Gender Prediction from Handwritten Text: a Case Study

No Institute Given

Abstract. Automatic author's gender prediction from a handwritten sample is an essential task in a wide range of areas, e.g., forensic analysis, psychology, and historical document analysis. Despite a high interest within a broad spectrum of research communities, comparably few works were published in this area while mostly being limited to English and Arabic languages. The difficulty of this task can be demonstrated by the relatively low (below 90%) performance of state-of-the-art systems and even lower performance of human examiners. This paper presents the results of a case study with multiple supervised models for gender prediction from the handwritten text. The study was performed on two datasets, the subset of the QUWI dataset, consisting of handwritten documents in English and Arabic, and a new dataset of documents in Hebrew script. We perform extensive experiments and analyze and compare the results obtained with different networks and datasets. The paper also describes a new dataset for gender classification, comprised of handwritten documents in Hebrew script, and compares the obtained results versus human-level performance.

Keywords: Gender Prediction · Handwritten Document Analysis · Supervised learning · Neural Network · Binary Classification.

1 Introduction

Handwritten documents are considered one of the oldest means of communication. A person learns to write by copying forms that are adapted and vary according to several factors, like geographical region, different times, as well as social and cultural features [16]. Handwriting gender classification is of great interest due to the broad range of areas it can be applied in, ranging from medicine to study human behavior and health to historical documents analysis and criminological investigations. Psychological studies of handwriting analysis have confirmed that classification of gender can be made according to several significant differences in handwriting [11, 12, 32]. While female's handwriting tends to be more uniform, ordered, and has greater circularity, male's handwriting tends to be more pointed, messy, and slanted. With technological advances in image analysis and computer vision, manual handwriting analysis is enhanced by automatic systems. Early approaches for handwriting gender classification used traditional machine learning algorithms [19, 13, 5, 31, 9], while more recent studies use deep learning networks [22, 17, 23] that do not require manual feature engineering.

This work describes a case study for gender classification from a handwritten text with multiple supervised approaches. We perform extensive experiments and analyze the results on two different datasets: a new dataset of Hebrew documents and the subset of the QUWI dataset [4] which consists of documents written in English and Arabic languages. Namely, this paper has two main contributions:
(1) It introduces and analyses a new Hebrew Gender (HebG) dataset of handwritten documents in Hebrew script. The dataset is freely available in its preprocessed format to the scientific community[1]. To the best of our knowledge, it is the first publicly available dataset in Hebrew script for gender classification.
(2) It reports comparative results for ten systems on two previously mentioned datasets and interprets results. Besides, the performances of the models are compared versus human-level performance.

2 Related Work

In forensic, psychology and historical documents communities, gender identification from handwritten samples has attracted significant attention over the years [11, 12, 32].

In the last decades, several studies have addressed the problem of automatic gender classification based on handwriting analysis.

Early methods used traditional machine learning approaches based on manual feature engineering, such as decision tree by making use of 133 graphology attributes [31]; ensemble of artificial neural networks, support vector machine, nearest neighbor classifier, decision trees, and random forests, applied on a set of textural features extracted from handwriting samples [2]; and support vector machine (SVM) employed on a set of multiple handwriting features, extracted by different methods and then selected by kernel mutual information [6], basic image features [10], Cloud of Line Distribution (COLD) and Hinge features [9], and features produced by the wavelet analysis [3]. Authors of [3] also employed artificial neural network as an additional classification model. Each handwritten image in this work was converted into a series of wavelet sub-bands, which were then extended into data sequences. Each data sequence was quantized to produce a probabilistic finite state automata that generated feature vectors.

ICDAR competitions on Gender Prediction from Handwriting [13, 8] gathered researchers from around a world to compare between different techniques in gender prediction from handwriting. The competition in ICDAR 2013 has attracted 194 teams from both academia and industry. All methods followed the standard approach of the traditional machine learning, where feature extraction must be performed prior to the classification. Eight teams competed in the ICDAR 2015 gender classification competition, which comprised four tasks: two mono-script scenarios, i.e., training and testing on the same language, and two cross-script scenarios, i.e., training on English and testing on Arabic, and

[1] Following the double-blind policy, we omit the link to the dataset at this stage

vice versa. All the methods used feature extraction followed by traditional machine learning algorithms. A highest classification rate of 65% was achieved in mono-script task on Arabic script.

