A Classification of Arab Ethnicity Based on Face Image Using Deep Learning Approach

NORAH A. AL-HUMAIDAN AND MASTER PRINCE

Department of Computer Science, Qassim University, Mulaydha 51452, Saudi Arabia
Corresponding author: Norah A. Al-Humaidan (noura.ah1493@gmail.com)

This work was supported by the Qassim University, Saudi Arabia, to complete Master Thesis under the course M.S. in computer science.

ABSTRACT

Human face and facial features gain a lot of attention from researchers and are considered as one of the most popular topics recently. Features and information extracted from a person are known as soft biometric, they have been used to improve the recognition performance and enhance the search engine for face images, which can be further applied in various fields such as law enforcement, surveillance videos, advertisement, and social media profiling. By observing relevant studies in the field, we noted a lack of mention of the Arab world and an absence of Arab dataset as well. Therefore, our aim in this paper is to create an Arab dataset with proper labeling of Arab sub-ethnic groups, then classify these labels using deep learning approaches. Arab image dataset that was created consists of three labels: Gulf Cooperation Council countries (GCC), the Levant, and Egyptian. Two types of learning were used to solve the problem. The first type is supervised deep learning (classification); a Convolutional Neural Network (CNN) pre-trained model has been used as CNN models achieved state of art results in computer vision classification problems. The second type is unsupervised deep learning (deep clustering). The aim of using unsupervised learning is to explore the ability of such models in classifying ethnicities. To our knowledge, this is the first time deep clustering is used for ethnicity classification problems. For this, three methods were chosen. The best result of training a pre-trained CNN on the full Arab dataset then evaluating on a different dataset was 56.97%, and 52.12% when Arab dataset labels were balanced. The methods of deep clustering were applied on different datasets, showed an ACC from 32% to 59%, and NMI and ARI result from zero to 0.2714 and 0.2543 respectively.

INDEX TERMS

Arab, convolutional neural network (CNN), deep learning, deep clustering, ethnicity.

I. INTRODUCTION

The research on human face ethnicity and gender recognition was initially studied by psychologists from the perspective of cognitive science [1].

In Computer Vision, human face and facial features gain a lot of attention from researchers and are considered as one of the most popular topics recently [2], [3]. Features and information extracted from a person such as age, gender, and ethnicity, are known as soft biometric. Other soft biometric examples are hair color, eye color, height, width, scars, marks, the shape of nose and mouth, etc. Soft biometrics can’t identify a person’s identity by itself. However, it has been used to improve the recognition performance by combining it with hard biometric [4]. Veropoulos et al. [5] used soft biometric classification as a filtering step to limit the search space in databases for user identification systems. Also Kumar et al. [6] used soft biometrics to enhance the search engine for face images. Besides, in Human-Computer Interface (HCI) field, the computer can provide speech recognition or offer options to the user based on soft biometrics [7], [8]. In surveillance videos, based on soft biometrics, suspects can be located [8]. Niinuma et al. [9] used it for continuous user authentication. It also can be used in law enforcement, advertisement, social media profiling [3], [10].

As mentioned before, ethnicity is a soft biometric. It is defined as: “the fact or state of belonging to a social group that has a common national or cultural tradition” [11]. It is multifaceted and keeps changing over time based on cultural or geographical factors. Therefore, it is used for people who share the same race, language, nationality, or religion, etc [12], [13]. In Computer Vision, ethnicity classification using face image has been getting a lot of attention that addresses two main topics; basic races such as Black, White,
Asian, etc. [3], [14]. The second type focuses on smaller ethnic groups or sub-ethnic groups, these groups can be people of the same nationality, for example [15] proposed a Myanmar / non-Myanmar classification method. [16] focused on East Asian countries: Vietnam, Burma, Thailand, China, Korea, Japan, Indonesia, and Malaysia, while [10] classified Bangladeshis, Chinese and Indian people. Besides, sub-ethnic groups can define smaller groups in the same country, for example, [17] perform classification on eight ethnic groups from China.

In this research, our focus is on Arab ethnicity. Arab refers to people who live in the Arab world, which consists of 22 countries from North Africa and Western Asia who have Arabic as their official language [18]. Our contribution is to provide an Arab face dataset that consists of three labels which are; Gulf Cooperation Council countries, Levant, and Egyptians, and perform supervised classification on it. And lastly, test some deep clustering methods on our dataset and a benchmark dataset. To the best of our knowledge, this is the first time that deep clustering is performed to solve the ethnicity classification problem.

The rest of the paper is divided into 5 sections. The second section contains literature reviewed. The methodology comes after that, in section four experimental results are discussed. Finally, the entire work is concluded and further developments are suggested.

II. RELATED WORK

The interest in CNN reached the field of ethnicity classification. In this section, we will summarize some studies that used CNN for ethnicity classification.

Heng et al. [10] hybridized a pre-trained CNN classifier (VGG-16), which was trained on ImageNet, output with image ranking engine and train Support Vector Machine (SVM) using the hybrid features. The approach evaluated on a new dataset which contains Bangladeshis, Chinese and Indian people faces. The result showed improvement when compared to Faster R-CNN and Wang’s method with an accuracy of 95.2%.

Narang and Bourlai [19] investigated the problems of distance, night time, and uncontrolled conditions of face images. They used NIR images taken from 30, 60, 90, and 120 meters at night and visible images at a distance of 1.5 meters from the Long Distance WVU Database, they used Long Distance Heterogeneous Face LDHF database as well. They used CNN (VGG architecture), to classify the gender and ethnicity of Asians and Caucasians under these environments, and the (VGG architecture), to classify the gender and ethnicity of Heterogeneous Face LDHF database as well. They used CNN for ethnicity classification. In this research, our focus is on Arab ethnicity. Arab refers to people who live in the Arab world, which consists of 22 countries from North Africa and Western Asia who have Arabic as their official language [18]. Our contribution is to provide an Arab face dataset that consists of three labels which are; Gulf Cooperation Council countries, Levant, and Egyptians, and perform supervised classification on it. And lastly, test some deep clustering methods on our dataset and a benchmark dataset. To the best of our knowledge, this is the first time that deep clustering is performed to solve the ethnicity classification problem.

