Scientific Research Publishing

Beijing Law Review, 2023, 14, 1000-1028
https://www.scirp.org/journal/blr
ISSN Online: 2159-4635
ISSN Print: 2159-4627

# Evaluating Open-Source Facial Recognition Software in Public Security: Effectiveness and Observations on Ethnicity

**Methanias Colaço Júnior[1,2*], Luan Bruno Barbosa de Souza Costa[2,3], Everton Carlos Santos Recchi[3], Ana Carla Bliacheriene[1], Fatima de L. S. Nunes[1], Luciano Vieira de Araújo[1]**

[1]School of Arts, Sciences and Humanities (EACH), University of São Paulo (USP), São Paulo, Brazil
[2]Postgraduate Program in Computer Science (PROCC), Federal University of Sergipe (UFS), São Cristóvão, Brazil
[3]Special Action Group to Combat Organized Crime (GAECO), Prosecution Office (MPSE), Aracaju, Brazil
Email: *mjrse@hotmail.com

**How to cite this paper:** Colaço Júnior, M., Costa, L. B. B. de S., Recchi, E. C. S., Bliacheriene, A. C., Nunes, F. de L. S., & Araújo, L. V. de (2023). Evaluating Open-Source Facial Recognition Software in Public Security: Effectiveness and Observations on Ethnicity. *Beijing Law Review, 14,* 1000-1028.
https://doi.org/10.4236/blr.2023.142054

**Received:** May 18, 2023
**Accepted:** June 23, 2023
**Published:** June 26, 2023



## Abstract

***Context***: In the criminal investigation environment, there is often a lack of information about a particular suspect, demanding instruments capable of searching for information based on limited evidence. Facial recognition, utilizing archived photos and/or real-time image capture, acts as one such instrument. ***Objective***: This study aims to analyze the facial recognition results of the an open-source and free product, evaluating its effectiveness through acceptable accuracy and sensitivity rates for investigators, with the intention of exploring its potential application in the field of public security. Additionally, an analysis of efficacy by ethnicity was performed, which discussed solutions to avoid racism. ***Method***: A controlled in vitro experiment was conducted, employing a dataset of approximately 20,000 authentic photographs of incarcerated individuals from the prison system of the state of Sergipe, Brazil. ***Results***: The effectiveness results obtained indicate that the open-source and free Face Recognition tool holds potential for identifying individuals in the context of front-view photos of inmates in the prison system and similar Public Security Management applications. Upon completion of the tests and taking into account statistical significance, the software successfully identified incarcerated individuals using their images, achieving an average accuracy rate of over 84.8%, a sensitivity rate of over 89.9%, and an fβ-measure of over 82.5%, in line with the established criteria. It is worth mentioning that replications of this experiment may also validate a better average for the accuracy rate, which reached, for many cases, the final level of 90% and even 100% precision. ***Conclusion***: The effectiveness results demonstrate the potential suitability of the facial recognition tool for identifying individuals, particular-

ly within the context of front-view photos of inmates in the prison system and similar Public Security Management applications. However, the study also revealed a higher rate of false positives among Black individuals, emphasizing the importance of addressing potential biases and refining the technology to ensure equitable treatment across all ethnic groups. Finally, this research contributes to ongoing discussions on harnessing the benefits of technology while mitigating potential negative consequences in law enforcement contexts.

## Keywords

Facial Recognition, Transparency, Racism, Public Agencies Security, Criminal Investigation

## 1. Introduction

In the context of criminal investigation, information is placed as a central asset in the workflow. Being crime a complex social phenomenon that acts in a continuous way through a web of multiple connections and interdependencies that perdures in time and space, the solution to a concrete case demands global knowledge of the phenomenon it emerges from (BRAZ, 2013). Just like human beings need oxygen to fulfill their vital functions, criminal investigation needs information to proceed with its purposes every moment. In this line of thought, we can quote the maxim by Sun Tzu (Tzu & Pin, 2015) "If you know neither the enemy nor yourself, you will succumb in every battle" to highlight the importance of information in the context of criminal investigation, being vital and decisive to interact with the socio-criminal environment to explore closed information networks, and to manage the sources properly (BRAZ, 2013).

As for the police investigation work, it is possible to verify the low capacity to produce the necessary proof for the conviction of the defendants. According to Adorno and Paisinato (Adorno & Paisinato, 2008), in a sociologic research carried out by the Center of Violence Studies of the University of São Paulo (NEV/Cepid/USP), aimed to answer questionings related to criminal impunity in the city of São Paulo, and starting from the hypothesis that the high rate of impunity compromises the credibility of the institutions in charge of enforcing the law to benefit citizens, it was possible to verify the weak police willingness to investigate crimes by unknown offenders. Therefore, we can observe that in fact few police reports become investigations (Adorno & Paisinato, 2008). This restricted course of action makes it impossible to really combat the increase of crime rates, and does not enable dealing with the changes provoked by the emergence of organized crime and the violation of human rights either (Vasconcellos, 2008). In this context, facial recognition presents itself as another tool for encouraging the initiation of well-supported investigations and for police engagement in the investigation of crimes by unknown offenders.

The literature has attested the beginning of efforts at researching the recognition of faces since the 19th century (Da Silva & Santa Rosa, 2004: p. 176). In 1878, the English scientist Sir Francis Galton presented an article at the Royal Anthropological Institute of Great Britain and Ireland in which he described his research involving the combination of photos of people by overlapping images of faces, one over the other, concluding that it could be possible to reach the photo that would present the typical traits he was searching for by reducing or eliminating the existing variations (Da Silva & Santa Rosa, 2004: p. 176).

Despite the fact that there are ethical, moral and constitutional discussions about the not-so-often consented use of the image of somebody else, as in (Conceição, Viana, & Rocha, 2019) and (Carvalho, 2020), several organs have been applying tools of this nature to identify potential criminals. More recently, the Secretariat of Public Security of the State of Bahia (SSP/BA), Brazil, found 42 fugitive inmates through its Facial Recognition System during the 2020 Salvador Carnival (Forato, 2020), an event that recorded the presence of approximately 16.5 million people during that year (Salvador Tourism Company, 2020). In all of the cases, the monitoring tool showed a similarity rate over 90%, according to the data presented by the headquarters of the State Police General Command, during a press conference about the popular street party. These numbers do not only point out the necessity, but also the effectiveness of the use of tools for this type of biometric identification.

In this article, a free and open source facial recognition tool is tested, in the contest of roughly 20,000 images of inmates in the prison system of the State of Sergipe. The application of this tool, given the proposed context, can result in benefits in the assistance to identify individuals within an ongoing investigative process, in case they have been incarcerated before. Even though the individual to be researched is not the final object of the investigation, understanding the relationship network in which he/she is inserted may be important for the success of the process (BRAZ, 2013), as in the identification of third parties in social media photos, for example. Another possible application, given the context, is the possibility to identify fugitives in raids or when entering public institutions that have access control, like courthouses, a fact that may prevent the occurrence of a crime or trigger an occasional apprehension, since the fugitive may go to such places in search of information.

Lastly, the fact that the tool is free and open source may signify a consistent cost reduction in the implementation of the aforementioned applications, besides enabling new implementations and evolution. After the execution of the experiment, it was possible to observe that the tool in question was able to detect faces with an average accuracy rate over 84.9%, sensibility rate over 89.9% and fβ-measure over 82.5%, given the established criteria. It is worth highlighting that replications of this experiment can also validate a better average for the accuracy rate, which reached the final level of 90% and even 100% precision.

Regarding the ethnicity viewpoint, there were more false positives in the images of Black individuals. It opens a debate about how to extract the most from

technology, by optimizing how police approach or non-approach people, without disadvantaging a specific ethnicity.

From now on, the article is organized in the following way: Section 2 encompasses two related works; Section 3 describes the work methodology; in Section 4 the conceptual basis for the development of the experimentation is presented; Section 5 describes the tool chosen as the object of the experiment; Section 6 presents the definition and planning of the experiment; the operation and results of the experiment are described in Sections 7 and 8, respectively; finally, Section 9 concludes with the findings of the experiment.

## 2. Related Works

Due to the variety of applications resulting from facial recognition activity, a series of studies have been elaborated to experiment and enhance their techniques. Kaufman and Breeding (Kaufman & Breeding, 1976: pp. 113-121) reported a recognition rate of 90% by using facial profiles. However, they used a database of only 10 individuals. Harmon et al. (Harmon et al., 1981: pp. 97-110) obtained a recognition accuracy rate of 96% in a database of 112 individuals by using a 17-dimensional feature vector to describe facial profiles and a Euclidian distance measure for the correspondence. More recently, Liposcak and Loncaric (Lipoščak & Loncaric, 1999: pp. 243-250) reported a precision rate of 90% in a database of 30 individuals by using subspace filtering to derivate a 21-dimensional feature vector, describing the profiles of the faces and deploying a Euclidian distance measure to match them.

