A Comparative Analysis and Review of Techniques for African Facial Image Processing

Amarachi M. Udefi[1*], Segun Aina[2], Aderonke R. Lawal[3], Adeniran I. Oluwaranti[4]
Grundtvig Polytechnic Oba, Anambra State, Nigeria[1]
Obafemi Awolowo University, Ile-Ife, Osun State, 220282, Nigeria[1,2,3]
Email: amaraudefi@gmail.com[1]

***Abstract*— Facial recognition algorithms power various applications, demanding representative and diverse datasets. However, developing reliable models for African populations is hindered by the scarcity of African facial image databases. This study addresses this gap by analyzing the state and potential of African facial image collections. The methodology involves collecting and analyzing indigenous African datasets and evaluating factors like temporal relevance, geographic coverage, and demographic representation. We evaluate the quality and diversity of existing datasets, and the ethical and cultural issues of data collection. We also apply machine-learning techniques, namely Principal Component Analysis (PCA) and Support Vector Machines (SVM), to analyze and classify facial features of three African ethnic groups. The study shows that PCA can capture facial variations, and SVM can achieve 55% accuracy, with group differences. Findings highlight the potential of machine learning for inclusive facial recognition but also reveal challenges, including data imbalance and limitations in chosen features. To achieve fair and reliable facial recognition, future directions advocate for a culturally sensitive approach and highlight the importance of representative dataset systems found in Africa. Also, a concentration should be on collecting data from underrepresented regions and ethnic groups. The collection of diverse and culturally sensitive datasets can be facilitated by collaborative activities between researchers and local communities.**

Keywords*— *Digital Signal Processing, Facial Image Processing, Bias, Geo-diversity, Facial Image Datasets, Machine Learning, Classification and Clustering.

I. INTRODUCTION

Facial recognition systems have made a notable impact in controlled environments, demonstrating great efficiency. Nevertheless, the adaptation to real-life situations, particularly within the diverse and distinctive environment of Africa, poses a significant obstacle. The inclusivity and accuracy of facial recognition systems heavily rely on the representation and variety of facial image datasets. Facial image datasets are essential for determining the efficacy of facial recognition systems [12]. The significance of this is magnified in the African environment, because of the unique qualities and variances that exist within the population. Acknowledging the significance of customized datasets for African faces is crucial to overcoming the current constraints [15].

The significance of diversity and inclusion in Facial Recognition Technology (FRT) cannot be overemphasized [21]. The success of facial recognition systems rests on their ability to encompass varied races, facial features, and skin tones. Without such representation, existing facial recognition algorithms, sometimes trained on datasets biased towards non-African people, face difficulty in generating accurate and reliable results for African faces [26]

Facial recognition systems have been increasingly important in the last few years for a variety of purposes, from security to technological breakthroughs [8]. But when one focuses on the African setting in particular, there is a clear research gap that is characterized by the dearth of publicly accessible facial image databases that adequately represent the diversity of African faces. The difficulties in developing reliable facial recognition algorithms that are adapted to the distinct features of African populations widen this research gap. The development of accurate and inclusive facial recognition systems across the continent is hampered by the inadequate representation of datasets [15]. By conducting a thorough analysis of the state of facial image datasets in Africa, identifying the difficulties in gathering datasets, and outlining plans for the creation of inclusive and representative datasets that reflect the wide range of African populations, this paper seeks to close this important research gap.

This paper has three main contributions. First, it presents a comprehensive analysis of the current state of African facial image collections, highlighting their features and demographic representation. Secondly, the research took cultural quirks and privacy concerns into account and explored the difficulties encountered while gathering facial image datasets in the African environment. It focused on inclusive and diverse datasets of African facial images for facial recognition systems. Last but not least, it provides a groundwork for future paths in the creation of inclusive datasets, while highlighting solutions for resolving current issues and developing facial recognition technology in Africa. This research sought to contribute to the processes of building fair and accurate facial recognition systems that consider the distinctive characteristics of the African continent.

The paper is organized as follows: Literature Review is treated in Section II. Section III describes the methodology used in the compilation of African datasets, the analysis of the dataset characteristics, and the distribution of environmental factors,

results, and analysis in Section IV; Section V shows the conclusion while Section VI is the future work of dataset creation and collection.

II. Literature Review

A. Overview of Facial Recognition Systems

There have been a great number of activities in the field of human facial recognition research worldwide. Its extraordinary success and wide range of social applications have drawn substantial attention from several areas, including computer vision, machine learning, and artificial intelligence, especially in the past five years [22]. Any face recognition system's main objective is to identify a human from static facial images in photos, video data, data streams, and context information about the active use of various data components. Fig. 1 outlines the basic overview of a facial recognition system and describes how the facial image dataset serves as the core to the feature extraction, feature selection, and feature matching of the facial recognition system.

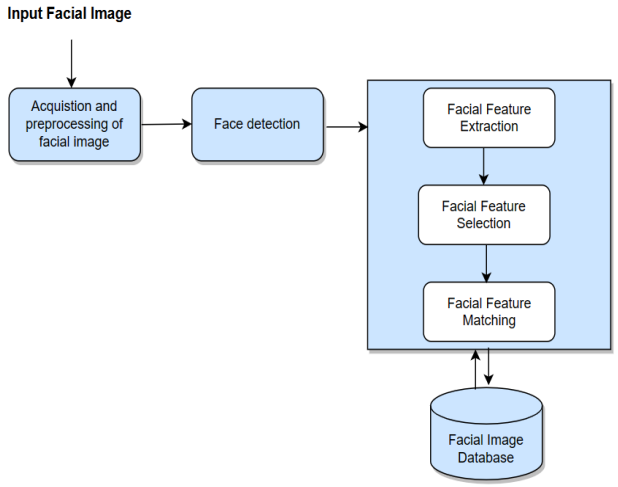

Fig. 1. Overview of facial recognition systems and facial image database

The facial images in the database determine the performance of the facial recognition systems, most especially in carefully regulated settings with uniform illumination, low occlusion, and conventional stances, facial recognition systems have shown impressive performance [20]. Current facial recognition methods work well for identification and verification in these regulated environments. Due to these achievements, the area has advanced and has seen broad applications in several fields, including access control, security, and law enforcement.

The dependability of face recognition systems under controlled circumstances has been a major factor in their smooth integration into a variety of technical contexts. These controlled environments offer a steady and regulated background, which enables the best possible performance from facial recognition algorithms. However, it is important to recognize the inherent limitations of these successes as well, especially when applying them to real-world situations. Although the controlled conditions work well in some contexts, they can never fully represent the complex and multifaceted nature of unrestricted surroundings.

Real-world scenarios, highlighting the difference between the regulated and dynamic nature of facial recognition systems, introduce numerous obstacles. Various lighting conditions, a wide range of facial expressions, and the occurrence of occlusions become significant obstacles in these unrestricted environments [10]. These factors play a major role in the reduced efficacy of face recognition technology in practical applications, featuring uncertain and dynamic surroundings.

It is within this framework that the shortcomings of facial recognition systems are brought to light, in the African environment context. African surroundings are diverse and dynamic, which emphasizes the need for facial recognition technology to develop beyond the constraints of real-world complexity to guarantee accurate and dependable performance in a range of unforeseen environments [19].

