Deep Convolution Neural Network Based Face Recognition

[1]Oluwole A. S., [2]Peter E. E., [3]Jenyo E. I., [4]Dada I. I.,[5]Fakiya W. I. and [6]Falola, O. E.

[1-6]Department of Electrical/Electronics Engineering, Federal University Oye Ekiti, Nigeria
[1]asoluwole@gmail.com , [2]eno.peter@fuoye.edu.ng

ABSTRACT

Face recognition is a critical and challenging task in the field of computer vision and biometrics, with applications ranging from security and surveillance to human-computer interaction. In recent years, deep learning techniques, particularly deep convolutional neural networks (CNNs), have revolutionized the face recognition landscape by achieving remarkable accuracy and robustness. It is a means of enhancing the quality of Object recognition; it plays a crucial role in various computer vision applications, such as autonomous driving, surveillance systems, and image retrieval. Deep Convolutional Neural Networks (CNNs) have emerged as a powerful approach for achieving state-of-the-art results in object recognition tasks. The core of this paper details the architecture and design of the proposed deep CNN-based face recognition system. It covers key components such as data preprocessing, network architecture selection, training strategies, and optimization techniques. An evaluation of various CNN architectures, including FaceNet, VGGNet, FACENet, ResNet, and EfficientNet, is presented along with their respective strengths and weaknesses in the context of face recognition. The research was aimed at creating an effective system that can accurately recognize student faces and display corresponding profile from the database before examination. FaceNet is the preferred CNN architecture for this work as it has comparatively better accuracy than other CNN architectures. It explores the challenges posed by variations in pose, illumination, and occlusion, and discusses techniques employed to mitigate these challenges within the proposed system. Additionally, the paper included fine-tuning strategies for adapting a pre-trained model to specific face recognition tasks. In creating this hardware, the model was pre-trained with student faces for recognition. 50 individuals, each with 20 images, were imputed as the dataset, giving a total of 1000 images in the dataset. These 1000 images were used to test the various popular CNN models given above. The test resulted in: VGGNet having an accuracy of 92%, precision of 90%, recall of 91%, and inference speed of 0.5: ResNet 50 having an accuracy of 93%, precision of 91%, recall of 92%, and inference speed of 0.3: MobileNet having an accuracy of 88, precision of 84, recall of 86, and inference speed of 0.2: FaceNet having an accuracy of 94%, precision of 92%, recall of 93, % and inference speed of 0.4. The resulted showed that optimized models designed for edge AI microprocessors significantly enhanced inference speed while maintaining similar levels of accuracy, as demonstrated by the FaceNet model.

Keywords: *Face recognition, FaceNet, EfficientNet, VGGNet, ResNet, MobileNet*

1.0 INTRODUCTION

Object recognition is a fundamental task in computer vision that involves the identification and classification of objects within images or video frames. Old methods of object recognition included manually created features and shallow learning algorithms, which often encountered difficulties in handling complex and diverse visual patterns. However, it was the introduction of deeper convolution neural networks (CNNs) that had dramatically altered this field and led to a significant improvement in object recognition accuracy and performance. The structure and function of the human visual cortex have inspired deep CNNs. For them to recover and learn significant features from input images, they constitute a series of layers of connected neurons which are performing complex operations like convolution, pooling or non-linear activation. These features capture local and global patterns, enabling the network to distinguish between different object classes.

The development of deep CNN-based systems for object recognition gained momentum with the introduction of the AlexNet architecture by Krizhevsky *et al*. in 2012. AlexNet demonstrated the potential of deep CNNs

by achieving a significant improvement in object recognition accuracy on the large-scale ImageNet dataset. This breakthrough sparked a wave of research and exploration into developing more sophisticated CNN architectures. Subsequent advancements in CNN architecture design included the VGGNet, GoogLeNet, and ResNet models. VGGNet, proposed by Simonyan and Zisserman in 2014, focused on using deeper networks with smaller filter sizes, leading to improved performance. GoogLeNet, introduced by Szegedy *et al.* in 2014, introduced the concept of inception modules that allowed for more efficient information flow within the network. ResNet, presented by He *et al*. in 2015, introduced residual connections to address the challenge of training very deep networks and achieved significant gains in accuracy.

The success of deep CNNs for object recognition can be attributed to their ability to learn hierarchical representations of visual features. The early layers of the network capture low-level features such as edges, corners, and textures, while deeper layers learn increasingly complex and abstract features. This hierarchical representation enables CNNs to recognize objects based on a combination of local and global visual patterns. Training deep CNN-based object recognition systems requires large, labelled datasets. The availability of large-scale datasets, such as ImageNet, has played a crucial role in training deep networks and facilitating comparative evaluations. Additionally, techniques such as data augmentation, transfer learning, and fine-tuning have been employed to mitigate overfitting and leverage pre-trained models for better performance. Evaluating the performance of deep CNN-based object recognition systems involves assessing metrics such as accuracy, precision, recall, and F1-score. These metrics quantify the system's ability to correctly classify objects and provide a benchmark for comparing different architectures and techniques. Benchmark datasets like ImageNet have been widely used to evaluate and compare the performance of various CNN models.

