ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 14 (2024) No. 5, pp. 113-130
DOI:10.33168/JSMS.2024.0507

Feature-Level Fusion of FaceNet and Wav2vec Representations for Multimodal Biometric Authentication

Novita Retno Puji Lestari, Gede Putra Kusuma

Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

novita.lestari001@binus.ac.id, inegara@binus.edu

Abstract. Biometrics are highly popular due to their ease and convenience in utilizing unique human physical characteristics. Biometrics has advanced, and multimodal biometrics are more accurate than unimodal biometrics. Multimodal biometrics involves the combination of two or more biometric modalities. This study investigates multimodal biometric authentication using face and voice modalities with deep learning models. The FaceNet and wav2vec models are utilized for feature extraction from face images and voice samples of 50 subjects from the MSU-AVIS dataset. Feature level and score level fusion strategies are implemented. Using the Euclidean distance and classifiers methods such as Decision Tree, Support Vector Machine (SVM), Random Forest, XGBoost, and Artificial Neural Network (ANN), performance is evaluated by correct acceptance rate (CAR) at 0.1 and 0.01 false acceptance rates and area under the ROC curve (AUC). The proposed feature level fusion approach with XGBoost obtained the highest CAR of 0.995 at 0.01 FAR and AUC of 0.996 on the test set, demonstrating effective multimodal biometric authentication.

Keywords: face authentication, voice authentication, multimodal biometric, fusion method, deep learning

1. Introduction

User authentication is the process carried out to verify the identity of a user who wants to access a system, ensuring that the user is authorized to use the application. The purpose of user authentication is to protect the security of the system and the data stored within it from unauthorized access. Typically, user authentication relies on usernames and passwords. However, it has several vulnerabilities (Farik et al., 2016), including susceptibility to physical attacks through password-recording cameras or keyloggers, vulnerability to brute force attacks for password guessing, the possibility of data leakage of stored passwords and usernames within databases, the potential for phishing attacks leading users to unintentionally disclose password and username information, the use of weak passwords, and vulnerability to password forgetfulness.

Biometric authentication is a method of verifying someone's identity using their unique physical or behavioural characteristics such as face, voice, fingerprint, iris, etc. Authentication using biometrics is highly popular as a more secure and convenient alternative to remembering passwords or PINs. Users only need to use their biological attributes without remembering and typing complex passwords. The development of biometric technology has advanced rapidly, and multimodal biometrics involves the combination of two or more biometric modalities, which is considered superior to unimodal biometrics. Multimodal biometrics involve the integration of multiple biometric modalities to enhance the accuracy and reliability of identification or authentication systems. The growth of multimodal biometrics stems from its ability to address the limitations of unimodal systems, offering more robust and secure solutions. As advancements in technology have made multimodal data acquisition and integration more feasible, researchers and practitioners increasingly recognize its potential for a wide range of applications. The market size of biometric usage was valued at USD 872.7 million in 2019 and is expected to increase at a compound annual growth rate (CAGR) of 24.4% from 2020 to 2027. The rise in online transactions and increase in fraudulent activities is driving the demand for digital authentication across world (Grand View Research, 2023). A study (Oloyede & Hancke, 2016) comparing the performance of unimodal and multimodal biometrics showed that multimodal biometrics provide higher accuracy compared to unimodal biometrics. Additionally, multimodal biometrics also offer higher security levels as it is challenging for attackers to deceive multiple biometric technologies simultaneously. Besides providing better accuracy, the possibility of fraud is reduced. Using more than one biometric data, less noise is generated from the input data, so the error rate is reduced (Baghelp et al., 2017).

The study (Zhang et al., 2020) applied biometric authentication using multimodal face and voice, resulting in accuracy. Face authentication, employing the LBP feature extraction method, achieved 99.92% accuracy, while voice authentication, using the MFCC feature extraction method, achieved 98.85% accuracy. Another recent study (Saleh & Jouny, 2022) on multimodal biometrics involved face feature extraction using Viola-Jones and voice feature extraction using MFCC. The fusion method employed was decision-level fusion. The experiment, which utilized 102 test cases, yielded 101 successful cases and 1 failed case.

In combining or integrating two or more biometric technologies, fusion is performed. Fusion in implementing multimodal biometrics is a technology combining multiple biometrics types. Several fusion techniques include sensor level fusion, feature level fusion, score level fusion, and decision level fusion. A study (Byahatti & Shettar, 2020) conducted experiments on multimodal face and voice with various fusion techniques, and the best result was achieved using feature level fusion. Deep learning is a branch of machine learning that utilizes large and complex neural network architectures to process data and automatically learn complex features. Over time, deep learning becomes more intelligent by training itself using vast amounts of data. Deep learning technology has been used for various tasks, such as facial and voice recognition. Deep learning provides more accurate results compared to traditional techniques.

Face recognition methods by comparing two deep learning models, namely FaceNet and Openface.

The FaceNet model performs 6.7% better than the OpenFace model (Cahyono et al., 2020). The study (Nyein & Oo, 2019) compares the performance of FaceNet, VGG16, and CNN model in face recognition. The results show that Facenet is the best model with an accuracy of 98.66%, while VGG16 achieves 94.66%, and CNN achieves 93.33%. Facenet as facial authentication in a presence system (Liu et al., 2019) produces 99.5% accuracy. The voice authentication system using wav2vec in the feature extraction process (Moustafa & Aly, 2021) produces 98% accuracy in the experimental process. Biometric fusion methods are very diverse, such as sensor level, feature level, score level and decision level. In research (Byahatti & Shettar, 2020) feature-level fusion produced the best performance.

