Automatic Gender Classification from Handwritten Images: a Case Study

Irina Rabaev[0000−0002−8542−8342]⋆, Marina Litvak[0000−0003−3044−3681]⋆, Sean Asulin, and Or Haim Tabibi

Shamoon College of Engineering, 56 Bialik St. Be'er Sheva 8410802, Israel
{irinar,marinal,shonas,orita4}@ac.sce.ac.il

Abstract. Using a handwritten sample to automatically classify the writer's gender is an essential task in a wide range of areas, e.g., psychology, historical documents classification, and forensic analysis. The challenge of gender prediction from offline handwriting is demonstrated by the relatively low (below 90%) performance of state-of-the-art systems. Despite a high interest within a broad spectrum of research communities, the published works in this area generally concentrate on English and Arabic languages. Most of the existing approaches focus on manual feature selection. In this work, we study an application of deep neural networks for gender classification, where we investigate cross-domain transfer learning with ImageNet pre-training. The study was performed on two datasets, the QUWI dataset, consisting of handwritten documents in English and Arabic, and a new dataset of documents in Hebrew script. We perform extensive experiments, analyze and compare the results obtained with different neural networks. We also compare the obtained results against human-level performance.

Keywords: Gender Classification · Offline Handwriting Analysis · Transfer learning · Deep Neural Network

1 Introduction

Handwriting gender classification is of great interest due to the broad range of areas it can be applied in, e.g., psychology, historical documents classification, and forensics [10, 11, 27]. Psychological studies of handwriting analysis have confirmed that gender classification can be made according to several significant differences in handwriting [10, 11, 27]. In general, while a female's handwriting tends to be more uniform, ordered, and has greater circularity, a male's handwriting tends to be more pointed, messy, and slanted.

With technological advances in image analysis and computer vision, manual handwriting analysis has become enhanced by automatic systems. The existing automatic methods can be classified into two categories: (1) traditional machine learning techniques that require prior features selection and (2) methods that

⋆ These authors contributed equally to this work

apply deep neural networks (DNNs), where features are learned within a network framework.

Many features were used for gender classification, including graphology features [26, 16, 9], textural features [1, 9], geometric features [4], Cloud of Line Distribution (COLD) and Hinge features [8], and wavelet-based features [2]. Traditional machine learning algorithms include k-Nearest Neighbours (kNN) [9], support vector machine (SVM) [5, 8], decision trees and random forests [26], Gaussian Mixture Models (GMM) [16], and combinations (ensembles) of several classifiers [17, 4, 1, 9, 2]. ICDAR competitions focusing on Gender Prediction from Handwriting [12, 7] gathered researchers from around the world to compare between different techniques. All systems that participated in the competitions followed the standard approach of traditional machine learning, where feature extraction must be performed prior to the classification.

The serious disadvantage of traditional machine learning approaches is that they are strongly dependent on manual feature engineering, which in the case of handwriting analysis often requires expert knowledge. Therefore, in our study we decided to lean on the most advanced deep learning approaches, those designed for image processing and published recently (or during the last decade). The authors of [15] applied a CNN-based model with a relatively small number of layers. In [21], gender classification was performed using advanced CNNs, such as DenseNet201, InceptionV3, and Xception. In this work, we report a similar study but extended it to an additional language (Hebrew) and the most recent networks. In [28], an attention-based two-pathway densely connected convolutional network was proposed to identify the gender of a handwritten document. Transfer learning was applied in [19] to detect the writer's gender from scanned handwritten documents. The authors used two pre-trained CNNs, GoogleNet (InceptionV1) and ResNet, as fixed feature extractors. For the classification stage, they applied SVM. In [18], pre-trained CNNs have been employed as feature extractors to discriminate between male and female handwriting, while classification is carried out using a number of classifiers, with Linear Discriminant Analysis being the most effective. Both works, [19] and [18], performed feature extraction and classification separately, by different models. In our study, the analyzed documents are submitted as input to a network, and the same network provides their classes.

This paper has two main contributions. First, we present the results of a case study for gender classification from handwritten document images with multiple DNNs. Because DNNs require a vast amount of training data, we investigate an application of cross-domain transfer learning with ImageNet pre-training. We perform extensive experiments and analyze the results on two different datasets: a newly introduced Hebrew Gender (HHD_gender) dataset of handwritten documents in Hebrew script, and the QUWI dataset [3], which consists of documents in English and Arabic languages. Second, we establish baseline results for the HHD_gender. The original collection of scanned images was introduced in [20]. Here, we preprocess and annotate the dataset for the gender classification task.

