KurdFace-1000: A New Multi-Attribute Dataset for Facial Beauty Prediction Using Multi-Task Learning

Ali H. Ibrahem[1†] and Adnan M. Abdulazeez[2]

[1]Department of Information Technology, Technical College of Informatics, Akre University for Applied Sciences, Akre, Kurdistan region – F.R. Iraq

[2]Department of IT, Technical College of Engineering, Duhok Polytechnic University, Duhok, Kurdistan region – F.R. Iraq

***Abstract*—In response to the demographic bias commonly observed in facial beauty prediction (FBP) models due to the underrepresentation of certain ethnic groups, KurdFace-1000 is introduced as a novel and balanced dataset developed as a field of research community. This dataset comprises 1000 color facial images of individuals from the Kurdistan Region of Iraq, evenly distributed across gender (500 male, 500 female) and facial expression (500 smiling, 500 non-smiling), and includes both frontal and profile views. Within the results of males, 200 are smiled and 300 are non-smiled and for the results of females, 300 are smiled and 200 are non-smiled. Moreover, each image is annotated with three key attributes: A beauty score on a [1–5] scale rated by five independent human raters, binary gender, and smile expression. KurdFace-1000 is the first dataset specifically designed to represent Kurdish facial features in FBP tasks, aiming to reduce ethnic bias and improve model performance for underrepresented populations. With a balanced structure and diverse annotations, the dataset supports various computational paradigms, including classification and regression, and serves as a critical step toward building culturally aware and inclusive deep learning models in FBP.**

***Index Terms*—Beauty regression, Demographic diversity, Emotion recognition, Facial dataset, Kurdish faces.**

I. Introduction

Facial beauty prediction (FBP) has become an increasingly important topic in computer vision and affective computing, driven by its potential applications in social media, cosmetic planning, entertainment, and personalized recommendation systems (Fan, et al., 2019; Zhang, et al., 2022; Zhuang, et al., 2018; Boukhari, et al., 2023). With the advent of deep learning, particularly convolutional neural networks (CNNs), significant progress has been made in automating aesthetic assessment of human faces (Saeed and Abdulazeez, 2021; Ibrahem and Abdulazeez, 2025). However, most existing FBP models are trained on datasets that exhibit notable demographic imbalances, especially in terms of ethnicity, gender, and age, resulting in biased predictions and limited generalizability to underrepresented populations (Krishnan, et al., 2020; Khalil, et al., 2020).

In recent years, several benchmark datasets have been introduced to facilitate FBP research. The SCUT-FBP5500 dataset (Xie, et al., 2015; Liang, et al., 2018) is one of the most widely used resources and includes 5500 facial images labeled with beauty scores from multiple raters, along with gender and ethnicity labels. However, its ethnic diversity is limited primarily to Asian and Caucasian subjects. Similarly, the HotOrNot dataset (Kagian, et al., 2008) relies on crowdsourced ratings of online celebrity photos, which often include social media artifacts and exhibit Western-centric biases. The Beauty 799 and Beauty 639 datasets also lack controlled demographic balance and are limited in scope, typically focusing on frontal female faces (Liu, et al., 2019). These limitations have prompted researchers to search for the creation of culturally inclusive datasets to ensure fairness in FBP systems (Raji, et al., 2020; Danner, et al., 2023). To mitigate such concerns, recent studies have begun incorporating multi-task learning (MTL) approaches that predict beauty scores alongside auxiliary attributes such as gender and smile (Gao, et al., 2018). MTL not only improves model performance through shared feature representations but also enriches facial understanding by simultaneously learning correlated tasks. For instance, (Vahdati and Suen, 2020) proposed an MTL framework that jointly predicts beauty, gender, and age using SCUT-FBP5500, demonstrating improved robustness. However, even in such studies, the training data lacks ethnic inclusivity and does not address fairness concerns from a representational standpoint. In response to this significant research gap, this study presents KurdFace-1000, a novel and demographically focused facial dataset developed to represent Kurdish individuals as an ethnic group largely absent from existing computer vision benchmarks. This dataset includes 1000 color images from individuals residing in the Kurdistan Region of Iraq and is meticulously balanced across gender (500 male, 500 female) and facial expression (500 smiling, 500 non-smiling). Different from the previously available datasets,

ARO-The Scientific Journal of Koya University
Vol. XIV, No.1 (2026), Article ID: ARO.12393. 11 pages
DOI: 10.14500/aro.12393
Received: 30 June 2025; Accepted: 19 October 2025
Regular research paper; Published: 20 January 2026
†Corresponding author's e-mail: ali.hikmat@auas.edu.krd

KurdFace-1000 includes both frontal and profile facial views captured in natural, unconstrained environments, as well as a realistic range of male facial hair styles (e.g., beards and mustaches). The above characteristics are important in the enhancement of intra-ethnic variability and better reflect real-world conditions. Each image is annotated with three essential attributes: (1) A continuous beauty score on a 5-point Likert scale (Likert, 1932), rated by five independent human raters, (2) binary gender, and (3) smile expression. To provide a visual overview of the dataset composition and diversity, Fig. 1 displays a selection of sample images from the KurdFace-1000 dataset, illustrating variations in gender, expression (smile vs. non-smile), and viewpoint (frontal vs. profile). The above examples reflect the balanced structure of the dataset and help emphasize its utility for developing MTL models that generalize across intra-ethnic variations successfully.

As tabulated in Table I, KurdFace-1000 stands apart from other facial beauty datasets by allowing them the first to target an underrepresented ethnic group (Kurdish), by enabling MTL setups, and by prioritizing balance across key facial attributes. The dataset is not publicly available due to regional privacy and cultural considerations but is structured to support both regression and classification paradigms in facial analysis.

To evaluate the utility of KurdFace-1000, this study adopts an MTL framework built upon the EfficientNetV2B0 architecture, which is fine-tuned to simultaneously perform three tasks: beauty score regression, gender classification, and smile recognition. The network comprises shared convolutional layers followed by three task-specific branches.