Based on previous research findings, there have been no studies utilizing deep learning in the feature extraction process for multimodal biometrics. Therefore, this research focuses on utilizing multimodal face and voice biometrics for authentication by employing deep learning in feature extraction. Fusion in multimodal biometrics is achieved through feature-level and score-level methods. Deep learning models, MTCNN and FaceNet, are utilized for face biometrics, while wav2vec is utilized for voice biometrics. Various algorithms, including decision tree, SVM, random forest, XGBOOST, and ANN, are applied for training and testing the data. Consequently, the experiment produces multiple outcomes that can be utilized to compare the best results. Additionally, evaluation is conducted using the Euclidean distance. The anticipated output of this paper is to contribute to the advancement of multimodal biometric systems for face and voice through the utilization of deep learning, ultimately aiming for superior performance.

2. Related Works

2.1. Multimodal Biometric Face and Voice

Several studies have been conducted on multimodal biometric face and voice. Research (X. Zhang et al., 2017) focuses on multimodal face and voice as an identification system. The fusion method employed is feature level fusion. The fusion process generates high-dimensional vectors that combine the face feature vector and the voice feature vector. Feature extraction for voice and face is carried out independently using the Haar wavelet transformation. After fusion, the matrix undergoes classification using SVM, with the output classification results being 1 and 0. This research achieves an accuracy of 93.6%, with a FAR value of 0% and an FRR value of 1.4%. The proposed system overcomes the limitations of single-face recognition and single-voice verification, as it exhibits significantly higher accuracy compared to unimodal systems.

Research (X. Zhang et al., 2020) explores multimodal face and voice as an authentication system for Android applications. The fusion method employed is score level fusion. The score level is determined by two calculations: the summation and multiplication of the face and voice scores, each with a specific weight range. Each calculation is then normalized to a value between 1 and 0. In the face modality, face detection is performed using a combination of the Haar cascade and AdaBoost algorithms to enhance accuracy, followed by normalization. Feature extraction is conducted using Local Binary Patterns (LBP), and the score for the face biometric is determined by calculating the Euclidean Distance. As for the voice modality, Voice Activity Detection (VAD) is performed, followed by feature extraction based on Mel Frequency Cepstral Coefficients (MFCC). The score calculation method for the voice biometric involves using the Maximum A Posteriori (MAP) approach. The research evaluates the system using cross-validation on two public datasets: XJTU and a combination of GT_DB and TIMIT. The XJTU dataset yields a FRR value of 0% and a FAR value of 0%, while the GT_DB and TIMIT dataset results in a FRR value of 9.72% and a FAR value of 9.29%.

In another recent study (Sujana & Reddy, 2021), multimodal biometric face and voice are investigated using the decision level fusion method. The MSU-AVIS dataset is utilized for the research. Face detection is performed using the Viola-Jones algorithm, and classification is carried out using the deep-learning AlexNet model. For the voice modality, feature extraction involves capturing two features,

namely pitch and MFCC, followed by classification using the deep learning VGGish model. Fusion is achieved by calculating the confidence score for each CNN-based face and voice classification. Three calculation techniques are implemented, namely raw confidence score, normalized confidence score, and entropy confidence score. Experimental results from 102 test cases indicate a success rate of 98 test cases and a failure rate of 4 test cases.

2.2. Biometric Face

Face recognition methods in an attendance system by comparing two deep learning models, namely FaceNet and Openface (Cahyono et al., 2020). A private dataset is used, and preprocessing is performed using MTCNN. In FaceNet, the input images are resized to 160x160 pixels and result in a 128-dimensional vector. In Openface, the input images are resized to 96x96 pixels and produce a 128-dimensional vector. Classification is carried out using SVM with a linear kernel and true probability. The data is divided into 80% for training and 20% for testing. The validation system utilizes 5-fold cross-validation, which yields a 100% accuracy for the FaceNet model and a 93.3% accuracy for the OpenFace model. The implementation of the FaceNet model achieves 100% accuracy using a threshold of 0.25. The study (Nyein & Oo, 2019) compares the performance of FaceNet, VGG16, and CNN model in face recognition. Face detection is carried out using PIL and the face recognition library, and the detected faces are resized to 160x160 pixels as input. The experiment uses a dataset consisting of facial images from 55 individuals, and SVM classification is applied. The results show that FaceNet achieves an accuracy of 98.66%, VGG16 achieves 94.66%, and CNN model achieves 93.33% accuracy.

2.3. Biometric Voice

Voice biometrics as an authentication method on the Android platform (X. Zhang et al., 2018). This study employs a random shuffling algorithm for voice cypher generation, resulting in a random sequence of 8-digit numbers. The voice features are extracted using the MFCC (Mel-Frequency Cepstral Coefficients) technique, and the GMM (Gaussian Mixture Model) method is utilized for classification. The training data consists of 5 samples, and the research achieved an accuracy ranging from 89% to 96%. The authentication process takes approximately 210ms to 320ms. The research (Arshad et al., 2022) utilized deep learning in Biometric Voice using the wav2vec model. The model underwent pre-training to extract useful features from audio data, such as frequency patterns and amplitude, resulting in an embedding vector representing the information within the audio data. In this research, Biometric Voice was applied to automatic speech recognition, achieving an error rate of 1.8 using the public dataset called Librispeech. This research highlights the potential of deep learning, particularly the wav2vec model, in enhancing the performance of Biometric Voice systems for tasks like automatic speech recognition. Utilizing such models enables more robust and accurate voice-based applications and contributes to the advancement of voice biometrics technology.