Despite the remarkable progress in deep CNN-based object recognition, several challenges remain. These include addressing issues of robustness to occlusion, viewpoint changes, and variations in scale and lighting conditions. Exploring interpretability and understanding the learned features of deep CNNs is another active research area. Additionally, deploying CNN models efficiently on resource-constrained devices and optimizing the training process for large-scale datasets are ongoing research topics. The development of deep CNN-based systems for object recognition has significantly advanced the field of computer vision. Through their ability to automatically learn and extract meaningful features from visual data, deep CNNs have revolutionized object recognition tasks. Ongoing research aims to address the remaining challenges and further enhance the capabilities of these systems, ultimately leading to improved object recognition in real-world scenarios.

2.0 REVIEW OF RELATED WORKS

Borra, (2020) developed a system with twenty-five layered CNN. As complexity increases, number of layers. For more complex operations multilayered CNN is required to perform with greater efficiency. There is a quicker retraining process to this face recognition device. The epochs used for training in this system are 20 epochs. The accuracy of the system is 100% for every 4 subjects. The program should be proficient in obtaining new images and saving them into theimage database. Bharadiya, (2023) Image classification in computer vision is important for our education, jobs, and daily life. Images are classified using a procedure that includes image pre-processing, image segmentation, key feature extraction, and matching identification. With the aid of the most modern image classification techniques, we are now able to acquire image data more quickly than ever before and put it to use in several fields, including face recognition, traffic identification, security, and medical equipment. This research's major goal is to comprehend how effectively networks

operate with both static and real-time video streams. Transfer learning on networks using picture datasets is the initial stage in the next process. The next stage is to execute transfer learning on networks with picture datasets.

Wang (2022) the purpose of target recognition is to automatically recognize the position information and the category of a single object in each image. In the target recognition technology, which takes image and video as the processing object, with the rapid development in recent years, convolutional neural network gradually shows its absolute advantage in the field of image target recognition. As one of the important methods of image processing, the basic idea of image recognition is to extract the features of a given image data and use a classification algorithm to recognize its features and predict the label category of the image. Knysh, (2021) Image processing is extremely important in modern science and practice, so it is constantly evolving and improving. Image processing can be used in many industries, namely precision farming (agricultural monitoring), safety systems, quality control, etc. One type of image processing is the recognition of objects in images, which is widely used in the industry, art, medicine, space technology, process management, automation, and many other fields. In autopilots, in collision avoidance systems with other UAVs, for machine vision, analysis of agricultural infrastructure. Recognition of objects in lower resolution images is difficult. To overcome the stated problem, CNN model for identifying objects in lower resolution images is proposed by Raghunathan, (2021) In object recognition datasets, this approach outperforms the high recognition accuracy. Raghunathan, (2021) developed CNN model built using simple 3 x 3 convolution layers. This is done because if a convolutional layer replaces max pooling, then 3 × 3 is the minimal filter size required to construct the model. This deployed model is called the "All Convolutional Neural Network" model. Anirudha, (2020) Concluded that Convolutional Neural Networks (CNN) has become state-of-the-art algorithm for computer vision, natural language processing, and pattern recognition problems. This CNN has been using to build many use cases models from simply digit recognition to complex medical image analysis. Anirudha, (2020) explain each component of a CNN, how it works to image analysis, and other relevant things. Deep convolutional neural networks (DCNNs) can learn to recognize objects as perfectly as human; yet it is unclear whether they can learn semantic relatedness among objects that is not provided in the learning dataset. This is important because it may shed light on how human acquire semantic knowledge on objects without top-down conceptual guidance. To do this Huang, (2021) explored the relation among object categories, indexed by representational similarity, in two typical DCNNs (AlexNet and VGG11). Huang, (2021) found that representations of object categories were organized in a hierarchical fashion, suggesting that the relatedness among objects emerged automatically when learning to recognize them. Object Detection, acknowledgment and distinguishing is the proof of explicit approach with the one of the key research domains for grouped areas including legal applications whereby the suspicious people or articles can be recognized utilizing their live highlights, conduct and attributes. There are numerous sections in the human face which can be prepared and further broke down for the acknowledgment in measurable applications (He, 2020; Wang, 2022). These items are lips, temple, cheeks, jaw, and numerous others which generally make the human grin and pushes forward to the face grin identification.