The main fundamental contributions of this study are outlined as follows:

1. A demographically novel dataset: KurdFace-1000 is the first facial beauty dataset focused on Kurdish individuals,

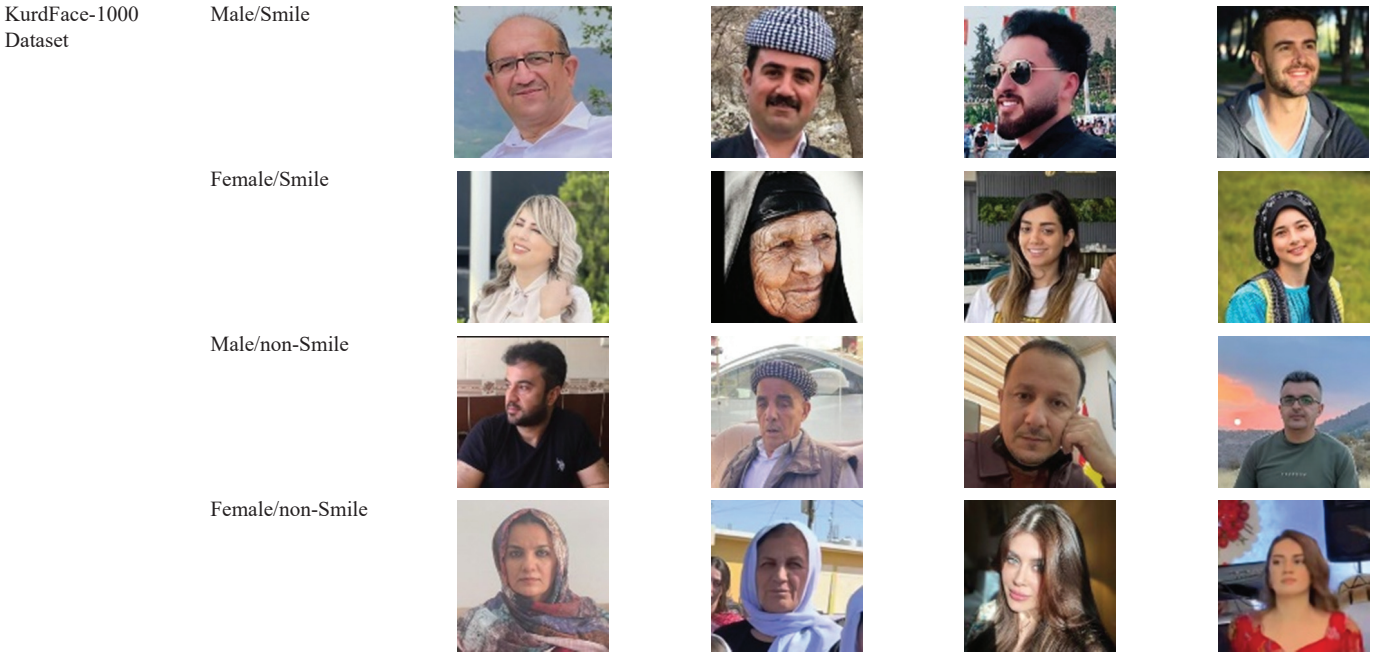

Fig. 1. Selected sample images from the KurdFace-1000 dataset.

TABLE I
Representative Database for Facial Beauty Prediction

Database	Image Number	Labelers/ image	Beauty class	Face property	Face landmarks	Publicly available
(Chen and Zhang, 2010) Beauty Face Database	23412	Unknown	2	Asian male/female	Available	Not available
(Fan, et al., 2012) Chicago face database	432	30	7	Generated female	Available	Not available
(Gunes and Piccardi, 2006) taken from the internet	48	46	10	Female	Available	Not available
(Liu, et al., 2015) CelebA (CelebFaces Attributes Dataset)	202,599	not for beauty scores	No beauty scores included	40 binary facial attributes: gender, smile, age, etc…	Available	Available
(Xie, et al., 2015) SCUT-FBP	500	70	5	Asian Female	Available	Available
(Liang, et al., 2018) SCUT-FBP5500	5500	60	5	Asian and Caucasian: Male and Female	Available	Available
(Abual-Rub, et al.) Arab face	2619	not for beauty scores	No beauty scores included	Arab Gulf States, Egypt, Levant, Maghreb, and North and East Arab African Countries	Not available	Not available
KurdFace-1000	1000	5	5 (rated 1–5)	Gender (M/F), Smile (Yes/No), Kurdish only	Not available	Not available

addressing the ethnic underrepresentation in existing FBP datasets.
2. Balanced and richly annotated data: The dataset is carefully balanced across gender and smile attributes and includes beauty scores from five raters, ensuring robust and interpretable annotations.
3. Support for MTL: KurdFace-1000 enables simultaneous training on beauty regression and attribute classification (gender and smile), aligning with current trends in MTL for facial analysis.
4. Resource for fair and inclusive AI: By introducing an underrepresented population into the FBP space, this dataset encourages the research field on culturally sensitive and fair model development.
5. Baseline structure for future expansion: This dataset represents a foundation for future extensions, including additional ethnic groups and age ranges, with the goal of creating a broader MultiCultural Beauty DB.

Despite its advantages, KurdFace-1000 has a few limitations. It currently lacks age annotations, which restricts age-conditioned beauty analysis. Furthermore, approximately 20% of images are side views, which may affect models sensitive to posing variations. Moreover, due to local cultural constraints, facial image sharing in Kurdistan is limited, reducing potential diversity in expressions and environments. While existing FBP studies have leveraged large datasets and deep learning techniques, they overwhelmingly focus on Western and East Asian populations. There is a clear research gap in datasets that offer cultural specificity and demographic balance, particularly for Middle Eastern. This lack of representation not only skews model predictions but also perpetuates exclusion in AI systems designed to be universally applicable (Zhang, et al., 2022; Buolamwini and Gebru, 2018; Zhao, et al., 2017). This study proposes KurdFace-1000 as a balanced, demographically specific, and multi-task enabled facial image dataset designed to support fair and inclusive artificial intelligence applications in FBP, gender classification, and smile recognition. By focusing on Kurdish faces, an ethnic group significantly underrepresented in current benchmarks, this dataset aims to mitigate cultural and demographic biases commonly found in existing models. Furthermore, the integration of MTL capabilities allows for joint optimization of aesthetic and attribute-based predictions, enhancing both performance and interpretability. KurdFace-1000 ultimately contributes to the development of culturally aware, ethically responsible, and technically robust deep learning systems in the facial analysis domain.

To demonstrate the utility and significance of the proposed dataset, the manuscript is organized as follows. Section 2 presents the constitution and annotation protocol of the KurdFace-1000 dataset, highlighting its demographic balance and multi-attribute structure. Section 3 describes the statistical analyses performed to assess annotation consistency and potential biases related to gender and facial expressions and outlines the proposed MTL framework built on the EfficientNetV2B0 architecture and details the experimental setup. Section 4 reports the performance results of both single-task and multi-task models across beauty prediction, gender classification, and smile recognition tasks. This is followed by a discussion of the findings, including a comparative evaluation of learning paradigms and the implications of MTL. Section 5 addresses key limitations of the dataset and the study, and Section 6 concludes with a summary of contributions and suggestions for future research directions.

