2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE) | 979-8-3503-2095-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCoSITE57641.2023.10127799

Comparison of Face Recognition Accuracy of ArcFace, Facenet and Facenet512 Models on Deepface Framework

Andrian Firmansyah
School of Industrial and System Engineering
Telkom University
Bandung, Indonesia
andrianfirmansyah@student.telkomuniversity.ac.id

Tien Fabrianti Kusumasari
School of Industrial and System Engineering
Telkom University
Bandung, Indonesia
tienkusumasari@telkomuniversity.ac.id

Ekky Novriza Alam
School of Industrial and System Engineering
Telkom University
Bandung, Indonesia
ekkynovrizalam@telkomuniversity.ac.id

***Abstract*— Face recognition is one of the biometric-based authentication methods known for its reliability. In addition, face recognition is also currently very concerning, especially with the growing use and available technology. Many frameworks can be used for the face recognition process, one of which is DeepFace. DeepFace has many models and detectors that can be used for face recognition with an accuracy above 93%. However, the accuracy obtained needs to be tested, especially when faced with a dataset of Indonesian faces. This research will discuss the accuracy comparison of the Facenet model, Facenet512, from ArcFace, available in the DeepFace framework. From the comparison results, it is obtained that Facenet512 has a high value in accuracy calculation which is 0.974 or 97.4%, compared to Facenet, which has an accuracy of 0.921 or 92.1%, and ArcFace, which has an accuracy of 0.878 or 87.8%. The benefit of this research is to test how high the accuracy of the existing model in DeepFace is if tested with the Indonesian dataset. In this test, Facenet512 is the model that has the highest accuracy when compared to ArcFace and Facenet. This research is expected to help DeepFace users determine the best model to use and provide references to DeepFace developers for future development.**

Keywords— biometric-based authentication, face recognition, confusion matrix, accuracy test

I. INTRODUCTION

Face recognition is one of the most widely accepted biometric-based authentication methods as a reliable biometric parameter [1]. Moreover, face recognition has received tremendous attention since more simplified image analysis, and pattern recognition applications existed. In addition, the demand for commercial applications and the availability of relevant technologies in development are factors supporting the development of face recognition today [2].

There are many implementations and research on face recognition today. One of the studies that have been done in the application of face recognition for monitoring employee attendance using Convolutional Neural Network (CNN) and CCTV cameras as face image input receivers [3]. In addition, other research is about the application of face recognition on faces that use masks in research using Principal Component Analysis and produces accuracy above 65% to recognize faces that use masks [4]. This shows that the development and research of face recognition are getting wider with different conditions.

There are several advantages of applying face recognition in everyday life, such as no need to memorize passwords or PINs during self-verification [5], does not require physical contact when capturing facial data, ease to use, and accessibility when an evaluation process needed for each person recorded and no special skills are needed to carry out the evaluation [2].

Several frameworks can be used to perform the face recognition process. Some examples of open source frameworks that are often used today are Facenet [6], InsightFace [7], face_recogntion [8], DeepFace [9], and CompreFace [10]. The Facenet framework uses Multi-task Cascaded Convolutional Network (MTCNN) and Inception Resnet-V1 in training its model and has an accuracy of 99.65% [6]. Face_recognition uses the Dlib model with an accuracy of 99.38% [7]. CompreFace uses Facenet and InsighFace models and has an accuracy of 99.83% [12][13]. DeepFace provides more varied models to be used in the face recognition process, such as Facenet, Facenet512, VGG-Face, SFace, OpenFace, ArcFace, and DeepFace, and has an accuracy above 93% depending on the model used [11].

The accuracy of the previously mentioned models depends on the type of dataset used. Because based on previous research, it was found that there are differences in accuracy in the face recognition process of African-American faces with Caucasian faces [12]. In addition, previous research states that there are biases and differences in accuracy between Caucasian face datasets with East Asian face datasets [13]. It can be concluded that each of the above models can have differences in accuracy if tested on the faces of other races.

