IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON BEHAVIORAL BIOMETRICS FOR EHEALTH AND WELL-BEING

Received September 16, 2020, accepted September 28, 2020, date of publication October 5, 2020, date of current version October 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3028694

Facial Anthropometric Measurements and Photographs – An Interdisciplinary Study

JASBIR DHALIWAL[1], JOHN WAGNER[2], (Member, IEEE), SHU LING LEONG[1], AND CHERN HONG LIM[1], (Member, IEEE)

[1]School of Information Technology, Monash University Malaysia, Subang Jaya 47500, Malaysia
[2]IBM Research-Australia, Southbank, VIC 3006, Australia

Corresponding author: Jasbir Dhaliwal (jasbirkaur.dhaliwal@monash.edu)

ABSTRACT In recent years, automatic facial analysis has attracted much interest among computer science researchers in the healthcare and computer vision fields studying facial anthropometric measurements using photographs. However, to date, there have been no healthcare or computer vision publications that use standardized photographs to differentiate features between sub-ethnic groups by leveraging the power of machine learning on two-dimensional computer vision benchmark data sets (2D CVBDs). Thus, the present work is an interdisciplinary study at the interface of healthcare and computer vision fields that attempts to fill this literature gap where we explore the use of machine learning on 2,789 photographs from eleven 2D CVBDs to identify k top discriminative features in major and sub-ethnic groups. These features are ranked based on information gain values and p-values. We also provide a comprehensive analysis of using information-gain-based and p-value-based features. Our machine learning model achieves an accuracy of 96-99%, and our findings reveal that information-gain-based features have the upper hand over p-value-based features. The top three information-gain-based features in sub-ethnic groups are: *dn* (distance from the tip of the nose to the center of the mouth), *hf* (face height) and *wn* (nose width), while the top three information-gain-based features in major ethnic groups are: *de* (distance between the inner corners of the eyelids), *hf* and *dn*. These results are then compared to the results obtained using standard deep learning techniques such as OxfordNet (VGG16), Residual Networks (ResNet50), and Inception-V3, where accuracy of 90-94% was seen. We hope that these findings will lead to future collaboration between computer vision and healthcare researchers studying facial anthropometric measurement studies.

INDEX TERMS Facial anthropometric measurements, computer vision benchmark data sets, machine learning, healthcare.

I. INTRODUCTION

In recent years, automatic facial analysis, which includes facial recognition and demographic classification (sex, age, and ethnicity estimation), has attracted much interest in the healthcare and computer vision fields and motivated computer science researchers to study facial anthropometric measurements. In this research, facial landmarks are annotated on two-dimensional objects, e.g., standardized photographs (2D face images), 3D representations of human faces, or even the skin of living humans. These landmarks are used when calculating measurements using traditional feature-based approaches (linear and vertical) that are handcrafted by researchers and have successfully been used in ethnicity and sex estimation studies. Linear measurement studies [1]–[4] compute horizontal and/or vertical distances between identified anthropometric landmarks using Euclidean distances. Whereas, angular measurement studies [5]–[7] generate facial angles from anthropometric landmarks instead. A recent trend is to combine both the measurements and was used in these two studies [8], [9]. Table 1 summarizes the data sets and the statistical tests of the discussed anthropometrical measurements-based classification studies. It is worthwhile to mention that in contrast to the anthropometrical measurements-based classification scheme, appearance-based classification schemes that utilize machine learning [10]–[13] or even deep learning paradigm [14]–[16] exists, and have also obtained success in ethnicity, sex and

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

TABLE 1. The facial anthropometric measurements to distinguish populations, and the tests to determine the statistical differences.

Literature	Data	Measurements		Comments
		Linear	Angular	
[17](2008)	110 Croatians (52 males and 58 females), aged between 23-28 years old with Class I occlusal relationship.	✓		Student t-test
[7](2008)	110 Croatians (52 males and 58 females), aged between 23 - 28 years old with a dental Class I occlusal relationship and harmonious soft tissue profile.		✓	Student t-test, Dahlberg's formula
[1](2009)	430 Turkish adults (149 males and 281 females), aged between 18-24 years old with no noticeable facial disfigurement and no history facial surgery in any of the subjects.	✓		t-test and Welch's t-test
[8] (2012)	120 Nigerians males of Ibo ethnicity within 18 and 28 years with no facial deformities and congenital facial abnormalities.	✓	✓	
[9] (2012)	107 Persian females between the age of 18 and 40 years old with no history of trauma, surgery or craniofacial syndromes.	✓	✓	unpaired t-test
[6](2013)	100 Bangladeshi Garos (50 males and 50 females), aged between 25 - 45 years old with without any history of acquired or genetic craniofacial anomalies.		✓	Student t-test.
[5] (2013)	477 Nigerians within 18-35 years range with no facial asymmetry, congenital abnormalities, facial fractures or maxillofacial surgeries. 276 (184 males and 92 females) of Ibo etnicity and 201 (106 males and 95 females) of Yoruba ethnicity.		✓	Student t-test
[2] (2018)	1349 Brazilians (660 males and 689 females), aged 30 around years, and classified according to regions of birth, as follows: south, southeast, midwest, northeast, and north from the regions.	✓		Student t-test, discriminant analysis and decision tree.

age estimation studies. Deep learning offers a radical alternative to traditional feature-based approaches as it performs automatic feature extraction on the facial images to obtain learned features.

The purpose of doing anthropometrical measurements-based classification is to determine reference ranges of the average soft tissue profile of human faces. A reference range is a set of values that includes upper and lower limits based on a group of healthy individuals. For example, the researchers in [18] found that individuals with major thalassemia have wider heads and faces by comparing their findings to the reference ranges of healthy individuals. Moreover, these values are crucial in healthcare applications such as the measurements of dental arch dimensions [19], diagnosis of craniofacial anomalies [20], setting standards for the planning of facial construction surgeries [21], and establishment of aging patterns [22].

However, it is well established in the healthcare field that a single reference range value cannot be applied to different ethnic and sex groups [4]. Nevertheless, few studies to date have used photography to find the distinguishing features for different ethnic groups, and therefore is the focus of the present study. Most healthcare literature [2], [4], [8], [23], [24] uses anthropometrical measurements-based classification scheme, while most computer vision literature [10]–[12] uses appearance-based classification schemes instead. Although both healthcare and computer vision literature use different classification schemes, they focus on distinguishing features in one or two major ethnic groups, with almost no attention given to sub-ethnic groups, such as Chinese, Indian, and Malay in the Asian category. It is worthwhile to mention that reference [25] is a recent survey that provides a detailed review of the state-of-the-art advances in face-race perception, principles, algorithms, and applications.

Moreover, finding discriminative features in multiple major ethnic and sub-ethnic groups has already been carried out on human faces by practitioners, for example, in [26]. This study was conducted on 1,470 subjects drawn from five regions of the world. We conclude that, to the best of our knowledge, there have been no reports in the healthcare and computer vision fields that use standardized photographs to differentiate features between sub-ethnic groups by leveraging the power of machine learning on 2D CVBDs at a large scale. Thus, this paper attempts to fill this literature gap by exploring the use of machine learning on 2,789 photographs from eleven 2D CVBDs [27]–[37] to identify k top discriminative features based on ranked information gain values and p-values. The contributions of this paper are summarized below:

- Provides the first evidence that two-dimensional benchmark data sets developed in the computer vision field can be used in facial anthropometric measurement studies of the healthcare field.
- Provides a comprehensive analysis of using information-gain-based and p-value-based features to find k top discriminative features in major and sub-ethnic groups.
- Proposes the top three features that may be useful in differentiating populations in major and sub-ethnic groups. These results are then compared to the results obtained using standard deep learning techniques such as VGG16 [38], ResNet50 [39], and Inception-V3 [40].

The rest of this paper is organized as follows. Section II provides an overview of the 2D CVBDs that we were able to access, as well as the dilemma faced while categorizing

TABLE 2. Summary of the characteristics of the 2D CVBDs that we used in this study.

Face Database	Full Name	Description
CaNAFF [27]	Caucasian and North African French Faces	Contains color images of 147 young males for three eye gazes, i.e., right, frontal and left, in France. https://osf.io/274ry/
CAS-PEAL [28]	CAS - Pose, Expression, Accessory, and Lighting	Contains grayscale images of 1,040 individuals (595 males and 445 females), applied in different sources of variations where five facial expressions, six accessories, and 15 lighting changes were captured, while focusing on frontal and nonfrontal eye gaze directions. http://www.jdl.ac.cn/peal/top.htm
CFD [29]	Chicago Face Database	Contains 608 individuals (307 males and 298 females), where various emotional expressions including neutral, fearful/afraid, angry and happy (both open and closed mouth smiles) were captured. Furthermore, there is extensive analysis of the images using 30 facial features. https://chicagofaces.org/default/
CUHK [30]	Chinese University of Hong Kong	Contains 188 individuals (54 females and 134 males) in neutral and frontal poses. http://mmlab.ie.cuhk.edu.hk/archive/facesketch.html
Color FERET [31]	Face Recognition Technology	Contains 994 individuals including frontal images with neutral and randomly different facial expression (usually a smile). Ground truth information includes year of birth, sex, ethnic background and accessory. https://www.nist.gov/itl/products-and-services/color-feret-database.
JAFFE [32]	Japanese Female Facial Expression	Contains ten female individuals with seven facial expressions. https://zenodo.org/record/3451524.XrtXnS1L2u4
MR2 [33]	Multi-racial, Mega-resolution Database	Consists of 74 individuals (41 females and 33 males) between the age of 18 and 25 with neutral expressions. http://ninastrohminger.com/the-mr2
Texas 3DFR [34]	Texas 3D Face Recognition Database	Contains 2D and 3D images of 115 individuals (84 males and 31 females). Ground truth information includes sex, ethnicity, facial expression and locations of 25 manually located anthropometric facial fiducial points. http://live.ece.utexas.edu/research/texas3dfr/
FEI [35]	-	Contains 200 Brazilians (100 males and 100 females) in the frontal position with a profile rotation of up to approximately 180 degrees, with neutral and smile expressions considered. https://fei.edu.br/ cet/facedatabase.html
TFEID [36]	Taiwanese Facial Expression Image Database	Contains 39 individuals with eight frontal facial expressions including neutral. http://bml.ym.edu.tw/tfeid/
IFD [37]	Indian Facial Database	Contains 66 Indians (43 males and 23 females) in the frontal positions that express the problems faced by individuals wearing spectacles.https://ieee-dataport.org/documents/indian-facial-database-highlighting-spectacle-problems

the population data into their major ethnic groups. Section III presents our methodology, describes the experiments, and discusses the obtained results using handcrafted features. We compare these features with the features learned automatically by standard deep learning algorithms in Section IV. Finally, we draw some conclusions in Section V.

II. OVERVIEW OF TWO-DIMENSIONAL COMPUTER VISION BENCHMARK DATA SETS

This section describes the 2D CVBDs that we managed to gain access to and the challenges faced when categorizing the population data into the appropriate major ethnic groups.

A. TWO-DIMENSIONAL COMPUTER VISION DATA SETS

There exist numerous 2D face databases for various purposes. Facial anthropometry researchers build 2D face databases in the healthcare field [1], [2], [8], [9], [17] to evaluate facial feature differences between populations. On the other hand, computer vision researchers build 2D face databases [27]–[37], [41] with various poses, illuminations, expressions as well as different accessories (e.g., spectacles, beard, mustache) to evaluate face recognition and detection algorithms. These images are captured in controlled conditions and are often referred to as the 'gold standard data sets' for testing computer vision algorithms.

We now describe the 2D CVBDs that we managed to obtain access (see Table 2 for details). We note that the list is not exhaustive. However, it is sufficient enough to demonstrate that benchmark data sets developed in the computer vision field can be used in the healthcare field to study the facial feature differences. One of the main challenges we faced while working on the 2D CVBDs was categorizing the ground truth information (GTI) labels into their major ethnic groups as they were developed separately in different regions of the world. The study in [42] noted that the present European and American studies on race, ethnicity, and health use poorly defined labels for population studies. However, the search for an accurate definition is controversial for scientific and social reasons as well as due to the changing meaning of ethnicity in the United Kingdom and the United States. We do not argue that an accurate definition is unnecessary. Instead, we seek to review the challenges faced in the next section to assist researchers in categorizing populations' GTI labels into the appropriate major ethnic groups.

