Smart Access Control System Based On Uncontrolled Environment Human Face Recognition Using Convolutional Neural Network

Muhammad Ikram Andrianur Akbar[1,a)], Anggi Rinaldi[1,b)], Iman Fahruzi[1,c)]

[1]*Department of Mechatronics Engineering, State Polytechnics of Batam, Batam, Indonesia*

[a)]Corresponding author: ikram.andrianur@gmail.com
[b)]anggirinaldi1207@gmail.com
[c)]iman@polibatam.ac.id

Abstract. Neural networks or other artificial intelligence methods have developed rapidly over the past four decades. Its use in various fields makes many people compete to develop it further. One of the applications of artificial intelligence is an automatic door opening and closing system. This development can provide many advantages for users, one of which is that there is no need to make direct contact with the door handle. Armed with a capable PC and an esp32 microcontroller, the system works by detecting images of user facial expressions approaching the object using a webcam. If the required input matches the system rules, the motor will move to open the door. By using convolutional neural network technique, the system can classify the image quickly. Several expressions such as angry, disgusted, scared, happy, sad, surprised, and normal can be the door-opening key of the system. The user can select one to use as the input key to drive the motor to open the door. The study outcomes for several predetermined facial expressions yielded an accuracy rate of 60% and a detection time of under 4 seconds. The detectable distance extends to ± 2 meters. Further study could enable the development of this autonomous door with an IoT-based system for enhanced efficiency. Hopefully, this research can influence the development of intelligent building systems and other fields of artificial intelligence technology.

Keywords: Automatic Door, Convolutional Neural Network, Face Recognition

INTRODUCTION

Automatic door opening systems have become popular in recent years. conventional doors usually consist of locks, handles, and mortise locks to become a whole unit. in general, office doors can become more practical with an automatic system. the door will open when it receives an input signal such as a sensor. for example, if someone wants to enter the room, the door will open by itself [1].

Although there are many sensors that can support this desire, still each thing has its own advantages and disadvantages. For example, proximity sensors [2] will detect objects that pass through its range, but not only detect human movement but also detect passing objects such as trolleys. To solve this problem, we can also replace the sensor with a type of infrared sensor that can capture the heat energy generated by each object. After overcoming the problem of sensors that detect inanimate objects, but other problems still exist. Infrared sensors can still detect moving animals with a distance of less than 1 meter.

Some previous research [3] has successfully developed a campus attendance system with student face recognition technology with an accuracy rate of 95%. However, this system must always be updated every time the student level increases.

In 2019 [4] has developed a facial recognition system using deep learning. This system uses Raspberry Pi as the main system controller. In the system the system uses two authentication methods, namely data training and actual data capture, but the system has a drawback where it takes a long time to capture images and process them into complete data.

Research [5] made a door opening system with Raspberry PI-based face recognition. The system has high accuracy but has the disadvantage that the person who wants to enter the door must have a key pin if the face is not detected in

the database. In addition to having to add other components as a substitute for automatic doors, new face registration is also more complicated and takes a long time. After that, an image testing is carried out first before the system can be used with good accuracy.

The design of a door opening system by detecting facial expressions using the CNN method aims to facilitate and provide comfort in accessing the door. By using facial expression detection technology, the system can identify individuals who are allowed to enter a room or building, and determine whether they have appropriate access based on a pre-set system.

With this background, it was thought to design a door opener system using the Convolutional Neural Network method, where this design functions as automation in the residential and industrial fields, and it is hoped that this new system can have a positive impact on personal and the surrounding environment, as well as provide a new innovation in the field of automation / door opener technology development.