The rest of the paper is divided into 5 sections. The second section contains literature reviewed. The methodology comes after that, in section four experimental results are discussed. Finally, the entire work is concluded and further developments are suggested.

II. RELATED WORK

The interest in CNN reached the field of ethnicity classification. In this section, we will summarize some studies that used CNN for ethnicity classification.

Heng et al. [10] hybridized a pre-trained CNN classifier (VGG-16), which was trained on ImageNet, output with image ranking engine and train Support Vector Machine (SVM) using the hybrid features. The approach evaluated on a new dataset which contains Bangladeshis, Chinese and Indian people faces. The result showed improvement when compared to Faster R-CNN and Wang’s method with an accuracy of 95.2%.

Narang and Bourlai [19] investigated the problems of distance, night time, and uncontrolled conditions of face images. They used NIR images taken from 30, 60, 90, and 120 meters at night and visible images at a distance of 1.5 meters from the Long Distance WVU Database, they used Long Distance Heterogeneous Face LDHF database as well. They used CNN (VGG architecture), to classify the gender and ethnicity of Asians and Caucasians under these environments, and the results 78.98% showed improvement from previous results.

Wang et al. [20] proposed a deep CNN classification model consists of three convolutional and pooling layers and end it with two fully-connected layers. The model was applied on several datasets in addition to two self-collected datasets. Some of the datasets were used only for testing to evaluate the classification of images that are not from the same trained dataset. The classification was done separately on white and black, Chinese and non-Chinese, finally on Han, Uyghurs, and Non-Chinese. The results were 100% vs 99.4%, 99.8 vs 99.9 and 99.4 vs 99.5 vs 99.9 respectively. They were compared to previous works and showed good improvements.

Anwar and Islam [14] used a CNN called VGG-face, which was pre-trained on a large face dataset of 2.6 million images to extract features then SVM with linear kernel is used as a classifier. It was trained on three classes: Asian, African-American, and Caucasian from ten different datasets they are Computer Vision Lab (CVL), Chicago Face Database (CFD), FERET, Multi-racial mega resolution (MR2) face database, UT Dallas face database, Psychological Image Collection at Stirling (PICS) Aberdeen, Japanese Female Facial Expression (JAFFE), CAS-PEAL-R1, Montreal Set of Facial Displays of Emotion Database (MSFDE) and Chinese University of Hong Kong face database (CUFC). Average classification accuracy over all databases is 98.28%, 99.66%, and 99.05% for Asian, African-American, and Caucasian respectively.

Masood et al. [3] also used pre-trained VGGNet, which is a 16-layer architecture in their proposed work. They attempted to classify the ethnicity of Mongolian, Caucasian, and the Negro using ANN and CNN. ANN was applied after calculating geometric features, calculating normalized forehead area, and extracting skin color. FERET database was used in the experiments. CNN achieved superior results with 98.6%, while ANN was good as compared to other works with 82.4%.

Srinivas et al. [16] presented a new dataset called WEAFD, which consists of constrained and unconstrained images of people from East Asian countries. Also, a CNN model that has three convolutional layers and several fully connected layers was applied on WEAFD to classify age, gender, and ethnicity. They presented two networks; one with full face images and the second with face divided into regions. Age and gender results were better in the first network, while ethnicity was better in the second one. However, the results of ethnicity (24.06%, 33.33%) and age (38.04%, 36.43%) were low in both networks compared to gender (88.02%, 84.70%). They explain that low results could be because of the lack of training data for age and ethnicity. Also, the quality of labels could be another reason as they mentioned that it is more likely for a human to make mistakes while labeling age and ethnicity data than gender.

In [21], the structure of Gudi’s CNN as follows: convolutional layer, local contrast normalization and a max-pooling layer, another two convolutional layers, after that a fully connected layer. For the preprocessing step they use Global Contrast Normalization on the VicarVision dataset. The classification accuracy is 92.24%. The model performed well on Caucasian and East Asian, which has a higher number of images (thousands). However, it didn’t perform well on the rest of the labels. The average precision for all labels was 61.52%. In this case, average precision was a better indicator to understand how well the model performed.

Chen et al. [22] used four different algorithms, which are k-Nearest Neighbor (kNN), SVM, Two-Layer Neural Network, and CNN to classify Korean, Japanese, and Chinese
TABLE 1. Summary of ethnicity classification studies in the literature.

