International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No. 1 (Mar-24)

http://dx.doi.org/10.12785/ijcds/150189

Comparison of YOLO (V3,V5) and MobileNet-SSD (V1,V2) for Person Identification Using Ear-Biometrics

Shahadat Hossain[1], **Humaira Anzum**[1] **and Shamim Akhter**[1]

[1]*AISIP Lab, Dept. Of Computer Science and Engineering Ahsanullah University of Science and Technology Dhaka, Bangladesh*

Received 18 Apr. 2023, Revised 11 Feb. 2024, Accepted 5 Jan. 2024, Published 10 Mar. 2024

Abstract: The ear is a visible organ with a unique structure for each person. As a result, it can be used as a biometric to circumvent the constraints of person identification. Deep learning methods like You Only Look Once (YOLO) and MobileNet have recently significantly aided real-time biometric recognition. As a result, in this paper, we approach identifying a person using YOLOV3, YOLOV5, MobileNet-SSDV1, and MobileNet-SSDV2 deep learning algorithms using their ear biometrics. The used ear biometric is a standard dataset (EarVN1.0 Dataset) from 164 individuals with a total of 27,592 images. We chose 10 people at random, totaling 2057 pictures. Of these, 85% were used for training, 5% for validation, and 10% for testing. The performance of the algorithms is determined based on their accuracy and how smoothly the ear of a person is detected. The training accuracy of the algorithms is thresholded at 99.87%. MobileNet-SSDV1, MobileNet-SSDV2, YOLOV3, and YOLOV5 have testing accuracy that is 88%, 91%, 95%, and 96%, respectively. We concluded that the YOLOV5 model outperforms the others in terms of accuracy and size (16MB) for person identification using ear biometrics.

Keywords: Person Identification, Ear Biometric, Deep Learning, YOLO, MobileNet, SSD

1. INTRODUCTION

Secure authentication is necessary for many applications, including identifying a person to access complicated systems, screening and detecting unauthorized users, safeguarding resources from unique threats, and protecting entities. Biometric verification examines biological qualities such as a person's retinas, irises, blood vessels, vocals, appearance, and patterns of fingers to identify them. This technique is realistic and very difficult to replicate. As a result, a lot of studies have been done on the use of biometric patterns to gain access to various valuable resources. Human ear shape and characteristics can be used as a biometric feature for the identification of individuals. A person can be distinguished from other people based on the biological and geometrical features of their ears. In addition, a person's ear outer structure stays the same over time or varies very little. To detect ear biometrics, several techniques can be used, such as feature-based template matching [1], [2], [3] and mathematical morphological operations [4], [5] including dilation, erosion, etc.

Cropping ear images can benefit from computer vision because required details like shape, contour, edge, curves, and graphs are extracted through image processing; however, this process still necessitates a lot of manual image editing.

The recent development of computer vision algorithms like different types of convolutional neural networks [6] has reduced the number of manual works required by turning image identification, classification, and object detection processes into a straightforward automated pipeline. In recent times YOLOV3 [7] and YOLOV5 [8] algorithms have been used for ear detection in real time. Due to its quick inference speed and high precision, YOLO has been acknowledged as one of the most reliable object detectors [9], [10], [11], [12], [13], [14]. MobileNet [15] is yet another straightforward, effective, and lightweight convolution neural network (CNN) for smartphone applications. Numerous real-world apps, such as object detection, fine-grained classifications, face attributes, and localization, make extensive use of MobileNet [16], [17], [18]. YOLOV5 [8] and MobileNetSSD [15] are used independently for detecting ear objects in person face images. In addition, YOLOV3 was evaluated in our previous works [19], [20] individually to recognize a person based on ear images.

An effective and fast biometric security system must be able to identify and authenticate people based on their unique biological characteristics. The performance of the YOLO and MobileNetSSD algorithms for person identification on the EarVN1.0 benchmark dataset is thus compared in this study [21]. The EarVN1.0 dataset consists of 28,412 ear images from 164 distinct people. Images are distinct according to their size, luminance, occlusion,

E-mail address: shahadathossain8721@gmail.com, humaira.snigdha.22@gmail.com, shamimakhter@gmail.com

resolution, lighting conditions, and other factors.
The following are the primary contributions of this research:

- Introducing YOLO (V3, V5) and MobileNet-SSD(V2, V3) models for identifying individual persons using ear biometrics.
- Selecting the best possible model with higher accuracy among the two-stage and one-stage deep learning models detecting the human ear in real time for its application in the biometric security system. We expect the best model will work effectively in embedded devices and authenticate people with higher precessions.

This section briefly introduces the overall objectives of our current study. A brief overview of relevant efforts on ear identification is provided in Section II. Section III describes the architectures of the models, and is also responsible for preparing the data sets, and models for evaluation. Section IV presents the findings and analyses. Finally, in Section V, we end this work with final thoughts and future work.

2. RELATED WORKS

French criminologists identified the distinctiveness of the ear shape for the first time in 1890 and proposed using it as a biometric for people [22]. A manual ear identification system [23] was developed in 1949. In their study in 2012[24], Elsayed Hemayed and M. Fayek used image processing to convert color pictures to grayscale values before removing noise. Additionally, edge detection was used to identify the closed boundary's roundness condition, height-to-breadth ratio, and other parameters.
This work could only identify static 2D images due to its inability to distinguish rotation and scale differences, which prohibited it from detecting 3D images. Additionally, the sample that was used was quite small.
On-ear detection from images, there are mainly two (2) directions for research. The first is to recognize the ear as an object in facial or video frames, and the second is to recognize a person from his or her ear images. Retinaface was proposed by Huy Nguyen Quoc and Vinh Truong Hoang in [25] as a way to identify ears from an image and train it using YOLOV3. This was developed to replace the time-consuming image processing from previous works and to speed up image-based training. To identify faces in images and segment the ear, Retinaface was used. YOLOV3 performed admirably in terms of precision and speed. A CNN was trained using a collection of 2735 manually annotated ear images, each with 45 interest points, for the third publication by Victor Acuna, Carolina Paschetta, et al. [26]. To avoid overfitting and achieve a high generalization rate, specialized learning techniques were used. The geometrical Features Extraction algorithm for ear recognition was developed in [27] and is invariant to scaling, translation, and rotation. Huy Nguyen Quoc and Vinh Truong Hoang recently released their second paper on ear detection using YOLOV5, which had an accuracy of 82.5% but was faster than YOLOV3[28]. Faster R-CNN was used for ear detection and localization [29]. Xuebin Xu et al [30] released a paper about human ear recognition based on MobileNet V2. They used AWE and EARVN1.0 human ear datasets. MobileNet V2 accuracy was 80.51% and 91.09%. SSD MobileNet V1 was applied to human ear image collection and recognized the ear images with 98% accuracy. Deep CNN was trained on the IIT Delhi ear image database with 100-200 different subjects in the second stage of research and obtained 96%-97% accuracy [31]. ResNet50 identified the individual in [32] using AWE data sets with left and right-side ear images. The YOLO is utilized to identify the ear images and identify the corresponding individual, and the model training accuracy is 82.5%.
Convolutional Neural Networks (CNN) are used to identify individual persons using font view and side view human ear images [33]. This model test accuracy is 84% for font view and 80% for side view. Convolutional Neural Network (CNN) and Shape Mapping technique (CCM) method using ear recognition techniques to improve the security and safety level[34], and this model's predicted accuracy is 85%. Namitha Santhosh et. al.[35] Implementation ear biometrics authentication system using deep learning model YOLOV3 in the web application, instead of traditional authentication system. The NASNet model used an unconstrained ear dataset [36], with an accuracy of 50.4%. For ear detection and ear identification tasks, Faster-RCNN and VGG-19 are employed [37]. Recently, Shamim Akhter et al [19], [20] used the YOLOV3 model with a standard dataset named EarVN1.0. This collection was made up of 28412 ear pictures from 164 different groups. The PCA, ICA method (a non-deep learning strategy), and YOLOV3 algorithm accuracy were then compared. The test accuracy for YOLOV3 to recognize people from their ear images was approximately 85%. The studies from [25], [26], [27], [28], [29] and [31] showed ML models for recognizing and identifying the ear from image data sets. However, [19], [20], [32], [37] studies use ear images to identify individuals.
A typical classification issue requires the detection and localization of objects. Because of the availability of huge amounts of data, optimized algorithms, and high-speed processing capabilities, ML models for object detection are produced. To identify objects from images, two types of ML models are commonly available: two-stage and one-stage. RCNN, Fast RCNN, Faster-RCNN, and other two-stage ML models identify objects in two steps. The first step is to define an area of interest that has a high chance of being an object, and the second step is to classify the objects or regress their bounding boxes. The one-stage ML models, such as YOLO, SSD, RetinaNet, and others, use image regression to produce class probability and bounding box coordinates. Due to the simple structure of single-stage detection techniques can be easily combined with the Internet of Things to handle real-time application scenarios [38]. Our experiments primarily concentrate on one-stage ML models due to their better performance than two-stage