Recent works mainly incorporate deep learning techniques—most are based on convolutional neural network (CNN)—where no feature engineering is required. In [24], gender and handedness classification have been examined by using advanced CNNs, such as DenseNet201, InceptionV3, and Xception. Two databases, IAM [20] (English texts) and KHATT [1] (Arabic texts) have been employed in this study. The authors achieved 84% accuracy on IAM and 75% accuracy on KHATT with the proposed methodology. In [33], attention-based two-pathway densely connected convolutional networks (ATP-DenseNet) is proposed to identify the gender of handwritten document. The proposed model performed with accuracy below 80% on two datasets, IAM and KHATT.

Authors of [17] introduced a private dataset of labeled handwritten samples, in Hebrew and English, collected from 405 participants. They applied a CNN, which performs automatic feature extraction from a given handwritten image, followed by classification of the writer's gender. Comparing the gender classification performance against human examiners, the authors showed that the proposed deep learning-based approach is substantially more accurate than that of humans. Namely, the automatic classification performed with 77% accuracy for both languages; and with 74.61% and 79.34% accuracy on Hebrew and English texts, respectively. In comparison, for human examiners, the average classification accuracy for English and Hebrew handwritings was 63.6% and 66.2%, respectively.

Transfer learning was applied in [23] to detect the writer gender from scanned handwritten documents. Authors used two pre-trained CNN, GoogleNet, and ResNet, as fixed feature extractors. For the classification stage, they applied SVM. The obtained classification accuracy was 80.05% and 83.32% for GoogleNet and ResNet, respectively.

In [21], pre-trained CNNs have been employed as feature extractors to discriminate male and female handwriting, while classification is carried out using a number of classifiers, with Linear Discriminant Analysis being the most effective. Feature extraction was performed over words, patches, and page images. Authors achieved a classification accuracy of 70% on the QUWI handwriting dataset, combining English and Arabic samples [4].

3 Case Study

3.1 Methods

All recent most successful models for image classification are CNN-based. It has been shown that shallow layers extract simple (low-level) features of an image, and deeper layers can extract more complex (high-level) features. Thus, to make CNN more accurate, researchers mainly increase their depth by adding more layers. One of the first successful products of this approach is VGGNet, introduced by Karen Simonyan and Andrew Zisserman [25] for large-scale image classification. VGGNet demonstrated that the representation depth is beneficial for the classification accuracy. VGGNet is composed of a sequence of convolutional and pooling layers, followed by three dense layers. In this work, we use **VGG16** and **VGG19**. The main difference between them is a number of convolutional layers. VGG16 has 16 layers with learnable weights: 13 convolutional layers, and 3 fully connected layers, while there are 19 layers with learnable weights—16 convolutional layers, and 3 fully connected layers—at VGG19.

Authors of [14] presented a residual learning framework to ease the training of networks that are substantially deeper than those used previously. They reformulated the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. The core idea of ResNet is introducing a so-called "identity shortcut connection" that skips one or more layers. The authors empirically showed that these residual networks (ResNets) are easier to optimize, and that networks can gain accuracy from considerably increased depth. On the ImageNet dataset[2], ResNet with a depth of 152 layers (8 times deeper than VGGNet) had lower runtime than VGGNet. An ensemble of these ResNets won the 1st place on the ILSVRC 2015 classification task. We use **ResNet50**, which is a variant of ResNet model which has 48 Convolution layers, 1 MaxPool, and 1 Average Pool layer.

The Inception deep convolutional architecture was introduced in [28] and was called GoogLeNet (or Inception-v1). Later, the Inception architecture was refined in various ways, first by the introduction of batch normalization [18] (Inception-v2) by Ioffe et al; subsequently, the architecture was improved by additional factorization ideas in the third iteration [29] which is referred to as **Inception-v3** in this paper. Next year, this architecture was refined again, in [27], and several architectures for Inception-ResNet, including **Inception-ResNet-v2** were proposed and evaluated. In our case study, we use two Inception-based models: Inception-v3 and Inception-ResNet-v2. The difference between them is that Inception-v3 is a deep CNN not utilizing residual connections, while Inception-ResNet-v2 is Inception style networks that utilize residual connections instead of filter concatenation. According to [27], Inception-ResNet-v2 is a hybrid Inception version which has a higher computational cost and significant improvement of recognition performance, in comparison to earlier versions.