Within the ambit of facial recognition for forensic activity, Da Silva and Santa Rosa (Da Silva & Santa Rosa, 2004: p. 176) proposed an automated facial recognition model based on Eigenfaces. The main purpose of that approach was to recognize people with semi-covered faces, wearing masks or another artifact, usually in crime scenes. Although in adverse conditions, the authors reported hit rates between 80% and 98%.

Still in the inmate recognition field, Celli (Celli, 1999) proposed an assistance system for the recognition of suspects by the victim, through ordering the presented photos, which means, in a similarity order, making the correct identification of the suspect less tedious and faster. The work consisted in the selection of techniques and in the creation of the system referred above, along with the modeling of a database, by including photos, in partnership with the State Police of the State of São Paulo. According to the authors, by using the algorithm PCA (Principal Component Analysis), the tests suggest a good degree of recognition and good adaptation to the task of ordering photo sets. However, there is not precision in the results of the tests, only presenting the description of the algorithm acceptance to the application, with the justification that the focus of the project is not to precisely determine if the photo is in the database, but only order the data by similarity. It is a situation similar to the application of this article, from the viewpoint of the investigation.

Besides not proposing any new techniques or methodologies for the facial

recognition task, this article differs from the mentioned works in the size of the database for the achievement of the research study. For example, whereas the work of Da Silva and Santa Rosa (Da Silva & Santa Rosa, 2004: p. 176) used a database of 120 images, and the work of Celli (Celli, 1999) used 640 images of inmates—according to **Table 1**—for the validation of the supposed hypotheses, the present article used a significant larger database, in the order of 20.000 individuals.

Another factor observed when analyzing **Table 1** is that the facial recognition accuracy rate observed in the found works is equal or superior to 90%, in their majority. The exception is the work of Da Silva and Santa Rosa (Da Silva & Santa Rosa, 2004: p. 176), which reported hit rates between 80% and 98%.

It is also worth mentioning that works that performed an analysis of the facial recognition tools for forensic application by considering an experimental approach with the statistical validation of the data significance, as it is proposed in this article, were not found. A robust knowledge base (Big Science and Big Data Science) can only be generated with the replications of real controlled experiments that statistically validate their works, which can be used as input for real data meta-analyses.

Lastly, regarding the fact that the current article may refer to a tool developed by third parties, contrary to other works with forensic application that developed their own tools, it will not result in any increase in cost in an occasional practical application, since it is a free tool and is available in open source, according to Section 5. Moreover, it eliminates conflict of interests by mitigating bias and strengthening evaluation without characterizing demand, which are confusing factors and serious threats to the internal validity of any experiments.

## 3. Methodology

At first, an exploratory research was carried out (Severino, 2017), in order to verify the state of the art of facial recognition and of the quality metrics used for the evaluation of the results of the experiment, as well as for listing free and open source solutions in this area. The synopsis of the material used as reference is shown in Section 4.

Next, the application to be submitted to the experiment was selected. The tool named Face Recognition, available at https://github.com/ageitgey/face_recognition, was chosen by meeting the simple choosing criteria determined by the investigators of the District Attorney Office of Sergipe and by the authors of this article:

**Table 1.** Comparison among recognition works.

| Work | Images | Accuracy |
|---|---|---|
| Kaufman & Breeding, 1976 | 10 | 90% |
| Harmon et al., 1981 | 112 | 96% |
| Lipoščak & Loncaric, 1999 | 30 | 90% |
| Da Silva & Santa Rosa, 2004 | 120 | 80% - 98%. |
| Celli, 1999 | 640 | Not informed |

open source, zero cost, the best evaluation from clients, and simple interface. More details about the application and the selection procedure will be described in Section 5.

The following step was to prepare the database. The photo set used in the experiment was obtained through the SAP—"Sistema de Administração Penitenciária"—Prison Administration System of the State of Sergipe. For each inmate, there were at least three images, being one front view and the others side view. As shown in (Mahoor & Abdel-Mottaleb, 2009), one of the factors that reduces the quality of the facial recognition process is the rotation of the head over its horizontal axis, called PM (pose measure) by the authors. Thus, the next stage was to remove the side-view photos of the inmates.

Still in the preparation of the database (Colaço et al., 2019), once the described set contains only one front-view photo of each inmate, front-view photos of other individuals that the authors had access to and could be used in the identification test of the experiment were inserted into the experimentation database. After selecting the images, they were divided into blocks and then uploaded into the tool in order to proceed with the training. The amounts and details of the images will be elucidated in the definition of the experiment planning in Section 6.

Afterwards, the tests were carried out in an in vitro environment, since the material was removed from the original system to be manipulated in a controlled environment. In this context, the research study presented in this article was also classified as a laboratory and experimental one, due to the planning and execution of a controlled experiment.

The experiment planning is better described in Section 6, while the preparation process and execution of the experiment is detailed in Section 7.

## 4. Conceptual Basis

Some concepts for the understanding of this article are presented in this section.

### 4.1. Criminal Investigation

The Brazilian Constitution of 1988 makes important constitutional guarantees in its article 9. Among them, item XXXV says that "the law shall not exclude injury or threat to the right from the assessment of the judicial power" (BRASIL, 1988). Thus, it is mandatory that the State assess infringements of the law, making itself a sovereign entity, holder of the power and the right to punish when someone violates a criminal rule, as long as the guarantee to the due legal process present in the Brazilian Constitution of 1988 is respected. And for this purpose, seeking elements that corroborate the existence of a crime and who violated the criminal rule (Tourinho Filho, 2013).

Once the judicial order is harmed when a crime or criminal offence is committed, it is up to the State to thoroughly investigate the occurred event, clarifying it in all its circumstances and unveiling all its consequences. Such procedure

to elucidate the facts is called Criminal Investigation (Oliveira, 2014). With the news of the crime, the investigative entity starts to perform investigations to recover the event and find the possible perpetrator (Tourinho Filho, 2013). The purpose of criminal investigation is to build a previous evidence frame, which justifies the penal action on behalf of the minimum safety that is required for the state activity against someone in the crime field. On behalf of the dignity of a human being, a rule-based Democracy in all areas, chiefly in Criminal Law and Criminal Procedural Law, is sought. That is the reason why it is neither permitted to attack an individual by investigating his/her private life—which is naturally guaranteed by the constitutional right to intimacy—and nor to institute legal proceedings against someone without minimum proof, in a way to instruct and sustain not only concreteness (existence proof of the penal infraction), but also sufficient evidence (reasonable proof that the individual is the perpetrator or reasonable proof of misdemeanor) (NUCCI, 2012).

Crime investigation has administrative nature and is conducted before the legal proceeding, which means, in the pre-procedural phase, being used to persuade the individual responsible for the criminal charge. The judge must not be biased and must not assess any evidence, and must intervene only to halt any threats or harm to individual rights and guarantees, to the effectiveness of jurisdiction, and must guarantee constitutional and legal principles (Oliveira, 2014). The State has the obligation to guarantee collective safety and order, but it is halted in all of its actions by the observance of legality and transparency, stated in the Brazilian Constitution of 1988 (NUCCI, 2012).

In summary, the purpose of criminal investigation is to investigate the existence of a crime and the perpetrator, elements that constitute just cause for the legal proceeding, corresponding to the procedural interest, which is the condition of the action and an essential element to the right of action (Greco Filho, 2015).

At this point, one of the contributions of the present article stands out, once it analyses a helping tool for the identification of the person who committed an offense. Moreover, as criminality acts through a web of multiple connections and interdependences—and whose resolutions of the concrete cases demand the global knowledge of the phenomenon from which they emerge—the identification of individuals who somehow are associated with an illegal practice, even in a passive manner, can help elucidate elements and settle down issues that were previously unidentified and that are necessary to close cases.

### 4.2. Facial Recognition

Biometry-based techniques have recently appeared as the most promising option for individual recognition, rather than password-based authentication, PIN (Personal Identification Number) or cards, which have to deal with problems like theft and loss. Biometry-based technologies include identification based on physiological features, such as face, fingerprints, finger geometry, hand geome-

try, palm, iris, retina, ear, and voice (Jafri & Arabnia, 2009).

Facial recognition appears to offer several advantages over other biometric methods. Almost all the other technologies require some voluntary action from the user. For example, recognizing the fingerprints, when the user needs to put his finger on a specific digital reader. Or in another similar situation, when the user must stand in front of a camera in a specific position, so the retina can be read. This characteristic is of the utmost relevance to the present article, considering that the right to not produce proof against yourself (nemo tenetur se detegere), consecrated by the Brazilian and international legislation (Queijo, 2017), can prevent the collection of some biologic data. In this context, facial recognition can be performed passively, without any explicit actions from the user, enabling the capture to be made from distance, though a camera.

Furthermore, facial recognition has a lower operational cost than iris, retina, and fingerprint analyses, which unlike the tool that is the object of study in this article, require the acquisition of additional equipment. Voice identification, despite having low cost of capture, tends to present many background noises, in public places mainly, or during the recording of a telephone call (Jafri & Arabnia, 2009).

