Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.10 No.1 January 2025, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Optimizing Face Recognition and Emotion Detection in Student Identification Using FaceNet and YOLOv8 Models

Ghina Fairuz Mumtaz[1], Junta Zeniarja[2], Ardytha Luthfiarta[3], Almas Najiib Imam Muttaqin[4]

[1,2,3,4]*Informatics Department, Universitas Dian Nuswantoro, Indonesia*

[1]ghinafairuz2321@gmail.com(*)

[2,3] [junta, ardytha.luthfiarta]@dsn.dinus.ac.id, [4]almasnajiib27@gmail.com

Received: 2024-12-01; Accepted: 2025-01-13; Published: 2025-01-21

***Abstract*—** The growing need for efficient student identification systems has driven advancements in face recognition and emotion detection technologies. This research presents a single-photo-based system that integrates YOLOv8 for face detection and FaceNet for generating unique facial embeddings, ensuring high-precision student identification and consistent emotion detection under diverse conditions. YOLOv8 localizes faces within images, while FaceNet processes them to generate embeddings for recognition. Emotion detection is performed using these embeddings or an auxiliary emotion classification model. The methodology includes pre-processing images into 64x64 grayscale format, employing image augmentation to enhance model generalization, and evaluating performance using accuracy, precision, recall, and F1 score metrics. The experimental dataset comprises 10 formal student photos, with testing conducted on 100 images. Results demonstrate 94% accuracy in face recognition with augmentation, surpassing 92% without it. Emotion detection achieves 95% accuracy in identifying seven emotions, including angry, happy, sad, neutral, fear, disgust, and surprise, despite variations in expression, lighting, and angle. This system provides a scalable and efficient solution for educational applications such as automated student identification, attendance monitoring, and emotion-based learning management. Its potential spans short-term automation in attendance and mental health monitoring, medium-term improvements in personalized learning and campus security, and long-term AI-driven educational advancements while addressing privacy and social acceptance challenges.

Keywords— Face Recognition; Emotion Detection; YOLOv8; Face Net; Image Augmentation; Student Identification

I. INTRODUCTION

In the growing digital era, face and emotion recognition technology have become very important in various fields, including security systems, health, and education [1][2]. In education, face identification has great potential to simplify the attendance process and monitor the emotional state of students, which can be an important indicator in the learning process [3]. This technology creates a more adaptive learning environment that responds to students [3].

Through the utilization of the Deep-Face Framework, this research endeavors to assist in the development of a learning environment that is more adaptable., which is designed to help overcome resource limitations in academic environments by supporting various superior models such as FaceNet [1][2][3], Arc-Face, VGG-Face, and Open-Face, giving users the flexibility to choose the model that best suits their needs without having to develop algorithms from scratch[4][5]. With the advantages offered by the DeepFace Framework, FaceNet stands out as one of the most prominent models in its application. According to studies [1][2], FaceNet demonstrates high performance in efficiently recognizing and comparing faces, even when faced with variations in lighting, facial expressions, or angles, making it highly suitable for attendance systems based on facial recognition. Furthermore, as highlighted in a study [3], FaceNet detects emotions through facial features in real-time by processing data locally, thereby enhancing privacy and operational efficiency. In addition to its robust face recognition capabilities, the DeepFace Framework includes emotion detection functionality, a key feature for applications requiring real-time emotional analysis. This capability is facilitated through a pre-constructed model and pre-trained weights designed explicitly for recognizing facial expressions. As detailed in the provided resources, users can leverage the emotion detection feature within the Deep-Face library to analyze facial expressions directly, enabling the identification of emotions such as angry, disgust, fear, happy, neutral, sad, and surprised. This integration enhances the adaptability of learning environments. It allows for a more nuanced understanding of user engagement and emotional responses during interactions, ultimately fostering a more responsive and supportive educational framework [4].

Building on these advantages, users can focus on applications and research without dealing with technical complexities, which is especially beneficial for academic environments with limited resources [5]. The framework provides high accuracy in face recognition and face emotion recognition, which is important in education [4][5]. In addition, the use of existing models and simple interfaces reduces development time and costs [4][5].

This research applies YOLOv8 as a face detector. YOLOv8 supports object detection in a single stage with high efficiency [8][11]. YOLOv8's capabilities in terms of speed and accuracy make it highly suitable for applications that require real-time response, which is important in the educational context to enhance student's learning experience by analyzing their behavior and emotions during the learning process [9][11]. With the capability of detecting faces in various sizes and

DOI : https://doi.org/10.25139/inform.v10i1.9304

orientations, YOLOv8 is an ideal choice to be combined with FaceNet in a deep learning-based student identity recognition system, which not only identifies students but also analyzes their emotions during the learning process[6][7][9].

The use of deep learning models for face recognition has shown promising results, with various models developed to improve the accuracy and efficiency of the face recognition process. One widely used model is FaceNet[6][7]. FaceNet is highly effective in face recognition systems, achieving high accuracy through an embedding method that allows comparison of distances between facial features in high dimensions [6]. FaceNet generates a vector representation or embedding of facial images, where identical faces are close to each other in high-dimensional space, while different faces are farther apart [6]. This approach allows the system to recognize faces with high precision, even when there are variations in lighting and viewing angle [6][7]. With a high degree of accuracy and consistency, FaceNet is an ideal choice for applications that require fast and reliable identity verification, such as automated attendance systems in educational institutions [7].

This research integrates FaceNet for face recognition [6][7] and YOLOv8 for face detection [11]. With the approach of using one reference photo and ten validation photos, this research aims to measure the accuracy and speed of the system in recognizing faces in various conditions [11]. Facial expression-based emotion recognition technology is also in the spotlight, especially in education, as it can detect student's emotional responses to customize learning methods [9]. The combination of these technologies is expected to support the engagement, well-being, and effectiveness of automated attendance systems on campus [3].

Previous research [12] utilizing YOLOv8 for object detection and FaceNet for face recognition has shown promising results in developing face recognition systems using a combination of YOLO for face detection and FaceNet for feature extraction, as well as improving the security of electronic systems through biometric identification. Although the research [12] opens up new exploration space in face recognition technology, it faces significant limitations regarding dataset diversity and real-world testing scenarios. These limitations become particularly apparent when examining the research methodology, as the study lacks detailed performance comparisons with relevant current face recognition models, such as Arc-Face, which could provide a more complete perspective on the effectiveness of the developed system [6]. Based on the literature analysis, the use of YOLOv8 and FaceNet in education, especially in student identification and monitoring student's emotional states, is still relatively minimal [8][9]. The implementation of these two technologies in the learning environment has not been widely explored, especially in the aspect of integration for attendance analysis and real-time monitoring of student's emotional state [8]. This point indicates a significant research opportunity in developing face recognition and facial emotion recognition designed for modern educational needs [9].

