*Electronic Letters on Computer Vision and Image Analysis 23(1):1-14, 2023*

# A Multimodal Biometric Authentication System: Using of Autoencoders and Siamese Networks for Enhanced Security

Théo GUEURET*, Leila KERKENI+

* *University Gustave Eiffel, France*
+ *Capgemini Engineering, France*

Received 12th of December, 2023; accepted 2nd of April 2024

**Abstract**

Ensuring secure and reliable identity verification is crucial, and biometric authentication plays a significant role in achieving this. However, relying on a single biometric trait, unimodal authentication, may have accuracy and attack vulnerability limitations. On the other hand, multimodal authentication, which combines multiple biometric traits, can enhance accuracy and security by leveraging their complementary strengths. In the literature, different biometric modalities, such as face, voice, fingerprint, and iris, have been studied and used extensively for user authentication.

Our research introduces a highly effective multimodal biometric authentication system with a deep learning approach. Our study focuses on two of the most user-friendly safety mechanisms: face and voice recognition. We employ a convolutional autoencoder for face images and an LSTM autoencoder for voice data to extract features. These features are then combined through concatenation to form a joint feature representation. A Siamese network carries out the final step of user identification. We evaluated our model's efficiency using the OMG-Emotion and RAVDESS datasets. We achieved an accuracy of 89.79% and 95% on RAVDESS and OMG-Emotion datasets, respectively. These results are obtained using a combination of face and voice modality.

*Key Words*: User Authentication, Multimodal Biometrics, Deep Learning, Siamese Neural Network, autoencoder, Face Recognition, Voice Recognition, Fusion.

## 1 Introduction

In today's ever-changing digital world, it is essential to have strong security measures for authentication. Traditional methods like passwords and PINs are no longer enough as they can be easily compromised. Hence, there is a pressing need for more reliable and efficient alternatives. Multimodal biometric technologies are emerging as promising solutions that utilize individuals' unique physical or behavioral traits. These technologies combine biometric modalities such as face, voice, fingerprint, and iris for reliable and precise identification.

Face and voice recognition have gained significant attention among various biometric modalities due to their widespread availability and non-intrusive nature [1]. However, each modality has its limitations and

Correspondence to: theo.gueuret@univ-lille.fr

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.1811
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

vulnerabilities. Facial recognition systems can be susceptible to spoofing attacks using photographs or videos, while voice recognition systems can be compromised through voice synthesis techniques. To address these challenges, researchers have turned to multimodal biometric systems that combine face and voice to enhance security and reliability [2] and [1].
In this paper, we introduce an innovative multimodal Siamese network that seamlessly integrates face and voice biometrics, setting a new benchmark for efficiency in authentication systems. The Siamese architecture also means that the model can recognize individuals without relying on prior examples, making it capable of identifying unseen individuals. Diverging from conventional approaches, our architecture leverages the complementary strengths of both modalities to deliver unparalleled accuracy and security. The dual-modality integration within the Siamese framework is what sets our solution apart from the state-of-the-art, providing a versatile platform poised for future innovation. Our choice of utilizing both face and voice modalities can be justified based on several factors. First, the complementary information provided by each modality enhances overall authentication accuracy and robustness. Second, combining face and voice modalities adds an extra layer of security against spoofing attacks, making it more challenging for attackers to deceive both modalities simultaneously. Third, leveraging familiar and non-intrusive modalities such as face and voice provides a user-friendly authentication experience. Lastly, the availability of large-scale datasets for face and voice allows for robust model development and benchmarking. Deep learning techniques have revolutionized the field of biometrics by enabling effective feature extraction and modeling of complex patterns in multimodal data [3]. In this research paper, we propose a multimodal biometric authentication system, leveraging the power of Siamese neural networks and autoencoders for feature extraction.

This research paper is organized as follows: Section 2 analyses the literature survey. In Section 3, our proposed model architecture is presented and used to explain the data, what we did to prepare it, and clarify the method we used to reach our goal. Experimental results are discussed in Section 4, and finally, the conclusion is explained in Section 5.

## 2 Literature survey

In recent years, biometric authentication has made significant advancements. In this paper, we present a comprehensive overview of the current status of multimodal authentication. We have thoroughly examined key studies, highlighted significant discoveries, and discussed researchers' techniques in this field. Furthermore, we have identified limitations and gaps in the existing research that have motivated our proposed approach. Our objective is to contribute to the advancement of multimodal authentication and provide context to our work within the broader research landscape.

In [1], the researchers proposed a multi-biometric authentication system by integrating face and voice biometric information. The system demonstrates promising authentication performance surpassing existing methods through cross-validation experiments using XJTU databases and the face and voice database ( GT_DB and TIMIT database). Authors of [4] presented an adaptive fusion strategy for the authentication process based on Face and Voice Using Matching Level Fusion. Through simulation experiments on a PC, the algorithm achieves a 100% authentication accuracy on the benchmark database with a training sample number of 5, and a single authentication time of approximately 341ms. This research paper [2] explores the effectiveness of combining face and voice biometrics for robust authentication. The study focuses on using the likelihood ratio (LR) classifier at the score level. By conducting experiments using the XM2VTS Benchmark database, the researchers demonstrate a consistent performance enhancement compared to the efficient sum rule, which is preceded by various score normalization techniques.