Xception model was proposed by Francois Chollet from Google in [7]. Xception is an extension of the Inception architecture which replaces the standard Inception modules with depthwise Separable Convolutions. Xception slightly outperformed Inception-v3 on the ImageNet dataset and significantly outperformed it on a larger image classification dataset. Since the Xception architecture has the same number of parameters as Inception-v3, the performance gains are due to a more efficient use of model parameters. The Xception architecture has 36 convolutional layers performing feature extraction from input pictures. These layers are arranged into 14 modules, each one having linear residual connections around it, except for the first and last modules. As such, the Xception

[2] http://www.image-net.org/

architecture forms a linear stack of depthwise separable convolutional layers with residual connections. This makes the architecture very easy to define and modify, similar to VGG-based models, but dissimilar to Inception-based models which are much more complex.

While CNNs go deeper and the path from the network input layer to its output layer becomes longer, the chance of information to reach the other side gets lower. Dense Convolutional Network (DenseNet) [15] solves this problem by ensuring maximum information flow. To do it, DenseNet connects each layer to every other layer in a feed-forward fashion. Whereas traditional convolutional networks with N layers have N connections – one between each layer and its subsequent layer – DenseNet has $N(N+1)/2$ direct connections. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers, so no need in learning redundant feature maps. DenseNet has several compelling advantages: it alleviates the vanishing-gradient problem, strengthens feature propagation, exploits the potential of the network through feature reuses, and substantially reduces the number of parameters. In our case study, we apply **DenseNet121** and **DenseNet169**, where 121 and 169 denote the depths of these models, respectively.

Authors of [34] introduced **NASNet** model. They proposed Neural Architecture Search (NAS) for Cells—to search for an architectural building block on a small dataset and then transfer the block to a larger dataset. Though the overall architecture is predefined, the blocks or cells are not predefined by authors. Instead, they are searched by reinforcement learning search method. Particularly, authors of NASNet searched for the best convolutional layer or cell on CIFAR-10[3] first, then apply this cell to the ImageNet by stacking together more copies of this cell. Authors also proposed a new regularization technique called ScheduledDropPath, which significantly improved the generalization in the NASNet model. NASNet achieved state-of-the-art results with smaller model size and lower complexity.

Authors of [30] proposed a new scaling method, **EfficientNet**, that uniformly scales all dimensions of depth, width,and resolution using a simple yet highly effective compound coefficient, based on an observation that carefully balancing network depth, width, and resolution can lead to better performance. Authors demonstrated the effectiveness of this method on scaling up MobileNets and ResNet models.

We used open-source implementations of all mentioned models using Keras and TensorFlow, are provided as part of the Keras Applications module[4].

3.2 Datasets

The experimental study is performed on two datasets that consist of documents written in three different languages, the HebG dataset and the subset of QUWI [4] dataset. In this section, we describe both datasets.

[3] https://www.cs.toronto.edu/ kriz/cifar.html

[4] https://keras.io/api/applications/

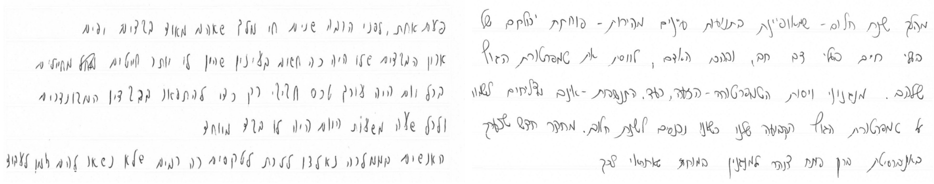

Fig. 1. Examples of the images from the HebG dataset; male - left image, female - right image.

The HebG Dataset The HebG dataset contains 819 handwritten forms written by volunteers of different educational backgrounds and ages (as young as 11 years old and as old as late 60s), both native and non-native Hebrew speakers. Each participant filled (voluntarily) statistical information, such as date, town, 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 62 words on average. The forms are scanned in color with the resolution of 600 dpi resolution in TIFF format.

The forms were preprocessed as follows. First, the handwritten paragraph was extracted using the corresponding text box's coordinates. 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. About 30 documents did not contain gender information (the participant preferred not to fill in personal information). These forms 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. The examples of the processed images are illustrated in Figure 1. For training, the HebG dataset was randomly subdivided into training (80%), validation (10%) and test (10%) sets.

The QUWI dataset The QUWI [4] dataset contains handwritten documents in Arabic and English languages. The documents were written by volunteers of different ages, 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. Figure 2 illustrates the sample pages from the QUWI dataset.