ML models[39], [40], [41] and evaluate their abilities to identify humans based on ear biometrics.

In our earlier research, we found that YOLOV3 (Darknet) outperformed CNN (VGGNet) at detecting human identification via ear biometrics. The CornerNet model will not be a good solution for sophisticated small items or multi-object groups [38] since it does not take the information inside the bounding box into account. RetinaNet and CornerNet were also tested in [25] with YOLOV3 performance. CornerNet performs worse than YOLOV3 in accuracy and timeliness. RetinaNet performed better in terms of accuracy, but training costs are three times as much. Similar outcomes also found in [42]. The mean average precision (MAP) of RetinaNet reached 82.89%, but the frames per second (FPS) is much higher than YOLOV3, making real-time performance challenging. SSD performs poorly on the MAP and FPS indicators. YOLOV3 has a little lower MAP than the others (80.69%), but it has a huge edge in terms of detecting speed. YOLOV3 also outperformed when tasked with hard sample detection, implying that YOLO models are better suited for deployment in real-time applications. Recently, Huy Nguyen Quoc and Vinh Truong Hoang published research outcomes on-ear detection using YOLOV5, which was quicker than YOLOV3 [28] and had an accuracy of 82.5%. Thus, our model should concentrate on accuracy, timeliness, and suitability for porting into embedded devices for real-time implementation. Thus, YOLO and MobileNet families are chosen for our experiments.

According to the related studies, there are no commercially accessible ear recognition systems. Human identification using ear biometrics or facial biometrics, on the other hand, has enormous potential. As a result, we intend to develop an effective and fast biometric security system for identifying humans based on real-time captured ear images. The first stage in building this type of embedded system is to upload a pre-trained recognition model. YOLO and MobileNet are two recent deep-learning techniques notable for their ability to detect objects quickly. In this study, we compare these two methods with their various versions and select the best one to use in our biometric security system.

3. DATASET AND MODEL CONSTRUCTION

A. Dataset for Benchmarking

EarVN1.0 is one of the largest publicly available ear datasets which was developed in 2018 and contains 28,412 ear images from 164 distinct individuals [3]. In this study, we choose 2057 (100%)images from random 10 people, among them 1642(80%) random images are taken for training, 305 random images for validation(15%), and 110 random(5%) images for testing. All images are renamed with a specific format like Person1_.....jpg, Person2_.....jpg, Person3_.....jpg. Person1's name starts Person1_...jpg, Person2's name starts Person2_....jpg, Person3's name starts Person3_....jpg, and continue till Person10's name starts Person10_....jpg. Before the first underscore (_) all image naming convention is matched with the Person's name like Person1 images name "Person1_1_jpg.rf.132f689363b0277d0b4b2c94c46b4155.jpg" and after the first underscore (_) naming convention with random character on a full naming convention every image.

B. Architecture of the Models

1) You Only Look Once (YOLO)

Object identification in the YOLO model happens reasonably quickly because only a single transmission through the Convolution Neural Networks (CNN) is required. The YOLO model is composed of three (3) operations: Residual Blocks, Bounding Box Regression, and Intersection Over Union (IOU), which simplify its design while increasing accuracy. In Residual Blocks operations, the entire image is partitioned into multiple grids of S x S size, and the grids are responsible for detecting objects within them. Following that, the Bounding Box Regression operations projected the Bounding Box parameters for each object, including height, width, center, class, and confidence scores. Intersection Over Union (IOU) operations are performed on the bounding boxes of the same objects with their confidence scores to determine the best fit. YOLO has several versions, including V1-V5, with V3-V5 being more stable and resolving overfitting issues than previous versions. YOLOV3 employs Darknet 53, Residual Block, skip connection, and up sampling to increase precision. YOLO V4 replaces Darknet53 as the backbone with CSPDarknet53, increasing speed and precision. The YOLOV5 is the latest and lightest variant of the previous one. The PyTorch framework is used instead of the Darknet framework (Figure 2), but the backbone is CSPDarknet53 (Figure 3). It is visible in Figure 2 and Figure 3 that CSPNet implements a similar fashion as DenseNet, however, the feature map is partitioned into two sections. One half is routed through the thick block and the transition layer, while the other is integrated with the transmitted feature map in the following stage. As a consequence, we chose V3 and V5 stable versions with two distinct backbone architectures for our application's implementation and compared the performances.

2) MobileNet-Single Shot Detector (SSD)

MobileNet is a portable, one-stage ML network designed for mobile and embedded system applications. To minimize model size and complexity, the MobileNet architecture employs depth-wise separable convolution (depth-wise convolution and pointwise convolution). The model's primary factors are the width multiplier and the resolution multiplier. It adjusts them during training to meet the speed and small size criteria. V1-V3 are the three (3) variants of MobileNet. V2 adds inverted residual blocks and linear bottlenecks to V1 architecture and the ReLU activation function is replaced by the ReLU6 activation function. V2 and V3 almost have identical designs. As a result, we decided to deploy V1 and V2 for our application and evaluate their performances on human identification.

A feed-forward convolution network called a Single Shot Multibox Detector (SSD) that generates multiple bounding boxes and scores them based on the existence of objects. The final detections are then output using a non-maximum suppression layer. SSD typically employs an additional

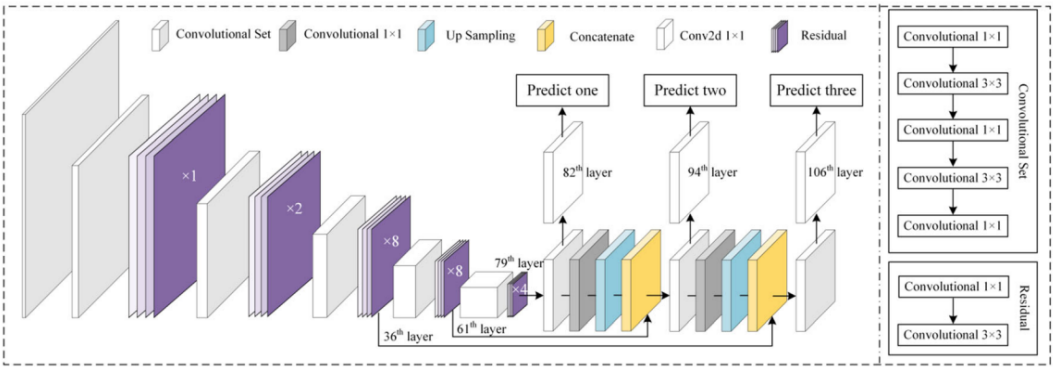


Figure 1. YOLOV3 network architecture [43]