Based on the above references, facial biometrics with the best performance is the Face Net deep learning model (William et al., 2019), while voice biometrics will try to use the wav2vec deep learning model. This study utilizes the MSU-AVIS dataset, where a previous study (Saleh & Jouny, 2022) achieved an 80% success rate using decision level fusion. Research (Byahatti & Shettar, 2020) reviewed fusion in multimodal biometrics and found that feature level fusion is the best method.

So, this research will rigorously evaluate the performance of the FaceNet deep learning model for facial biometrics and the wav2vec deep learning model for voice biometrics. To comprehensively assess the efficacy of multimodal biometrics fusion, we will explore two fusion methods: feature-level fusion and score-level fusion. Feature level involves the merging of face and voice feature vectors using concatenation. Score level is the average score from the best classification results of face and voice biometrics. In terms of data analysis, we will not only leverage machine learning algorithms such as decision trees, support vector machines (SVM), random forests, XGBOOST, and artificial neural networks (ANN) but also incorporate evaluation using the Euclidean distance. This comprehensive approach will allow us to benchmark our results against a diverse set of methods commonly used in the

field of multimodal biometrics.

3. Theory and Methods

3.1. Multimodal Biometrics

Multimodal biometrics can provide better accuracy than single biometrics. Therefore, multimodal authentication schemes have become a new trend in technology. Multimodal biometrics is a system that utilizes more than one physiological or behavioral characteristic of individuals for verification or identification purposes, often referred to as the fusion of two or more biometric techniques (Dargan & Kumar, 2020). Research (Gudavalli et al., 2012) on multimodal biometrics involves reviewing multimodal sources, architectures, and fusion techniques. Multimodal sources include multiple sensors, multiple classifiers, multiple instances (different aspects), multiple units, and multiple biometrics. Multimodal architectures can be categorized as serial (using one biometric modality followed by another) or parallel (combining the results). The study concludes that combining multiple biometric modalities in a multimodal system can improve performance and prevent spoofing attacks (Dhole et al., 2012).

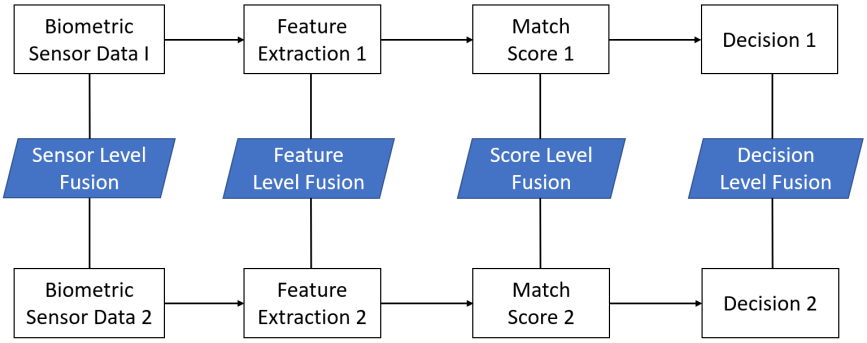

Fig 1: Fusion in Multimodal Biometrics

There are several fusion methods (Oloyede & Hancke, 2016) in multimodal biometrics, namely sensor level, feature level, score level, and decision level, as shown in Fig 1. Sensor level is a fusion technique that collects information from multiple samples of the same biometric, such as face and iris images taken from different cameras and angles or faces images taken after recording voice within a certain interval (Srivastava, 2007). This information is integrated to build new information for the next process.

Feature level is a fusion technique that combines features from multiple biometrics. Combining features from multiple biometrics into a combined vector results in a high-dimensional vector, so reduction techniques are needed to select the most useful information (Byahatti & Shettar, 2020). The combined features are then classified to obtain a model that can recognize individuals based on the features that have been combined previously (Sarangi et al., 2022).

The Score level technique calculates the fusion of optimal scores from multiple biometrics (Walia et al., 2019). Scores can be obtained from the prediction results of classification. The scores generated from the classification represent an individual. The scores from each biometric have different value ranges, so they cannot be integrated directly. It is important to normalize the scores first before the fusion process at the score level (Sujana & Reddy, 2021). Decision level is a technique that combines the decisions from two separate biometrics and then combines them to claim the final decision. Several techniques can be used in decision level fusion, including voting, Bayesian, decision, AND, and OR (Cherrat et al., 2020).

Decision level depends on the final decision of each biometric. Several modes can be used at the decision level: parallel, serial, and hierarchical (Sanjekar & Patil, 2019). Serial mode is a one-by-one mode, if the decision of one biometric is accepted, it will be accepted, but if it is rejected, it will be

passed on to the next biometric, so the final decision depends heavily on the last biometric. The parallel mode determines decisions based mostly on decisions from multiple biometrics. If more decisions are accepted than rejected, they will be considered accepted. The hierarchical mode is a combination of serial and parallel modes.

3.2. MTCNN and FaceNet

MTCNN (Multitask Cascaded Convolutional Neural Network) is a face detection method with deep learning which consists of three parts, namely P-Net, R-Net and O-Net (K. Zhang et al., 2016). The first stage, face detection, uses a convolutional neural network to search for areas in the image that may contain faces. The second stage, feature point determination, uses an advanced convolutional neural network to identify key points on the face, such as the eyes, nose, and mouth. The final stage, determining the bounding box, fixes the location and size of the detected face bounding box by utilizing the information from the previous stage.