The preprocessed dataset can be downloaded from the Zenodo repository[1]. To the best of our knowledge, this is the first publicly available dataset in Hebrew script that can be used for gender classification. In addition, we compare the performances of the models against human-level performance on the HHD_gender dataset.

2 Case Study

The most successful recent image classification models are CNN-based. They show that shallow layers extract simple (low-level) features of an image, and deeper layers extract more complex (high-level) features. Thus, to make CNN more accurate, researchers mainly increase their depth by adding more layers. In this study we used the following models: VGGNet [22] (VVG16 and VGG19), ResNet [13], Inception [24] (Inception-v3 and Inception-ResNet-v2), Xception [6], DenseNet [14] (DenseNet121 and DenseNet169), NASNet [29], and EfficientNet [25]. All the networks are pre-trained on ImageNet. For comparison, we also report the results for a simple CNN trained from scratch.

2.1 Datasets

The experimental study is performed on the HHD_gender dataset and the QUWI dataset, together containing documents written in three different languages.

The HHD_gender dataset is a subset of the HHD dataset introduced in [20]. It contains 819 handwritten forms written by volunteers of different educational backgrounds and ages, both native and non-native Hebrew speakers. Each participant voluntarily provided demographic information, such as date, gender, and age, written at the top of the form, and copied a text paragraph printed above the text field. There are 50 variations of the forms; each form contains a text paragraph with an average of 62 words. The forms were scanned in color at a resolution of 600 dpi.

The forms were preprocessed as follows. First, the handwritten paragraph was extracted using coordinates of the corresponding text box. Then, the ground truth labeling (male and female) was performed automatically, based on the number of foreground pixels in the corresponding field. Finally, the labels were proof-read manually. Documents that did not contain gender information (the participant preferred not to fill in personal information) were withdrawn from the dataset. This process yielded 368 forms written by males and 461 by females. Finally, the images were converted to grayscale since the color caries no important information for the gender classification. Examples of the processed images are illustrated in the top row in Figure 1. For the experiments, the HHD_gender dataset was randomly subdivided into training (80%), validation (10%), and test (10%) sets.

The QUWI [3] dataset contains handwritten documents in Arabic and English languages. The documents were written by volunteers of different ages,

[1] https://zenodo.org/record/4729908#.YIvT7meJGhe

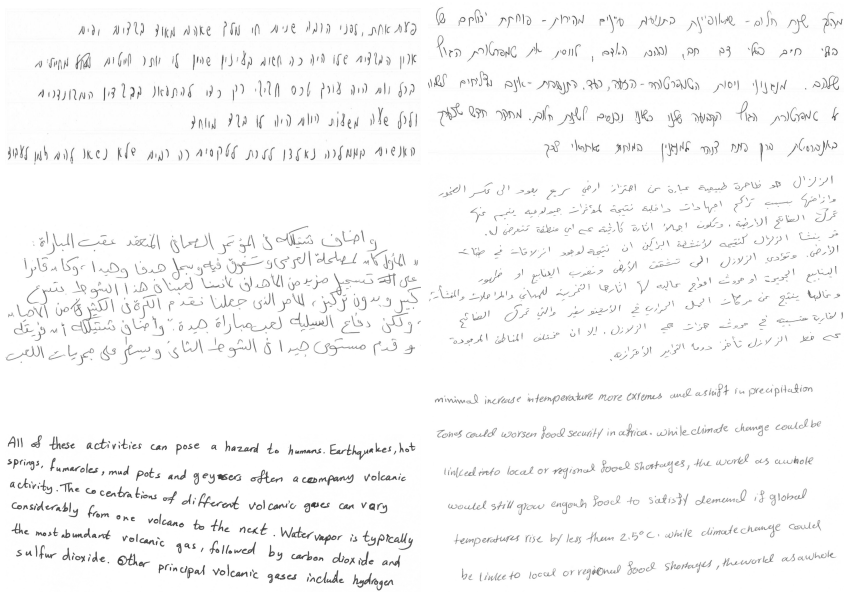

Fig. 1. Top row: Examples of the images from the HHD_gender dataset; male - left image, female - right image. Middle and bottom rows: Examples of the images from the QUWI dataset; male - left, female - right; different writers.

nationalities, and education levels. Each writer produced four handwritten documents: two in Arabic and two in English. One page in Arabic and one page in English contain the same text for all writers; the text on two other pages varies from writer to writer. Images are scanned with a 600 dpi resolution in JPG format. The middle and bottom rows in Figure 1 illustrate the samples from the QUWI dataset.