The ICDAR 2013 and 2015 competitions on Gender Prediction from Handwriting [13, 8] used a subset of the QUWI dataset. ICDAR 2013 dataset is composed of handwritten documents of 475 writers and is divided into training (282 writers) and test (193 writers) sets. Three classification schemes are 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

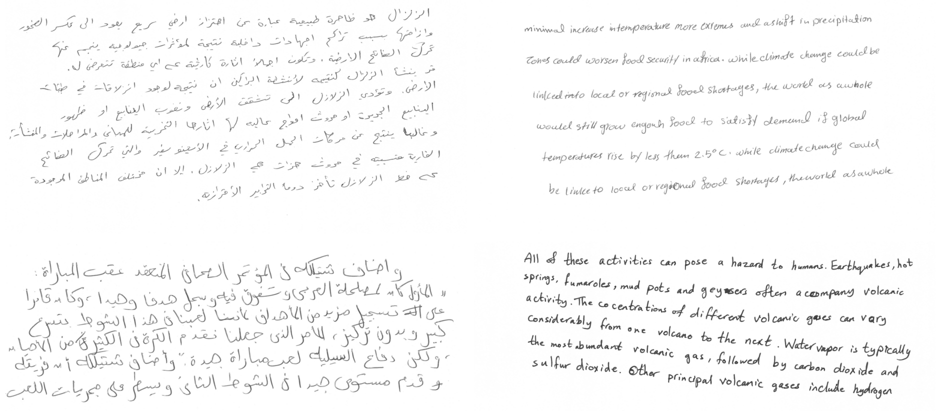
minimal increase intemperature more extremes and a shift in precipitation
zones could worsen food security in africa. while climate change could be
linked into local or regional food shortages, the world as awhole
would still grow engouh food to satisfy demand if global
temperatures rise by less them 2.5°C. while climate change could
be linke to local or regional food shortages, the world as a whole
All of these activities can pose a hazard to humans. Earthquakes, hot
springs, fumaroles, mud pots and geysers often accompany volcanic
activity. The cocentrations of different volcanic gases can vary
considerably from one volcano to the next. Water vapor is typically
the most abundant volcanic gas, followed by carbon dioxide and
sulfur dioxide. Other principal volcanic gases include hydrogen

Fig. 2. Examples of the images from the QUWI dataset. Female - first row, male - second row. Different writers.

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 test; gender classification using English samples in training and Arabic samples in test. Unfortunately, we were unable to get the exact 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 the conditions as similar as possible to the conditions in the ICDAR 2015 competition, we 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.

3.3 Experiment settings

For the classification, we investigated ten architectures described in Section 3.1. 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 are trained until convergence. We compared the results of all models with the results of a baseline system adopted from Kaggle competition on image classification[5] – CNN with 11 (including two convolutional and consequent pooling) layers and 39257 parameters (39193 of them are trainable).

3.4 Data Preparation

The networks were trained using two different modes, which we refer to as *page-based mode* and *patch-based mode*. In the page-based mode, the input to the

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

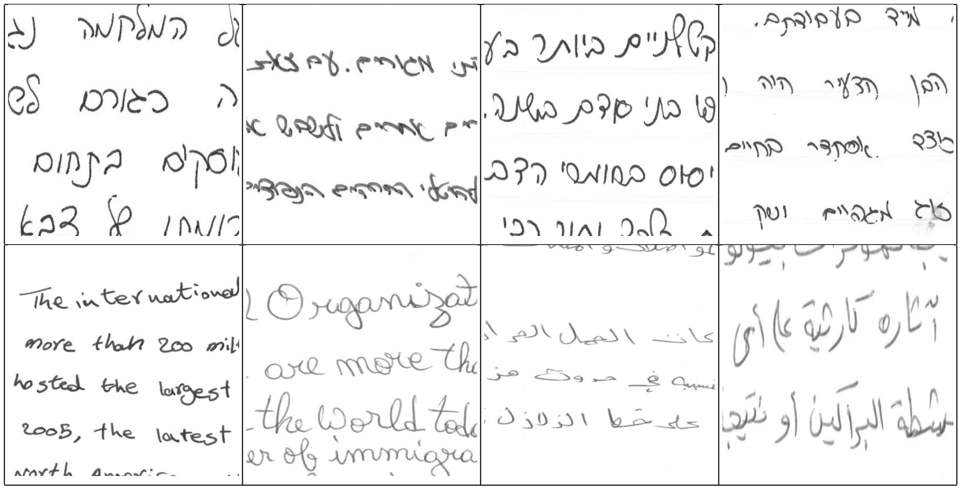

Fig. 3. Samples of the extracted patches; top row - the HebG dataset, bottom row - the QUWI dataset.