METHODS

A. Software Design

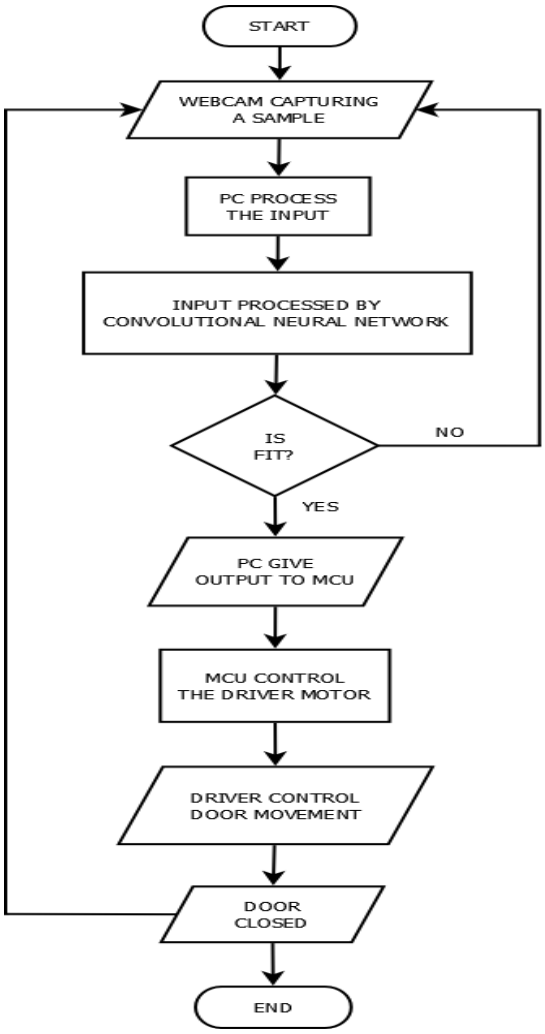

FIGURE 1. Software Design

There are several steps in determining the structure of the device programming. The first step is for the user to look at the webcam that has been provided. The second step is that the PC will process the input in the Convolutional Neural Network (CNN) method program. If the input is fit with the system, then PC giving a signal to ESP32. The

next step is the driver will forward the output signal to the motor. If the facial expression given is in accordance with the required input, the door will move to open and after a delay of 10 seconds the door will close again, then the system returns to the beginning (webcam).

B. Convolutional Neural Network

I. What Is Convolutional Neural Network?

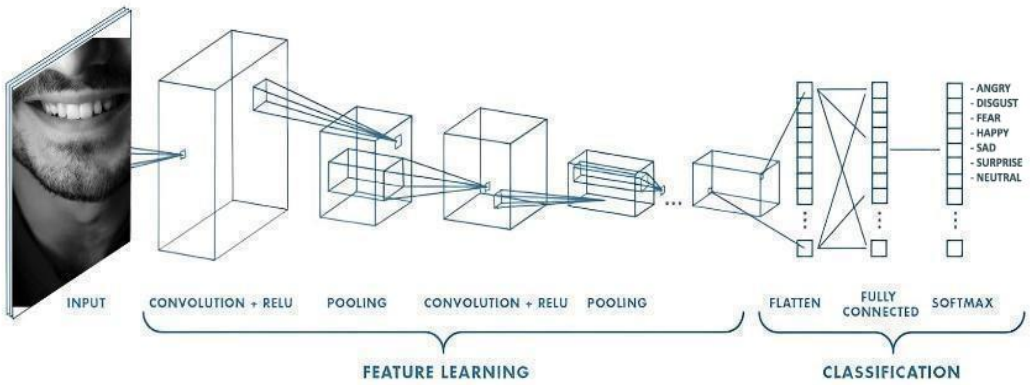

FIGURE 2. Convolutional Neural Network

Convolutional Neural Network is an artificial intelligence neural network that can be used to detect and recognize objects in a digital image and the development of Multilayer Perceptron (MLP). Developed for the first time [6] which was inspired by the way mammals - humans see things. CNN also differs from other types of artificial neural networks in that it retains the special nature of images, which is to recognize patterns in images locally and continue to learn more complex patterns at a higher level. CNNs have become the first choice in various image processing applications, such as face detection object recognition, and image segmentation. Due to their ability to automatically learn important features in images. CNNs are very effective in reducing the need for manual feature extraction, thus speeding up processing and improving pattern recognition accuracy.

Broadly speaking, there are two stages of CNN processing, namely feature learning and classification. The first stage is feature learning. At this stage, there are several layers that are tasked with receiving input images to produce the desired output [7]. In this section, there are several processes that are carried out:

- Input Layer
  This layer functions as a container for pixel valuations from input images, both those that have 2 channels (grayscale) and 3 channels (RGB).
- Convolutional Layer
  This layer is tasked with performing feature extraction on the input image to get the right results for the next process. This layer consists of neurons arranged in such a way as to create pixel filters in the form of length and height [7]. In each layer there are parameters that can be changed, namely filter size, stride, and padding. Stride is a parameter that determines the amount of filter movement. Smaller strides can potentially extract more information from the input, but require more computation compared to larger steps [8]. Padding is an additional layer that can be added to the edges of the image by increasing the pixel size by a certain amount around the input data, ensuring the resulting receptive field is not too small and not too much information is lost [8].
- Rectified Linier Unit (ReLU)
  ReLU (Rectified Linear Unit) activation is the activation layer of a Convolutional Neural Network model that applies the function f(x)= max (0, x), which means it applies a zero value threshold to the pixel values in the input image. This activation causes all pixel values less than zero in the image to become zero [10].