The work in [5] focuses on robust multimodal biometric authentication algorithms that combine fingerprint, iris, and voice features. Three different algorithms have been proposed, each utilizing distinct feature extraction techniques for the three modalities. The algorithms were evaluated based on classification accuracy, equal error rate (EER), and ROC curves. The second algorithm, which uses the SVM classifier and sum fusion of features, achieves a perfect classification rate of 100%. On the other hand, the first algorithm exhibits the shortest computational time, while the lowest EER is achieved by the first algorithm using features from the Karhunen-Loeve transform (KLT). In [6], the authors presented a novel approach for a multimodal biometric system using transfer learning convolutional neural networks (TL-CNN) and a modified Lion optimization algorithm (MLOA). Their system strives to attain multi-level security through biometric verification utilizing eight different types of biometric data. It has been proven to provide special recognition and authentication performance compared to traditional methods. In the meantime, [7] proposed presents an efficient multimodal biometric system that combines face, left palm print, and right palm print matching scores using score fusion. The system utilizes convolutional neural networks (CNN) and k-nearest neighbors (KNN) algorithms to recognize and identify individuals based on these multimodal biometric scores. Popular benchmarks such as the FEI face dataset, and IITD palm print database are used to train the system and create a strong and secure verification/identification system.

# 3 Proposed Method

This paper describes a multimodal system for user authentication. When designing this model, the main constraint was to authenticate a user with a single sample of data, which led us to consider Siamese Networks as a potential solution.

## 3.1 Main architecture

Our architecture, described in Figure 1, aims to authenticate users with a single sample of data. Our Siamese multimodal system uses faces and voices as modalities. It is designed to simultaneously learn representations from face and voice modalities through a fusion of learned representations.

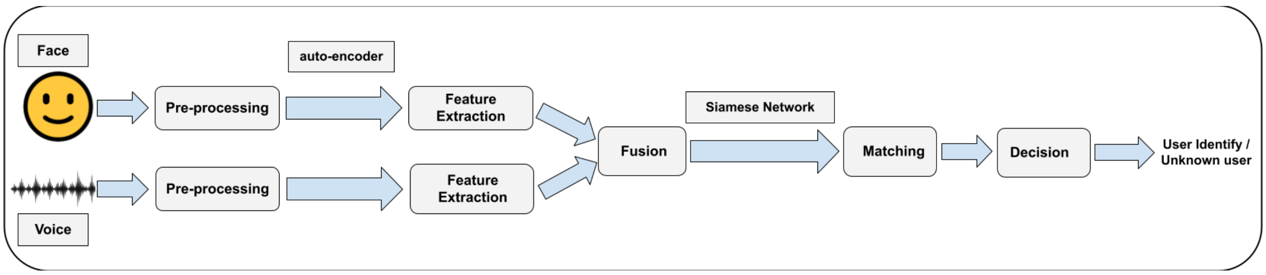


Figure 1: Biometric authentication process

The Siamese architecture consists of two or more identical sub-networks, each dedicated to processing a different modality, in this case, face and voice. These sub-networks share weights and parameters, facilitating learning a common representation space for the two modalities.
The face component of the network analyses facial images and extracts features that include important facial characteristics such as shape, texture, and landmarks. This can involve using techniques such as convolutional neural networks (CNNs) [8] for image analysis. In our system, we have used a self-trained autoencoder to improve the quality of the results.
The speech component processes audio data, usually in speech or voice recordings. Techniques such as recurrent neural networks (RNNs) [9] [10] or convolutional neural networks (CNNs) [11] [12] are typically used to extract relevant features from the audio signals, such as pitch, intensity, or spectrogram representations. In our specific case, we have used MFCC [13] features in conjunction with an autoencoder. MFCC stands for Mel-Frequency Cepstral Coefficients and is a widely used feature for analyzing audio signals, particularly

in speech and music processing. They capture a sound's spectral characteristics by transforming the signal's power spectrum into a more compact representation in the frequency domain. MFCCs are often used to extract important acoustic features that can be used for tasks such as speech recognition, speaker identification, and emotion detection.

The extracted features from the face and voice modalities are combined and integrated into the network to learn joint representations that capture their underlying relationships. This joint representation aims to capture complementary information from both modalities, allowing the network to harness the combined power of face and voice data for the authentication task at hand.

During training, the Siamese network is typically trained with pairs of examples, where each pair consists of samples from the face and voice modalities, corresponding either to the same individual or to different individuals. The network learns to map similar face and voice inputs to close points in the joint representation space while increasing the distance between dissimilar inputs. This encourages the network to acquire a discriminative embedding space in which similar face and voice features are closely clustered, facilitating subsequent matching and authentication processes. To achieve this goal, we have used the contrastive loss function.

## 3.2 Dataset preprocessing

We relied on different datasets tailored to different process stages to develop our authentication system. Specifically, we used two datasets for training the autoencoders and a final dataset for training the Siamese model. Using different datasets for each stage of the system ensured that the autoencoders were trained on relevant and specialized data. Furthermore, the dedicated dataset for training the Siamese model enabled the network to learn the intricate relationships between merged face and voice features, facilitating accurate user authentication.

### 3.2.1 Labeled Face in the Wild dataset Preprocessing

Face recognition is widely used in various industries, resulting in the availability of numerous datasets specifically designed for this purpose. In our study, we chose to use a well-known and widely used dataset considered one of the pioneering datasets in face recognition. [3] and [14]. The Labeled Face in the Wild dataset [15] comprises 13,233 faces collected from the web. This dataset consists of the 5749 identities of 1680 people with two or more images.