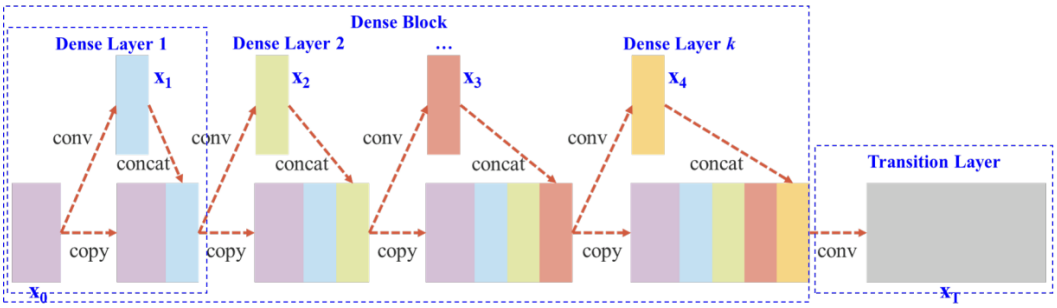


Figure 2. Illustrations of DenseNet[44]

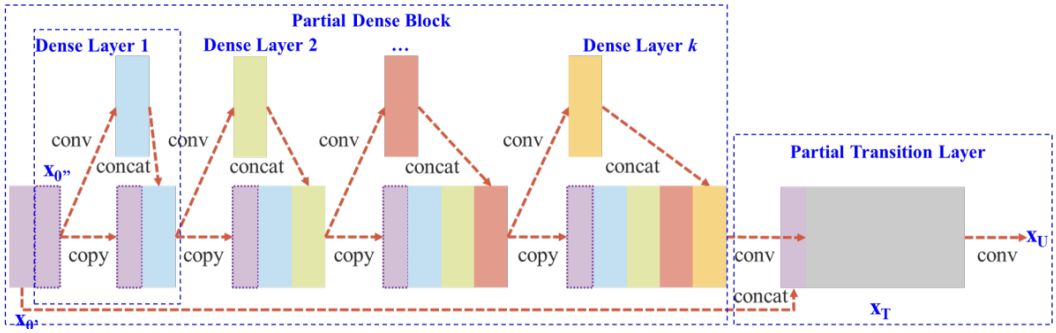


Figure 3. Illustrations of CSPDenseNet[44]

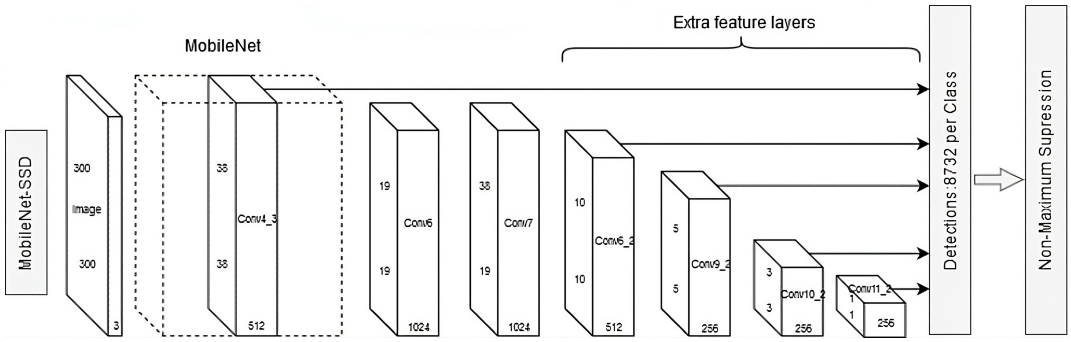


Figure 4. MobileNetSSD V1 network architecture

framework as a base network, such as Faster-RCNN or MobileNet, to extract features and build an ensembled form for significantly faster and more accurate detections. For our comparative purposes, we merge the SSD and MobileNetV1 in Figure 4 as an ensemble detector or classifier.

C. Setup the Models

First, we must make annotations for particular objects inside the images. To generate annotation files for our ML models, we use the online Roboflow app [45]. YOLOV3 and YOLOV5 are annotated on txt file formats that match with image file names, such as imagename.txt, and annotated file individual names in data.ymal file for training purposes. Pretrained models for YOLOV3 and YOLOV5 were downloaded from [46] and [47] respectively. The default hyperparameters are shared from the data/hyps/hyp.scratch-low.yaml and execute the following command to train the model.
!python train.py –img 640 –batch 16 –epochs 260 –data data.yaml –weights yoloV5s.pt
The model inherits only three(3) parameters from the users. The rest of the hyperparameters are initialized automatically from default settings.

MobileNet-SSDV1 and MobileNet-SSDV2 are annotated on Pascal VOC mean xml file format match with the image file name, like imagename.xml. By default, the Roboflow software resizes images to 640x640 pixels. The 640x640 image format is used by the YOLOV3, YOLOV5, and MobileNet-SSDV2 devices. Images for MobileNet-SSDV1 are reduced to 300x300 pixels. MobileNet-SSDV1 and MobileNet-SSDV2 pre-trained models were obtained from [48]. The models were built in Pytorch and trained on the Open Images dataset. The models were trained using transfer learning methods with labeled training data from the EarVN1.0 dataset.
Figure 5 presents the overall methodological steps to train the models.
We wrote a script that reads testing ear images sequentially and identifies the individual. Figure 6 depicts the image and its associated findings. Figure 7 represents the classes name and image quantity for the training of the dataset. Figure 8 represents the classes name and quantity for validation of the dataset. Figure 9 represents classes name and image quantities for the test of a dataset. Figure 7, Figure 8, and Figure 9 present which class was used and how many images were used for training, validation, and testing.

4. RESULT AND ANALYSIS

A. Applying Transfer Learning on Pre-Trained Models

The four ML models including YOLO (V3, V5) and MobileNet-SSD (V1, V2) were trained on Pascal VOC, COCO, and Open Image datasets. All the models are retrained with the EarVN1.0 dataset. Table I. presents the training status of the models. The models are trained with 1642 images of 10 individuals. MobileNet-SSDV1 completes 200 epochs in 4.19 hours, MobileNet-SSDV2 completes 200 epochs in 3.40 hours, YOLOV3 completes 30000 epochs in 11.20 hours, and YOLOV5 completes 260 epochs in 4.50 hours to achieve the predetermined training accuracy (99.5%). It appears that YOLOV3 is the slowest and MobileNet-SSDV2 is the fastest training model to achieve the desired precision.

TABLE I. Training Analysis of Models

Model	Epochs	Time	Retrained Model Size
MobileNet-SSDV1	200	4.19 hours	32 MB
MobileNet-SSDV2	200	3.40 hours	16 MB
YOLOV3	30000	11.20 hours	247 MB
YOLOV5	260	4.50 hours	16 MB

When the overall number of epochs is 200, Figure 10 shows the loss value of the MobileNet-SSDV1 model after retraining. During the training time, the validation epoch is set to 1. So, after the completion of every epoch, the model demonstrates the training time to continue till it reaches the 200 epochs. It is noted that the Loss value is decreasing with the increment of the Epochs. The minimum Loss value is achieved in 197 epoch points. Similarly, Figure 11 presents the Loss value of the re-train MobileNet-SSDV2 model where the total epoch number is 200. The graph shows the 152 epoch point as the lowest Loss. The YOLOV3 model's Loss value curve is shown in Figure 12 with a total training period count of 30000. The average loss over 30000 number epochs is 0.0414. The YOLOV5 model's training behavior is depicted in Figure 13. Over the increase in epochs, the accuracy, recall, and mAP values rise while the box_loss, obj_loss, and cls_loss values fall.