3.0 DESIGN OF THE FACIAL RECOGNITION SYSTEM USING OPTIMIZED CNN MODEL

The system was designed using different pre-trained existing CNN models (VGG Face, ResNet, MobileNet, FaceNet) for facial recognition systems as shown in table 1. It encompasses the area of edge computing and AI for facial recognition. The husky lens machine vision sensor is trained using multiple picture samples of students registered for examination. The machine vision sensor interfaces with a powerful microprocessor

which processes all the data from the husky lens. 20 samples pictures of each student registered for examination are inputted into the CNN model. The Husky lens machine vision sensor is built upon the Kendryte K210 Edge AI Microprocessor and the Ov2640 camera for image processing. Each picture samples in table 1 were edited before being inputted into the CNN model for memory optimization so as to prevent bugs while running the CNN modes. After developing the CNN Models using the student image samples and processing their predictions using the microprocessor the starting phase of enrollment process is achieved. The systems also comprise of a lightweight database for storage of students' data including names, matric number, session, level and the courses they registered embedded into the flash memory of the microcontroller. After inputting the student verification credentials the ending phase of enrollment is achieved. The enrollment purpose is done via program using the C, C++ and Python programming language. The microprocessor device is also interfaced with the display unit consisting of a 7.0-inch TFT display. After enrollment the system is tested on user faces to make predictions on if the user is registered or not. Whenever the user face is recognized by the CNN model the display unit displays the student's credentials. Whenever the user face is unrecognized by the CNN model the system remains in the home page. The pre-trained existing CNN models VGG Face, MobileNet, ResNet and FaceNet models were used for the facial recognition and the one which produced the best result was adopted for the system. The figures 1 and 2 show the working principle and the design implementation of the system.

Table 1: Some examples of the Student's Image Dataset

S/N	IMAGE TYPE	DATA SIZE (PIXELS)	NUMBER OF DATA (SAMPLE)	TRAINING (%)	TESTING (%)
1.		480 x 480	20	80	20
2.		480 x 480	20	80	80
3.		480 x 480	20	80	20