B. ETHNIC GROUPS CATEGORIZATION

There seems to be a great deal of confusion surrounding the definition of the term ''ethnic'' or ''ethnicity''. One such example is found in the Webster dictionary [43], which defines ''ethnic'' as ''of or relating to large groups of people

classed according to common racial, national, tribal, religious, linguistic, or cultural origin or background''. We categorize the GTI labels from the 2D CVBDs based on their ancestry. Here, ancestry refers to the origins of a population, i.e., place of birth of the person or the person's parents or ancestors prior to migrating to a new country.

The US Office of Management and Budget (OMB) defines a *WHITE* person as having origins from Europe, North Africa, or the Middle East. However, in Britain, Middle Easterners and North Africans are not considered *WHITE*. We categorize these two conversational populations into the *MIDDLE EAST* ethnicity as they share the same ancestry. Furthermore, the study in [27] showed there are distinct differences between North Africans and whites in France. Likewise, the *ASIAN* ethnic group refers to persons of Asian origins, and thus, *Indian*, *Chinese* and *Japanese* people are categorized as Asians. Furthermore, the Taiwanese population is categorized as Chinese, as the authors in [44] showed that more than 95% of Taiwanese people are from China.

Likewise, the *BLACK* ethnic group is used to categorize people of African origin. However, there have been disputes regarding whether to categorize the Brazilian population into *LATINO* ethnicity. For example, OMB does not consider Brazilian Americans to be Latinos as they define the term *LATINO* to be synonymous with the term *Hispanic*, implying a Spanish-speaking society, while Brazil is a Portuguese-speaking society. We have chosen to categorize Brazilians under the *LATINO* ethnicity based on their Latin American origins. Second, we also categorize populations that have *Hispanic* labels from the benchmark data sets to be in the *LATINO* category, as the data sets were created in the US. As noted earlier, the Hispanic and Latino categories are used synonymously in the US.

TABLE 3. Population categorization of the 2D CVBDs based on ethnicity: major and sub-ethnic groups.

Major ethnic groups	Description	Sub-ethnic groups
WHITE	People of European origin.	*White*, *White French*
MIDDLE EAST	People of Middle Eastern origin.	*North African French*, *Middle Eastern*
LATINO	People of Latin American origin.	*Latino*, *Brazilian*
ASIAN	People of Asian origin.	*Asian*, *Indian*, *Chinese*, *Japanese*
BLACK	People of African origin.	*Black*

We summarize the ethnicity categorization used in the paper in Table 3. We indicate major ethnic groups using italicized uppercase letters. On the other hand, we refer to the GTI labels from the 2D CVBDs as sub-ethnic groups and indicate them using italicized lowercase letters. The importance of having sub-ethnic categorization cannot be ignored because, as pointed out by the researchers in [45], their results are weak as they selected subjects randomly from three sub-ethnic groups, and categorized them into a single major ethnic group (i.e., *ASIAN*).

III. EXPERIMENTAL ANALYSIS

This section describes the methodology and provides a comprehensive analysis of the results.

A. METHODOLOGY

In this section, we present our experimental framework comprising data acquisition, preprocessing, feature extraction methods, and classification.

1) DATA ACQUISITION

For this study, we used the face images from the 2D CVBDs described in Section II-A. As our results depend heavily on the accuracy of the feature extraction step, we only consider frontal and neutral face images under controlled lighting with indoor environments. Moreover, we manually removed images of subjects wearing accessories, including spectacles and long beard, to prevent occlusions from affecting our final results. Table 4 summarizes the number of images used from each database.

2) PREPROCESSING

This phase consists of detecting and normalizing face images to eliminate noise or inconsistencies that may lead to misleading results.

Face detection We begin face detection by correcting the illumination of the images using the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm from the Open Source Computer Vision (OpenCV) library before locating the head's position to reduce the search space. We applied a pre-trained histogram of gradient (HoG) features and the support vector machine (SVM) algorithm from the Dlib library for face detection over the Haar Cascade face detector. The former works very well for frontal images and is considerably faster and more lightweight than the latter. The output of this step is a bounding box around the detected face in the input image.

Face normalization Even though the images share similar properties such as controlled lighting with indoor environments, some differences will still be present, in particular, the sizes and positions of faces in each database. We believe this discrepancy may be attributed to the differences in the distance between the camera and the subjects. Therefore, face normalization is an essential step before feature extraction, where the original image is warped and transformed into the desired output coordinate space. To do so, we first feed the output from the previous step to the Dlib facial landmark predictor to retrieve the eye regions as they are used as reference points. Thus, the center of the eyes and the angle between eye centroids are computed to allow for rotational correction. Next, we apply affine

TABLE 4. The number of images used in this study based on major and sub-ethnic groups.

	CANAFF	CAS-PEAL	CFD	CUHK	Color FERET	FEI	MR2	TFID	Texas 3DFR	JAFFE	IFD	Total
ASIAN												
Asian	0	0	109	0	78	0	20	0	27	0	0	234
Chinese	0	1024	0	187	0	0	0	39	0	0	0	1,250
Indian	0	0	0	0	0	0	0	0	7	0	51	58
Japanese	0	0	0	0	0	0	0	0	0	9	0	9
LATINO												
Latino	0	0	108	0	28	0	0	0	0	0	0	136
Brazilian	0	0	0	0	0	190	0	0	0	0	0	190
WHITE												
White	0	0	181	0	214	0	22	0	76	0	0	493
White French	71	0	0	0	0	0	0	0	0	0	0	71
BLACK												
Black	0	0	193	0	35	0	32	0	3	0	0	263
MIDDLE EAST												
Middle Eastern	0	0	0	0	21	0	0	0	0	0	0	21
North African French	64	0	0	0	0	0	0	0	0	0	0	64
Total	135	1,024	591	187	376	190	74	39	113	9	51	2,789

FIGURE 1. Subjects 564 and 576 from the Color FERET database. Top to bottom: before and after face normalization.

TABLE 5. Linear features used in the study to represent the four regions of interest: left eye, nose, mouth and chin. The positions of the landmarks are shown in Figure 2b.

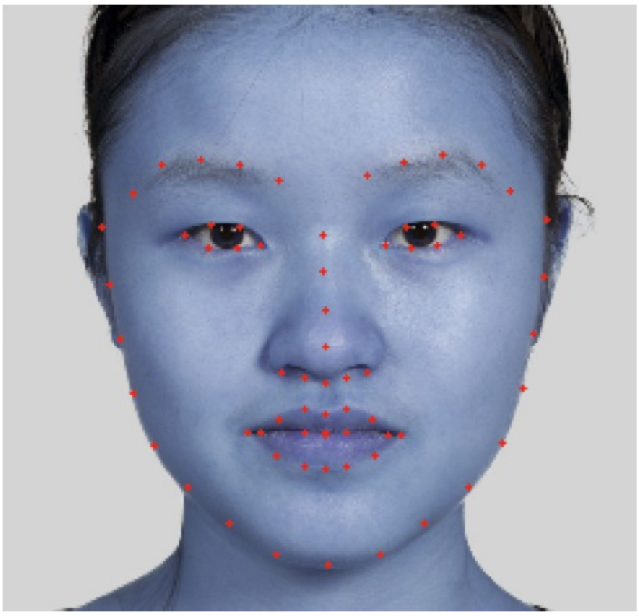
Features	Position of Landmarks	Description
he	45 and 47	Height of eye
we	43 and 46	Width of eye
hn	28 and 34	Height of nose
wn	32 and 36	Width of nose
hm	52 and 58	Height of mouth
wm	49 and 55	Width of mouth
hf	9 and 28	Height of face
wf	3 and 15	Width of face
dn	52, 63, 67, 58 and 34	Distance from the tip of the nose to the center of the mouth
de	40 and 43	Distance between the inner corners of the eyelids

transformations so that both eyes are on the same horizontal line before scaling the face's size to be approximately identical. However, if a failure is detected during this step, we exclude the image from our data set. Figure 1 shows the facial images before and after normalization. We note that a similar face normalization approach was used in [46].

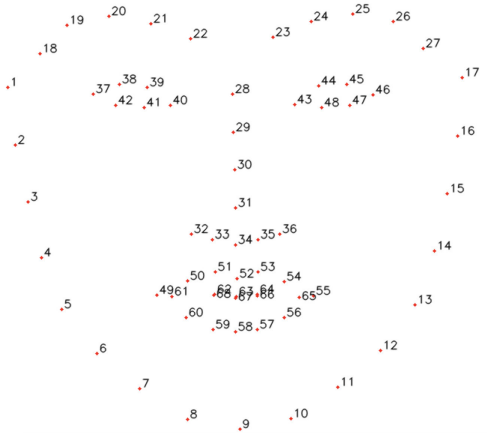
3) FEATURE EXTRACTION METHOD

This phase extract features using the facial landmark detector of [47], which consists of an ensemble of regression trees that have been pre-trained using the 68 landmark positions annotated on each image from the iBUG 300-W database for direct estimation from pixel intensities. This phase's output is 68 (x, y)-coordinates that highlight the contours of seven facial regions: the chin, left eye, left eyebrow, right eye, right eyebrow, nose, and mouth. For example, Figure 2a shows how the 68 annotated landmark positions overlap on an image. Whereas, the contours, indicated by the landmarks, are shown in Figure 2b. Of note, we disregard the eyebrow as a prominent facial region and focus on the following four regions: chin, left eye (as eyes are generally shaped identically), nose and mouth as subjects may shave their eyebrows. We then represent these regions using linear measurements, with Euclidean distances computed between identified landmark positions. We used five horizontal and five vertical features. More details are given in Table 5, and these features are used throughout the study.

The descriptive data of the mean, standard deviation, and maximum and minimum dimensions of our features are presented in Appendix V-A for sub-ethnic groups. Whereas, Appendix V-B report them for the major ethnic groups instead. As we have multiple features with varying ranges of value, feature scaling is necessary when using machine learning algorithms. Therefore, we consider two sets of features, described below, where the statistical significance is tested using Welch's t-test.



(a) Original image of the subject annotated with 68 landmark positions.



(b) Contours of the facial regions of eyebrows, eyes, nose, mouth and chin indicated by the 68 landmark positions.

FIGURE 2. The figure on top shows the annotated landmark positions on the original image of subject af01 from the MR2 database, while the figure below shows the contours of these positions.

Raw feature set This set of features uses the raw data as it is. The results of the Welch's t-test are reported in Appendix V-C for the sub-ethnic groups, and in Appendix V-D for the major ethnic groups. The results take on the form of the following format: ''*he-hn-hm-hf-we-wn-wm-wf-dn-de*'', where we specify the feature if statistical differences exist between population's mean data of the feature. However, if no statistical differences exist, we indicate it with the string ''xx''.

Normalized feature set This set of features applies min-max normalization on the raw data. The results of the Welch's t-test are reported in Appendix V-E (for sub-ethnic groups) and Appendix V-F (for major ethnic groups). We chose the min-max normalization instead of any other normalization as it mimics the way healthcare researchers work on facial features. For example, the researchers in [48] only considered females regardless of their ethnicity with the upper lip length between 18 - 22 mm range, where the minimum value is 18 mm, while the maximum value is 22 mm. They obtained these ranges that we refer to as reference range by studying 60 females aged between 18 and 35 years. Their technique assumes the upper lip length of an average female will be in the reference range, and it will be rare to find a female that has an upper lip length of more than 22 mm. However, in our study, instead of normalizing features based on gender, we normalized them based on major and sub-ethnic groups to obtain the reference range. Thus, we note that our model will fail to classify an individual that is not within the reference range until it retrains with the new reference range.

4) CLASSIFICATION

To test the effectiveness of features ranked with information gain values compared to the traditional p-values on the raw and normalized feature sets, we employed the extreme gradient boosting (XGBoost) and Select K Best (SKB) feature selection algorithms. The XGBoost library implements the gradient boosting decision tree algorithm for feature selection and classification. The feature selection algorithm ranks the importance of a feature using a value known as information gain; the more prominent the feature is, the higher the information gain value is. These ranked features are fed to the XGBoost classifier for classification. By contrast, the employed SKB feature selection algorithm implements the ANOVA F-value that generates the p-values to order the features. These ranked features are fed to the SVM for classification.

Moreover, to find the k top discriminative features for major and sub-ethnic groups, another experiment that assesses the number of features is also designed, where k varies from 1 to 10. As we have imbalanced data sets, we employed stratified K-fold as the cross-validation technique for both the XGBoost and SVM classifiers. We set K to three as the lowest number of samples in the sub-ethnic groups is nine.