The ICDAR 2013 and 2015 competitions on Gender Prediction from Handwriting [12, 7] used a subset of the QUWI dataset. ICDAR 2013 dataset is composed of handwritten documents of 475 writers: 221 males and 254 females, and is divided into training (282 writers) and test (193 writers) sets. Three classification schemes were applied: training and testing on samples in Arabic, training and testing on samples in English, and training and testing on samples in both languages. ICDAR 2015 competition dataset is composed of documents written by 500 writers and is divided into training (300 writers), validation (100 writers), and test (100 writers) sets. This competition comprised four tasks: gender classification on Arabic handwriting; gender classification on English handwriting; gender classification using Arabic samples in training and English samples in testing; gender classification using English samples in training and Arabic samples in testing. Unfortunately, we were unable to get the dataset used in the ICDAR 2015 competition. However, we had access to the ICDAR 2013 dataset. Hence, we used it in our experiments. In order to create conditions as similar as

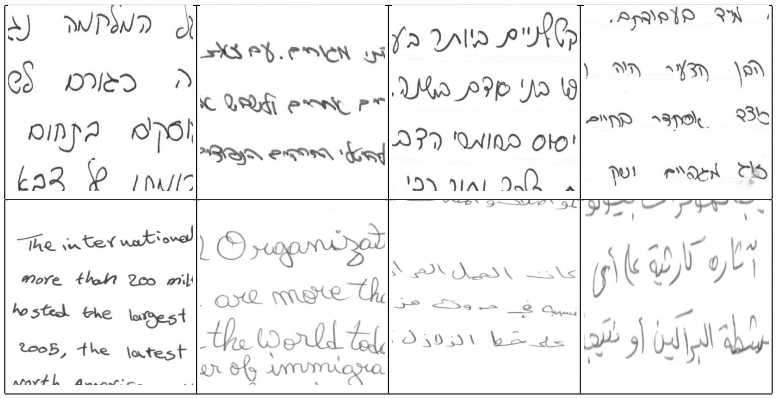

Fig. 2. Samples of the extracted patches; top row - the HHD_gender dataset, bottom row - the QUWI dataset.

possible to the conditions in the ICDAR 2015 competition, we randomly divided the 475 writers into training (300 writers), test (100 writers), and validation (75 writers) sets. This division is comparable to the division used in the ICDAR 2015 competition.

2.2 Experiment settings

For the classification, we investigated ten architectures mentioned at the beginning of Section 2. We performed fine-tuning by replacing the last fully connected layer by fully connected layer with two neurons and freezing all other layers. The models were trained until convergence. We compared the results of all models with the results of a baseline system trained from scratch. This system was adopted from the Kaggle competition on image classification[2] – CNN with 11 (including two convolutional and consequent pooling) layers and 39257 parameters (39193 of them are trainable).

The networks were trained using patches extracted from each document image. The patches were extracted by moving a sliding window of size 400×400 pixels at a stride 200 pixels in vertical and horizontal directions. The patch size was chosen experimentally to include three to four text lines. Figure 2 illustrates the examples of the resulted patches from the QUWI and HHD_gender datasets.

In each experiment, the input images were resized to the input dimensions of the respective network. For prediction, the manuscript image was also cut into overlapping 400×400 pixels patches at a stride of 200×200 pixels, and each was classified. The resulting page-level classification was obtained by the majority voting scheme over all patches from the same page.

[2] https://www.kaggle.com/sujoykg/keras-cnn-with-grayscale-images/

2.3 Results and Discussions

We applied the same models on both HHD_gender and QUWI datasets. For the QUWI dataset, we used the classification scenarios employed in ICDAR 2013 and 2015 competitions, as described in Section 2.1.