B. Performance Analysis of the Retrained Models

There are various methods for improving object recognition algorithms, including more accurate placement, faster speed, and more accurate classification. Three popular metrics used to analyze the model's prediction behavior, sensitivity to recognizing the object of interest and avoiding false alarms, and processing time are intersection over Union (IoU), mean average precision (MAP), and rendered frames per second (FPS). Inter-models performance comparisons, on the other hand, entail the selection of a common and easy method for assessing the models' performances. The F-measure or F1 score based on the confusion matrix can be used to evaluate algorithm performance, especially when the dataset is imbalanced. A higher confidence level and F1 score are frequently recommended [49]. Thus, we choose F1 score for inter-models performance comparisons.
The confusion matrices produced by the MobileNet-SSDV1, MobileNet-SSDV2, YOLOV3, and YOLOV5 are shown in Figures 11, 12, 13, and 14, respectively. The model's expected values are defined by the x-axis, and the

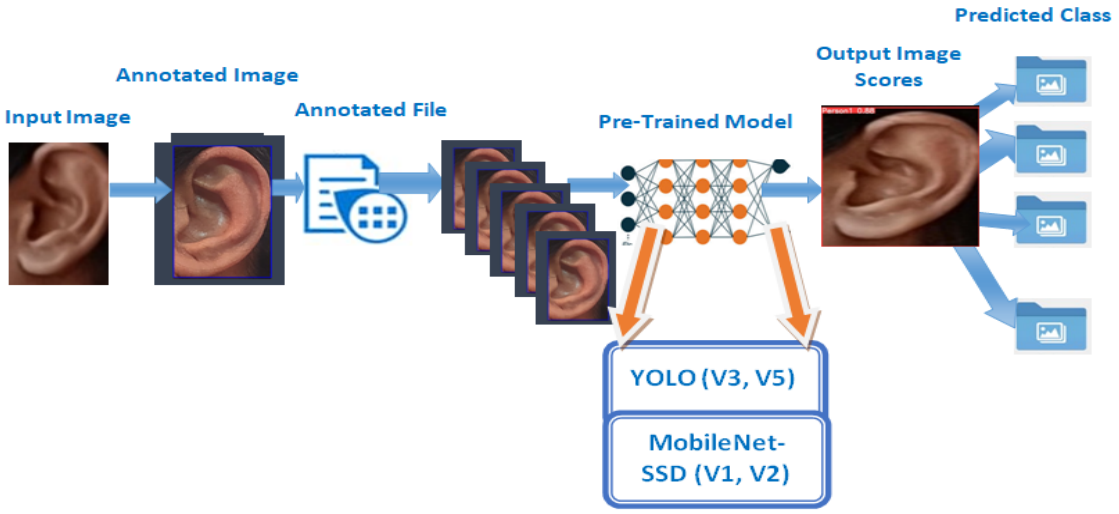
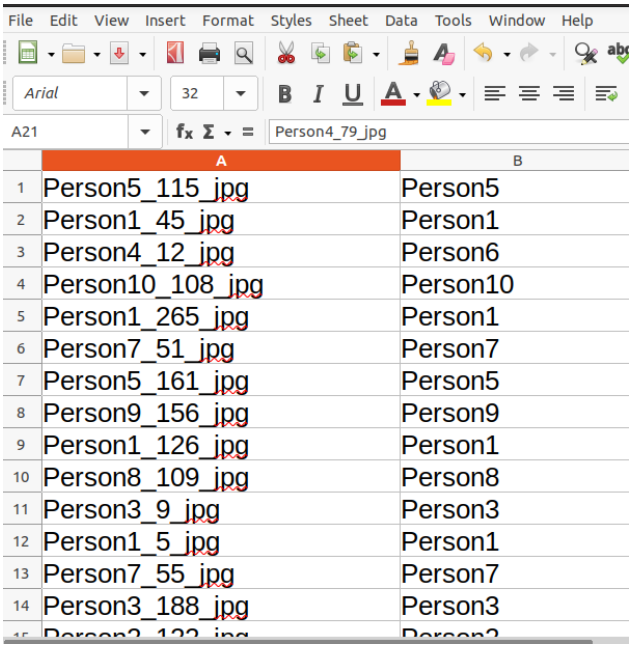


Figure 5. Presents the overall methodological steps to train the models.



	A	B
1	Person5_115_jpg	Person5
2	Person1_45_jpg	Person1
3	Person4_12_jpg	Person6
4	Person10_108_jpg	Person10
5	Person1_265_jpg	Person1
6	Person7_51_jpg	Person7
7	Person5_161_jpg	Person5
8	Person9_156_jpg	Person9
9	Person1_126_jpg	Person1
10	Person8_109_jpg	Person8
11	Person3_9_jpg	Person3
12	Person1_5_jpg	Person1
13	Person7_55_jpg	Person7
14	Person3_188_jpg	Person3

Figure 6. Results of the script

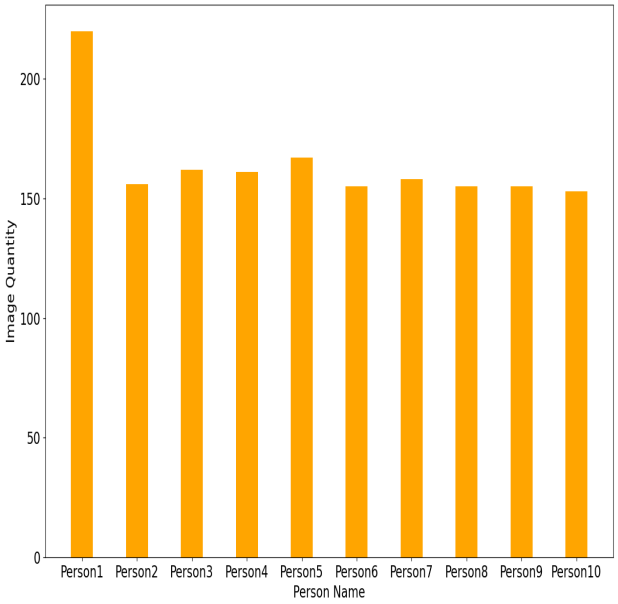


Figure 7. Classes and images quantity of training

actual values are defined by the y-axis. The 10x12(MxN) matrix structure was used to build this matrix. According to our model, some images have numerous objects detected while others have no objects detected at all. As a result, the matrix has two additional columns labeled False Negative and No Object Detection, respectively. The number of non-object detection images in MobileNet-SSDV1 (Figure 14) is zero, but four images are identified as multiples of person for Person2, Person3, Person6, and Person10. The number of non-object detection pictures in MobileNet-SSDV2 (Figure 15) is three for Person 5, Person 9, and Person 10. Person3 is multiplied by a single image detection. In YOLOV3 (Figure 16), the number of non-object detection images is zero, and the number of detection multiples person images is zero. In YOLOV5(Figure 17), the number of non-object detection images for Person1 and Person8 is two, while the number of detection multiples of person images for Person1 is one.

Tables II, III, IV, and V demonstrate the performance evaluation results for MobileNet-SSDV1, MobileNet-SSDV2, YOLOV3, and YOLOV5 using the test datasets. To compare the performance of individual identification, the precision, recall, and F1-Score were calculated using the One-vs-