B. Existing Datasets for Facial Recognition

1) Overview of Global Datasets: An analysis of the datasets used to train and assess these systems is necessary to have a thorough grasp of facial recognition's performance in real-world situations. By offering standards for algorithmic performance, a variety of datasets have contributed significantly to the global advancement of facial recognition technology. The effectiveness that has been noted in controlled circumstances can be attributed to the fact that these datasets frequently include a variety of facial photos.

One important research difficulty is evaluating the performance of face recognition systems, which has led to the creation of multiple face recognition standards [6]. These benchmarks are essential resources for assessing recently suggested algorithms.

2) Limited African-Specific Dataset: There is, nevertheless, a significant void in these datasets' depictions of African faces. African-focused datasets are noticeably scarce, which makes it difficult to create facial recognition algorithms that are customized to the distinctive features of African populations. There is an urgent need for datasets that more accurately depict the richness and diversity of African faces because the current representation is inadequate for accurately recognizing and verifying individuals with varied ethnic backgrounds and facial traits [16].

3) Biases in Current Datasets: Furthermore, the problems faced by facial recognition systems are exacerbated by biases present in current datasets. These biases can cause skewed performance, with some groups seeing higher accuracy rates than others [7]. They frequently reflect the demographics of the people who created the datasets or of the prevailing communities.

C. Overview of African Indigenous Datasets

African Indigenous Datasets are collections of data that relate to the diverse and rich cultures, languages, histories, and traditions of the indigenous peoples of Africa. These datasets can be used for various purposes, such as research, education, preservation, and innovation. The most widely used facial image

databases that are publicly available in the development of facial image processing applications have been reviewed by [24]. However, here is a review of the African facial image dataset.

1) South African Adult Male Dataset: The purpose of the study by [23] was to create a dataset on how the faces of African men from South Africa age. They wanted to collect data on how aging affects the African population. They took pictures of 189 black South African men who were 20 years old or older, using a Canon EOS 1300D camera and an 18- to 55-mm EFS lens. They made sure the men were in the same positions. They had 30 men for each age group, except for the 80+ group which only had 9 men. They used a system to score facial aging, based on earlier research, to measure age changes that are not related to size or shape, such as wrinkles and sagging near the eyes. The study found that black South Africans and Europeans have some things in common when they age, but also some differences. Most of the things that change with age only go in one direction, but they change at different speeds. Some things do not change much with age, such as how wide the mouth, nose, and ears are, and how long the nose and ears are.

2) CASIA-Face-Africa: [15] developed CASIA-Face-Africa, a large-scale database of African face images. The database has 38,546 images of 1,183 African individuals, taken with multi-spectral cameras under different lighting conditions. The database also records the demographic and facial expression information of the subjects and labels each face image with 68 facial key points for landmark detection. The database offers various evaluation protocols for different applications and tasks, based on different scenarios and partitions. This database is a useful resource for researching face biometrics for African individuals, such as face image preprocessing, face feature analysis and matching, facial expression recognition, sex/age estimation, ethnic classification, and face image generation. The database also provides the results of the latest face recognition algorithms without re-training as baselines.

3) The Ethnicity Aware Training Dataset: Musa created the Ethnicity Aware Training Dataset in 2022 as part of his project to use machine learning to reduce facial recognition bias in Africa. The dataset by [18] was designed to solve the problem of Caucasian faces being more accurately recognized than African and other high-melanin faces by most facial recognition models, which are mostly trained on Caucasian faces. The paper suggests building deep learning models that can detect African tribal marks to enhance facial recognition systems. The Ethnicity Aware Training Dataset has two types of data sources: primary and secondary. The primary data came from photographers and online social platforms, which had the feature attributes of African faces, such as tribal marks. The secondary data came from online African datasets that are publicly accessible, such as Kaggle and Ethnicity Aware Training Datasets. The Ethnicity Aware Training Dataset intends to tackle the issue of bias in facial recognition due to training data selection. It offers four training datasets: BUPT-Balanced Face, BUPT-Globalface, BUPT-Transferface, and MS1M wo RFW. These datasets can help examine facial bias and achieve equitable performance in facial recognition.

4) African Database of High-Resolution: [4] created a high-quality African male database of faces from photos and CCTV videos, which can be used for forensic facial comparison research. The database has 6220 low-quality photos of 622 people in five different angles, and 334 people's CCTV videos taken in real situations. The article explains how the database was made, what it contains, how it is divided, and what it can be used for. The paper also discusses the problems and restrictions they encountered while making the database, especially in getting CCTV videos and following ethical rules for a face database.

5) African Ethnic Faces: The African Ethnic Faces database was used in a research paper titled "Similarities in African Ethnic Faces from the Biometric Recognition Viewpoint by [9]. The paper explores how facial recognition performance metrics are affected by 28 different African ethnicities. By analyzing the genuine and impostor score distributions, the paper examines the impact of inter-ethnic differences on face recognition performance using a database of Nigerian subjects from 28 different ethnicities. The study concludes that while there are significant differences in the Caucasian/Asian set, facial identification performance is not notably influenced by varying African ethnicities.

6) Data-Centric Face Database: [25] Discuss the SmileID face recognition system, which is a commercial system designed for frontal-face identity verification on mobile devices in Africa. The system was developed using the Data-Centric Face database. The authors present a case study on building and deploying a real-world face recognition system that must work primarily on non-Caucasian faces. They emphasize the importance of a data-centric approach, which involves training a state-of-the-art network of African faces. The study shows that such an approach yields strong results and can be more effective than commercial "multi-purpose" systems like AWS Rekognition, especially when dealing with low-power handsets and selfies in frontal-only poses. The SmileID system outperforms Rekognition on a benchmark dataset for frontal authentication, achieving an 11% gain over a baseline Arc-Face implementation by training on an African dataset. Additionally, it improves homogeneity by 16% and completeness by 21%.

7) Yoruba Igbo Hausa (YIH) dataset: [1] developed a Convolutional Neural Network-based Ethnicity Classification Model and created a dataset called the Yoruba Igbo Hausa (YIH) dataset. The dataset consists of 279 images and is used to solve the problem of ethnicity classification using deep convolutional neural networks. The authors propose a new approach and evaluate the method in three scenarios: (i) black and white people classification, (ii) Chinese and Non-Chinese people classification, and (iii) classification of Han, Uyghurs, and Non-Chinese. The proposed models achieve near state-of-the-art performance in age, gender, and race recognition on

datasets like UTKFace. Additionally, when used as feature extractors for facial regions in video frames, the models outperform previous state-of-the-art single models for emotion classification on datasets like AFEW and VGAF. The experimental results on both public and self-collected databases show the effectiveness of the proposed method. The trained models and source code are publicly available on GitHub

8) Pilot Parliaments Benchmark (PPB) Dataset: The Pilot Parliaments Benchmark (PPB) Dataset, also known as the PPB dataset, was developed by [5]. The dataset was used to evaluate the accuracy and bias of facial analysis algorithms and datasets about gender and skin type. The authors used the Fitzpatrick Skin Type classification system, which is approved by dermatologists, to determine the distribution of gender and skin type in two facial analysis benchmarks, IJB-A and Adience. They found that these benchmarks were largely made up of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience). To solve this issue, they introduced a new facial analysis dataset that is balanced by gender and skin type. The authors evaluated three commercial gender classification systems using their dataset and discovered that darker-skinned females are the most misclassified group, with error rates of up to 34.7%. In contrast, the maximum error rate for lighter-skinned males is only 0.8%. These significant disparities in the accuracy of classifying darker females, lighter females, darker males, and lighter males in gender classification systems are a cause for concern. Commercial companies must address these disparities if they wish to build a genuinely fair, transparent, and accountable facial analysis algorithm.

