*ESP Journal of Engineering & Technology Advancements (ESP-JETA)*
*ISSN : 2583-2646*
*Special Issue*

# Ai Based Home Security System with Recognition

Ms. V. Merlin Freeda
Assistant professor Of CSE
Sree Sowdambika College Of Engineering
Metlinfreeda96@gmail.com
C. Indra U. Sakthi Meena
Final Year CSE Department Final Year CSE Department
Sree Sowdambika College Of Engineering Sree Sowdambika College Of Engineering
Smeriindra0913@gmail.com usakthimeena@gmail.com

**ABSTARCT:**

**In this paper, a holistic solution for Smart Home Security is implemented which helps in improving privacy and security using two independent and emerging technologies of facial authentication and speech recognition. With the help of the proposed application, the user will be able to monitor his home through his mobile phone/tablet/PC. This method involves facial recognition by taking a real-time feed of the person at the door and then analysis of the live feed is carried out where the face recognised is authenticated with the data of owners in the database which matches the face to a name. Speech recognition has been used to doubly check the output of facial authentication. The entire process is carried out with the help of neural networks. If there's an unauthorized person at the door, an alert will be triggered and the owner will get a notification of this unauthorized access and would get to choose whether they want to add the person to their database or not. The overall accuracy of the proposed model is 82.71% with an accuracy of 87.5% for Facial Authentication and 84.62 % for Speaker Authentication. Along with this, the main novelty for the research is to identify faces through masks which will help to properly verify the identity of the person and would prove to be beneficial not only in the current COVID-19 scenario but also in cases of thefts and burglaries by alerting the owner about the anomaly. Thus this smart security system can be extended to applications like banks, malls, offices, etc., and shall not be limited to only homes.**

## 1. Introduction

Crimes such as burglary and theft are serious concerns for any household. To be free from constant worrying especially at night, people can install Smart Home Security systems which will be accessible to them through a single device. Usually, people have a camera built-in on all entrances of their homes for security and are able to see the person visiting their home. Some advanced Home Security systems allow face recognition as well. Facial recognition can be carried out by a process similar to a method which involves a combination of geometrical feature points and low-level visual features (Klobas et al., 2019). But this can be surpassed by criminals by showing the image of the owner or a member of the household.

The invention involves a Home Security method that will not only authenticate using face authentication but also using speech recognition. This will authenticate every person visiting the house twice through a live feed and will compare the input with the names stored in the owner's database. Speech recognition is being included so that no one will be able to just show a picture of one of the identified members and enter, instead they will also need to speak a particular passphrase which will run through the speech recognition system to authenticate the person twice (Arif et al., 2020). The existing members will easily be authenticated and allowed to enter the house by giving notification to the owner's device. When there is an unidentified person at any of the entrances, an alert will be sent to the owner along with their picture. If the user wants to add the person to their database, they can simply add them with a click of a button, or just allow them to enter once without adding them to the database thereby not giving them lifelong permission to enter.

Since the dawn of the COVID-19 pandemic, whenever people meet up, they must wear a mask. The mask has made it difficult for facial recognition systems to recognise the person behind the mask. The research incorporates the study of how the facial recognition system can recognise the person even if they are wearing a mask. The study includes various methods possible for facial recognition through minimum feature extraction and specially focusing on recognising through the eye-region of the face. This research will not only help during COVID-19, but will also help if there is a burglar or thief wearing a mask. The recognition system can be run by authorities to identify a criminal.

The research includes finding an accurate model for speaker recognition, an accurate artificial neural network for face recognition, and combining the two to form an efficient home security system. The whole

system will be able to work through a single application by using the Internet of Things. It will consist of a network connected to the application on a personal device assistant such as a mobile or tablet, cameras and voice capturing devices.

Even though there are many ways to implement face authentication, artificial neural network is chosen because it has the ability to learn and model non-linear and complex relationships which are needed when it comes to pictorial inputs. Artificial neural network converts the input image into a vector and is then mathematically designated by a notation. Not only that, but the artificial neural network also gives high accuracy which is very important when it comes to authentication for security. It can prove to be very useful for multimodal facial biometrics recognition (Jose and Malekian, 2017).

Security and privacy are important for every single individual. Having a local database of identified people helps with both security and privacy. The data in the security system will travel on the local network itself keeping the users" data and them safe.

A Smart Home Security system will not only make the housemates feel safer at their own home but also make them technologically stronger. This idea can be extended as a security system for banks, malls, and other places requiring security.

## 2. Literature survey

Smart home devices account for a large portion of the consumer IoT market, but they pose security risks. Nothing is understood about how homeowners" views of security risk affect their decisions to use smart home technology. (Seng et al., 2021) evaluated a new model of how perceived security risk influences intention to use smart devices. Another method is presented by (Mokhayeri and Granger, 2019). In presenting their findings, they have described a smart home safe framework based on a refined version of the blockchain called Consortium blockchain. (Yin and Liu, 2017) carried out the classification of natural access points in the home as primary and secondary access points according to their use. Logic-based sensing is made by pointing to the normal user performance of these access points and requesting user verification wherever necessary. The gaps, which can be seen here, would be to improve the user behaviour prediction which can be done by analysing different user actions at home to make smart home security better.

Facial recognition technology is being used in both the private and public sectors for a variety of uses, ranging from physical security to customised shopping experiences. However, it is unclear how consumers interpret this new technology in terms of utility, danger, and comfort. (Yang et al., 2017) address these questions. Deep Siamese networks have recently been used for pair-wise face matching to increase robustness to intra-class variations. Although these networks can increase state-of-the-art accuracy, the lack of prior information from the target domain necessitates the collection of a large number of images to account for all potential capture conditions, which is impractical in many real-world surveillance applications (Kim and Ro, 2018). Previous research presented by (Banerjee and Yu, 2020) proposed a multi-task convolutional neural network (CNN) for face recognition, with identity classification as the primary task and pose, illumination, and expression (PIE) estimations as side tasks. A dynamic-weighting scheme was also devised for automatically assigning loss weights to each side mission, solving MTL" s critical task balancing issue. A weighted mixture deep neural network (WMDNN) is proposed to automatically extract features that work with FER functions. Many pre-processing methods, such as face detection, rotation adjustment, and data addition, are used to limit FER regions (Castiglione et al., 2020). (Zhou et al., 2021) proposed a comprehensive network framework for capturing identity information from facial strengths and their relationships. In another proposed method, facial expressions from the smile were analysed and used for facial verification. The 3D authentication feature presented by (Ntalianis and Tsapatsoulis, 2015) is as important to the user as security and provides an easy way to authenticate the right user. (Tarannum et al., 2020) proposed a method that uses a powerful appearance and time-dependent local features that express a person's face during a speech in relation to its temporary and temporary elements. (Natheem et al., 2013) proposed a user-friendly authentication system for the EchoPrint novel, which incorporates the acoustics and concept of secure and easy -to-use authentication, without requiring any special hardware. Remote authentication includes the transmission of encrypted information, as well as visual and audio signals (photos/videos, personal voice, etc.). (He and Dong, 2020) suggested a strong authentication method based on the semantic phase, chaotic encryption, and data encryption. A new multi-user-based framework had been developed and used for large types of image data (Rahmani et al., 2018). In this framework, various biometric features such as IRIS, facial and finger features are used to determine the user's unique data validity and security process. (Dey et al., 2017) designed a model to use Fourier Optics and Neural Networks which uses an advanced optical Fourier plane correlator (real-time) for face recognition and feature extraction.