This research creates superior face recognition and emotion recognition specifically for educational situations to overcome the constraints present in previous studies regarding these areas. This research utilizes comprehensive Image Augmentation techniques to improve FaceNet's ability to recognize various student facial characteristics. This research utilizes Image Augmentation, including horizontal inversion, brightness adjustment, Gaussian noise addition, and random cropping to improve the model's robustness to variations in real conditions in learning environments [9][12]. Image augmentation is an important technique to improve FaceNet's ability to recognize individuals, especially when identifying students with various facial variations. This technique helps FaceNet to overcome differences such as facial expressions, different lighting conditions, and only sometimes good camera quality [10][12]. By using this technique, FaceNet can better detect faces and emotions with high accuracy [6][7][9], thus providing an effective technological solution for recognizing and analyzing students. Through this innovative strategy, research proves that image augmentation is not just an additional technique but a key element in developing a reliable face recognition system that is reliable and responsive to the complexities of educational environments [9][12]. The result is a technological solution providing deep insights into student's identities and emotional states with improved accuracy.

This research aims to develop and apply face recognition and emotion detection technologies, particularly in the educational context, to improve the efficiency and accuracy of student identification. By utilizing FaceNet techniques supported by image augmentation, the proposed system can handle a wide variety of facial characteristics and different environmental conditions [6][7][9]. The main objective of this research is to automate the process of verifying attendance and monitoring student's emotional state. The system is expected to improve school operational efficiency and support student well-being in the short term. In the medium term, the system can reinforce personalized learning, improve campus security, and provide a more comprehensive learning analysis [9]. In the long term, integration with AI Education technology will enable more personalized learning, early detection of student mental health issues, soft skills development, and improved school management.

Despite challenges such as infrastructure, regulatory, and social acceptance issues, the opportunities offered by these technologies are enormous [9][3]. These systems not only have the potential to change the way schools manage and support learning but also bring education closer to student's needs. This research is expected to serve as a foundation for further studies on integrating face and emotion detection models for wider applications, such as attendance management, mental health analysis, and more responsive learning.

II. RESEARCH METHODOLOGY

This research methodology is designed to develop a student face recognition and emotion detection system by combining FaceNet and YOLOv8 models in Fig.1. The method includes a series of interrelated stages to ensure the system can work

DOI : https://doi.org/10.25139/inform.v10i1.9304

accurately and optimally, from data collection to final evaluation [6][13]. The research begins with preparing a dataset containing student face images and applying them for the initial detection process [8][11]. Furthermore, the data obtained through the detection process is reprocessed in the pre-processing stage to improve the model's efficiency [9]. Each step in this research aims to ensure that the system can overcome the challenges of real environments, such as uneven lighting and variations in facial expressions [9].

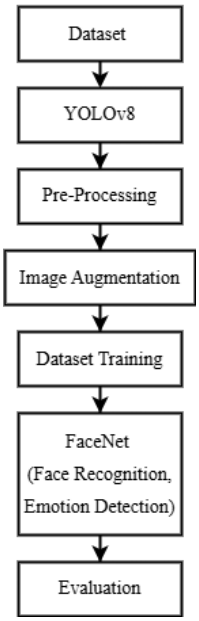

Fig.1. Research Framework

The stages of this methodology are outlined comprehensively in the research flowchart, beginning with dataset collection. The data undergoes pre-processing, including converting images to grayscale format [9]. Data augmentation techniques enhance dataset diversity and ensure robust model performance across various conditions [6][7]. The FaceNet model is employed for face recognition through vector representation, while YOLOv8 is an advanced object detector that efficiently and accurately identifies facial regions [8][13]. This potent combination establishes a foundational framework for subsequent analysis, where the procedure for recognizing facial expressions is initiated by YOLOv8, which detects and extracts the region of interest (ROI) containing the face [8]. This cropped facial region is then passed to FaceNet, which transforms the facial images into high-dimensional vector representations, facilitating the comparison of facial similarities. The identified face undergoes expression analysis via a trained emotion recognition model leveraging the DeepFace framework [4]. Specifically, DeepFace extracts pertinent features from the image, feeding them into an emotion classifier that discerns one of seven potential emotions, including anger, disgust, fear, happiness, sadness, surprise, or neutrality [4][9]. This integrated methodology guarantees swift and precise identification of facial expressions, thereby creating a robust pipeline for effectively recognizing and analyzing emotional states.

Upon completing the training phase with the extended dataset, the system undergoes a comprehensive evaluation using accuracy, precision, recall, and F1-score metrics to thoroughly assess its capability in recognizing faces and detecting student emotions [9]. This evaluation process highlights the flow of a structured and systematic approach, essential for designing a robust and education-ready system [9].

A. Dataset

The dataset used in this study consists of formal photographs of students from Universitas Dian Nuswantoro taken during their first-year enrolment. These photographs depict students in formal poses wearing the university uniform. Each individual in the dataset is represented by a single photo, resulting in 10 photos from 10 students. The formal photographs shown in Fig.2 exhibit variations in lighting and size to reflect conditions encountered in real-world applications, thereby adding challenges for the model to achieve accurate identification. These formal photographs are used as the training dataset, totalling 10 photos.

After training, testing is conducted using testing data to evaluate face recognition and emotion detection. In the testing dataset, each student is represented by 10 additional photos, with the example that person 1 has 10 photos, person 2 has 10 photos, and so on. Thus, the total testing dataset consists of 100 photos with various emotions, incorporating variations in lighting and size to reflect conditions encountered in real-world applications, thereby increasing the challenges for the model to achieve accurate identification.

Fig.2. Academic Student Photos

B. YOLOv8

YOLOv8 is an advanced object detection model designed to detect objects in images or videos with high speed and accuracy [8],[15]. YOLOv8 is an object detection model to recognize student faces in camera images [11]. The YOLOv8 architecture uses the basic principle of a single-shot detector, which allows detection in a single-pass inference process [15]. In this process, YOLOv8 can immediately generate a bounding box, class label, and confidence level for each detected object [11]. One of the

DOI : https://doi.org/10.25139/inform.v10i1.9304

outstanding features of YOLOv8 is its anchor-free approach, which distinguishes it from previous generations of YOLO models [15]. This approach allows YOLOv8 to perform predictions without configuring anchor points, simplifying the detection process, reducing calculation complexity, and improving the model's efficiency and accuracy in various scenarios, including object detection in complex environments [8][15].