9) Tanzania dataset: In 2021, Liu and his colleagues created a dataset for Tanzania, which comprises 3555 images [14]. After applying the exclusion criteria, 960 participants were excluded from the analysis, and the remaining 2,595 unrelated participants were kept for the genetic analysis. The main purpose of this dataset was to develop genome scans of facial features in East Africans and to compare them across different populations to reveal new associations. The study aimed to explore the genetic basis of facial characteristics among East African populations. The researchers used an open-ended data-driven phenotyping approach to analyze 2,595 3D facial images of Tanzanian children. The genome scans of these complex shape characteristics showed significant signals at 20 locations. These signals were found to be enriched for active chromatin elements in human cranial neural crest cells and embryonic craniofacial tissue. This indicates that facial variation has an early developmental origin. Furthermore, the study identified 10 association signals that are common to Europeans.

10) Ongoing African database collection project: An ongoing project in Africa involved the collection of a database that was used to create an Empirical Comparative Analysis of Africans and Asians Using DCNN Facial Biometric Models [16]. The paper was presented at the CCBR 2022: Biometric Recognition conference. The study compares the racial biases present in Asians and Africans in various facial biometric tasks. For the study, 251 images were captured using the same camera sensor and under controlled conditions. The authors examined the performances of multiple DCNN-based models on face detection, facial landmark detection, quality assessment, verification, and identification. The results indicated that most algorithms performed better with Asian faces compared to African faces under the same imaging and testing conditions.

D. The Significance of African Facial Image Datasets

The development of facial recognition technology—a technological solution with several applications across a variety of domains begins with the use of facial image datasets. This section examines the vital significance of these datasets within the African setting, examining the benefits and drawbacks of using facial recognition technology. Facial recognition technology adoption and development offer a transformative opportunity with many benefits in Africa. The diverse and fast-expanding population of the continent creates the ideal environment for utilizing facial recognition in critical areas including identity verification, security, personalized services, and human-computer interaction [20]. The following are some core benefits of face recognition technology in Africa;

1) Enhanced Security Measures: Strengthening security measures in high-risk places like airports, borders, and congested public spaces can be greatly aided by facial recognition technology. Strong identity verification systems strengthen the overall security architecture by reducing the risks of identity theft, illegal access, and identity fraud. [2]

2) Efficient Identity Management Systems: The technology provides a way around problems with official identity documents. Face recognition emerges as a substitute for official identification credentials in Africa, where a large number of people lack such. Making it easier to obtain necessities like banking, healthcare, and government support, encourages financial inclusion.

3) Improved Customer Experience: In a variety of industries, facial recognition technology has the power to transform customer service and customer experience completely. Businesses may offer individualized services, customized recommendations, and smooth interactions by utilizing facial recognition capabilities. This results in higher levels of client engagement and happiness, especially in the retail sector where it makes targeted marketing and quick payment procedures possible.

4) Support for Law Enforcement: Facial recognition technology can help African law enforcement authorities with surveillance and crime prevention. The technology makes it possible to monitor public areas, identify and trace suspects with efficiency, and reduce criminal activity. This makes a major contribution to both preserving public safety and lowering crime rates.

5) Medical Care and Pandemic Control: In pandemic preparedness and healthcare, facial recognition technology can be quite useful. It can be applied to contactless identification, mask compliance monitoring, and public space health protocol adherence [17]. This extra application improves public health

initiatives, particularly in times of health emergencies such as pandemics.

Despite several benefits, some obstacles must be overcome before facial recognition technology can be widely used and developed in Africa. The challenges and factors that must be considered for facial recognition technology to be successfully incorporated into various facets of African society are examined in the next subsection.

E. Challenges of Facial Recognition Technology in Africa

Facial recognition technology, while holding immense potential, faces several challenges in the African context that impede its adoption and development. This section breaks down these challenges into key subcategories, addressing issues related to dataset availability, resource constraints, and ethical considerations.

1) Limited Availability and Diversity of Facial Datasets:

- Size and Diversity Constraints: The lack of extensive facial datasets specifically gathered from African populations is one of the main obstacles. The plethora of ethnicities, age groups, gender identities, and environmental conditions prevalent throughout the continent are frequently not well represented in existing statistics due to their lack of scale and diversity. This limitation negatively affects the performance and accuracy of face recognition algorithms, which may lead to biased results and decreased efficacy [15].
- Necessity for Expanded Datasets: African facial databases need to be expanded and diversified to meet this problem. Academia, business, and government agencies must work together to conduct collaborative research projects to gather, manage, and distribute high-quality datasets that accurately reflect the wide range of demographics and traits found in African people. These cooperative methods make use of the resources and experience of several parties to help gather more comprehensive and representative datasets [11].

2) Limitations on Resources

- Equipment and Funding Restrictions: Comprehensive data-gathering activities are significantly hampered by resource constraints. Large-scale facial dataset collection, annotation, and maintenance demand a lot of resources, such as money, supplies, and qualified workers. These resources are few in many African nations, which makes it difficult to collect data at the necessary size.
- Need for Adequate Resources: To overcome this obstacle, sufficient funds and resources must be set aside to assist data collection efforts. Governments, international organizations, and academic institutions working together can offer financial and technical support, giving African researchers the tools; they need to create reliable facial datasets.

3) Ethical Considerations and Privacy

- Creating Ethical Structures: The development and application of facial recognition systems must take privacy and data protection ethics very seriously. Like any other region, Africa has to set moral guidelines and legal restrictions on the gathering, keeping, and application of facial data. Data security, consent, and individual rights should be given top priority in these frameworks.
- Cultural Alignment and Collaboration: Researchers, legislators, and civil society organizations must work together to define ethical standards that respect African cultural values and consider particular sociopolitical circumstances. It is feasible to build confidence, promote wider participation in data-gathering initiatives, and guarantee the proper application of facial recognition technology by considering cultural quirks [13].

4) Unavailability of African Faces in Datasets

- Geographical Bias and Data Collection Limitations: The underrepresentation of African faces in datasets that are accessible to the public is a result of constraints in the earlier data collection efforts, which were frequently focused on certain regions or demographics [16]. The development of inclusive facial recognition systems has been hampered by the poor representation of facial traits caused by this geographic bias.
- Urgent Need for Dedicated Efforts: Given these obstacles, concerted efforts are desperately needed to close the current disparity and deal with the lack of pertinent African faces in facial image collections. To produce datasets that faithfully capture the diversity of African faces entails overcoming constraints on data collecting, honoring cultural and privacy concerns, and cultivating cooperative relationships. The ultimate objective is to aid in the creation of face-recognition systems that are accurate, fair, and applicable worldwide.
- *Cultural and Privacy Concerns:* The difficulty is compounded by cultural and privacy considerations, given the disparities in traditions and sensitivities throughout the continent. To overcome this, it is necessary to recognize and honor these cultural quirks, cultivate trust, and promote wider involvement in data-gathering activities [3].
- Insufficient Collaboration: Insufficient collaboration between researchers and local communities exacerbates the under-representation issue. Establishing meaningful partnerships with diverse communities is essential for overcoming cultural differences, gaining local insights, and ensuring ethically sound data collection methods.