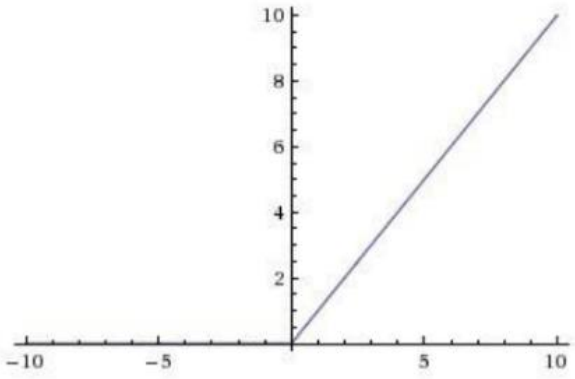

FIGURE 3. ReLU Activation

- Pooling Layer
  It is a layer used to reduce the number of parameters in the feature map and obtain the most important information through down sampling operations. The most commonly used pooling methods are max pooling and average pooling [11].

The second stage is classification. This layer consists of several layers containing neurons that are fully connected to other layers. This layer receives input from the output layer of the feature learning section, is processed by flattening, and adding several hidden layers to the fully connected units in the layer to produce output in the form of classification accuracy for each class [7].

- Flatten
  Flatten has the task of reshaping the feature map into a vector so that it can be used as input for the fully connected layers. In the flatten process, the n×n (3x3 from the figure illustration) pooling layer result matrix will be converted into an n×1 matrix. This matrix will be used as input in the next process, namely classification.

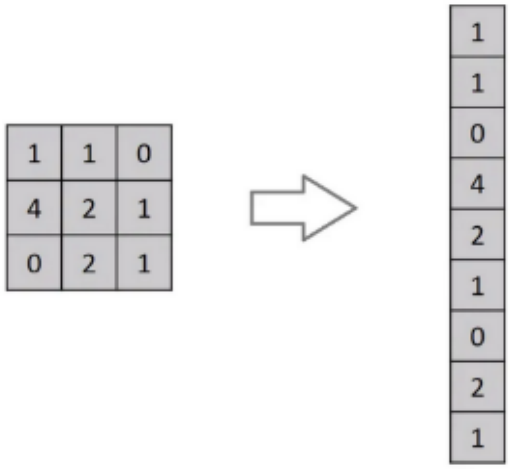

FIGURE 4. Flatten Illustration

- Fully Connected Layer
  A fully connected layer or better known as a dense layer is a layer where each active neuron in the previous layer is connected to the neurons in the next layer, just like an artificial neural network [7].

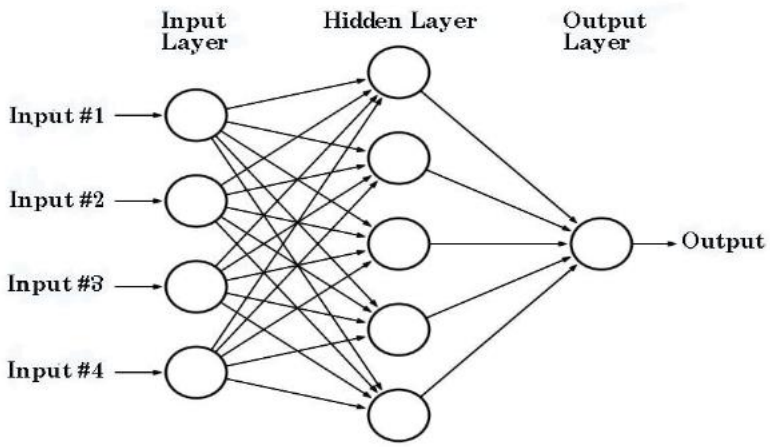

FIGURE 5. Fully Connected Layer [7]

- Softmax
  The SoftMax activation function is used to obtain classification results. The SoftMax function calculates the probability of each target class over all possible target classes and will help to determine the target class for a given input [8].

II. Method Process

It should be noted that in the design of the system this time there are two main processes in the success of CNN, namely the training process and the validation process [9]. Prepare data sets in the form of images in jpg, png, or others extension. Images that have been prepared should be grayscale color to facilitate training later. The prepared facial expression image is inserted into the dataset, then convert it to 48x48 pixels, do data normalization for better results. then preprocessing will be done on the image.