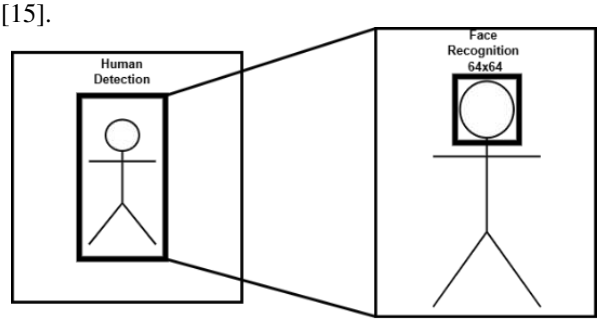

Fig.3. Conceptual Diagram of Human Detection and Face Recognition

The processing stage of YOLOv8 starts from the input stage, where an input image with a consistent size, such as the one in Fig.3, of 64x64 pixels, is fed into the model. This fixed input size is essential to ensure dimensional consistency during the detection process and makes it easier for the model to handle various input data with a consistent scale [17]. With the same input size, YOLOv8 can optimize calculations and reduce computational load in the early processing stages [14][15].

After the input is received, the image data is passed to the backbone part of the model. The YOLOv8 backbone is designed to extract key features from images, and in this architecture, the CSPNet (Cross Stage Partial Network) element is used as the core structure. CSPNet allows YOLOv8 to reduce information redundancy and improve data flow between network layers [16]. This allowed YOLOv8 to extract object features more efficiently without losing important information, thus improving the accuracy of the model in detecting objects [8].

YOLOv8 also utilizes the C2f (Cross Stage Partial with Bottleneck - CSP) layer, which aims to strengthen the feature extraction process in more detail. C2f helps to improve the feature representation of smaller, detailed objects in the image [13]. This layer works by breaking features into stages and processing them incrementally to retrieve richer information, especially for complex images or objects that are difficult to recognize [15].

Furthermore, YOLOv8 uses the SPPF (Spatial Pyramid Pooling-Fast) module to extend the spatial context of the extracted features. SPPF allows the model to simultaneously identify features from different scales and areas in the image, which is crucial for detecting objects of different sizes [13][15]. It adds more details about the location and size of objects in the image, thus improving detection accuracy on different objects in a single image [15].

At the Neck stage, YOLOv8 implements the Path Aggregation Network (PANet) to combine feature information from different scales. PANet facilitates the detection of small objects and overlapping objects in a single image because it can effectively integrate features from large to small scales [16]. With this Neck module, YOLOv8 becomes more sensitive to object details of different sizes and positions in the image, contributing to an overall improvement in detection accuracy [15].

The final part of YOLOv8 is the head, which generates the final prediction through bounding boxes, class labels, and confidence scores for each object. The head combines information from the backbone and neck and performs classification and localization of objects in the image. With the anchor-free configuration, the YOLOv8 head can generate predictions more quickly and accurately without the need to adjust anchor points as in previous YOLO versions [8][13].

In the final stage, the YOLOv8 detect module compiles and displays the detection results in a bounding box with each object's label and confidence level. This complete process flow, as illustrated in Fig.4, shows how YOLOv8 provides structured information about the detected objects in the image through its various architectural components. With its efficiently designed architecture and stages, YOLOv8 can handle object detection tasks with high speed and accuracy, making it ideal for various applications [15][16].

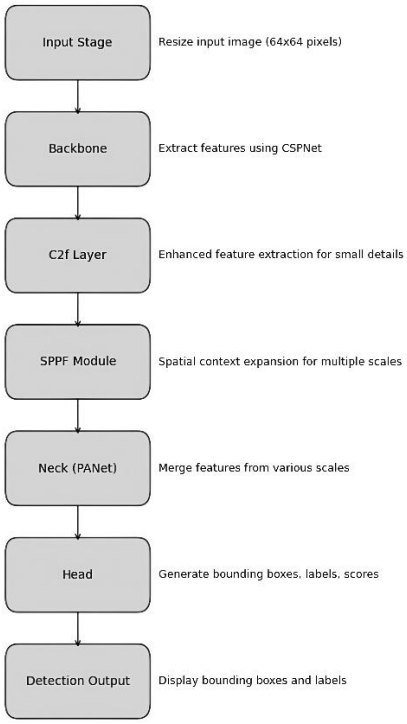

Fig.4. Architecture of YOLOv8 Detection Model

C. Pre-Processing

In the pre-processing stage, after the face area is successfully detected, the image is cropped according to the identified region to ensure that the focus is only on the face, as shown in Fig.5, where a sample image undergoes grayscale conversion. The image is then resized to a standard resolution of 64x64 pixels. This step aims to create consistency in the input size fed to the model during the training process. This size consistency is very important so that the model can recognize patterns more

DOI : https://doi.org/10.25139/inform.v10i1.9304

accurately without being affected by variations in the original size of the image [10][19].

After the image is cropped, the next step is to convert the image to grayscale format. Grayscale processing was chosen because it only requires one colour channel, unlike colour images that require three channels (RGB). This results in more efficient memory usage and processing. In the context of face detection, colour details are usually not very significant. What is more necessary are geometric features, such as the distance between the eyes, nose, and mouth, as these features are more effective in improving face recognition accuracy than colour information susceptible to lighting changes [22][19]. Grayscale processing is highly efficient for image classification tasks, where the focus is on geometric features rather than colour information [18].

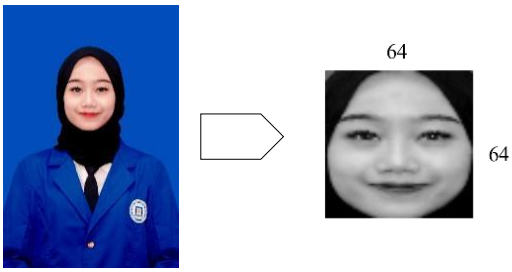

Fig.5. Face Image After Being Grayscale

The next stage involves several steps to prepare the image data to be more varied and consistent during model training. The first step is determining the augmentation performed on each original image. This amount of augmentation can vary, for example, by 20, 30, or even up to 40 new images for each original image. These new images are created with various modifications, such as rotation, brightness adjustment, zoom, or horizontal inversion. The addition of these variations aims to enrich the dataset, enable the model to be more robust in recognizing faces under various lighting conditions, viewpoints, and expressions, and improve its ability to generalize to new data [10][16].