In recent years, real-time speech recognition technology has been commonly used in the fields of intelligent voice toys, industrial control, and intelligent rehabilitation as a primary cross-technology in the field of artificial intelligence. Since real-time speech recognition based on embedded technology has obvious advantages in terms of device scale, power consumption, and R&D costs, it has become a hot carrier for achieving efficient speech recognition technology. (Zeinali et al., 2017) build a basic framework of machine learning based on Markov random field theory combined with machine learning theory, and research the algorithm of real-time speech vocabulary matching recognition based on this framework in order to realise a simple and functional real-time speech recognition system based on embedded systems. Another common method is using artificial neural networks. Some carefully built deep autoencoders are proposed by (Liu et al., 2018) to generate effective bimodal features from audio and visual stream inputs. Speaker recognition can be divided into text-dependent speaker verification and text-independent speaker verification. (Stafylakis et al., 2016) show that deep neural networks based systems have significantly outperformed GMM for text-dependent speaker verification. (Zeinali et al., 2017) and (Chai et al., 2020) have recently demonstrated that similar ideas can be applied to the text-dependent speaker verification mission, inspired by the success of Deep Neural Networks (DNN) in text-independent speaker recognition. They discussed new developments in their state-of-the-art i-vector-based approach to text-dependent speaker verification, which also employs various DNN techniques. (Meng et al., 2018) proposed a novel model to improve the recognition accuracy of the short utterance speaker recognition system in this paper. On the other hand, (Asaei et al., 2017) investigate the use of joint factor analysis (JFA) for text-dependent speaker recognition with random digit strings. A unique method of using facial expressions with voice is used by (Din et al., 2020). A new data rating is designed by (Royer et al., 2018) to detect deviations from optimal speech quality. In (Chen and Sang, 2018), a new cross-entropy-guided measure is proposed to indirectly evaluate the details of automatic speech recognition for discounted speech with speech enhancements before and without performing ASR tests directly.

(Ma et al., 2019) removed masked objects from facial images. This problem is challenging because a facial mask often covers a large area of the face that extends beyond the lower border of the chin, and mask and without mask facial pairs do not exist for training. Findings of (Lyamin and Cherepovskaya, 2016) at the community level corroborated previous findings showing the significance of the eye area for face recognition. They also showed that, from the observer's perspective, face processing capacity is linked to a systematic increase in the use of the eye region, especially the left eye. (Chamikara et al., 2020) proposed a face-mask recognition approach based on the Gaussian Mixture Model for fraud prevention. In comparison to other conventional face recognition approaches, their methodology has been designed to improve the ability to identify abnormal faces such as sunglasses, masks, and respirators, as well as reduce the risk of these rare faces in the se-

curity sector. (Nautsch et al., 2019) proposed eye movements and authorization based on iris recognition (EmIr-Auth), a proven authentication system for biometrics-based novel operators. The idea proposed by (Vasanthi and Seetharaman, 2020) related to the biometric diagnostic method and the problem of providing accurate diagnostic results using an eye-frequency eye tracker. (Mahmood, 2019) proposed a privacy-preserving technique for "controlled information release" in which they mask an original face image and prevent biometric feature leakage while identifying an individual.

## 3. Problem statement

The research done on Smart Home security using facial authentication and speaker recognition using artificial neural networks is limited in comparison to the research done on each of them individually. This restraint can be seen because of the challenges faced by the researchers while performing such tasks independently. When it comes to Face Authentication, researchers prefer to use Support-Vector Machines (SVM) or Convolutional Neural Networks (CNN). A gap that can be seen here is that when the number of features is greater than the number of samples, SVM does not perform well. Even though a standard CNN is considered to be better than SVM when it comes to image inputs since CNN detects important features automatically without any human supervision, it may falter when the dataset is too small. The problem that may occur when it comes to Smart Home Security is that the dataset would be too small as there would be a limited number of images for each individual and the CNN must be able to detect the required features from the dataset available. To tackle this problem, Siamese neural networks help as they can work with a limited dataset using the one-shot learning approach. One important gap that can be seen while identifying an individual using face authentication is that if the individual is wearing a mask, the system is unable to identify them as most of their facial features are hidden. This creates a problem for individuals trying to enter a home where face authentication is required, especially during the COVID-19 pandemic, as the system will not accept them. To tackle this problem, the system can focus on the eye region and run the biometrics-based on the individual" s iris using modified neural network models. On the other hand, when it comes to speaker recognition, most systems use text-independent methods using Gaussian Mixture Model or i-vectors. Even though these methods are effective, these can be surpassed by an unidentified individual if they have a voice recording of an identified individual. This can be solved by using a specified passphrase which must be spoken by the individual for recognition. This will help in dual authentication as the system will first use speech recognition to verify the phrase and then use voice authentication to identify the individual.

The main target of the proposed idea would be to fill the above gaps to provide better home security by contributing the below works. The following aims are to be kept in mind for building a smart Solution using facial authentication and speaker recognition through artificial neural networks for Home Security.

- To propose an effective facial authentication system using Siamese neural networks.
- To intend GMM (Gaussian Mixture Model) to train on extracted Mel Frequency Cepstral Coefficients (MFCCs) features from an audio wav file for speech authentication and speaker recognition.
- To perform masked facial recognition through minimum feature extraction by focussing more on the eye region of the face.

The above efforts would be taken to achieve a high classification and accuracy for facial authentication, masked facial recognition and speaker recognition to provide a better home security solution.

## 4. Research framework

### 4.1. Overall architecture

As shown in the proposed model in Figure 1, an image of the individual will be captured at the entrance using a camera placed at the door. The captured image will be sent to the face recognition and authentication system. The system will first detect whether the person is wearing a mask or not. Once that is detected, the image is sent to „Face Recognition and Authentication System" if the person is not wearing a mask or to „Masked-Face Recognition and Authentication System" if the person is wearing a mask. If the face recognition system does not recognise the individual, then the system doesn't move on to the speaker recognition system, instead, it breaks out of the system and immediately sends a notification to the owner along with the individual" s image. On the other hand, if the system recognises the individual, then the individual is asked to speak the passphrase which is taken as an input for the speaker recognition system. Once the user speaks the passphrase, the speaker recognition system works on text-dependent voice authentication. The System will use a speech recognition model to identify the speaker and it compares the spoken phrase with the security system passphrase. If the speaker isn't identified or if the phrase spoken does not match the passphrase, then the system breaks and sends an alert to the owner along with the individual" s image. On the other hand, if the individual is identified as an existing speaker and says the correct passphrase, the individual's name is compared with the name recognised using the facial authentication system. If both the systems have identified the same person, then the individual is allowed to enter into the house by sending a notification to the owner. Else, if both the systems identify the individual as two different persons, then also an alert is sent to the owner along with the individual" s image.