II. Constitution of KurdFace-1000 Dataset

A. Dataset Overview

The KurdFace-1000 dataset is a recently curated resource specifically developed to mitigate the demographic imbalance observed in the current FBP field. It comprises 1,000 colored images of individuals from the Iraqi Kurdistan region, with a careful attention to balance and diversity. The dataset is evenly distributed across two binary attributes: Gender with 500 images each for male and female subjects and facial expression (smiling), comprising 500 smiling and 500 non-smiling images. The dataset includes both frontal and profile views, reflecting natural diversity in facial pose and orientation. All images are in RGB-colored format and were manually collected by the authors from a combination of public photography and private sources, ensuring both authenticity and cultural relevance.

To standardize the dataset for machine learning applications, most of the images were cropped to highlight the facial region and subsequently resized to 224×224 pixels. The dataset spans individuals aged 18 years and above, offering a broad representation of adult facial features, although explicit age annotations are not provided, constituting a known limitation. Each image is annotated with three primary labels: a beauty score on a scale from 1 to 5, assessed by five independent human raters; gender, encoded as 1 for male and 0 for female; and smile expression, encoded as 1 for smiling and 0 for non-smiling. With its balanced structure, demographic specificity, and multi-attribute annotations, KurdFace-1000 supports a range of predictive tasks such as beauty regression, gender classification, and smile detection, making it an ideal foundation for developing fair, culturally aware, and multi-task deep learning models.

To further illustrate the characteristics of the KurdFace-1000 dataset, Fig. 2. presents the distribution of final beauty scores across four demographic subgroups, defined by gender and smile expression. These histograms visually demonstrate the relative balance and natural spread of perceived attractiveness within each subgroup.

Beauty rating and image annotation protocol

Each image in the KurdFace-1000 dataset was independently rated by five human evaluators, both male and female, using a discrete beauty scale from 1 to 5. Although no explicit rating guidelines were imposed on the human evaluators, this was an intentional design choice aimed at preserving the natural subjectivity inherent to facial beauty perception. Raters were encouraged to use their own aesthetic judgment, rooted in shared cultural norms, without being constrained by fixed criteria such as facial symmetry, proportions, or skin clarity. This approach aligns

http://dx.doi.org/10.14500/aro.12393

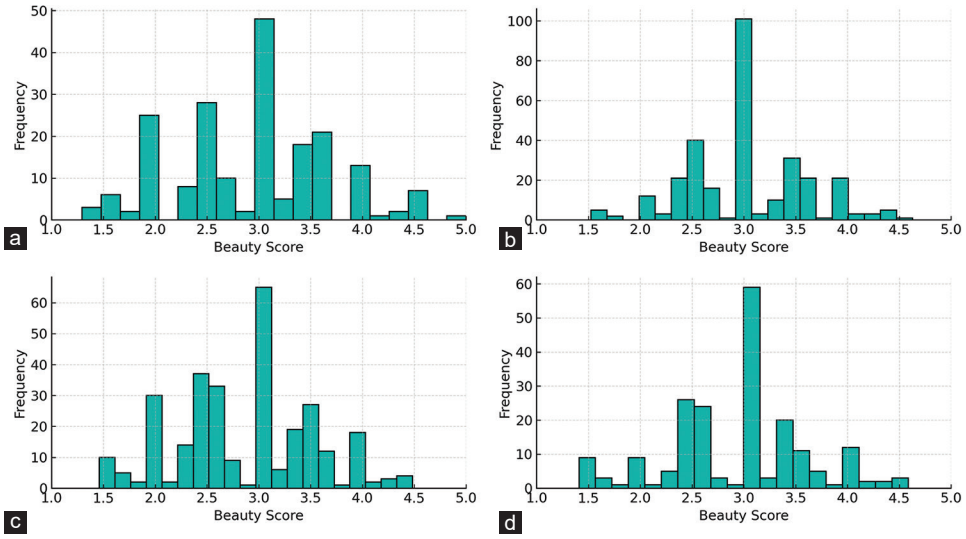

Fig. 2. (a) Female (Non-smile), (b) Female (Smile), (c) Male (Non-smile), and (d) Male (Smile). Each histogram shows the frequency of final averaged beauty ratings (ranging from 1 to 5), as assigned by five human raters. The distributions indicate Gaussian patterns in all subgroups, confirming the validity and consistency of the collected annotations.

with previous facial beauty datasets such as SCUT-FBP5500, which also relied on intuitive assessments to reflect real-world variability. Despite the absence of formal guidelines, the reliability and consistency of ratings were carefully assessed. All raters were Kurdish, ensuring cultural homogeneity in aesthetic perception. The dataset was subjected to multiple statistical evaluations, including standard deviation (SD) analysis, Shapiro–Wilk normality testing, and Fleiss' Kappa coefficient computation. These metrics revealed a high level of inter-rater consistency, with 91.7% of images receiving an SD below 0.5 and only 4 images identified as outliers (SD ≥ 1.0). Fleiss' Kappa yielded a value of 0.3525, suggesting fair agreement among the five raters. Informal post-rating interviews indicated that raters subconsciously evaluated beauty using common aesthetic markers such as facial harmony, skin smoothness, symmetry, and the presence of a pleasant expression. These insights, although anecdotal, help contextualize the rating process and underscore the dataset's cultural coherence.

The final beauty score for each image was calculated as the arithmetic mean of the five ratings, ensuring a representative aggregate measure of perceived attractiveness. To assess the statistical characteristics of these ratings, a distribution analysis was conducted. The resulting histogram, fitted with a Gaussian curve, as illustrated in Fig. 3 reveals that beauty scores approximately a normal distribution, reflecting natural variation in aesthetic perception. Furthermore, a Shapiro–Wilk (Shaphiro and Wilk, 1965) test for normality yielded a W statistic of 0.9792 with a $p = 9.51 \times 10^{-11}$. While this result indicates a statistically significant deviation from perfect normality, the visual analysis confirms that the distribution is Gaussian like.

The observations suggest a sufficient degree of diversity and consistency in the ratings, supporting their use in predictive modeling tasks. In addition to the beauty scores, each image was manually annotated with binary labels for gender (1 = male, 0 = female) and smile expression

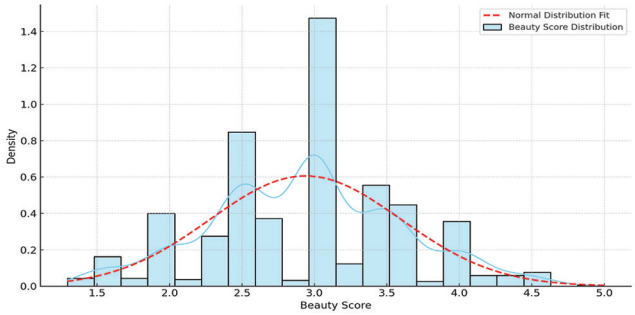

Fig. 3. Distribution of beauty scores in the KurdFace-1000 dataset with a fitted normal distribution curve. The histogram illustrates the overall spread of aesthetic ratings, showing a rough Gaussian shape centered near score 3. A red dashed line represents the best-fit normal distribution.