In addition, to ensure consistency in the augmentation process and other processing involving random numbers, an initial value for the random number generator (seed) with a value of 42 was set. By setting the seed value, the random results generated will be the same in each experiment so that experiments can be repeated with consistent and reliable results [17][19].

D. Image Augmentation

In the Image Augmentation stage, this process is performed to increase the variation in the dataset and train the face detection and recognition model more effectively. Augmentation techniques aim to mimic various real-life conditions that may affect the appearance of faces, such as changes in lighting, viewing angle, contrast, and image quality. These are important so the model can recognize faces in various situations [10][22]. Augmenting the dataset has been proven to improve the accuracy of face recognition models by enriching the original dataset through various transformations, such as rotation, brightness change, and noise addition [16][19].

E. Dataset Training

The dataset Training stage is an important step in preparing the data to train the model in college student's face recognition and emotion detection. This stage involves processing the augmented data to produce an adequate training dataset in quantity and quality. This process is designed to improve the model's ability to generalize so that it not only memorizes patterns in the training data but can also recognize faces in various real conditions, such as differences in lighting, viewing angles, and facial expressions.

The augmented image becomes the main component in creating the training dataset at this stage. The augmentation techniques applied include rotation, brightness adjustment, zoom, horizontal inversion, and noise addition. These techniques produce visual variations that aim to replicate various real-world conditions, making the model more resilient to challenges that may occur in real-world applications [16][19]. The total number of augmented images is adjusted according to the number of original images and the amount of augmentation per image. For illustration, if there are ten original images and each image is augmented with ten augmentations, the training dataset includes 100 face images [13].

Each augmented image is structured in a specific folder. The file naming is customized with the individual identity, augmentation number, and relevant labels to make tracking easier during training. In addition, the augmented images are converted to a standard size, such as 64x64 pixels, to ensure uniformity of input dimensions during training. This consistent size not only supports the efficiency of the computational process but also improves the model's accuracy [10][19].

After processing, this training dataset can use the FaceNet model in the facial feature extraction stage. This model generates a numerical representation (embedding) of each face image, which is then used as the main input in model training. This stage ensures that the training data has optimal diversity, quality, and structure to support the model's performance in recognizing faces and accurately detecting emotions in face-based student identification research [19][22].

F. FaceNet (Face Recognition, Emotion Detection)

In this research, FaceNet is a deep learning model used for face recognition and emotion detection of each individual's face in the dataset. The dataset was obtained from formal photos of the Universitas Dian Nuswantoro students in 2021. Each individual in the dataset is represented by one formal photo, which shows a formal appearance with a university uniform. There are 10 photos from 10 different students, as shown in Fig.2.

After the 10 photos of 10 students shown in Fig.2 were processed in the image augmentation and training dataset, testing was conducted using a testing dataset to evaluate face recognition and emotion detection. In the testing dataset, each student is represented by 10 additional photos. For example, person 1 has 10 photos, person 2 has 10 photos, and so on. This

DOI : https://doi.org/10.25139/inform.v10i1.9304

results in a total of 100 photos in the testing dataset. These additional photos include variations in lighting conditions, facial expressions, and image sizes to simulate real-world challenges that may arise in practical applications.

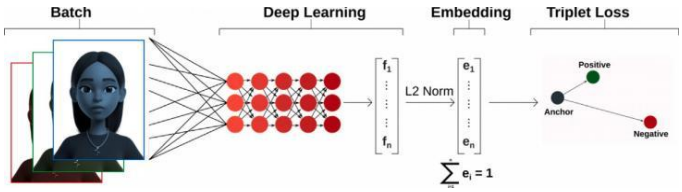

Fig.6. FaceNet Architecture

FaceNet, whose architecture is illustrated in Fig.6, is used as the primary pre-trained model in this research to create face embeddings that serve as a unique numerical representation of each individual. The FaceNet model is loaded using the Deep-Face Framework, which supports various pre-trained models such as VGG-Face, Open-Face, Arc-Face, and others [4][5]. FaceNet was chosen due to its ability to generate a fixed-dimensional embedding (128 dimensions), which contains specific information about the unique features of the face [10][22]. FaceNet is designed using CNN (Convolutional Neural Networks) to extract unique features from faces and generate face embedding that can be used for various identification and classification tasks [19][20].

The embedding extraction process using FaceNet starts from the convolutional layers, which extract the face's main spatial features, such as contours and textures, and the position of important elements, such as eyes, nose, and lips [22]. Input images with a standard size of 64x64 pixels provide uniform dimensions, ensuring the model can work efficiently without losing important details on the face. Once the spatial features are extracted, this data is passed to the fully connected layer, where the features are converted into a fixed-dimensional vector of 128 dimensions. This embedding vector represents the unique characteristics of each face and allows the model to distinguish individuals with a high degree of accuracy[19][22].

Fig.7. Training Facenet With Triplet Loss Function For Face Recognition

During training, FaceNet uses a loss function called Triplet Loss, as shown in Fig.7. This function compares three elements: anchor (main face image), positive (another face image from the same individual), and negative (face image from a different individual).

In Fig.8, Triplet Loss ensures that the embedding of anchor and positive has a small distance, while the embedding of anchor and negative has a large distance. This approach allows FaceNet to produce consistent and reliable embedding despite variations in lighting, viewing angle, or facial expression [22].

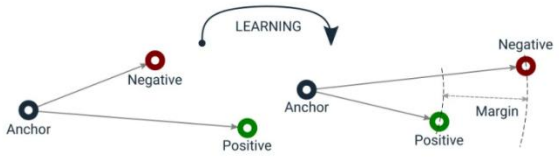

Fig.8. Triplet loss — Learning process

The embedding generated by FaceNet is not only used to distinguish individual identities but is also utilized in facial emotion analysis. By utilizing the embedding, the system can analyze specific facial patterns, such as lip expression, eyebrow position, or patterns around the eyes, which are associated with various emotions [3][22]. FaceNet works with a Convolution Neural Networks (CNN) based architecture, which consists of several important layers [20][22].

These layers extract key features from the face, filter out noise, and represent the face as a fixed-dimensional numerical vector. This approach makes FaceNet excellent at handling various real-world facial conditions, such as lighting changes or dynamic expressions [19][22].