### 4.2. Proposed methodology

The entire project can be divided into three components – Face Recognition and Authentication System, Masked-Face Recognition and Authentication System, and Speaker Recognition and Authentication System.

#### 4.2.1. Face recognition and authentication system

The system first captures the image of the individual at the entrance. If the person is not wearing a mask, then the captured image is sent to the „Face Recognition and Authentication' system which uses Siamese neural network and one-shot learning using FaceNet model. It can be seen that the dataset for a household will have limited images or maybe even only one image per person. Siamese neural network works most efficiently in such a case as it distinguishes two different classes by measuring their similarity instead of detecting characters of a class. It can tell whether a pair of pictures belong to two separate classes or to the same class by studying pairs of pictures. This authentication system is implemented by feeding a pair of images: the captured image and an image from the dataset, to the same Siamese neural network. Their features are given as the two outputs. Then the distance of the two outputs is calculated which means comparing the features. The calculated distance indicates the similarity between the two images. The similarity scores of the same class pairs are low, while the similarity scores of different class pairs are high. In addition to the Siamese neural network, FaceNet model is used for faster computation. It computes two pairs simultaneously by calculating the distance between the captured image with one image in the dataset and the captured image with another image in the dataset. The two distances calculated show which image is similar to the captured image. Keeping this distance as a base, the system compares all captured images with other images in the datasets. When the similarity score is lower for another pair, it is taken as the base for further computation. If the system provides a recognised name