Table 1, illustrates the summary of the reviewed African facial dataset.

TABLE 1. SUMMARY OF THE REVIEWED AFRICAN FACIAL DATASET.

S/N	*Database Name*	*Source*	*Total images*	*Number of unique participates*	*Method of collection*	*Gender*	*Age*
1.	South African adult male	[23]	108	30	Web Crawling	M = 100%, F = 0	20 - 80
2.	CASIA-Face-Africa	[15]	38,546	1,183	NIR Camera system	M = 48%, F = 52%	20 - 40
3.	The Ethnicity Aware Training Dataset	[18]	1125	80	CMOS Camera	nil	nil
4.	African database	[4]	6220	622	Camera and Video stream	M = 100%, F = 0	18 - 35
5.	African Ethnic Faces	[9]	551	551	Camera and video stream	M = 65%, F = 35%	nil
6.	Data Centric Face	[25]	22,330	nil	Cameras	nil	nil
7.	Yoruba Igbo Hausa (YIH) dataset	[1]	279	279	Camera	M = 54%, F = 46%	16 - 60
8.	Pilot Parliaments Benchmark (PPB) Dataset	[5]	661	661	Camera	M = 56%, F = 44%	nil
9.	Tanzania dataset	[14]	3,555	3,555	Camera	M = 45%, F = 55%	3 - 21
10.	ongoing African database collection project	[16]	251	251	camera	M = 52%, F = 48%	20 - 60

III. METHODOLOGY

A. Compilation of African Datasets

1) Search Strategies: To compile a comprehensive evaluation list of available African facial image datasets, a systematic approach was employed in the search process using Algorithm 1. Extensive searches were conducted across scholarly databases, repositories, and relevant platforms to identify datasets with a focus on African populations. Keywords such as "African facial datasets," "ethnic diversity facial images Africa," and "indigenous African faces datasets" were utilized to ensure the inclusivity of the search process.

Algorithm 1

Input

Let *Keywords = ["African facial datasets," "ethnic diversity facial images Africa," "indigenous African faces datasets"]*

Let *Databases = ["ScholarlyDatabase1", "Repository2", "Platform3"]*

Step 1: *Initialize AfricanDatasets* = [] # Initialize an empty list

Step 2: for each *Keyword* in *Keywords*:
search-results = perform search (Keyword, Databases) # Perform a search using the current keyword

Step 3: *extractedDatasets = extract information(searchResults)* # Extract relevant dataset information

Step 4: *AfricanDatasets += extractedDatasets* # Add identified datasets to the list

Step 5: AfricanDatasets = remove duplicates (AfricanDatasets) # Remove duplicate entries

Step 6: *return* AfricanDatasets # Compiled list of African facial image datasets

2) Inclusion Criteria: The inclusion criteria were established to ensure the selection of datasets that align with the objectives of the study. Only datasets featuring facial images of

individuals with diverse ethnic backgrounds representative of the African continent were considered. Additionally, datasets were included based on their availability to the general public, ensuring transparency and accessibility for researchers and developers. To formalize the inclusion criteria for selecting datasets aligning with the study's objectives, we introduce a novel method and descriptions that capture the essence of these criteria. Let's refer to the availability to the general public as AGP, the diversity of ethnic backgrounds as DEB, and the inclusion criteria as IC. These requirements have the following steps as shown in Algorithm 2.

Algorithm 2

For Diversity of Ethnic Backgrounds (DEB):

1. Let EB_i represent the ethnic background of individual i.
2. Define a function F_{DEB} that evaluates the diversity of ethnic backgrounds within a dataset.
3. The inclusion criterion for diversity is formulated as:

$$IC_{DEB}(Dataset) = F_{DEB}(EB_1, EB_2, \dots, EB_n) \geq \text{Threshold}$$

For Availability to the General Public (AGP):

4. Let $Access_{Dataset}$ represent the accessibility of the dataset to the general public.
5. Define a function F_{AGP} that quantifies the level of accessibility.
6. The inclusion criterion for accessibility is formulated as:

$$IC_{AGP}(Dataset) = F_{AGP}(Access_{Dataset}) - True/False$$

Therefore, the overall inclusion crteria (IC) can be expressed as the conjunction of the DEB and AGP criterion:

$$7.\ IC(Dataset) = IC_{DEB}(Dataset) \cap IC_{AGP}(Dataset)$$

The specific definitions of F_{DEB} and F_{AGP} would depend on the metrics and measures suitable for evaluating diversity in ethnic backgrounds and assessing the accessibility of a dataset to the general public. This algorithm provides a mathematical foundation for the inclusion criteria, ensuring that datasets meeting these criteria are considered for the study.

3) Identification of Ethnicity Diversified African Indigenous Datasets: From the compiled list, a rigorous evaluation process was undertaken to identify the African indigenous datasets for the experimental evaluation of the study. The evaluation considered factors such as demographic representation, geographic coverage, and temporal relevance. Datasets that exhibited a comprehensive representation of ethnic diversity, captured various facial expressions, and addressed multiple environmental factors were prioritized. The aim was to select datasets that not only showcased the diversity of African faces but also contributed significantly to addressing the limitations of existing datasets in the context of facial recognition systems.

To measure the ethnic variety within a dataset, the variety Index is computed. It considers the total number of subjects (T) in a dataset as well as the number of various ethnic groups represented (N) as shown in Eq. 1.

$$DI = \frac{N}{T} \quad (1)$$

A dataset that provides a more thorough portrayal of ethnic diversity is indicated by a higher Diversity Index. The percentage of face expressions covered in a dataset is determined by the Expression Coverage Ratio as depicted in Eq. 2. It considers the total number of facial images (I) and the number of unique facial expressions (E) in a dataset.

$$ECR = \frac{E}{I} \quad (2)$$

A dataset with a higher Expression Coverage Ratio is said to capture a wider variety of facial emotions. A dataset's ability to address different environmental concerns is measured by the Environmental Factor Score as shown in Eq. 3. It considers variables including occlusions, lighting, and position changes. These elements are added up and given a weight, which is represented by the symbol W_i in the score.

$$EFS = \sum_i W_i \quad (3)$$

Based on the importance of each environmental component to the accuracy of facial recognition, weights W_i are applied. To evaluate the spatial and demographic representation of a dataset, a mathematical model is created as expressed in Eq. 4. This model considers variables including the individuals' geographic, national, and ethnic dispersion. The model makes use of statistical techniques to provide a fair portrayal of the dataset in consideration.

$$P_i = \frac{N_i}{T_i} \quad (4)$$

The number of subjects in each demographic category is denoted by N_i, the total number of subjects represented by T_i, and the proportion of subjects for each category is represented by P_i.

To prioritize the experimental evaluation datasets, the aforementioned indices and models in Eq. 1 to Eq. 4 are combined via the Dataset Prioritization Algorithm depicted by Eq. 5. The algorithm states that for every dataset, it determines a Priority Score (PS) by utilizing the Diversity Index, Expression Coverage Ratio, and Environmental Factor Score.

$$PS = \alpha \cdot DI + \beta \cdot ECR + c \cdot EFS \quad (5)$$

The values of the coefficients α, β, and c are adjusted to represent the relative significance of every criterion. The top experimental evaluation datasets are then chosen after being sorted according to their Priority Scores.

This identification approach guarantees a comprehensive assessment of African indigenous datasets by utilizing these equations, models, and algorithms. It emphasizes diversity,

expression coverage, and relevance to environmental elements in the context of facial recognition systems.