FaceNet is the name of a face recognition system proposed by Google researchers in 2015. The study (Schroff et al., 2015) used the benchmark datasets Labeled Faces in the Wild (LFW) and the YouTube Face Database. FaceNet extracts facial features into a vector using a deep Convolutional Neural Network (deep CNN) architecture. FaceNet utilizes deep CNN architectures such as ZF-Net and Inception Network. FaceNet takes the image of a person's face as input and produces a 128-dimensional vector as output. In machine learning, this vector is an embedding and is of great importance. All the essential information from an image is embedded in this vector. FaceNet takes a person's face and compresses it into a 128-dimensional vector. The resulting embedding vector can map the similarity between faces with close positions in the embedding space. Embedding is a vector that can interpret points in a Cartesian coordinate system, allowing us to plot faces in the coordinate system using embeddings. The FaceNet architecture employs a deep learning approach called Convolutional Neural Network (CNN) for extracting facial features. This process involves a series of convolutional layers, pooling layers, and batch normalization, iteratively applied to generate more abstract facial features, which are referred to as embeddings. FaceNet utilizes the triplet loss function during the training process. The concept of triplet loss involves three facial images: anchor (reference face), positive (same face as an anchor), and negative (different face from anchor). The goal is to ensure that the distance between the anchor and the positive is smaller than the distance between the anchor and the negative.

3.3. Wav2vec

Wav2vec2.0 is an automatic speech recognition (ASR) model developed by Facebook AI Research (FAIR) in 2020. It is a pre-trained model used to learn representations of speech in the form of vectors (Baevski et al., 2020). The goal is to make speech processing more efficient, especially in speech recognition tasks. The pre-training technique used in the wav2vec model can be applied to various speech-related tasks, such as speech recognition, speech-to-text conversion, and speech transcription. The architecture of wav2vec2.0 is used only to obtain features that will be used for feature level fusion. The extractor aims to capture important features from the audio signal and encode them into a more abstract and complex vector representation. The feature encoder consists of multiple transformation blocks based on the Transformer architecture. These transformation blocks include self-attention layers and feed-forward layers.

The feature encoder of wav2vec 2.0 is a convolutional neural network (CNN) that transforms raw audio signals into feature vectors. The architecture of the feature encoder consists of several convolutional layers along with batch normalization and Rectified Linear Unit (ReLU) activation functions. The output from each block undergoes a pooling layer to reduce the resolution of the feature map. Overall, the feature encoder is designed to extract meaningful features from raw audio signals that can be used for verification purposes (Fan et al., 2020).

4. Proposed Method

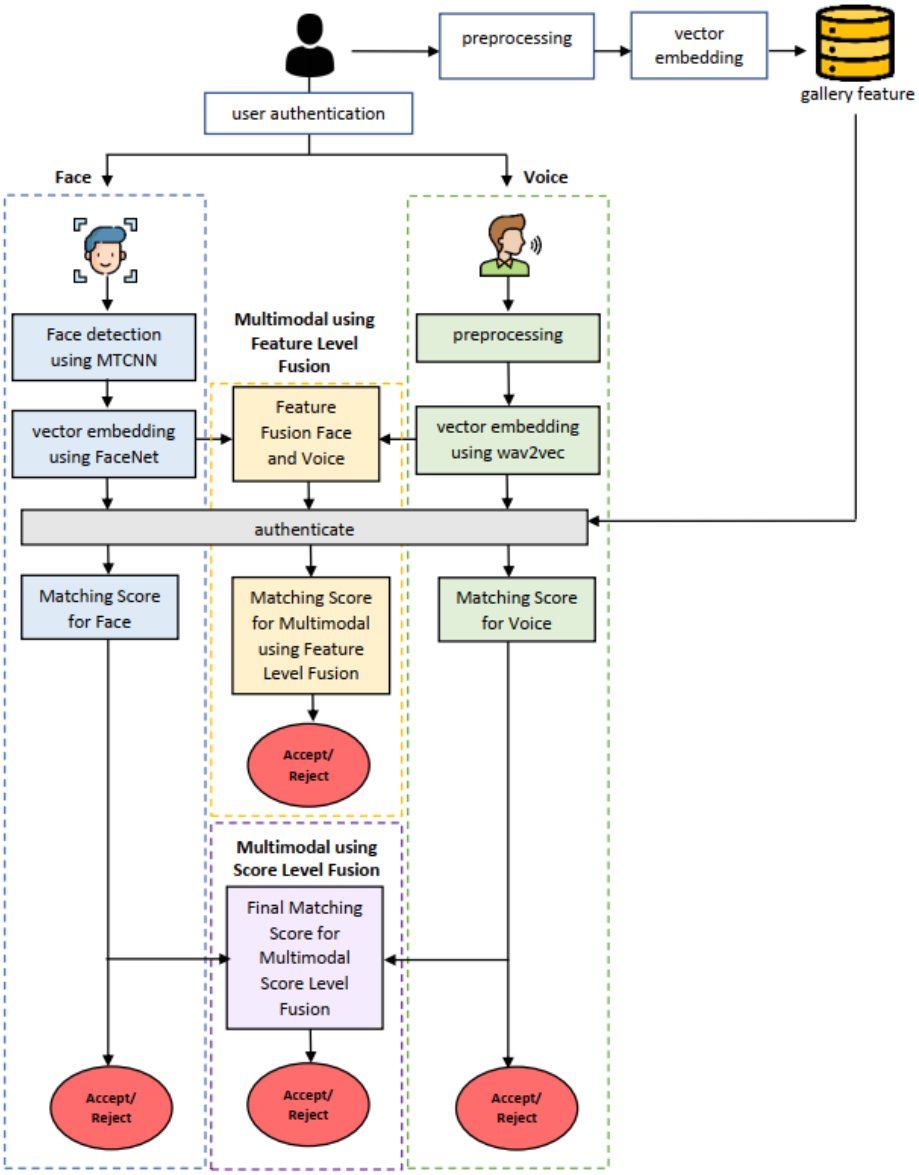

Fig 2: Proposed Method

The methodology used in this study is the system architecture shown in Fig 2. The biometrics used are face and voice. The fusion used in multimodal biometrics is feature level fusion and score level fusion. The evaluation utilized Euclidean Distance and several classification methods such as Decision Tree, Support Vector Machine (SVM), Random Forest, XGBOOST, and Artificial Neural Network (ANN). Besides that, hyperparameter tuning is also carried out on the algorithm to get the most optimal results. The evaluation process calculates the CAR value at FAR 0.1, CAR at FAR 0.01, and the AUC value.