In general, the problem with biometric identification based on facial recognition can be formulated as: given a facial image as input and an image database of known individuals, how to verify or determine the identity of the person in the input image? The main difficulty of this problem results from the fact that, in the most common types of pictures (front-view photos), the faces seem to be reasonably alike and the differences among them are very subtle. Consequently, front-view images form an extremely dense cluster in an image space (through proximity), which makes it virtually impossible for traditional standard recognition techniques to discriminate them with accuracy and a high degree of success (Nastar & Mitschke, 1998).

Besides, the human face is not a unique and rigid object. Actually, there are several factors that make the appearance of the face vary. The sources of facial appearance variation can be categorized in two groups: intrinsic and extrinsic factors (Gong, Mckenna, & Psarrou, 2000). 1) Intrinsic factors are related to purely physical nature and are independent of the observer. These factors can be divided into two groups: intrapersonal and interpersonal (Jebara, 1995). Intrapersonal factors are responsible for varying the personal appearance of the same person, being age, facial expression and facial aspects (facial hairs, glasses, cosmetics, etc.) some examples. On the other hand, interpersonal factors are responsible for the differences in the facial appearance of different people, being ethnicity and gender some examples. 2) Extrinsic factors make the aspect of the face alter through the interaction of light with the face and the observer. These factors include illumination, pose, scales, and imaging parameters (for example, resolution, focus, image, noise, etc.). Even though the majority of the recent facial recognition systems works well under restrict conditions (it means, scena-

rios in which at least some of the factors that contribute to the variability among facial images are controlled), the performance of the majority of these systems degrades rapidly when they start working under conditions in which none of these factors are regulated (Yang, Chen, & Kunz, 2002).

Mohamed Abdel-Mottaleb and Mohammad H. Mahoor (Mahoor & Abdel-Mottaleb, 2009) conducted a work in which three factors that negatively affect the performance of facial recognition systems were analyzed. The first one of these factors is connected to the blurring degree. The second factor is linked to the effect of luminosity. The third factor used by the authors refers to the pose of the head, represented by the measure PM (pose measure). From this measure, it is possible to estimate the angle regarding the rotation of the face around a vertical axis that passes through the center of the head. This last factor represents a critical attribute in the context of facial recognition, due to the fact that the bigger this angle, the more the image of the face will be uncharacterized. For this reason, side-view photos of the inmates were removed from the experiment described in the present article, as explained in Section 3.

Depending on the methodology for the collection of facial data, facial recognition techniques can be widely divided into three categories: methods that operate on the intensity images, the ones that deal with video sequences, and the ones that demand other sensorial data, like 3D information or infrared images. In this article, we emphasize the group of techniques that operate on the intensity images.

The facial recognition method through intensity image can be divided into two main categories: feature-based and holistic (Brunelli & Poggio, 1993), (Maxim, 2000) and (Heisele et al., 2003). The holistic approaches try to identify faces by using global representations, which means, descriptions based on the entire image and not on local features of the face, and can be subdivided into two groups: statistical approaches and AI.

Firstly, feature-based approaches process the input image to identify and extract (and measure) distinct facial features such as eyes, mouth, nose, and other fiducial marks by calculating the geometric relations among these facial points, and then reducing the input facial image to a geometric-feature vector. Standard techniques for the recognition of statistical patterns are then employed to match faces by using this measure.

One of the most known techniques in the feature-based approach is the Elastic Bunch Graph Matching, proposed by Wiskott (Wiskott, 1997). A graph for an individual face is generated this way: a set of fiducial points on a face is chosen.

Each fiducial point is a completely connected graph node, according to **Figure 1**, and is labeled with the answers from the Gabor filters applied to a window around the fiducial point. Each arch is labeled with the distances between the correspondent fiducial points. A representative set of such graphs is matched with a battery-like structure called face graph. Once the system has a face graph, the graphs for new facial images can be automatically generated by the Elastic Bunch Graph Matching. The recognition of a new image of the face is performed

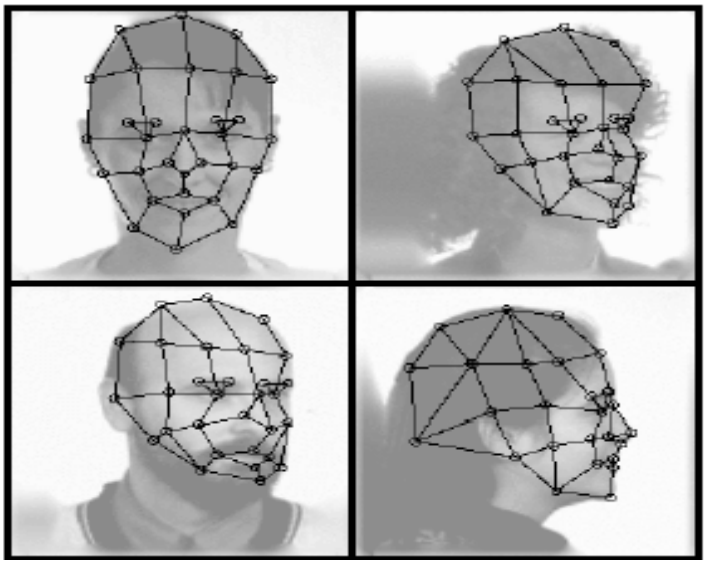

**Figure 1.** Facial recognition grids (Fellous, Kruger & Von Der Malsburg, 1996).

by comparing the graph of the image with the one of all the known facial images, and then by choosing the one with higher similarity value. By using this architecture, the recognition rate can reach 99% (Jafri & Arabnia, 2009).

The main advantage provided by the feature-based techniques, such as the technique used by the tool analyzed in this article, is that as the extraction of the points precedes the analysis performed to match the image with the one of a known individual, these methods are relatively robust to position variations in the input image (Jebara, 1995). On the other hand, the main disadvantage is the difficulty of automatically detecting features, besides the fact that the implementer of any of these techniques needs to make arbitrary decisions about which features are important (Phillips et al., 1998). For example, it is necessary that the implementer determine that the distance between the eyes is a relevant feature in the context of identification, because it presents itself in different forms for each individual. Having done it, this distance will also be accounted for in the similarity calculation between the two faces.

As shown in Section 5, the tool used for the experimentation already has the definition regarding which features will be identified, as well as their detection in the images. Moreover, as it is an open source tool, these implementation details can possibly be revised by the institution that adopts it. The section below summarizes the factors that must be overcome by this type of tool.

### Factors That Make the Appearance of the Face Vary

Understanding the factors that influence the appearance of the face and their effects on facial recognition algorithms is crucial for developing more reliable and robust systems. By addressing the challenges posed by lighting, facial expressions, makeup, aging, and pose variations, researchers strive to enhance the precision and performance of facial recognition algorithms.

#### Lighting

Lighting plays a crucial role in facial appearance, as it can create shadows, highlight specific features, and alter the overall appearance of the face. Changes in lighting conditions can pose challenges to accurate face recognition by algorithms. Yang et al. (2017) discuss the importance of addressing pose robustness in face recognition systems, emphasizing the impact of lighting variations.

**Facial Expressions**

Facial expressions, such as smiles, frowns, and eye blinks, can modify the shape of the face and potentially affect the accuracy of facial recognition algorithms. Lyons et al. (1998) explore the coding of facial expressions using Gabor wavelets, highlighting the relevance of expressions in facial analysis.

**Makeup and Accessories**

The application of makeup, wearing glasses, hats, and other accessories can alter the appearance of the face, making it more challenging for facial recognition algorithms to accurately identify individuals. Patel et al. (2016) investigate the sensitivity of automatic recognition systems to facial cosmetics, shedding light on the impact of makeup on recognition performance.

**Aging**

As individuals age, significant changes occur in facial features, such as the emergence of wrinkles, loss of skin elasticity, and overall facial shape alterations (Fu et al., 2010).

**Pose Variations**

Varied poses, including different viewing angles, head tilts, and rotations, can impede precise feature matching between extracted facial characteristics and trained models (Gross et al., 2010).

## 4.3. Quality Metrics

In this article, the metrics accuracy, sensitivity, precision, and fβ-measure were used. For the comprehension of these measures, it is necessary to assume that:

1) TP (True Positive)—The total of instances of individuals present in the database that were recognized correctly;

2) TN (True Negative)—The total of instances of individuals **not** present in the database that were **not** recognized;

3) FP (False Positive)—The total of instances of individuals that were recognized, but are **not** present in the database;

4) FN (False Negative)—The total of instances of individuals that are present in the database, but were **not** recognized.