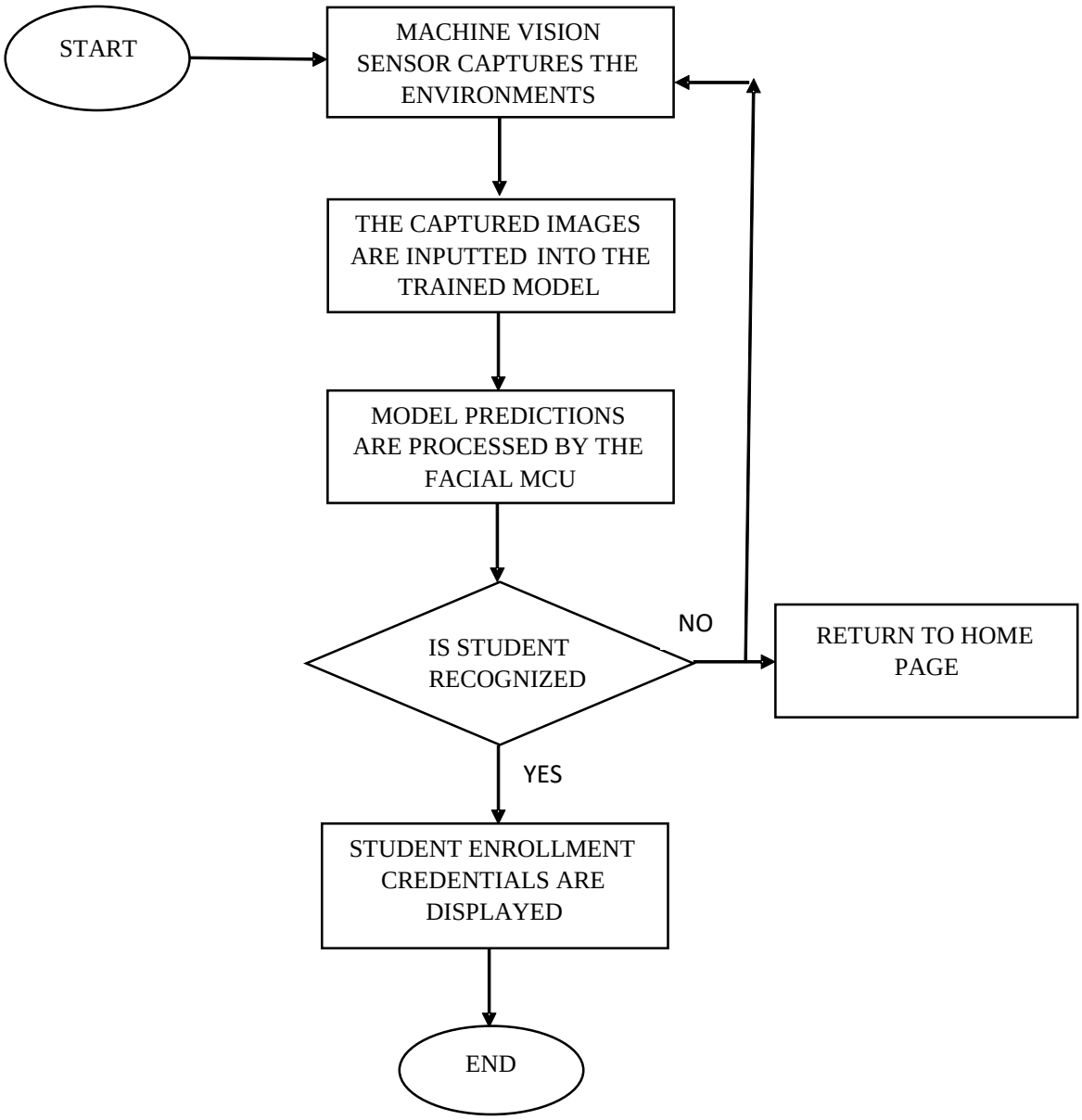

Figure 1: Systems Flowchart

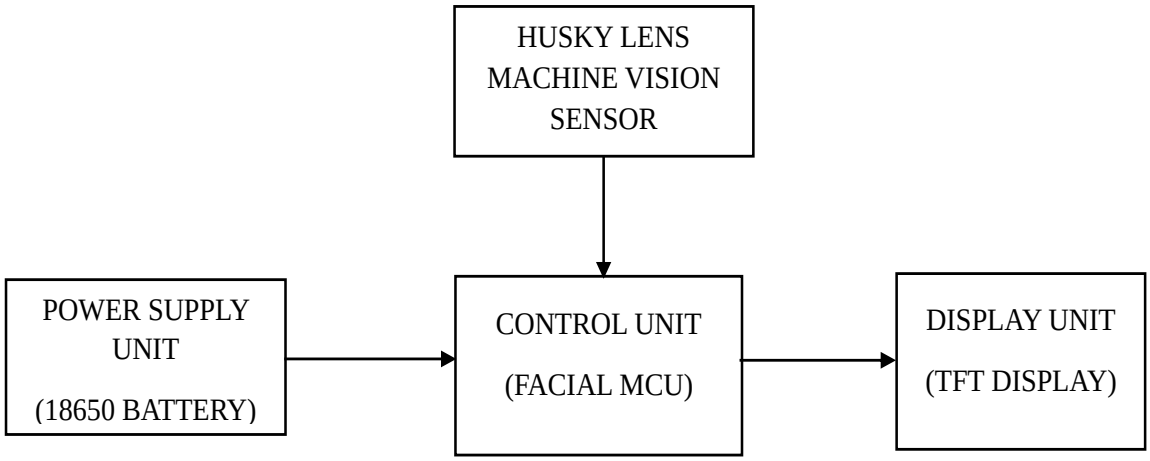

Figure 2: Systems Block Diagram

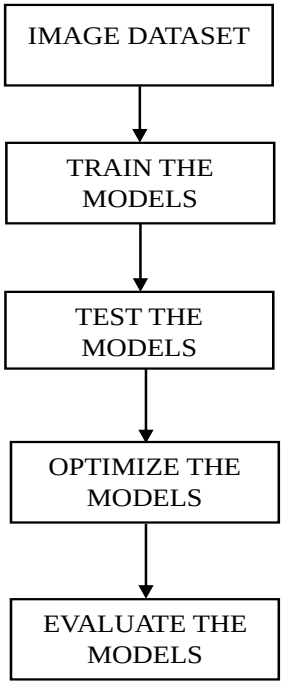

Figure 3: CNN Model Implementation

3.1 FACIAL RECOGNITION CNN MODELS

3.1.1 VGG Face

VGG Face is a deep convolutional neural network architecture specifically designed for face recognition tasks. It is based on the VGGNet architecture, which was originally proposed for image classification. VGG Face, however, has been adapted and fine-tuned for the purpose of extracting facial features and enabling accurate facial recognition.

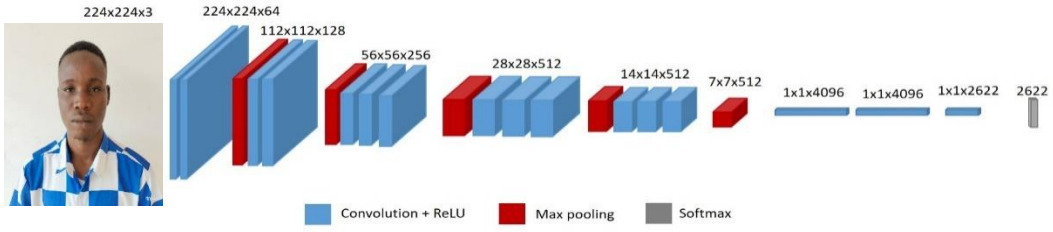

Figure 3: VGG Face for Facial Recognition Model Layers.