1. Decision Tree

   Decision Tree is a classification method commonly used to recognize or verify individual identities based on their biometric features. It is a model resembling a tree that makes decisions by recursively dividing data into subsets based on the values of input features. The leaf nodes in the Decision Tree represent the corresponding classification labels for the identified individuals. For example, if the decision tree recognizes specific facial features as belonging to individual A, then the label at the leaf node will indicate the identity of individual A. The process of building the decision tree involves selecting the best feature at each node to divide the data. The "best" feature is chosen based on specific criteria. The criterion used in this research is the Gini criterion because it results in maximal

homogeneity or purity of the subsets.

2. SVM
   SVM functions to separate two classes of data by finding an optimal hyperplane (a separating plane) that maximizes the distance between the classes with a maximum margin. The margin is the distance between the hyperplane and the closest support vectors, which are data points from each class that are closest to the hyperplane. By seeking the hyperplane with the maximum margin, SVM can more easily predict different target data points since the larger the distance between the different classes, the higher the likelihood that data points from each class will be well-separated and distinct. This separation makes the model more robust and generalizable to new data. This research uses the "C" parameter, which typically takes values between 1 and 10, to control the trade-off between maximizing the margin and minimizing the classification errors. The “rbf” (radial basis function) kernel allows SVM to handle non-linearly separable data by transforming the data into a higher-dimensional space where a linear hyperplane can effectively separate the classes. SVM with RBF kernel and value “C” produces the optimal classification algorithm in this study.
3. Random Forest
   Random Forest is an extension of the Decision Tree algorithm and operates by building multiple decision trees during the training phase. Random Forest takes a data point input into the model and runs it through each pre-determined number of decision trees that have been created. Each tree then makes its own class label prediction for the data point, and then all the predictions made by the trees are "voted" to see which class is the most common, which is then output as the result. The optimal number of trees in this research was found to be between 50 to 100, and two criteria, “gini” and “entropy”, have been chosen. Gini measures the impurity or disorder of a node, while entropy measures the information gained by a split. Using Random Forest helps to mitigate the overfitting problem often encountered with individual decision trees and provides more robust and accurate predictions by aggregating the results from multiple trees.
4. XGBOOST
   This algorithm is based on ensemble learning techniques, specifically gradient boosting, which aims to combine predictions from several weak learner models to create a stronger model. XGBOOST utilizes multiple decision trees known as "CART" (Classification and Regression Trees) as its weak learners. The training process of XGBOOST is performed iteratively, where each iteration focuses on improving the prediction errors from the previous iteration. This is achieved by assigning higher weights to data points that were mis predicted, allowing the subsequent models to focus on the more challenging cases. The training process of XGBOOST continues for a specified number of iterations, and in this research, the optimal number of iterations found was between 100 to 200. By iteratively updating the model and refining its predictions, XGBOOST can gradually improve its performance and achieve higher accuracy.
5. ANN
   ANN is a computational model inspired by the structure and function of neural networks in the human brain. It is used for data separation and classification into different classes. The ANN model consists of layers of interconnected neurons or nodes. The first layer is called the input layer, which receives data as input features. The last layer is the output layer, which generates predictions or classification labels. In addition to the input and output layers, the ANN model can also have hidden layers. Hidden layers act as feature processors and enable the model to discover more complex patterns in the data. The number of hidden layers and the number of neurons in each layer are parameters that can be adjusted during the training process. In this research, the ANN structure consists of an input layer, 2 hidden layers, and an output layer. The input layer has the same number of neurons as the biometric features used. The hidden layers are responsible for extracting features and learning patterns from biometric data. They use the ReLU activation function, which helps introduce non-linearity and allows the model to capture more complex relationships in the data. The

output layer is the last layer with a single output neuron that produces the prediction. It uses the Sigmoid activation function, which squashes the output value between 0 and 1, representing the probability of the data belonging to a certain class in binary classification problems.

Table 1: Hyperparatamer Tuning

Model	Tuned Parameters	Value	Techniques
Decision Tree	criterion	gini, entropy	Grid Search
SVM	C	1, 5, 10	Grid Search
	kernel	linear, rbf, poly, sigmoid	
Random Forest	n_estimators	50, 100, 150, 200	Grid Search
	criterion	gini, entropy	
XGBoost	n_estimators	50, 100, 150, 200	Grid Search

5. Experiments

5.1. Dataset

The MSU-AVIS dataset was collected by Michigan State University (Chowdhury et al., 2018), is a valuable resource for research in multimodal biometrics. The dataset has a total of 50 subjects, with 16 of them being females and 34 males. The data was collected using a Logitech C920 HD Pro webcam equipped with dual built-in stereo microphones that are used to record the audio in the room. It is positioned 240 cm from the ground. The facial data records subjects looking straight ahead at the camera, then turning their head to one side and back again. Meanwhile, the voice data records subjects speaking randomly selected short scripts. Therefore, this study is considered text independent. This dataset also presents challenges in the identification process. When acquiring some videos in the dataset, the internal microphone was replaced with a lower-quality microphone that captured surrounding noise at a higher rate. Approximately 30% of the data records some background noise. There is also significant variation in the subjects' poses relative to the camera, with some subjects spending extended periods with the back of their heads facing the camera rather than facing the camera directly.