(1 = smiling, 0 = non-smiling), providing rich multi-attribute supervision for downstream learning frameworks.

Ethics statement and informed consent

The data collection protocol for the KurdFace-1000 dataset was reviewed and approved by the Ethics Committee of Akre University for Applied Sciences, under reference number 525 on 23rd February 2025. All participants voluntarily contributed their facial images after being fully informed about the purpose of the research, the scope of data use, and the non-public nature of the dataset. Written informed consent was obtained from each individual included in the dataset. Participants were informed that their data would be used exclusively for academic research in facial analysis and that their identities would remain confidential. All identifiable metadata were excluded. In addition, participants were given the right to withdraw their consent at any time, and no minors were included in the dataset. The clarity and the visibility of ethical procedures in accordance with publication standards are ensured.

To ensure consistency and quality in image collection, a semi-structured image acquisition protocol was adopted

during the development of the KurdFace-1000 dataset. The images were gathered using digital cameras and smartphones with a standard resolution. Whenever possible, subjects were photographed under natural daylight or soft indoor lighting to reduce harsh shadows and enhance visibility of facial features. Artificial lighting was only used when consistent ambient conditions could not be achieved, and no flash was used to avoid overexposure.

Facial orientation was controlled to capture both frontal and profile views, with participants instructed to face the camera directly or turn sideways from 30° to 45°. All images were captured with the subject's face clearly visible, and attempts were made to minimize occlusions such as sunglasses, scarves, or heavy hair covering facial features. However, to maintain realism and cultural authenticity, natural facial attributes such as beards, mustaches, and light headscarves were retained where applicable.

The backgrounds were varied but mostly neutral, including plain walls or natural scenery. In cases where the background contained identifiable objects or people, manual cropping was applied to center and isolate the face. All images were further processed to ensure uniform size (224 × 224 pixels) and cropped to focus on the facial region.

B. Statistical Analysis

Beauty rating consistency analysis

To evaluate the reliability and consistency of human annotations in the KurdFace-1000 dataset, the SD of beauty scores across the five raters was computed for each image. This metric quantifies the level of agreement: a lower SD reflects stronger consensus among raters, while a higher SD suggests subjective divergence (Rosenthal and Rosnow, 2008).

The results obtained from this study show a high level of consistency. Specifically, 917 out of 1000 images had an SD below 0.5, indicating a strong inter-rater agreement for the vast majority of samples. In contrast, only 4 images exhibited an SD ≥ 1.0, suggesting outstanding disagreement on perceived beauty. The remaining 79 images fell within the range 0.5 ≤ SD < 1.0, indicating a moderate level of variation among raters. A total of 395 images exhibited an SD of exactly 0.000, reflecting complete consensus among all five raters on these samples. The SD statistics are presented in Table II. Moreover, the distribution is visualized in Fig. 4.

.As illustrated in Fig. 4, the histogram of SD values is highly concentrated near zero, confirming the overall consistency in beauty assessments. Moreover, Table III reports the representative images with the highest disagreement among raters (SD ≥ 1.0), which may help identify cases of subjective divergence and support potential rater of reassessment or further investigation of ambiguous facial characteristics.

The five individuals who rated the images were native Kurdish adults, aged between 25 and 45 years, residing in the Kurdistan Region of Iraq. Due to privacy constraints and the informal volunteer-based nature of the annotation process, further demographic details such as gender or socioeconomic background were not formally documented. All raters were culturally embedded in the region and were selected to reflect the local understanding of aesthetic preferences. This cultural consistency was intended to reduce variability in cross-cultural interpretations and to ensure the ratings reflected a localized aesthetic norm. It is acknowledged that the absence of broader demographic diversity among raters limits the generalizability of the dataset to other populations. This represents a deliberate methodological choice aligned with the study's aim to create a culturally specific dataset for underrepresented ethnic groups in FBP research.

TABLE II
SUMMARY STATISTICS OF SD ACROSS ALL 1000 IMAGES IN KURDFACE-1000

Metric	Value
Minimum SD	0.000
Maximum SD	1.289
Mean SD	0.282
Images with SD <0.5	917
Images with 0.5 ≤SD <1.0	79
Images with SD ≥1.0	4

SD: Standard deviation

TABLE III
IMAGES WITH THE HIGHEST DISAGREEMENT AMONG RATERS (STANDARD DEVIATION ≥1.0)

Image name	Mean score	Standard deviation
0074_M_S.jpg	3.153	1.289
0136_F_S.jpg	3.816	1.008
0166_F_NS.jpg	3.894	1.014
0285_M_NS.jpg	2.849	1.05

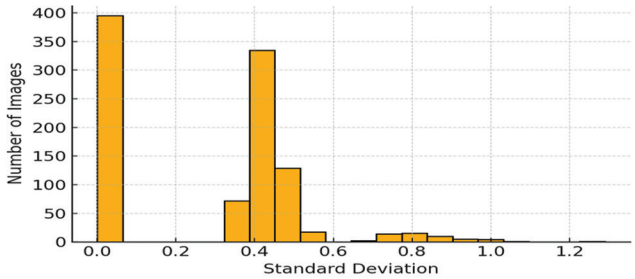

Fig. 4. Distribution of standard deviation values across beauty ratings in the KurdFace-1000 dataset. Most images show low variability, indicating high rater agreement.

Outlier distribution analysis

This section investigates the distribution of outliers in beauty ratings across gender and smile expression categories. An outlier is defined as an image for which the SD of the five beauty scores provided by human raters is ≥1.0. Table IV summarizes the total number of images, the number of identified outliers, and their corresponding percentage within each gender expression group.

As shown in Table IV, only four outlier images were identified across the entire dataset, and they are uniformly distributed between the gender and smile expression groups. This feature indicates a high level of rating consistency and suggests that no specific subgroup is disproportionately affected by extreme variability in beauty assessments.

Inter-rater agreement–Fleiss' Kappa

To assess the reliability of human annotations in the KurdFace-1000 dataset, we employed Fleiss' Kappa, a statistical measure specifically designed to evaluate inter-rater agreement for categorical data when more than two raters are involved. Different from Cohen's Kappa, which is restricted to pairwise comparisons, Fleiss' Kappa generalizes to multi-rater settings and is therefore well-suited for our scenario, where five independent raters evaluated each facial image on a beauty scale ranging from 1 to 5. The computed Fleiss' Kappa score was 0.3525, which falls into the "fair agreement" Table V. category based on the widely accepted Landis and Koch (Landis and Koch, 1977) interpretation scale. This score indicates a moderate level of consistency among raters: while individual differences in aesthetic judgment naturally exist, the overall rating process provides a sufficiently reliable foundation for training predictive models. The calculation was performed using a library in Python, a widely used statistical computing framework. The finding supports the use of these annotations in machine learning applications, though it also emphasizes the inherent subjectivity in human beauty evaluation.