The HHD_gender column of Table 1 (left) contains the classification accuracy results for all models on the HHD_gender dataset. As can be seen, Xception provided the best accuracy. However, it consumes the largest number of epochs (40 vs. 15-20 for other models). EfficientNet and NasNet follow with the second best accuracy score.

We compared the results of all the models with the results of a baseline system - CNN with 11 layers. As can be seen, despite its considerably simpler structure and random initialization (the baseline is not pre-trained on ImageNet, in contrast to other models), its performance is comparable to the performance of the best models, even outperforming several networks — both VGGNets, both DenseNets, and Inception-v3. We explain this by a different nature of general pictures and handwriting images. As other studies show [23], pre-training on ImageNet is helpful when not enough training data is provided. However, in the case of the HHD_gender dataset, the training data we have is quite large, of high quality, and is sufficient for accurate learning (for moderate size CNN).

We have also experimented with various types of augmentation using the HHD_gender dataset, applying rotation between -30 to 30 degrees, scaling by a random factor from 0.8 to 1.2, and adding noise to the document. We experimented with these augmentation methods separately and in combination but did not observe any improvement in accuracy rates. Moreover, in some experiments, the augmentation actually harmed the performance. The most likely explanation for this outcome is that the HHD_gender dataset is consistent. Augmentation such as rotation did not help because the forms were aligned horizontally as a part of their preprocessing. Similarly, adding noise did not help because the test set includes clean images. Adding augmentation forced the network to learn examples that are not present in the test set, thereby wasting its predictive resources on irrelevant scenarios.

We performed experiments on the QUWI dataset using both ICDAR 2013 and ICDAR 2015 settings, and compared the results against the top results from these competitions. The ICDAR 2013 column of Table 1 (middle) contains the accuracy scores for the QUWI dataset splittings used in the ICDAR 2013 competition: (1) mono-script English handwriting, (2) mono-script Arabic handwriting, and (3) multi-script handwriting with both languages. In the mono-script experiments, training and testing were performed only on documents in one language; in the multi-script experiment, training and testing were run on handwriting documents in both languages[3]. The accuracy scores of all the models on the QUWI dataset are lower than on the HHD_gender. We explain this by the smaller number of samples used in training in a mono-script setting, and a much

[3] Because only LogLoss scores were reported in the ICDAR 2013 competition, the accuracy scores for the winning system were retrieved from [2].

	HHD_gender	ICDAR 2013			ICDAR 2015			
Models	Hebrew	English	Arabic	Both	2A	2B	2C	2D
Top ICDAR results	-	***0.79***	***0.74***	***0.76***	*0.65*	*0.60*	*0.63*	***0.58***
Baseline	0.81	0.52	0.61	0.56	0.61	0.62	0.65	0.52
VGG16	0.79	0.70	0.59	0.65	0.56	0.67	0.64	0.51
VGG19	0.74	0.69	0.66	0.67	0.55	0.60	0.60	0.52
Xception	**0.85**	0.75	0.68	0.68	0.64	**0.75**	**0.66**	0.55
EfficientNet	0.84	0.75	**0.74**	0.67	**0.67**	0.69	0.59	0.55
Inception-ResNet-v2	0.81	0.71	0.65	0.75	0.65	0.68	0.64	0.54
Inception-v3	0.77	0.73	0.69	0.71	0.61	0.74	0.70	0.54
DenseNet121	0.74	0.68	0.69	0.69	0.61	0.71	0.64	0.51
DenseNet169	0.74	0.71	0.66	0.67	0.63	0.72	0.62	0.50
ResNet50	0.81	0.67	0.50	0.67	0.58	0.65	0.57	0.55
NasNet	0.84	0.60	0.64	0.66	0.66	0.74	0.62	0.55

Table 1. Models' performance on the HHD_gender and QUWI—ICDAR 2013 and 2015 splits—datasets.

more challenging scenario in the multi-script setting. Although the DNNs did not outperform the top results of the ICDAR 2013 competition, they achieved close results. EfficientNet has a clear advantage over the other systems in mono-script learning in both languages, except that Xception has the same score for English. Inception-ResNet-v2 has the best accuracy in multi-script gender detection.