The LFW dataset underwent several preprocessing steps to prepare it for the face autoencoder. First, a face detection algorithm was applied to identify and extract facial regions from the images. This step, shown in Figure 2, ensured that only relevant face information was retained for subsequent analysis. Next, the pixel values of the images were normalized to a range between 0 and 1. This normalization step facilitated consistent autoencoder training by ensuring all images fell within the same intensity range. In addition, the images were converted from the BGR (blue-green-red) color space to the RGB (red-green-blue) color space. This conversion ensured that the color channels were in the correct order for subsequent processing and analysis. Finally, the images were resized to a resolution of (128, 128) pixels. This resizing step balanced retaining sufficient facial information and minimizing computational complexity. The autoencoder could extract essential features by reducing the image size without processing large and computationally intensive images. These preprocessing steps optimized the LFW dataset for subsequent training and efficient feature extraction in the face autoencoder, improving performance and accuracy.

### 3.2.2 RAVDESS dataset Preprocessing

In addition to datasets specifically curated for speaker identification tasks, it is worth noting that several datasets collected for automatic speech recognition (ASR) purposes can also be leveraged for training or evaluating speaker recognition systems. These ASR datasets provide valuable resources for studying speaker characteristics and developing robust recognition algorithms. [3]. In this work, we opted to use a database originally

Figure 2: Faces preprocessing

collected for speech emotion recognition tasks to train and evaluate our multimodal biometric authentication system.

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [16] is a well-known database that contains both audio and video recordings of emotional speech and singing. We used this database for speech data processing and also for validating our developed model. This dataset contains 1440 utterances of sentences pronounced with different emotions. Before computing the Mel Frequency Cepstrum Coefficients (MFCC) features, we applied a series of preprocessing steps to ensure a standardized and optimal speech data representation. These steps included normalizing the audio length by removing silence or selecting the middle segment of the audio if it exceeded a certain duration. And we added padding where necessary. By normalizing the audio samples in this way, we achieved consistency in the temporal aspect of the data, which enabled reliable feature extraction. The computed MFCC features then captured the salient characteristics of each voice, providing a concise and informative representation for further analysis and integration into our authentication system.

#### 3.2.3 OMG-Emotion dataset Preprocessing

As mentioned above, speaker recognition researchers have often used existing speech recognition datasets for training and evaluating their speaker recognition systems [3]. Our study used the One-Minute Gradual-Emotional Behavior dataset (OMG-Emotion dataset), pre-processed in previous work, as an evaluation set for our multimodal biometric authentication task. This dataset provides a valuable resource for evaluating the performance of our system in a multimodal setting.

The OMG-Emotion dataset [17] contains videos collected from popular online platforms, such as YouTube, and covers a wide range of emotions expressed by individuals in various real-life scenarios. It consists of 567 emotional videos, totaling approximately 15 hours of content. Each video has an average length of 1.6 minutes. We divide these videos into smaller clips of approximately 8 seconds each to facilitate further processing. This results in 7,371 clips, with an average of 12.96 utterances per video. Although the clips are initially labeled with different emotions, we focus only on the videos and do not use the emotion labels for our specific task.

The videos serve as a starting point for our input data in our approach. However, before using them, we apply several transformations to obtain the final version of our data. First, we separate the 567 videos into three sets: the training, validation, and test sets. This separation is necessary to ensure the data remains separate when creating pairs for the Siamese network. The allocation of videos follows an 80/10/10 split ratio for the train/validation/test sets.

Next, for each video clip, we perform the following steps: First, we save the clip's original audio without any modifications in a separate folder. Then we extract frames from the video. Using a face detection algorithm, we locate faces within each frame. If multiple faces are detected, we keep the face that is most centrally positioned in the frame. From the faces detected in the clip, we randomly selected five faces. Finally, these faces are resized to a shape of (128, 128, 3). The 3 represents the 3 colors channels of the image.

Using these data preparation steps, we ensure that our input data is appropriately processed and organized to enable the subsequent creation of the pairs required for Siamese network training.

Once the preprocessing step is complete, the videos are transformed into separate audio and face data, which simplifies the subsequent steps of our process. Our next task is to create pairs for our multimodal Siamese network. We need correct pairs, consisting of two faces and two audios from the same individual, and incorrect pairs, consisting of a face and voice from one individual paired with a face and voice from another individual. We follow the steps below:
First, for each individual in the dataset and each face extracted from the corresponding video, we create a correct pair. This involves selecting another face extracted from the same individual and two voice samples also extracted from the same individual. In addition, we create an incorrect pair by pairing a voice sample from the same individual with a face and voice sample from a different individual, ensuring that the data remains properly separated according to the predefined sets, such as training, validation, or test.
This process results in the creation of pairs, which are meticulously organized into separate folders, thereby establishing a well-defined folder tree structure, as illustrated in Figure 3. For further clarification, the dataset is systematically divided into three distinct subsets: training, validation, and testing. It is important to note that the individuals do not overlap across these subsets to eliminate any potential bias. Each subset comprises two specific folders. The first folder contains 'correct pairs,' where both the images and voice samples correspond to the same individual. On the other hand, the second folder houses 'incorrect pairs,' where the images and voice samples do not match, indicating that they belong to different individuals.

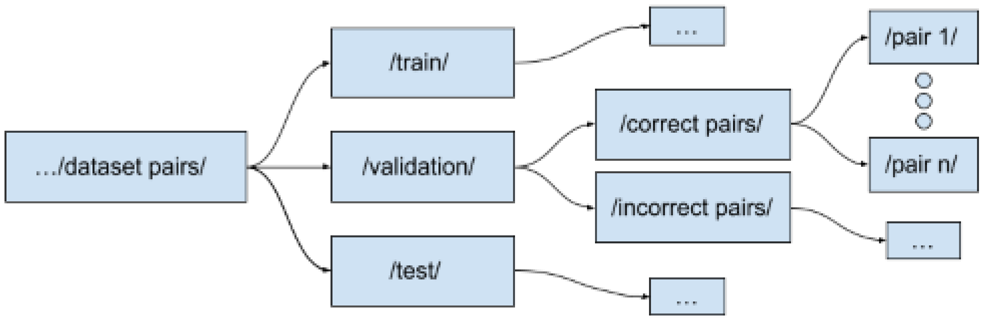


Figure 3: Dataset as a tree architecture

The next and final stage is to prepare the pairs for training, which involves similar steps to those used for the previous two datasets. The image data is normalized between 0 and 1 and converted from BGR to RGB format. It is important to note that the images have already been resized. As for the audio files, size normalization is performed. This involves eliminating any initial or final silence and selecting the middle segment if the audio duration exceeds a specified threshold. Padding is introduced when the audio duration falls below the threshold. In addition, 40 Mel Frequency Cepstrum (MFCC) features are extracted from the normalized audio data.