Because there can be differences in accuracy if tested using faces of different races, in this paper, the author will compare the accuracy of the Facenet512, Facenet, and ArcFace models on the DeepFace framework, especially on the face dataset of Indonesian people. Tests will be carried out by testing the accuracy of each image in the dataset and then will be calculated using the classification performance matrix on the accuracy, precision, sensitivity, specificity, and F1 score parameters based on data on the confusion matrix [14].

II. LITERATURE REVIEW

Face recognition is one of the current biometric authentication methods. The concept of face recognition was first introduced in 1964 where in that study the face recognition method was still semi-automatic where there were operators who entered twenty assessment parameters which included the position of the mouth or face and then experienced developments to achieve the current performance [2].

In face recognition, there are three main stages performed, namely face detection, features extraction and face recognition. Fig. 1 illustrates the stages of face recognition from the first stage to the last stage [5]. The first process is face face detection, this process is done to determine the location and size of the face in a digital image. If a facial feature is detected, then other images such as buildings or human bodies will be ignored. After the face is successfully detected and mapped its position, the next step is features extraction. This process represents the face into a feature vector that describes the position of the eyes, nose and mouth of the detected face [5]. In addition, at this stage, checking is carried out between the properties that become the identity of the face and the distinguishing properties [15]. In the last stage, the face recognition process will be carried out. At this stage is the process of comparing the image that has been processed with the image in the available database.

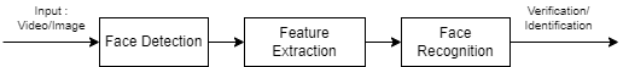

Fig. 1. Stages of Face Recognition [7]

In Fig. 2, it is illustrated that the face recognition stage is divided into two methods, namely verification and identification. These methods can be used depending on the conditions at hand. In the verification method, the process is to compare the captured face with the existing face model registered in the database. This method is one-to-one which will result in a decision whether the identity of the captured face matches the data being compared. This verification is usually used to prevent other people from using the same identity and prevent other people from accessing someone's data. While the identification method is the process of finding a face model in the database that best matches the captured face. This method is one-to-many, that is, the image will and the result will be an individual who is considered to be in accordance with the captured face. This process will be considered a failure if there are no faces that are considered suitable without stating their identity [2].

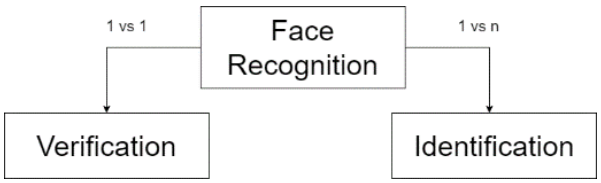

Fig. 2. Face Recognition Mode [2]

DeepFace is a machine learning-based face recognition framework that helps bridge the gap between machine learning and software engineering [16]. In the DeepFace framework, there are models of face recognition, such as Facenet [17]. ArcFace [18]. SFace [19], VGG-Face, OpenFace [20], Dlib, Human-Beings, and DeepID [16]. DeepFace uses the Labelled Face in the Wild (LFW) and Youtube Face (YTF) datasets for training existing models. Table II contains the accuracy of each model in DeepFace. All available models can have an accuracy above 90% in testing with LFW and YTF datasets. Only OpenFace has an accuracy below 95%, at 93.80%. Facenet512 is the model that has the best accuracy of 99.65% using the LFW dataset [9].

TABLE I. MODEL ACCURATION ON DEEPFACE [9]

Model	LFW Score	YTF Score
Facenet512	99.65%	-
Facenet	99.20%	-
SFace	99.60%	-
ArcFace	99.41%	-
Dlib	99.38%	-
VGG-Face	98.78%	97.40%
Human-Beings	97.53%	-
OpenFace	93.80%	-
DeepID	-	97.05%

In machine learning, a confusion matrix is widely used to calculate the success rate of classification based on algorithms or human observations and compared with accurate measurements [21]. Fig. 3 is an overview of the confusion matrix for binary classification. The confusion matrix is divided into two parts: actual class and predicted. An actual class describes the real condition of the data under study. At the same time, the predicted class describes the conditions resulting from the prediction process [22].