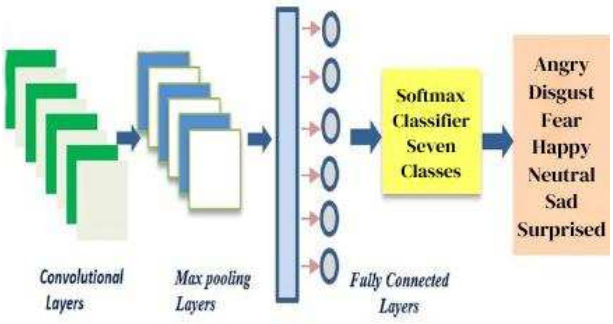

Fig.9. CNN (Convolution Neural Network) Architecture

The CNN workflow process is depicted in Fig.9, where the Convolutional Layers extract spatial features relevant to emotions from the input image. Examples of these layers include smiles, which convey cheerful feelings, and furrowed brows, which convey angry emotions [15][21]. Furthermore, the Pooling layer reduces the data size while retaining important features, thus improving computational efficiency without sacrificing accuracy [15]. The output of the Convolution Layers is passed to the Fully Connected layer, which converts the features into numerical representations that can be used for emotion classification [20][22].

In the final stage, the Output layer uses the Soft-max function to generate probabilities for the six emotion categories listed in Fig.10: angry, disgust fear, happy, sad, surprised, and neutral [15][22].

Fig.10. Sample Image from JAFFE Dataset

DOI : https://doi.org/10.25139/inform.v10i1.9304

G. Evaluation

The evaluation stage in this research aims to measure the system's performance in face recognition and emotion detection using FaceNet and YOLOv8 models. The evaluation dataset used in this study consists of 100 images, which are evenly divided into ten classes, each representing one student. Each student has ten images, each equipped with a ground truth label that indicates the student's true identity. The dataset is organized in a directory format, where each folder represents the student's name as the ground truth label. The arrangement of the dataset in this structured format aims to facilitate the validation and identity-matching process [5].

This research uses various evaluation metrics to measure the performance of the model's Face Recognition and Emotion Recognition. These evaluation methods include Accuracy, Precision, Recall, F1-Score, and a Confusion Matrix [6][9]. In addition, the Confusion Matrix is used to visualize the model's performance by comparing the predicted and actual labels. This matrix consists of four main components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), which are visualized as heat maps for in-depth analysis [6][8].

Accuracy measures the percentage of correct predictions against the total number of test images using Equation (1). This metric provides an overview of how often the model makes correct predictions. In the context of accuracy, the Unknown category affects the calculation because the model assigns this label when the tested face cannot be recognized, either because the embedding distance exceeds a specified threshold, the individual does not exist in the database, or face detection fails. Therefore, the Unknown label is treated as an incorrect prediction in the accuracy calculation [7]. Precision is important to evaluate the accuracy of predictions in each class, especially in situations with data imbalance. It measures the proportion of correct positive predictions to all positive predictions using Equation (2). Recall, as outlined in Equation (3), focuses on the ability of the model to detect all positive cases. Recall is a metric used to measure the model's ability to detect all data that belongs to a particular class. To provide a balance between Precision and Recall, F1-Score is used, which is an integrated average of the two. To provide a balance between Precision and Recall, the F1-Score is used, as shown in Equation (4). This metric ensures that the model focuses on prediction accuracy and effectively recognizes all relevant data.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Images} \quad (1)$$

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Positive\ (FP)} \quad (2)$$

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+FalseNegative\ (FN)} \quad (3)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (4)$$

III. Result and Discussion

This research uses 64x64 pixel images for all feature detection and extraction processes, including emotion detection. This size was chosen to maintain computational efficiency without sacrificing accuracy. The Cosine Similarity metric is used for face recognition to measure the angle between two face representation vectors, as defined in Equation (5). Where X and Y are face representation vectors, this research uses a Threshold of 0.6 to determine face recognition. If the Cosine Similarity value is less than 0.6, the face is considered recognized, while a value greater than 0.6 indicates that the face is not recognized [5][7]. The selection of this threshold aims to minimize detection errors, such as false positives, so that the system can provide more accurate identification [7].

$$Cosine = \frac{x \cdot y}{\|x\|\|y\|} \quad (5)$$

Emotion detection is performed after face location detection by *YOLOv8*, using the generated feature representation [6]. With an image size of 64x64 pixels, this system maintains consistency in feature processing for identity and emotion recognition [6][9].

A. Image Augmentation

Table I presents the various image augmentation techniques applied to increase the variety of the dataset to strengthen the model's ability to recognize faces and detect emotions.

TABLE I
AUGMENTATION TECHNIQUE

Type	Value	Effect
Filter Detail	Applied	Enhance Image sharpness
Multiply And Add To Brightness	Multiply (0.5, 1.5), Add (-30, 30)	Simulates lighting variations by adjusting image brightness
Affine	Scale: (0.8, 1.2)	Changes in object size in the image scale the image.
Log Contrast	Gain: (0.8, 1.2)	Adjust contrast
Multiply	Multiply: (0.9, 1.2)	Adjusts colour intensity
Poisson Noise	Lambda (0, 30)	Adds random noise to simulate noise effects in the image
Sharpen	Alpha: (0.2, 0.8), Lightness: (0.8, 1.2)	Enhances image details with increased sharpness
Flip Horizontal	Probability: 0.5	Flips the image horizontally with a 50% probability
With Brightness Channels	Add: (-50, 50)	Adjusts the brightness in color channels
Gaussian Blur	Sigma: (0, 1.0)	Adds blur effect to the image.

DOI : https://doi.org/10.25139/inform.v10i1.9304

The first technique, Detail Filter, highlights important details such as facial lines and skin texture by adding sharpness to the image. Multiply and Add to Brightness adjusts the brightness level by a certain factor, creating lighting variations that resemble real-life conditions [7]. Furthermore, Affine Transformation changes the size or position of objects in the image, ensuring that the model remains effective despite face shifting or rotation [7]. Other techniques, such as Log Contrast and Multiply, focus on adjusting the contrast and intensity of colours to improve the

clarity of image features, allowing the model to recognize faces better even if the image has varying luminance or saturation. Poisson Noise adds noise that resembles real environmental conditions, helping the model handle low-quality data. Meanwhile, Sharpen accentuates important lines such as nose and lip contours, improving the model's ability to recognize facial features.