networks are the original document images; in the patch-based mode, the networks are trained using patches extracted from each document image instead of training a network on a whole image. The patches are extracted by moving a sliding window of size 400 × 400 with the stride 200 in vertical and horizontal directions. It is important to emphasize that in both modes, page-based and patch-based, the prediction is made on the page level. The patch size was chosen experimentally to include 3-4 text lines. Figure 3 illustrates the examples of the resulted patches from the QUWI and HebG datasets. As we show in Section 3.5, the patch-based mode obtains better performances.

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

3.5 Results

We applied the same models on both, HebG and QUWI, datasets. For the QUWI dataset, we used the classification scenarios employed in ICDAR 2013 and 2015 competitions, as described in Section 3.2. The best number of epochs for each model was obtained by incremental increasing its number from the initial one (10) until a model's performance stopped to improve. We report accuracy rates for a patch-based mode classification and a page-based mode classification, where the former is produced by using patches classification and then majority voting, and the latter is produced by using a page itself as input, without patch extraction.

Table 1. Models' accuracy on the HebG dataset.

Model	patch-based mode	page-based mode
Baseline	*0.81*	*0.70*
VGG16	0.79	0.75
VGG19	0.74	0.75
Xception	**0.85**	**0.79**
EfficientNet	0.84	0.75
Inception-ResNet-v2	0.81	0.69
Inception-v3	0.77	0.68
DenseNet121	0.74	0.66
DenseNet169	0.74	0.70
ResNet50	0.81	0.75
NasNet	0.84	0.73

Table 1 contains the results (classification accuracy) for all models on the HebG dataset. As can be seen, Xception has the best accuracy for both modes, with and without patch extraction. However, it consumes the largest number of epochs (40 vs. 15-20 for other models). EfficientNet produces the second best accuracy. We compared the results of all models with the results of a baseline system - CNN with 11 layers. As can be seen, despite considerably simple structure and random initialization (the baseline is not pre-trained on ImageNet, in contrast to other models), the baseline's performance is comparable to the performance of the best models, and it even outperforms several networks — both VGGNets, both DenseNets, and Inception-v3 — in patch-based mode. We explain it by a different nature of general pictures and handwriting pages and a need to train the classification models on the original training data instead of pre-training and consequent fine-tuning. As other studies show, pre-training on ImageNet is helpful when not enough training data is provided. However, in our case, we have quite large, high-quality (and sufficient for accurate learning) training data.

We have also experimented with various types of augmentation using the HebG dataset: 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 even harmed the performance. The possible explanation for this outcome is that the HebG dataset is consistent, meaning that all forms are similar and were scanned in the same conditions. Adding augmentation forced the network to learn the scenarios that are not present in the test set and waste its predictive resources on irrelevant scenarios.

We have performed experiments on the QUWI dataset using both ICDAR 2013 and ICDAR 2015 settings, and compared the results versus the top results from these competitions. Table 2 contains the results (classification accuracy) for the QUWI dataset splittings used in ICDAR 2013 competition, including

Table 2. Models' performance on the QUWI dataset. ICDAR 2013 split. Patch-based mode.

Model	English	Arabic	Both languages
Top ICDAR'13 results	***0.79***	***0.74***	***0.76***
Baseline	0.52	0.61	0.56
VGG16	0.70	0.59	0.65
VGG19	0.69	0.66	0.67
Xception	0.75	0.68	0.68
EfficientNet	0.75	**0.74**	0.67
Inception-ResNet-v2	0.71	0.65	0.75
Inception-v3	0.73	0.69	0.71
DenseNet121	0.68	0.69	0.69
DenseNet169	0.71	0.66	0.67
ResNet50	0.67	0.50	0.67
NasNet	0.60	0.64	0.66