In the confusion matrix, the test results are divided into four parts, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [21]. True Positive (TP) is a condition where the actual positive data is predicted as positive. In contrast, if the data is predicted to be negative, it is called False Negative (FN). True Negative (TN) is a condition where the actual data is negative and predicted as negative data. In contrast, if the data is predicted as positive data, the condition is called False Positive (FP) [22][23].

		Predicted Class		
		P	***N***	**Instances**
Actual Class	***P***	**TP** $\lambda_{PP} m_P$	**FN** $(1-\lambda_{PP}) m_P$	m_P
	N	**FP** $(1-\lambda_{NN}) m_N$	**TN** $\lambda_{NN} m_N$	m_N
Estimations		e_P	e_N	m

Fig. 3. Confusion Matrix Structure [21]

Classification performance metrics can be used to see how the test results have been carried out. Table II shows some of the calculations in the Classification Performance Matrix, namely sensitivity, specificity, accuracy, precision, and F1 score [21]. Sensitivity is the True Positive rate that occurs in the testing. Specificity is the True Negative rate that occurs in the testing, accuracy is the True Positive and True Negative rate that occurs compared to the total number

of testing, precision is the positive predictive value that occurs, and F1 score is a weighted comparison of precision and sensitivity [22].

TABLE II. CLASSIFICATION PERFORMANCE MATRIX [21]

Symbol	Matrix	Formula
SNS	Sensitivity	$\frac{\text{True Positive}}{\text{True Positive + False Negative}}$
SPC	Specificity	$\frac{\text{True Negative}}{\text{True Negative + False Positive}}$
ACC	Accuracy	$\frac{\text{True Positive + True Negative}}{\text{Total Data}}$
PRC	Precision	$\frac{\text{True Positive}}{\text{True Positive + False Positive}}$
F1	F1 Score	$2*(\frac{\text{Sensitivity*Precision}}{\text{Sensitivity + Precision}})$

III. METHODOLOGY

This research uses a comparative method by comparing the accuracy of three existing algorithms in DeepFace: Facenet, Facenet512, and ArcFace. Figure 4 explains the researcher's method of comparing the accuracy of existing models in DeepFace. In stage 1, the process is to determine the limitations of the test. The researcher determines the limitations where the models used are Facenet, Facenet512, and ArcFace and use the Multi-task Cascaded Convolutional Network (MTCNN) face detector.

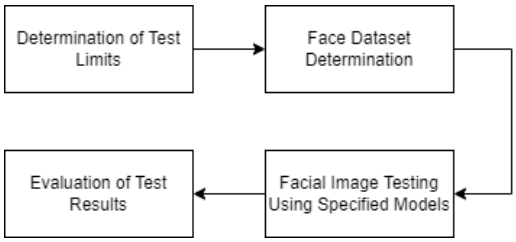

Fig. 4. Research Stages in This Paper

After the limitations and framework used are successfully obtained, the next phase is the selection of the dataset to be used during testing. The dataset used is the face dataset of Indonesian people that have existed in previous research [24]. The photos used in this study consist of 10 people, each with two photos with the provisions of one formal photo and one with a free pose. Each person is distinguished using letter labels from A to J to distinguish between one person and another. In addition, to distinguish between formal photos and photos with free poses, each picture is distinguished by numbers. Number 1 is for formal photos, and number 2 is for photos with free poses. This is chosen because formal photos have more information about a person's face when compared to other photo poses [25]. Fig. 5 is the shape and structure of the dataset used in this research.

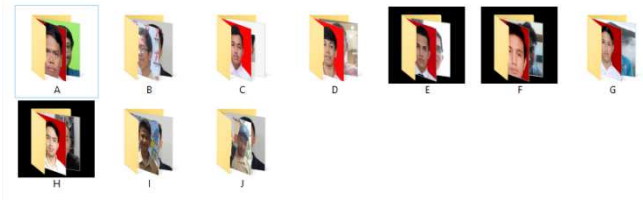

Fig. 5. Datasets Used in Research

In the third stage, researchers will conduct tests to obtain prediction results on an image by comparing it to each existing image. The image will be tested with the three models and will be seen and recorded for each prediction result. Each prediction result will be recorded on the confusion matrix according to the category [21]. Fig. 6 contains the test architecture in this research based on the architecture used in DeepFace [16]. First, both images will be detected whether there is a face image or not, and this process will use the Multi-task Cascaded Convolutional Network (MTCNN) detector. After that, a representation process will be carried out to adjust the image to the model's desired input. After the process is complete, the next step is to perform the verification process. This process is used because the process is carried out to compare one photo with another photo with a one-to-one scheme [2]. Furthermore, in the end, the test results will be obtained.