| Author             | Year | Databases                                           | No. of images/subjects for each label used in experiments | Classifier | Accuracy   |
|--------------------|------|-----------------------------------------------------|----------------------------------------------------------|------------|------------|
| Naraqi and Boultif [19] | 2016 | WVU Database, LMIF database                         | 103 subjects, 100 subjects (Asian and Caucasian)         | CNN        | 78.98%     |
| Wang et al. [20]   | 2016 | MORPH-II, SAR Image, CASA+WebFace MORPH-II          | 10,530 white, 10,530 black                               | CNN        | 100% vs 99.4% |
|                    |      | IDPhotos (Han), IDPhotos (Uyghur)                    | 40,000 Chinese, 71,319 Chinese, 4,886 Chinese, 91,594 Non-Chinese, 38,817 Non-Chinese | CNN        | 99.8% vs 99.9% |
|                    |      |                                                    | 100,000 Han, 100,000 Uyghur, 91,594 Non-Chinese, 38,817 Non-Chinese | CNN        | 99.4% vs 99.5% vs 99.9% |
| Srinivas et al. [16] | 2017 | the new dataset called WEAFERD                      | 8706 Chinese, 1047 Japanese, 960 Korean, 795 Filipino, 210 Indonesian, 589 Vietnamese, 48 Burmese, 2 Thai | CNN: full image | 24.06% |
|                    |      |                                                    |                                                        | CNN: divided into regions | 33.33% |
| Hong et al. [10]   | 2018 | Self-collected dataset                              | 1000 images/1000 subjects Bangladesh, 1520 images/1022 subjects Chinese, 1078 images/1000 subjects Indian | pre-trained CNN hybridized with image ranking engine and the hybrid features classified by SVM | 95.2% |
| Anwar and Islam [14] | 2017 | CVL, CFD, FERET, MRJ, UT Dallas face database, PCs, Abderdeen JAFFE, CASA+WebFace MRJ, and OMEAC | 8291 Asian, 1635 Black, 4668 White | SVM (pre-trained CNN used for feature extraction) | An average accuracy of all 10 datasets: 97.72% ± 100% ± 99.01% |
| Chen et al. [22]   | 2016 | self-collected with 1380 image                      | 418 Chinese, 400 Japanese, 1066 Korean                  | kNN, SVM, Two-Layer NN, CNN | 89.2%, and 61.3% when tested on a different dataset |
| Guo [21]           | 2016 | VicarVision dataset                                 | 3266 Cezamen, 4556 East Asian, 161 African, 254 South Asian, 47 others | CNN | 92.24% (average precision 61.52%). |
| Musso et al. [5]   | 2018 | FERET database                                     | 1418 Asian, 132 Korean                                 | ANN, pre-trained CNN | 82.4%, 98.6% |

with and without identified gender. CNN architecture consists of two convolutional layers followed by one fully-connected layer and a dropout layer. The dataset used for experiments is self-collected. CNN was the best with 89.2% accuracy for 3 classes and 83.5% for 6 classes, which includes gender as well. Despite the good accuracy, CNN only predicts 61.3% accuracy on other new images, which means overfitting occurred.

Table 1 summarizes all the studies. However, the Arab world ethnicity has not been considered in the literature yet. One reason could be the lack of a dataset. Hence, a dataset containing images of people from different parts of the Arab world and label them accordingly is needed. In this study, we aim to create a dataset that contains Arab images with labels. However, labeling people according to their countries could be hard to classify due to the closeness of a lot of countries. Therefore, we have decided to classify according to a wider range (regions) such as Gulf Cooperation Council countries (GCC), which contains Kuwait, Oman, Qatar, Saudi Arabia and the United Arab Emirates. Another known region is Al-Sham (the Levant) which consists of four countries; Syria, Palestine, Jordan, Lebanon. The last label is Egypt, which has the largest population in Arab world with over 100 million inhabitants [23]. It has more population than both CGG and the Levant, therefore we decided to give it its own label. The map in Figure 1 illustrates the distribution of countries for each label.

All related studies mentioned previously are performed under supervised learning. Labeling data to perform supervised learning methods take a lot of time and effort. Therefore, there is a need to develop unsupervised learning methods to deal with unlabeled data. Clustering is one of the most popular unsupervised methods. It means grouping data...
that are more similar to each other to form a cluster [24]. Chen et al. [22], attempted to apply unsupervised learning to ethnicity classification by using k-means clustering. They did not report experiment details nor results of the experiment, they only mentioned that the algorithm failed to cluster labels due to background noises and similarity between labels.

To our knowledge, deep clustering has not been used to solve ethnicity classification problems before. Deep clustering becomes an interesting field to researchers after Deep Embedded Clustering (DEC) [25] was proposed by Min et al. [26]. In this study, deep clustering methods will be applied to the labeled datasets, to conclude the effectiveness of the methods.

Methods that will be used are DEC, the blueprint of a lot of deep clustering methods. Also, we will use Improved Deep Embedded Clustering (IDEC) [27], which is slightly different than DEC by keeping the decoder after the pre-training phase. It showed improvement in results from DEC [27]. The last method is Dynamic Autoencoder (DynAE) [28]. DynAE has a bit similar structure to the other two; however, its main contribution is the dynamic loss function. It achieved state of art results in image clustering of three benchmark datasets (MNIST-full, MNIST-test, and USPS) according to [29].

III. METHODOLOGY

This work aims to classify sub-ethnic groups of Arabs. To accomplish this, an Arab dataset is introduced by collecting images from the internet that belongs to specific subjects. Then some pre-processing and data cleaning tasks have been performed on the collected images. Once the dataset was ready, supervised and unsupervised models are applied to solve the ethnicity classification problem. CNN was chosen for the supervised learning due to its amazing performance in solving ethnicity classification problems in [2], [10], [14], [20]. In unsupervised learning, three deep clustering methods will be used. To evaluate models, our Arab dataset, as well as other datasets are used. Several matrices are used to report evaluation results. The summary of the methodology is shown in Figure 2.

### A. DATASETS

Our dataset provides labeled images from the Arab world. We decided to choose three labels as explained in the related work section. The process to create an Arab dataset follows:

1. Collect names of subjects. Subjects are public figures with accessible background information, such as Actors, Actresses, Singers, Football players, Social media influencers, Writers, Announcers, Journalists, Businessmen / Businesswomen, Ministers, Producers, Filmmakers, Artists, etc.

2. Check their nationalities, origins if available. They must belong to only one of the labels and no others:
   - Saudi, Kuwaiti, Qatari, Emirati, Omani, Bahraini are called: GCC.
   - Syrian, Palestinian, Jordanian, Lebanese are called: The Levant
   - Egyptian

3. Download images from google search using a modified python script from [30] accordingly. We download 10 images for each subject.