### 4.3.1. Accuracy

Accuracy is the proportion of correct predictions, by not taking what is positive and what is negative into consideration (Zhu et al., 2010). It is calculated as the number of all the correct predictions divided by the total number of the data set (Saito & Rehmsmeier, 2017), and is defined by:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

#### 4.3.2. Precision

Precision id the proportion of true positives in relation to all the positive predictions (Zhu et al., 2010). It is calculated by the number of correct positive predictions divided by the total number of positive predictions (Saito & Rehmsmeier, 2017), and is defined by:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

#### 4.3.3. Sensitivity

Sensitivity is the proportion of true positives, which means, the capacity of the system to predict the condition for cases that really have it correctly (Zhu et al., 2010). It is calculated by the number of correct positive predictions divided by the total number of positives (Saito & Rehmsmeier, 2017), and is defined by:

$$\text{sensibility} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

#### 4.3.4. F-Measure

The F-measure is defined as the harmonic mean of the precision and sensibility rates of the test. In fact, the F-measure can also be called the F1-measure, because it is fβ-measure (see nest subsection), with Beta equal to 1. This implies a score calculated according to:

$$F = \frac{2 \cdot \text{precision} \cdot \text{sensibility}}{\text{recision} + \text{sensibility}}$$

#### 4.3.5. fβ-Measure

The calculation of the fβ-measure is similar to the one of the F1-measure. However, in this case, it enables the users the possibility to attribute a different relative meaning to precision and sensibility. It measures information retrieval effectiveness in relation to a user that attributes ß times more importance to sensibility than to precision (Van Rijsbergen, 1979). A lower beta value, such as 0.5, gives more weight to precision and less to sensitivity, whereas a higher beta value, such as 2.0, gives less weight to precision and more weight to sensitivity in the calculation of the score. The measure is calculated according to the formula defined by:

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{\text{precision} \cdot \text{sensibility}}{\left(\beta^2 \cdot \text{precision}\right) + \text{sensibility}}$$

## 5. Face Recognition

The Face Recognition application is a free, open source software, available at https://github.com/ageitgey/face_recognition. Besides facial recognition, the application still provides other tools for the manipulation of facial images. For the current article, the source code of the software was not analyzed or modified, which means, the application will be treated as a black box, in order to concentrate the efforts in the comprehension of its user interface.

Some search criteria were established in the selection process of this tool. As seen in Section 1, a free and open source tool might signify a consistent reduction in the implementation cost, besides enabling new implementations and evolutions. Moreover, having a simple interface also reduces its implementation complexity and use. Therefore, a tool on the GitHub platform that met the criteria above was sought: 1) free, 2) open source, and 3) documentation on how to use the interface. GitHub, in turn, is a development platform where developers provide application source codes in a free and open source way, besides enabling them to review and manage projects of third parties (GitHub, 2020). Once reviewed, it is possible for more than 50 million of users to review the hosted applications.

Once the search on GitHub would already result in applications that met the two first criteria, the research was carried out by using the search string "Face Recognition". The resulting search list was headed by the selected application, since it answered the string search perfectly, besides the fact that GitHub uses the number of positive reviews of each application as a display criterion. Nowadays, the selected software has more than 35 thousand positive reviews, whereas the second best reviewed software resulting from the search has a little more than 13 thousand. Immediately afterwards, through the documentation available on GitHub itself, it was also verified that Face Recognition met the third requirement, which is having a documentation on how to use the interface.

The system, developed by using the Python language adopts feature-based recognition, according to Section 4, being able to identify 128 fiducial points and perform the Euclidian distance calculation in the image passed as parameter, and the vectors in the pre-logged images. The system performs the distance calculation for each face in the database, whose result is a real number between 0 and 1, being 0 the minimum distance and 1 the maximum distance.

Once the distance between the faces is calculated, the system returns the ones whose result of the calculation was below a pre-established threshold, denominated **tolerance**. Then, tolerance value can be defined by the user of the system. By the definition of the own author of the software, the tolerance value is 0.5. Defining inferior values will implicate in having a more restrict software in relation to proximity, which means, the system will only return faces that are relatively closer to the original one, a fact that might result in the increase of occurrences of False Negatives. On the other hand, increasing tolerance level might implicate in an increase of False Positives, once the number of returned faces will be higher. For the present experiment, test runs were performed by using tolerance values close, inferior or superior to the tolerance level suggested by the author.

Among the main limitations of the application, according to the author of the system (GitHub, 2020), is the low accuracy rate for non-European individuals, especially Asians. Another reservation expressed by the author is the deficiency of the model to recognize infant faces.

## 6. Definition and Planning of the Experiment

In this and in the following section, this article is presented as an experimental process. It follows the guidelines of Colaço Júnior et al. (2022).

### 6.1. Definition of the Purpose

The purpose of the experiment, formalized by using the GQM model (Goal, Question, Metric), proposed by Basili and Weiss (Basili & Weiss, 1984), is: To **Analyze** the facial recognition results of the product Face Recognition through a controlled **experiment in order to** evaluate it, **regarding** its effectiveness **from the viewpoint** of programmers, police investigators and researchers, **in the context of** the front-view photos of the inmates in the prison system of the State of Sergipe, by considering a similarity measure threshold that represents acceptable accuracy and sensibility rates for the investigative process.

### 6.2. Planning

#### 6.2.1. Context Selection

The context chosen to carry out the experiment is initially a database that contains 40.226 images of inmates in the prison system of the State of Sergipe, as well as their identification. They were collected at the moment of imprisonment, being front-view captured in a similar way, and apparently by different cameras, since the images have different resolutions.

The images were collected via the SAP (Prison Administration System) of the Department of Public Safety of the State of Sergipe and provided to the GAECO (Special Action Group to Combat Organized Crime) of the MPSP (Prosecution Office of the State of Sergipe) through a cooperation agreement. At the same time the present experiment seeks to evaluate the capacity of the tool Face Recognition to be used in the investigatory proceedings at the GAECO of the MPSP, its immediate aptitude to be used in other organs of the same nature, such as the GAECOS of other states and the investigative police, is expected.

#### 6.2.2. Hypotheses Formulation

By aiming at the real use of the algorithm in a real criminal investigation environment, this article seeks to verify how precious its proximity calculations are, in order to determine the viability of its use in an application. Bearing in mind that, many times in the context of criminal investigation, the available information about a specific individual—the object of investigation—is scarce, it would already be acceptable that the application module that will implement the facial recognition algorithm was capable of reducing the search space for individuals drastically, based only on a photograph. For this experiment, it is considered acceptable that the recognition process results in a list of up to **ten individuals**, because the investigator usually searches for more detailed information about the suspects in other databases. Therefore, the module should provide an image as input and return a reduced list of people, whose facial similarity calculations present high proximity rates, over a previously established threshold.

As observed in the related works in Section 2, the majority of the related works found reported accuracy rates equal or superior to 90%. Moreover, since the primary purpose in the real use of the application is the reduction of the search space by the investigator, a facial recognition software that does not have an accuracy rate necessarily close to or equal to 100% will not be discarded. This condition would be necessary if the application were used for control access, for example. Nevertheless, through well-defined procedures and polished and polite approaches within legal standards and within each organization, an accuracy rate of about 90% can be interesting for access, accompanied by a more accurate precision rate.

Due to this information, the hypotheses formulated in this experiment consider the threshold observed in the related works, which means, the experiment intends to refute the following null hypotheses (H0), accompanied by their respective alternate hypotheses (H1), which seek the minimum goal of 90%, by considering only one decimal point:

- $H_0$: Given an input image, the Face Recognition software is capable of identifying the associated person, within the **ten** most similar faces to the original one, with an accuracy rate equal to **89.9%**.
- $H_1$: Given an input image, the Face Recognition software is capable of identifying the associated person, within the **ten** most similar faces to the original one, with an accuracy rate superior to **89.9%**.

At the same time, given the main purpose of the application—which is the reduction of the search space for inmates—it is important that the application return results in a way to reduce the occurrences of False Negatives (FN) to the detriment of a possible increase in cases of False Positives (FP). It means, given the context, similar to what occurs in medical applications—in which it is more important to not neglect a sick patient than to reduce the number of false positives—sensitivity (the most sensitive classifier) has a higher importance degree than the one of precision, the reason why a precision hypothesis will not be tested separately in this experiment. To gauge this measurement, we will also analyze the results of the sensitivity calculations through the following hypotheses:

- $H_0$: Given an input image, the Face Recognition software is capable of identifying the associated person, within the **ten** most similar faces to the original one, with a sensitivity rate equal to **89.9%**.
- $H_1$: Given an input image, the Face Recognition software is capable of identifying the associated person, within the **ten** most similar faces to the original one, with a sensitivity rate superior to **89.9%**.

As mentioned before, given the main purpose of the application, the occurrence of False Positives in the facial recognition process is acceptable to a determined degree in a way to minimize the occurrence of False Negatives. In order to evaluate precision and sensitivity rates thoughtfully, we will also analyze the results of the fβ-measure calculations, with beta equal to 3 (it means, sensitivity was considered three times more important than precision, see section 4.2.5),

through the following hypotheses:

- $H_0$: The Face Recognition software is capable of identifying an associated person within the **ten** most similar faces to the original one, by reaching fβ-measure with beta equal to 3, with a rate equal to **89.9%**.
- $H_1$: The Face Recognition software is capable of identifying an associated person, within the **ten** most similar faces to the original one, by reaching fβ-measure with beta equal to 3, with a rate equal superior to **89.9%**.