## 3.3 Feature extraction

As mentioned earlier, the Siamese network uses feature extractors to acquire embedded representations of voices and faces. There are several methods available to achieve this objective.

### 3.3.1 Face features extractor

In the case of images, Convolutional Neural Networks (CNNs) are widely utilized, and pre-trained networks like ResNets have demonstrated notable proficiency in feature extraction [18]. However, given our specific task of user authentication, we exclusively focus on faces. Therefore, instead of employing a large-scale CNN pre-trained on general datasets like ImageNet [19], we have opted to train a lighter autoencoder on a more targeted face-specific dataset. This approach allows us to develop a feature extractor specifically tailored for faces, enabling more streamlined and efficient face processing. We conducted comparisons to identify the most suitable model for our task, and the results are presented in the subsequent sections of this paper. Our autoencoder consistently outperformed other models, establishing it as the optimal choice for face feature extraction. As mentioned above, we have used the dataset LFW (Labeled Faces in the Wild) for the training of this model. The preprocessing steps of the data have already been described.

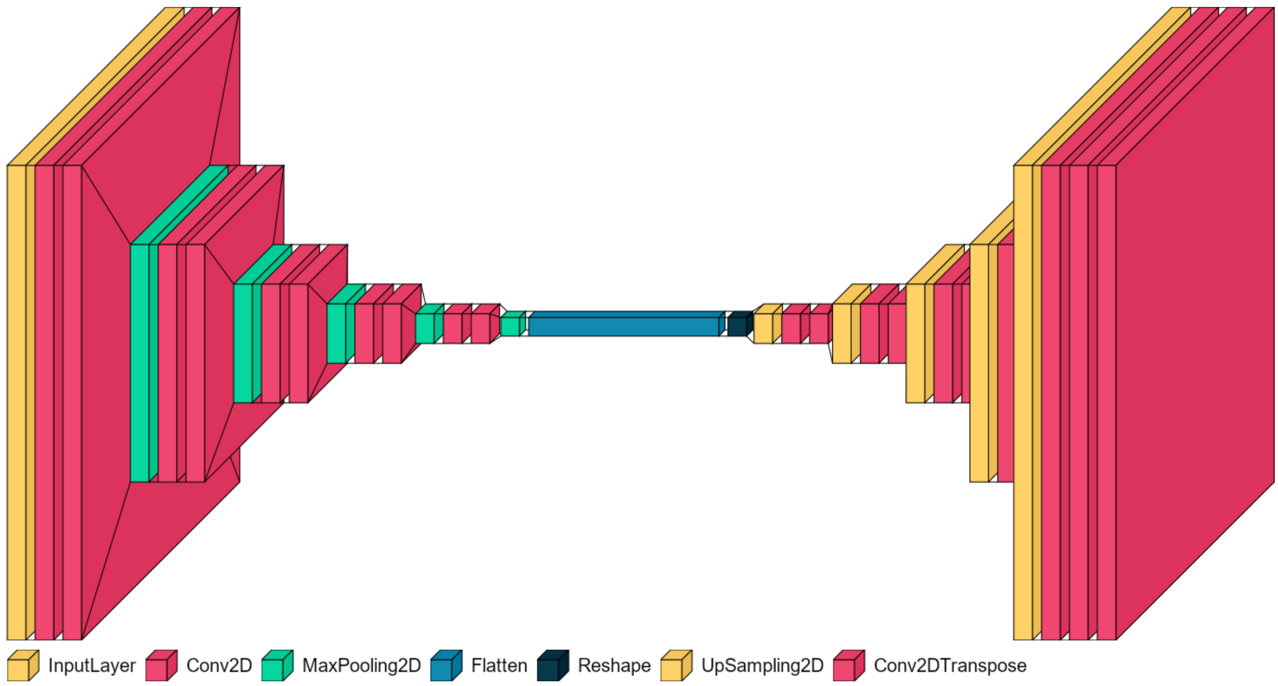


Figure 4: Faces autoencoder Architecture

The Face autoencoder architecture, as shown in Figure 4, consists of approximately 1.9 million parameters. The encoder section is structured with a pattern of two convolutional layers followed by a max-pooling layer using the ReLU activation function. This configuration is repeated 5 times for a total of 877k parameters. The kernel size is set to (3, 3) with padding preserved. After dimension reduction, the resulting feature map is flattened to produce a vector of size 2048, which serves as the desired output of the encoder. This encoder becomes the face feature extractor after training.
The decoder section, on the other hand, follows a similar pattern, but to increase dimensions. It uses UpSampling2D layers instead of max-pooling and Conv2DTranspose layers instead of regular convolutional layers. It is important to note that the decoder part is not used in the final Siamese network and therefore does not contribute to the subsequent steps of the model.