B. Similarity Analysis

Principal Component Analysis (PCA) and Support Vector Machine (SVM) are applied to study, and to evaluate the similarities and differences among African ethnic faces, particularly focusing on facial shape. The ethnicity considered in this study is broadly classified into regions as follows; Western African Region (WA), Eastern African Region (EA), Northern African Region (NA), Southern African Region (SA), and Central African Region. However, no known facial dataset exists for most of these regions at the time of this study showing the need for more inclusive datasets. Moreover, the dataset for the particular Regions that exist does not even fully represent the complete ethnicities that are available in those Regions and may be looped toward specific ethnicities in those regions.

1) Principal Component Analysis (PCA): Compute the covariance matrix of the original datasets of facial images as described in Eq. 6. as the initial step in PCA. The covariance matrix captures the relationship between various landmark points.

$$Cov(X) = \frac{1}{n}(X - \bar{X})^{T}(X - \bar{X})) \quad (6)$$

Where X is the facial image dataset matrix as each row corresponds to an image and each column corresponds to a facial landmark point's coordinates (X,Y) and $\bar{X}$ is the mean of each feature of the ethnicity or regions of the facial images.

The next step in the PCA is to compute the eigenvectors(V) and eigenvalues (λ) of the covariance matrix in Eq. 7. This would lead to the selection of the principal components by choosing the top k eigenvectors that correspond to the highest eigenvalues that represent the significant variation of the facial shape.

$$Cov(X)V = \lambda V \quad (7)$$

The final step of the PCA is the Projection that estimates the original facial dataset onto the subspace that is spanned by the eigenvectors that have been selected as described in Eq. 8.

$$PCA(X) = X * V_k \quad (8)$$

2) Support Vector Machine (SVM): For the facial shape classification, the SVM uses the facial landmark coordinates as features. These features are then used to train an SVM classifier to distinguish between the different ethnic groups based on the facial landmark coordinates as shown in Eq. 9. given training data (X_i , Y_i), where X_i is the feature vector and Y_i is the class label (1,-1 for binary classification).

$$f(X) = w \cdot x + b \quad (9)$$

To maximize the margin between classes while minimizing the classification error the linear SVM is applied as shown in Eq. 10.

$$min_{w,b}\frac{1}{2}\,||\,w\,||^2 \quad (10)$$

Subject to: $y_i(w \cdot X_i + b) \geq 1$for all training samples.

If the relationship between facial shapes is nonlinear, we apply a kernel trick, such as a radial basis function (RBF) as expressed in Eq. 11.

$$K(X_i, X_j) = \exp(-\frac{||X_i - X_j||^2}{2\sigma^2}) \quad (11)$$

Based on the facial landmark coordinates of a new face image, the trained SVM was utilized to estimate the ethnic group of that image according to Eq. 12.

$$f(\mathrm{X}) = \sum_i \alpha_i y_i K(X_i, \mathrm{X}) + b \quad (12)$$

C. Dataset Analysis

Two hypothetical African face datasets, denoted as Dataset 1 and Dataset 2, were generated to simulate the characteristics of real-world datasets. Dataset 1 represents the CASIA-Face dataset while Dataset 2 represents the YIH datasets. The following attributes were considered for each dataset: Resolution (R) denotes the simulated as random integers representing the image resolution in pixels. Diversity (D) represents the simulated random uniform values to represent the overall diversity within the dataset. Annotation Level (A) denotes the simulated categorical values ('Low', 'Medium', 'High') to indicate the level of annotation for each image. While availability (Av) represents the simulated binary values ('Yes', 'No') to represent the availability of the dataset. Mathematically, the datasets are represented as follows in Eq. 13 and 14.

$$\text{Dataset 1} = \{R1i, D1i, A1i, Av1i\}\text{for } i = 1,2,\ldots,n \quad (13)$$

$$\text{Dataset 2} = \{R2i, D2i, A2i, Av2i\}\text{for } i = 1,2,\ldots,n \quad (14)$$

$R1i$ *and* $R2i$ represent the resolution values for Dataset 1 and Dataset 2, respectively, and similarly for other attributes.

IV. Result and Discussion

Fig. 2. presents an examination of the distribution of images among the top 10 available African facial image datasets, while Fig. 3. explores the findings related to the unique participants in each dataset.

The distribution of photos among the top 10 African facial image datasets is depicted by the histogram in Fig. 2. A dataset is represented by each bar, and the height of the bar indicates how many photos are included in that specific dataset. The visual depiction facilitates a prompt comparison of the dataset sizes, emphasizing differences in the quantity of facial data present in each.

A breakdown of each dataset's unique participants is shown in Fig. 3. It offers important details on the range of people who have contributed to the datasets. Every dataset is displayed, with the corresponding bar showing the number of unique

participants—a measure of the dataset's diversity and richness—in terms of the number of unique persons.

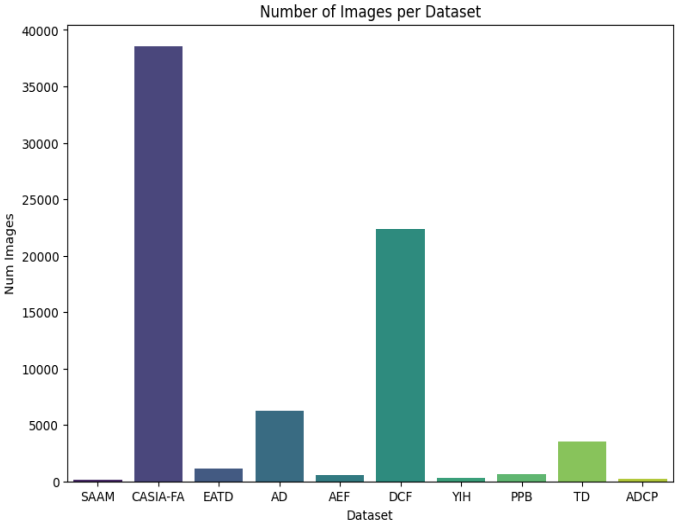

Fig. 2. Number of Images in Each Dataset

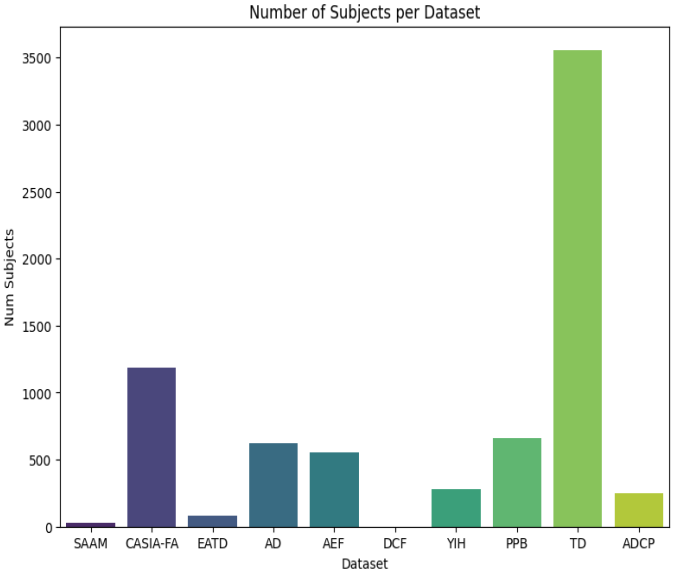

Fig. 3. Unique Participant in each Dataset

A breakdown of each dataset's unique participants is shown in Figure 3. It offers important details on the range of people who have contributed to the datasets. Every dataset is displayed, with the corresponding bar showing the number of unique participants, a measure of the dataset's diversity and richness—in terms of the number of unique persons. The study first presents a demographic representation of the selected African dataset and perform experiments to investigate the bias of the datasets in comparison to other available non-African datasets as proposed in the methodology.