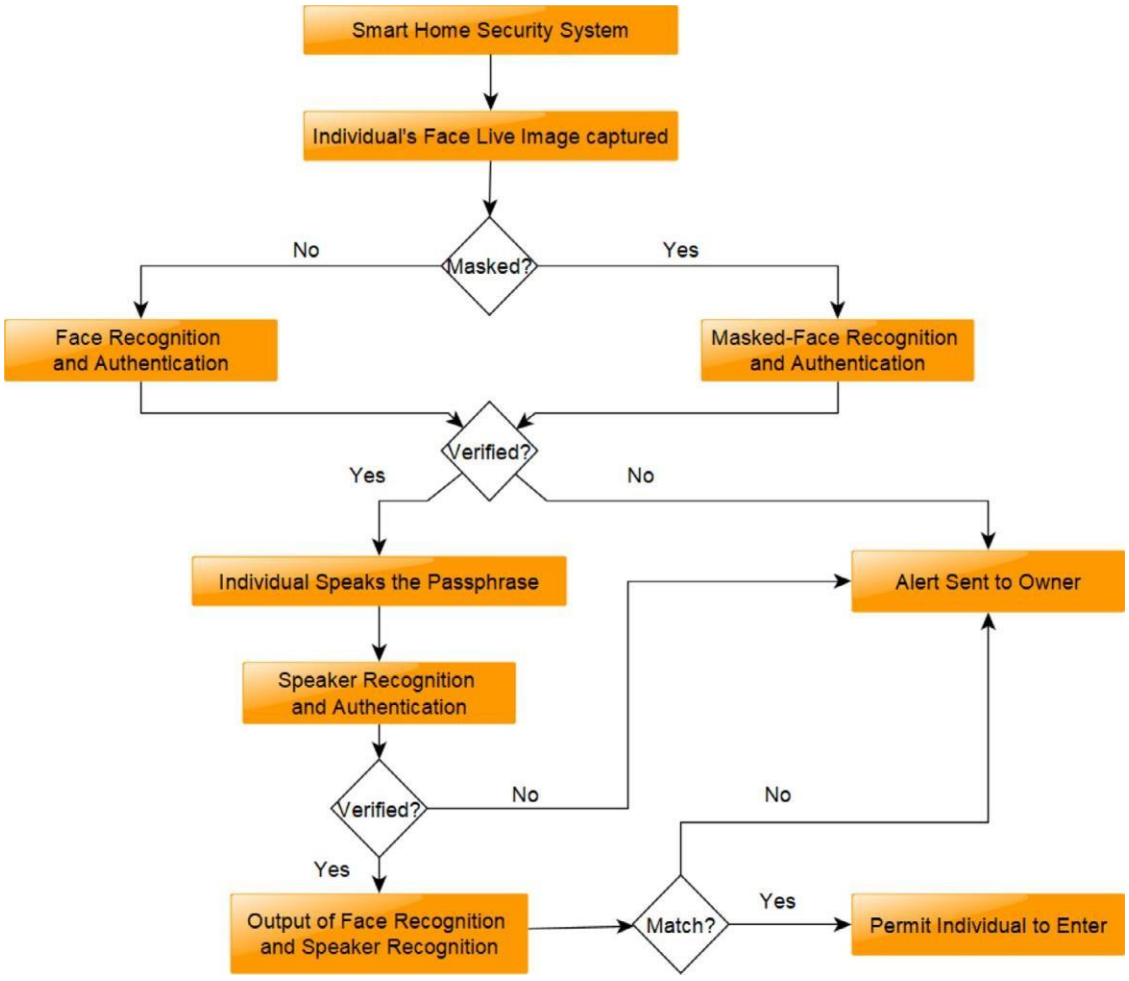


**Fig. 1.** Overall system architecture.

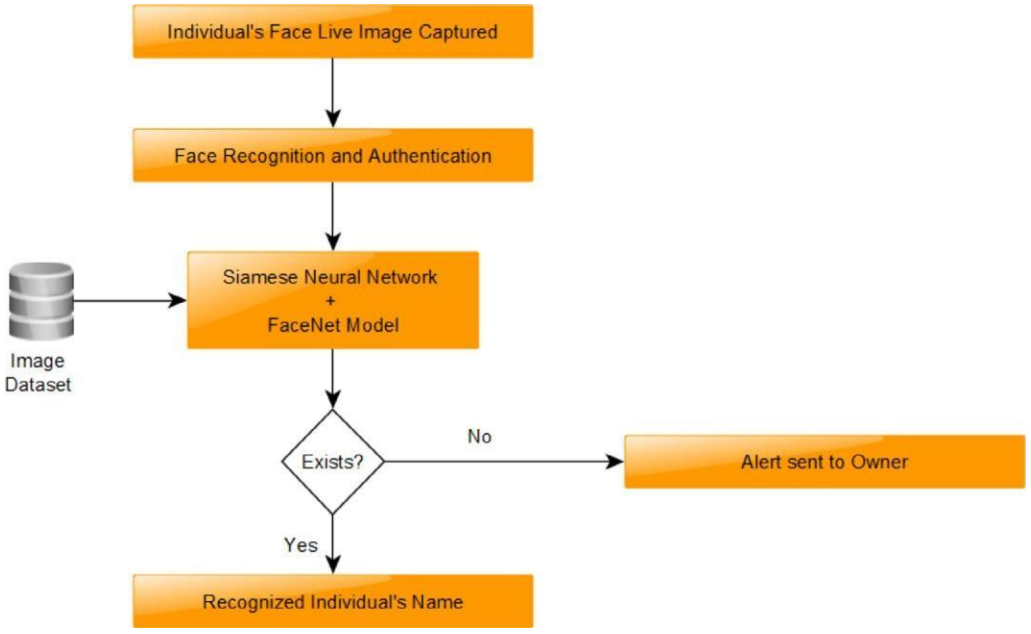


**Fig. 2.** Face recognition and authentication system architecture.

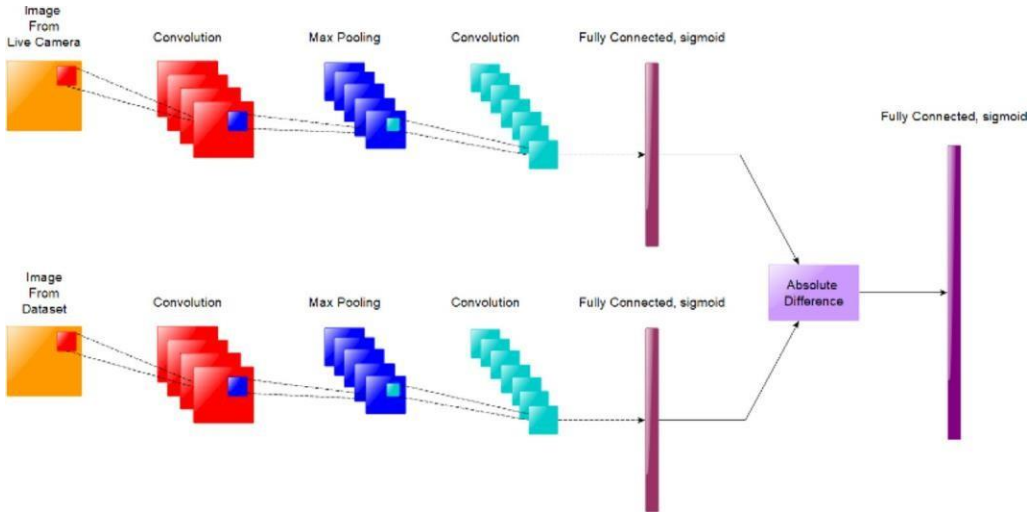


**Fig. 3.** Siamese neural network architecture.

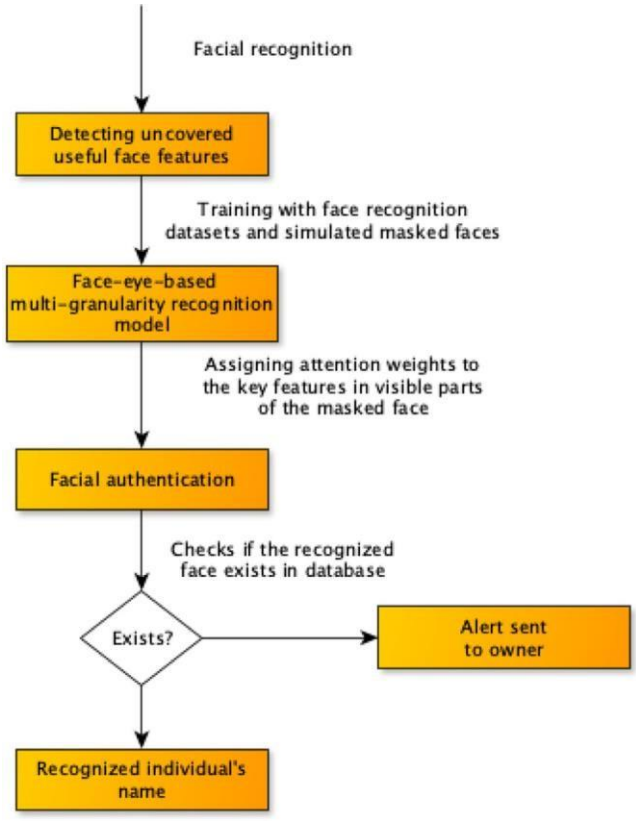


**Fig. 4.** Masked-face recognition and authentication system architecture.

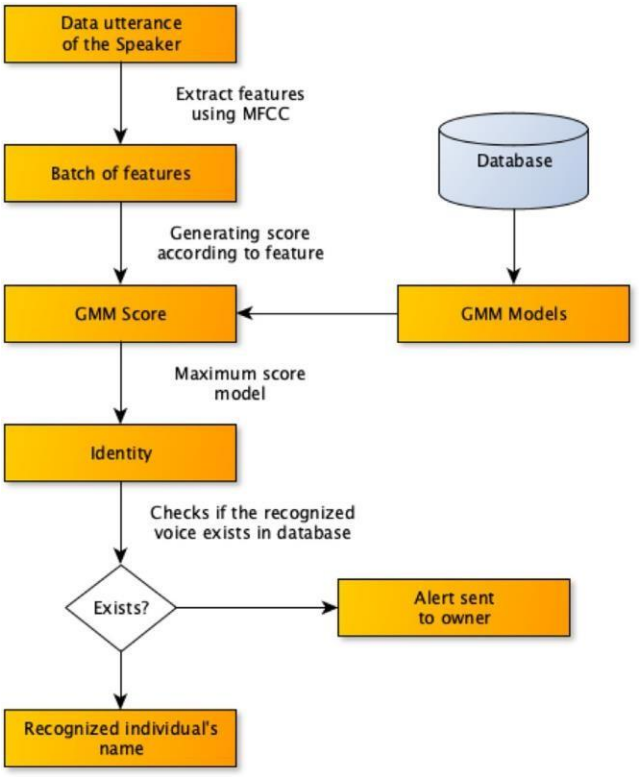


**Fig. 5.** Speaker recognition and authentication system architecture.

for the individual, it moves on to the speaker authentication. If not, then an alert is sent to the owner.

*4.2.2. Masked face recognition and authentication system*

In the captured real-time image, if the person is wearing a mask, traditional facial authentication systems would not work and would give less accuracy. In these cases, frontal face images are extracted at a high quality so that the task of identifying the concealed face is no longer so difficult. Although the mask covers most parts of the face, features on the upper part, such as the eye and eyebrows, can still be used to improve the availability of a face recognition system. The basic premise would be to remove the distortion of the mask and give priority to the exposed face features which are useful. The proposed masked facial authentication approach will be on 2 main aspects. One is a built-in database in which both masked and unmasked faces have been included, and the other is the appropriate use of uncovered facial features which are useful. Various attention weights have been applied to the important aspects of the visible facial features such as nose, right eye, left eye, right eyebrow and left eyebrow, etc., which effectively address the problem of unequal distribution of discriminatory facial information. If the identified face is registered in the database, then it moves on to the speaker authentication. Else, if the identified face is not registered in the database, an alert would be sent to the owner.