5) EVALUATION METRICS

We evaluated the described experiments using three machine learning evaluation metrics: accuracy, F1-Score, and confusion matrix. All values were between 0 and 1, and the standard deviations are given in brackets. A perfect 1 is the best score, while a perfect 0 is the worst score.

Accuracy is a metric tied to precision and recall. High precision and recall scores show that the classifier is giving accurate results (high precision), and the majority of the results are positive (recall). F1-Score considers both precision and recall by taking the harmonic mean of them. We report micro- and macro-averages for F1-Score, whereby each score has a different interpretation. A micro-average considers each sample equally, whereas a macro-average considers each class equally. The former is preferable for imbalanced data

sets, and the latter for balanced data sets. Both the micro- and macro-averages will report the same scores if the data sets are balanced.

This paper reports the micro-average F1-Score due to the class imbalance problem, and the macro-average F1-Score to show the skewed class distribution. On the other hand, we used the normalized confusion matrix to summarize the performance of a classifier where each row represents an actual class, while each column represents a predicted class. A satisfactory confusion matrix will have most of its instances on its main diagonal.

B. RESULTS

This section describes the experimental results obtained using handcrafted features on XGBoost and SVM classifiers based on major and sub-ethnic groups.

1) SUB-ETHNIC GROUPS

This section describes the experimental results of using raw and normalized feature sets on the sub-ethnic groups. For the raw feature set, we summarize the top ten information-gain-based features selected by the XGBoost feature selection algorithm in Table 6, and the top ten p-value-based features selected by the SKB feature selection algorithm in Table 7. Of note, these algorithms ranked *hn* in the same order.

TABLE 6. The k top raw-information-gain-based features in the sub-ethnic groups. Here k is 10, and the features are ordered based on the highest information gain values.

	k top features	Information gain values
1	*de*	1.64
2	*we*	1.13
3	*wn*	1.06
4	*hm*	0.97
5	*wm*	0.79
6	*hf*	0.71
7	*he*	0.60
8	*wf*	0.59
9	*hn*	0.58
10	*dn*	0.58

TABLE 7. The k top raw-p-value-based features in the sub-ethnic groups. Here k is 10, and the features are ordered based on the smallest p-values.

	k top features	p-values
1	*we*	1.32E-280
2	*de*	3.26E-269
3	*wm*	1.53E-178
4	*wn*	2.25E-176
5	*hm*	2.39E-173
6	*he*	1.01E-164
7	*hf*	4.01E-116
8	*dn*	1.20E-82
9	*hn*	1.77E-54
10	*wf*	4.33E-49

On the other hand, Tables 8 and 9 show the effect of varying the number of k top features on the raw feature set, where k varies from 1 to 10. The information-gain-based features achieved the best accuracy and micro-average F1-Score

TABLE 8. Effect of variation on the k top raw-information-gain-based features in the sub-ethnic groups. The standard deviation given in brackets.

k top features	Accuracy	Micro F1-Score	Macro F1-Score
1	0.51 (0.01)	0.51 (0.01)	0.12 (0.01)
2	0.51 (0.01)	0.51 (0.01)	0.15 (0.00)
3	0.52 (0.01)	0.52 (0.01)	0.20 (0.01)
4	0.54 (0.01)	0.54 (0.01)	0.25 (0.02)
5	0.55 (0.01)	0.55 (0.01)	0.27 (0.02)
6	0.59 (0.01)	0.59 (0.01)	0.30 (0.01)
7	0.60 (0.01)	0.60 (0.01)	0.31 (0.01)
8	0.61 (0.01)	0.61 (0.01)	0.33 (0.01)
9	0.62 (0.01)	0.62 (0.01)	0.35 (0.02)
10	0.62 (0.00)	0.62 (0.00)	0.35 (0.02)

TABLE 9. Effect of variation on the k top raw-p-value-based features in the sub-ethnic groups. The standard deviation given in brackets.

k top features	Accuracy	Micro F1-Score	Macro F1-Score
1	0.51 (0.01)	0.51 (0.01)	0.10 (0.00)
2	0.52 (0.00)	0.52 (0.00)	0.10 (0.00)
3	0.53 (0.01)	0.53 (0.01)	0.12 (0.01)
4	0.57 (0.01)	0.57 (0.01)	0.18 (0.01)
5	0.61 (0.01)	0.61 (0.01)	0.22 (0.00)
6	0.61 (0.01)	0.61 (0.01)	0.24 (0.01)
7	0.63 (0.01)	0.63 (0.01)	0.26 (0.01)
8	0.64 (0.01)	0.64 (0.01)	0.29 (0.01)
9	0.65 (0.01)	0.65 (0.01)	0.29 (0.01)
10	0.67 (0.01)	0.67 (0.01)	0.31 (0.01)

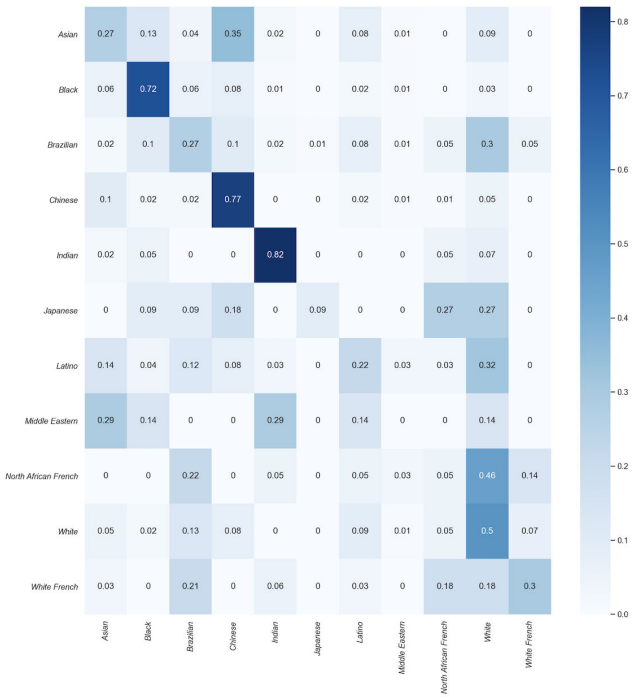


FIGURE 3. Normalized confusion matrix of using the top nine raw-information-gain-based features in the sub-ethnic groups.

(i.e., 0.62) using all the nine features. Similarly, p-value-based features obtained the best accuracy and micro-average F1-Score (i.e., 0.67) using all the ten features. Of note, Figure 3 shows the confusion matrix of using the top nine information-gain-based features, while Figure 4 shows the confusion matrix of using the top ten p-value-based features.

FIGURE 4. Normalized confusion matrix of using the top ten raw-p-value-based features in the sub-ethnic groups.

TABLE 10. The k top normalized-information-gain-based features in the sub-ethnic groups. Here k is 9, and the features are ordered based on the highest information gain values.

	k top features	Information gain values
1	*dn*	4.26
2	*hf*	2.67
3	*wn*	1.47
4	*wf*	1.30
5	*wm*	1.05
6	*hn*	0.95
7	*we*	0.82
8	*hm*	0.80
9	*he*	0.75

TABLE 11. The k top normalized-p-value-based features in the sub-ethnic groups. Here k is 9, and the features are ordered based on the smallest p-value.

	k top features	p-values
1	*dn*	3.25E-124
2	*wm*	1.29E-74
3	*hn*	4.69E-67
4	*wf*	8.25E-57
5	*hm*	1.05E-53
6	*wn*	6.72E-44
7	*he*	2.50E-40
8	*hf*	3.99E-40
9	*we*	8.53E-38

TABLE 12. Effect of variation on the k top normalized-information-gain-based features in the sub-ethnic groups. The standard deviation given in brackets.

k top features	Accuracy	Micro F1-Score	Macro F1-Score
1	0.84 (0.00)	0.84 (0.00)	0.55 (0.02)
2	0.94 (0.01)	0.94 (0.01)	0.80 (0.02)
3	0.96 (0.00)	0.96 (0.00)	0.80 (0.01)
4	0.95 (0.00)	0.95 (0.00)	0.79 (0.02)
5	0.94 (0.01)	0.94 (0.01)	0.78 (0.03)
6	0.94 (0.01)	0.94 (0.01)	0.77 (0.04)
7	0.94 (0.01)	0.94 (0.01)	0.75 (0.04)
8	0.94 (0.01)	0.94 (0.01)	0.74 (0.03)
9	0.94 (0.00)	0.94 (0.00)	0.75 (0.04)

TABLE 13. Effect of variation on the k top normalized-p-value-based features in the sub-ethnic groups. The standard deviation given in brackets.

k top features	Accuracy	Micro F1-Score	Macro F1-Score
1	0.45 (0.00)	0.45 (0.00)	0.08 (0.02)
2	0.53 (0.00)	0.53 (0.00)	0.14 (0.00)
3	0.54 (0.01)	0.54 (0.01)	0.14 (0.01)
4	0.56 (0.01)	0.56 (0.01)	0.17 (0.01)
5	0.59 (0.01)	0.59 (0.01)	0.20 (0.01)
6	0.60 (0.00)	0.60 (0.00)	0.22 (0.00)
7	0.61 (0.01)	0.61 (0.01)	0.23 (0.02)
8	0.64 (0.00)	0.64 (0.00)	0.26 (0.01)
9	0.66 (0.01)	0.66 (0.01)	0.30 (0.02)

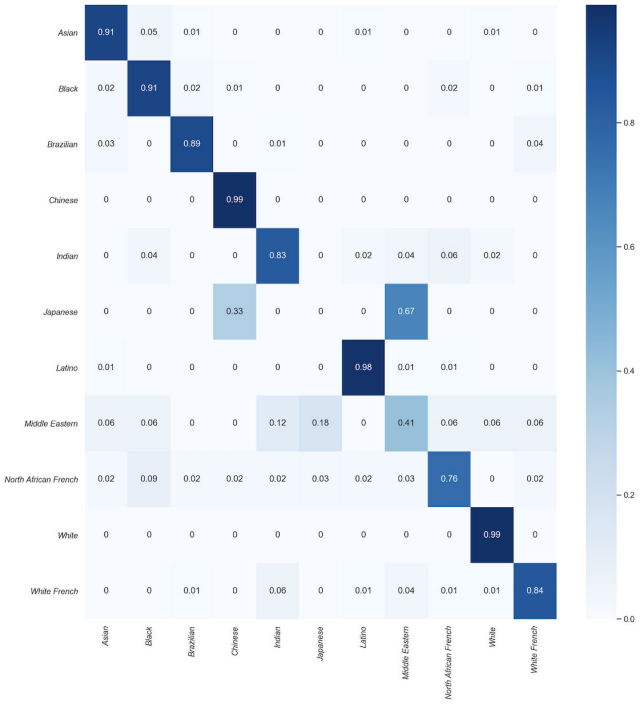


FIGURE 5. Normalized confusion matrix of using the top three normalized-information-gain-based features in the sub-ethnic groups.

Hence, we can conclude that regardless of the order the feature selection algorithms ranked the features, they yielded almost similar prediction results.

Likewise, in Tables 10, 11, 12 and 13, we show the same statistics but on the normalized feature set. Table 10 shows the top nine normalized information-gain-based features, whereas Table 11 shows the top nine p-value-based features. Both the feature selection algorithms discarded *de* as a prominent feature as the value became zero after normalization. The algorithms ranked *dn* on the first position and *wf* on the fourth position. As indicated in Table 12, the information-gain-based features obtained the best accuracy and micro-average

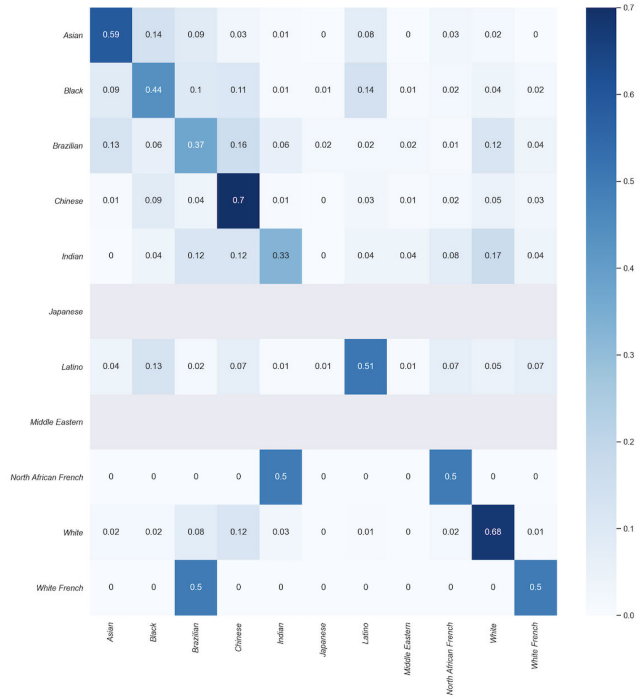


FIGURE 6. Normalized confusion matrix of using the top nine normalized-p-value-based features in the sub-ethnic groups.