After all the data regarding the test is obtained, to see how each model obtains the performance, the next step is to evaluate and calculate using the classification performance matrix. In this study, researchers used five calculation components: accuracy, specificity, sensitivity, precision, and F1-Score. The final results of this research are the accuracy, specificity, sensitivity, precision, and F1-Score values, which will be compared to assess which model has good results in testing using the Indonesian face dataset.

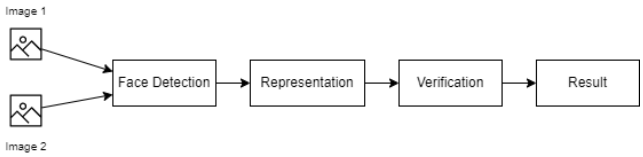

Fig. 6. Photo Testing Flow

IV. RESULT AND DISCUSSION

This research uses Indonesian datasets to compare the accuracy of ArcFace, Facenet, and Facenet512 models in the DeepFace framework. The output of this research is a comparison of the accuracy of each model. In the first stage, researchers determine the limits of testing and research in this paper. This research uses Multi-task Cascaded Convolutional Network (MTCNN) as the face detector used in testing. This detector was chosen because it performs well compared to other detectors in DeepFace [9]. The testing technique compares each image with another image 190 times and is carried out on the three models that will be compared in accuracy. The test was conducted using a Jupyter Notebook and used the DeepFace.verify() function for the testing process. Fig. 7 contains an example of the

output generated from each test. The results obtained from each verification process are verified distance and threshold. The threshold value can change depending on the model and metric used [16]. Since this research uses a cosine metric, the threshold value for each model is 0.68 for ArcFace, 0.4 for Facenet, and 0.3 for Facenet512. The verified value depends on the distance value; if the value is greater than the threshold, the value will be False, and vice versa. If the value is smaller, it will be True.


```
{'verified': True,
 'distance': 0.3651015173853437,
 'threshold': 0.68,
 'model': 'ArcFace',
 'detector_backend': 'retinaface',
 'similarity_metric': 'cosine'}
```


Fig. 7. Example of Testing Output

After the testing process is complete, the next step is to enter and classify the test result data based on its type in the confusion matrix. In table II, it is obtained that from 190 tests using the ArcFace model, 10 prediction results are True Positive (TP), 0 prediction results are False Negative (FN), 23 prediction results are False Positive (FP), and 169 other prediction results are True Negative (TN).

TABLE III. CONFUSION MATRIX OF ARCFACE MODEL TESTING

n=190	**Predicted (+)**	**Predicted (-)**
Actual (+)	10	0
Actual (-)	23	157

In table IV, it is obtained that from 190 trials using the Facenet model, 6 prediction results are worth True Positive (TP), 4 prediction results are worth False Negative (FN), 11 prediction results are worth False Positive (FP), and 169 other prediction results worth True Negative (TN).

TABLE IV. CONFUSION MATRIX OF FACENET MODEL TESTING

n=190	**Predicted (+)**	**Predicted (-)**
Actual (+)	6	4
Actual (-)	11	169

In table V, it is obtained that from 190 tests using the Facenet512 model, 6 prediction results are worth True Positive (TP), 4 prediction results are worth False Negative (FN), 1 prediction result is worth False Positive (FP), and 179 other prediction results worth True Negative (TN).