Before conducting the final stage of analyzing system performance, CNN model training is first carried out. This aims as a weight mapping of input and output. The method used in training this model is a basic CNN using a modified Sequential model [9] with the aim of classifying human facial expressions into seven classes. In this model training, the pixels of the image will be resized into 64×64 pixels, 128×128 pixels, 512×512 pixels, and 512x512 pixels.

After both processes are carried out, the results obtained will be divided into seven classes. There is angry, disgust, fear, happy, neutral, sad, and surprise. Each class will be carried out training tests with the aim of training the CNN model to recognize patterns that will be executed next. For more details of the process can be seen in the figure 6 and 7 bellow.

The final stage of system development is performance analysis. This aims to evaluate the system that has been made. The analysis stage can be done through Google Colabs or Jupyter Notebook. In this stage there are five things that can be analyzed, namely accuracy, loss, precision, recall, and f1-score [9] but in this study only measurements will be made from accuracy and loss.

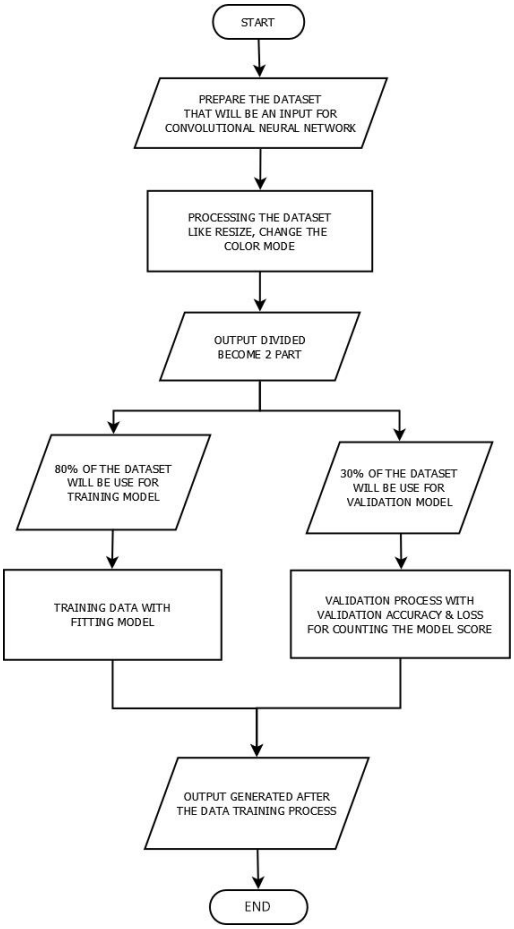

FIGURE 6. Method Process CNN System Diagram

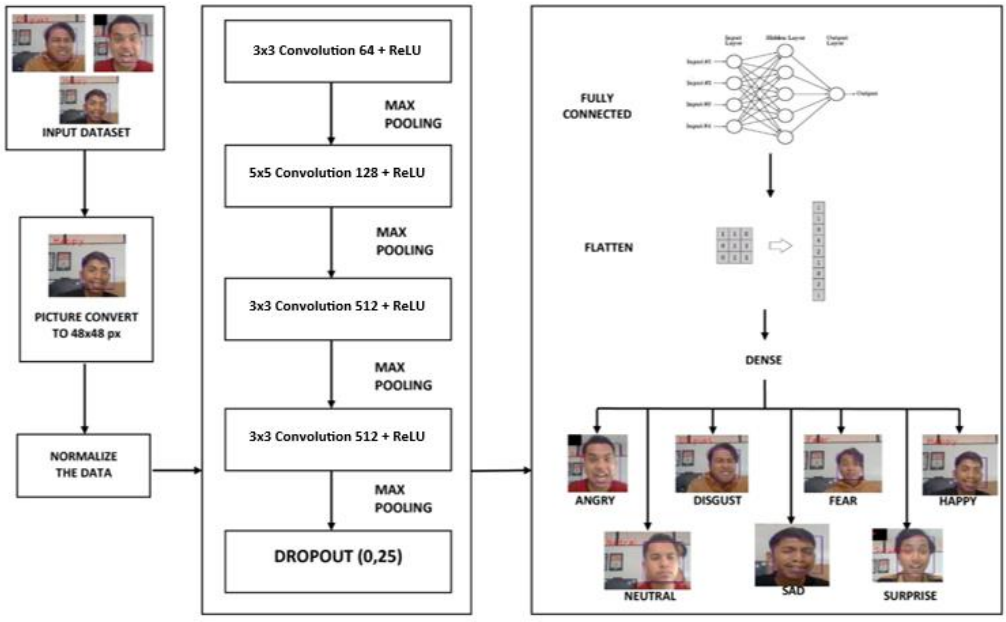

FIGURE 7. CNN Model Training

RESULTS AND DISCUSSION

A. Convolutional Neural Network

The PC specifications used in this training are as follows:

TABLE 1. PC Specification

Component	Specification
Processor	Intel(R) Core(TM) i3-4005U CPU @ 1.70GHz (4 CPUs), ~1.7GHz
RAM	8192MB
System Type	64-bit operating system, x64-based processor
Card Name	Intel(R) HD Graphics Family
Display Memory	2160 MB
Dedicated Memory	112 MB
Software	Jupyter Notebook

It takes adequate specifications to train convolutional neural network models. In this case the applications used are jupyter notebook and google colabs. The processor in processing this training is Intel (R) Core i3-400 with a total RAM of 8 GB. With a 64 bit operating system, the training model is more optimized.

TABLE 2. Dataset

Expression	Training	Validation
Angry	3993	960
Disgust	436	111
Fear	4103	1018
Happy	7164	1825
Neutral	4982	1216
Sad	4938	1139
Surprise	3205	797
Total	**28821**	**7066**

The dataset is divided into two parts, namely training and testing with a total of ±36000 images. In each part there are the same seven classes, but with a different number of images. The comparison in the division is 80 percent of the data for the training process, and the remaining 20 percent is used as material for final validation.

The next process is to train the previously designed model. In this process, three trainings were conducted to see how many epochs are good in training this model. The first training was conducted with 10 epochs, and the second training used 25 epochs. The third training was conducted with 40 epochs.

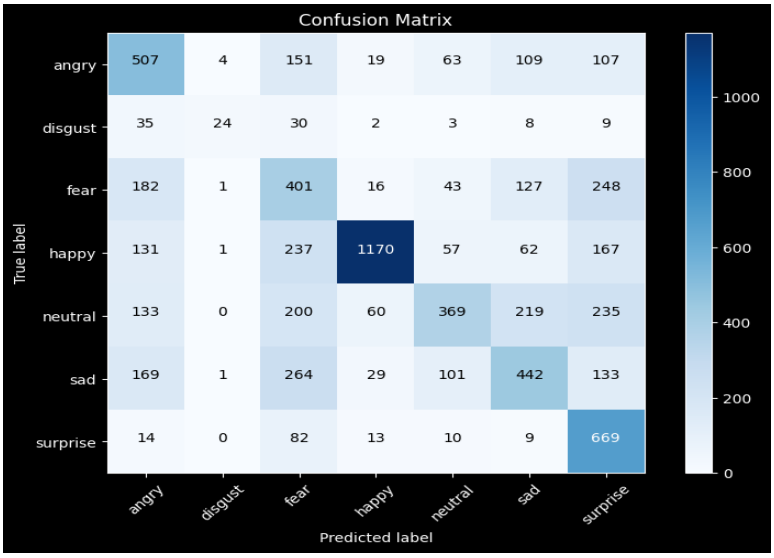

FIGURE 8. 10 Epoch Confusion Matrix

	precision	recall	f1-score	support
angry	0.43	0.53	0.48	960
disgust	0.77	0.22	0.34	111
fear	0.29	0.39	0.34	1018
happy	0.89	0.64	0.75	1825
neutral	0.57	0.30	0.40	1216
sad	0.45	0.39	0.42	1139
surprise	0.43	0.84	0.57	797
accuracy			0.51	7066
macro avg	0.55	0.47	0.47	7066
weighted avg	0.56	0.51	0.51	7066

FIGURE 9. 10 Epoch Confusion Matrix Table

The figure above shows the confusion matrix display of the results of training methods using 10 epochs. Of the total 7066 data used as validation training, there are 507 data “angry” predicted “angry”, 24 data “disgust” predicted “disgust”, 401 data “fear” predicted “fear”, 1170 data “happy” predicted “happy”, 369 data “neutral” predicted “neutral”, 442 data “sad” predicted “sad”, and finally 669 data “surprise” detected “surprise”. The last stage of validation can be seen from the training report results in Figure 9 where the accuracy of epoch 10 is 0.51 or 51% accurate.