#### 6.2.3. Participants Selection

The database to conduct the experiment comes from the SAP (Prison Administration System), which stores data of the detained and incarcerated individuals in the State of Sergipe. At the moment of conducting the tests, the database had 40.266 images of inmates, being a part of them front-view photos, and the other part side-view ones. Therefore, it was necessary to select the images that would be used in the experiment. The side-view photos were removed for identification, since one of the factors that reduces the quality of the facial recognition process is the rotation of the head over its horizontal axis, as shown in section 4. After removing the photos, a total of **20.557** images was accounted for.

Due to the fact that there is only one front-view photo of each inmate, these images cannot be reused in the facial recognition tests, what would generate misleading accuracy bias because of the use of the same photos from the training database for identification. In other words, verifying facial similarity from two identical photos returns extremely close to or equal to 0 distance data. In this context, at least two photos for the execution of the algorithm are necessary: one for the training and the other for the test. Therefore, the images of the inmates from the SAP will be used in the training phase of the application and will be used as a simulation of a real application, in which images of ordinary citizens are compared with the inmate databases.

Due to the lack of photos of the inmates for the algorithm test phase, **50** images of **50** people were artificially inserted into the training database, being one front-view photo in the same angle and facial expression as the ones from the SAP for each individual. These new front-view images were personal photos of collaborators, spontaneously provided by them, or photos of public figures gathered via Google Images (Google Images, 2019), being the criterion for choosing the public figures random. Google Images is a search engine belonging to Google that enables the users to search for image content on the Web. At the time, other **50** non-front-view images of the same people (collaborators and public figures) were collected from social media to be used in the test phase. Choosing photos from social media in the test phase is coherent with one of the possible applications of the recognition software, according to Section 1. Thus, the front-view photos were inserted into the photo database of the SAP to train the algorithm, and the photos from social media were used for the similarity calculation in the algorithm, which means, the test.

Lastly, 50 images of other public figures that were not inserted into the train-

ing database were also collected from social media, according to the previous paragraph. This image set was also used for facial recognition tests, but for these cases the algorithm should return a null list of people as an ideal answer, once there are no images of these people in the original database. Just like when choosing the images to compose the training database, the criterion for the selection of public figures to compliment the test database was random. Considering the rigor and embarrassment that a False Positive can generate, mainly in applications to control access to public institutions, the return of any individual from the input of any one of these 50 images will characterize a False Positive.

Even though the training database has a majority of Black and brown-skinned men, this ethnic and gender proportion was neither replicated in the selection of the additional training images nor in the selection of the test images. Ideally, a test should be performed (according to) the proportions of data input occurred in a real environment. However, it was neither possible to use the investigation database for tests and nor to automatically define the input percentages by ethnicity, sex, and culpability, since that not all input information currently contains structured digital data on the ethnicity and/or pictures with validated identities and reclusion time.

On the other hand, although it is not the purpose of this work, and aiming to foment new works and to investigate, at initial levels, the possibility of disfavoring the Black ethnicity regarding the presence of more False Positives for this ethnicity because of disproportion in its favor in the inmate training database, the test images were labeled and divided between white and Black people (or brown-skinned). This division was made after the analysis of each image individually and manually, without the assistance of an automated classifier for example. Through this division, it was also possible to analyze the quality metrics by ethnicity.

#### 6.2.4. Independent Variables

According to the selection criteria described in Section 5, the Face Recognition software, as well as different values for the tolerance level, were selected for this experimentation. Besides these variables, the 20.577 photos of the inmates are also Independent Variables, besides the 50 photos inserted artificially, and the 100 photos used for testing.

#### 6.2.5. Dependent Variables

They refer to the classifications of the algorithm (list of people for each input photo) from which the metrics Accuracy, Precision, Sensitivity, F-measure, and fβ-measure can be derived.

#### 6.2.6. Project of the Experiment

The stages of the experiment can be visualized in **Figure 2**. Firstly, the 20.577 images from the Prison Administration System will be selected. The 50 images of third parties will be added to them, according to Section 6.2.3. Then, the image set will be uploaded into the software for training.

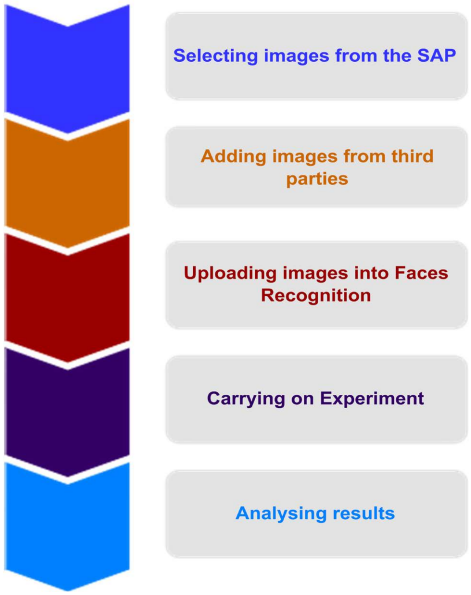


**Figure 2.** Stages of the experiment.

After selecting 50 images that are not in the system knowledge model, but are of the same individuals that were artificially added to the training base, these were joined to 50 images of unknown individuals, not present in the knowledge and training base, according to section 6.2.3. The 100 test images were divided into 10 clusters with 10 images each one: 5 different images of known individuals and 5 images of unknown individuals to the system. This way, the dependent variables will be extracted from each cluster, considering the calculation of the metrics based on the accuracy rate of the predictions in each cluster, and consequently will generate 10 data points, or 10 measures for each metric. The division in 10 clusters is considered a minimum size (Nadeau & Bengio, 2000) for the statistical significance of the difference calculation between the goal of 90% and the averages of each metric.

After that, the experiment was conducted and then the results could be analyzed.

#### 6.2.7. Instrumentation

The Face Recognition application will be used for the conduction of the facial recognition process, as described in Section 5.

A small Java software was developed to initially feed the application with the photos of the inmates by using the IDE Eclipse (Eclipse Foundation, 2019), which inserted the **20.577** images interactively.

The Postman application (Postman Inc., 2019) was used to insert the remaining artificial images, as well as to execute the algorithm. Postman is a platform for assisting in the development and use of APIs tests via HTTP requisitions.

## 7. Operation of the Experiment

### 7.1. Preparation

The preparation consisted in the training of the 20.577 images of inmates added

to the 50 images of third parties, according to Section 6.2.3. Also, initially, just for ambiance and to calibrate the best tolerance for each metric, using the final 100 test images and another 50 random images, metrics were collected for various tolerance values.

### 7.2. Execution

Once the application was trained, the testing image clusters were subjected to the evaluation of the algorithm, with the calculation of the respective similarities between the testing images and the images from the training database.

As mentioned in Section 5, the identification of the individual in the Face Recognition App admits a tolerance value as input, which is a real value between 0 (zero) and 1 (one), and this value will determine the minimum proximity that an image must have in relation to the input image in order to be considered similar by the algorithm. 0.5 is the standard value. Therefore, the tests were carried out several times, evaluating the best results for the different tolerance values. The results of this process will be described in Section 8.

### 7.3. Validation of Data

After carrying out the experiments, and in order to verify the supposed hypotheses, it was necessary to evaluate the samples obtained from the resulting dependent variables. To this end, the Shapiro-Wilk test (Shapiro & Wilk, 1965) was used. This test verifies whether a specific sample follows a normal distribution.

Once data normality was detected, two tests were performed to verify the supposed hypotheses. The t-Student test (Mankiewicz, 2000) was used for the sample sets with normal distribution. The t-Student test is a hypothesis test that uses statistical concepts to reject or not a null hypothesis, being robust for small or large samples.

The Wilcoxon test (Wilcoxon, 1992) was used to validate the sample tests with abnormal distribution. The Wilcoxon test is a non-parametric method to compare two paired samples, and can be also used as an alternative for the t-Student test when it is not possible to assume that the sample is distributed normally. The whole experiment adopted an α significance level of 0.05, which means, a confidence level of 95%. The results of this analysis are described in Section 8.

## 8. Results

### 8.1. Analysis and Interpretation of Data

Once the application calibration is executed with the balanced test cases mentioned plus another 50 random images (total of 150), the graph for the measures of the dependent variables in relation to the tolerance values is shown in **Figure 3**. It is worth highlighting that values under 0.4 and over 0.65 produced terrible results, and will not be listed.