TABLE 14. The k top raw-information-gain-based features in the major ethnic groups. Here k is 10, and the features are ordered by the highest information gain values.

	k top features	Information gain values
1	*we*	2.14
2	*hm*	1.25
3	*de*	1.18
4	*wn*	1.14
5	*wm*	0.90
6	*hf*	0.71
7	*hn*	0.67
8	*wf*	0.62
9	*dn*	0.58
10	*he*	0.57

TABLE 15. The k top raw-p-value-based features in the major ethnic groups. Here k is 10, and the features are ordered by the smallest p-values.

	k top features	p-values
1	*we*	5.57E-198
2	*de*	2.08E-180
3	*wm*	6.82E-154
4	*hm*	2.57E-153
5	*wn*	3.28E-122
6	*he*	9.17E-64
7	*hf*	2.62E-34
8	*wf*	1.68E-32
9	*dn*	4.10E-30
10	*hn*	1.34E-26

F1-Score (i.e., 0.96) using the top three features. By contrast, the p-value-based features achieved the best accuracy and micro-average F1-Score (i.e., 0.66) using all the nine features, as shown in Table 13. Of note, Figure 5 shows the confusion matrix of using the top three information-gain-based features,

TABLE 16. Effect of variation on the k top raw-information-gain-based features in the major ethnic groups. The standard deviation given in brackets.

k top features	Accuracy	Micro F1-Score	Macro F1-Score
1	0.59 (0.00)	0.59 (0.00)	0.23 (0.01)
2	0.63 (0.01)	0.63 (0.01)	0.35 (0.01)
3	0.62 (0.01)	0.62 (0.01)	0.38 (0.02)
4	0.64 (0.01)	0.64 (0.01)	0.43 (0.01)
5	0.67 (0.01)	0.67 (0.01)	0.45 (0.02)
6	0.69 (0.01)	0.69 (0.01)	0.47 (0.00)
7	0.70 (0.00)	0.70 (0.00)	0.48 (0.01)
8	0.72 (0.00)	0.72 (0.00)	0.50 (0.01)
9	0.73 (0.01)	0.73 (0.01)	0.52 (0.01)
10	0.73 (0.01)	0.73 (0.01)	0.52 (0.01)

TABLE 17. Effect of variation on the k top raw-p-value-based features in the major ethnic groups. The standard deviation given in brackets.

k top features	Accuracy	Micro F1-Score	Macro F1-Score
1	0.59 (0.01)	0.59 (0.01)	0.23 (0.01)
2	0.59 (0.00)	0.59 (0.00)	0.23 (0.00)
3	0.61 (0.01)	0.61 (0.01)	0.27 (0.02)
4	0.67 (0.01)	0.67 (0.01)	0.37 (0.01)
5	0.69 (0.01)	0.69 (0.01)	0.40 (0.01)
6	0.69 (0.01)	0.69 (0.01)	0.40 (0.01)
7	0.72 (0.01)	0.72 (0.01)	0.48 (0.01)
8	0.73 (0.01)	0.73 (0.01)	0.49 (0.01)
9	0.74 (0.01)	0.74 (0.01)	0.52 (0.02)
10	0.75 (0.01)	0.75 (0.01)	0.52 (0.01)

TABLE 18. The k top normalized-information-gain-based features in the major ethnic groups. Here k is ten, and the features are ordered based on the highest information gain values.

	k top features	Information gain values
1	*de*	13.12
2	*hf*	4.33
3	*dn*	3.42
4	*we*	1.55
5	*wn*	1.23
6	*wf*	0.75
7	*wm*	0.57
8	*hm*	0.49
9	*he*	0.40
10	*hn*	0.35

TABLE 19. The k top normalized-p-value-based features in the major ethnic groups. Here k is 10, and the features are ordered based on the smallest p-values.

	k top features	p-values
1	*de*	8.09E-254
2	*hf*	3.55E-148
3	*wf*	2.57E-112
4	*wn*	6.68E-86
5	*we*	2.14E-80
6	*dn*	2.48E-65
7	*hm*	3.70E-48
8	*hn*	4.39E-13
9	*wm*	2.05E-11
10	*he*	2.81E-02

while Figure 6 shows the confusion matrix of using all the nine p-value-based features. Therefore, we can conclude that information-gain-based features selected by the XGBoost feature selection algorithm yielded better prediction results than the p-value-based features selected by the SKB feature selection algorithm on the normalized feature set.

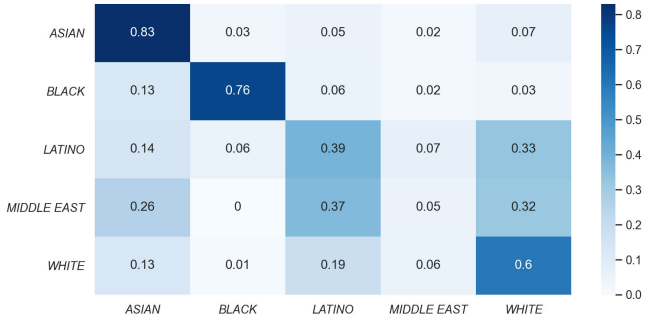


FIGURE 7. Normalized confusion matrix of using the top nine raw-information-gain-based features in the major ethnic groups.

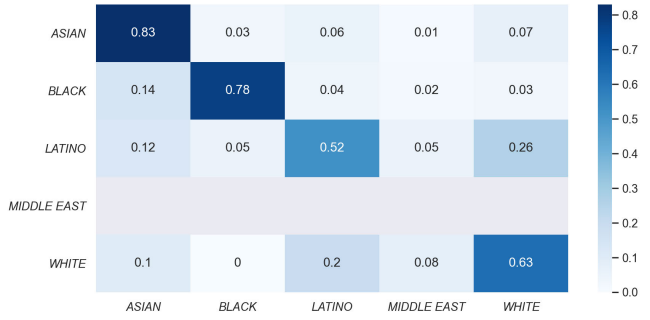


FIGURE 8. Normalized confusion matrix of using the top ten raw-p-value-based features in the major ethnic groups.

TABLE 20. Effect of variation on the k top normalized-information-gain-based features in the major ethnic groups. The standard deviation given in brackets.

k top features	Accuracy	Micro F1-Score	Macro F1-Score
1	0.65 (0.00)	0.65 (0.00)	0.30 (0.00)
2	0.97 (0.00)	0.97 (0.00)	0.93 (0.01)
3	0.99 (0.00)	0.99 (0.00)	0.97 (0.01)
4	0.99 (0.00)	0.99 (0.00)	0.97 (0.00)
5	0.98 (0.00)	0.98 (0.00)	0.95 (0.01)
6	0.98 (0.00)	0.98 (0.00)	0.95 (0.01)
7	0.98 (0.00)	0.98 (0.00)	0.95 (0.01)
8	0.98 (0.00)	0.98 (0.00)	0.94 (0.01)
9	0.98 (0.01)	0.98 (0.01)	0.94 (0.01)
10	0.98 (0.00)	0.98 (0.00)	0.94 (0.01)

TABLE 21. Effect of variation on the k top normalized-p-value-based-features in the major ethnic groups. The standard deviation given in brackets.

k top features	Accuracy	Micro F1-Score	Macro F1-Score
1	0.63 (0.00)	0.63 (0.00)	0.29 (0.00)
2	0.68 (0.01)	0.68 (0.01)	0.40 (0.01)
3	0.70 (0.00)	0.70 (0.00)	0.41 (0.00)
4	0.71 (0.01)	0.71 (0.01)	0.42 (0.00)
5	0.73 (0.01)	0.73 (0.01)	0.49 (0.01)
6	0.81 (0.01)	0.81 (0.01)	0.57 (0.01)
7	0.83 (0.01)	0.83 (0.01)	0.62 (0.02)
8	0.83 (0.01)	0.83 (0.01)	0.62 (0.02)
9	0.83 (0.00)	0.83 (0.00)	0.63 (0.01)
10	0.84 (0.00)	0.84 (0.00)	0.64 (0.02)

2) MAJOR ETHNIC GROUPS

This section describes the experimental results of using raw and normalized feature sets based on the major ethnic groups.

TABLE 22. Evaluation metrics results of the deep learning techniques on sub-ethnic groups.

Deep learning techniques	Loss	Accuracy	Precision	Recall
VGG16	0.70 (0.39)	0.91 (0.04)	0.91 (0.04)	0.91 (0.04)
ResNet50	0.58 (0.08)	0.90 (0.02)	0.91 (0.01)	0.90 (0.01)
Inception-V3	0.32 (0.06)	0.94 (0.01)	0.94 (0.01)	0.94 (0.01)

TABLE 23. Evaluation metrics results of the deep learning techniques on major ethnic groups.

Deep learning techniques	Loss	Accuracy	Precision	Recall
VGG16	0.56 (0.32)	0.91 (0.04)	0.91 (0.04)	0.91 (0.04)
ResNet50	0.58 (0.13)	0.90 (0.01)	0.90 (0.01)	0.90 (0.01)
Inception-V3	0.39 (0.02)	0.93 (0.00)	0.93 (0.00)	0.93 (0.00)

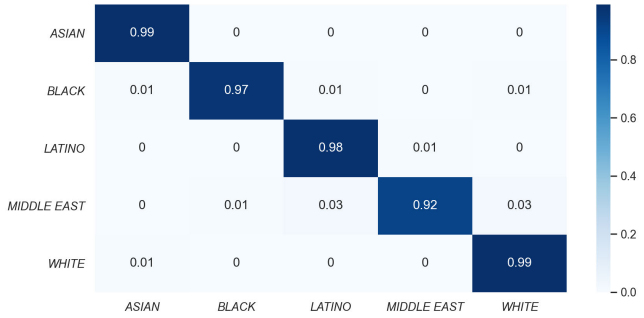


FIGURE 9. Normalized confusion matrix of using the top three normalized-information-gain-based features in the major ethnic groups.