mono-script English handwriting, mono-script Arabic handwriting, and multi-script handwriting with both languages. In mono-script experiments, training and testing were performed only on documents in one language, in multi-script experiment training and testing were run on handwriting documents in both languages. We report only patch-based mode scores due to the obvious advantage of this mode over page-based mode in all our experiments. The ICDAR 2013 winning system used a Gradient Boosting Machine for both feature selection and classification. Despite our top models did not outperform the top results of the ICDAR 2013 competition, they achieved very close results. EfficientNet has a clear advantage in mono-script learning in both languages over the other systems, except Xception having the same score for English. Inception-ResNet-v2 has the best accuracy in multi-script gender detection. As can be seen, the superiority of the advanced models over simpler baseline CNN model is obvious. Table 3 shows the comparative results on the QUWI dataset, using the splitting ratios and scenarios as in ICDAR 2015 competition. The following scenarios were employed: (2A) Gender classification on Arabic writings; (2B) Gender classification on English writings; (2C) Gender classification using Arabic samples in training and English samples in test; (2D) Gender classification using English samples in training and Arabic samples in test. As can be seen, in 3 out of 4 scenarios Xception and EfficientNet outperform the best results from the ICDAR 2015 competition. The best ICDAR 2015 competition system – CVC system – employed a variant of Local Binary Patterns (LBPs) developed specifically for textual content. LBPs are concatenated to form a feature vector. Principal Component Analysis (PCA) is then applied for dimensionality reduction and extracting the relevant features for characterizing the writer and the gender of the writer.

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

Table 3. Models' accuracy on the QUWI dataset. Comparison with the results of ICDAR 2015 competition on gender prediction.

Model	2A	2B	2C	2D
Top ICDAR'15 results	*0.650*	*0.600*	*0.630*	***0.580***
Baseline	*0.605*	*0.615*	*0.645*	*0.520*
VGG16	0.560	0.665	0.635	0.510
VGG19	0.545	0.600	0.600	0.520
Xception	0.635	**0.745**	**0.655**	0.545
EfficientNet	**0.665**	0.685	0.588	0.550
Inception-ResNet-v2	0.650	0.680	0.640	0.540
Inception-v3	0.605	0.740	0.695	0.540
DenseNet121	0.605	0.705	0.640	0.505
DenseNet169	0.625	0.715	0.615	0.500
NasNet	0.660	0.740	0.615	0.550
ResNet50	0.575	0.650	0.570	0.550

Arabic samples in test). We discuss their architectures and possible advantages over other models in Section 3.6.

Ensembles of models We have also experimented with different combinations of ensembles of models. In each experiment, the final prediction is assigned using the majority voting over all components. Surprisingly, we found that the ensemble of models does not improve the classification results. We have performed an error analysis and found that the networks gave wrong predictions on the same documents in most cases. Since we were using the majority scheme over networks predictions, 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 There is a theoretical bound on the lowest possible error rate for any classifier – the Bayes error. In many cases, a human-level performance is very close to the Bayes error. To compare the models' performance to those of humans, we compiled four online questionnaires[6]. The questionnaires include 70 manuscript images in total and are divided into 18, 18, 17, 17 groups. Each participant can answer questions from one to four groups, using links from each questionnaire to the consequent one. Each participant was asked to predict the writer's gender of handwritten text samples. Upon completion, a participant can see the classification results. Comparing the results of human examiners, presented in Table 4, with the results of automated classification from Table 1, we can notice that the best deep-learning models have surpassed human-level performance. The human-level results are consistent with that reported in [17].

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

	# 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 4. The results of human examiners on the samples from the HebG dataset.

3.6 Results Analysis

In this case study, we treat handwritings as images and gender identification as a binary classification task. That explains why we decided to apply the most advanced state-of-the-art networks, originally designed for image classification, to our task. The following observations can be made from the results:

- Half of advanced (very deep) networks applied to our task perform better that the baseline CNN model. This outcome supports the hypothesis that increasing neural network's depth refines features extraction, which is not less crucial and challenging in document image domain than in general image classification. However, the baseline's superiority over another half of pre-trained networks supports another hypothesis that training from scratch may have advantage if we have enough training data of good quality, like in our case.
- Based on the results on the QUWI dataset, we can conclude that advanced deep-learning models outperform conventional machine learning approaches where features need to be designed manually. Extracting features from handwriting is a very challenging task. As such, deep neural networks save us much work while providing very reasonable results.
- Two networks had an exceptional performance in most scenarios: Xception and EfficientNet. 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 plenty of features, the greater number of feature maps, which produce the greater number of features, is beneficial to our task. Also, inventors of Xception model clearly showed the benefit of depthwise separable convolutions in neural computer vision architectures, offering similar properties as Inception modules but easily used as regular convolutional layers.
  EfficientNet is the most advanced CNN, with significant improvement in both accuracy and efficiency over most current models. Its "secret" is in 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.
  Both networks are pre-trained on ImageNet, and their superiority comes in line with the observation made in [26] that pre-training networks on a rich external dataset (ImageNet) may have a positive effect on several document image analysis tasks.