After the result analysis per cluster, the best accuracy rates were found while executing the algorithm with a tolerance value equal to 0.5, being its measure

equal to 90%. A better F1-measure was also observed in this scenario, with an average equal to 88.67%. However, the best sensitivity rate and fβ-measure, whose averages reached 96% e 89% respectively, were observed when defining a tolerance level of 0.55. Lastly, when reducing the tolerance level to the value of 0.45, we found the best average precision rate possible, equal to 100%.

Table 2 also illustrates the results of the dependent variables per cluster, as well as the value of W and the calculated *p*-values to verify normality, according to the Shapiro-Wilk test (Shapiro & Wilk, 1965).

It is possible to observe that the normality test for the accuracy samples resulted in a *p*-value under 0.05, which is necessary for the rejection of the hypothesis that supposed the normal distribution. The same thing happened to the sensibility samples. However, it is possible to consider the normality in the sample sets of the fβ-measure through the Shapiro-Wilk test, once the result of its *p*-value was over the threshold of 0.05.

Once normality is verified, the next step is the previously mentioned scientific hypothesis test. As for accuracy, the null hypothesis assumes that the software is capable of identifying an individual through his/her image with an accuracy rate equal to 89.9%, while the alternative hypothesis assumes a higher accuracy rate. As seen in the previous paragraph, the accuracy samples resulting from the experiment in the ten tested clusters did not follow a normal distribution.

Therefore, the Wilcoxon test (Wilcoxon, 1992) was used to validate the hypothesis. This test uses the median as operational variable, or in other words, the distribution of data around a median. The operational null hypothesis starts to be H0: Median (metric) = 89.9.

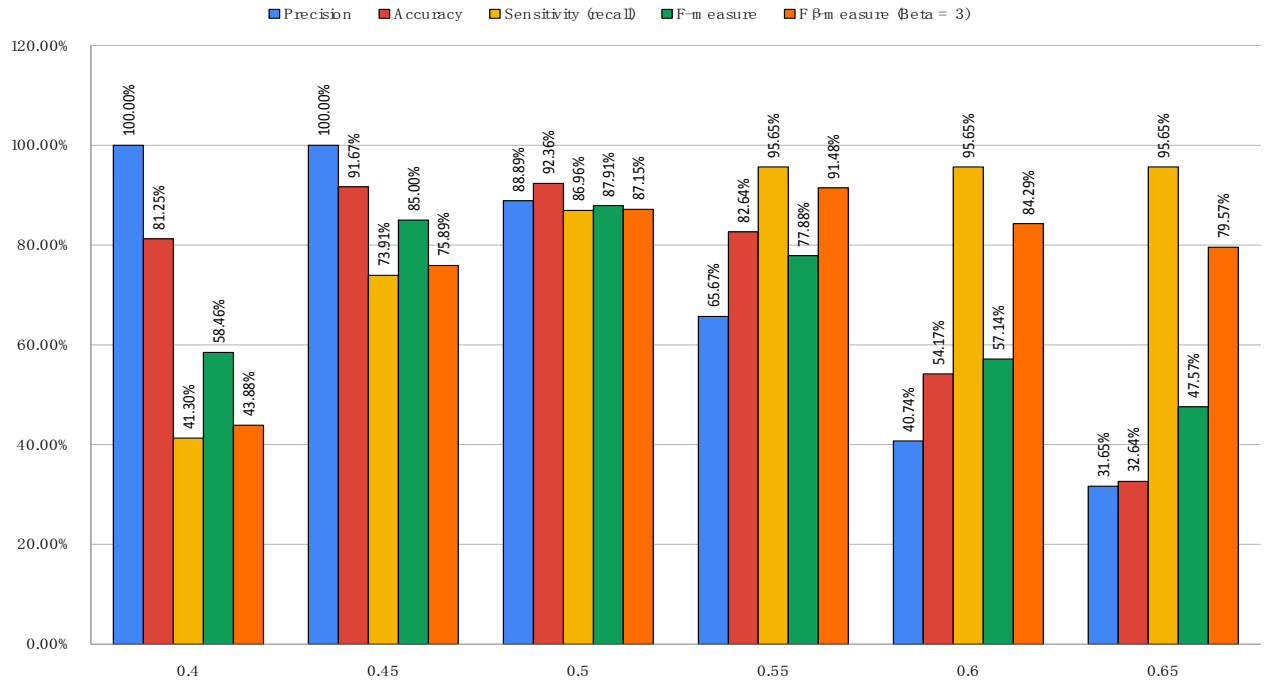


**Figure 3.** Precision, accuracy, sensitivity, F1-measure, and F-beta measure.

Table 2. Results by cluster and normality test of the data to be evaluated.

| Cluster | Precision (threshold = 0.45) | Accuracy (threshold = 0.5) | Sensibility (threshold = 0.55) | F1-measure (threshold = 0.5) | F-beta measure (β = 3) (threshold = 0.55) |
|---|---|---|---|---|---|
| 1 | 100% | 80% | 60% | 75% | 63.16% |
| 2 | 100% | 90% | 100% | 90.91% | 80% |
| 3 | 100% | 100% | 100% | 100% | 100% |
| 4 | 100% | 70% | 100% | 72.73% | 86.96% |
| 5 | 100% | 100% | 100% | 100% | 95.24% |
| 6 | 100% | 70% | 100% | 57.14% | 95.24% |
| 7 | 100% | 100% | 100% | 100% | 90.91% |
| 8 | 100% | 100% | 100% | 100% | 95.24% |
| 9 | 100% | 90% | 100% | 90.91% | 83.33% |
| 10 | 100% | 100% | 100% | 100% | 100% |
| **average** | **100%** | **90%** | **96%** | **88.67%** | **89%** |
| **W** | - | **0.777** | **0.365** | **0.78** | **0.862** |
| ***p*-value** | - | **0.0077** | **0.00** | **0.008** | **0.081** |

After subjecting the accuracy samples to the one-sided Wilcoxon test, establishing the value of 89.9 as threshold, the *p*-value found was 0.287, which means, over the value of 0.05, which did not enable us to reject the null hypothesis, with a confidence level of 95%. Thus, the test enabled us to infer that **there is no evidence** that the obtained accuracy sample set is distributed around a median over 89.9%, **not rejecting the null hypothesis**. However, after simulations performed from the same test, it was possible to establish a maximum rejection threshold for the null hypothesis equal to **84.9%**, enabling us to infer that the observed accuracy data are distributed over a median equal to 84.9%, with a confidence level of 95%.

The second dependent variable to be analyzed is sensitivity. The null hypothesis assumes that the software is capable of identifying an individual through his/her image with a sensitivity rate equal to 89.9%, while the alternate hypothesis assumes that the sensitivity rate is over this value. As seen before through the Shapiro-Wilk test, the sensitivity sample set found is not distributed normally, which means, the Wilcoxon test (Wilcoxon, 1992) was not used to verify the supposed hypotheses again.

After subjecting the sensitivity samples to the one-sided Wilcoxon test, the *p*-value found was 0.037, which means, under the value of 0.05, which enables us to **reject the null hypothesis** with a confidence level of 95%. Thus, the test enables us to infer that there is evidence that the sensitivity samples found are distributed around a median **over** 89.9%, **not rejecting** the alternate hypothesis, with a confidence interval of 95%.

The third dependent variable to be analyzed is the fβ-measure, with beta equal to 3. The F-beta measure sample set found is distributed normally, which means,

the t-Student test (Mankiewicz, 2000) was used to verify the supposed hypotheses. The operational null hypothesis becomes: H0: μ (metric) = 89.9. After calculating the t-Student statistic, the value T equal to −0.2498 was obtained, below the critical rejection value of the null hypothesis (1833), with a confidence level of 95%. Therefore, the test enabled us to infer that **there is no evidence** that the average of F-beta measures obtained is higher than **89.9**, **not rejecting the null hypothesis**. However, after simulations performed from the same test, it was possible to establish a maximum rejection threshold for the null hypothesis equal to **82.5%**, enabling us to infer that the F-beta measure data have an average superior to 82.5%, with a confidence level of 95%.

Therefore, after conducting the experiment, it was possible to infer that the Face Recognition software is capable of reducing the search space for individuals from a same image, based on the established context and with a confidence level of **95%**, with an **accuracy** rate superior to **84.9%**, defining a tolerance level equal to **0.5%**; **sensitivity** rate superior to **89.9%**, defining a tolerance level equal to **0.55%**: and F-beta measure, with beta equal to 3, superior to **82.5%**, defining a tolerance level equal to **0.5**.