Before we divided the dataset for experiment, we must split the data. The facial data quality is 1920 x 1080 pixels, and the audio sampling rate is 48kHz.The facial videos have an average duration of 3-11 seconds. It is cropped at a frame rate of 0.1 seconds, resulting in 40 facial images per subject. Meanwhile, the voice videos have an average time of 10-26 seconds. It is split with overlapping intervals ranging from 0.128 to 0.53 seconds, depending on the video length, resulting in 40 video data files containing audio with a duration of 5 seconds per subject. The facial dataset consists of 40 files in .jpg format per subject, and the voice dataset consists of 40 files in .mp4 format per subject.

Table 2: Data Experiment

Data	Positive Sample	Negative Sample	Total
Training	1200	1200	2400
Validation	400	400	800
Testing	400	400	800

For experiment, the researchers divided the dataset into three parts: 60% training data, 20% validation data, and 20% testing data. So, each subject has 24 training data, 8 validation data, and 8 testing data. The dataset contains distance vector data. Positive samples are labelled 1, representing the distance between the data and the average of the data itself. Meanwhile, negative samples are labelled 0 which represents the distance between the data and the average data of other individuals. Each subject

has 24 training data, so there are 24 positive data samples. For 50 subjects, the number of positive samples in training data is 1200 data. For negative samples, the distance between the data and the average training data of other subjects is calculated. The data was sorted, and 24 pieces of data were taken that had the smallest distance. So, the negative samples in the training data for 50 subjects are 1200 data. In validation data, each subject has 8 data. For 50 subjects, the number of positive samples in the validation data is 400. For negative samples, the distance between the data and the average training data of other subjects is calculated. The data was sorted, and 8 pieces of data were taken that had the smallest distance. So, there are 400 negative samples in the validation data for 50 subjects. The process for testing data is the same as validation data. Each training, validation and testing data is randomized so that the data labels are not sequential. In summary, training data teaches the model, validation data helps optimize it, and testing data assesses its real-world performance. These three datasets are essential components in the development and evaluation of machine learning models, ensuring their effectiveness and generalization ability.

5.2. Experimental Design

5.2.1. Biometric Face

Face detection using the deep learning MTCNN. The FaceNet model performs the embedding. The output of the embedding process is a vector with a size of 128. The verification process utilizes a machine learning model by adjusting the threshold to determine accept and reject decisions. The results are used to construct the ROC curve for performance evaluation. The biometric face process has been illustrated in Fig 3.

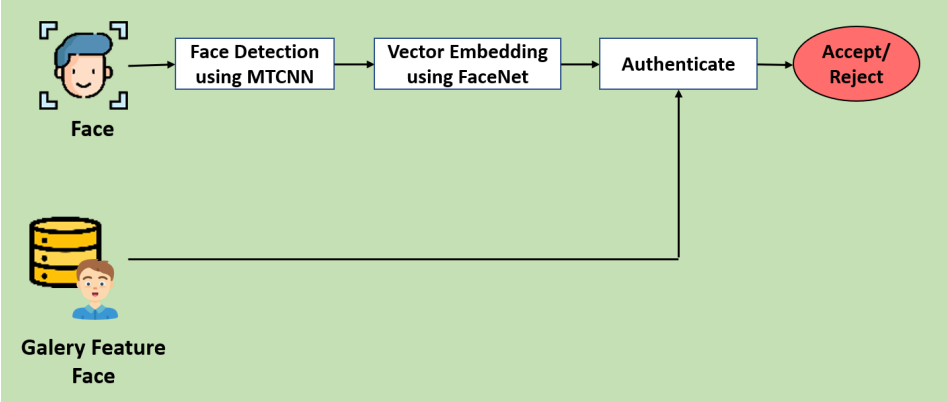

Fig 3: Biometric Face Architecture

The implementation was carried out using the MSU-AVIS dataset of 50 subjects. Preprocessing on facial data starts from 1920x1080 pixel data converted to RGB. MTCNN was used to detect faces and obtain the bounding box, confidence, and key points, which include the coordinates of the left eye, right eye, nose, mouth left, and mouth right. The data was then reshaped to 160x160 pixels, as shown in Fig 4. The 160x160 images were converted into a face array and then embedded using the FaceNet model. The embedding process resulted in a 1x128 vector.

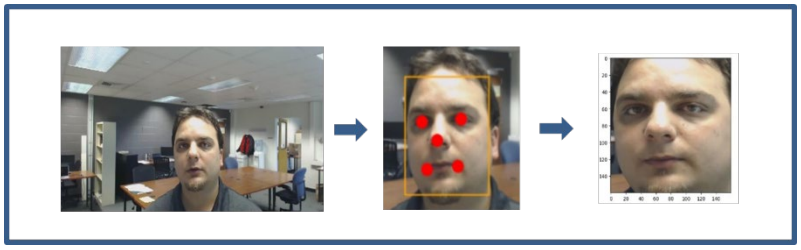

Fig 4: Face Detection using MTCNN

FaceNet consistently shows state-of-the-art performance in face recognition tasks. Its deep neural network is designed to learn highly discriminative facial features, making it an attractive choice for our facial biometric tasks. FaceNet extracts facial features into a vector using a deep Convolutional Neural Network (deep CNN) architecture. FaceNet utilizes deep CNN architectures such as ZF-Net and Inception Network. FaceNet produces fixed-length embeddings for each face, ensuring that regardless of the size or orientation of the input face image, we obtain a consistent representation. FaceNet takes the image of a person's face as input and produces a 128-dimensional vector as output that representation face inportance information. This is important for effective fusion in multimodal biometrics. In FaceNet, triplet loss is used as a special function among other loss functions. The Triplet Loss decreases the distance between an anchor and a positive input when both of which have the same identity, and increases the distance between the anchor and a negative input for the different identities. FaceNet is open source, widely adopted, and has a robust implementation, making it accessible to researchers.