Chi-square tests for gender and smile-related bias in beauty annotations

To investigate the presence of potential biases in beauty score annotations, Chi-square tests of independence were conducted to examine the association between beauty categories and two categorical attributes: gender and smile expression. For analytical purposes, continuous beauty scores were discretized into three categories: Low (score <2.5), Medium (2.5≤ score <3.5), and High (score ≥3.5) (Mining, 2006).

In gender against Beauty Category, the first test assessed whether beauty ratings were associated with gender. The contingency Table VI summarizes the counts of beauty categories for male and female subjects. The computed Chi-Square statistic was 5.56, with a p = 0.062, which is above the 0.05 significance threshold. Statistically, this method indicates that there is no significant association between gender and beauty scores in the KurdFace-1000 dataset, suggesting that beauty ratings were distributed fairly across male and female images.

The second test explored the relationship between smile expression and beauty ratings in Smile versus Beauty Category. It is clearly seen that the Chi-Square test reveals statistics of 19.09 and a p = 0.00007, which is significantly below 0.05. The obtained result indicates statistically significant association between smiling and beauty perception in Table VII. Specifically, images with smiling faces are more likely to receive higher beauty ratings compared to non-smiling ones, highlighting a potential perceptual bias favoring facial expressions associated with positive effects. These findings underscore the importance of accounting for facial expressions as potential confounding factors in beauty prediction models, while also affirming gender neutrality in the rater annotations for this dataset.

T-Test analysis of facial beauty ratings

To evaluate whether beauty scores differ significantly based on gender or facial expression, independent two-sample T-Tests were performed. This statistical method evaluates whether the means of two independent groups are statistically distinct, assuming a continuous dependent variable and a binary independent variable (Moore, Mccabe and Craig, 2009). A p-value below 0.05 is typically considered evidence of a statistically significant difference, as illustrated in Table VIII.

For the Beauty Scores by Gender, the mean beauty score for male images was 2.887, while female images received a mean score of 2.989. The t-test yielded a t-statistic of −2.4626 and a p = 0.01396, indicating a statistically significant difference in average beauty ratings between genders. This group suggests that, within this dataset, female faces were rated as slightly more attractive on average than male faces.

Concerning the Beauty Scores by Smile Expression, images with smiling expressions received a mean score of

TABLE IV
Outlier Statistics by Gender and Smile Expression in KurdFace-1000

Gender	Smile expression	Total images	Outlier images (standard deviation ≥1.0)	Outlier percentage
Female	Non-Smile	200	1	0.5
Female	Smile	300	1	0.33
Male	Non-Smile	300	1	0.33
Male	Smile	200	1	0.5

TABLE V
Interpretation Scale for Fleiss' Kappa Agreement Levels (Landis and Koch, 1977)

Fleiss' Kappa range	Level of agreement
<0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

TABLE VI
Contingency Table for Gender and Beauty Category

Gender	Low	Medium	High
Female	112	279	109
Male	144	261	95

TABLE VII
Contingency Table for Smile and Beauty Category

Smile	Low	Medium	High
Non-Smiling	153	237	110
Smiling	103	303	94

TABLE VIII
Summary of T-Test Results for Gender and Smile Comparisons

Comparison	Group 1 mean	Group 2 mean	T-statistic	p-value
Gender (male vs. female)	2.887	2.989	−2.4626	0.01396
Smile (smiling vs. non-smiling)	2.984	2.893	2.1882	0.02889

2.984, compared to 2.893 for non-smiling faces. The t-test produced a t-statistic of 2.1882 and a p = 0.02889, also confirming a statistically significant difference. These findings suggest that smiling positively influences the perceived beauty in the KurdFace-1000 dataset. Computed results reinforce the necessity of controlling such attributes in FBP models, as both gender and smile can affect the perception of the assessment.

III. Methodology

This study adopts an MTL framework to simultaneously predict facial beauty scores and classify auxiliary facial attributes such as gender and smile expression. The custom-built dataset used in this research, detailed in Section 2, was randomly split into 80% for training, 10% for validation, and 10% for testing. MTL was selected for its capacity to exploit task-relatedness by enabling shared representation learning, which often leads to improved generalization and robustness across tasks.

The model architecture is based on EfficientNetV2B0, a high-performance CNN pretrained on ImageNet. The base model was fine-tuned on the target dataset after replacing the classification head with three task-specific branches. All facial images were resized to 384 × 384 pixels and augmented using horizontal flipping during training to enhance diversity and reduce overfitting.

The first output branch was designed for beauty score prediction and modeled as a regression task. It comprised two fully connected layers with ReLU activation, followed by a sigmoid-activated neuron. The resulting output, constrained to the [0, 1] interval, was linearly scaled to match the target beauty score range of [1, 5]. The second and third branches were constructed for binary classification of gender and smile expression. Both followed a similar architectural structure with two dense layers and a final softmax layer to produce class probabilities, as shown in Fig. 5.

To train the model, a composite loss function was employed: the Huber loss (with $\delta = 0.5$) for the regression task and sparse categorical cross-entropy for both classification tasks. The training process used the AdamW optimizer, combined with a triangular cyclical learning rate schedule. This dynamic learning rate strategy oscillated between a minimum of 1e-5 and a maximum of 1e-3, with a step size based on the number of training iterations per epoch, promoting faster convergence and improved generalization. Evaluation metrics included root mean squared error (RMSE) for the beauty prediction task and classification accuracy for both gender and smile recognition.

The model was trained end-to-end for 100 epochs with a batch size of 32. This integrated MTL configuration allowed the network to capture shared and task-specific patterns effectively, leading to enhanced performance across all three target outputs.

IV. Results and Discussion

A. Experimental Setup

All experiments were implemented using Python 3.10.12, TensorFlow 2.15, and Keras 2.15.

All input images were resized to 384 × 384 × 3. A simple data augmentation strategy was employed using horizontal flipping to increase variability and robustness. The dataset was randomly divided into 80% training, 10% validation, and 10% testing splits.