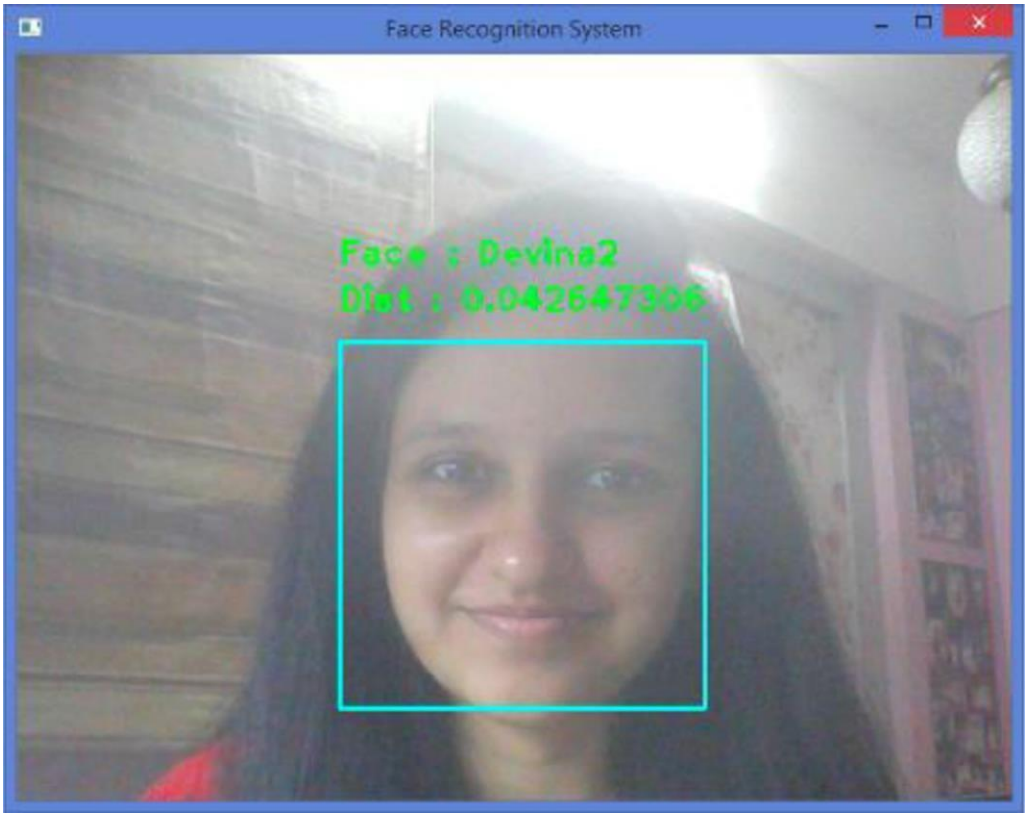


**Fig. 6.** Facial authentication of unmasked faces.

*4.2.3. Speaker recognition and authentication system*

Once the face is authenticated, the system moves to speaker authentication. For adding a new user, they must record 3 audio clips of 2 seconds each. The audios are converted to .wav files. Their features are extracted using Mel Frequency Cepstral Coefficients (MFCC). A Gaussian Mixture Model (GMM) is created based on the features from all three audios and stored in the database. For authentication, the user speaks for 2 seconds. The audio is converted to a test.wav file. Its features are extracted using MFCC. The extracted features are put in all the GMM models in the database and their scores are calculated. The model with the maximum score is chosen as the identified speaker. If the user is identified, it will be compared with the identity of the user recognised in facial authentication. If the two identities match, the individual will be allowed to enter the house. If they do not match then an alert will be sent to the owner. If the user does not exist in the database, the system will output unknown and an alert will be sent to the owner. They will have an option to add the individual to the database.

## 5. Results and discussions

### *5.1. Face recognition and authentication*

Dataset: The training dataset for the facial authentication model consists of 50 images each of 8 people. The testing dataset includes both masked and unmasked 8 images of 3 people.

Input: Input has been taken from the live camera feed wherein a masked/unmasked person stands close to the camera. The facial features are extracted from the live input and fed into the FaceNet model. In the case of masked facial recognition, the features from the upper region of the face like eyes and eyebrows are extracted and taken as input for the FaceNet model for authentication.

Output: If the user is registered, the model gives the output as his corresponding name. If the user is not registered in the database, the model gives an output of "Not Found".

Figure 8 shows the confusion matrix of the facial recognition and authentication model. It shows that for User 0 and User 2 the model gives a true positive value (i.e. predicted the actual user itself). However, for User 1, the model gave 1 true positive and 1 false negative value (i.e. predicted another user instead of the actual user).

Figure 9 shows the classification report for the proposed model. The individual precision, recall and f1-score for each user are shown along with the overall values. For testing - 4 images of user „Devina', 2 images of user „Emilia Clarke' and 2 images of user „Gauri' were used. According to the weighted average, the total precision for the proposed model is 0.92, recall is 0.88 and f1-score is 0.87. The facial recognition and authentication model reported an accuracy of 87.5%.

Table I shows the comparison between state-of-the-art methods and the proposed model based on one short learning. It can be seen that usually classification requires a large number of images of each class for training. Also, if users need to test the model on another class apart from the classes given in training, they cannot expect an accurate result. If a new class has to be added to the training dataset, a lot of images of the class are required and the model has to be re-trained again. This poses a problem when the number of classes is dynamically changing which in turn increases the computation and training costs. However, one-shot classification in the proposed model only requires one training example for each class. This network doesn't learn to classify an image to any class, instead, it uses a similarity function to check how similar the two input images are.

### *5.2. Speaker recognition and authentication system*

Dataset: Since the database for Smart Home Security will have a limited dataset, the prototype contains the voice recordings of 9 speakers. Each speaker records 3 voice recordings of 2 seconds each for training. While for testing, the speaker speaks for 2 seconds and then the prediction is done by comparing it with the existing GMM models.