### 3.3.2 Voice features extractor

For speech analysis, several techniques have shown promising results, including spectrogram transformation and Mel Frequency Cepstral Coefficients (MFCC). In our research, we specifically explored the use of MFCC features for speech data, which are the most commonly used features for speaker recognition [20]. These features effectively capture important voice characteristics, making them suitable for our purposes. To train our speech feature extractor, we used the RAVDESS Emotional speech audio dataset. By computing the MFCC features from the speech samples, we obtained meaningful representations of the speech data, allowing us to work with more informative features rather than raw audio signals.
Unlike the face autoencoder, the speech autoencoder has a different architecture and approach. Given the nature

of speech analysis, we used Long Short-Term Memory (LSTM) [21] layers and dropout layers to construct the autoencoder described in Figure 5.

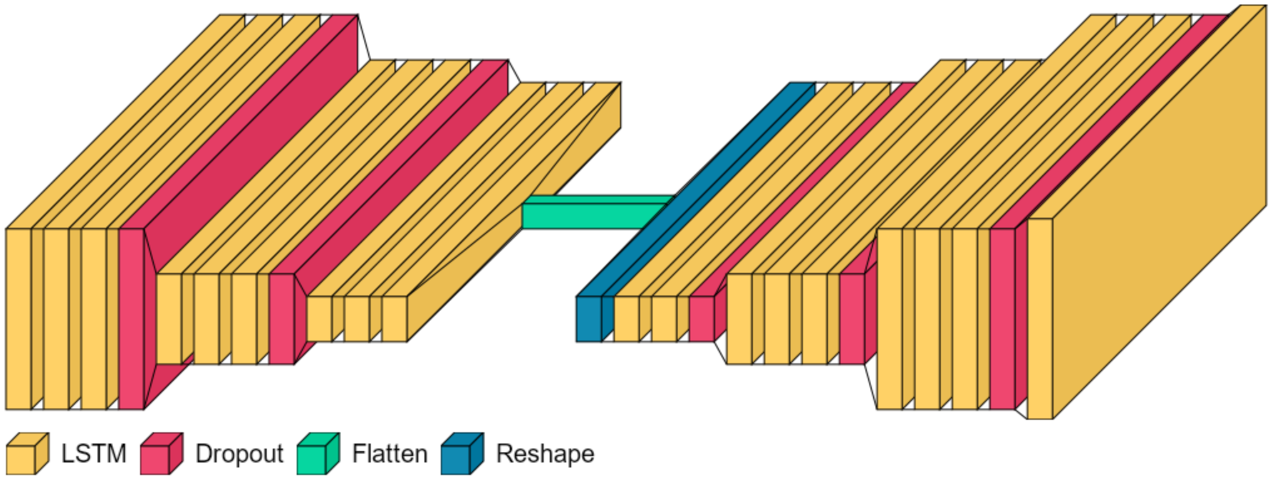


Figure 5: Voices autoencoder Architecture

The goal was to create a 2048-dimensional vector representation that effectively encodes the relevant and discriminative voice features. These features were then reconstructed using additional LSTM layers to ensure that the reconstructed speech data closely matched the original input.
The encoder component of the trained autoencoder serves as the feature extractor within our Siamese network. This feature extraction process enables the Siamese network to effectively compare and match voice samples for user verification and authentication.

## 3.4 Fusion

Fusion techniques in deep learning aim to combine information from multiple biometric modalities to improve the performance of biometric systems. Various fusion strategies have been developed [22], including fusion of raw sensor data, fusion of features from many modalities, fusion of comparison scores, and fusion on the decision level. Fusion of raw sensor data combines the raw data collected from multiple biometric sensors to create a unified representation for authentication. The fusion of features from many modalities integrates the extracted features from different biometric modalities, such as fingerprints, iris scans, and voice patterns, to create a comprehensive feature set for authentication.
Fusion of comparison scores combines the similarity or dissimilarity scores obtained from individual biometric comparisons to make a final decision. Fusion on the decision level fuses the decisions made by individual biometric classifiers or systems to reach a final authentication outcome.
The choice of fusion technique depends on the specific task and data characteristics and requires careful consideration for optimal performance. In addition, there are different methods for feature fusion, such as concatenation, element-wise fusion, and weighted fusion.
In the context of this paper, we employed a feature-level fusion to enhance the process of biometric authentication. This method follows the features extraction section of the architecture. We proceeded to concatenate these features, effectively combining the facial and vocal characteristics into a single, comprehensive feature vector. This unified feature vector then served as the input for the Siamese authentication section of the model.

### 3.5 Siamese Network

A Siamese network [23] is a type of neural network architecture that consists of multiple identical subnetworks called twins or branches. It is commonly used for tasks that compare similarities or distances between inputs. In [24], Siamese networks have been used to perform cross-system high-assurance authentication of users in virtual reality (VR) environment. By sharing weights and learning to map pairs of inputs onto a shared feature space, Siamese networks enable effective comparison and classification of similar and dissimilar instances. In the final step of our approach, we combine each individual's feature vectors of the face and voice modalities before calculating the distance metric. This fusion process allows us to capture complementary information from both sources and increase the discriminative power of our Siamese network. To achieve this, we concatenate each individual's face feature vector and voice feature vector into a single merged vector. This merging operation creates a joint representation that encapsulates the distinctive features of the individuals' facial and vocal attributes. By integrating these modalities, we aim to exploit the unique patterns and correlations between faces and voices to improve the accuracy and robustness of our authentication system. The merged vectors, now representing the combined face and voice features, are the input for calculating the distance between the two individuals. We use the Euclidean distance metric, which measures the dissimilarity between the merged vectors in feature space. By calculating the Euclidean distance, we obtain a quantitative measure of the dissimilarity or similarity between individuals based on their fused face and voice information. Through extensive experimentation and validation, we have found that the Euclidean distance provides effective discrimination and separation capabilities, enabling reliable differentiation between genuine and impostor identities.
We use the Contrastive Loss function [25] during the training phase to optimize the Siamese network.