4. Use a face detector to detect faces in images of each subject, mentioned in detail in preprocessing section.

5. Cleaning: Remove unrelated images to subjects (images of objects or other people), duplicated images, and images that were mistaking by the detector as faces.

Table 2 illustrate the number of subjects and images for each label in the Arab dataset. A subject is a unique person; that means, the dataset may have more than one image for one subject. As shown in Figure 3, the dataset is unbalanced. GCC has 70% of the subjects. In addition to the Arab dataset, A Private Arab dataset was collected from non-public figures, will be used to evaluate models. The private dataset consists of 88 Egyptian subjects, 104 GCC subjects, and 91 from the Levant.

| Label   | No. of Subjects | No. of Images | No. of subjects per Gender (male, female) |
|---------|----------------|---------------|------------------------------------------|
| GCC     | 784            | 5598          | (718, 66)                                |
| Levant  | 195            | 1665          | (141, 54)                                |
| Egyptians | 169         | 1555          | (100, 69)                                |
| Total   | 1148           | 8818          | (959, 189)                               |

![FIGURE 2. Methodology summary.](image)

![FIGURE 3. The chart illustrates the dataset's labels and percentage of gender for each label.](image)
Racial Faces in-the-Wild (RFW) [31], [32], [33] was collected from MS-Celeb-1M, and have four labels in which, each label contains 10K images of 3K subjects [31]. RFW is similar to the Arab dataset in terms of data source (internet). Moreover, it identifies subjects individually. Therefore, it is suitable to be combined with the Arab dataset to classify four labels (Arab, Asian, Black, White) without concern of subjects overlapping between train and test sets. Figure 4 shows samples from the Arab dataset and RFW.

FIGURE 4. Samples of all datasets labels.

BUPT-Transferface has 50K images of African, Asian and Indian, also over 460K images and 10K subjects of White.

FERET dataset [33], [34] is a well-known benchmark dataset of facial images used to report and compare results of different methods. It contains high-quality images for individual subjects with different poses, expressions, and lighting. Furthermore, it provides gender, age, and ethnicity information about subjects.

Lastly UTK dataset [35]. It contains images collected from the internet and provides their age, gender, and ethnicity. However, the dataset does not provide information about the individual identity of subjects. Therefore, we cannot be sure if overlapping occurs between train and test sets if it is used to train a classification model.

B. PREPROCESSING

Dlib’s pre-trained face detector based on a modification to the standard Histogram of Oriented Gradients + Linear SVM method for object detection from [36] was used to detect faces from all images. Then, the detected face is cropped and resized to 224 x 224. The size was chosen based on pre-trained CNN that used the same size. The steps are illustrated in Figure 5.

FIGURE 5. An illustration of pre-processing steps.

After that, cleaning images is implemented. Images such as duplicated, unrelated results to the subject (due to google search error or other faces in the same image as the subject) and detected errors will be removed.

C. DATA AUGMENTATION

Data augmentation (DA) is a way to reduce overfitting [37] by applying certain methods to images during the training process. In our experiments, Different data augmentation method sets will be used in different experiments to find suitable for Arab dataset. The methods that will be used in our experiments: flip horizontally and/or vertically (FL), multiply all pixels by random values to make them brighter or darker(ML), increase or decrease hue and saturation by random values (HS), rescaling, blur images, adjust image contrast, dropout (set some pixels to zero) and convert to grayscale.

D. CLASSIFICATION MODEL

In this section, we represent classification models that will be used to solve the ethnicity classification problem. Two types of learning are going to be used separately to solve the problem, supervised learning, and unsupervised learning.

1) SUPERVISED LEARNING

After pre-processing, a CNN model will be trained on the Arab dataset. Convolutional layers in CNN work as feature extractors, so there is no need for a separate feature extraction step [38]. The model which we use is a pre-trained CNN model. Usually, pre-trained models are trained on millions of images then used to train on small datasets (thousands in our case) [39], [40]. It was proven that pre-trained models improve results and outperform newly trained CNN from scratch [39], [40].

The model architecture to be used is ResNet-50 layers that were created by He et al. [41]. They were motivated by the degradation problem, which can be explained as; when the
network depth increased, accuracy gets saturated and then degrades rapidly after the saturation region. This degradation is not caused by overfitting, and adding more layers to a deep model that leads to higher training error was unexpected since theoretically, the network was supposed to perform better while going deeper [41].

Shortcut connections between blocks differentiate ResNet from other models [41]. Two types of shortcuts are used in ResNet-50 layers: Identity shortcuts are used when input/output have the same dimensions, while projection shortcuts are used to match dimensions [41]. Figure 6 shows more details about ResNet-50 layers architecture. Downsampling is performed between blocks with a stride of 2 [41].

ResNet50 was trained by Cao et al. [42] on the VGGface2 dataset, which contains 3,31 million images of 9131 subjects.

We will use the pre-trained ResNet50 model. However, the last layer of the model is going to be replaced with a new fully-connected layer that has an output of three classes. Then we start training with categorical cross-entropy loss function ($L_n$) for the $n^{th}$ training sample which is given by the equation:

$$L_n = - \sum_{i=1}^{3} x_i \log_2 (s_i)$$  \hspace{1cm} (1)

where $x_i$ is the truth label (0 or 1) of class $i$ and $s_i$ (0 to 1) is the probability of an object classified as a member of class $i$.

Categorical cross-entropy loss function or sometimes called softmax loss function was used to train the ResNet50 pre-trained model [42]. It is a common and popular choice for classification problems (multiclass classification) [43]. The model aims to minimize the loss function to improve the performance.

We tune the model with different hyperparameters. The total number of experiments is over 60. The hyperparameters were:

- Learning rate (LR): constant LR 0.01, 0.001, 0.0001, and automatic LR that increase at certain values of the epoch.
- Optimizer: SGD and Adam
- Freezing layers (blocks): there are 4 blocks in ResNet-50 explained in Figure 6. Switch freezing between blocks in different experiments and sometimes made all the network trainable. Freezing layers in pre-trained CNN means that the layer will not learn from the training process and only use what it learned from training before.
- Data augmentation: different sets of methods or not use at all in some experiments.