FIGURE 10. 25 Epoch Confusion Matrix

	precision	recall	f1-score	support
angry	0.59	0.47	0.52	960
disgust	0.84	0.57	0.68	111
fear	0.54	0.40	0.46	1018
happy	0.79	0.86	0.82	1825
neutral	0.57	0.60	0.58	1216
sad	0.48	0.58	0.52	1139
surprise	0.76	0.77	0.77	797
accuracy			0.64	7066
macro avg	0.65	0.61	0.62	7066
weighted avg	0.64	0.64	0.63	7066

FIGURE 11. 25 Epoch Confusion Matrix Table

The figure above shows the confusion matrix display of the results of the second training method using 25 epochs. Of the total 7066 data used as validation training, there are 453 data "angry" predicted "angry", 63 data "disgust" predicted "disgust", 405 data "fear" predicted "fear", 1568 data "happy" predicted "happy", 732 data "neutral" predicted "neutral", 661 data "sad" predicted "sad", and finally 617 data "surprise" detected "surprise". The last stage of validation can be seen from the training report results in Figure 11 where the accuracy of epoch 25 is 0.64 or 64% accurate.

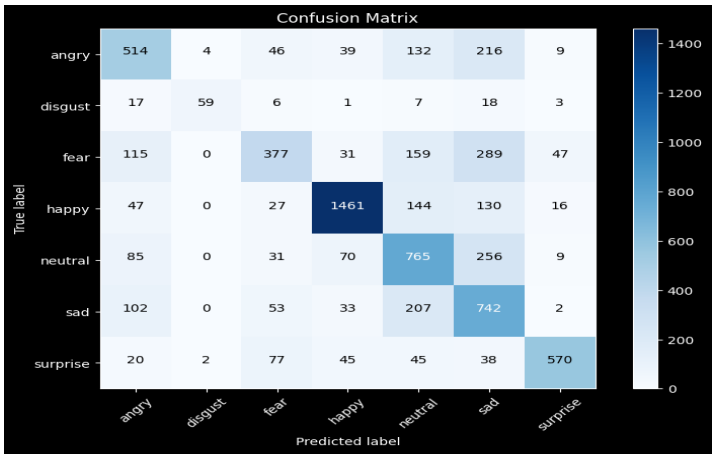

FIGURE 12. 40 Epoch Confusion Matrix

	precision	recall	f1-score	support
angry	0.57	0.54	0.55	960
disgust	0.91	0.53	0.67	111
fear	0.61	0.37	0.46	1018
happy	0.87	0.80	0.83	1825
neutral	0.52	0.63	0.57	1216
sad	0.44	0.65	0.52	1139
surprise	0.87	0.72	0.78	797
accuracy			0.64	7066
macro avg	0.68	0.60	0.63	7066
weighted avg	0.66	0.64	0.64	7066

FIGURE 13. 40 Epoch Confusion Matrix Table

And the last is a confusion matrix display of the results of training methods using 40 epochs. From a total of 7066 data used as validation training, there are 514 data "angry" predicted "angry", 59 data "disgust" predicted "disgust", 377 data "fear" predicted "fear", 1461 data "happy" predicted "happy", 765 data "neutral" predicted "neutral", 742

data “sad” predicted “sad”, and finally 570 data “surprise” detected “surprise”. The last stage of validation can be seen from the training report results in Figure 13 where the accuracy of epoch 40 is 0.64 or 64% accurate.