$$\mathcal{L}_{\text{contrastive}} = \underbrace{yd^2}_{\text{Term for similar samples}} + \underbrace{(1-y)\max(\mathrm{m}-d,0)^2}_{\text{Term for dissimilar samples}} \tag{1}$$

Where $y$ represents the labels, 0 being similar and one being dissimilar. $d$ represents the distance between the two representations, and $m$ is the margin, a hyper-parameter that controls the separation between similar and dissimilar samples.
This loss function uses the calculated distance between the merged feature vectors to generate a loss value propagated back through the network during the learning process. By minimizing this loss, the network learns to better discriminate between matching and non-matching pairs, thereby improving its ability to identify individuals based on their facial and vocal features accurately.
In the testing phase, we use a threshold-based approach to decision-making. Specifically, if the computed distance between the merged feature vectors of two individuals falls below a predefined threshold (typically set at 0.5), we infer that the individuals are the same person. Conversely, if the distance is above or equal to the threshold, we infer that the individuals differ. This choice of threshold allows us to control the trade-off between false acceptance and rejection rates, allowing the system to be tailored to specific requirements and desired performance levels.
In summary, we have developed a robust and effective approach to user authentication by merging the face and voice feature vectors and using the Euclidean distance metric within a Siamese network framework. Fusing face and voice modalities enhances the system's discriminative power, while distance calculation and threshold-based decision-making provide reliable identity verification capabilities. Through extensive experimentation and evaluation, we have demonstrated the effectiveness of our approach in accurately distinguishing between individuals and ensuring secure access control.

## 4 Results and Discussion

In this section, we present the training results and evaluate three Siamese networks using the OMG-Emotion dataset. According to our knowledge, the databases employed in our work are used for the first time in user

authentication research. Therefore, comparison with other articles is limited.
Many studies have focused on multimodal user authentication, including the fusion of fingerprint and iris. However, there needs to be more research on the fusion of voice and face modalities. We conducted a study where we trained three different networks for user authentication. The first only uses face data, the second only uses voice data, and the third combines face and voice data. We aimed to show that the 'Merged' model performs better than the other two separate models, showcasing the benefits of combining multiple modalities for improved user authentication.

| **Metric** | **Face** | **Voice** | **Merged** |
|---|---|---|---|
| Loss | 0.06 | 0.11 | 0.05 |
| Accuracy (%) | 94.23 | 87.18 | 95.46 |
| False Positive (%) | 2.47 | 9.81 | 2.47 |
| False Negative (%) | 3.31 | 3.01 | 2.07 |
| Precision (%) | 97.45 | 89.88 | 97.48 |
| Recall (%) | 96.61 | 96.66 | 97.87 |
| F1 Score (%) | 97.03 | 93.15 | 97.68 |

Table 1: Performances of the different models according to several metrics. The model 'Merged' means the merge of the Face and Voice modalities

In table 1, we utilized various metrics to evaluate the performance of our models. In the subsequent section, we will elaborate on the significance of each metric and how it helps in determining the effectiveness of our approach.
When a model mistakenly assigns a positive label to a negative instance, it's called a false positive. This error exposes the model's tendency to misidentify something as positive when it's actually not. Conversely, a false negative occurs when a model wrongly assigns a negative label to a positive instance, indicating that the model failed to recognize a positive instance. Precision (see equation 2) refers to the proportion of true positive predictions compared to all positive predictions made by the model. A higher precision value means fewer false positives, which shows the model's capability to accurately identify positive instances.

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

Recall (see equation 3) is a metric that determines the percentage of accurate positive predictions in the dataset. A higher recall value implies a reduced number of false negatives, which showcases the model's capability of accurately identifying positive instances.

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

finally, the F1 (see equation 4) score is a comprehensive evaluation of a model's performance as it combines precision and recall, offering a balanced measure. It takes into account the accuracy of positive predictions and the model's capacity to detect positive instances, resulting in an overall assessment of its effectiveness.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The loss is calculated using the validation data from the best epoch, while the remaining metrics are evaluated using the test data. It is important to note that the validation and test datasets are completely distinct and separate from the training data.
The results presented in Table 1 clearly show that the merged model outperforms the single modality systems

in all evaluated metrics, demonstrating a significant performance improvement. This superiority of the merged model underscores the effectiveness of incorporating multiple modalities to achieve superior results in our analysis.
To confirm our analysis, we tested our model on another dataset. For this second application, we used the RAVDESS dataset [16]. We used the same preprocessing as before, which is to extract faces and audio from the videos, and then normalize and use the data. For this dataset, we achieved an accuracy of 89.79%.