Flip Horizontal creates a variety of face orientations by flipping the image horizontally, allowing the model to recognize faces from different viewpoints. The With Brightness Channel technique provides additional adjustments to the brightness level to produce richer lighting variations. Finally, Gaussian Blur is applied to reduce certain details in the image, testing the model's ability to recognize faces even when the image is blurry. This combination of augmentation strategies, as seen in Figure 11, results in a dataset that is more diverse and adaptable to obstacles encountered in the real world. These challenges include differences in image size, noise, and lighting.

Fig.11. Augmentation Results

B. Testing Result

Based on Table II, tests were conducted on the face recognition and face emotion recognition models using a dataset consisting of 100 photos from 10 students, where each student has ten formal photos. This test aims to evaluate the performance of the model with and without the application of augmentation, as well as the effect of the amount of augmentation on accuracy, precision, recall, and F1 score.

TABLE II
TESTING RESULTS

The Number of Augmented Photos generated per formal photo	Application Augmentation	Accuracy	Precision	Recall	F1-Score
1x1	Not Applied	92/95	92/95	100/100	96/97
1x1	Applied	39/95	54/95	58/100	56/97
1x10	Applied	77/95	81/95	94/100	87/97
1x20	Applied	80/95	81/95	99/100	89/97
1x30	Applied	89/95	90/95	99/100	94/97
1x40	Applied	90/95	90/95	100/100	95/97
1x50	Applied	89/95	89/95	100/100	94/97
1x60	Applied	91/95	91/95	100/100	95/97
1x70	Applied	93/95	93/95	100/100	96/97
1x80	Not Applied	92/95	92/95	100/100	96/97
1x80	Applied	94/95	94/95	100/100	97/97

In the condition without augmentation (first row), the model showed high performance in face recognition with 92% accuracy, 92% precision, 100% recall, and 96% F1 score.

As for emotion detection, the model is consistent with 95% accuracy, 95% precision, 100% recall, and 97% F1 score. This indicates that the model works well on formal data with no additional variations.

When augmentation is applied to one image per student (second row), the face recognition model's performance drops significantly, with an accuracy of 39%, precision of 54%, recall at 58%, and F1 score of 56%. However, emotion recognition performance remains stable, with an accuracy of 95%, precision of 95%, recall of 100%, and F1 score of 97%. This decrease occurs because the model faces new variations that require adaptation, while more augmentation data is needed to help the model recognize new patterns.

When the number of augmentations increases to 10 images per student (third row), the face recognition performance increases with 77% accuracy, 81% precision, 94% recall, and 87% F1 score, while emotion recognition remains stable at the same metrics (95% for all metrics except recall remains 100%). Increasing the augmentation provides more sufficient variation, helping the model recognize more complex patterns on faces.

DOI : https://doi.org/10.25139/inform.v10i1.9304

In the ninth row of Table II, where 80 images per formal photo are used without applying augmentation, the face recognition performance achieves an accuracy of 92%, precision of 92%, recall of 100%, and an F1 score of 96%, while emotion recognition consistently performs at 95% accuracy, precision, recall, and an F1 score of 97%.

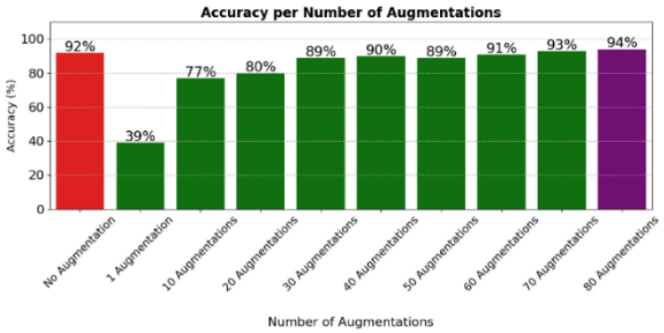

Fig.12. Comparison of Accuracy

Increasing the amount of augmentation to 80 images per student (last row) results in the best performance for face recognition, with an accuracy of 94%, precision of 94%, recall of 100%, and F1 score of 97%, which is even better than without augmentation. As illustrated in Fig.12, the diagram compares face recognition performance with and without augmentation, highlighting how augmentation significantly improves the model's ability to recognize patterns. This indicates that sufficiently large augmentation helped the model recognize patterns better without losing performance on face recognition. Emotion recognition performance remained consistent with 95% accuracy, 95% precision, 100% recall, and 97% F1 score, indicating whether augmentation was applied did not affect emotion detection.

Overall, image augmentation improved the generalization ability of the face recognition model, especially when the amount of augmentation is large enough. However, for emotion recognition, the model's performance is unaffected by augmentation or changes in the number of images, indicating that emotion detection is more stable against dataset variations. This means that augmentation is very important for face recognition but does not affect facial emotion detection.

Fig.13 visualizes the process and results of face and emotion recognition using the FaceNet model applied to student data. The Fig.13 at the top shows the initial stages of the face recognition process, where the original image of the student is converted into a 64x64 pixel grayscale image before going through the feature extraction process using FaceNet. This process produces an embedding vector representing the unique characteristics of the student's face, which is the model.

Then, it is used to predict the identity and emotion of the face. At the bottom, the facial recognition results and emotions are shown for various scenarios, such as happy, fear, neutral, surprise, disgust, sad, and angry expressions. Prediction of the student's identity (Actual: Almas, predicted: Almas) is achieved with high accuracy despite variations in facial expressions and lighting conditions, as evidenced by the model's performance in the previous test results table. This shows that the applied image augmentation helps the model better recognize facial patterns over various data [10].