TABLE V. CONFUSION MATRIX OF FACENET512 MODEL TESTING

n=190	**Predicted (+)**	**Predicted (-)**
Actual (+)	6	4
Actual (-)	1	179

After recapitulating the test results data in the confusion matrix, calculations are performed using the Classification Performance Matrix. Table IV describes the calculations from the research results for each model. The results show that ArcFace has an accuracy value of 0.878 from 190 tests. Besides that, the specificity value is 0.872, sensitivity is 1, precision is 0.303, and F1-score is 0.465. The Facenet model has an accuracy value of 0.921, a specificity value of 0.938, a sensitivity of 0.6, a precision of 0.353, and an F1-score of 0.444. Facenet512 has an accuracy value of 0.974 out of 190 trials. In addition, a specificity value of 0.994 was obtained, a sensitivity of 0.6, a precision of 0.857, and an F1-Score of 0.706. From these results, it can be concluded that the Facenet512 model is superior to the ArcFace and Facenet models in terms of accuracy, specificity, precision, and F1-Score, and ArcFace has a better sensitivity value.

TABLE VI. CLASSIFICATION PERFORMANCE MATRIX

Metrics	**Model**		
	ArcFace	**Facenet**	**Facenet512**
Accuracy	0.878	0.921	0.974
Specificity	0.872	0.938	0.994
Sensitivity	1	0.6	0.6
Precision	0.303	0.353	0.857
F1-Score	0.465	0.444	0.706

Furthermore, to find out the comparison between the accuracy of the research results and the accuracy of the model previously studied by the creators of the DeepFace Framework [9]. The researcher will compare the accuracy of the two results and illustrate this in a bar chart in Fig 8. Based on Fig 8, the accuracy obtained from the results of this study is lower than the accuracy possessed by the model. According to the researchers, this occurred because there were differences in the type of dataset used where DeepFace used the LFW dataset in the training process. In contrast, this study used Indonesian facial datasets. This condition occurred in previous studies where racial differences in the dataset affected the level of accuracy in the face recognition process [12][13].

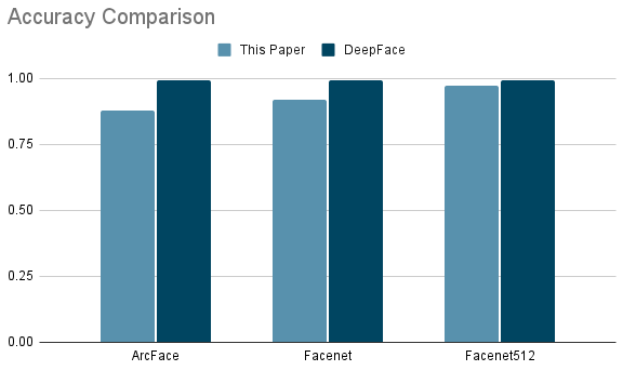

Fig. 8. Model Accuracy Comparison Diagram between Test Results and Accuracy on DeepFace

V. CONCLUSION

This research compares several models in the DeepFace framework using a dataset of Indonesian faces. Tests were conducted 190 times on each model, with the technicality of each image compared to one another. When tested, the Facenet512 model has higher accuracy than the Facenet and ArcFace models, which is 0.974 or 97.4%. However, there is a weakness where the sensitivity level of Facenet512 is only 0.66 or 66% compared to ArcFace, which can predict 100% of the actual faces that are true. From the results of this study, researchers concluded that the ArcFace model contained in the DeepFace framework could be used if the desired objective prefers false positives to false negatives because it has high sensitivity when compared to Facenet and Facenet512. However, if the objective is other than that, the Facenet512 model is ideal because it has better accuracy, specificity, precision, and F1 score values [26].

REFERENCES

[1] P. Kaur, K. Krishan, S. K. Sharma, and T. Kanchan, "Facial-recognition algorithms: A literature review," Medicine, Science and the Law, vol. 60, no. 2, p. 002580241989316, Jan. 2020, doi: 10.1177/0025802419893168

[2] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, Present, and Future of Face Recognition: A Review," Electronics, vol. 9, no. 8, p. 1188, Jul. 2020, doi: 10.3390/electronics9081188.