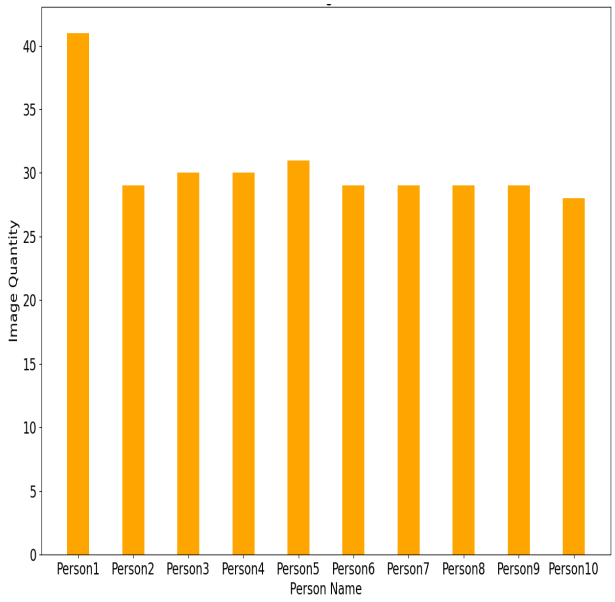


Figure 8. Classes and images quantity of validation

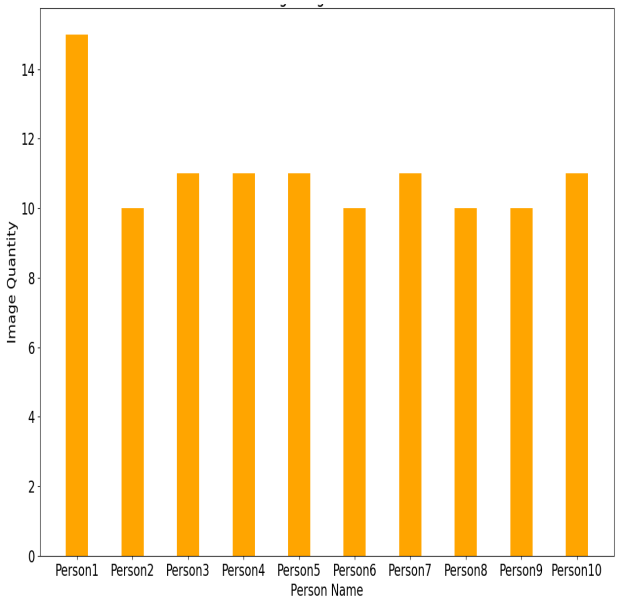


Figure 9. Classes and images quantity of testing

Rest (OvR) approach. However, rather than comparing numerous per-class values, it is preferable to compare them overall using their average values. The average precision number for MobileNet-SSDV1 is 0.902, MobileNet-SSDV2 is 0.926, YOLOV3 is 0.957, and YOLOV5 is 0.974, as shown in Table II-V. It seems YOLOV5 provides the best correct prediction of the identification classes among the positive class predictions. The average recall number for MobileNet-SSDV1 is 0.864, 0.895 for MobileNet-SSDV2, 0.956 for YOLOV3, and 0.952 for YOLOV5. So, all models

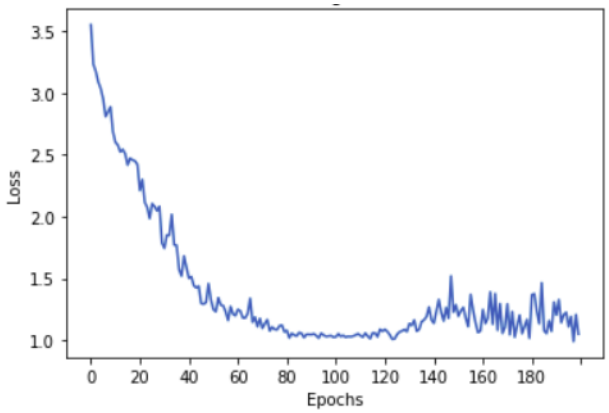


Figure 10. Loss value of MobileNet-SSDV1

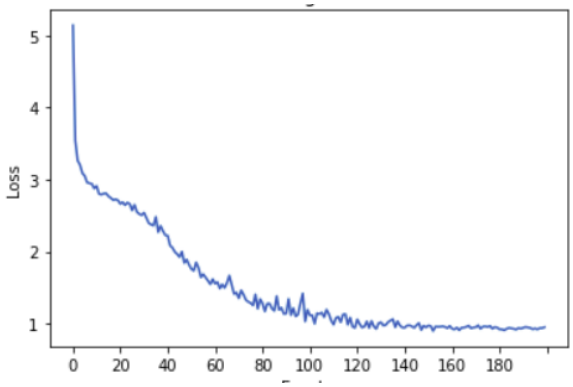


Figure 11. Loss value of MobileNet-SSDV2

can identify classes with their ear images in a proper way-however, YOLOV3 is the best to arrange the ear images into their right classes. The F1 scores are used to assess actual test accuracy. The average F1 score (Macro) for MobileNet-SSDV1 is 0.878, MobileNet-SSDV2 is 0.907, YOLOV3 is 0.954 and YOLOV5 is 0.961. Thus, in terms of precision and recall, YOLOV5 is the top performer, with 96.1% accuracy. So far, we have compared the models' recognition accuracy based on their average precision, recall, and F1-score values. We can also compare the models' success to some combined F1 scores, such as Micro F1 and Weighted F1. As a result, Table VI displays the worth of Micro F1 and Weighted F1 scores, as well as Macro scores, for all models. The results for MobileNet-SSDV1 Micro F1 and Weighted F1 are 88% and 88%, respectively. The results for MobileNet-SSDV2 Micro F1 and Weighted F1 are 91% and 91%, respectively. Scores for YOLOV3 Micro F1 and Weighted F1 are 95% and 95%, respectively. Scores for YOLOV5 Micro F1 and Weighted F1 are 96% and 96%, respectively. As a result of the above analysis, we can infer that the YOLOV5 model outperforms the others in terms of accuracy and size (16MB) for person identification using ear biometrics.