The ICDAR 2015 column of Table 1 (right) shows the comparative results on the QUWI dataset, using the splitting ratios and scenarios as used in the ICDAR 2015 competition. The following scenarios were employed in the competition: (2A) Gender classification on Arabic writings; (2B) Gender classification on English writings; (2C) Gender classification using Arabic samples in training and English samples in testing; (2D) Gender classification using English samples in training and Arabic samples in testing. As can be seen, in 3 out of 4 scenarios Xception and EfficientNet outperform the best results from the ICDAR 2015 competition.

In general, we can see that two networks — Xception and EfficientNet — are the best-performing systems in most scenarios (except for the QUWI with mixed languages and cross-language classification using English samples in training and Arabic samples in testing). Their superiority can be explained by their advanced architectures. The Xception architecture has 36 convolution layers, which form a very strong basis for feature extraction from input handwritings. Because handwriting can be described by numerous features, the greater number of feature maps, which produce the greater number of features, is beneficial to our task. Also, inventors of the Xception model clearly showed the benefit of depthwise separable convolutions in neural computer vision architectures. EfficientNet's "secret" is in its synergy in scaling multiple dimensions together. The authors produced the theoretically optimal formula of "compound scaling" by an extensive grid search and used it to scale up the EfficientNet.

	# participants	accuracy
Questionnaire 1	166	0.623
Questionnaire 2	109	0.632
Questionnaire 3	89	0.739
Questionnaire 4	86	0.707
Average over all questionnaires		0.675

Table 2. The results of human examiners on the samples from the HHD_gender dataset.

We also experimented with different combinations of ensembles of models. In each experiment, the final prediction was assigned using the majority voting over all components. Surprisingly, we found that the ensemble of models does not improve the classification results. We performed an error analysis and found that in most cases the networks gave incorrect predictions on the same documents. Because we were using the majority scheme over predictions by the networks, this resulted in lower classification rates compared to using the best single network. Interestingly, most human examiners have also wrongly classified the same documents on which the networks failed.

Human-Level Performance To compare the models' performance to those of humans, we compiled four online questionnaires[4]. The questionnaires include 70 images from the HHD_gender dataset, divided into 18, 18, 17, 17 groups. Each participant can answer questions from one to four groups, using links from each questionnaire to the following one. Each participant was asked to classify the writer's gender, based on handwritten text samples. Upon completion, a participant can see the classification results. Comparing the results of human examiners, as presented in Table 2, with the results of automated classification from Table 1 (left), we can notice that the best deep-learning models have surpassed human-level performance.

The possible limitation of this experiment is that the classification was carried by the general public. However, the results of the experiment show that the problem is challenging for non-experts, and the automatic classification system can be used in cases when an expert in handwriting analysis is unavailable.

3 Conclusions and Future Work

This paper reports the results of an extensive empirical case study for gender classification from offline handwritten images, performed on different datasets, with various deep CNNs designed for the image classification task. The aim of this study was an extensive investigation of the effect of cross-domain transfer learning with ImageNet pre-training for the gender classification task. We can conclude that advanced deep-learning models outperform conventional machine learning approaches where features must be designed manually. Also, pre-training networks on a rich external dataset (ImageNet) has a positive effect on

[4] https://forms.gle/Ay7XV9CX61fkU6qt7

the gender classification task. For comparison, we also report the results for a simple CNN trained from scratch.

In addition, we established baseline results for a new dataset of handwritten images in Hebrew script, preprocessed and annotated with a writer's gender. The HHD_gender dataset is publicly available for the research community together with the partitions for the training and test sets. To the best of our knowledge, this is the first publicly available dataset in Hebrew script for a gender classification task.

We also show that the problem is challenging for non-experts, and that the automatic classification system can be used in cases when an expert in handwriting analysis is unavailable. In the future, we intend to perform a similar experiment with a graphology expert.