3.1.2 ResNet50

ResNet-50, short for Residual Network with 50 layers, is a deep convolutional neural network architecture that has been widely used for various computer vision tasks, including image classification and feature extraction. It is part of the ResNet family, which was introduced to address the challenge of training very deep neural networks by using residual connections.Facial recognition is a biometric technology that involves identifying and verifying individuals based on their facial features. ResNet-50 can be used as a key component in building a facial recognition system due to its ability to extract high-level features from images.

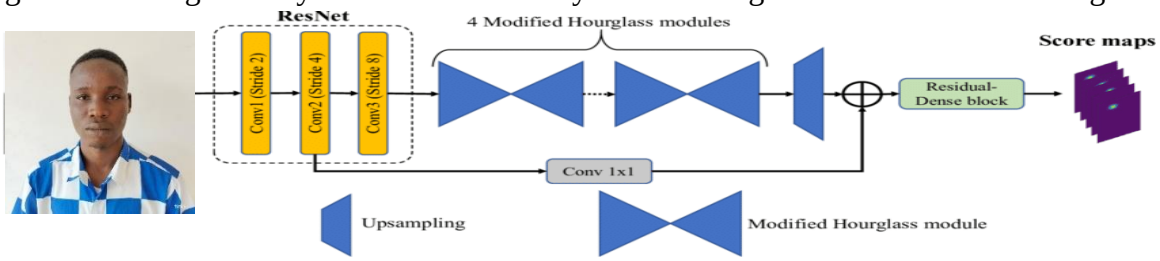

Figure 4: ResNet50 for Facial Recognition.

3.1.3 MobileNet

MobileNet is a family of lightweight neural network architectures designed for efficient deployment on resource-constrained devices, such as mobile phones and embedded systems. These architectures are known for their compact size, low computational requirements, and good trade-off between accuracy and model size. MobileNet can also be used as a component in a facial recognition system, particularly when efficiency and real-time processing are important.

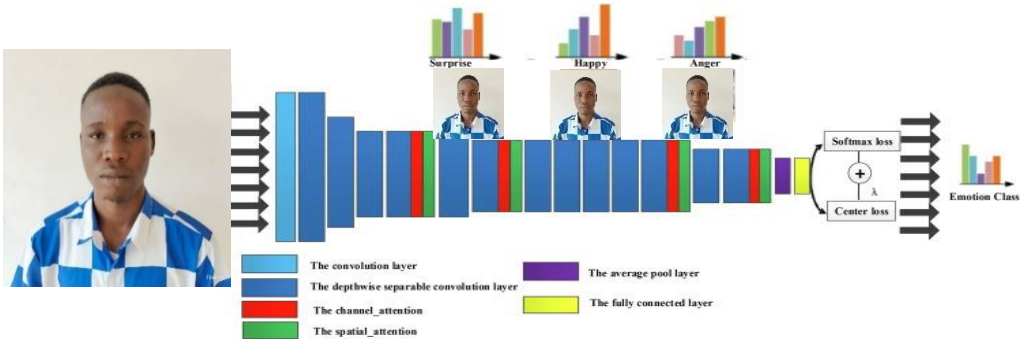

Figure 5:MobileNet for Facial Recognition

3.1.4 FaceNet

FaceNet is a deep learning model developed for facial recognition that aims to directly learn a mapping from face images to a high-dimensional feature space, where the distances between the embeddings (feature vectors) of different faces reflect their similarity. This approach was introduced to address the limitations of traditional methods that relied on handcrafted features and pairwise comparisons. FaceNet revolutionized facial recognition by enabling the learning of robust and discriminative feature representations directly from

data. Its concepts have inspired subsequent research in the field, contributing to advancements in face recognition accuracy and real-world applications.