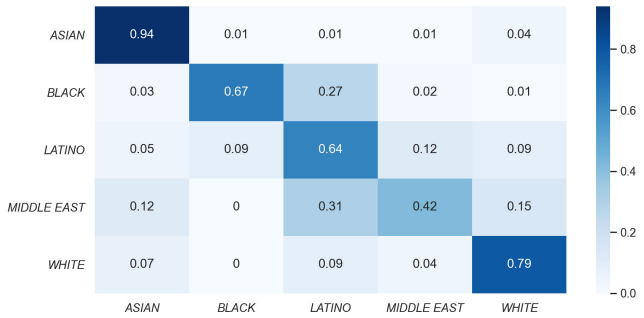
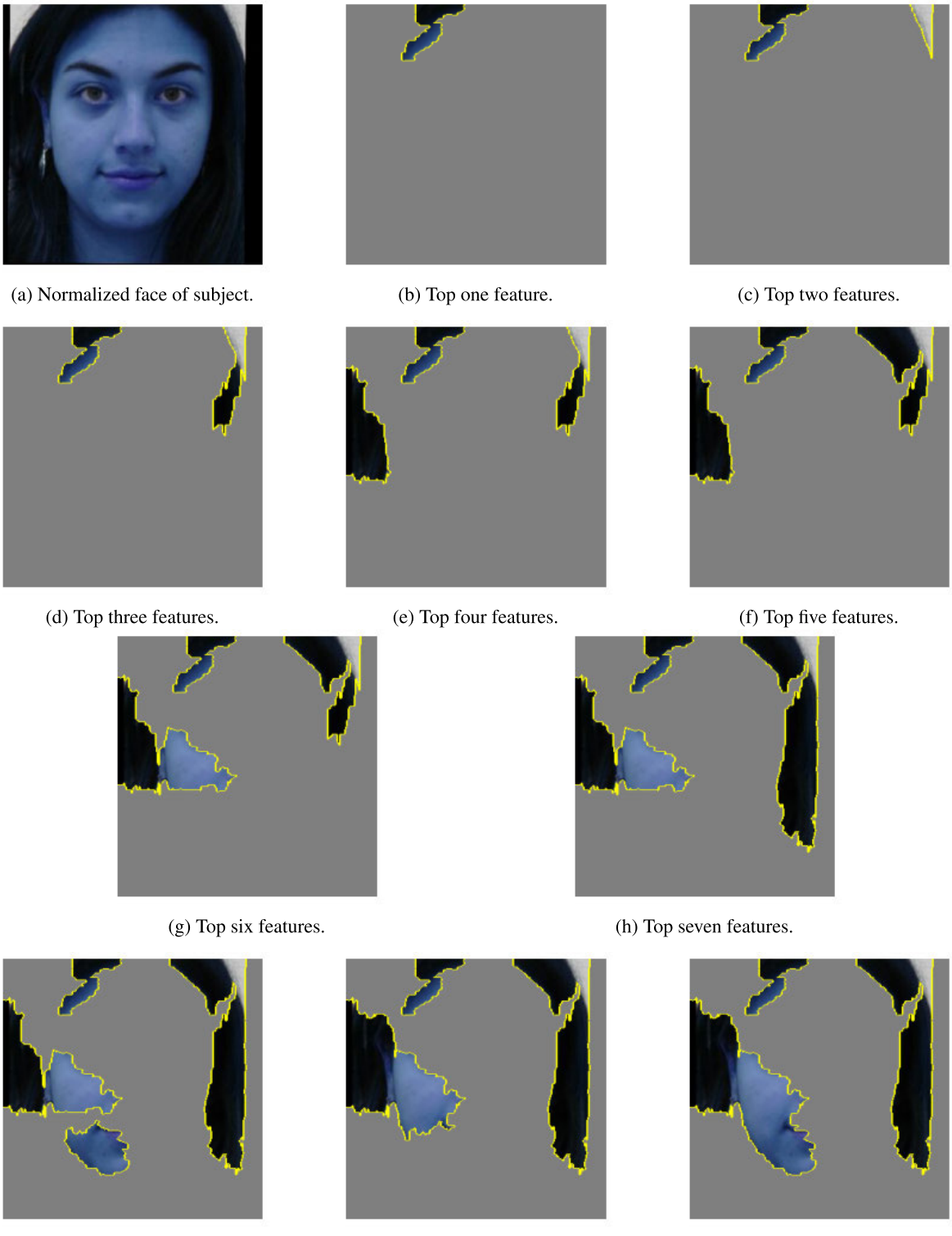


FIGURE 10. Normalized confusion matrix of using the top ten normalized-p-value-based features in the major ethnic groups.

For the raw feature set, Table 14 summarizes the top ten information-gain-based features and Table 15 summarizes the top ten p-value-based features. The XGBoost and SKB feature selection algorithms ranked *we*, *wf* and *dn* in the same order. As indicated in Table 16, the information-gain-based features obtained the best accuracy and micro F1-Score (i.e., 0.73) using the top nine features. By contrast, p-value-based features achieved the best accuracy and micro F1-Score (i.e., 0.75) using all the ten features (see Table 17). Of note, Figure 7 shows the confusion matrix of using the top nine information-gain-based features, while Figure 8 shows the confusion matrix of using the top ten p-value-based features.



(a) Normalized face of subject. (b) Top one feature. (c) Top two features.

(d) Top three features. (e) Top four features. (f) Top five features.

(g) Top six features. (h) Top seven features.

(i) Top eight features. (j) Top nine features. (k) Top ten features.

FIGURE 11. The k top features of subject 69a from the FEI database.

Thus, we can conclude that regardless of the order the feature selection algorithms ranked the features, they yielded almost similar prediction results.

Similarly, Tables 18, 19, 20 and 21 show the same comparisons, but on the normalized feature set. Table 18 summarizes the top ten information-gain-based features, whereas Table 19

TABLE 24. Descriptive statistics of the raw feature set in the sub-ethnic groups.

	he		*hn*		*hm*		*hf*		*we*		*wn*		*wm*		*wf*		*dn*		*de*	
Sub-ethnic groups	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**
Asian (234)	9.3 (1.6)	14.0 (3.0)	59.4 (3.9)	74.0 (51.0)	21.7 (3.4)	33.0 (10.0)	138.6 (8.7)	175.0 (117.0)	29.4 (1.7)	34.0 (26.0)	35.0 (3.0)	47.0 (28.0)	61.8 (4.7)	81.0 (50.0)	166.1 (8.6)	200.0 (147.0)	27.8 (2.8)	36.0 (18.0)	46.3 (2.1)	59.0 (39.0)
Black (263)	10.3 (1.7)	15.0 (3.0)	57.0 (4.2)	71.0 (46.0)	25.6 (3.5)	39.0 (14.0)	138.0 (8.8)	163.0 (119.0)	30.5 (1.6)	36.0 (26.0)	37.6 (3.0)	46.0 (31.0)	67.0 (4.5)	86.0 (57.0)	158.2 (8.8)	188.0 (136.0)	29.6 (2.9)	36.0 (23.0)	45.0 (1.7)	50.0 (40.0)
Brazilian (190)	10.6 (1.5)	15.0 (5.0)	60.0 (4.2)	74.0 (50.0)	21.1 (3.4)	32.0 (13.0)	142.1 (8.3)	165.0 (122.0)	31.4 (1.4)	35.0 (27.0)	32.4 (2.6)	39.0 (27.0)	64.4 (4.0)	76.0 (55.0)	160.6 (8.0)	190.0 (144.0)	26.7 (2.7)	34.0 (21.0)	44.3 (1.6)	49.0 (41.0)
Chinese (1250)	9.0 (1.4)	14.0 (4.0)	59.8 (3.8)	74.0 (47.0)	22.1 (3.4)	37.0 (11.0)	137.8 (7.6)	164.0 (114.0)	28.7 (1.4)	33.0 (25.0)	33.3 (2.4)	42.0 (27.0)	59.5 (4.1)	75.0 (48.0)	162.6 (6.8)	187.0 (138.0)	27.0 (2.7)	38.0 (19.0)	47.3 (1.6)	52.0 (41.0)
Indian (58)	13.4 (2.4)	18.0 (6.0)	66.5 (6.5)	78.0 (50.0)	27.3 (5.7)	37.0 (14.0)	161.2 (16.0)	183.0 (107.0)	32.6 (1.8)	36.0 (25.0)	36.0 (2.4)	42.0 (31.0)	63.8 (3.8)	71.0 (55.0)	164.3 (7.8)	182.0 (145.0)	33.3 (5.4)	44.0 (18.0)	42.5 (1.6)	46.0 (40.0)
Japanese (9)	11.0 (1.1)	12.0 (9.0)	58.7 (3.4)	65.0 (54.0)	21.8 (5.1)	31.0 (15.0)	129.8 (9.2)	150.0 (121.0)	29.8 (1.5)	32.0 (28.0)	32.1 (1.4)	34.0 (30.0)	57.1 (3.8)	64.0 (52.0)	151.1 (6.6)	165.0 (145.0)	25.7 (2.9)	33.0 (24.0)	45.3 (1.7)	48.0 (42.0)
Latino (136)	10.1 (1.4)	14.0 (7.0)	60.0 (4.0)	70.0 (49.0)	20.4 (3.3)	27.0 (9.0)	141.8 (7.3)	160.0 (123.0)	31.0 (1.5)	35.0 (28.0)	34.3 (2.3)	40.0 (29.0)	64.8 (4.1)	79.0 (55.0)	165.9 (7.5)	184.0 (150.0)	27.4 (2.9)	33.0 (19.0)	44.8 (1.6)	49.0 (40.0)
Middle Eastern (21)	9.4 (1.5)	13.0 (7.0)	59.4 (5.3)	71.0 (49.0)	21.1 (3.8)	30.0 (14.0)	141.0 (8.4)	158.0 (123.0)	30.7 (1.4)	34.0 (28.0)	36.0 (2.6)	40.0 (31.0)	65.1 (3.7)	71.0 (59.0)	163.6 (7.5)	180.0 (150.0)	27.0 (2.5)	32.0 (24.0)	45.4 (1.6)	48.0 (42.0)
North African French (64)	11.0 (1.3)	14.0 (8.0)	60.2 (4.8)	74.0 (50.0)	20.7 (3.6)	29.0 (11.0)	145.2 (9.0)	170.0 (125.0)	31.7 (1.4)	34.0 (29.0)	31.8 (2.7)	38.0 (27.0)	61.9 (4.4)	73.0 (49.0)	165.8 (7.8)	185.0 (146.0)	29.2 (2.7)	35.0 (23.0)	44.1 (1.7)	48.0 (41.0)
White (493)	10.1 (1.7)	14.0 (1.0)	60.3 (4.3)	76.0 (48.0)	18.3 (4.0)	34.0 (4.0)	143.7 (9.3)	176.0 (108.0)	30.8 (1.6)	37.0 (24.0)	33.1 (2.8)	48.0 (26.0)	63.8 (4.9)	85.0 (44.0)	165.6 (8.8)	214.0 (135.0)	27.7 (3.2)	39.0 (18.0)	44.8 (2.2)	60.0 (33.0)
White French (71)	11.1 (1.3)	16.0 (8.0)	59.8 (4.3)	67.0 (50.0)	18.2 (3.2)	25.0 (10.0)	144.2 (9.0)	165.0 (125.0)	31.4 (1.4)	34.0 (28.0)	30.1 (2.1)	35.0 (26.0)	60.8 (4.2)	73.0 (49.0)	164.5 (7.2)	183.0 (150.0)	28.1 (2.9)	35.0 (22.0)	44.4 (1.4)	48.0 (40.0)

TABLE 25. Descriptive statistics of the raw feature set in the major ethnic groups.

	he		*hn*		*hm*		*hf*		*we*		*wn*		*wm*		*wf*		*dn*		*de*	
Major ethnic groups	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**	**Mean (Std)**	**Max (Min)**
ASIAN (1551)	9.2 (1.7)	18.0 (3.0)	60.0 (4.1)	78.0 (47.0)	22.2 (3.7)	37.0 (10.0)	138.8 (9.3)	183.0 (107.0)	28.9 (1.7)	36.0 (25.0)	33.7 (2.6)	47.0 (27.0)	60.0 (4.3)	81.0 (48.0)	163.1 (7.3)	200.0 (138.0)	27.3 (3.1)	44.0 (18.0)	47.0 (1.9)	59.0 (39.0)
BLACK (263)	10.3 (1.7)	15.0 (3.0)	57.0 (4.2)	71.0 (46.0)	25.6 (3.5)	39.0 (14.0)	138.0 (8.8)	163.0 (119.0)	30.5 (1.6)	36.0 (26.0)	37.6 (3.0)	46.0 (31.0)	67.0 (4.5)	86.0 (57.0)	158.2 (8.8)	188.0 (136.0)	29.6 (2.9)	36.0 (23.0)	45.0 (1.7)	50.0 (40.0)
LATINO (326)	10.4 (1.5)	15.0 (5.0)	60.0 (4.1)	74.0 (49.0)	20.8 (3.4)	32.0 (9.0)	141.9 (7.9)	165.0 (122.0)	31.2 (1.4)	35.0 (27.0)	33.2 (2.6)	40.0 (27.0)	64.5 (4.1)	79.0 (55.0)	162.8 (8.2)	190.0 (144.0)	27.0 (2.8)	34.0 (19.0)	44.5 (1.6)	49.0 (40.0)
MIDDLE EAST (85)	10.6 (1.5)	14.0 (7.0)	60.0 (4.9)	74.0 (49.0)	20.8 (3.7)	30.0 (11.0)	144.2 (9.0)	170.0 (123.0)	31.4 (1.5)	34.0 (28.0)	32.9 (3.2)	40.0 (27.0)	62.7 (4.4)	73.0 (49.0)	165.2 (7.8)	185.0 (146.0)	28.6 (2.8)	35.0 (23.0)	44.4 (1.7)	48.0 (41.0)
WHITE (564)	10.2 (1.7)	16.0 (1.0)	60.2 (4.3)	76.0 (48.0)	18.3 (3.9)	34.0 (4.0)	143.8 (9.3)	176.0 (108.0)	30.9 (1.5)	37.0 (24.0)	32.8 (2.9)	48.0 (26.0)	63.4 (4.9)	85.0 (44.0)	165.4 (8.7)	214.0 (135.0)	27.8 (3.2)	39.0 (18.0)	44.8 (2.1)	60.0 (33.0)

summarizes the top ten p-value-based features. The feature selection algorithms ranked *de* and *hf* in similar positions. The p-value-based features required all the ten features to obtain the best accuracy and micro F1-Score (i.e., 0.84). By contrast, the information-gain-based features only required the top three features to achieve the best accuracy and micro F1-Score (i.e., 0.99). See Tables 20 and 21 for details. Furthermore, Figure 9 shows the confusion matrix of using the top three information-gain-based features, while Figure 10 shows the confusion matrix of using all the ten p-value-based features. Therefore, we can conclude that information-gain-based features selected by the XGBoost feature selection algorithm yielded better prediction results than the p-value-based features selected by the SKB feature selection algorithm in the normalized feature set.