The number of epochs is 30 and the batch size is 64 for all experiments. We test all models on a private dataset and represent the top five accuracies in the results and discussion section.

2) DEEP CLUSTERING MODELS (UNSUPERVISED LEARNING)

The general idea of deep clustering consists of two stages; pre-training an autoencoder, which allows the network to learn features that are used to initialize the cluster centers [44], and fine-tuning, where clustering and feature learning are jointly performed [44]. The methods we will use for clustering are as mentioned before; DEC [25], IDEC [27], and DynAE [28].

The first two methods were implemented with a convolutional network that was introduced in [44]. The difference between DEC and IDEC is that DEC discards the decoder after pre-training and fine-tune the encoder with clustering loss, while IDEC keeps the decoder. Figures 7 and 8 illustrate each method architecture.
In DynAE [28], they overcame the trade-off between clustering and reconstruction by using dynamic loss function. Figure 9 shows the general architecture of DynAE.

For all methods, the number of clusters is a priori knowledge given before the start of clustering.

### E. EVALUATION METRICS

For supervised model (classification) evaluation we evaluate it using accuracy metric:

\[
\text{accuracy} = \frac{M}{N}
\]

Equation 2, where \( M \) is the number of correct samples and \( N \) is the number of all samples. Besides, we use two metrics that are widely used to evaluate deep clustering methods [26]. The first one is unsupervised clustering accuracy (ACC):

\[
\text{ACC} = \max_m \sum_{i=1}^{n} \frac{1\{k_i = m(r_i)\}}{n}
\]

Equation 3, where \( k_i \) is the ground-truth label, \( r_i \) is the clustering algorithm result, and \( m \) ranges over all one-to-one mappings that are possible between clusters and true labels. The metric takes cluster results from the clustering algorithm and a ground-truth label and then discovers the best matching between them, which can be computed by the Hungarian algorithm [45]. The second metric is Normalized Mutual Information (NMI):

\[
\text{NMI} (l,c) = \frac{M (k,r)}{\frac{1}{2} [H (k) + H (r)]}
\]

Equation 4, where \( M \) is the mutual information metric, \( H \) is entropy, \( k \) is the ground-truth label and \( r \) is the clustering result. Mutual information measures the mutual dependence of two groups, which are ground-truth and clustering results. NMI a normalized version of it and permutations does not affect its results [46]. When NMI equal to 0, it means the two are independent. And if it equal to 1, that means the two are identical.

The last metric is the Adjusted Rand Index (ARI), which is the chance-corrected version of the Rand Index (RI). RI focus on the pairwise agreement. For each possible pair, it evaluates how similar the two clusters treat them [47]. RI is calculated by:

\[
\text{RI} = \frac{a + b}{a + b + c + d}
\]

where \( a \) and \( b \) are pairs that both ground truth and clustering results agree. \( c \) and \( d \) represent the disagreement, on one side they are put together, where they are separated on the other [47]. And ARI is calculated using Equation 5 by:

\[
\text{ARI} = \frac{\text{RI} - E(\text{RI})}{\text{RI}_{\text{max}} - E(\text{RI})}
\]

### IV. RESULTS AND DISCUSSION

Experiments were done using Google Colab and Deep Learning AMI (Ubuntu 18.04) Version 28.1 and g3s.xlarge from Amazon Web Services (AWS).

#### A. CLASSIFICATION RESULTS

Arab dataset subjects were divided into 80% training set and 20% validation set without subjects overlapping, i.e. images of subjects used in train sets are not used in the test set. 60 Experiments were done on the Arab dataset using the ResNet50 pre-trained model to tune hyperparameters, please refer to the supervised learning section for more details about hyperparameters. After that, all models were evaluated on a different dataset to determine the best model. Accuracy was calculated according to equation 1. Table 4 represents only top-5 accuracy results that were obtained by testing on the private dataset, last two have equal results. As we can see in Table 3, all top-5 results used SGD as the optimizer. 5 out of 6 used data augmentation but different sets of methods, and 3 had no frozen layer, while one had the first block frozen, another had the first two blocks and the last had the first three blocks frozen. 5 out of 6 used the same learning rate method, which starts from 0.01 and exponentially increasing by a factor of 0.1 every 5 epochs. The last model used a learning rate that starts from 0.01, increasing by a factor...
of 0.1 at epoch 5, epoch 10, and epoch 20. The best model was optimized with SGD, has its first block froze, has FL, HS, and rescaling as DA, and LR 0.01 increase exponentially by 0.1 at every 5 epochs.

Results of all models tested on Arab dataset and private dataset are represented in Table 4. The accuracies of testing in the Arab dataset were between 0.72 and 0.76. However, it drops to 0.56 when testing on a different dataset. The best accuracy was 0.5697 by model-1 and comes close to it 0.5606 by model-2.

We will look deeper into model-1 prediction results (exp1). Confusion matrix of model 1 evaluated on a private Arab dataset shown in Figure 10. As we can see 75% of GCC were predicted correctly. Levant and Egyptian labels have 43% and 48% of images predicted correctly, respectively. Another thing we noticed, over 30% of Levant and Egyptians were predicted as GCC. We were concern if GCC dominating the dataset by 70% had caused the model to be biased toward GCC.