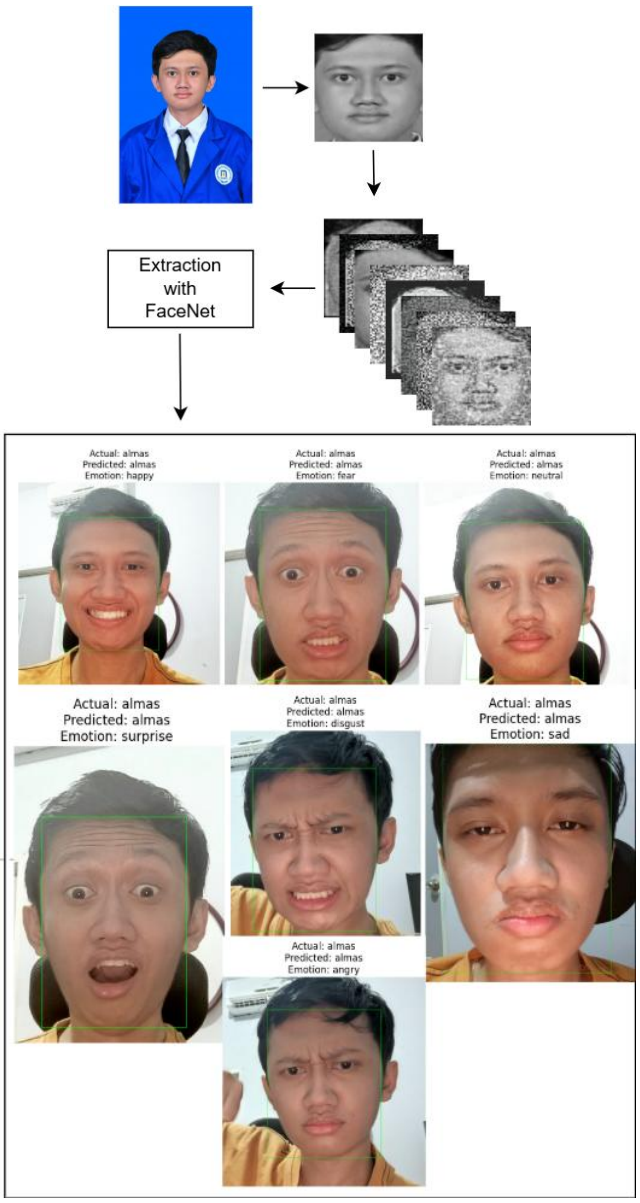

Fig.13. Face and Emotion Recognition Results

Tests on face recognition and face emotion recognition using FaceNet proved to be highly accurate [6][7][22], as reflected in Fig.13. The model was able to identify faces with an accuracy of up to 94% and detect emotions with a stable accuracy of 95%, as seen in Table II. Emotion prediction remained accurate even when there were changes in expression, lighting, or viewing angle in the image.

C. Comparison of FaceNet with Other Models

Table III shows that FaceNet consistently outperforms Arc-Face in face recognition. FaceNet achieves 92% accuracy without augmentation, much higher than Arc-Face's 37%. With

DOI : https://doi.org/10.25139/inform.v10i1.9304

augmentation on 80 student images, FaceNet improved its accuracy to 94%, while Arc-Face remained low at 38%. This result confirms that FaceNet is more adaptive to variable data than Arc-Face.

TABLE III
FACENET COMPARISON WITH ANOTHER MODEL

Number of Photos Augmentation	Model	Application Augmentation	Accuracy (%)
1x1	Arc-Face	Not Applied	37
	FaceNet		92
1x1	Arc-Face	Applied	8
	FaceNet		39
1x80	Arc-Face	Applied	38
	FaceNet		94

Overall, the integration of FaceNet [6][7][13] and YOLOv8[14][16][23] offers an efficient and accurate solution for face and emotion recognition. With further optimizations, such as more effective use of augmentation and computational efficiency, this combination has great potential to be applied in various scenarios, including automatic attendance systems and emotion monitoring in education. The comparison with Arc-Face confirms FaceNet's superiority as a more robust and adaptive model to data variations.

IV. CONCLUSION

Integrating FaceNet and YOLOv8 models provides an efficient and accurate face recognition and emotion detection solution in student identification systems. FaceNet generates embedding-based facial feature representations, achieving a face recognition accuracy of 94%. Table III highlights that FaceNet consistently outperforms Arc-Face in face recognition tasks. FaceNet achieves 92% accuracy without augmentation, significantly higher than Arc-Face's 37%. When augmentation is applied to 80 Photos Augmentation, FaceNet improves its accuracy to 94%, while Arc-Face remains at a much lower 38%. This result confirms FaceNet's superior adaptability to variable data compared to Arc-Face.

Additionally, the emotion detection capability of the system demonstrates stable performance, reaching 95% accuracy under various conditions, including changes in lighting, expression, and viewing angle. The system effectively detects emotions such as angry, fear, neutral, happy, sad, and surprise. YOLOv8, the face detector, supports FaceNet through fast and accurate face detection, making the system suitable for real-time scenarios like automatic attendance or emotion monitoring in educational environments.

This research also identified the importance of image augmentation in improving the model's generalization ability. Tests show that a large amount of augmentation helps FaceNet better recognize patterns, significantly improving its performance compared to scenarios without augmentation. Table II indicates that FaceNet achieves 94% accuracy in face recognition when augmentation is applied to 80 images, compared to 92% without augmentation. Moreover, the experimental results demonstrate that the accuracy remains stable without augmentation at 92%. In contrast, emotion detection showed consistent robustness to data changes, confirming FaceNet's stability for this research. Compared to other models such as Arc-Face, the quantitative analysis presented in Table III confirms that FaceNet proved superior regarding data utilization efficiency and ability to deal with complex pattern variations.

The integration of *FaceNet* and *YOLOv8* for optimizing face recognition and emotion detection has successfully achieved its research objectives, demonstrating high performance across various test scenarios and proving its real-world applicability. While the system shows significant promise, notable challenges were identified, particularly regarding *YOLOv8*'s substantial computational requirements and hardware demands, which could limit deployment in resource-constrained environments. Additionally, *FaceNet*'s reliance on data augmentation techniques highlights areas for potential improvement in model robustness. Despite these limitations, the research findings confirm that the integration of *FaceNet* and *YOLOv8* holds considerable potential for educational systems and similar applications, with future work focused on enhancing computational efficiency and reducing system complexity while maintaining high-performance standards.

The integration of FaceNet and YOLOv8 for optimizing face recognition and emotion detection has successfully achieved its research objectives, demonstrating high performance across various test scenarios and proving its real-world applicability. While the system shows significant promise, notable challenges were identified, particularly regarding YOLOv8's substantial computational requirements and hardware demands, which could limit deployment in resource-constrained environments. Additionally, FaceNet's reliance on data augmentation techniques highlights areas for potential improvement in model robustness. Despite these limitations, the research findings confirm that the integration of FaceNet and YOLOv8 holds considerable potential for educational systems and similar applications, with future work focused on enhancing computational efficiency and reducing system complexity while maintaining high-performance standards.

ACKNOWLEDGMENT

The authors sincerely thank the Department of Informatics Engineering for their invaluable support and guidance throughout the research process. Their unwavering commitment to promoting innovation and academic excellence has been instrumental in completing this project.