IV. AUTOMATICALLY-LEARNED FEATURES

This section describes the experimental results obtained using features learned automatically by standard deep learning techniques such as VGG16, ResNet50, and Inception-V3. For all the three deep learning techniques, the epoch size is 50, the batch size is 32, the target size (i.e., height and width) of the images is 128, pooling is avg and the optimizer is Adam. The default learning rate is used for ResNet50, while the other two techniques used the learning rate of 3E-4. These techniques were implemented using the Keras library.

Table 22 shows the evaluation metrics results on the sub-ethnic groups. Cross-entropy loss increases as the predicted labels continue to differ from the predicted labels. As evaluation metrics F1-Score has been removed from Keras, we report our findings using precision and recall metrics. Likewise, in Table 23,we show the same statistics but on the major ethnic groups. Therefore, we can conclude that Inception-V3 gave the best accuracy score (i.e., 0.93-0.94) with the lowest cross-entropy loss (i.e., 0.32-0.39).

To understand the features used by deep learning techniques to classify major and sub-ethnic groups, we used Local Interpretable Model-Agnostic Explanations (LIME) [49]. LIME is a library that can determine the set of features used for the classification. For example, we show the k top features used by the Inception-V3 model to classify subject 69a from the FEI database in Figure 11. Here k is ten. Of note, the LIME library interpreted similar features for the VGG16 and ResNet50 models as well.

V. CONCLUDING REMARKS

We conclude from the above discussion that compared to handcrafted features that are Euclidean distance-based, automatically learned features are shape-based. The shape of the

TABLE 26. Welch's t-test results on the raw feature set for the sub-ethnic groups.

Sub-ethnic groups	*Asian*	*Black*	*Brazilian*	*Chinese*	*Indian*	*Japanese*	*Latino*	*Middle Eastern*	*North African French*	*White*	*White French*
Asian	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-hn-hm-xx-we-wn-wm-wf-dn-de*	*he-xx-xx-hf-we-wn-wm-wf-dn-de*	*he-xx-xx-xx-we-wn-wm-wf-dn-de*	*he-hn-hm-hf-we-wn-wm-xx-dn-de*	*he-xx-xx-hf-xx-wn-wm-wf-xx-xx*	*he-xx-hm-hf-we-wn-wm-xx-xx-de*	xx-xx-xx-xx-*we-xx-wm-xx*-xx-*de*	*he-xx-hm-hf-we-wn-xx-xx-dn-de*	*he-hn-hm-hf-we-wn-wm-xx-xx-de*	*he-xx-hm-hf-we-wn-xx-xx-xx-de*
Black	*he-hn-hm-xx-we-wn-wm-wf-dn-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-hn-hm-hf-we-wn-wm-wf-dn-de*	*he-hn-hm-xx-we-wn-wm-wf-dn-de*	*he-hn-hm-hf-we-wn-wm-wf-dn-de*	xx-xx-xx-*hf-xx-wn-wm-wf-dn-xx*	xx-*hn-hm-hf-we-wn-wm-wf-dn-xx*	*he-hn-hm-xx-xx-wn-wm-wf-dn-xx*	*he-hn-hm-hf-we-wn-wm-wf-xx-de*	xx-*hn-hm-hf-we-wn-wm-wf-dn-xx*	*he-hn-hm-hf-we-wn-wm-wf-dn-de*
Brazilian	*he-xx-xx-hf-we-wn-wm-wf-dn-de*	*he-hn-hm-hf-we-wn-wm-wf-dn-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-xx-hm-hf-we-wn-wm-wf-xx-de*	*he-hn-hm-hf-we-wn-xx-wf-dn-de*	xx-xx-xx-*hf-we-xx-wm-wf-xx-xx*	*he-xx-xx-xx-we-wn-xx-wf-dn-de*	*he-xx-xx-xx-we-wn-xx-xx-xx-de*	xx-xx-xx-*hf-xx-xx-wm-wf-dn-xx*	*he-xx-hm-hf-we-wn-xx-wf-dn-de*	*he-xx-hm-xx-xx-wn-wm-wf-dn-xx*
Chinese	*he-xx-xx-xx-we-wn-wm-wf-dn-de*	*he-hn-hm-xx-we-wn-wm-wf-dn-de*	*he-xx-hm-hf-we-wn-wm-wf-xx-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-hn-hm-hf-we-wn-wm-xx-dn-de*	*he-xx-xx-hf-xx-wn-xx-wf-xx-de*	*he-xx-hm-hf-we-wn-wm-wf-xx-de*	xx-xx-xx-xx-*we-wn-wm-xx-xx-de*	*he-xx-hm-hf-we-wn-wm-wf-dn-de*	*he-hn-hm-hf-we-xx-wm-wf-dn-de*	*he-xx-hm-hf-we-wn-wm-wf-dn-de*
Indian	*he-hn-hm-hf-we-wn-wm-xx-dn-de*	*he-hn-hm-hf-we-wn-wm-wf-dn-de*	*he-hn-hm-hf-we-wn-xx-wf-dn-de*	*he-hn-hm-hf-we-wn-wm-xx-dn-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-hn-hm-hf-we-wn-wm-wf-dn-de*	*he-hn-hm-hf-we-wn-xx-xx-dn-de*	*he-hn-hm-hf-we-xx-xx-xx-dn-de*	*he-hn-hm-hf-we-wn-wm-xx-dn-de*	*he-hn-hm-hf-we-wn-xx-xx-dn-de*	*he-hn-hm-hf-we-wn-wm-xx-dn-de*
Japanese	*he-xx-xx-hf-xx-wn-wm-wf-xx-xx*	xx-xx-xx-*hf-xx-wn-wm-wf-dn-xx*	xx-xx-xx-*hf-we-xx-wm-wf-xx-xx*	*he-xx-xx-hf-xx-wn-xx-wf-xx-de*	*he-hn-hm-hf-we-wn-wm-wf-dn-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-xx-xx-hf-we-wn-wm-wf-xx-xx*	*he-xx-xx-hf-xx-wn-wm-wf-xx-xx*	xx-xx-xx-*hf-we-xx-wm-wf-dn-xx*	*he-xx-xx-hf-xx-xx-wm-wf-xx-xx*	xx-xx-xx-*hf-we-wn-wm-wf-dn-xx*
Latino	*he-xx-hm-hf-we-wn-wm-xx-xx-de*	xx-*hn-hm-hf-we-wn-wm-wf-dn-xx*	*he-xx-xx-xx-we-wn-xx-wf-dn-de*	*he-xx-hm-hf-we-wn-wm-wf-xx-de*	*he-hn-hm-hf-we-wn-xx-xx-dn-de*	*he-xx-xx-hf-we-wn-wm-wf-xx-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-xx-xx-xx-*wn-xx-xx-xx*	*he-xx-xx-hf-we-wn-wm-xx-dn-de*	xx-xx-*hm-hf-xx-wn-wm-xx-xx-xx*	*he-xx-hm-hf-xx-wn-wm-xx-xx-xx*
Middle Eastern	xx-xx-xx-xx-*we-xx-wm-xx-xx-de*	*he-hn-hm-xx-xx-wn-wm-wf-dn-xx*	*he-xx-xx-xx-we-wn-xx-xx-xx-de*	xx-xx-xx-xx-*we-wn-wm-xx-xx-de*	*he-hn-hm-hf-we-xx-xx-xx-dn-de*	*he-xx-xx-hf-xx-wn-wm-wf-xx-xx*	xx-xx-xx-xx-xx-*wn-xx-xx-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-xx-xx-xx-we-wn-wm-xx-dn-de*	xx-xx-*hm-xx-xx-wn-xx-xx-xx*	*he-xx-hm-xx-we-wn-wm-xx-xx-de*
North African French	*he-xx-hm-hf-we-wn-xx-xx-dn-de*	*he-hn-hm-hf-we-wn-wm-wf-xx-de*	xx-xx-xx-*hf-xx-xx-wm-wf-dn-xx*	*he-xx-hm-hf-we-wn-wm-wf-dn-de*	*he-hn-hm-hf-we-wn-wm-xx-dn-de*	xx-xx-xx-*hf-we-xx-wm-wf-dn-xx*	*he-xx-xx-hf-we-wn-wm-xx-dn-de*	*he-xx-xx-xx-we-wn-wm-xx-dn-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-xx-hm-xx-we-wn-wm-xx-dn-de*	xx-xx-*hm-xx-xx-wn-xx-xx-dn-xx*
White	*he-hn-hm-hf-we-wn-wm-xx-xx-de*	xx-*hn-hm-hf-we-wn-wm-wf-dn-xx*	*he-xx-hm-hf-we-xx-wn-wf-dn-de*	*he-hn-hm-hf-we-xx-wm-wf-dn-de*	*he-hn-hm-hf-we-wn-xx-xx-dn-de*	*he-xx-xx-hf-xx-xx-wm-wf-xx-xx*	xx-xx-*hm-hf-xx-wn-wm-xx-xx-xx*	xx-xx-*hm-xx-xx-wn-xx-xx-xx*	*he-xx-hm-xx-we-wn-wm-xx-dn-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-xx-xx-xx-we-wn-wm-xx-xx-de*
White French	*he-xx-hm-hf-we-wn-xx-xx-xx-de*	*he-hn-hm-hf-we-wn-wm-wf-dn-de*	*he-xx-hm-xx-xx-wn-wm-wf-dn-xx*	*he-xx-hm-hf-we-wn-wm-wf-dn-de*	*he-hn-hm-hf-we-wn-wm-xx-dn-de*	xx-xx-xx-*hf-we-wn-wm-wf-dn-xx*	*he-xx-hm-hf-xx-wn-wm-xx-xx-xx*	*he-xx-hm-xx-we-wn-wm-xx-xx-de*	xx-xx-*hm-xx-xx-wn-xx-xx-dn-xx*	*he-xx-xx-xx-we-wn-wm-xx-xx-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx

TABLE 27. Welch's t-test results on the raw feature set for the major ethnic groups.

Major ethnic groups	*ASIAN*	*BLACK*	*LATINO*	*MIDDLE EAST*	*WHITE*
ASIAN	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-hn-hm-xx-we-wn-wm-wf-dn-de*	*he-xx-hm-hf-we-wn-wm-xx-xx-de*	*he-xx-hm-hf-we-wn-wm-wf-dn-de*	*he-xx-hm-hf-we-wn-wm-wf-dn-de*
BLACK	*he-hn-hm-xx-we-wn-wm-wf-dn-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-*hn-hm-hf-we-wn-wm-wf-dn-de*	xx-*hn-hm-hf-we-wn-wm-wf-dn-de*	xx-*hn-hm-hf-we-wn-wm-wf-dn-xx*
LATINO	*he-xx-hm-hf-we-wn-wm-xx-xx-de*	xx-*hn-hm-hf-we-wn-wm-wf-dn-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-xx-*hf-xx-xx-wm-wf-dn-xx*	xx-xx-*hm-hf-we-wn-wm-wf-dn-de*
MIDDLE EAST	*he-xx-hm-hf-we-wn-wm-wf-dn-de*	xx-*hn-hm-hf-we-wn-wm-wf-dn-de*	xx-xx-xx-*hf-xx-xx-wm-wf-dn-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-*hm-xx-we-xx-xx-xx-dn-xx*
WHITE	*he-xx-hm-hf-we-wn-wm-wf-dn-de*	xx-*hn-hm-hf-we-wn-wm-wf-dn-xx*	xx-xx-*hm-hf-we-wn-wm-wf-dn-de*	xx-xx-*hm-xx-we-xx-xx-xx-dn-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx

learned features includes the hair, forehead, and neck. These features are disregarded in the handcrafted features as during the preprocessing phase, a bounding box around the detected face is fed to the feature extraction method.