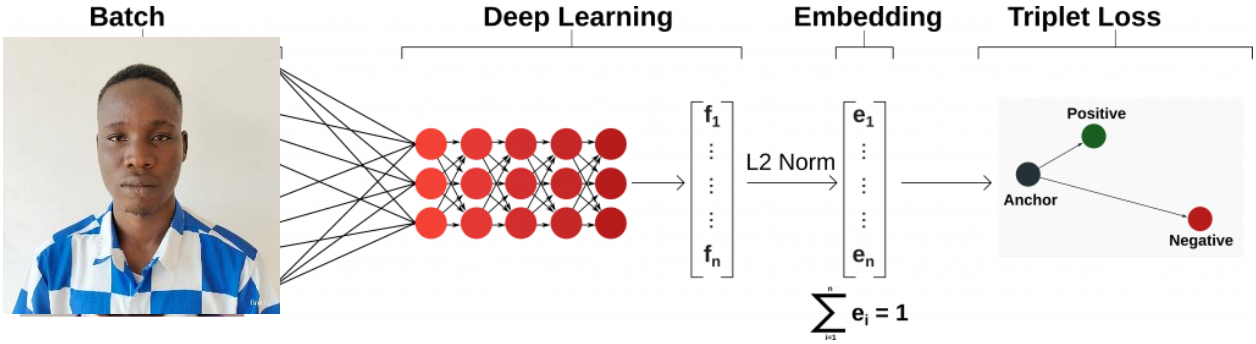

Figure 6: FaceNet for Facial Recognition

4.0 ELECTRICAL DESIGN IMPLEMENTATION

The systems electrical circuit was designed, simulated, and tested using proteus software. The electrical design started with the power unit design which is built using 12pcs of 18650 lithium-ion batteries connected in series. The battery banks are connected to the power control board which controls the charging, battery level control and indication. The power control board supplies an output of 5v 2.4A on 2 different channels used to power the various other units. The husky lens machine vision sensor, 7.0-inch TFT display and the edge AI facial MCU board. All components are place on the PCB and joined via soldering and the jumper wires were used for circuit connections implementation.

4.1 Power Supply Unit Calculation

Battery Specifications = 2500mAh x 12 = 30,000mah = 30Ah at 5.0V

$$Q_{battery} = \frac{E_{battery}}{V_{battery}} \quad (P_{device} = 10\text{W}) \tag{1}$$

$$T = Q \times \frac{V}{P} = 30 \times \frac{5}{10} = 15 hours$$

Device operating time = 15 hours

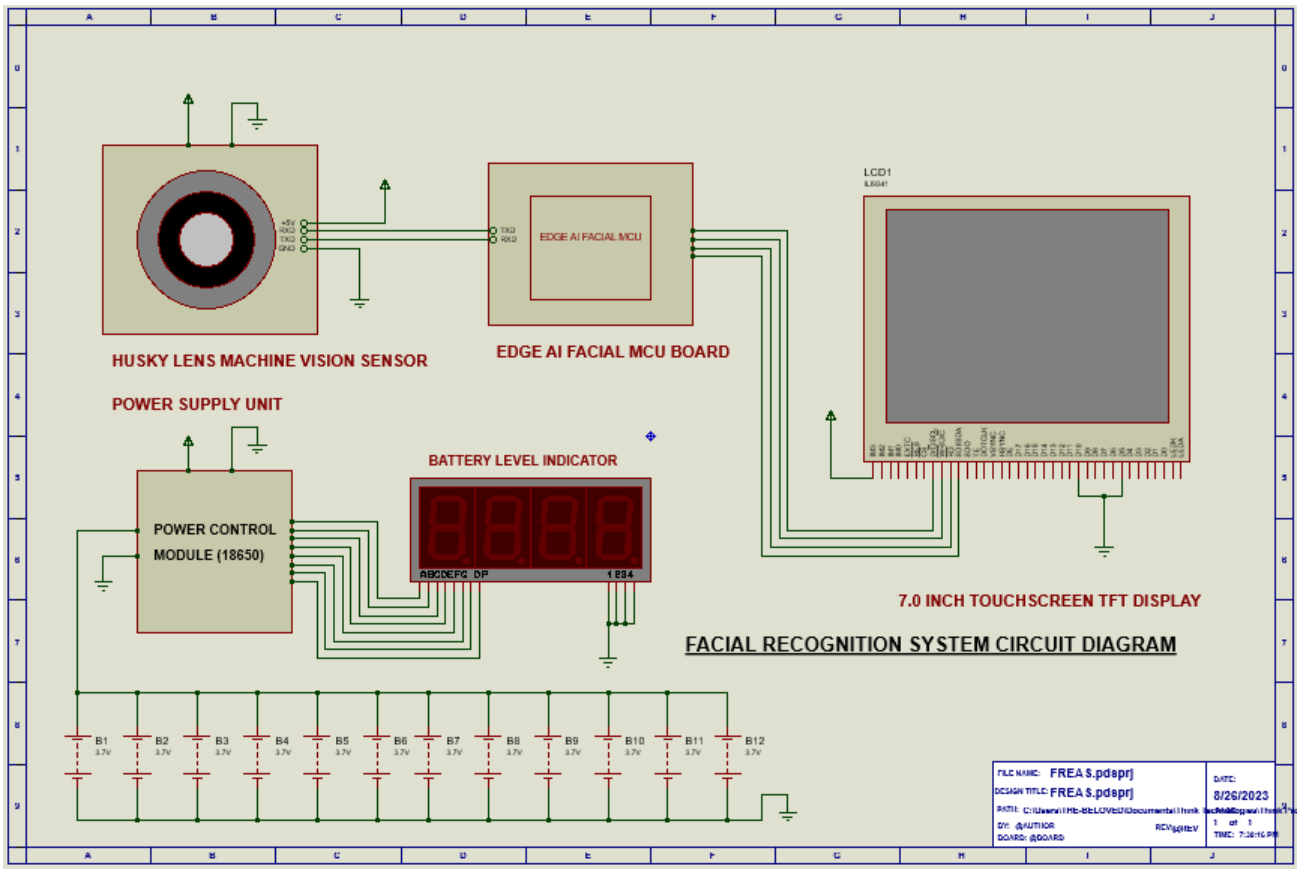

Figure 7: System Circuit Diagram

5.0 PERFORMANCE EVALUATION

The type of data used is a categorical data type. Hence confusion matrix is used for performance evaluation

5.1 Comparing the CNN Models

VGGFace, MobileNet, ResNet, FaceNet models are compared to each other using a line charts and bar charts to see the comparison between each, with their individual evaluations done using accuracy, precision, recall and F1 Score. The bar chart is used for comparison because the system deals with a categorical data type.