The authenticate process uses euclidean distance and a machine learning: decision tree, SVM, random forest, xgboost, and ANN. We conducted a search over various hyperparameters. We provide a detailed explanation of this process along with the specific parameters tuned for each classifier as show in Table 1. We employed a grid search technique, which systematically explores a predefined range of hyperparameters to identify the best combination. It demonstrates a systematic approach to fine-tuning the models for optimal performance.

5.2.2. Biometric Voice

Preprocessing on voice data involves converting data with the mp4 file extension into wav format and then transforming the wav data into a waveform. After that, voice embedding is performed using the pre-trained wav2vec2.0 model to obtain a voice vector. The authenticate process uses a machine learning model by adjusting the threshold to determine the accept and reject decisions used to build the ROC curve as a performance evaluation. The biometric voice process has been illustrated in Fig 5.

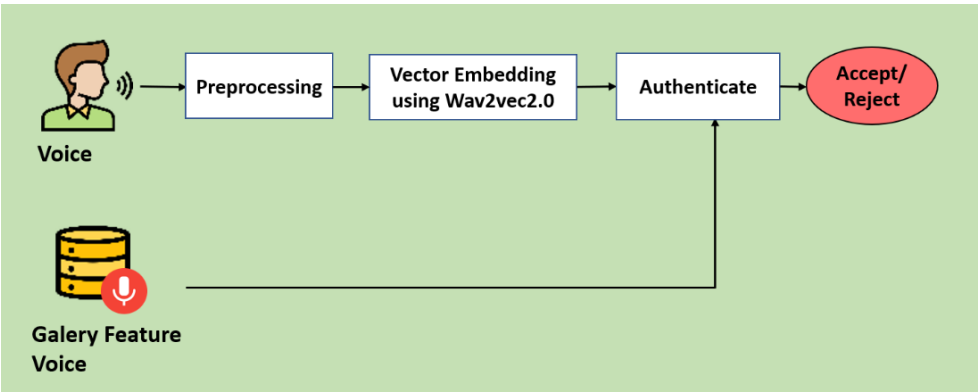

Fig 5: Biometric Voice Architecture

The MSU-AVIS dataset consists of audio videos in mp4 format with durations ranging from 10-26 seconds. The data was split with overlapping distances between 0.128-0.53 seconds, depending on the original audio duration, resulting in 40 audio files, with each file having a duration of 5 seconds.

Preprocessing on voice data starts from converted the mp4 audio data to wav format. The audio waveforms were then converted into arrays using the touch audio library with a sampling rate of 16000 Hz, resulting in 2 channels. To convert the two channels into a single channel, the torch mean function was used to average the channels, as shown in Fig 6.

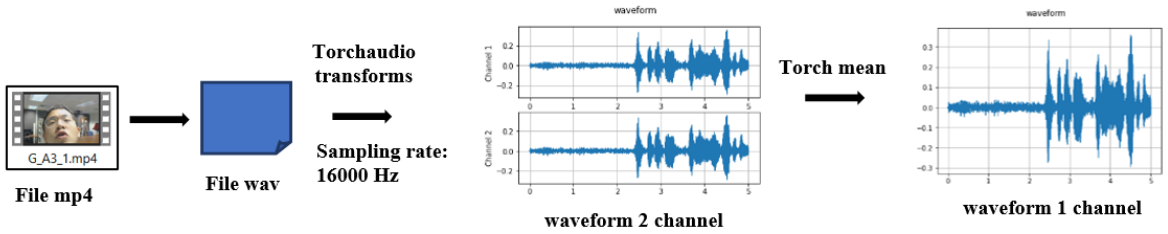

Fig 6: Voice Preprocessing

Wav2vec excels in feature extraction from audio data. Its self-supervised learning approach allows it to learn rich representations from voice samples, which is particularly beneficial when labeled voice data is limited. Wav2vec's ability to capture contextual information from audio signals makes it robust to background noise, reverberations, and variations in pronunciation. This aligns well with the challenges encountered in real-world voice recognition scenarios. The feature encoder of wav2vec 2.0 is a convolutional neural network (CNN) that transforms raw audio signals into feature vectors. The architecture of the feature encoder consists of several convolutional layers along with batch normalization and Rectified Linear Unit (ReLU) activation functions. The output from each block undergoes a pooling layer to reduce the resolution of the feature map. Overall, the feature encoder is designed to extract meaningful features from raw audio signals that can be used for verification purposes.

The authenticate process uses euclidean distance and a machine learning: decision tree, SVM, random forest, xgboost, and ANN. We conducted a search over various hyperparameters. We provide a detailed explanation of this process along with the specific parameters tuned for each classifier as show in Table 1. We employed a grid search technique, which systematically explores a predefined range of hyperparameters to identify the best combination. It demonstrates a systematic approach to fine-tuning the models for optimal performance.

5.2.3. Multimodal Biometric using Feature Level Fusion

The biometric fusion is performed by comparing two levels: feature level and score level. At the feature level, as shown in Fig 7, the merging method involves concatenating the face vector and the voice vector. The face vector has a size of 1x128, and the voice vector has a size of 1x128. When concatenated, the resulting vector has a length of 1x256.