Of note, both the XGBoost and SKB feature selection algorithms gave similar prediction results with the raw feature set in the major and sub-ethnic groups. However, using the normalized feature set, information-gain-based features yielded the best prediction results in both the groups. Hence, we have demonstrated that machine learning can be used to obtain ethnicity-based reference ranges for the features. The current healthcare literature [1], [6]–[9], [48] uses statistical tests, as indicated in Table 1, to obtain the reference ranges.

The top three information-gain-based features in the sub-ethnic groups are: *dn* (distance from the tip of the nose to the center of the mouth), *hf* (face height) and *wn* (nose width), while the top three information-gain-based features in the major ethnic groups are: *de* (distance between the inner corners of the eyelids), *hf* and *dn*. Of note, the researchers in [50] observed that Indian American females have a smaller *de* but a larger *wn* than North American White females. On the other hand, the study in [51] revealed that African American females have a longer lower face and wider nose than the North American Caucasian females. However, to the best of our knowledge, we are not aware of any ethnicity studies that use the nose's tip to the mouth's center (*dn*) as a differentiating feature. Nevertheless, such a measurement has been used by plastic surgeons to calculate the golden ratio of the human face [52].

In this paper, we have 1) provided the first evidence that two-dimensional benchmark data sets developed in the computer vision field can be used in facial anthropometric measurement studies of the healthcare field; 2) provided a comprehensive analysis of information-gain-based and p-value-based features to find the k top discriminative features in major and sub-ethnic groups; and 3) proposed the top three features that may be useful in differentiating populations in major and sub-ethnic groups and compared them to the features obtained automatically from standard deep learning techniques.

TABLE 28. Welch's t-test results on the normalized feature set for the sub-ethnic groups.

Sub-ethnic groups	Asian	Black	Brazilian	Chinese	Indian	Japanese	Latino	Middle Eastern	North African French	White	White French
Asian	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-hn-hm-xx-we-wn-wm-wf-dn-xx*	*he-hn-hm-hf-we-wn-xx-wf-dn-xx*	*he-hn-hm-hf-we-wn-wm-wf-dn-xx*	*he-hn-hm-hf-we-wn-wm-xx-xx-xx*	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	*he-hn-hm-xx-we-xx-wm-wf-dn-xx*	xx-xx-hm-xx-we-xx-xx-wf-dn-xx	*he-xx-hm-hf-we-wn-wm-xx-dn-xx*	*he-hn-hm-hf-we-wn-wm-wf-dn-xx*	*he-xx-hm-hf-we-wn-wm-wf-dn-xx*
Black	*he-hn-hm-xx-we-wn-wm-wf-dn-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-hn-xx-hf-xx-wn-wm-wf-dn-xx	*he-hn-hm-xx-xx-wn-wm-wf-dn-xx*	*he-hn-hm-hf-xx-wn-wm-wf-dn-xx*	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	*he-hn-hm-xx-xx-wn-xx-wf-xx-xx*	xx-xx-xx-xx-xx-wn-xx-xx-xx	xx-hn-xx-hf-xx-wn-wm-xx-xx-xx	xx-xx-hm-hf-we-wn-wm-wf-dn-xx	*he-hn-xx-hf-xx-wn-xx-xx-dn-xx*
Brazilian	*he-hn-hm-hf-we-wn-xx-wf-dn-xx*	xx-hn-xx-hf-xx-wn-wm-wf-dn-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-xx-xx-hf-xx-xx-wm-xx-dn-xx*	*he-hn-hm-hf-xx-wn-wm-wf-dn-xx*	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	*he-hn-xx-hf-xx-xx-wm-wf-dn-xx*	xx-xx-xx-xx-xx-xx-xx-dn-xx	xx-hn-xx-xx-xx-wn-wm-wf-xx-xx	xx-hn-hm-xx-we-wn-xx-xx-xx	*he-hn-xx-xx-xx-xx-wm-xx-dn-xx*
Chinese	*he-hn-hm-hf-we-wn-wm-wf-dn-xx*	*he-hn-hm-xx-xx-wn-wm-wf-dn-xx*	*he-xx-xx-hf-xx-xx-wm-xx-dn-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-hn-hm-hf-xx-wn-xx-wf-dn-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-hn-xx-xx-we-xx-xx-wf-dn-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-hn-xx-hf-xx-wn-xx-wf-dn-xx	*he-hn-hm-hf-we-wn-wm-wf-dn-xx*	*he-hn-xx-hf-xx-wn-wm-xx-xx-xx*
Indian	*he-hn-hm-hf-we-wn-wm-xx-xx-xx*	*he-hn-hm-hf-xx-wn-wm-wf-dn-xx*	*he-hn-hm-hf-xx-wn-wm-wf-xx-xx*	*he-hn-hm-hf-xx-wn-xx-wf-dn-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-hm-xx-xx-xx-xx-wf-xx-xx	xx-xx-hm-hf-xx-wn-xx-xx-dn-xx	xx-hn-hm-hf-xx-xx-xx-wf-dn-xx	*he-xx-hm-hf-xx-xx-xx-xx-xx-xx*	*he-hn-hm-hf-we-wn-wm-wf-xx-xx*	xx-xx-hm-hf-xx-xx-wm-wf-dn-xx
Japanese	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-*hm*-xx-xx-xx-xx-wf-xx-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx
Latino	*he-hn-hm-xx-we-xx-wm-wf-dn-xx*	*he-hn-hm-xx-xx-wn-xx-wf-xx-xx*	*he-hn-xx-hf-xx-xx-wm-wf-dn-xx*	*he-hn-xx-xx-we-xx-xx-wf-dn-xx*	xx-xx-*hm*-hf-xx-wn-xx-xx-dn-xx	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-*hn*-xx-xx-xx-xx-xx-wf-xx-xx	*he-hn-xx-hf-xx-wn-xx-wf-dn-xx*	*he-hn-xx-hf-we-wn-wm-wf-dn-xx*	xx-hn-xx-hf-xx-wn-xx-wf-xx-xx
Middle eastern	xx-xx-*hm*-xx-we-xx-xx-wf-dn-xx	xx-xx-xx-xx-xx-xx-wn-xx-xx-xx	xx-xx-xx-xx-xx-xx-xx-xx-dn-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-*hn*-hm-hf-xx-xx-xx-wf-dn-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-*hn*-xx-xx-xx-xx-xx-wf-xx-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-xx-xx-we-xx-xx-xx-dn-xx	xx-xx-xx-*hf*-xx-xx-xx-xx-xx
North African French	*he-xx-hm-hf-we-wn-wm-xx-dn-xx*	xx-*hn*-xx-hf-xx-wn-wm-xx-xx-xx	xx-*hn*-xx-xx-xx-wn-wm-wf-xx-xx	xx-*hn*-xx-hf-xx-wn-xx-wf-dn-xx	*he-xx-hm-hf-*xx-xx-xx-xx-xx	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	*he-hn-xx-hf-*xx-wn-xx-wf-dn-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-*hn-hm*-xx-we-wn-wm-xx-xx-xx	*he-xx-xx-xx-we-xx-xx-xx-dn-xx*
White	*he-hn-hm-hf-we-wn-wm-wf-dn-xx*	xx-xx-*hm-hf-we-wn-wm-wf-dn-xx*	xx-*hn-hm*-xx-we-wn-xx-xx-xx-xx	*he-hn-hm-hf-we-wn-wm-wf-dn-xx*	*he-hn-hm-hf-we-wn-wm-wf-*xx-xx	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	*he-hn-xx-hf-we-wn-wm-wf-dn-xx*	xx-xx-xx-xx-we-xx-xx-xx-dn-xx	xx-*hn-hm*-xx-we-wn-wm-xx-xx-xx	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	*he-hn-xx-xx-we-xx-xx-xx-dn-xx*
White French	*he-xx-hm-hf-we-wn-wm-wf-dn-xx*	*he-hn-xx-hf-*xx-wn-xx-xx-*dn-xx*	*he-hn-xx-xx-*xx-xx-*wm-xx-dn-xx*	*he-hn-xx-hf-*xx-wn-wm-xx-xx-xx	xx-xx-*hm-hf-*xx-xx-*wm-wf-dn-xx*	xx-xx-xx-xx-xx-xx-xx-wf-xx-xx	xx-*hn-xx-hf-*xx-wn-xx-wf-xx-xx	xx-xx-xx-*hf-*xx-xx-xx-xx-xx	*he-xx-xx-xx-*xx-xx-xx-xx-*dn-xx*	*he-hn-xx-xx-we-xx-xx-xx-xx-dn-xx*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx

TABLE 29. Welch's t-test results on the normalized feature set for the major ethnic groups.

Major ethnic groups	ASIAN	BLACK	LATINO	MIDDLE EAST	WHITE
ASIAN	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-xx-*hf-we-*xx-xx-*wf-dn-de*	*he-hn-hm-hf-we-wn-*xx-*wf-dn-de*	xx-*hn*-xx-*hf-we-wn-wm-wf-dn-*xx	*he-hn-hm-hf-we-wn-wm-wf-dn-de*
BLACK	xx-xx-xx-*hf-we-*xx-xx-*wf-dn-de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-*hn-hm-hf-*xx-*wn-*xx-xx-xx-*de*	xx-*hn*-xx-*hf-*xx-*wn-wm-*xx-xx-*de*	xx-xx-*hm-hf-we-wn-wm-wf-dn-de*
LATINO	*he-hn-hm-hf-we-wn-*xx-*wf-dn-de*	xx-*hn-hm-hf-*xx-*wn-*xx-xx-xx-*de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-xx-xx-xx-*wn-wm-wf-*xx-*de*	xx-*hn-hm-*xx-*we-wn-wm-wf-dn-de*
MIDDLE EAST	xx-*hn*-xx-*hf-we-wn-wm-wf-dn-*xx	xx-*hn*-xx-*hf-*xx-*wn-wm-*xx-xx-*de*	xx-xx-xx-xx-xx-*wn-wm-wf-*xx-*de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx	xx-xx-*hm-*xx-*we-wn-wm-*xx-xx-*de*
WHITE	*he-hn-hm-hf-we-wn-wm-wf-dn-de*	xx-xx-*hm-hf-we-wn-wm-wf-dn-de*	xx-*hn-hm-*xx-*we-wn-wm-wf-dn-de*	xx-xx-*hm-*xx-*we-wn-wm-*xx-xx-*de*	xx-xx-xx-xx-xx-xx-xx-xx-xx-xx

APPENDIX

A. DESCRIPTIVE STATISTICS OF THE RAW FEATURE SET IN THE SUB-ETHNIC GROUPS. POPULATION SIZE IS INDICATED IN BRACKETS IN SUB-ETHNIC GROUPS COLUMN

See Table 24.

B. DESCRIPTIVE STATISTICS OF THE RAW FEATURE SET IN THE MAJOR ETHNIC GROUPS. POPULATION SIZE IS INDICATED IN BRACKETS IN MAJOR ETHNIC GROUPS COLUMN

See Table 25.

C. WELCH's T-TEST RESULTS ON THE RAW FEATURE SET FOR THE SUB-ETHNIC GROUPS

See Table 26.

D. WELCH's T-TEST RESULTS ON THE RAW FEATURE SET FOR THE MAJOR ETHNIC GROUPS

See Table 27.

E. WELCH's T-TEST RESULTS ON THE NORMALIZED FEATURE SET FOR THE SUB-ETHNIC GROUPS

See Table 28.

F. WELCH's T-TEST RESULTS ON THE NORMALIZED FEATURE SET FOR THE MAJOR ETHNIC GROUPS

See Table 29.

ACKNOWLEDGMENT

The authors thank Kok Sheik Wong for fruitful discussions.

REFERENCES

[1] S. T. Ozdemir, D. Sigirli, I. Ercan, and N. S. Cankur, "Photographic facial soft tissue analysis of healthy Turkish young adults: Anthropometric measurements," *Aesthetic Plastic Surg.*, vol. 33, no. 2, pp. 175–184, Mar. 2009.

[2] P. S. Gonzales, C. E. P. Machado, and E. Michel-Crosato, "Photoanthropometry of the face in the young white Brazilian population," *Brazilian Dental J.*, vol. 29, no. 6, pp. 619–623, Dec. 2018.

