networks,” U.S. Patent Appl. 20200118554A1, Apr. 16, 2020.
- [3] K. A. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. International Conference on Learning Representations (ICLR)*, 2015, pp. 1–14.
- [4] B. Shrestha, Y. Kwon, D. Chung and W. Gal, “The atrous cnn method with short computation time for super-resolution,” *International Journal of Computing and Digital Systems*, vol. 9, no. 2, pp. 221–227, 2020.
- [5] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li and G. Hua, “Labeled faces in the wild: A survey,” in *Advances in Face Detection and Facial Image Analysis*, 1st ed., Cham, Switzerland: Springer, 2016, pp. 189–248.
- [6] N. Van Balen, C. Ball and H. Wang, “Analysis of targeted mouse movements for gender classification,” *EAI Endorsed Transactions on Security and Safety*, vol. 4, no. 11, 2017.
- [7] J. E. Tapia and C. A. Perez, “Gender classification from nrr images by using quadrature encoding filters of the most relevant features,” *IEEE Access*, vol. 7, pp. 29114–29127, 2019.
- [8] D. Yaman, F. I. Eyiokur and H. K. Ekenel, “Multimodal age and gender classification using ear and profile face images,” in *Proc. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 2414–2421.

- [9] N. A. Nazifa, C. Y. Fook, L. C. Chin, V. Vijejan and E. S. Kheng, “Gender prediction by speech analysis,” *Journal of Physics: Conference Series*, vol. 1372, no. 012011, 2019.
- [10] B. Kaur, D. Singh and P. P. Roy, “Age and gender classification using brain–computer interface,” *Neural Computing and Applications*, vol. 31, no. 10, pp. 5887–5900, 2019.
- [11] C. Bhagvati, “Word representations for gender classification using deep learning,” *Procedia Computer Science*, vol. 132, pp. 614–622, 2018.
- [12] Z. Q. Mawlood and A. T. Sabir, “Human gait-based gender classification using neutral and non-neutral gait sequences,” *Revista Innovaciencia*, vol. 7, no. 1, pp. 1–13, 2019.
- [13] J. A. Polin and O. Khan, “Gender identification from smart phone usage using machine learning algorithm,” B.Sc. Report, Department Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh, 2019.
- [14] M. Afifi, “11K Hands: gender recognition and biometric identification using a large dataset of hand images,” *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20835–20854, 2019.
- [15] S. Bentaieb, A. Ouamri and M. Keche, “SAX2SEX: Gender classification on 3d faces using symbolic aggregate approximation,” in *Proc. 6th International Conference on Image and Signal Processing and their Applications (ISPA)*, 2019, pp. 1–5.
- [16] V. Muthukumar, T. Pedapati, N. Ratha, P. Sattigeri, C. Wu, B. Kingsbury, A. Kumar, S. Thomas, A. Mojsilovic and K. Varshney, “Understanding unequal gender classification accuracy from face images,” 2018, *arXiv:1812.00099*.
- [17] F. Bougourzi, S. Bekhouche, M. Zighem, A. Benlamoudi, T. Ouafi and A. Taleb-Ahmed, “A comparative study on textures descriptors in facial gender classification,” presented at 10^{ème} Conférence sur le Génie Electrique, Bordj El Bahri, Algeria, 2017.
- [18] O. Arriaga, M. Valdenegro-Toro and P. Plöger, “Real-time convolutional neural networks for emotion and gender classification,” in *Proc. European Symposium on Artificial Neural Networks (ESANN)*, 2019, pp. 221–226.
- [19] G. Levi and T. Hassner, “Age and gender classification using convolutional neural networks,” in *Proc. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 34–42.
- [20] J. Zhang, Y. Xie, Q. Wu and Y. Xia, “Medical image classification using synergic deep learning,” *Medical Image Analysis*, vol. 54, pp. 10–19, 2019.