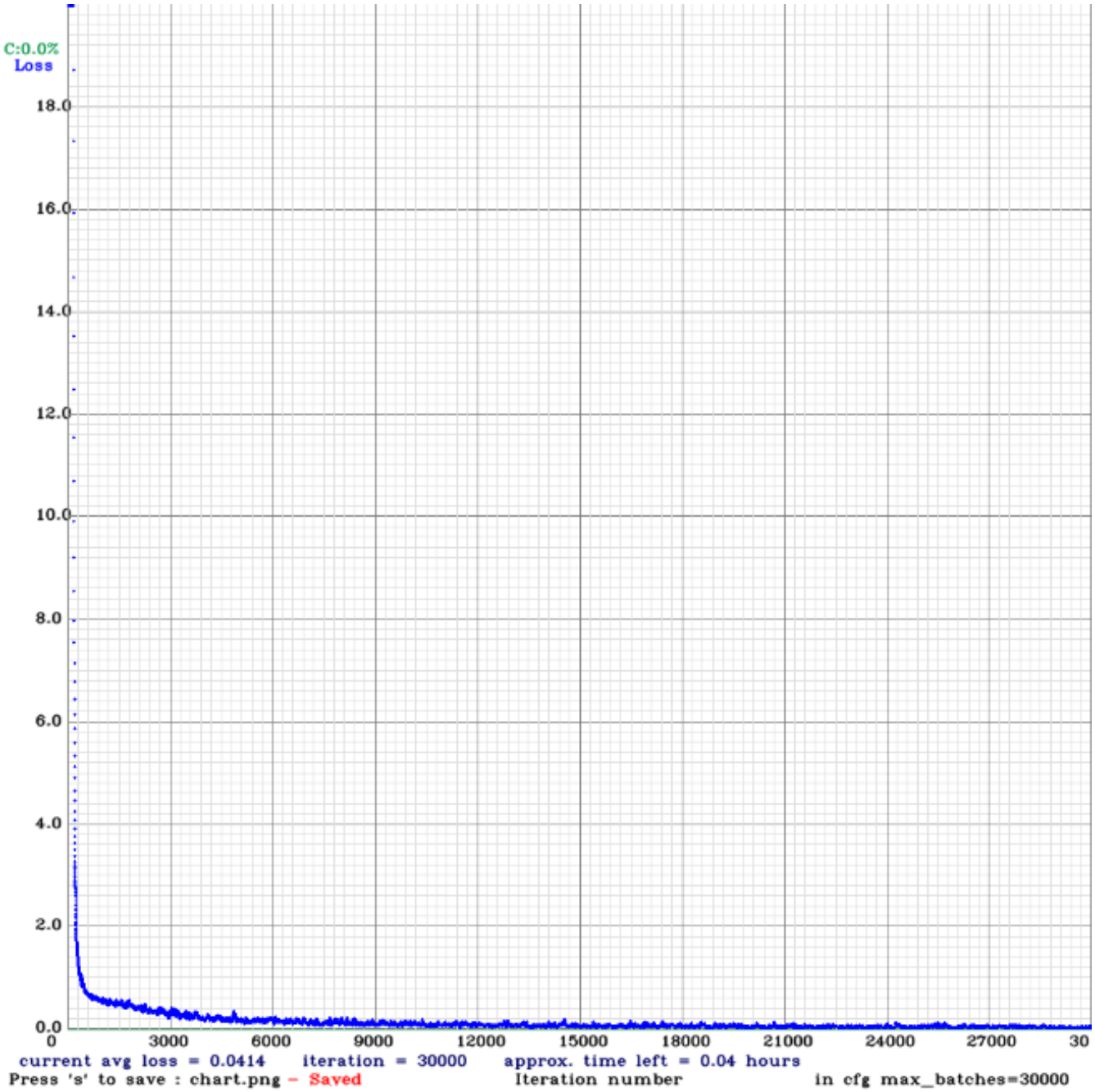


Figure 12. Loss value of YOLOV3

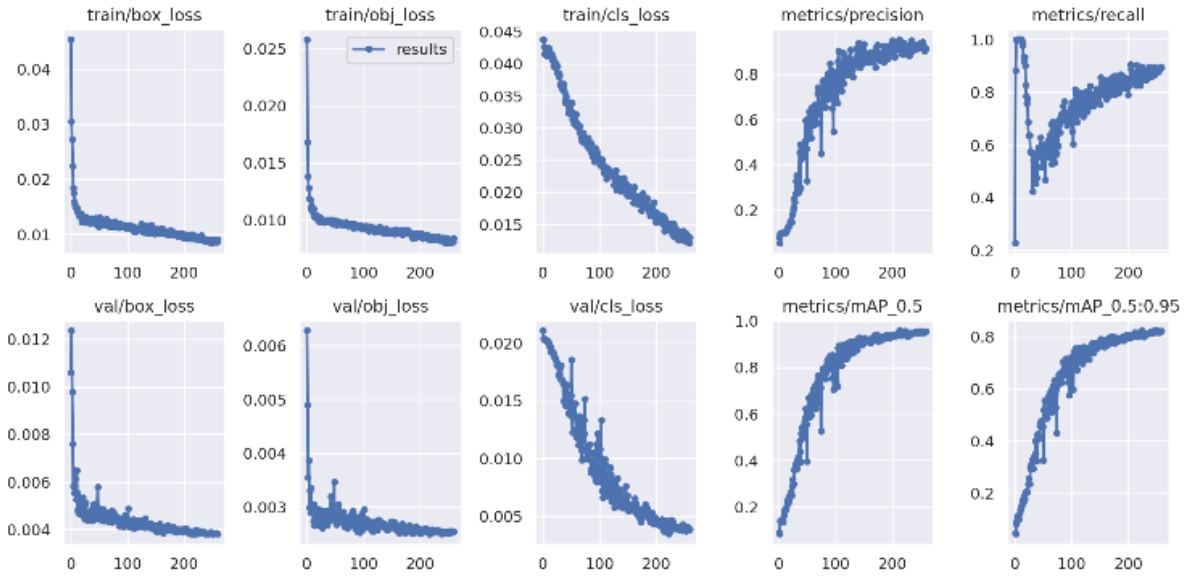


Figure 13. Training Result of YOLOV5

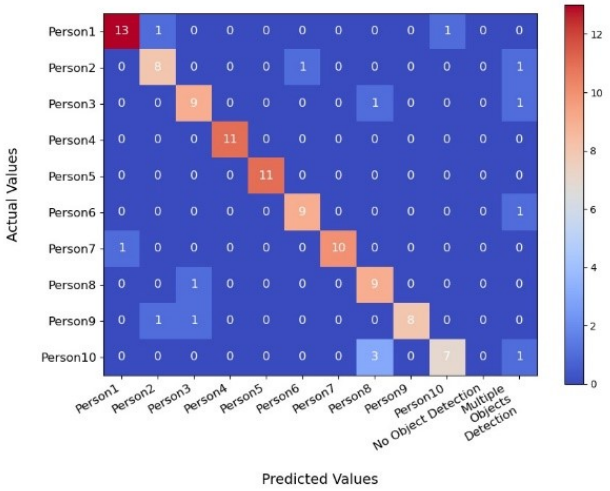


Figure 14. Matrix of MobileNet-SSDV1 Test Result

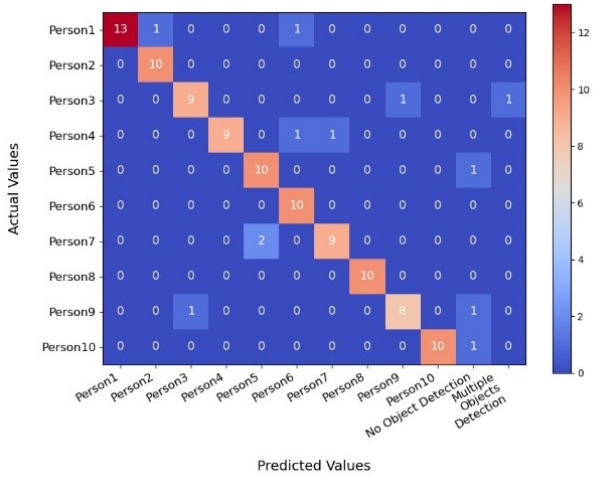


Figure 15. Matrix of MobileNet-SSDV2 Test Result

TABLE II. Test Results Analysis of MobileNet-SSDV1

Class	Precision	Recall	F1-Score
Person1	0.93	0.87	0.90
Person2	0.80	0.80	0.80
Person3	0.82	0.82	0.82
Person4	1.00	1.00	1.00
Person5	1.00	1.00	1.00
Person6	0.90	0.90	0.90
Person7	1.00	0.91	0.95
Person8	0.69	0.90	0.78
Person9	1.00	0.80	0.89
Person10	0.88	0.64	0.74

TABLE III. Test Results Analysis of MobileNet-SSDV2

Class	Precision	Recall	F1-Score
Person1	1.00	0.87	0.93
Person2	0.91	1.00	0.95
Person3	0.90	0.82	0.86
Person4	1.00	0.82	0.90
Person5	0.83	0.91	0.87
Person6	0.83	1.00	0.91
Person7	0.90	0.82	0.86
Person8	1.00	1.00	1.00
Person9	0.89	0.80	0.84
Person10	1.00	0.91	0.95