Training was conducted on a Tesla P100 GPU, a widely adopted hardware platform in deep learning research due to its high memory bandwidth and computational efficiency. The model was trained for 100 epochs using the AdamW optimizer with a triangular cyclical learning rate policy implemented through TensorFlow Addons. The learning rate ranged between 1e−5 (min) and 1e−3 (max). The model checkpoints were configured to restore weights yielding the best validation accuracy. In this study, various prediction metrics, Pearson correlation (PC), RMSE, and mean absolute error (MAE) (Kukharev and Kaziyeva, 2020; Ibrahem, Saeed, and Abdulazeez, 2025) are used to evaluate the performance of the automatic rater. Indeed, accurate predictions yield larger PC values, as well as smaller errors. PC (Eq. 1), RMSE (Eq. 2), and MAE (Eq. 3) are used to find single and multi-task learning. The three-prediction metrics are as follows.

$$PC = \frac{\sum_{i=1}^{N}(l_{i-}\bar{l})(o_{i-}\bar{o})}{\sqrt{\sum_{i=1}^{N}(l_{i-}\bar{l})^2}\sqrt{\sum_{i=1}^{N}o_{i-}\bar{o})^2}} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(l_{i-}o_i)^2} \quad (2)$$

$$MAE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(l_{i-}o_i)^2} \quad (3)$$

The PC of (*l1, l2,..., ln*) and (*o1, o2,..., on*) are calculated using (1), where *l1, l2., ln* are the ground-truth scores and

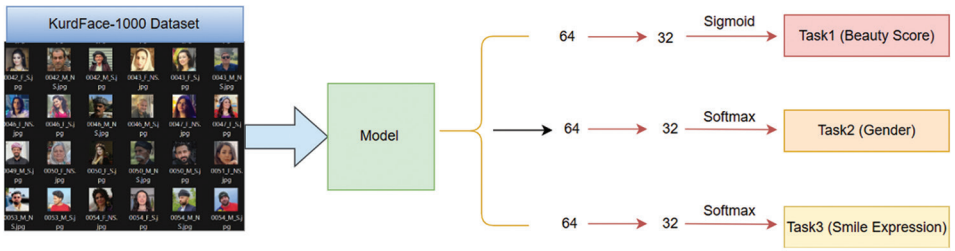

Fig. 5. Transfer learning process for EfficientNetV2B0 model.

http://dx.doi.org/10.14500/aro.12393

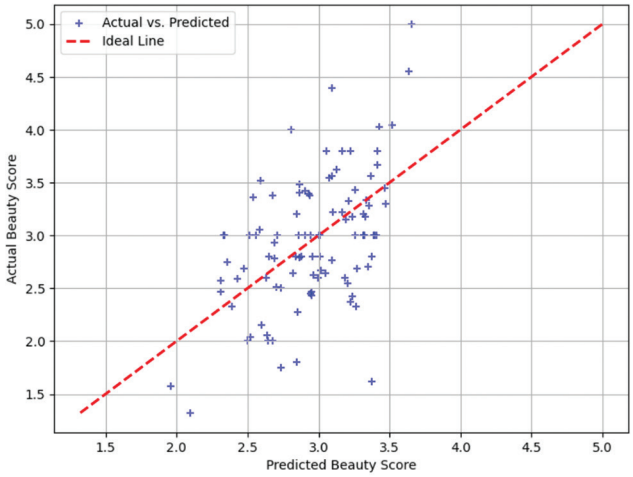

Fig. 6. Comparison of actual versus predicted beauty scores for the single-task model on the KurdFace-1000 dataset.

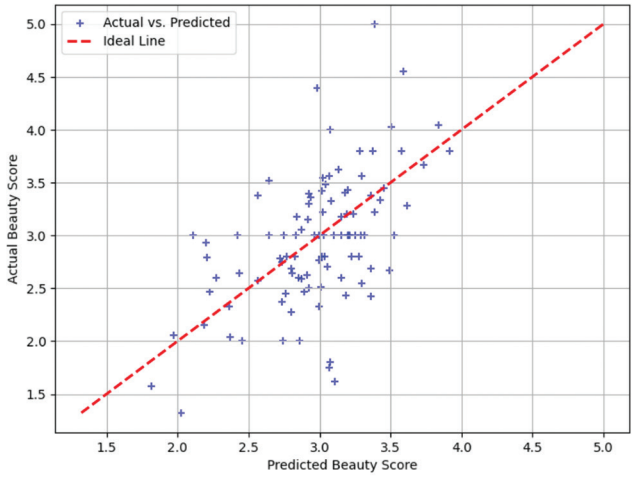

Fig. 7. Comparison of actual versus predicted beauty scores for the multi-task learning model on the KurdFace-1000 dataset

o1, o2,..., on are the scores predicted by our model. Where $\bar{}$l and $\bar{}$p are the mean of (*l1, l2,..., ln*) and (*o1, o2,..., on*), respectively. Importantly, "PC" is between −1 and 1, as the significant positive linear association is signified by "1."

B. Performance of Single-Task Learning

In the single-task learning baseline, the EfficientNetV2B0 model was fine-tuned exclusively for facial beauty score regression using the KurdFace-1000 dataset. As shown in Fig. 6, the scatter plot compares the predicted beauty scores with the ground truth ratings. The model achieved an RMSE of 0.5386, a MAE of 0.4300, and a PC of 0.5164. The results indicate a moderate positive correlation between predicted and actual scores. While the predictions follow the general trend of human assessments as evidenced by the proximity of many points to the ideal line noticeable deviations still exist, particularly for images at the higher and lower ends of the beauty scale. This suggests that the model captures broad aesthetic patterns but may struggle with finer-grained distinctions in beauty perception.

C. Performance of MTL

In the MTL setup, the EfficientNetV2B0 model was designed to predict facial beauty scores while simultaneously performing gender and smile classification. The primary task of beauty score prediction was formulated as a regression problem. Here, the model achieved an RMSE of 0.5262, an MAE of 0.3997, and a PC of 0.5548. These results indicate a moderate correlation between the predicted and actual beauty scores, reflecting the model's capability to approximate human perception of attractiveness more accurately than the single-task baseline. Fig. 7 shows a scatter plot comparing the actual beauty scores to the predicted values for the multi-task model. The red dashed line represents the ideal 1:1 correlation, while the distribution of blue points around this line illustrates the model's prediction consistency. The tighter clustering of points near the diagonal suggests improved predictive performance and reduced error variance.

TABLE IX
PERFORMANCE COMPARISON BETWEEN SINGLE-TASK LEARNING AND MULTI-TASK LEARNING MODELS ON THE KURDFACE-1000 DATASET. THE MULTI-TASK MODEL DEMONSTRATES IMPROVED PREDICTIVE ACCURACY, AS INDICATED BY A HIGHER PC AND LOWER MAE AND RMSE

Method	↑ PC	↓ MAE	↓ RMSE
Single-task learning	0.5164	0.4300	0.5386
Multi-task learning	0.5548	0.3997	0.5262

PC: Pearson correlation, MAE: Mean absolute error, RMSE: Root mean squared error

In addition to beauty prediction, the model also performed auxiliary tasks of gender and smile classification. An accuracy of 98.00% for gender classification and 90.00% for smile recognition was achieved. The high yielded accuracy demonstrates the effectiveness of shared representation learning, where the joint training of related facial attributes improves the generalization of the model and robustness across the tasks.