[3] S. K. Jilani, H. Ugail, A. M. Bukar, A. Logan, and T. Munshi, "A machine learning approach for ethnic classification: The British Pakistani face," in *Proc. Int. Conf. Cyberworlds (CW)*, Sep. 2017, pp. 170–173.

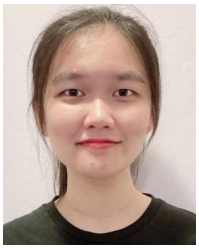
[4] Z. Akhter, M. Banu, M. Alam, S. Hossain, and M. Nazneen, "Photo-anthropometric study on face among Garo adult females of Bangladesh," *Bangladesh Med. Res. Council Bull.*, vol. 39, no. 2, pp. 61–64, Jul. 2014.

[5] C. Eliakim-Ikechukwu, A. S. Ekpo, M. Etika, C. Ihentuge, and O. Mesembe, "Facial aesthetic angles of the Ibo and Yoruba ethnic groups of Nigeria," *IOSR J. Pharmacy Biol. Sci.*, vol. 5, no. 5, pp. 14–17, Jan. 2013.
[6] M. A. Ferdousi, A. A. Mamun, L. A. Banu, and S. Paul, "Angular photogrammetric analysis of the facial profile of the adult Bangladeshi Garo," *Adv. Anthropol.*, vol. 3, no. 4, pp. 188–192, 2013.
[7] S. Anicy-Milosevicy, M. Lapter-Varga, and M. Slaj, "Analysis of the soft tissue facial profile by means of angular measurements," *Eur. J. Orthodontics*, vol. 30, pp. 135–140, Apr. 2008.
[8] U. U. Ukoha, O. O. Udemezue, C. K. Oranusi, A. L. Asomugha, U. Dimkpa, and L. C. Nzeukwu, "Photometric facial analysis of the Igbo Nigerian adult male," *Nigerian Med. J.*, vol. 53, no. 4, pp. 240–244, 2012.
[9] A. Sepehr, P. J. Mathew, J.-P. Pepper, K. Karimi, Z. Devcic, and A. M. Karam, "The Persian woman's face: A photogrammetric analysis," *Aesthetic Plastic Surg.*, vol. 36, no. 3, pp. 687–691, Jun. 2012.
[10] S. Hosoi, E. Takikawa, and M. Kawade, "Ethnicity estimation with facial images," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Jun. 2004, pp. 195–200.
[11] X. Lu and A. K. Jain, "Ethnicity identification from face images," *Proc. SPIE, Int. Soc. Opt. Eng.*, vol. 5404, pp. 114–123, May 2004.
[12] Y. Ou, X. Wu, H. Qian, and Y. Xu, "A real time race classification system," in *Proc. IEEE Int. Conf. Inf. Acquisition*, Jun. 2005, p. 6.
[13] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Dec. 2014.
[14] 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.
[15] R. Achkar, G. Haidar, M. El Assal, D. Habchy, D. Al Ashi, and T. Maylaa, "Ethnicity recognition system using back propagation algorithm of an MLP," in *Proc. 4th Int. Conf. Adv. Comput. Tools Eng. Appl. (ACTEA)*, Jul. 2019, pp. 1–5.
[16] S. K. Jilani, H. Ugail, A. M. Bukar, and A. Logan, "On the ethnic classification of Pakistani face using deep learning," in *Proc. Int. Conf. Cyberworlds (CW)*, Oct. 2019, pp. 191–198.
[17] S. A. Milošević, M. L. Varga, and M. Šlaj, "Analysis of the soft tissue facial profile of croatians using of linear measurements," *J. Craniofacial Surg.*, vol. 19, no. 1, pp. 251–258, Jan. 2008.
[18] A. J. Naimi, S. Bolourian, M. Mohammadzadeh, M. Farahmand, F. Ghanbari, and S. Samiee, "Investigating the relationship between major Thalassemia diseases with anthropometric sizes of head and facial soft tissue," *Biosci. Biotechnol. Res. Commun.*, vol. 10, pp. 233–240, 2017.
[19] D. Normando, P. Lima da Silva, and A. M. Mendes, "A clinical photogrammetric method to measure dental arch dimensions and mesio-distal tooth size," *Eur. J. Orthodontics*, vol. 33, no. 6, pp. 721–726, Dec. 2011.
[20] L. Guyot, M. Dubuc, O. Richard, N. Philip, and O. Dutour, "Comparison between direct clinical and digital photogrammetric measurements in patients with 22q11 microdeletion," *Int. J. Oral Maxillofacial Surgery*, vol. 32, no. 3, pp. 246–252, Jun. 2003.
[21] W. C. Ngeow and S. T. Aljunid, "Craniofacial anthropometric norms of Malays," *Singap. Med. J.*, vol. 50, no. 5, pp. 525–528, 2009.
[22] L. G. Farkas, O. G. Eiben, S. Sivkov, B. Tompson, M. J. Katic, and C. R. Forrest, "Anthropometric measurements of the facial framework in adulthood: Age-related changes in eight age categories in 600 healthy white north Americans of European ancestry from 16 to 90 years of age," *J. Craniofacial Surg.*, vol. 15, no. 2, pp. 288–298, Mar. 2004.
[23] G. D. Kilci, E. Başer, A. Verim, Ö. F. Çalim, B. Veyseller, O. Özturan, A. Altintaş, and M. Çelik, "Outcomes of external septorhinoplasty in a turkish male population," *Brazilian J. Otorhinolaryngol.*, vol. 84, no. 4, pp. 426–434, Jul. 2018.
[24] H. N. D. R. Fortes, T. C. Guimarães, I. M. L. Belo, and E. N. R. D. Matta, "Photometric analysis of esthetically pleasant and unpleasant facial profile," *Dental Press J. Orthodontics*, vol. 19, no. 2, pp. 66–75, Apr. 2014.
[25] S. Fu, H. He, and Z.-G. Hou, "Learning race from face: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2483–2509, Dec. 2014.
[26] L. G. Farkas, M. J. Katic, and C. R. Forrest, "International anthropometric study of facial morphology in various ethnic groups/races," *J. Craniofacial Surg.*, vol. 16, no. 4, pp. 615–646, Jul. 2005.
[27] R. Courset, M. Rougier, R. Palluel-Germain, A. Smeding, J. M. Jonte, A. Chauvin, and D. Müller, "The Caucasian and north African French faces (CaNAFF): A face database," *Int. Rev. Social Psychol.*, vol. 31, no. 1, pp. 1–10, Jul. 2018.
[28] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.
[29] D. S. Ma, J. Correll, and B. Wittenbrink, "The Chicago face database: A free stimulus set of faces and norming data," *Behav. Res. methods*, vol. 47, pp. 1122–1135, Jan. 2015.
[30] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.
[31] 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, 1998.
[32] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Jun. 1998, pp. 200–205.
[33] N. Strohminger, K. Gray, V. Chituc, J. Heffner, C. Schein, and T. B. Heagins, "The MR2: A multi-racial, mega-resolution database of facial stimuli," *Behav. Res. Methods*, vol. 48, no. 3, pp. 1197–1204, Sep. 2016.
[34] S. Gupta, K. R. Castleman, M. K. Markey, and A. C. Bovik, "Texas 3D face recognition database," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation (SSIAI)*, May 2010, pp. 97–100.
[35] C. E. Thomaz. (2012). *Fei Face Database*. Accessed: Apr. 12, 2020. [Online]. Available: https://fei.edu.br/~cet/facedatabase.html
[36] The Brain Mapping Laboratory (National Yang-Ming University) and Integrated Brain Research Unit (Taipei Veterans General Hospital). (2007). *Taiwanese Facial Expression Image Database (TFEID)*. Accessed: Apr. 12, 2020. [Online]. Available: http://bml.ym.edu.tw/tfeid/
[37] Z. Lazarus, S. Gupta, and N. Panda, "An Indian facial database highlighting the spectacle problem," in *Proc. IEEE Int. Conf. Innov. Technol. Eng.*, Hyderabad, India, Apr. 2018.
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds. USA, 2015.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.
[41] D. Riccio, G. Tortora, M. D. Marsico, and H. Wechsler, "EGA—Ethnicity, gender and age, a pre-annotated face database," in *Proc. IEEE Workshop Biometric Meas. Syst. Secur. Med. Appl. (BIOMS)*, Sep. 2012, pp. 1–8.
[42] R. Bhopal and L. Donaldson, "White, European, Western, Caucasian, or what? Inappropriate labeling in research on race, ethnicity, and health," *Amer. J. Public Health*, vol. 88, pp. 1303–1307, Oct. 1998.
[43] Webster. (2020). *Merriam-Webster*. Accessed: Jun. 12, 2020. [Online]. Available: https://www.merriam-webster.com/dictionary/ethnic
[44] C.-H. Chen *et al.*, "Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan biobank project," *Hum. Mol. Genet.*, vol. 25, no. 24, pp. 5321–5331, Dec. 2016.
[45] C. S. Lin, R. Shaari, M. K. Alam, and S. A. Rahman, "Photogrammetric analysis of nasolabial angle and mentolabial angle norm in Malaysian adults," *Bangladesh J. Med. Sci.*, vol. 12, no. 2, pp. 209–214, 2013.
[46] C. Wang, Q. Zhang, W. Liu, Y. Liu, and L. Miao, "Facial feature discovery for ethnicity recognition," *WIREs, Data Mining Knowl. Discovery*, vol. 9, no. 1, p. e1278, Aug. 2018.
[47] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.
[48] P. Roe, K. Runcharassaeng, J. Y. K. Kan, R. D. Patel, W. V. Campagni, and J. S. Brudvik, "The influence of upper lip length and lip mobility on maxillary Incisal exposure," *Amer. J. Esthetic Dentistry*, vol. 2, pp. 116–125, Jun. 2012.
[49] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. KDD*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144.
[50] O. F. Husein, A. Sepehr, R. Garg, M. Sina-Khadiv, S. Gattu, J. Waltzman, E. C. Wu, M. Shieh, G. M. Heitmann, and S. E. Galle, "Anthropometric and aesthetic analysis of the Indian American woman's face," *J. Plastic, Reconstructive Aesthetic Surg.*, vol. 63, no. 11, pp. 1825–1831, Nov. 2010.

[51] J. P. Porter and K. L. Olson, "Anthropometric facial analysis of the African American woman," *Arch. Facial Plastic Surg.*, vol. 3, pp. 191–197, Jul. 2001.
[52] A. Reddy. (2020). *The Golden Ratio of a Beautiful Face*. Accessed: Sep. 14, 2020. [Online]. Available: https://www.medisculpt.co.za/golden-ratio-beautiful-face/

JASBIR DHALIWAL received the B.Sc. degree (Hons.) in computer science from the University of Malaya, Malaysia, in 2004, and the M.Sc. degree in applied science and the Ph.D. degree from the Royal Melbourne Institute of Technology (RMIT) University, Australia, in 2008 and 2013, respectively. She has worked as a Software Engineer with Motorola, Malaysia. She was a Postdoctoral Researcher with IBM Research - Australia. She was a Data Scientist with FTI Consulting Australia. She is currently a Lecturer with Monash University Malaysia. Her research interests include bioinformatics, computer vision, machine learning, deep learning, and string related algorithms.

JOHN WAGNER (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in applied mathematics from the University of California at Davis. He worked with the Institute of Theoretical Dynamics, where worked on cellular calcium regulation (computational cell biology) with Joel Keizer. He received an NIH NRSA Postdoctoral Fellowship, working with Les Loew at the National Resource for Cellular Analysis and Modeling, University of Connecticut Health Center. He is currently a Research Staff Member with the Blockchain Research and Development Team, IBM Research-Australia. He is also the Former Manager of the IBM Research Collaboratory for Life Sciences-Melbourne. His research interests include blockchain, cancer biology, modeling, simulation, and biological applications for high-performance computing.



SHU LING LEONG received the B.Sc. degree (Hons.) in computer science from Monash University Malaysia, in 2019. She was a Research Assistant attached to the project. Her research interest includes computer vision.

CHERN HONG LIM (Member, IEEE) received the B.Sc. (Hons) and Ph.D. degrees in computer science majoring in artificial intelligence from the University of Malaya, in 2010 and 2015, respectively. He is currently with the School of Information Technology, Monash University Malaysia. His research interests include artificial intelligence, machine learning, computer vision, image processing, and fuzzy logic, with focus on image or video content analysis and human motion inference. He is an Active Member. He also serves as the Chair of the IEEE Computational Intelligence Society (CIS) Malaysia Chapter.