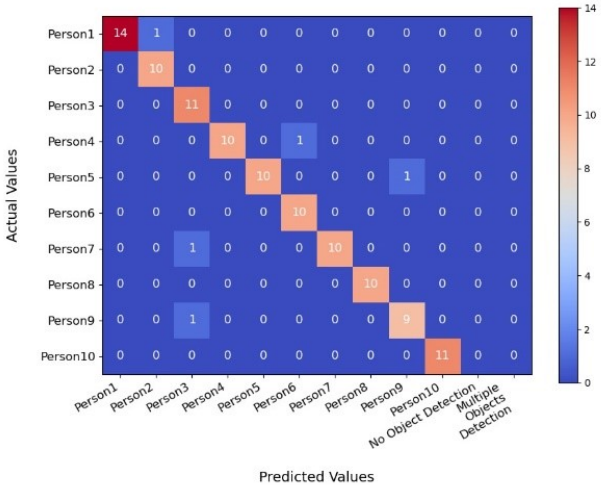


Figure 16. Matrix of YOLOV3 Test Result

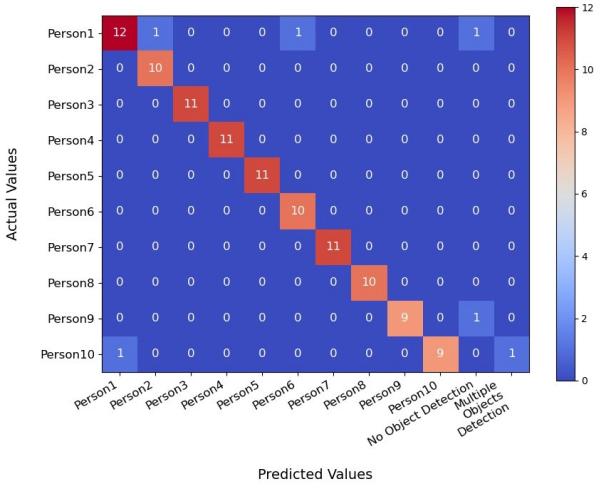


Figure 17. Matrix of YOLOV5 Test Result

TABLE IV. Test Results Analysis of YOLOV3

Class	Precision	Recall	F1-Score
Person1	1.00	0.93	0.97
Person2	0.91	1.00	0.95
Person3	0.85	1.00	0.92
Person4	1.00	0.91	0.95
Person5	1.00	0.91	0.95
Person6	0.91	1.00	0.95
Person7	1.00	0.91	0.95
Person8	1.00	1.00	1.00
Person9	0.90	0.90	0.90
Person10	1.00	1.00	1.00

TABLE V. Test Results Analysis of YOLOV5

Class	Precision	Recall	F1-Score
Person1	0.92	0.80	0.86
Person2	0.91	1.00	0.95
Person3	1.00	1.00	1.00
Person4	1.00	1.00	1.00
Person5	1.00	1.00	1.00
Person6	0.91	1.00	0.95
Person7	1.00	1.00	1.00
Person8	1.00	1.00	1.00
Person9	1.00	0.90	0.95
Person10	1.00	0.82	0.90

TABLE VI. Performance Comparison of YOLO (V3, V5) and MobileNet-SSD (V1, V2)

Model	Micro F1	Macro F1	Weighted F1
MobileNet-SSDV1	0.88	0.88	0.88
MobileNet-SSDV2	0.91	0.91	0.91
YOLOV3	0.95	0.95	0.95
YOLOV5	0.96	0.96	0.96

5. CONCLUSION

Ear biometrics for identity verification is growing in popularity every day. Different human biometrics components are identified using various algorithms (CNN, fast-CNN, faster-CNN). For our experiments, we used YOLO and two MobileNet models for ear-biometric person recognition. Both versions are widely utilized in embedded systems and are renowned for their quicker object detection. The performance of each model with different versions is measured using Precision, Recall, and F1-Score to identify certain class groups. The performance matrices between the models are then defined by calculating Macro, Micro, and Weighted F1 scores.Our findings show that the YOLOV5 Micro F1, Macro F1, and Weighted F1 are, respectively, 96%, 96%, and 96%. And among all the models, this is the best performance. Consequently, Macro, Micro, and Weighted F1-scores are computed to define the performance matrices of the models. Our findings show that the YOLOV version 5 Micro F1, Macro F1, and Weighted F1 are, respectively, 96%, 96%, and 96%. And among all the models, this is the best performance. As a result, we can conclude that for person identification using ear biometrics, the YOLO version 5 model performs better than the others in terms of accuracy and size (16MB).

Our study is limited by the lack of an automated cropping technique for selecting ear objects from images, as well as the processing time required for training models. In the future, we will use Jetson Nano or Raspberry Pi to create an effective and quick biometric security system capable of identifying and authenticating people based on their unique ear biometrics. Jetson Nano can work with the GPU

platform to overcome the training time constraint.

References

[1] M. K. Mohanapriya, "Ear biometric recognition using feature extraction for user identification," *International Journal of Scientific Development and Research (IJSDR)*, vol. 2, pp. 260–266, March 2017. [Online]. Available: https://www.ijsdr.org/papers/IJSDR1703042.pdf

[2] M. A. A. Tariq, M.A. Anjumt, "Personal identification using computerized human ear recognition system," *International Conference on Computer Science and Network Technolog*, pp. 50–54, December 2011. [Online]. Available: http://biomisa.org/uploads/2013/07/ear_2.pdf

[3] J. C. A. Kohlakala, "Ear-based biometric authentication through the detection of prominent contours," *SAIEE Africa Research Journal*, vol. 112, no. 2, June 2021. [Online]. Available: http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S1991-16962021000200005

[4] P. H.-D. M. Mursalin, M. Ahmed, "Biometric security: A novel ear recognition approach using a 3d morphable ear model," *Sensors*, 2022. [Online]. Available: https://www.mdpi.com/14248220/22/22/8988

[5] M. M. S. A. M. Mayya, "Human recognition based on ear shape images using pca-wavelets and different classification methods." [Online]. Available: https://oatext.com/Human-Recognition-Based-On-Ear-Shape-Images-Using-PCA-\Wavelets-and-Different-Classification-Methods.php

[6] R. B. Girshick, "Fast r-cnn," in *ICCV '15: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, December 2015, pp. 1440–1448.

[7] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint*, 2018. [Online]. Available: https://pjreddie.com/media/files/papers/YOLOv3.pdf

[8] H. N. Quoc and V. T. Hoang, "Human ear-side detection based on YOLOv5 detector and deep neural networks," in *2021 International Conference on Decision Aid Sciences and Application*, 2021.

[9] J. P. Lin and M. T. Sun, "A yolo-based traffic counting system," in *Proceedings of the 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. Taichung, Taiwan: IEEE, November 2018, pp. 82–85.

[10] R. Laroca, E. Severo, L. A. Zanlorensi *et al.*, "A robust real-time automatic license plate recognition based on the yolo detector," in *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)*. Rio, Brazil: IEEE, July 2018, pp. 1–10.

[11] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "YOLO-Face: A real-time face detector," *The Visual Computer*, vol. 37, 2021.

[12] Z. Qiu, S. Wang, Z. Zeng, and D. Yu, "Automatic visual defects inspection of wind turbine blades via YOLO-based small object detection approach," *Journal of Electronic Imaging*, vol. 28, no. 4, p. 043023, 2019.

[13] Z. Du, J. Yin, and J. Yang, "Expanding receptive field yolo for small object detection," in *Journal of Physics: Conference Series*, vol. 1314. IOP Publishing, 2019.

[14] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "Uav-yolo: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, p. 2238, 2020.

[15] Y. Lei, B. Du, J. Qian, and Z. Feng, "Research on ear recognition based on SSD_MobileNet_v1 network," in *2020 Chinese Automation Congress (CAC)*, 2020.

[16] A. G. Howard, M. Zhu, B. Chen *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 4510–4520.

[18] A. Howard, M. Sandler, G. Chu *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019, pp. 1314–1324.

[19] S. Hossain, S. S. Mitu, S. Afrin, and S. Akhter, "A real-time machine learning-based person recognition system with ear biometrics," *University of Bahrain Journal of Engineering and Technology*, vol. 2, no. 1, pp. 18–25, 2022.

[20] S. Hossain and S. Akhter, "Realtime person identification using ear biometrics," in *2021 International Conference on Information Technology (ICIT)*. IEEE, 2021, pp. 12–16.

[21] V. T. Hoang, "Earvn1.0," 2020. [Online]. Available: http://dx.doi.org/10.17632/yws3v3mwx3.4

[22] A. Bertillon, *La photographie judiciaire: avec un appendice sur la classification et l'identification anthropométriques*. Paris: Gauthier-Villars, 1890.

[23] A. Lannarelli, *Ear identification. forensic identification series*, 1989.

[24] E. Hemayed and M. Fayek, "Heard: An automatic human ear detection technique," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 5, no. 5, pp. 205–218, 2012.