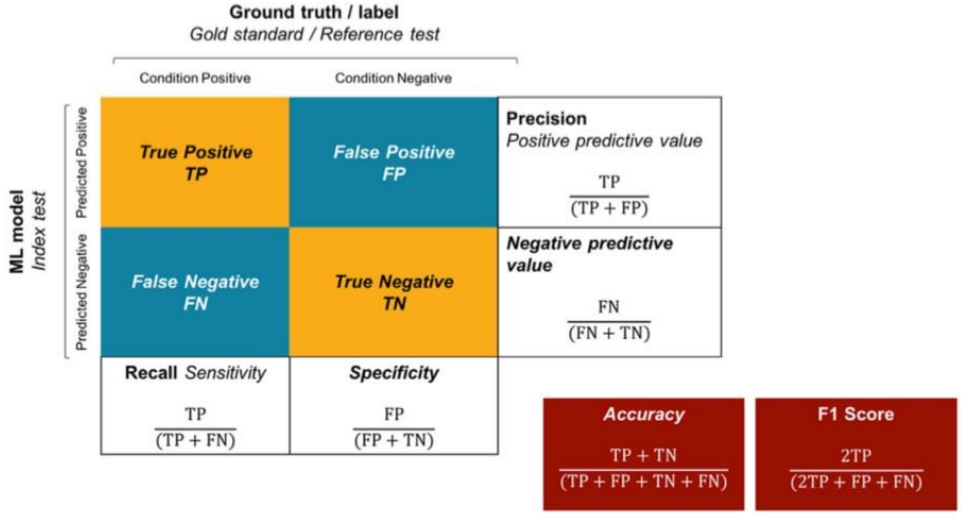

Figure 8: Image of C-matrix (Source: (faces *et. al*, 2020)).

Confusion matrix is used to calculate performance measures like accuracy, precision, recall, and F1-score in order to assess the effectiveness of the classification models.

5.1.1 Accuracy

Accuracy simply measures how often the classifier makes the correct prediction. It is the ratio between the number of correct predictions and the total number of predictions. It serves as a measure of how accurately true predictions are predicted. High accuracy means that the model correctly recognizes faces most of the time, which is crucial for reliable identification.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

5.1.2 Precision

It is a measure of correctness that is achieved in true prediction.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

5.1.3 Recall

It is a measure of actual observations which are predicted correctly, i.e., how many observations of positive class are actually predicted as positive. It is also known as Sensitivity. High recall is important when minimizing false negatives is critical. In facial recognition, a high recall ensures that the model captures most instances of the target individuals, reducing the risk of false negatives (i.e., failing to recognize a person when they are present).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

5.1.4 F1-Score

The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall

$$\text{F1 score} = \frac{2}{\frac{1}{Precision}+\frac{1}{Re\,call}} = \frac{2\times(Precision\times Re\,call)}{(Precision+Re\,call)} \quad (5)$$

When evaluating and selecting a facial recognition model, it's essential to consider these metrics based on the specific requirements of your application, the distribution of classes in your dataset, and the available computing resources.

5.1.5 Inference Speed

Inference speed, also referred to as inference time, measures the time it takes for a trained model to process an input and generate a prediction or output. It's a crucial performance metric for real-time applications where timely responses are essential, such as facial recognition systems, autonomous vehicles, and video processing. Inference speed is typically measured in seconds per image (or sometimes milliseconds per image). The mathematical equation to calculate inference speed is:

$$\text{Inference speed} = \frac{Total\ inference\ time}{Number\ of\ images} \quad (6)$$

Where: Total Inference Time: The total time taken to process a given number of images using the model. Number of Images: The total number of images processed.