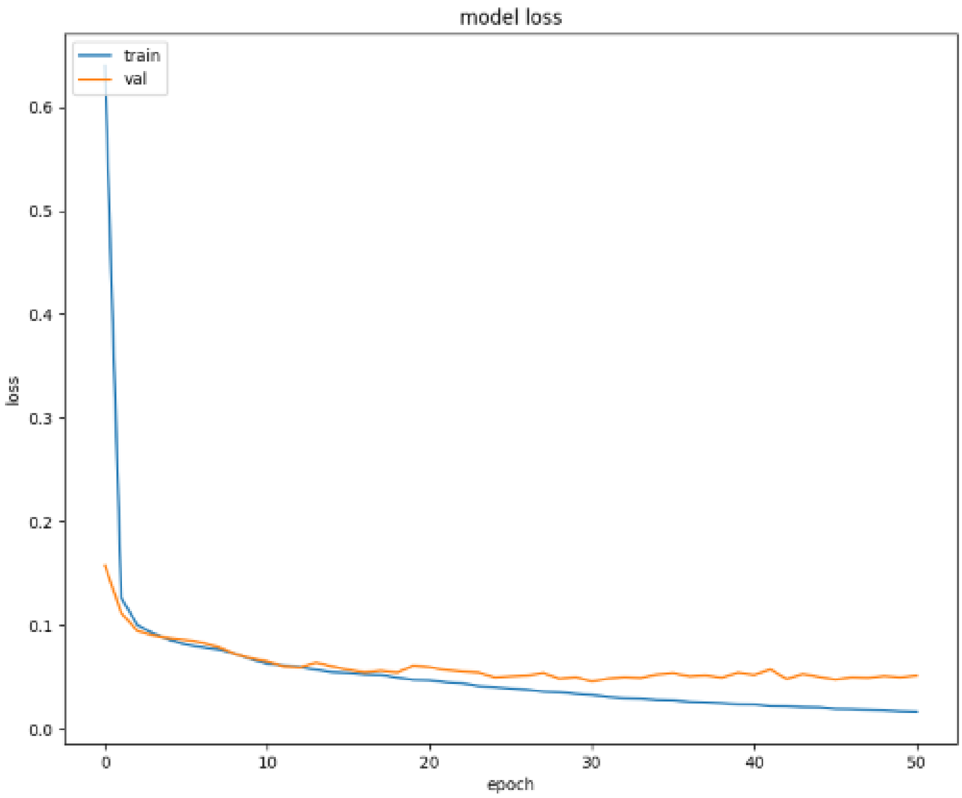


Figure 6: Merged Model's Loss during Training

In the training, validation, and test phases, the hyperparameters used are displayed in Table 2. One of these hyperparameters is the threshold, a predetermined value used to distinguish between positive and negative instances in a binary classification task. In our case, the hall serves as the cut-off point. We evaluated different thresholds ranging from 0.1 to 1 and selected the one that offered the best outcomes on the test set based on the metrics we evaluated. However, it's essential to remember that the threshold can be adjusted to meet the specific requirements of your data, such as during experiments or demonstrations.
Figure 6 illustrates the training and validation loss curves. The specific loss function utilized was the *contrastive loss*. The curves show a good fit, as the training loss plot decreases to a point of stability, and the validation loss plot reaches a point of stability with a small gap from the training loss. It is important to mention that an *EarlyStopping* callback with patience of 20 was incorporated, which is why the figure displays 51 epochs. It is worth mentioning that the top-performing model was saved during epoch 31, coinciding with the lowest validation loss.
We thoroughly searched for audio-visual databases that could help us evaluate and compare our model. Unfortunately, despite our exhaustive efforts and different search strategies, we couldn't find many suitable databases

| Hyperparameters | Face | Voice | Merged |
|---|---|---|---|
| Train Samples | 14,366 | 14,366 | 14,366 |
| Validation Samples | 1774 | 1774 | 1774 |
| Test Samples | 1774 | 1774 | 1774 |
| Epochs | 27 | 44 | 31 |
| Batch Size | 32 | 32 | 32 |
| Test Batch Size | 32 | 32 | 32 |
| Optimizer | Adam | Adam | Adam |
| Learning Rate | 1e-4 | 1e-4 | 4e-5 |
| Beta 1 | 0.9 | 0.9 | 0.9 |
| Beta 2 | 0.999 | 0.999 | 0.999 |
| Epsilon | 1e-07 | 1e-07 | 1e-07 |
| Loss | Constrastive Loss | Constrastive Loss | Constrastive Loss |
| Number Total Parameters | 877,344 | 43,776 | 921,120 |
| Number Trainable Parameters | 877,344 | 43,776 | 921,120 |
| Number Non-Trainable Parameters | 0 | 0 | 0 |
| Best Threshold (Inference) | 0.45 | 0.5 | 0.4 |

Table 2: Trained Models Hyperparameters

that met our criteria for accessibility, compatibility, and relevance to our research objectives. We explored several well-known repositories and online platforms like the XJTU multimodal database, the XM2VTS Database, the BANCA database, BioSoft, etc. However, none provided openly accessible audio-visual datasets for researchers in our field. This lack of publicly available databases restricted our ability to conduct comprehensive analyses and validate our model against existing standards.

## 5 Conclusion and Future Work

In this study, we introduced a multimodal authentication system that sets a new standard for efficiency and reliability in the domain of biometric security. By combining facial and vocal biometrics, our system addresses and surmounts the inherent limitations found within single-modality frameworks, thereby enhancing overall system performance. The employment of advanced deep learning techniques, alongside the strategic training of a Siamese model using the OMG-emotion and RAVDESS databases, has propelled our research to the forefront of multimodal biometric authentication. Our contributions not only improve accuracy and robustness but also pave the way for more secure and dependable real-world applications.

Our research reveals that although single-modality (uni-modal) systems provide adequate results on their own, combining them into a unified multimodal system allow to fully use the pertinent features of both modalities to enhance the accuracy of the global recognition rate. This synergistic effect significantly lowers the rate of errors, clearly demonstrating the superiority of a multimodal approach in achieving not only greater accuracy but also in promoting fairness across different biometric analyses.

Looking forward, we are committed to further refining the efficiency of our proposed model. We aim to extend our research to include additional benchmark databases and to explore diverse fusion techniques, especially in scenarios where modalities exhibit independence. This ongoing pursuit of improvements positions our work not just as a significant contribution to the field, but as a landmark for future advancements in multi-modal biometric authentication technology.