[25] H. N. Quoc and V. T. Hoang, "Real-time human ear detection based on the joint of yolo and retinaface," 2021.

[26] C. Celia, M. Quinto-Sanchez, V. Acuna *et al.*, "Automatic ear detection and feature extraction using geometric morphometrics and convolutional neural networks," *Forensic science international*, vol. 272, pp. 14–22, 2017.

[27] A. S. Anwara, K. K. A. Ghany, and H. Elmahdyc, "Human ear recognition using geometrical features extraction," in *International Conference on Communication, Management and Information Technology (ICCMIT 2015)*. Elsevier, 2015.

[28] H. N. Quoc and V. T. Hoang, "Human ear-side detection based on yolov5 detector and deep neural networks," *arXiv preprint arXiv:2201.06570*, 2022.

[29] A. M. Alkababji and O. H. Mohammed, "Real-time ear recognition using deep learning," *TELKOMNIKA Telecommunication, Computing, Electronics, and Control*, vol. 19, no. 5, pp. 2298–2305, 2021. [Online]. Available: https://www.academia.edu/46876194/Real_time_ear_recognition_using_deep_learning

[30] X. Xu, Y. Liu, S. Cao, and L. Lu, "An efficient and lightweight method for human ear recognition based on mobilenet."

[31] R. A. Priyadharshini, S. Arivazhagan, and M. Arun, "A deep learning approach for person identification using ear biometrics," *Applied Intelligence*, vol. 50, no. 12, pp. 4269–4280, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s10489-020-01995-8

[32] K. R. Resmi, G. Raju, V. Padmanabha, and J. Mani, "Person identification by models trained using left and right ear images independently," in *Proceedings of the 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022)*. Atlantis Press, 1 2023, pp. 1–10.

[33] N. Petaitiemthong, P. Chuenpet, and T.-U. N. Auephanwiriyakul, S. and, "Person identification from ear images using convolutional neural networks."

[34] S. Pandiarajan, J. S. Kumar, A. R. Kumar, and B. Parkavi, "Authentication for door lock system through bio scanning ear image processing system," *Journal of Data Acquisition and Processing*, vol. 38, no. 2, pp. 4397–4404, 2023. [Online]. Available: https://sjcjycl.cn/article/view-2023/pdf/02_4397.pdf

[35] N. Santhosh, N. Afsal, S. Franco, T. Roshan, and L. Joseph, "Biometric authentication system using human ear," *International Journal Of Innovative Research In Technology (IJIRT)*, vol. 9, pp. 1257–1261, May 2023. [Online]. Available: https://ijirt.org/master/publishedpaper/IJIRT160077_PAPER.pdf

[36] K. Radhika, K. Devika, T. Aswathi, P. Sreevidya, V. Sowmya, and K. Soman, "Performance analysis of nasnet on unconstrained ear recognition," *Nature Inspired Computing for Data Science*, vol. 871, pp. 57–82, 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-33820-6_3

[37] A. Kamboj, R. Rani, and A. Nigam, "A comprehensive survey and deep learning-based approach for human recognition using ear biometric," *The Visual Computer*, pp. 1–19, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s00371-021-02119-0

[38] R. S. C. H. Zhang, "Review on one-stage object detection based on deep learning." [Online]. Available: https://www.researchgate.net/publication/361200268_Review_on_One-Stage_Object_Detection_Based_on_Deep_Learning

[39] M. Li, Z. Zhang, X. Wang, X. Guo, and L. Lei, "Agricultural greenhouses detection in high-resolution satellite images based on convolutional neural networks: Comparison of faster r-cnn, yolov3, and ssd," *Sensors*, vol. 20, no. 17, p. 4938, 2020.

[40] K. Zhao and X. Ren, "Small aircraft detection in remote sensing images based on yolov3," in *IOP Conference Series: Materials Science and Engineering*, vol. 525, no. 2. IOP Publishing, 2019, p. 022053.

[41] M. G. Dorrer and A. E. Tolmacheva, "Comparison of the yolov3 and mask r-cnn architectures' efficiency in the smart refrigerator's computer vision," *Journal of Physics: Conference Series*, vol. 1679, no. 4, p. 042022, 2020.

[42] A. D. M. R. Prusty, V. Tripathi, "A novel data augmentation approach for mask detection using deep transfer learning." [Online]. Available: https://www.researchgate.net/publication/352648228_A_novel_data_augmentation_approach_for_mask_detection_using_deep_transfer_learning/figures?lo=1

[43] J. Du, "Understanding of object detection based on cnn family and yolo," *IOP Conf. Series: Journal of Physics: Conf. Series*, vol. 1004, no. 1, p. 012029, 2018. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1004/1/012029/pdf

[44] C. Wang, H. M. Liao, I. Yeh, Y. Wu, P. Chen, and J. Hsieh, "Cspnet: A new backbone that can enhance learning capability of cnn," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3818–3830, 2020. [Online]. Available: https://arxiv.org/pdf/1911.11929.pdf

[45] "Roboflow App." [Online]. Available: https://app.roboflow.com

[46] "YOLOV3 Pretrained Model." [Online]. Available: https://pjreddie.com/media/files/darknet53.conv.74

[47] "YOLOV5 Pretrained Model." [Online]. Available: https://github.com/ultralytics/yolov5/releases/download/v7.0/yolov5s.pt

[48] "MobileNet-SSDV1, MobileNet-SSDV2 Pretrained Models." [Online]. Available: https://github.com/qfgaohao/pytorch-ssd?fbclid=IwAR19eQtHONVh3qkVeBLbZpNbPkASBDhq89yKRRyh_toTlm99AVnTst4fu_0

[49] T. Wong, "Linear approximation of f-measure for the performance evaluation of classification algorithms on imbalanced data sets," *IEEE Transactions on Knowledge and Data Engineering:1-12*. [Online]. Available: https://doi.org/10.1109/TKDE.2020.2986749

Shahadat Hossain Shahadat Hossain received his BSc in Computer Science degrees from IUBAT University. He is currently working as a research associate at the AISIP lab, where he is developing deep learning models for various new intelligent applications.
ORCID iD: https://orcid.org/0009-0001-3998-9982

Humaira Anzum Humaira Anzum is a computer science and engineering (CSE) undergraduate student at AUST in Bangladesh. She is currently a research assistant at the AISIP lab, where she is working on the Jetson Nano device to port machine learning models. Data mining and big-data applications are among her research interests.

Dr. Shamim Akhter Dr. Shamim Akhter is a distinguished computer scientist and academic currently affiliated with the Department of Computer Science and Engineering at Ahsanullah University of Science and Technology (AUST) as Professor. He earned his undergraduate degree from the American International University of Bangladesh in 2001 and completed his master's at the Asian Institute of Technology, Thailand, in 2005. His academic journey culminated in a Ph.D. in information processing from the Tokyo Institute of Technology, Japan, in 2009. Throughout his career, Dr. Akhter has held various roles, including Research Associate, Assistant/Associate Professor, and leadership positions at institutions such as the American International University of Bangladesh and Bangladesh Army University of Science and Technology. His international experience includes a JSPS postdoctoral fellowship in Japan and a role as a contact faculty at Thompson Rivers University, Canada. As a prolific author, Dr. Akhter has contributed significantly to the field, writing a book and publishing over 60 articles. His research has garnered support from esteemed organizations, and he holds the distinction of being an IEEE senior member actively involved in conferences in various capacities. Dr. Akhter's research interests are diverse, covering AI and machine intelligence tools in the CSP domain, data science strategies for business models, IoT and embedded-based solutions for IR4 constraints, evolutionary algorithms focusing on communication and distribution issues, parallel or distributed processing, data distribution, and optimization. His multifaceted contributions underscore a dedication to advancing computer science and technology, emphasizing international collaboration and knowledge dissemination.

ORCID iD: https://orcid.org/0000-0003-1408-9133