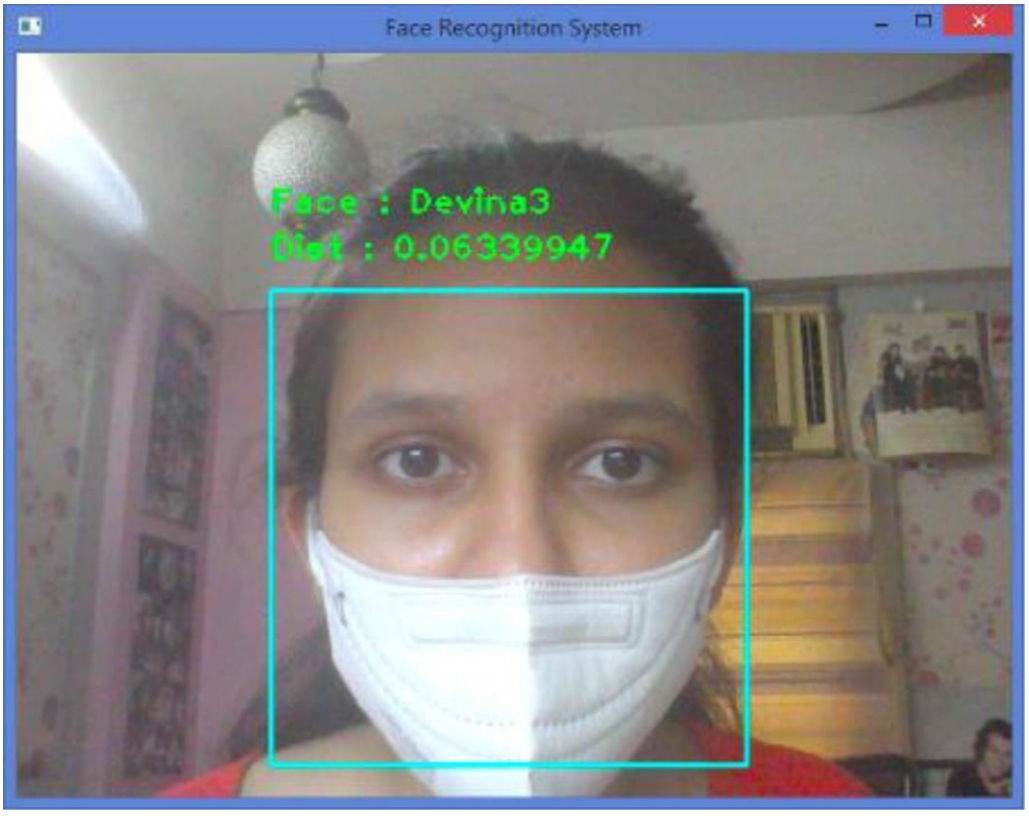


**Fig. 7.** Facial authentication of masked faces.

**Table I**
Comparison with existing models for facial authentication.

| Model | Methodology | Training Dataset | Accuracy ± Std(%) |
|---|---|---|---|
| DeepID2 (Sun et al., 2014) | Using deep learning and both face identification and verification signals have been used for supervision. The face identification process extracts DeepID2 features from different identities by using inter-personal variations, and the face verification task takes the DeepID2 features extracted from the same identity and reduces intra-personal variations. Used Joint-Bayes method. | 202599 images of 10177 subjects, private on LFW dataset | 95.43 |
| DeepFace (Taigman et al., 2014) | Utilized 3D face modelling to apply a piecewise affine transformation using a nine-layer deep neural network which involves using about 120 million parameters. This method involves using locally connected layers without weight sharing, instead of the standard convolutional layers. | 4.4M images of 4030 subjects, private on LFW dataset | 95.92 ± 0.29 |
| FaceNet (Schroff et al., 2015) | Uses a deep convolutional network trained to directly learn a mapping from face images to a small Euclidean space where distances correspond to a measure of face similarity by using only 128 bytes per face. | 260M images of 8M subjects, private on LFW dataset | 99.63 ± 0.09 |
| Sparse Representation-based Classification (Wright et al., 2008) | Using sparse representation computed by l-minimization, a general classification algorithm for object recognition has been devised which handles two issues in face recognition: feature extraction and robustness to occlusion. | 48 video sequences and 64204 face images on Chokepoint database | 52.4±0.32 (pAUC) 47.5±0.031 (AUPR) |
| **Proposed model** | **Fed a pair of images - the captured image and an image from the dataset, to the same Siamese neural network. Along with this, the FaceNet model computes two pairs simultaneously by calculating the distance between the captured image with one image in the dataset and the captured image with another image in the dataset.** | **Custom, private dataset (50 images of 8 users for training, 8 images for testing)** | **87.5** |

Input: Speaker speaks for 2 seconds which is recorded and converted into a test.wav file.

Output: This .wav file is compared with the existing Gaussian Mixture Models and the one with the maximum score is chosen as the identity which is given as the output.

Figure 10 shows the Confusion matrix for predicting the speaker. It is an 8 × 8 matrix. The confusion matrix compares the actual target values with those predicted by the proposed model. It can be seen that for Speakers 0,1,2,3,4,5, and 6, the model was able to predict the true positive values (i.e. predicted the actual speaker itself). However, for Speaker 7, the model gave only 1 true positive and 2 false negative values (i.e. predicted another speaker instead of the actual speaker).

Figure 11 shows the classification report for the proposed model. The individual precision, recall and f1-score for each actual speaker are shown along with the overall values. For calculating the values – Devina, Navya, Speaker27,Speaker28, Speaker29, Speaker30, Speaker34 and unknown had 1,1,1,2,12,2,3 testing cases respectively. The total precision for the model is 0.91, recall is 0.85 and f1-score is 0.83. The accuracy of the proposed model is 84.62%.

**Table II**
Comparison with existing models for speaker authentication.

| Model | Methodology | Training Dataset | Metrics |
|---|---|---|---|
| GMM and CNN Hybrid (Liu et al., 2018) | It achieves successful training and identification for a limited number of biometric samples by training the pre-processed speech spectrum in deep networks to extract the deep features of the full frequency spectrum of short utterances. | Speed spectrum images of the 50 individual speech samples of people, Each sample contains 10 speech spectrum pictures. | Equal Error Rate = 2.5% For 5000 iterations, Accuracy = 87% |
| Joint Factor Analysis (Stafylakis et al., 2016) | Testing the model based on many systems and their fusion. It was found that the best EER was possible when all 6 systems were fused together. | RSR2015 (part III) dataset | Male = 2.01% and Female = 3.19% Equal Error Rates |
| HMM-Based i-Vector Extractor (Zeinali et al., 2017) | Preconditioned i-vectors with a regularized version of within-class covariance normalization, which can be robustly calculated phrase-dependently on the minimal datasets available for the text-dependent task. | RSR2015 (143 female and 157 male speakers, includes more than 60 different utterances spoken by all speakers) | Equal Error Rate Male = 1.11 |
| **Proposed model** | **Implemented using MFCC (Mel Frequency Cepstral Coefficient) feature extraction which passes its output to the GMM (Gaussian Mixture Model). The prediction is based on the maximum score calculated with respect to the model.** | **Custom, private dataset (9 speakers for training, 13 speakers for testing)** | **Accuracy = 84.62%** |