## References


[1] X. Zhang, D. Cheng, P. Jia, Y. Dai, X. Xu, An efficient android-based multimodal biometric authentication system with face and voice, IEEE Access 8 (2020) 102757–102772. doi:https://doi.org/10.1109/ACCESS.2020.2999115.

[2] M. Bengherabi, L. Mezai, F. Harizi, A. Guessoum, M. Cheriet, Robust authentication using likelihood ratio and gmm for the fusion of voice and face, in: 2009 3rd International Conference on Signals, Circuits and Systems (SCS), IEEE, 2009, pp. 1–6. doi:https://doi.org/10.1109/ICSCS.2009.5412538.

[3] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, D. Zhang, Biometrics recognition using deep learning: A survey, Artificial Intelligence Review (2023) 1–49doi:https://doi.org/10.1007/s10462-022-10237-x.

[4] X. Zhang, D. Cheng, Y. Dai, X. Xu, Multimodal biometric authentication system for smartphone based on face and voice using matching level fusion, in: 2018 IEEE 4th International Conference on Computer and Communications (ICCC), IEEE, 2018, pp. 1468–1472. doi:https://doi.org/10.1109/CompComm.2018.8780935.

[5] M. S. El_Tokhy, Robust multimodal biometric authentication algorithms using fingerprint, iris and voice features fusion, Journal of Intelligent & Fuzzy Systems 40 (1) (2021) 647–672. doi:https://doi.org/10.3233/JIFS-200425.

[6] A. Gona, M. Subramoniam, R. Swarnalatha, Transfer learning convolutional neural network with modified lion optimization for multimodal biometric system, Computers and Electrical Engineering 108 (2023) 108664. doi:https://doi.org/10.1016/j.compeleceng.2023.108664.

[7] C. Medjahed, A. Rahmoun, C. Charrier, F. Mezzoudj, A deep learning-based multimodal biometric system using score fusion, IAES Int. J. Artif. Intell 11 (1) (2022) 65. doi:https://doi.org/10.11591/ijai.v11.i1.pp65-80.

[8] M. Asim, Z. Ming, M. Y. Javed, Cnn based spatio-temporal feature extraction for face anti-spoofing, in: 2017 2nd International Conference on Image, Vision and Computing (ICIVC), IEEE, 2017, pp. 234–238. doi:https://doi.org/10.1109/ICIVC.2017.7984552.

[9] Y. Miao, M. Gowayyed, F. Metze, Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding, in: 2015 IEEE workshop on automatic speech recognition and understanding (ASRU), IEEE, 2015, pp. 167–174. doi:https://doi.org/10.1109/ASRU.2015.7404790.

[10] A. Shewalkar, D. Nyavanandi, S. A. Ludwig, Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru, Journal of Artificial Intelligence and Soft Computing Research 9 (4) (2019) 235–245. doi:https://doi.org/10.2478/jaiscr-2019-0006.

[11] Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using cnn, in: Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 801–804. doi:https://doi.org/10.1145/2647868.2654984.

[12] V. Passricha, R. K. Aggarwal, A hybrid of deep cnn and bidirectional lstm for automatic speech recognition, Journal of Intelligent Systems 29 (1) (2019) 1261–1274. doi:https://doi.org/10.1515/jisys-2018-0372.

[13] C. Ittichaichareon, S. Suksri, T. Yingthawornsuk, Speech recognition using mfcc, in: International conference on computer graphics, simulation and modeling, Vol. 9, 2012.

[14] X. Zhang, W. Gong, X. Xu, Y. Zhang, Biometric recognition databases: A survey, in: Proceedings of the 2018 VII International Conference on Network, Communication and Computing, 2018, pp. 77–81. doi:https://doi.org/10.1145/3301326.3301384.

[15] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, G. Hua, Labeled faces in the wild: A survey, Advances in face detection and facial image analysis (2016) 189–248doi:https://doi.org/10.1007/978-3-319-25958-1_8.

[16] S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one 13 (5) (2018) e0196391. doi:https://doi.org/10.1371/journal.pone.0196391.

[17] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, S. Wermter, The omg-emotion behavior dataset, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–7. doi:https://doi.org/10.1109/IJCNN.2018.8489099.

[18] B. Li, D. Lima, Facial expression recognition via resnet-50, International Journal of Cognitive Computing in Engineering 2 (2021) 57–64. doi:https://doi.org/10.1016/j.ijcce.2021.02.002.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255. doi:https://doi.org/10.1109/CVPR.2009.5206848.

[20] V. Tiwari, Mfcc and its applications in speaker recognition, International journal on emerging technologies 1 (1) (2010) 19–22.

[21] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780. doi:https://doi.org/10.1162/neco.1997.9.8.1735.

[22] P. Szczuko, A. Harasimiuk, A. Czyżewski, Evaluation of decision fusion methods for multimodal biometrics in the banking application, Sensors 22 (6) (2022) 2356. doi:https://doi.org/10.3390/s22062356.

[23] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a" siamese" time delay neural network, Advances in neural information processing systems 6 (1993). doi:https://doi.org/10.1142/s0218001493000339.

[24] R. Miller, N. K. Banerjee, S. Banerjee, Using siamese neural networks to perform cross-system behavioral authentication in virtual reality, in: 2021 IEEE Virtual Reality and 3D User Interfaces (VR), IEEE, 2021, pp. 140–149. doi:https://doi.org/10.1109/VR50410.2021.00035.

[25] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), Vol. 2, IEEE, 2006, pp. 1735–1742. doi:https://doi.org/10.1109/CVPR.2006.100.