To solve this concern, we did another experiment (exp2) with a modified Arab dataset (Arab balanced dataset). The modified dataset has a similar number of subjects and images for each label. Hyperparameters used in this experiment are the same as model-1. The accuracy result on the Arab balanced dataset was 0.5349. When evaluated on a private Arab dataset, the accuracy was 0.5212. In the confusion matrix Figure 11, GCC again had the highest correct predictions by 65%, lower than the model-1 result by 10%. Levant and Egyptian have 42% and 45% respectively. This experiment shows that the model can identify GCC better than the others even with a similar number of subjects/images. 31% of Levant were predicted as Egyptian, which is greater by 9% than model-1 results. We can see in both exp 1 and 2; models struggle in classification. Especially in classifying Levant and Egyptian.

The third experiment (exp 3) had four labels, three labels from the RFW dataset (Black, Asian and White) and one Arab label from Arab dataset labels combined. The dataset is divided into 80% of subjects for the train set and 20% test set. The hyperparameters were the same as model-1. Testing on the same dataset results in high accuracy of 0.9663. However, we had two tests on two different datasets (BUPT-TRANSFERFACE, UTK) combined with the Arab private dataset as one label. The two tests achieved 0.9675 and 0.6995 respectively. Figures 12 and 13 show the confusion matrix of both tests. 88% of Arab label was predicted correctly in the two tests, there is 9% of Arabs were wrongly predicted as White. As for other labels, there was a wide
gap between Black and White results in test-1 (BUPT-TRANSFERFACE dataset) and test-2 (UTK dataset). Test-1 has almost all labels predicted correctly while in test-2, 30% of Black and 36% of White were predicted as Arab.

Through these experiments, we noticed that the model can successfully identify Arabs up to 88% when put with others. Even though around 30% of Black and White were mistaken as Arab in test-2. However, the model does not give good classification performance if we classify Arab labels together, it probably because the similarity between Arab classes is higher.

B. DEEP CLUSTERING RESULTS

Experiments were done using DEC, IDEC, and DynAE. The size of images used is 60 × 60 for all datasets. The parameters used are the same as the implementation in their respective papers for all three methods. Adam was the optimizer for DEC and IDEC and pre-training phase in DynAE while SGD was used for the clustering phase. In DEC and IDEC CNN was used while in DynAE it was a fully connected network. Three metrics were used to evaluate experiments: ACC (Equation 3) measures how many individuals clustered correctly. NMI (Equation 4) focuses on partitioning and distribution of ground truth and clusters. ARI (Equation 6) considers counting all pairs that are assigned to the same or different clusters in predicted and ground truth. We did some experiments with balanced and unbalanced datasets because according to [48], the cluster size could affect the results.

Table 5 shows ACC, NMI, and ARI for each method with different datasets. All experiments have three labels except the last two experiments, one had four labels which is a combination of RFW dataset (Black, White Asian) and Arab label from Arab dataset. And the last experiment had five classes, the Indian class from RFW is added. The best ACC was 0.5955 in FERET by DynAE. In Figure 14, most images are clustered in White, when we look at the statistics of the FERET dataset (Asian: 952, Black: 257, White: 2883), the White class consists of 70% of total images which affect the ACC.

Worst ACC was 0.3206 in RFW (4 labels) + Arab by DynAE too. Figure 15 shows that the correct prediction of all labels is low, with Black having ACC of 44% being the highest, while the rest were from 36% to 22%.

NMI and ARI consider the unmatched parts of clusters, the distribution of images, and pairing [48]. Their best results were 0.2714 and 0.2543 respectively, by DEC applied to RFW (3 labels) + Arab, while there are several lows, most notable in all experiments on the FERET dataset. DynAE with FERET dataset achieved NMI of 0.0012 and ARI of −0.0008. Figure 14 shows that images of each label were distributed throughout the clusters by the same percentage, which means there is no specific relation between cluster items.

The second one is Figure 16 which an NMI of 0.0740 and an ARI of 0.0565. Around half of GCC is in one cluster while Egyptian and Levant were similarly distributed. Figure 17 has better results than the previous ones with an NMI of 0.1902 and ARI of 0.1938. Black is dominating one cluster. While White and Asian are similarly distributed in
TABLE 5. ACC, NMI, and ARI results of deep clustering methods for each dataset.

| Method | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| DynAE  | 0.453 | 0.961 | 0.696 | 0.471 | 0.691 | 0.610 | 0.414 | 0.524 | 0.574 | 0.372 | 0.096 | 0.057 |
| IDEC   | 0.408 | 0.001 | 0.488 | 0.408 | 0.001 | 0.001 | 0.405 | 0.050 | 0.012 | 0.516 | 0.054 | 0.001 |
| DEC    | 0.619 | 0.001 | 0.698 | 0.619 | 0.001 | 0.001 | 0.530 | 0.054 | 0.001 | 0.661 | 0.056 | 0.001 |

FIGURE 15. Normalized confusion matrix of DynAE clusters on RFW(4 labels) + Arab dataset.

FIGURE 16. Normalized confusion matrix of DEC clusters on Arab dataset.

the other two clusters. However, the correct clusters here are higher than the previous one.

Figure 18 has the best results for NMI and ARI. Arab and Black are both dominating one cluster for each. Asian and White are similarly distributed, with half of Asians been clustered correctly.

Experiments on the FERET dataset and the balanced version of it have similar results. Even though the ACC is between 40% and 59%, NMI and ARI are between 0 and 0.02. These results tell that partitions are random, that ground truth and clusters are independent, and the model is uncertain about the clusters.

Experiments on the Arab dataset and the balanced version results are also similar. ACC is between 37% and 47%. NMI and ARI are slightly better here, NMI achieved from 0.03 to 0.07 while ARI from 0.01 to 0.08. Only one experiment, DynAE on Arab balanced dataset, performed worse than the rest. It had NMI and ARI near zero. The rest of the results showed a small improvement. However, they are still low. Ground-truth and clusters are nearly independent and not similar.