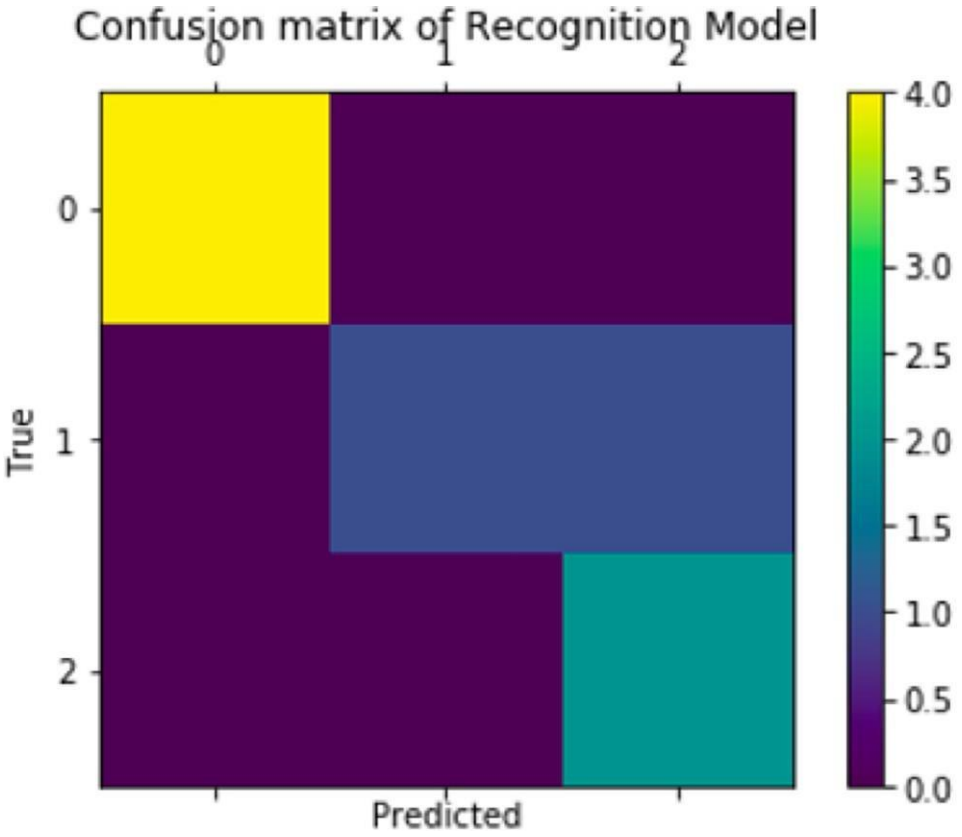


**Fig. 8.** Confusion matrix for facial recognition and authentication.

Table II shows the comparison of the proposed model with existing models for speaker authentication. It can be seen that even with a larger dataset than the one used in the proposed model, the accuracy of GMM and CNN is 87%. The importance of testing the proposed model with a smaller dataset is that a household has three to four members on average. So the system should be able to train and test the model based on the minimum availability of sound recordings. For this purpose, an accuracy of 84.62% of the proposed model for a small dataset works well.

It can be seen that the proposed model is 87.5% accurate for facial recognition and authentication and 84.62% accurate for speaker authentication. Once the system recognises the user through facial authentication, the system moves to speaker recognition. If the output of both the systems match, then the authentication is successful, or else the system denies access to the user. Holistically combining the facial and speaker authentication systems, an overall accuracy of 82.71% is achieved. The existing models have been trained and tested on large datasets, but for Smart Home Security, the database will include limited datasets. For small datasets as per the requirement for the Home Security System, the existing models might have lesser accuracy as those methods require multiple training data for each user. On the other hand, the proposed model gives a higher accuracy even with smaller datasets with less training data. Hence it can be seen that the proposed model is more efficient for a Smart Home Security Solution.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Devina | 1.00 | 1.00 | 1.00 | 4 |
| Emilia Clarke | 1.00 | 0.50 | 0.67 | 2 |
| Gauri | 0.67 | 1.00 | 0.80 | 2 |
| accuracy | | | 0.88 | 8 |
| macro avg | 0.89 | 0.83 | 0.82 | 8 |
| weighted avg | 0.92 | 0.88 | 0.87 | 8 |

Accuracy: 87.5%

**Fig. 9.** Classification report for facial recognition and authentication.

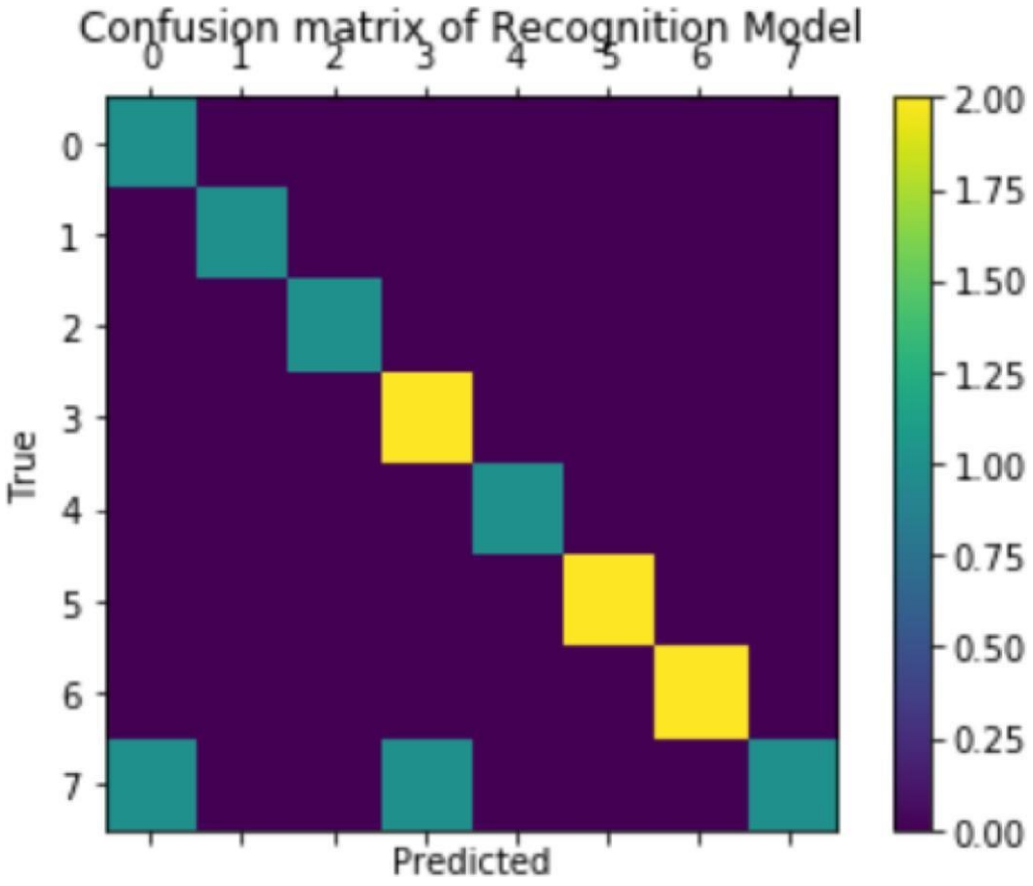


**Fig. 10.** Confusion matrix for speaker authentication.

## 6. Conclusion

This paper presents a solution for Smart Home Security. Models for facial and speaker recognition have been proposed for user authentication. Siamese neural network with FaceNet based on one-shot learning is used for facial authentication and Gaussian Mixture Model with MFCC feature extraction is used for speaker authentication. Pre-processing is done for the captured image and the audio of the user. Based on the features extracted, the minimum distance for facial recognition and the maximum score for speaker recognition is taken. Using these parameters, the user is classified as either a member in the database or unidentified. For small datasets, the proposed models are more efficient than the state-of-the-art models which require larger datasets for training. Apart from this, the model not only recognises the identities of unmasked faces but also recognises masked faces. For a masked user, their eye and nose region should be clearly visible. The proposed model reports a final accuracy of 82.71% for the entire Home Security system. As a future scope, masked users could be recognised solely on the basis of their eye region. This Smart Security solution can also be extended to malls, offices and other places requiring security (Figs. 2, 3, 4, 5, 6, 7).