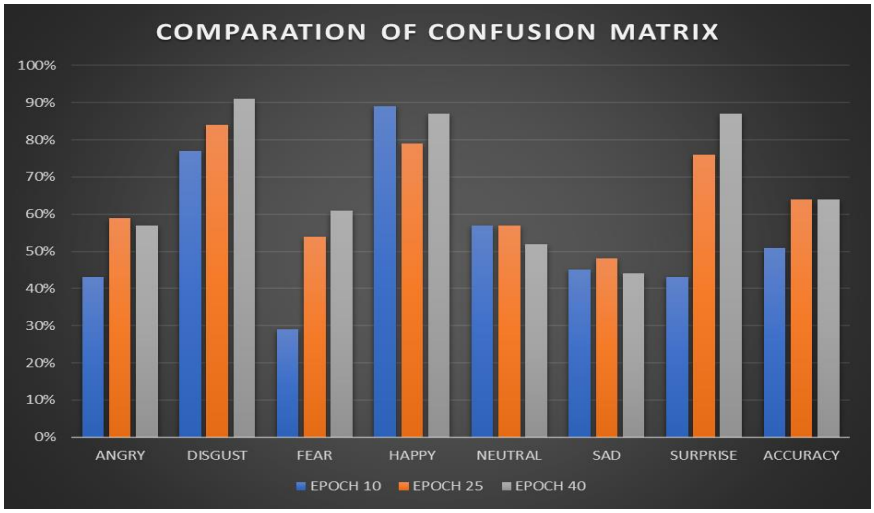

FIGURE 14. Comparation Of Epoch Result

TABLE 3. Comparation Of Epoch Result Table

EXPRESSION	EPOCH 10	EPOCH 25	EPOCH 40
ANGRY	43%	59%	57%
DISGUT	77%	84%	91%
FEAR	29%	54%	61%
HAPPY	89%	79%	87%
NEUTRAL	57%	57%	52%
SAD	45%	48%	44%
SURPRISE	43%	76%	87%
ACCURACY	51%	64%	64%

After conducting 3 training trials, it can be seen that each epoch increase can make the accuracy level increase. Although different for each class trained, epoch 10 has an advantage over the other two trainings. It can be seen that the “happy” class detected “happy” is quite high at around 89 percent, while at epoch 25 it is around 79 percent, and epoch 40 is around 87 percent. This is certainly better if the system expression that we will use later is “happy”. For a clearer comparison, it can be seen from the figure above.

B. Sample Reading With Condition

TABLE 4. Sample Reading With Condition

INPUT	CONDITION	RESULT
Angry	Indoor + Lamp	Detected
Disgust	Indoor + Lamp	Not Detected
Fear	Indoor + Lamp	Detected
Happy	Indoor + Lamp	Detected
Neutral	Indoor + Lamp	Detected
Sad	Indoor + Lamp	Detected
Surprise	Indoor + Lamp	Detected
Angry	Indoor + Lamp	Detected

Disgust	Indoor + No Lamp	Not Detected
Fear	Indoor + No Lamp	Detected
Happy	Indoor + No Lamp	Detected
Neutral	Indoor + No Lamp	Detected
Sad	Indoor + No Lamp	Detected
Surprise	Indoor + No Lamp	Detected
Angry	Outdoor + Noon	Detected
Disgust	Outdoor + Noon	Detected
Fear	Outdoor + Noon	Detected
Happy	Outdoor + Noon	Detected
Neutral	Outdoor + Noon	Detected
Sad	Outdoor + Noon	Detected
Surprise	Outdoor + Noon	Detected
Angry	Outdoor + Night	Detected
Disgust	Outdoor + Night	Not Detected
Fear	Outdoor + Night	Detected
Happy	Outdoor + Night	Detected
Neutral	Outdoor + Night	Detected
Sad	Outdoor + Night	Detected
Surprise	Outdoor + Night	Detected

The first test was conducted by reading the sample using a webcam under four predetermined conditions. Some of the conditions were indoors with the lights on and off, then in an open space during the day and also at night. Each test was conducted 15-30 times.

This test was carried out with 28 target experiments, where each experiment was carried out one to ten repetitions with each sample reading test condition using seven facial expression classifications. In the 28 trials, there were four trials that could not detect samples that had been expressed, namely the expression of disgust. This could be due to the fact that the feature expression of the input is difficult to do directly and also the camera resolution is of standard quality. This statement is based on the fact that the expression of disgust was still detected if a direct test was conducted using the photo samples from the previously available validation. It may also be based on the lack of training data from the disgust class which has a 1:10 data ratio with other classes so that the model recognizes similar expressions rather than the disgust expression itself.

C. Sample Reading With Distance

In the second test, a sample reading experiment was conducted with a distance range of 30 - 250 centimeters and the input given was a happy expression. This is based on the fact that the main input used in the system is only one expression. The distance will be calculated using a measuring tape and starts from where the camera is placed until it touches the user's face. Testing was also carried out during the day and in indoor conditions.

TABLE 5. Sample Reading With Distance

DISTANCE (CM)	INPUT	RESULT
30	Happy	Detected
50	Happy	Detected
80	Happy	Detected
100	Happy	Detected
130	Happy	Detected
150	Happy	Detected
180	Happy	Detected
200	Happy	Detected
230	Happy	Detected
250	Happy	Detected

It can be seen from the table above that at every distance tested it successfully detects the input given. But when the test was carried out with a distance of more than 2.5 meters, the camera had a little difficulty capturing the user's face and could not identify the given expression. User crowd conditions also greatly affect this detection. If there are two or more people in one camera frame, the system will only detect up to two people and the maximum distance is only 1.5 meters.