Another experiment was done on the RFW dataset with three labels (Black, White, Asian). Results of NMI and ARI are much better than previous experiments. THE best NMI was 0.2071 by DynAE, while the lowest was 0.1897 by DEC. and best ARI was 0.1938 by IDEC, while worst 0.1854 by DynAE. ACC was near 53% for all methods.

The last two experiments were done on RFW + Arab, one has four labels: Black, White and Asian from RFW and Arab, while the other has Indian as an addition. DynAE performed the worst in terms of all metrics in both experiments. DEC and IDEC for the first experiment had ACC of 52% for both, NMI of 0.2714 and 0.2682 which is the best of all experiments, and ARI of 0.2543 and 0.2459 respectively. As for the last experiment, DEC and IDEC had ACC of 44% and 42%, NMI of 0.2451 and 0.2366, ARI of 0.2024 and 0.1818 respectively.
We can see a similarity in the confusion matrix in Figures 18 and 19, 78% and 71% of Black were grouped respectively. Almost half of Asians and White were grouped in one cluster in both experiments, while Arabs were divided into two clusters in Figure 18, one of the clusters has also around 20% of Asians and White as well. Then in Figure 19 Arab and Indian are separated into three clusters.

Based on the discussion above no consistency has been seen in the performance of any of the methods considered for the experiment as far as deep clustering is concern. Moreover, the lower score of NMI and ARI confirms low intra and high inter-cluster similarity as well. So it can be said that facial features are much similar across the borderline of different ethnic groups. And this can be one of the reasons for the poorer performance of clustering. In support of this conclusion, it is noticeable that all the models across supervised and unsupervised provide the best accuracy with RFW dataset, the NMI and ARI score is higher for this dataset as well.

The conclusion can be drawn based on experiments. Table 6 shows a comparative accuracy of supervised and unsupervised learning models with two datasets Arab and RFW (3 labels) + Arab (1 label). Supervised learning witnesses better results whereas unsupervised methods that were used could not match the performance level of the supervised learning model yet.

### V. CONCLUSION AND FUTURE WORK

In this study, we investigate the possibility of a CNN model to classify sub-ethnic groups of Arabs. First, we create an Arab dataset with three labels chosen according to countries’ distribution into regions. Then a pre-train ResNet50 model was used to classify the Arab dataset. Over 60 experiments were done to fine-tune hyperparameters, explained in more detail in the supervised learning section. After that, the models have evaluated on a different dataset and the best accuracy result was 0.5697. Another experiment was done after balancing the number of subjects in each class. The accuracy after evaluation in a different dataset was 0.5212. From both experiments, the model is struggling to identify between labels, which can be due to the strong similarity between them.

A third experiment was done to classify Arabs as a whole and the other three ethnicities (Black, White, Asian) from the RFW dataset. The model was evaluated two times with two datasets (BUPT-TRANSFERFACE, UTK) each combined with our private Arab dataset. The results were 0.9675 and 0.6995 respectively.

For deep clustering experiments. ACC results were between 59% and 32%. However, NMI and ARI vary according to each dataset and method. They were Zero in FERET dataset experiments. And the best was experiments on a combination of three labels from RFW and one label Arab. The best was NMI of 0.2714 and ARI of 0.2543 by DEC. In the future, we would like to investigate more methods regarding ethnicity classification.

This study has some limitations; first, our Arab dataset does not cover all countries of the Arab world. The limited time and lack of knowledge about public figures in other countries made it hard to collect a proper amount of subjects. Moreover, the Arab dataset is unbalanced with GCC countries made it hard to collect a proper amount of subjects. Regarding age, the Arab dataset does not have people under 17, we are not sure if the same results can be applied to them. Also, we resized images to a small size (60 × 60) while performing Deep clustering methods, due to limited memory. We are concerned about how the size could affect the quality of the performances.

**LINK OF THE DATASET**

https://www.dropbox.com/sh/j4kjs9z9qkqekwad/AABixLKwaME-3YiCfqKOdmlSa?dl=0

**ACKNOWLEDGMENT**

 Portions of the research in this article use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

**REFERENCES**

1. C. Yu, Y. Fang, and Y. Li, “Multi-task learning for face ethnicity and gender recognition,” in Proc. Chin. Conf. Biometric Recognit., 2014, pp. 136–144.
2. H. Ding, D. Huang, Y. Wang, and L. Chen, “Facial ethnicity classification based on boosted local texture and shape descriptions,” in Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG), Apr. 2013, pp. 1–6.
3. S. Masood, S. Gupta, A. Wajid, S. Gupta, and M. Ahmed, “Prediction of human ethnicity from facial images using neural networks,” in Advances in Intelligent Systems and Computing. Singapore: Springer, 2018, pp. 217–226.
4. A. K. Jain, S. C. Dass, and K. Nandakumar, “Soft biometric traits for personal recognition systems,” in Biometric Authentication. Berlin, Germany: Springer, 2004, pp. 731–738.
5. K. Veropoulos, G. Bebis, and M. Webster, “Investigating the impact of face categorization on recognition performance,” in Advances in Visual Computing. Berlin, Germany: Springer, 2005, pp. 207–218.
[6] N. Kumar, P. Belhumeur, and S. Nayar, “FaceTracer: A search engine for large collections of faces with images,” in Computer Vision—ECCV. Berlin, Germany: Springer, 2008, pp. 340–353.

[7] Y. Hu, Y. Fu, U. Taqi, and T. S. Huang, “Subjective experiments on gender and ethnicity recognition from different face representations,” in Advances in Multimedia Modelling. Berlin, Germany: Springer, 2010, pp. 66–75.

[8] A. Dantcheva, P. Elia, and A. Ross, “What else does your biometric data reveal? A survey on soft biometrics,” IEEE Trans. Inf. Forensics Security, vol. 11, no. 3, pp. 441–467, Mar. 2016.