– Patch-based mode always improves the classification rate, which can be naturally explained by larger training data. Usually, with the appropriate patch size, we can multiply training samples without information loss. An additional advantage of employing patches is a reduction in memory utilization.
– Despite our expectations, the augmentation of training data did not improve the models' accuracy. The possible explanation for this is that the HebG dataset is consistent. All forms have been scanned in the same conditions with the same resolution. The augmentation like rotation did not help because the forms were aligned horizontally as a part of their preprocessing. The rotation forced the network to learn the examples that are not present in the test set because, at the test time, we only use horizontally aligned images. Similarly, adding noise did not help since the test set includes clean images.
– All deep networks distinguished between authors' genders based on their handwriting better than humans. However, the results are still relatively low – the best accuracy on the HebG dataset is 85%, and the best accuracy on the QUWI is 74.5% in the mono-script Arabic scenario. This outcome demonstrates the real challenge in the gender identification task.

4 Conclusion

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

In addition, we introduce a new dataset of handwritten images in Hebrew script, preprocessed and annotated with a writer gender. We show on that dataset that deep networks can distinguish between writers' genders with reasonable accuracy, even better than humans. The HebG dataset is publicly available for the research community together with the train and set partitions. To the best of our knowledge, this is the first publicly available dataset in Hebrew script for a gender classification task.

References

1. Ahmad, R., Naz, S., Afzal, M.Z., Rashid, S.F., Liwicki, M., Dengel, A.: KHATT: A deep learning benchmark on Arabic script. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 7, pp. 10–14. IEEE (2017)
2. 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)
3. 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)

4. Al Maadeed, S., Ayouby, W., Hassaine, A., Aljaam, J.M.: QUWI: an Arabic and English handwriting dataset for offline writer identification. In: 2012 International Conference on Frontiers in Handwriting Recognition. pp. 746–751. IEEE (2012)
5. 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)
6. 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)
7. 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)
8. 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: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1191–1195. IEEE (2015)
9. 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. Springer (2020)
10. 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)
11. Goodenough, F.L.: Sex differences in judging the sex of handwriting. The Journal of Social Psychology **22**(1), 61–68 (1945)
12. Hamid, S., Loewenthal, K.M.: Inferring gender from handwriting in Urdu and English. The Journal of social psychology **136**(6), 778–782 (1996)
13. Hassaïne, A., Al Maadeed, S., Aljaam, J., Jaoua, A.: ICDAR 2013 competition on gender prediction from handwriting. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1417–1421. IEEE (2013)
14. 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)
15. 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)
16. Huber, R.A., Headrick, A.M.: Handwriting identification: facts and fundamentals. CRC press (1999)
17. 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. Springer (2018)
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)
19. Liwicki, M., Schlapbach, A., Loretan, P., Bunke, H.: Automatic detection of gender and handedness from on-line handwriting. In: Proc. 13th Conf. of the Graphonomics Society. pp. 179–183 (2007)
20. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**(1), 39–46 (2002)

21. Moetesum, M., Siddiqi, I., Djeddi, C., Hannad, Y., Al-Maadeed, S.: Data driven feature extraction for gender classification using multi-script handwritten texts. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 564–569. IEEE (2018)
22. Morera, Á., Sánchez, Á., Vélez, J.F., Moreno, A.B.: Gender and handedness prediction from offline handwriting using convolutional neural networks. Complexity **2018** (2018)
23. 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. Springer (2020)
24. Rahmanian, M., Shayegan, M.A.: Handwriting-based gender and handedness classification using convolutional neural networks. Multimedia Tools and Applications pp. 1–24 (2021)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. 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: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 720–725. IEEE (2019)
27. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
28. 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)
29. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
30. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
31. Topaloglu, M., Ekmekci, S.: Gender detection and identifying one's handwriting with handwriting analysis. Expert Systems with Applications **79**, 236–243 (2017)
32. Upadhyay, S., Singh, J., Shukla, S.: Determination of sex through handwriting characteristics. Int J Cur Res Rev— Vol **9**(13), 11 (2017)
33. 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)
34. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8697–8710 (2018)