D. Sample Reading To Output Motor

TABLE 6. Sample Reading To Output Motor

INPUT	OUTPUT
Angry	Off
Disgust	Off
Fear	Off
Happy	On
Neutral	Off
Sad	Off
Surprise	Off

The last test is a motor movement experiment against the received sample. In seven facial expressions that have been put together into a whole model, the system will provide an output signal to the ESP32 if the input received matches the desired output. Happy expression is the main key for the PC to provide output. This process occurs because of the serial communication between the PC and ESP32 that drives all electronic components. When the camera identifies that the user is happy, the system will run as it should, but when the camera detects other expressions, the PC will not provide any data to the ESP32 so that the response to the door does not occur. However, the main input of the system can still be changed through the main program. If the user wants to use expressions such as anger, fear, etc. then the system will be setup and will replace the previous input that has been set.

CONCLUSIONS

After designing, testing, and analyzing this research, it can be concluded that the CNN method can be applied to automatic door opening. The test results of several predetermined facial expressions resulted in an accuracy value of more than 60% and in less than 4 seconds of detection. The distance that can be detected reaches ± 2 meters. With further research, this automatic door can also be developed using an IoT-based system for better efficiency.

REFERENCES

1. Yousif, M., Hewage, C., & Nawaf, L. (2021). IOT technologies during and beyond COVID-19: A comprehensive review. In Future Internet (Vol. 13, Issue 5). MDPI AG. https://doi.org/10.3390/fi13050105
2. Syrlybayev, D., Nauryz, N., Seisekulova, A., Yerzhanov, K., & Ali, M. H. (2022). Smart Door for COVID Restricted Areas. Procedia Computer Science, 201(C), 478–486. https://doi.org/10.1016/j.procs.2022.03.062
3. Budiman, A., Fabian, Yaputera, R. A., Achmad, S., & Kurniawan, A. (2023). Student attendance with face recognition (LBPH or CNN): Systematic literature review. Procedia Computer Science, 216, 31–38. https://doi.org/10.1016/j.procs.2022.12.108
4. Syafeeza, A. R., Mohd Fitri Alif, M. K., Nursyifaa Athirah, Y., Jaafar, A. S., Norihan, A. H., & Saleha, M. S. (2020). IoT based facial recognition door access control home security system using raspberry pi. International Journal of Power Electronics and Drive Systems, 11(1), 417–424. https://doi.org/10.11591/ijpeds.v11.i1.pp417-424
5. Krishna Vamsi, T., & Charan Sai, K. (2019). Face recognition based door unlocking system using Raspberry Pi. In International Journal of Advance Research. www.IJARIIT.com
6. K. Fukushima, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," Biological Cybernetics,1980

7. Azmi, K., Defit, S., & Putra Indonesia YPTK Padang Jl Raya Lubuk Begalung-Padang-Sumatera Barat, U. (2023). Implementasi Convolutional Neural Network (CNN) Untuk Klasifikasi Batik Tanah Liat Sumatera Barat. 16(1), 2023.
8. Paraijun, F., Nur Aziza, R., Kuswardani, D., Teknologi PLN Menara PLN, I., Lkr Luar Barat, J., Kosambi, D., Cengkareng, K., Jakarta Barat, K., & Khusus Ibukota Jakarta, D. (2022). Implementasi Algoritma Convolutional Neural Network Dalam Mengklasifikasi Kesegaran Buah Berdasarkan Citra Buah. 11(1). https://doi.org/10.33322/kilat.v11i1.1458
9. Sar, D. H. A., Sa'idah, S., & Pratiwi, N. K. C. (2022). Klasifikasi Jenis Kulit Wajah Menggunakan Modifikasi Convolutional Neural Network (CNN) Facial Skin Type Classification Using Modified Convolutional Neural Network (CNN).
10. Ilahiyah, S., & Nilogiri, A. (2018). Implementasi Deep Learning Pada Identifikasi Jenis Tumbuhan Berdasarkan Citra Daun Menggunakan Convolutional Neural Network.
11. ANHAR, A., & PUTRA, R. A. (2023). Perancangan dan Implementasi Self-Checkout System pada Toko Ritel menggunakan Convolutional Neural Network (CNN). ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika, 11(2), 466. https://doi.org/10.26760/elkomika.v 11i2.466https://doi.org/10.26760/elkomika.v11i2.466