[3] F. Hamami, I. A. Dahlan, S. W. Prakosa, and K. F. Somantri, "Implementation Face Recognition Attendance Monitoring System for Lab Surveillance with Hash Encryption," Journal of Physics: Conference Series, vol. 1641, p. 012084, Nov. 2020, doi: 10.1088/1742-6596/1641/1/012084.

[4] Ejaz, Md. Sabbir, et al. "Implementation of Principal Component Analysis on Masked and Non-Masked Face Recognition." 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), May 2019, 10.1109/icasert.2019.8934543.

[5] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face Recognition Systems: A Survey," Sensors, vol. 20, no. 2, p. 342, Jan. 2020, doi: 10.3390/s20020342

[6] D. Sandberg, "davidsandberg/facenet," GitHub, May 03, 2021. https://github.com/davidsandberg/facenet (accessed May 03, 2021).

[7] D. DeepInsight, "deepinsight/insightface," GitHub, Jun. 04, 2021. https://github.com/deepinsight/insightface

[8] ageitgey, "ageitgey/face_recognition," GitHub, Jun. 29, 2019. https://github.com/ageitgey/face_recognition

[9] S. I. Serengil, "serengil/deepface," GitHub, Aug. 30, 2020. https://github.com/serengil/deepface

[10] exadel-inc exadel-inc, "Exadel CompreFace is a leading free and open-source face recognition system," GitHub, Jan. 07, 2023. https://github.com/exadel-inc/CompreFace (accessed Jan. 07, 2023)

[11] "CompreFace — Face Recognition Service | Exadel," exadel.com. https://exadel.com/solutions/compreface/

[12] K. Vangara, M. C. King, V. Albiero, K. Bowyer, and Others, 'Characterizing the variability in face recognition accuracy relative to race', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0

[13] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?," IEEE Transactions on Biometrics, Behavior, and Identity Science, pp. 1–1, 2020, doi: 10.1109/TBIOM.2020.3027269.

[14] I. Düntsch and G. Gediga, "Confusion Matrices and Rough Set Data Analysis," Journal of Physics: Conference Series, vol. 1229, p. 012055, May 2019, doi: 10.1088/1742-6596/1229/1/012055.

[15] A. Kumar, A. Kaur, and M. Kumar, "Face detection techniques: a review," Artificial Intelligence Review, vol. 52, no. 2, pp. 927–948, Aug. 2018, doi: 10.1007/s10462-018-9650-2.

[16] S. I. Serengil and A. Ozpinar, 'LightFace: A Hybrid Deep Face Recognition Framework', in 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020, pp. 23–27, doi: 10.1109/ASYU50717.2020.9259802

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, doi: 10.1109/cvpr.2015.7298682.

[18] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, doi: 10.1109/cvpr.2019.00482.

[19] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen, "SFace: Sigmoid-Constrained Hypersphere Loss for Robust Face Recognition," IEEE Transactions on Image Processing, vol. 30, pp. 2587–2598, 2021, doi: 10.1109/tip.2020.3048632.

[20] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar. 2016, doi: 10.1109/wacv.2016.7477553.

[21] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

[22] F. Rahmad, Y. Suryanto, and K. Ramli, "Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification," IOP Conference Series: Materials Science and Engineering, vol. 879, p. 012076, Aug. 2020, doi: 10.1088/1757-899x/879/1/012076.

[23] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," BioData Mining, vol. 14, no. 1, Feb. 2021, doi: 10.1186/s13040-021-00244-z.

[24] T. Maulana Firdaus, T. Fabrianti Kusumasari, S. Suakanto, and O. Nurul Pratiwi, "In-House Facial Data Collection for Face Recognition Accuracy: Dataset Accuracy Analysis," The 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE 2023), Jan. 2023.

[25] D. Cui, G. Zhang, K. Hu, W. Han, and G.-B. Huang, "Face recognition using total loss function on face database with ID photos," Optics & Laser Technology, vol. 110, pp. 227–233, Feb. 2019, doi: 10.1016/j.optlastec.2017.10.016.

[26] Sarang Narkhede, "Understanding Confusion Matrix," Medium, May 09, 2018. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62