The authors also extend their heartfelt appreciation to their families for their unwavering love, encouragement, and understanding during the countless hours dedicated to this research endeavour. Their steadfast support has been a constant source of motivation and inspiration, enabling the authors to overcome the challenges and achieve this milestone.

REFERENCES

[1] Tej Chinimilli, A. Anjali, A. Kotturi, V. Reddy Kaipu, and J. Varma Mandapati, "Face Recognition based Attendance System using Haar

DOI : https://doi.org/10.25139/inform.v10i1.9304

Cascade and Local Binary Pattern Histogram Algorithm," in Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020, Institute of Electrical and Electronics Engineers Inc., Jun. 2020, pp. 701–704. doi: 10.1109/ICOEI48184.2020.9143046.

[2] E. P. Sochima, O. E. Taylor, P. S. Ezekiel, O. E. Taylor, and F. B. Deedam-Okuchaba, "Smart Attendance Monitoring System Using Facial Recognition," International Journal of Computer Techniques, vol. 8, no. 2, 2021.

[3] V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network," IEEE Trans Affect Comput, vol. 13, no. 4, pp. 2132–2143, 2022, doi: 10.1109/TAFFC.2022.3188390.

[4] S. Serengil and A. Özpınar, "A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules," Bilişim Teknolojileri Dergisi, vol. 17, no. 2, pp. 95–107, Apr. 2024, doi: 10.17671/gazibtd.1399077.

[5] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," IEEE Access, Nov. 2020, doi: 10.1109/ASYU50717.2020.9259802.

[6] S. Qi, X. Zuo, W. Feng, and I. G. Naveen, "Face Recognition Model Based On MTCNN And FaceNet," IEEE Access, Feb. 2023, doi: 10.1109/ICMNWC56175.2022.10031806.

[7] F. Cahyono, W. Wirawan, and R. Fuad Rachmadi, "Face recognition system using FaceNet algorithm for employee presence," in 4th International Conference on Vocational Education and Training, ICOVET 2020, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 57–62. doi: 10.1109/ICOVET50258.2020.9229888.

[8] M. Talib, A. H. Y. Al-Noori, and J. Suad, "YOLOv8-CAB: Improved YOLOv8 for Real-time Object Detection," Karbala International Journal of Modern Science, vol. 10, no. 1, pp. 56–68, 2024, doi: 10.33640/2405-609X.3339.

[9] Y. Chen and J. He, "Deep Learning-Based Emotion Detection," Journal of Computer and Communications, vol. 10, no. 02, pp. 57–71, 2022, doi: 10.4236/jcc.2022.102005.

[10] Zhuchkov, "Analyzing the Effectiveness of Image Augmentations for Face Recognition from Limited Data," in 2021 International Conference "Nonlinearity, Information and Robotics", NIR 2021, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/NIR52917.2021.9666135.

[11] K. P. Hasaraddi, G. Thambkar, A. P. Bidargaddi, A. P. Patil, and G. Betageri, "YOLOv8-Enhanced Facial Recognition for One-Shot Learning Attendance System," in 2024 5th International Conference for Emerging Technology, INCET 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/INCET61516.2024.10593178.

[12] P. Prayogo, Hendrawan, E. Mulyana, and W. Hermawan, "A Novel Approach for Face Recognition: YOLO-Based Face Detection and FaceNet," in Proceeding of 2023 9th International Conference on Wireless and Telematics, ICWT 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICWT58823.2023.10335263.

[13] J. Wang and H. Zhao, "Improved YOLOv8 Algorithm for Water Surface Object Detection," Sensors, vol. 24, no. 15, Aug. 2024, doi: 10.3390/s24155059.

[14] F. Feng, Y. Hu, W. Li, and F. Yang, "Improved YOLOv8 algorithms for small object detection in aerial imagery," Journal of King Saud University - Computer and Information Sciences, vol. 36, no. 6, Jul. 2024, doi: 10.1016/j.jksuci.2024.102113.

[15] Yi, B. Liu, B. Zhao, and E. Liu, "Small Object Detection Algorithm Based on Improved YOLOv8 for Remote Sensing," IEEE J Sel Top Appl Earth Obs Remote Sens, vol. 17, pp. 1734–1747, 2024, doi: 10.1109/JSTARS.2023.3339235.

[16] L. Shen, B. Lang, and Z. Song, "DS-YOLOv8-Based Object Detection Method for Remote Sensing Images," IEEE Access, 2023, doi: 10.1109/ACCESS.2023.3330844.

[17] O. Rukundo, "Effects of Image Size on Deep Learning," Electronics (Switzerland), vol. 12, no. 4, Feb. 2023, doi: 10.3390/electronics12040985.

[18] Fein-Ashley, S. Wickramasinghe, B. Zhang, R. Kannan, and V. Prasanna, "A Single Graph Convolution is All You Need: Efficient Grayscale Image Classification," Abu Dhabi, United Arab Emirates: IEEE International Conference on Image Processing (ICIP), Oct. 2024. doi: 10.1109/ICIP51287.2024.10647347.

[19] Rahmad, K. Arai, R. A. Asmara, E. Ekojono, and D. R. H. Putra, "Comparison of Geometric Features and Color Features for Face Recognition," International Journal of Intelligent Engineering and Systems, vol. 14, no. 1, pp. 541–551, Feb. 2021, doi: 10.22266/IJIES2021.0228.50.

[20] P. Lu, B. Song, and L. Xu, "Human face recognition based on convolutional neural network and augmented dataset," Systems Science and Control Engineering, vol. 9, no. S2, pp. 29–37, 2021, doi: 10.1080/21642583.2020.1836526.

[21] O. E. Gundersen, S. Shamsaliei, and R. J. Isdahl, "Do machine learning platforms provide out-of-the-box reproducibility?," Future Generation Computer Systems, vol. 126, pp. 34–47, Jan. 2022, doi: 10.1016/j.future.2021.06.014.

[22] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," Electronics (Switzerland), vol. 10, no. 9, May 2021, doi: 10.3390/electronics10091036.

[23] Paramita, C. Supriyanto, and K. Rahmyanto Putra, "Comparative Analysis of YOLOv5 and YOLOv8 Cigarette Detection in Social Media Content," Scientific Journal of Informatics, vol. 11, no. 2, 2024, doi: 10.15294/sji.v11i2.2808

DOI : https://doi.org/10.25139/inform.v10i1.9304