D. Discussion

The integration of transfer learning with an MTL framework proved more effective than single-task training, as shown in Table IX in modeling FBP for the Kurdish demographic. Although the correlation score (PC = 0.5548) for beauty prediction using MTL is modest compared to models trained on larger public datasets, it still outperforms the single-task baseline (PC = 0.5164), despite the limited dataset size. This improvement highlights the benefits of shared representation learning, where auxiliary tasks such as gender and smile classification (achieving 98.00% and 90.00% accuracy, respectively) contribute to richer and more generalizable feature extraction.

In addition, the performance of the model on the KurdFace-1000 dataset demonstrates that high-capacity architectures such as EfficientNetV2B0 can be effectively trained on culturally specific and moderately sized datasets when combined with transfer learning, data augmentation, and multi-task optimization.

Due to the private and culturally restricted nature of the KurdFace-1000 dataset, and considering time and regional resource constraints, it was unable to perform direct experimental comparisons using identical model settings on public datasets such as SCUT-FBP5500. To contextualize the performance of the proposed MTL model on KurdFace-1000, the results were compared with those reported in previous studies on established datasets such as SCUT-FBP5500. For instance, (Xie, et al., 2015) reported a PC of 0.6482 and an MAE of 0.3931 on the SCUT-FBP dataset using a single-task Gaussian regression approach, while (Liang, et al., 2018) achieved a PC of 0.8298 and MAE of 0.2938 on the SCUT-FBP5500 dataset using MTL with an AlexNet architecture. The current model, despite being trained on a smaller and culturally specific dataset, achieved a PC of 0.5548 and MAE of 0.3997. These results suggest that KurdFace-1000, when used in conjunction with efficient architecture and MTL, can yield competitive performance and demonstrate the dataset's viability for advancing fair and representative beauty prediction models.

E. Comparative Evaluation with Additional Architectures

To further evaluate the robustness of the KurdFace-1000 dataset in an MTL setting, we extended our experiments to include two widely adopted CNN backbones: MobileNet, EfficientNetV2B1, and ResNet50. All hyperparameters, training configurations, and preprocessing steps were kept identical to those used for EfficientNetV2B0 in the previous MTL experiments.

Table X summarizes the beauty score regression results obtained within the MTL framework. While all alternative architectures captured general aesthetic patterns in the dataset, their performance varied notably. MobileNet achieved an RMSE of 0.5780, MAE of 0.4544, and a PC of 0.4046, reflecting a weak correspondence with human ratings. ResNet50 yielded improved results over MobileNet, with RMSE of 0.5743, MAE of 0.4319, and PC of 0.4490. EfficientNetV2B1 demonstrated further gains, achieving an RMSE of 0.5527, MAE of 0.4359, and PC of 0.5200, indicating a stronger alignment with subjective assessments compared to MobileNet and ResNet50. Nonetheless, EfficientNetV2B0 remained the top-performing model, with RMSE = 0.5262, MAE = 0.3997, and PC = 0.5548.

These results confirm that EfficientNetV2B0, when trained in an MTL setting, is more effective in extracting culturally specific aesthetic features from KurdFace-1000 compared to other popular deep learning backbones. The findings also suggest that deeper or more parameter-efficient architectures, combined with MTL, are better suited for capturing subtle cues in beauty perception for demographically specific datasets.

TABLE X
Comparative Performance of Additional Architectures on Beauty Prediction using KurdFace-1000

Model	↑ PC	↓ MAE	↓ RMSE
MobileNet (MTL)	0.4046	0.4544	0.5780
ResNet50 (MTL)	0.4490	0.4319	0.5743
EfficientNetV2B1 (MTL)	0.5200	0.4359	0.5527
EfficientNetV2B0 (MTL)	0.5548	0.3997	0.5262

V. Limitations KurdFace-1000

Despite the valuable contributions of the KurdFace-1000 dataset, several limitations are acknowledged. First, the dataset lacks age group annotations, and no age-related labels were provided in this dataset. This limits the ability to analyze or model age-related variations in facial beauty perception and restricts the generalizability of models across different age demographics. Second, despite the balance of the dataset in terms of gender and smile expression, age diversity is not exhibited, which may impact the robustness of predictive models. Further limitation in this field is that the dataset contains only 1000 samples, which may be limiting for training complex deep learning models without overfitting. Moreover, demographic details of the human rates were not comprehensively recorded, limiting in-depth analysis of inter-rater bias across age or gender. While the raters were all native Kurdish adults familiar with regional beauty standards, this cultural specificity, while intentional, may constrain the generalizability of the dataset to other cultural contexts. The dataset is not publicly available due to the cultural and ethical principles in the Kurdistan region, where publishing personal facial images, especially of women, is often socially discouraged. They highlighted limits to open access and reproducibility, but they reflect important contextual challenges that researchers need to respect and navigate.

Beyond its academic and technical contributions, the KurdFace-1000 dataset holds significant potential for real-world applications, particularly in promoting fairness, regional relevance, and cultural sensitivity in artificial intelligence systems. It can serve as a benchmarking tool for assessing demographic bias in existing FBP models, many of which are trained predominantly on Western or East Asian datasets. By evaluating model performance on this underrepresented ethnic group, researchers can identify and address fairness issues. The dataset supports the development of regionally tailored recommendation systems, such as cosmetic or aesthetic advisory applications by aligning AI outputs with the aesthetic norms of the Kurdish population. Finally, KurdFace-1000 enables the training of culturally sensitive beauty analysis models, where beauty judgments reflect local values and perceptions rather than globalized or externally imposed standards. These applications illustrate the broader societal relevance of the dataset and highlight its potential to inform ethical, inclusive, and culturally aware AI development.

VI. Conclusion

This study designed KurdFace-1000, a novel and culturally tailored facial image dataset developed to advance FBP research for the underrepresented Kurdish demographic. The dataset consists of 1,000 images, equally balanced

across gender and smile expression, with beauty ratings provided by five independent human raters using a 5-point Likert scale. To evaluate the predictive capacity of deep learning models on the KurdFace-1000 dataset, the EfficientNetV2B0 architecture was utilized under two learning paradigms: Single-task and MTL. In the multi-task setting, the model was designed to simultaneously perform beauty score prediction and classification of gender and smile attributes. This approach outperformed the single-task model, achieving an RMSE of 0.5262 and a PC coefficient of 0.5548, compared to 0.5386 and 0.5164, respectively. These results underscore the advantage of multi-task supervision in enhancing feature representation and overall prediction accuracy, particularly in scenarios with moderately sized and culturally specific datasets. The findings of this study contribute to the field of computational aesthetics by providing a benchmark dataset that bridges a notable gap in demographic representation. Despite these promising results, the study faces certain limitations, including the lack of age annotations and the relatively small dataset size, which may constrain generalization to broader populations.