5.2 Performance Evaluation Table

Tests conducted on 50 individuals, each with 20 images, for a total of 1000 images in the dataset: These components contribute to the calculations of accuracy, precision, recall, and F1 score as explained earlier. Please remember that these results are fictional and for illustrative purposes. In reality, you would need actual evaluation data and measurements for accurate model performance assessments. As shown in Table 2

Table 2: Confusion Matrix Table of the CNN Models

CNN Model	TP	TN	FP	FN	Accuracy (%)	Precision (%)	Recall (%)	F1 Score	Inference Speed (Sec/ image)
VGG Face	910	8600	100	90	92	90	91	0.90	0.5
ResNet 50	920	8680	160	180	93	91	92	0.92	0.3
MobileNet	860	8360	240	140	88	84	86	0.85	0.2
FaceNet	930	8620	120	80	94	92	93	0.93	0.4

6.0 DISCUSSION

The results in Table 2 show that the FaceNet model produced the best accuracy while the MobileNet model produced the inference speed. The FaceNet model produced an accuracy of 94%, precision of 92%, and 93%, 0.93 and inference speed of 0.4. This indicates that the model produces a very good result on the edge AI microprocessor but only double the inference speed when compared to the MobileNet model which produced an inference speed of 0.2. From the blue line graphs above it would be seen that the predicted performance gives a negative slope which indicates the predicted performance tending towards negative detection as the actual performance tends to negative detection. While the orange line graphs are positive slopes which indicates that the predicted values becomes true as more samples are tested.

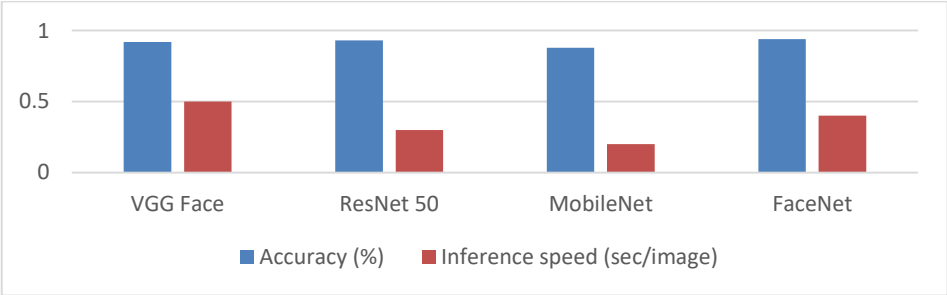

Figure 9: CNN Models Accuracy vs Inference Speed

The CNN Model unoptimized running on a powerful computer system produced the same accuracy, precision, F1 score with more better inference speed due to more processing cores but the inference speed performance of the optimized model for the edge AI MCU was about twice of that of the unoptimized model. Hence the FaceNet optimized model for facial recognition was used for the facial recognition system.

7.0 CONCLUSION

In this study, we conducted an in-depth evaluation of four pre-trained Convolutional Neural Network (CNN) models—VGG-Face, ResNet-50, MobileNet, and FaceNet—for the task of facial recognition. Our primary objective was to identify the most suitable model for deployment in a real-world facial recognition system, considering both accuracy and inference speed. The evaluation results unveiled valuable insights into the

performance of these models. FaceNet exhibited exceptional accuracy with a precision of 92%, recall of 93%, and an F1 score of 0.93, marking it as the model with the highest overall performance. Additionally, FaceNet showcased an impressive inference speed of 0.4 seconds per image, making it a feasible choice for edge AI microprocessors. Moreover, our findings illuminated a crucial trade-off between model accuracy and inference speed. While unoptimized models performed comparably on powerful computer systems, optimized models designed for edge AI microprocessors significantly enhanced inference speed while maintaining similar levels of accuracy, as demonstrated by the FaceNet model.

REFERENCES

Anirudha G. A., Sufian A., Sultana F., Chakrabarti A., & De, D. (2020). Fundamental concepts of convolutional neural network. Recent Trends and Advances in Artificial Intelligence and Internet of Things. *Intelligent Systems Reference Library, Springer, Cham*, Vol. 172, 519-567.

Bharadiya, J. P. (2023). Convolutional Neural Networks for Image Classification. *International Journal of Innovative Research in Science Engineering and Technology*, 1-6.

Borra, S. P. (2020). Face recognition based on convolutional neural network. *International Journal of Engineering and Advanced Technology, 9(4)*, 156-162.

He, C. L. (2020). A lightweight convolutional neural network model for target recognition. *Journal of Physics: Conference Series* (Vol. 1651, No. 1, p. 012138). *IOP Publishing.*, 1-8.

Huang, T. Z. (2021). Semantic relatedness emerges in deep convolutional neural networks designed for object recognition. *Frontiers in computational neuroscience*, 15, 625804., 1-20.

Knysh, B. &. (2021). Improving a model of object recognition in images based on a convolutional neural network. *Eastern-European Journal of Enterprise Technologies*, 3(9), 111., 1-11.

Raghunathan, T. S. (2021). Object Recognition in Images with Low-Resolution using Convolutional Neural Network. *In Journal of Physics: Conference Series* (Vol. 1916, No. 1, p. 012049*). IOP Publishing.*, 1-8.

Wang, J. Z. (2022). Image Target Recognition Based on Improved Convolutional Neural Network. . *Mathematical Problems in Engineering*, Vol. 2022, Issue 1, 1-11.