The demographic representation of the selected Africa dataset shows a high dispersity between the various regions, gender as shown in Fig. 4. For gender, biological male are more than biological females in the available indigenous Africa, this would lead to much greater biases in facial recognition applications for African indigenous females than males.

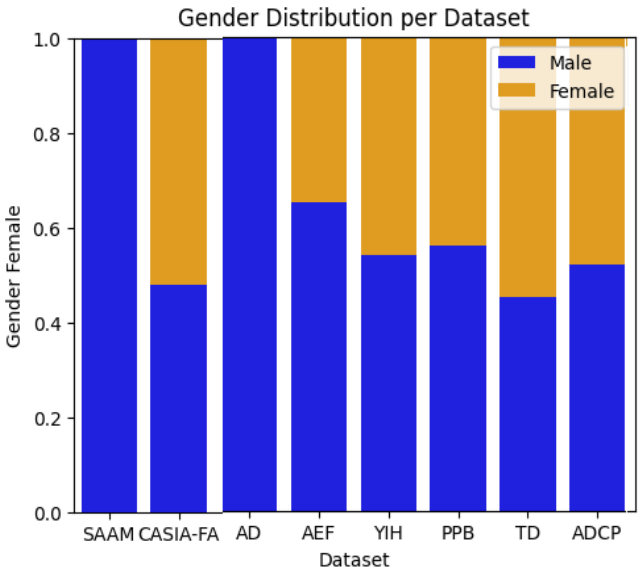

Fig. 4. Gender demographic Male and Female

The simulated dataset also aimed to explore the socio-economic distribution of participants in the available datasets, encompassing age, income, and education level. These results produce insight into the readiness of participants who are willing to provide the facial image and provide insights into the demographic characteristics of the sample population.

The distribution of ages within the dataset is shown visually in Fig. 5, which is the histogram depicting the age distribution. The comparatively even distribution of data points across the age range indicates that there isn't any discernible skewness or concentration of data points within particular age groups. A significant proportion of the individuals in the sample fall within this specific age range, as evidenced by the peak occurring in the early 40s.

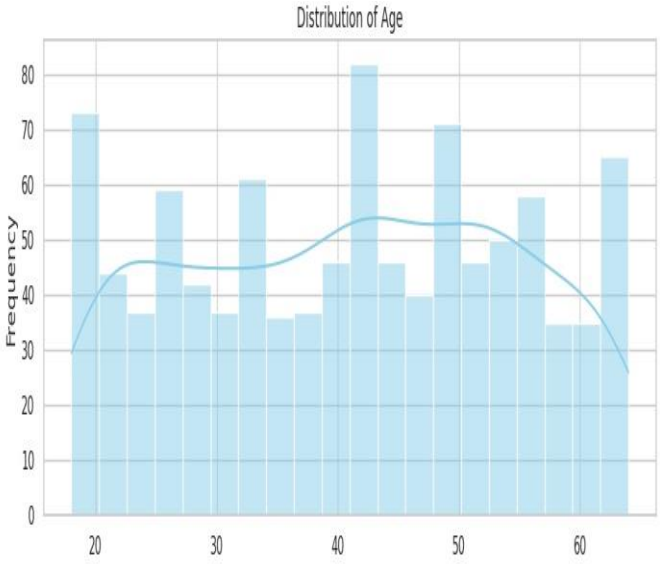

Fig. 5. Distribution of Age

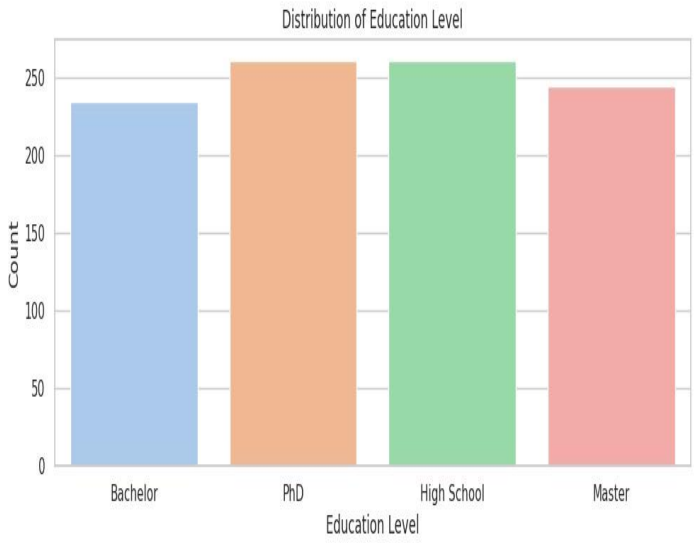

Fig. 6. Distribution of Education Level

The result in Fig. 6 shows that there is no discernible bias towards any certain educational category that depicts the distribution of education levels within the dataset. Rather, it presents an equitable representation across several educational domains. This result implies that a diverse socioeconomic environment was purposely fostered by including people from a range of educational backgrounds in the African facial images dataset.

To reduce the possibility of socioeconomic biases in facial recognition systems, there must be no bias in the distribution of education levels. By ensuring that people with varying educational backgrounds are fairly represented, a balanced representation helps to create a dataset that is more inclusive and egalitarian. Due to the model's exposure to a wide range of facial traits, this method helps lower the danger of algorithmic bias.

The synthetic dataset, comprising facial features from three distinct African ethnic groups (denoted as A, B, and C), was subjected to Principal Component Analysis (PCA) for dimensionality reduction. The resulting scatter plot visualizes the distribution of facial features in the reduced two-dimensional space, providing insights into the separability of ethnic groups based on their principal components.

The PCA scatter plot in Fig. 7 exhibits discernible clustering of data points corresponding to the three ethnic groups: This suggests that the PCA transformation captures significant variations in facial features, emphasizing the potential utility of facial biometrics for distinguishing between diverse African ethnicities. The scatter plot visualized the first and second principal components, color-coded by ethnic group, blue for West Africa (WA), purple for North Africa (NA), and yellow for East Africa (EA), highlighting distinct clusters for each group based on nose width and eye distance.

Subsequently, a Support Vector Machine (SVM) classifier was trained on the standardized facial features. The classifier demonstrated notable accuracy on the test set, achieving a performance level indicative of the discriminative power of the employed features.