## Declaration of Competing Interest

None.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Devina | 0.50 | 1.00 | 0.67 | 1 |
| Navya | 1.00 | 1.00 | 1.00 | 1 |
| Speaker27 | 1.00 | 1.00 | 1.00 | 1 |
| Speaker28 | 0.67 | 1.00 | 0.80 | 2 |
| Speaker29 | 1.00 | 1.00 | 1.00 | 1 |
| Speaker30 | 1.00 | 1.00 | 1.00 | 2 |
| Speaker34 | 1.00 | 1.00 | 1.00 | 2 |
| unknown | 1.00 | 0.33 | 0.50 | 3 |
| avg / total | 0.91 | 0.85 | 0.83 | 13 |

Accuracy: 84.62%

**Fig. 11.** Classification report for speaker authentication.

## References


Arif, S., Khan, M. A., Rehman, S. U., Kabir, M. A., & Imran, M. (2020). Investigating smart home security: Is blockchain the answer? *IEEE Access, 8*, 117802–117816.

Asaei, A., Cernak, M., & Bourlard, H. (2017). Perceptual information loss due to impaired speech production. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(12), 2433–2443.

Banerjee, D., & Yu, K. (2020). 3D face authentication software test automation. *IEEE Access, 8*, 46546–46558.

Castiglione, A., Nappi, M., & Ricciardi, S. (2020). Trustworthy Method for Person Identification in IIoT Environments by Means of Facial Dynamics. *IEEE Transactions on Industrial Informatics, 17*(2), 766–774.

Chai, L., Du, J., Liu, Q. F., & Lee, C. H. (2020). A Cross-Entropy-Guided Measure (CEGM) for Assessing Speech Recognition Performance and Optimizing DNN-Based Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 106–117.

Chamikara, M. A. P., Bertok, P., Khalil, I., Liu, D., & Camtepe, S. (2020). Privacy preserving face recognition utilizing differential privacy. *Computers & Security, 97*, Article 101951.

Chen, Q., & Sang, L. (2018). Face-mask recognition for fraud prevention using Gaussian mixture model. *Journal of Visual Communication and Image Representation, 55*, 795–801.

Dey, S., Motlicek, P., Madikeri, S., & Ferras, M. (2017). Template-matching for text-dependent speaker verification. *Speech communication, 88*, 96–105.

Din, N. U., Javed, K., Bae, S., & Yi, J. (2020). A novel GAN-based network for unmasking of masked face. *IEEE Access, 8*, 44276–44287.

He, Y., & Dong, X. (2020). Real time speech recognition algorithm on embedded system based on continuous Markov model. *Microprocessors and Microsystems, 75*, Article 103058.

Jose, A. C., & Malekian, R. (2017). Improving smart home security: Integrating logical sensing into smart home. *IEEE Sensors Journal, 17*(13), 4269–4286.

Kim, S. T., & Ro, Y. M. (2018). Attended relation feature representation of facial dynamics for facial authentication. *IEEE Transactions on Information Forensics and Security, 14*(7), 1768–1778.

Klobas, J. E., McGill, T., & Wang, X. (2019). How perceived security risk affects intention to use smart home devices: A reasoned action explanation. *Computers & Security, 87*, Article 101571.

Liu, Z., Wu, Z., Li, T., Li, J., & Shen, C. (2018). GMM and CNN hybrid method for short utterance speaker recognition. *IEEE Transactions on Industrial informatics, 14*(7), 3244–3252.

Lyamin, A. V., & Cherepovskaya, E. N. (2016). An approach to biometric identification by using low-frequency eye tracker. *IEEE Transactions on Information Forensics and Security, 12*(4), 881–891.

Ma, Z., Yang, Y., Liu, X., Liu, Y., Ma, S., Ren, K., & Yao, C. (2019). EmIr-Auth: eye movement and iris-based portable remote authentication for smart grid. *IEEE Transactions on Industrial Informatics, 16*(10), 6597–6606.

Mahmood, A. (2019). A Solution to the Security Authentication Problem in Smart Houses Based on Speech. *Procedia Computer Science, 155*, 606–611.

Meng, Z., Han, S., Liu, P., & Tong, Y. (2018). Improving speech related facial action unit recognition by audiovisual information fusion. *IEEE transactions on cybernetics, 49*(9), 3293–3306.

Mokhayeri, F., & Granger, E. (2019). Video face recognition using siamese networks with block-sparsity matching. *IEEE Transactions on Biometrics, Behavior, and Identity Science, 2*(2), 133–144.

Natheem, M. S., Narayanan, R., & Nagaiyan, P. N. (2013). Advanced face recognition system using Fourier Optics and Neural Networks. In *2013 Tenth International Conference on Wireless and Optical Communications Networks (WOCN)* (pp. 1–3). IEEE.

Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., . . . Busch, C. (2019). Preserving privacy in speaker and speech characterisation. *Computer Speech & Language, 58*, 441–480.

Ntalianis, K., & Tsapatsoulis, N. (2015). Remote authentication via biometrics: a robust video-object steganographic mechanism over wireless networks. *IEEE Transactions on Emerging Topics in Computing, 4*(1), 156–174.

Rahmani, M. H., Almasganj, F., & Seyyedsalehi, S. A. (2018). Audio-visual feature fusion via deep neural networks for automatic speech recognition. *Digital Signal Processing, 82*, 54–63.

Royer, J., Blais, C., Charbonneau, I., Déry, K., Tardif, J., Duchaine, B., . . . Fiset, D. (2018). Greater reliance on the eye region predicts better face recognition ability. *Cognition, 181*, 12–20.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Seng, S., Al-Ameen, M. N., & Wright, M. (2021). A First Look into Users' Perceptions of Facial Recognition in the Physical World. *Computers & Security*, Article 102227.

Stafylakis, T., Alam, M. J., & Kenny, P. (2016). Text-dependent speaker recognition with random digit strings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(7), 1194–1203.

Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation by joint identification-verification. arXiv preprint arXiv:1406.4773.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701–1708).

Tarannum, A., Rahman, Z. U., Rao, L. K., Srinivasulu, T., & Lay-Ekuakille, A. (2020). An efficient multi-modal biometric sensing and authentication framework for distributed applications. *IEEE Sensors Journal, 20*(24), 15014–15025.

Vasanthi, M., & Seetharaman, K. (2020). *Facial image recognition for biometric authentication systems using a combination of geometrical feature points and low-level visual features.* Journal of King Saud University-Computer and Information Sciences.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2008). Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence, 31*(2), 210–227.

Yang, B., Cao, J., Ni, R., & Zhang, Y. (2017). Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access, 6*, 4630–4640.

Yin, X., & Liu, X. (2017). Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing, 27*(2), 964–975.

Zeinali, H., Sameti, H., & Burget, L. (2017). HMM-based phrase-independent i-vector

extractor for text-dependent speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(7), 1421–1435.

Zeinali, H., Sameti, H., & Burget, L. (2017). Text-dependent speaker verification based on i-vectors, neural networks and hidden Markov models. *Computer Speech & Language*, *46*, 53–71.

Zhou, B., Xie, Z., Zhang, Y., Lohokare, J., Gao, R., & Ye, F. (2021). Robust Human Face Authentication Leveraging Acoustic Sensing on Smartphones. *IEEE Transactions on Mobile Computing*.