It is also valid to observe the behavior of two other measures: precision and F1-measure. Even though it is not the purpose of this analysis and considering the motivations described in section 1, it was possible to observe a precision average equal to 100% when establishing a lower tolerance level equal to 0.45. This configuration can be applicable in a context in which it is desirable to minimize the occurrence of false positives (cases where police officers can approach people on the street, for example), considering that their occurrence was equal to 0. As in accuracy, the best F1-measure (or fβ-measure with beta equal to 1) was found when a tolerance level equal to 0.5 was established, since this measure is calculated by the harmonic average between precision and sensitivity, without giving more attention to any of these metrics. The accuracy and F-measure averages for

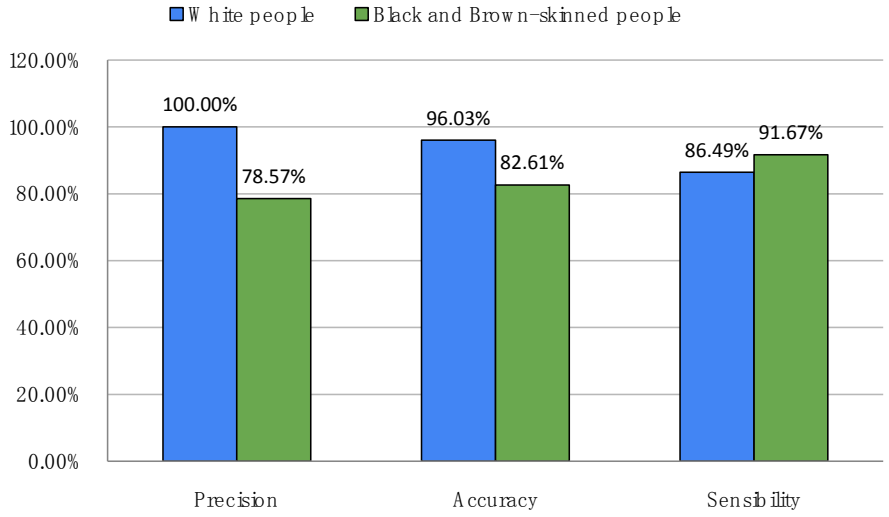


**Figure 4.** Precision, accuracy and sensitivity rates per ethnicity.

a tolerance level equal to 5 indicate that this is the best scenario to use the application in contexts with equal relevance degrees for false positives and false negatives.

Lastly, in order to investigate at initial levels the possibility of disfavoring the Black ethnicity regarding the presence of more False Positives for this ethnicity, precision, accuracy and sensitivity rates were calculated per ethnicity. The tolerance level equal to **0.5** was used for this specific analysis because, according to the previously made observation, the best accuracy rate of the algorithm and the highest balance rate among all the metrics was reached through it. The results are illustrated in Figure 4.

It is possible to notice a large discrepancy in the observed precision—100% for the white people and 78.57% for the Black and brown-skinned people. This difference is justified by the higher number of **False Positives** among the labeled images belonging to the Black people. From all the images that were not trained by the algorithm and consequently, should not be classified, **27.3%**—all of them of Black people—pointed out similarity with some individual in the database. Among the test images regarding white people, this number was **0%**, which means, there were no occurrences of **False Positives** among white individuals, defining a tolerance level equal to **0.5**.

At the same time, it is also possible to observe a higher sensitivity rate from the application in the identification of Black and brown-skinned individuals, a fact that was observed due to a higher incidence of False Negatives for white individuals. **13.5%** out of the individuals previously trained were not identified in the test phase, reducing the sensitivity rate from the application for individuals of such ethnicity. In other words, the observed sensitivity rate was equal to **86.49%** for white individuals, whereas it was equal to **91.67%** for Black and brown-skinned individuals. It confirmed the need for social discussions on this theme, that is, minimally; tolerance levels must be stringent for surveillance applications. The conclusion section summarizes what these results tell us.

## 8.2. Threats to Validity

Regarding the threats to the validity of the mentioned experiment, it is possible to classify them as construction threats and external threats.

Threats to construction validity:

1) The nature of the used images: There was no access to other varied images of the inmates, which means, the ones not belonging to the prison system database. The insertion of artificial images of third parties was necessary for the conduction of the tests. In order to mitigate this threat, the angle and facial expression patterns in the inmate database were kept in the inserted images, which were always front-view images and lacked facial expression, similar to images used on identification cards.

2) Software source code: The results reflect the knowledge model implemented by the algorithm of the Face Recognition application. This threat was mitigate

through the evaluation of several tolerance measures, having the best results for each metric as cutoff point.

Threats to External validity:

The described test images were balanced according to the ethnicity of the individuals, a fact that does not happen in the prison database, which is largely formed by the Black ethnicity. In a real application, depending on the context, data input can be unbalanced toward any ethnicity, or according to crime suspicion. For a test with an unbalanced database, the statistics for investigative photos is necessary if it is possible to confirm the higher number of input photos of culprits in investigative applications, as well as the higher number of input photos of victims in the applications to control access and events.

In view of this context, the test database load balancing in the two dimensions (ethnicity and suspicion) was a restriction and a threat mitigation at the same time, because the possibility of more false positives regarding Black people—due to this natural unbalancing of the training—could be evaluated through a specific analysis of these numbers per ethnicity. The purpose was to verify if the system is biased against an ethnicity, considering a majority of Black or brown-skinned people in the training database. It is worth highlighting that it was not possible to use the investigative database for tests and for the definition of the input percentages per ethnicity of culpability automatically, since not all the input information really contains structured digital data on ethnicity and/or photos with confirmed identification and incarceration time.

## 9. Conclusion

The improvement of the fight against crime goes through the modernization of the organizational processes and enhancement of the tools used by the responsible institutions, since crime investigation demands a volume of information that is frequently hard to access. Therefore, the use of technological tools that assist with the acquisition of information must be common among control and inspection institutions, such as GAECOS and law enforcement institutions, by improving the solution of crimes and identification of suspects. It is hard to understand that there are still institutions indifferent to the advancement of technology, and consequently to the expansion of the methods to fight illicit activities.

Aiming at the collective growth of these organs and the consequent benefits to society, this article described an experimental process to assess the feasibility of using a Facial Recognition tool in real life, since crime also improves with technological advances. The experiment enabled us to analyze three hypothesis sets. In the first one, despite the fact that the accuracy rate reached 90% and even 100% precision—considering the statistical significance—it was not possible to assume a value higher than 89.9% in the identification of inmates in the prison system of the State of Sergipe. However, significance of the accuracy rate of over **84.9%** was verified after simulations, accrediting the software for investigative

applications. New experiments may confirm an accuracy rate of 90% or greater.

According to the investigator's context of search space reduction, the sensitivity rate of the software was analyzed through the second hypothesis set. This time, it was possible to evidence the **non-rejection of the alternative hypothesis,** which was supposed a sensitivity rate **over 89.9%**, with an average sensitivity rate equal to 96%, when defining a tolerance level of **0.55**. This number reveals itself as being quite significant if we consider that the investigator has high tolerance for the occurrence of false positives in the facial detection process.

Subsequently, the two previous measures were analyzed together via the fβ-measure, with beta equal to 3. As in the accuracy analysis, it was not possible to reject the null hypothesis and consider the fβ-measure over 89.9%. However, with a confidence interval of 95%, it was possible to observe that the average of the fβ-measures found is statically significant over **82.5%**.

Simultaneously, by considering only the precision rate, the establishment of a tolerance level equal to 0.45 enabled to observe an average of 100%, ideal for a context in which the occurrence of False Positives must be minimized, as in access control applications. In other words, the establishment of different tolerance levels can capacitate the software to be used in different contexts by different applications.

Regarding future works, a suggestion is the continuation of this article by using images from surveillance cameras, for example. It is also valid to execute the experiments by altering the distance measure among the fiducial points, since these variations were not the object of study in this article.

From an ethnic point of view, there is much discussion about the possibility that the facial recognition solutions are racist, however solid experimental evidence is necessary to substantiate the non-use of facial recognition in some contexts, or even encourage the appearance of technologies that may match the numbers, even with an unbalanced training database. For an Aristotelian reflection and new hypothesis tests: Does the problem lie in the technology on in the type of approach toward the individual? Would polish and polite police approaches, including all the people who are in the environment, solve it?

The premise is: with a margin of error as rigorous as that of medical experiments, precision for one or more ethnicities cannot be less.

For now, to mitigate ethnic discrimination in facial recognition, as this is a structural social issue and much deeper than the simple use of technology, the minimum required should be: 1) The basic training for creating the initial models of the algorithms must be performed with balanced ethnicity data; 2) For street and surveillance applications, confirmation can be made on other photos by a human. However, just like machines, people can also misrecognize and papilloscopy can help the process; 3) Still in the context of surveillance applications, scientific experiments like this one must prioritize precision and approve tools with precision above 99%, for all ethnicities.

On the other hand, technology use decisions are contextual. For example, there

may be an existing human-powered facial recognition process in an organization that makes more mistakes than an AI. In this case, AI would help mitigate or eliminate racism. In other words, great care must be taken that good solutions for society are not invalidated, always under the justification of racism, if this is not on the agenda or if the balanced precision experimental results are transparent and approved by the competent bodies.

In the case of the software analyzed in this article, it has a low implementation cost, being adequate for the reality of many public Brazilian controlling institutions, in addition to not having the need the need to acquire additional biometric equipment. Moreover, the biometric identification through facial recognition can be performed in a passive way, without the voluntary action of the user, which qualifies it to be used in the context of criminal investigation and in the context of other public security applications.