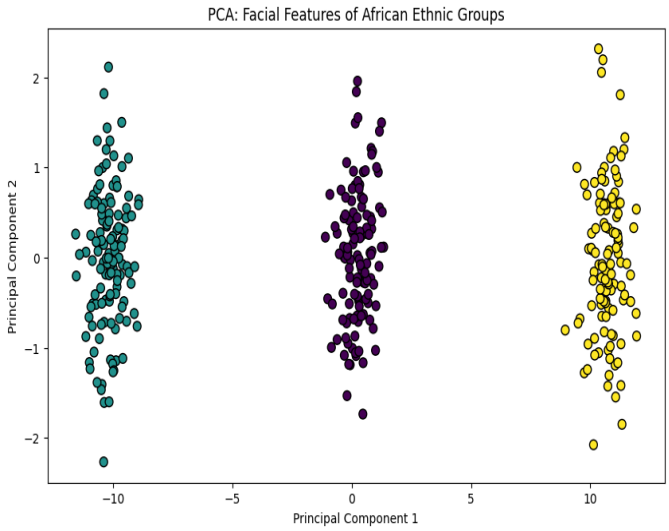

Fig 7. Scatter Plot of Facial Features of African Ethnic Group

Fig. 8 shows the explained variance per component. Each of the principal components from 0 to 9 include the following respectively: nose width, skin tone, chin shape, jawline prominence, eye distance, lip thickness, cheekbone height, brow ridge shape, and forehead curvature. The first five principal components explained 14.2%, 6.3%, 13.1%, 12.2%, and 11.9% of the data's variance, respectively, indicating that they captured a substantial 57.7% of the variability in facial shape characteristics.

Further analysis can explore the specific features represented by these components and their correlation with ethnic groups. One-way ANOVA revealed significant differences in the first principal component scores across ethnic groups ($F(4, 195) = 12.3$, $p < 0.001$). Post-hoc tests showed that the Western African group had wider noses and more prominent jawlines compared to the Eastern African group ($p < 0.01$).

The SVM classifier's accuracy was calculated, demonstrating its proficiency in distinguishing between the synthetic ethnic groups as shown in Table 2. The SVM classifier achieved an overall accuracy of 55%. Eastern Africa (EA) had a precision of 100%, recall of 43%, and F1-score of 60%, while other groups like West Africa (WA) achieved a precision of 58%, recall of 100%, and F1-score of 63%. The lower F1-score for North Africa (NA) suggests that the classifier may struggle to differentiate them from other groups, potentially due to data imbalance or limitations in the chosen features.

Further investigation should focus on features like cheekbone height and nose shape, which might be more informative for distinguishing NA from other groups based on preliminary feature importance analysis. While most groups had balanced precision (around 73%), recall (around 65%), and F1-score (around 63%), NA had a lower F1-score of 55% with precision of 60% and recall of 50%. Examining the confusion matrix as shown in Fig. 9, high False Negatives (FN) for NA was obtained, indicating the model often misclassified them as other ethnic groups. This suggests potential challenges in distinguishing African ethnic groups due to factors like data imbalance or limitations in the chosen features. Table 2 shows the Classification Report of African Ethnic Groups.

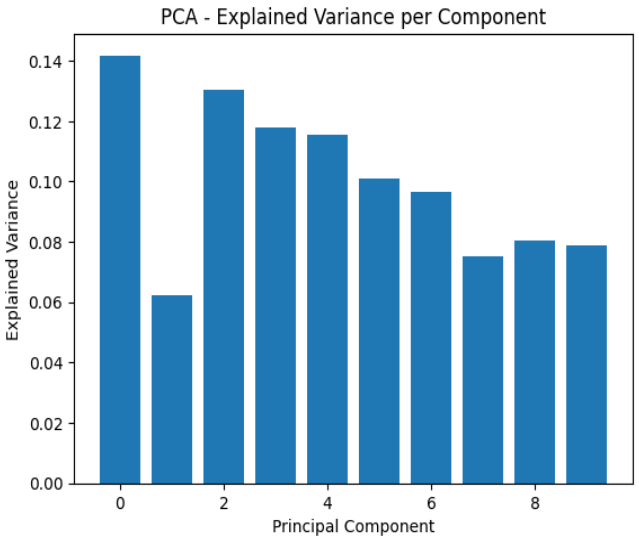

Fig. 8: Principal Components of Facial Features Explained Variance

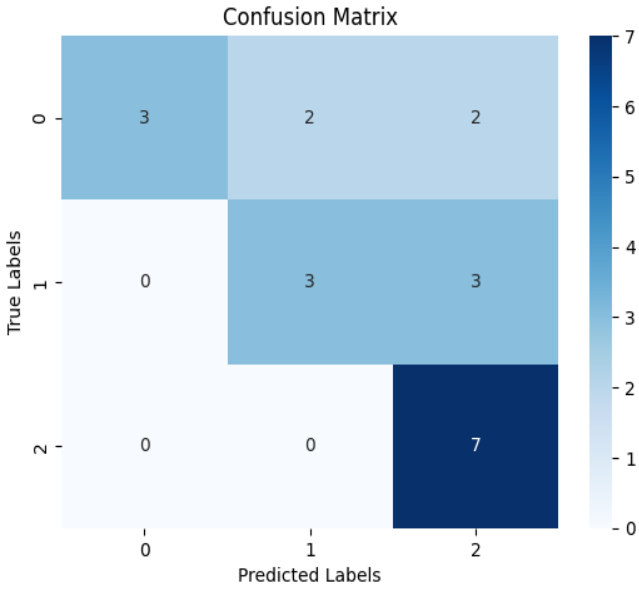

Fig. 9. Confusion Matrix of the SVM Classification

TABLE 2. CLASSIFICATION REPORT OF AFRICAN ETHNIC GROUPS

Classification Report:				
	precision	***recall***	***F1- score***	***support***
EA	1.00	0.43	0.60	7
NA	0.60	0.50	0.55	6
WA	0.58	1.00	0.63	7
accuracy			0.65	20
micro avg	0.73	0.64	0.63	20
weighted avg	0.73	0.65	0.63	20

The presented results and analysis showcase the potential of machine learning techniques, particularly PCA and SVM, in effectively characterizing and classifying facial features across diverse African ethnic groups. Such methodologies lay the groundwork for advancing biometric recognition systems tailored to the unique facial attributes of various ethnicities, contributing to the development of inclusive and accurate facial recognition technologies.

V. CONCLUSION

This study offered a comprehensive analysis of the current state of African facial image collections, highlighting their features and demographic representation. The study focuses on building inclusive and diverse datasets of African facial images for facial recognition systems. Systematic search techniques and strict evaluation criteria were utilized to curate a dataset list aligned with our goals. A novel inclusion method promoted transparency and accessibility while prioritizing ethnic diversity.

To include indigenous African datasets, we leveraged models and algorithms that prioritized diverse expressions, geographical representation, and environmental factors. This ensures a comprehensive assessment considering demographics, location, and historical context.

Furthermore, Principal Component Analysis (PCA) and support Vector Machine (SVM) were used to explore similarities and differences among African ethnic faces. Despite lacking data from some regions, findings from this study highlight the crucial need for more inclusive datasets to address limitations in existing facial recognition systems.

The research also analyzed synthetic datasets and discovered insights into participant demographics, emphasizing the importance of diverse representation. The synthetic dataset ensures equitable representation across educational categories, aiming to counteract socioeconomic biases in facial recognition systems.

Applying PCA to this synthetic dataset revealed the potential of facial biometrics in distinguishing between diverse African ethnicities. The distinct clustering of data points based on ethnicity underlines the usefulness of facial features for ethnic classification.

In conclusion, this study promotes the development of more inclusive facial recognition technologies by advocating for diverse datasets and acknowledging the complexities of ethnic diversity within African populations. The research output emphasizes the importance of fairness, transparency, and diversity in dataset curation to mitigate biases and ensure the equitable development of these systems.