[9] K. Niinuma, U. Park, and A. K. Jain, “Soft biometric traits for continuous user authentication,” IEEE Trans. Inf. Forensics Security, vol. 5, no. 4, pp. 771–780, Dec. 2010.

[10] Z. Heng, M. Dipu, and K.-H. Yap, “Hybrid supervised deep learning for ethnicity classification using face images,” in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), May 2018, pp. 1–5.

[11] E. J. Jewell and F. R. Abate, The New Oxford American Dictionary. London, U.K.: Oxford Univ. Press, 2001.

[12] D. J. Da Silva Santos, D. B. Palomares, D. Normando, C. Cardoso, and A. Quintão, “Race versus ethnicity: Differing for better application,” Den- tal Press J. Orthodontics, vol. 15, no. 3, pp. 4–121, 2010.

[13] G. Muhammad, M. Hassin, F. Alenezy, G. Bebis, A. M. Mirza, and H. Aboualsamh, “Race classification from face images using local descriptors,” Int. J. Artif. Intell. Tools, vol. 21, no. 5, Oct. 2012, Art. no. 1250019.

[14] I. Anwar and N. Ul Islam, “Learned features are better for ethnicity clas- sification,” Cybern. Inf. Technol., vol. 17, no. 3, pp. 152–164, Sep. 2017.

[15] H. H. K. Tin and M. M. Sein, “Race identification from face images,” Proc. Int. Conf. Adv. Comput. Eng. (ACE), 2011, pp. 1–4.

[16] S. Srinivas, H. Atwal, D. C. Rose, G. Mahalingam, K. Ricanek, and D. S. Bolme, “Age, gender, and fine-grained ethnicity prediction using convolutional neural networks for the east asian face dataset,” in Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2017, pp. 953–960.

[17] C. Wang, Q. Zhang, X. Duan, and J. Gan, “Multi-ethnic Chinese facial characterisation and analysis,” Multimedia Tools Appl., vol. 77, no. 23, pp. 30311–30329, Dec. 2018.

[18] N. Mrabah, N. M. Khan, R. Ksantini, and Z. Lachiri, “Deep clustering with local structure preservation,” in Advances in Multimedia Modelling. Berlin, Germany: Springer, 2017, pp. 1753–1759.

[19] N. Grira, M. Crucianu, and N. Boujemaa, “Unsupervised and semi-supervised clustering: A brief survey,” in A Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (6th Framework Programme). Citeseer, 2004, pp. 1–12.

[20] A. Gudi, “Recognizing semantic features in faces using deep learning,” 2016, arXiv:1512.00743. [Online]. Available: https://arxiv.org/abs/1512.00743

[21] H. Chen, Y. Deng, and S. Zhang, “Where am i from?-east asian ethnicity classification from facial recognition,” Project Study Stanford Univ., San Francisco, CA, USA, Tech. Rep., 2016.

[22] H. W. Kuhn, “The hungarian method for the assignment problem,” Nav. Res. Logistics Quart., vol. 2, nos. 1–2, pp. 53–58, Mar. 1955.

[23] L. A. Rowan, “Clustering—Scikit-Learn 0.24.1 Documentation.” Accessed: Mar. 26, 2021. [Online]. Available: https://scikit-learn.org

[24] N. A. AL-Humaidan, M. Prince: Classification of Arab Ethnicity Based on Face Image Using Deep Learning Approach. N. A. AL-HUMAIDAN received the B.S. degree in computer science from Qassim University, Saudi Arabia, in 2016, where she is currently pursuing the master’s degree with the Department of Computer Science.

[25] J. Dobies. Simple Image Download. Accessed: Mar. 26, 2021. [Online]. Available: https://github.com/RiddlerQ/simple_image_download

[26] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, “Racial faces in the wild: Reducing racial bias by information maximization adaptation network,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 692–702.

[27] M. Wang and W. Deng, “Mitigating bias in face recognition using skews-aware reinforcement learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9322–9331.

[28] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, “The FERET database and evaluation procedure for face-recognition algorithms,” Image Vis. Comput., vol. 16, no. 5, pp. 295–306, Apr. 1998.

[29] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET evaluation methodology for face-recognition algorithms,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 10, pp. 1090–1104, 2000.

[30] Github.io. UTKFace. Accessed: Mar. 26, 2021. [Online]. Available: https://susannq.github.io/UTKFace/

[31] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection to cite this version: HAL Id?: inria-00548512 histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 886–893.

[32] A. Krichhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Proc. Syst. (NIPS), 2012, pp. 1097–1105.

[33] W. Rawat and Z. Wang, “Deep convolutional neural networks for image classification: A comprehensive review,” Neural Comput., vol. 29, no. 9, pp. 2352–2449, Sep. 2017.

[34] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[35] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” in Proc. Brit. Mach. Vis. Conf., 2014, pp. 1–12.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[37] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2018, pp. 67–74.

[38] B. Barz and J. Denzler, “Deep learning on small datasets without pre- training using cosine loss,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020, pp. 1360–1369.

[39] X. Guo, E. Zhu, X. Liu, and J. Yin, “Deep embedded clustering with data augmentation,” in Proc. 10th Asian Conf. Mach. Learn., vol. 95, 2018, pp. 550–565.

[40] H. K. Kuhn, “The hungarian method for the assignment problem,” Nav. Res. Logistics Quart., vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[41] Scikit-learn.org. 2.3. Clustering—Scikit-Learn 0.24.1 Documentation. Accessed: Mar. 26, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html

[42] V. Labatut, “Generalised measures for the evaluation of community detec- tion methods,” Int. J. Social Netw. Mining, vol. 2, no. 1, pp. 44–65, 2015.

[43] M. Rezaei and P. Franti, “Set matching measures for external cluster validity,” IEEE Trans. Knowl. Data Eng., vol. 28, no. 8, pp. 2173–2186, Aug. 2016.

NORAH A. AL-HUMAIDAN