Finally, it is expected that this article will serve as a model for the improvement of the methods to fight crime and for the existing access control systems in several public organs, bearing in mind the purpose to facilitate the identification of individuals who transcend the law. The current experiment also increments the knowledge base of the existing facial recognition techniques, especially in the forensic area, because it was formally described and conducted according to an experimental methodology (Colaço Júnior et al., 2022), which enables transparency and replications that may serve as input for real data meta-analyses.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

Adorno, S., & Pasinato, W. (2008). Crime, violência e impunidade. *Com Ciência, Revista Eletrônica de Jornalismo Científico, No. 98,* 1-4.

Basili, V. R., & Weiss, D. M. (1984). A Methodology for Collecting Valid Software Engineering Data. *IEEE Transactions on Software Engineering, SE-10,* 728-738. https://doi.org/10.1109/TSE.1984.5010301

BRASIL [Constituição (1988)] (2002). *Constituição da República Federativa do Brasil. Organizado por Cláudio Brandão de Oliveira* (320 p.). Roma Victor.

BRAZ (2013). *José Alberto Campos. Investigação criminal.* Leya.

Brunelli, R., & Poggio, T. (1993). Face Recognition: Features versus Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15,* 1042-1052. https://doi.org/10.1109/34.254061

Carvalho, W. A. E. (2020). *Vigilância das forças de segurança através de câmeras de reconhecimento facial e o conflito com o direito à privacidade*. Tese de Doutorado.

Celli, P. L. F. (1999). *Um sistema de apoio à identificação de suspeitos com reconhecimento automático de faces*. Tese de Doutorado, Universidade de São Paulo.

Colaço Júnior, M., Cruz, R. F., & Lima, A. S. (2019). Proposal and Evaluation of a Strategy-Driven Business Intelligence Applications Development. In *International Confe-*

*rence on Information Systems and Technology Management* (pp. 1-29). CONTECSI Publisher.

Colaço Júnior, M., Cruz, R. F., Araújo, L. V., Bliacheriene, A. C., & Nunes, F. L. S. (2022). Evaluation of a Process for the Experimental Development of Data Mining, AI and Data Science Applications Aligned with the Strategic Planning. *JISTEM—Journal of Information Systems and Technology Management, 19,* e202219018. https://doi.org/10.4301/S1807-1775202219018

Conceição, V. S., Viana, C. C., & Rocha, A. M. (2019). Reconhecimento Facial e a Relativização do Direito de Imagem. *Revista INGI-Indicação Geográfica e Inovação, 3,* 436-450.

Da Silva, P. Q., & Santa Rosa, A. N. C. (2004). Reconhecimento facial aplicado a pericia criminal. In *ICCyber'2004* (p. 176). ABEAT Publisher.

Eclipse Foundation (2019). https://www.eclipse.org

Fellous, J.-M., Kruger, N., & Von der Malsburg, C. (1996). *Face Recognition by Elastic Bunch Graph Matching.*

Forato, F. (2020). *Reconhecimento facial ajuda a capturar 42 foragidos no Carnaval de Salvador*. Canal Tech, São Bernardo do Campo. https://www.canaltech.com.br/seguranca/reconhecimento-facial-ajuda-a-capturar-42-foragidos-no-carnaval-de-salvador-161020

Fu, Y., Guo, G. D., Huang, D. W., & Wang, Y. H. (2010). Age Synthesis and Estimation via Faces: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32,* 1955-1976. https://doi.org/10.1109/TPAMI.2010.36

GitHub (2020). https://www.github.com

Gong, S. G., Mckenna, S. J., & Psarrou, A. (2000). *Dynamic Vision: From Images to Face Recognition*. World Scientific. https://doi.org/10.1142/p155

Google Images (2019). https://www.google.com/imghp

Greco Filho, V. (2015). *Manual de processo penal*. Saraiva Educação SA.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-Pie. *Image and Vision Computing, 28,* 807-813. https://doi.org/10.1016/j.imavis.2009.08.002

Harmon, L. D. et al. (1981). Machine Identification of Human Faces. *Pattern Recognition, 13,* 97-110. https://doi.org/10.1016/0031-3203(81)90008-X

Heisele, B. et al. (2003). Face Recognition: Component-Based versus Global Approaches. *Computer Vision and Image Understanding, 91,* 6-21. https://doi.org/10.1016/S1077-3142(03)00073-0

Jafri, R., & Arabnia, H. R. (2009). A Survey of Face Recognition Techniques. *Journal of Information Processing Systems, 5,* 41-68. https://doi.org/10.3745/JIPS.2009.5.2.041

Jebara, T. S. (1995). *3D Pose Estimation and Normalization for Face Recognition*. Centre for Intelligent Machines, McGill University.

Kaufman, G. J., & Breeding, K. J. (1976). The Automatic Recognition of Human Faces from Profile Silhouettes. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-6,* 113-121. https://doi.org/10.1109/TSMC.1976.5409181

Lipoščak, Z., & Loncaric, S. (1999). A Scale-Space Approach to Face Recognition from Profiles. In F. Solina, & A. Leonardis (Eds.), *International Conference on Computer Analysis of Images and Patterns* (pp. 243-250). Springer. https://doi.org/10.1007/3-540-48375-6_30

Lyons, M. J., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding Facial Expressions with Gabor Wavelets. In *Proceedings Third IEEE International Conference on Auto-*

*matic Face and Gesture Recognition* (pp. 200-205). IEEE.

Mahoor, M. H., & Abdel-Mottaleb, M. (2009). Face Recognition Based on 3D Ridge Images Obtained from Range Data. *Pattern Recognition, 42,* 445-451. https://doi.org/10.1016/j.patcog.2008.08.012

Mankiewicz, R. (2000). *The Story of Mathematics*. Cassell.

Maxim, A. G. (2000). On Internal Representations in Face Recognition Systems. *Pattern Recognition, 33,* 1161-1177. https://doi.org/10.1016/S0031-3203(99)00104-1

Nadeau, C., & Bengio, Y. (2000). Inference for the Generalization Error. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems* (pp. 307-313). MIT Press.

Nastar, C., & Mitschke, M. (1998). Real-Time Face Recognition Using Feature Combination. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 312-317). IEEE.

NUCCI (2012). *Guilherme de Souza. Manual de processo penal e execução penal.*

Oliveira, E. P. (2014). *Curso de processo penal. rev., ampl. e atual* (pp. 100-101). Atlas.

Patel, V. M., Gopalan, R., Li, R., & Chellappa, R. (2016). Sensitivity to Facial Cosmetics in Automatic Recognition Systems. *IEEE Transactions on Information Forensics and Security, 11,* 698-711.

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET Database and Evaluation Procedure for Face-Recognition Algorithms. *Image and Vision Computing, 16,* 295-306. https://doi.org/10.1016/S0262-8856(97)00070-X

Postman Inc. (2019). https://www.getpostman.com

Queijo, M. E. (2017). *Direito de não produzir prova contra si mesmo*. Saraiva Educação SA.

Saito, T., & Rehmsmeier, M. (2017). *Basic Evaluation Measures from the Confusion Matrix*. https://classeval.wordpress.com/introduction/basic-evaluation-measures

Salvador Tourism Company (2020). *Carnaval 2020 teve 16,5 milhões de pessoas nas ruas*. Salvador. http://www.saltur.salvador.ba.gov.br/index.php/noticias/344-carnaval-2020-teve-16-5-milhoes-de-pessoas-nas-ruas

Severino, A. J. (2017). *Metodologia do trabalho científico*. Cortez editora.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika, 52,* 591-611. https://doi.org/10.1093/biomet/52.3-4.591

Tourinho Filho, F. C. (2013). *Processo Penal*.

Tzu, S., & Pin, S. (2015). *A arte da guerra*. WWF Martins Fontes.

Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). Newton, MA.

Vasconcellos, F. B. et al. (2008). *A prisão preventiva como mecanismo de controle e legitimação do campo jurídico*.

Wilcoxon, F. (1992). Individual Comparisons by Ranking Methods. In S. Kotz, & N. L. Johnson (Eds.), *Breakthroughs in Statistics* (pp. 196-202). Springer. https://doi.org/10.1007/978-1-4612-4380-9_16

Wiskott, L. et al. (1997). Face Recognition by Elastic Bunch Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19,* 775-779. https://doi.org/10.1109/34.598235

Yang, J., Chen, X. L., & Kunz, W. (2002). A PDA-Based Face Recognition System. In *Sixth IEEE Workshop on Applications of Computer Vision* (pp. 19-23). IEEE.

Yang, S. C., Li, Y., & Hospedales, T. M. (2017). Towards Pose Robust Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39,* 1583-1596.

Zhu, W. et al. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations. In *NESUG Proceedings: Health Care and Life Sciences* (Vol. 19, p. 67). NESUG Publisher.