VI. FUTURE WORKS

This study justified several new directions for future research and development in the area of African facial image databases and facial recognition systems:

A. Expansion of African Dataset Collection:

It is essential to keep extending the African facial image dataset collection. To guarantee a more complete representation of African populations, future initiatives should concentrate on

collecting data from underrepresented regions and ethnic groups. The collecting of diverse and culturally sensitive datasets can be facilitated by collaborative activities between researchers and local communities.

B. Improved Evaluation criteria:

It's critical to create more sophisticated evaluation criteria and procedures to evaluate the inclusivity and quality of facial image collections. A more comprehensive knowledge of dataset biases and limitations can be obtained by incorporating additional criteria like age representation, gender diversity, and cultural relevance.

C. Ethical Considerations

It is critical to address ethical issues about the gathering, storing, and use of data. To guarantee the appropriate and courteous management of face image data, future studies should examine ethical frameworks and rules unique to African contexts, especially about permission, privacy, and data security.

D. Algorithmic Fairness and Bias Mitigation

To overcome potential biases in facial recognition systems, more research is required into algorithmic fairness and bias mitigation strategies. Future research, considering the particular difficulties presented by varied African communities, should investigate novel strategies for reducing prejudices and advancing justice.

E. Real-world application and Deployment

Considering the practical difficulties and societal ramifications is crucial when facial recognition systems go from research to real-world application and deployment. To guarantee the ethical and appropriate application of facial recognition technologies in African situations, future study endeavors ought to concentrate on converting scientific discoveries into practicable tactics for legislators, practitioners, and technology developers.

F. Community Involvement and Capacity Building

Building trust, accountability, and inclusivity requires including stakeholders and local communities in the creation and application of facial recognition technology. To enable African communities to actively participate in influencing the development of facial recognition technology, future efforts should place a high priority on community involvement and capacity-building programs.

REFERENCES

[1] Aina, S., Adeniji, M. O., Lawal, A. R., & Oluwaranti, A. I. (2020). Development of a convolutional neural network-based ethnicity classification model from facial images. *International Journal of Innovative Science and Research Technology*, 7(4), 1216-1221.

[2] Ali, W., Tian, W., Din, S. U., Iradukunda, D., & Khan, A. A. (2021). Classical and modern face recognition approaches a complete review. Multimedia tools and applications, 80(4), 4825-4880.

[3] Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19 (pp. 289-295). ACM

[4] Bacci, N., Davimes, J., Steyn, M., & Briers, N. (2021). Development of the Wits Face Database: An African database of high-resolution facial photographs and multimodal closed-circuit television (CCTV) recordings. *F1000Research*, *10*.

[5] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

[6] Dhamecha, T. I., Singh, R., Vatsa, M., & Kumar, A. (2014). Recognizing disguised faces: Human and machine evaluation. PloS one, 9(7), e99212.

[7] Garcia, R. V., Wandzik, L., Grabner, L., & Krueger, J. (2019, June). The harms of demographic bias in deep face recognition research. In 2019 International Conference on Biometrics (ICB) (pp. 1-6). IEEE.

[8] Górriz, J. M., Ramírez, J., Ortíz, A., Martínez-Murcia, F. J., Segovia, F., Suckling, J., Leming, M., Zhang, Y.-D., Álvarez-Sánchez, J. R., Bologna, G., Bonomini, P., Casado, F. E., Charte, D., Charte, F., Contreras, R., Cuesta-Infante, A., Duro, R. J., Fernández-Caballero, A., Fernández-Jover, E., & Gómez-Vilda, P. (2020). *Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications*. *Neurocomputing*, *410*, 237-270.

[9] Iloanusi, O., Flynn, P. J., & Tinsley, P. (2022). Similarities in African Ethnic Faces from the Biometric Recognition Viewpoint. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 419-428).

[10] Kalaiselvi, P., & Nithya, S. (2013). Face recognition system under varying lighting conditions. IOSR Journal of Computer Engineering, 14(3), 79-88.

[11] Krishnan, A., Neas, B., & Rattani, A. (2022). Is facial recognition biased at near-infrared spectrum as well? In 2022 IEEE International Symposium on Technologies for Homeland Security (HST) (pp. 1-6). IEEE.

[12] Li, K., Chen, H., Huang, F., Ling, S., & You, Z. (2021). Sharpness and brightness quality assessment of face images for recognition. Scientific Programming, 2021, 1-21.

[13] Lin, C. (2020). Understanding cultural diversity and diverse identities. Quality education, 929-938.

[14] Liu, C., Lee, M. K., Naqvi, S., Hoskens, H., Liu, D., White, J. D., Indencleef, K., Matthews, H., Eller, R. J., Li, J., Mohammed, J., Swigut, T., Richmond, S., Manyama, M., Hallgrímsson, B., Spritz, R. A., Feingold, E., Marazita, M. L., Wysocka, J., Walsh, S., Shriver, M. D., Claes, P., & Shaffer, J. R. (2021). Genome scans of facial features in East Africans and cross-population comparisons reveal novel associations. *PLoS Genetics*, 17(8), e1009695.

[15] Muhammad, J., Wang, Y., Wang, C., Zhang, K., & Sun, Z. (2021). Casia-face-africa: A large-scale african face image database. IEEE Transactions on Information Forensics and Security, 16, 3634-3646.

[16] Muhammad, J., Wang, Y., Wang, L., Zhang, K., & Sun, Z. (2022, October). An Empirical Comparative Analysis of Africans with Asians Using DCNN Facial Biometric Models. In Chinese Conference on Biometric Recognition (pp. 138-148). Cham: Springer Nature Switzerland.

[17] Mundial, I. Q., Hassan, M. S. U., Tiwana, M. I., Qureshi, W. S., & Alanazi, E. (2020, September). Towards facial recognition problem in COVID-19 pandemic. In 2020 4th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM) (pp. 210-214). IEEE.

[18] Musa, Sadeeq. (2022). Using Machine Learning to Overcome Facial Recognition Bias in Africa. Global Scientific Journal 10(11), 1517-1526.

[19] Nothias, T. (2023). Facial recognition technologies in Africa: From deployment to advocacy. African Journalism Studies, 44(1), 1-18. https://pacscenter.stanford.edu/event/facial-recognition-technologies-in-africa-from-deployment-to-advocacy/

[20] Oloyede, M. O., Hancke, G. P., & Myburgh, H. C. (2020). A review on face recognition systems: recent approaches and challenges. Multimedia Tools and Applications, 79, 27891-27922.

[21] Perkowitz, S. (2021). The bias in the machine: Facial recognition technology and racial disparities, https://doi.org/10.21428/2c646de5.62272586.

[22] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications, and research directions. SN computer science, 2(3), 160.

[23] Schmidlin, E. J., Steyn, M., Houlton, T. M., & Briers, N. (2018). Facial aging in South African adult males. Forensic Science International, 289, 277-286.

[24] Udefi, A. M., Aina, S., Lawal, A. R., & Oluwarantie, A. I. (2023). An Analysis of Bias in Facial Image Processing: A Review of Datasets. *International Journal of Advanced Computer Science and Applications*, *14*(5).

[25] Uwizera, D., Bares, W., & Voss, C. (2024). Data-centric face recognition for African face authentication. *Journal of Computer Vision and Image Processing*, 14(1), 23-37. https://doi.org/10.1007/s00138-024-0112-5.

[26] Yucer, S., Akçay, S., Al-Moubayed, N., & Breckon, T. P. (2020). Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 18-19).