References

1. Ahmed, M., Rasool, A.G., Afzal, H., Siddiqi, I.: Improving handwriting based gender classification using ensemble classifiers. Expert Systems with Applications **85**, 158–168 (2017)
2. Akbari, Y., Nouri, K., Sadri, J., Djeddi, C., Siddiqi, I.: Wavelet-based gender detection on off-line handwritten documents using probabilistic finite state automata. Image and Vision Computing **59**, 17–30 (2017)
3. Al Maadeed, S., Ayouby, W., Hassaine, A., Aljaam, J.M.: QUWI: an Arabic and English handwriting dataset for offline writer identification. In: International Conference on Frontiers in Handwriting Recognition. pp. 746–751 (2012)
4. Al Maadeed, S., Hassaine, A.: Automatic prediction of age, gender, and nationality in offline handwriting. EURASIP Journal on Image and Video Processing **2014**(1), 1–10 (2014)
5. Bi, N., Suen, C.Y., Nobile, N., Tan, J.: A multi-feature selection approach for gender identification of handwriting based on kernel mutual information. Pattern Recognition Letters **121**, 123–132 (2019)
6. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
7. Djeddi, C., Al-Maadeed, S., Gattal, A., Siddiqi, I., Souici-Meslati, L., El Abed, H.: ICDAR 2015 competition on multi-script writer identification and gender classification using 'QUWI' database. In: International Conference on Document Analysis and Recognition. pp. 1191–1195 (2015)
8. Gattal, A., Djeddi, C., Bensefia, A., Ennaji, A.: Handwriting based gender classification using cold and hinge features. In: International Conference on Image and Signal Processing. pp. 233–242 (2020)
9. Gattal, A., Djeddi, C., Siddiqi, I., Chibani, Y.: Gender classification from offline multi-script handwriting images using oriented basic image features (oBIFs). Expert Systems with Applications **99**, 155–167 (2018)
10. Goodenough, F.L.: Sex differences in judging the sex of handwriting. The Journal of Social Psychology **22**(1), 61–68 (1945)
11. Hamid, S., Loewenthal, K.M.: Inferring gender from handwriting in Urdu and English. The Journal of social psychology **136**(6), 778–782 (1996)

12. Hassaïne, A., Al Maadeed, S., Aljaam, J., Jaoua, A.: ICDAR 2013 competition on gender prediction from handwriting. In: International Conference on Document Analysis and Recognition. pp. 1417–1421 (2013)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
15. Illouz, E., David, E.O., Netanyahu, N.S.: Handwriting-based gender classification using end-to-end deep neural networks. In: International Conference on Artificial Neural Networks. pp. 613–621 (2018)
16. Liwicki, M., Schlapbach, A., Bunke, H.: Automatic gender detection using on-line and off-line information. Pattern Analysis and Applications **14**(1), 87–92 (2011)
17. Maken, P., Gupta, A.: A method for automatic classification of gender based on text-independent handwriting. Multimedia Tools and Applications pp. 1–30 (2021)
18. Moetesum, M., Siddiqi, I., Djeddi, C., Hannad, Y., Al-Maadeed, S.: Data driven feature extraction for gender classification using multi-script handwritten texts. In: International Conference on Frontiers in Handwriting Recognition. pp. 564–569 (2018)
19. Najla, A.Q., Suen, C.Y.: Gender detection from handwritten documents using concept of transfer-learning. In: International Conference on Pattern Recognition and Artificial Intelligence. pp. 3–13 (2020)
20. Rabaev, I., Kurar Barakat, B., Churkin, A., El-Sana, J.: The HHD dataset. In: International Conference on Frontiers in Handwriting Recognition. pp. 228–233 (2020)
21. Rahmanian, M., Shayegan, M.A.: Handwriting-based gender and handedness classification using convolutional neural networks. Multimedia Tools and Applications pp. 1–24 (2021)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
23. Studer, L., Alberti, M., Pondenkandath, V., Goktepe, P., Kolonko, T., Fischer, A., Liwicki, M., Ingold, R.: A comprehensive study of ImageNet pre-training for historical document image analysis. In: International Conference on Document Analysis and Recognition. pp. 720–725 (2019)
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
25. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114 (2019)
26. Topaloglu, M., Ekmekci, S.: Gender detection and identifying one's handwriting with handwriting analysis. Expert Systems with Applications **79**, 236–243 (2017)
27. Upadhyay, S., Singh, J., Shukla, S.: Determination of sex through handwriting characteristics. Int J Cur Res Rev— Vol **9**(13), 11 (2017)
28. Xue, G., Liu, S., Gong, D., Ma, Y.: ATP-DenseNet: a hybrid deep learning-based gender identification of handwriting. Neural Computing and Applications pp. 1–12 (2020)
29. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: IEEE conference on computer vision and pattern recognition. pp. 8697–8710 (2018)