- [21] F. V. Massoli, G. Amato, F. Falchi, C. Gennaro and C. Vairo, “Improving multi-scale face recognition using VGGFace2,” in *Proc. International Conference on Image Analysis and Processing (ICIAP)*, 2019, pp. 21–29.
- [22] F. Juefei-Xu, E. Verma, P. Goel, A. Cherodian and M. Savvides, “Deepgender: occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention,” in *Proc. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 68–77.
- [23] H. Mliki, S. Dammak and E. Fendri, “An improved multi-scale face detection using convolutional neural network,” *Signal Image and Video Processing*, vol. 14, no. 7, pp. 1345–1353, 2020.

- [24] M. Masud, G. Muhammad, H. Alhumyani, S. S. Alshamrani, O. Cheikhrouhou, S. Ibrahim and M. S. Hossain, “Deep learning-based intelligent face recognition in IoT-cloud environment,” *Computer Communications*, vol. 152, pp. 215–222, 2020.
- [25] S. M. González-Lozoya, J. de la Calleja, L. Pellegrin, H. J. Escalante, M. A. Medina and A. Benitez-Ruiz, “Recognition of facial expressions based on CNN features,” *Multimedia Tools and Applications*, vol. 79, pp. 13987–14007, 2020.
- [26] A. Chaudhuri, “Deep learning models for face recognition: A comparative analysis,” in *Deep Biometrics*, 1st ed, Cham, Switzerland: Springer, 2020, pp. 99–140.
- [27] Y. Luo, Y. Shao, H. Chu, B. Wu, M. Huang and Y. Rao, “CNN-based blade tip vortex region detection in flow field,” in *Proc. International Conference on Graphics and Image Processing (ICGIP)*, 2019, vol. 11373.
- [28] Y. Wang, M. Liu, P. Zheng, H. Yang and J. Zou, “A smart surface inspection system using faster R-CNN in cloud-edge computing environment,” *Advanced Engineering Informatics*, vol. 43, no. 101037, 2020.
- [29] O. A. Aghdam, B. Bozorgtabar, H. K. Ekenel, J. Thiran, “Exploring factors for improving low resolution face recognition,” in *Proc. CVPR Workshops*, 2019, pp. 2363–2370.
- [30] G. Toliás, R. Sivic and H. Jégou, “Particular object retrieval with integral max-pooling of CNN activations,” in *Proc. International Conference on Learning Representations (ICLR)*, 2016, pp. 1–12.
- [31] X. Liang, X. Wang, Z. Lei, S. Liao and S. Li, “Soft-margin softmax for deep classification,” in *Proc. NIPS*, 2017, pp. 413–421.
- [32] A. Meliboev, J. Alikhanov and W. Kim, “1D CNN based network intrusion detection with normalization on imbalanced data,” in *Proc. International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, 2020, pp. 218–224.
- [33] O. M. Parkhi, A. Vedaldi and A. Zisserman, “Deep face recognition,” in *Proc. The British Machine Vision Conference (BMVC)*, 2015, pp. 1–12.
- [34] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, “Vggface2: A dataset for recognising faces across pose and age,” in *Proc. International Conference on Automatic Face & Gesture Recognition (FG)*, 2018, pp. 67–74.
- [35] E. Eiding, R. Enbar and T. Hassner, “Age and gender estimation of unfiltered faces,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.
- [36] S. C. Mynepalli, P. Hu and D. Ramanan, “Recognizing tiny faces,” in *Proc. International Conference on Computer Vision Workshops (ICCVW)*, 2019, pp. 1121–1130.
- [37] Z. Liao, S. Petridis and M. Pantic, “Local deep neural networks for age and gender classification,” 2017, *arXiv:1703.08497*.
- [38] T. Hassner, S. Harel, E. Paz and R. Enbar, “Effective face frontalization in unconstrained images,” in *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4295–4304.
- [39] G. Levi and T. Hassner, “Emotion recognition in the wild via convolutional neural networks and mapped binary patterns,” in *Proc. ACM on International Conference on Multimodal Interaction*, 2015, pp. 503–510.