While this study presents a statistically balanced dataset and evaluates demographic biases through traditional analyses (e.g., T-tests and Chi-square tests), no explicit fairness-aware modeling strategies such as adversarial debiasing, sample reweighting, or fairness-regularized loss functions were implemented at this stage. In future work, it is aimed to explore the incorporation of fairness-aware learning frameworks to further mitigate perceptual biases in beauty prediction, ensuring greater fairness and ethical reliability in downstream applications. In addition to the above research directions, it can include expanding the KurdFace-1000 in size and diversity, age annotation, incorporating geometric facial annotations, and applying fairness-aware learning techniques to further enhance the model's robustness, cultural sensitivity, and ethical alignment. Furthermore, all models can be further optimized to achieve better performance.

References

Abual-Rub, M., Nahar, K.M.O., Almomani, A., and Alzobi, F., 2025. Arab face recognition and identification based on ethnicity and gender using machine learning. *The International Arab Journal of Information Technology*, 22(4), pp, 694-708.

Boukhari, D.E., Chemsa, A., Taleb-Ahmed, A., Ajgou, R., and Bouzaher, M.T., 2023. Facial beauty prediction using an ensemble of deep convolutional neural networks. *Engineering Proceedings*, pp.56, 125.

Buolamwini, J., and Gebru, T., 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Conference on Fairness, Accountability and Transparency*. PMLR, pp.77-91.

Chen, F., and Zhang, D., 2010. A Benchmark for Geometric Facial Beauty Study. *In: International Conference on Medical Biometrics*. Springer, Berlin, pp. 21-32.

Danner, M., Hadžić, B., Radloff, R., Su, X., Peng, L., Weber, T., and Rätsch, M., 2023.Overcome Ethnic Discrimination with Unbiased Machine Learning for Facial Data Sets. In: *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2023)*. Vol. 5. SciTePress, Lisbon, Portugal, pp.464-471.

Fan, D., Kim, H., Kim, J., Liu, Y., and Huang, Q., 2019. Multi-task learning using task dependencies for face attributes prediction. *Applied Sciences*, 9(12), p.2535.

Fan, J., Chau, K., Wan, X., Zhai, L., and Lau, E., 2012. Prediction of facial attractiveness from facial proportions. *Pattern Recognition*, 45, pp.2326-2334.

Gao, L., Li, W., Huang, Z., Huang, D., and Wang, Y., 2018. Automatic Facial Attractiveness Prediction by Deep Multi-Task Learning. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, pp.3592-3597.

Gunes, H., and Piccardi, M., 2006. Assessing facial beauty through proportion analysis by image processing and supervised learning. *International Journal of Human-Computer Studies*, 64, pp.1184-1199.

Ibrahem, A., Saeed, J., and Abdulazeez, A., 2025. Insights into automated attractiveness evaluation from 2D facial images: A comprehensive review. *International Arab Journal of Information Technology (IAJIT)*, 22, pp.77-98

Ibrahem, A.H., and Abdulazeez, A.M., 2025. A comprehensive review of facial beauty prediction using multi-task learning and facial attributes. *ARO the Scientific Journal Of KOYA University*, 13, pp.10-21.

Kagian, A., Dror, G., Leyvand, T., Meilijson, I., Cohen-Or, D., and Ruppin, E., 2008. A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision Research*, 48, pp.235-243.

Khalil, A., Ahmed, S.G., Khattak, A.M., and Al-Qirim, N., 2020. Investigating bias in facial analysis systems: A systematic review. *IEEE Access*, 8, pp.130751-130761.

Krishnan, A., Almadan, A., and Rattani, A., 2020. Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups. In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp.1028-1035.

Kukharev, G., and Kaziyeva, N., 2020. Digital facial anthropometry: Application and implementation. *Pattern Recognition and Image Analysis*, 30, pp.496-511.

Landis, J.R., and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, pp.159-174.

Liang, L., Lin, L., Jin, L., Xie, D., and Li, M., 2018. SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, pp.1598-1603.

Likert, R., 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22, pp.140-155.

Liu, X., Li, T., Peng, H., Chuoying Ouyang, I., Kim, T., and Wang, R., 2019. Understanding Beauty Via Deep Facial Features. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Liu, Z., Luo, P., Wang, X., and Tang, X., 2015. Deep Learning Face Attributes in the Wild. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp.3730-3738.

Mining, W.I.D. 2006. Data mining: Concepts and techniques. *Morgan Kaufinann*, 10, pp. 559-569.

Moore, D.S., Mccabe, G.P., and Craig, B.A., 2009. *Introduction to the Practice of Statistics*. WH Freeman, New York.

Raji, I.D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, R., 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. pp.145-151.

Rosenthal, R., and Rosnow, R.L., 2008. *Essentials of Behavioral Research: Methods and Data Analysis*. McGraw-Hill Companies, Incorporated, United States.

Saeed, J., and Abdulazeez, A.M., 2021. Facial beauty prediction and analysis based on deep convolutional neural network: A review. *Journal of Soft Computing and Data Mining*, 2, pp.1-12.

Shaphiro, S.S, and Wilk, M.B, 1965. An analysis of variance test for normality. *Biometrika*, 52, pp.591-611.

Vahdati, E., and Suen, C.Y., 2020. Facial beauty prediction using transfer and multi-task learning techniques. In: *International Conference on Pattern Recognition and Artificial Intelligence*. Springer, Berlin, pp.441-452.

Xie, D., Liang, L., Jin, L., Xu, J., and Li, M., (2015), Scut-fbp: A Benchmark Dataset for Facial Beauty Perception. In: *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, pp.1821-1826.

Zhang, X., Zhang, Y., Zhang, G., Qiu, X., Tan, W., Yin, X., and Liao, L., 2022. Deep learning with radiomics for disease diagnosis and treatment: Challenges and potential. *Frontiers in Oncology*, 12, p.773840.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.W., 2017. *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-Level Constraints*, [arXiv Preprint].

Zhuang, N., Yan, Y., Chen, S., and Wang, H., 2018. Multi-Task Learning of Cascaded cnn for Facial Attribute Classification. In: *24th International Conference on Pattern Recognition (ICPR)*. IEEE, pp.2069-2074.