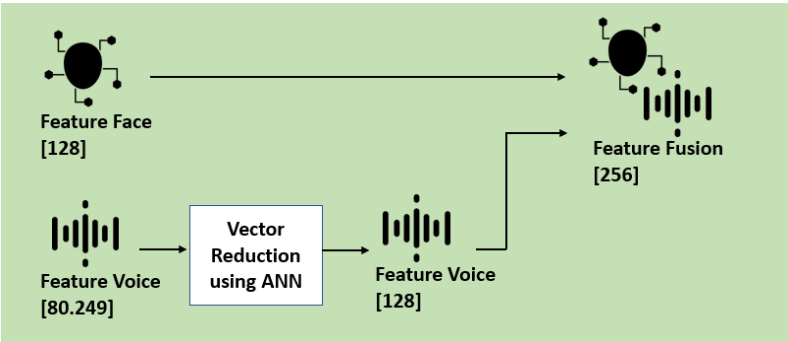

Fig 7: Feature Fusion Process

The multimodal feature level fusion is implemented by combining the face vector and the voice

vector. An adjustment must be made since the voice vector has a different length than the face vector. The voice vector will dominate the classification results if this adjustment is not made. The adjustment is made by reducing the vector dimension using an ANN. The ANN architecture used for vector reduction is shown in Fig 8. The input vector has a length of 80249 and is linearly reduced to an output vector with a length of 4096 using the ReLU activation function. A dropout layer is added during training. The output vector is then linearly reduced to an output of 512, which becomes the input for the next layer, resulting in an output vector with a length of 128. After the face and voice vectors have been adjusted to have the same length, they are combined using the numpy concatenate function. The resulting vector has a length of 256. The combined vector is then used for next process.

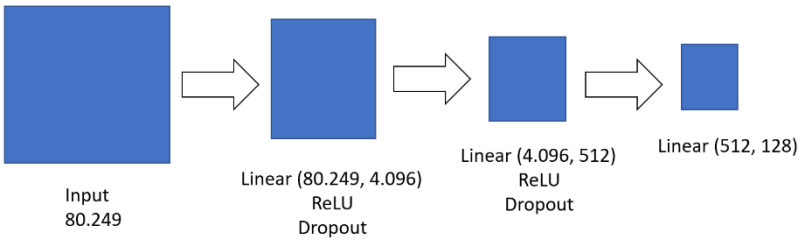

Fig 8: ANN Architecture

The next process is authentication. The authenticate process uses euclidean distance and a machine learning: decision tree, SVM, random forest, xgboost, and ANN. We conducted a search over various hyperparameters. We provide a detailed explanation of this process along with the specific parameters tuned for each classifier as show in Table 1. We employed a grid search technique, which systematically explores a predefined range of hyperparameters to identify the best combination. It demonstrates a systematic approach to fine-tuning the models for optimal performance.

5.2.4. Multimodal Biometric using Score Level Fusion

Implementation of multimodal score level fusion by combining the best face and voice classification results. It is important to note that the ANN model achieves the best performance. Therefore, in score-level fusion, we calculate final score from best score face using ANN and best score voice using ANN. The evaluation process uses two scenarios, namely calculating the average score of each biometric, and FLD (Fisher Linear Discriminant). FLD utilizes AUC in calculating its weight. The weights of the first biometric are calculated using the formula (1), which is obtained by dividing the AUC of the first biometric by the total AUC of both the first and second biometrics. Similarly, the weights of the second biometric are calculated using the formula (2), which is obtained by dividing the AUC of the second biometric by the total AUC of both the first and second biometrics. The final score is obtained by multiplying each biometric score by its respective weight and then dividing by the total weight, as seen in formula (3). This final score is then used to decide based on a threshold.

$$W_1 = \frac{AUC_1}{AUC_1 + AUC_2} \quad (1)$$

$$W_2 = \frac{AUC_2}{AUC_1 + AUC_2} \quad (2)$$

$$Score = \frac{W_1 S_1 + W_2 S_2}{W_1 + W_2} \quad (3)$$

5.2.5. Performance Measures

In this research, the performance of the biometric system is measured using the following parameters:

- CAR (Correct Acceptance Rate): It measures the percentage of genuine users who are correctly accepted by the system. CAR is calculated at different FAR (False Acceptance Rate) levels, such as FAR 0.1 and FAR 0.01. CAR at FAR 0.1 represents the system's ability to correctly accept genuine users while keeping the false acceptance rate at 0.1%. Similarly, CAR at FAR 0.01 represents the system's performance at a stricter threshold.
- AUC (Area Under the ROC Curve): It is a measure of the overall performance of the biometric system. The ROC (Receiver Operating Characteristic) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different decision thresholds. AUC represents the area under this curve and provides an aggregate measure of the system's performance across all possible decision thresholds.

By evaluating the CAR at different FAR levels and the AUC value, researchers can assess the effectiveness and robustness of the biometric system in terms of accuracy.

5.3. Experimental Results and Discussion

The methods used in this research are face biometrics, voice biometrics, multimodal biometrics using feature-level fusion, and multimodal biometrics using score-level fusion. In facial biometrics, 6 scenarios were conducted on validation data and testing data, as show in Table 2. The evaluation was based on the CAR at FAR 0.1 and FAR 0.01 and the area under the curve (AUC). The experiment results on the validation and testing data showed that the best model was achieved using ANN, which resulted in a CAR value of 0.9975 at FAR 0.1 and FAR 001, as well as an AUC value of 0.999.

Table 3: Experiment Result Face Only

Model	Validation			Testing		
	CAR at FAR 0.1	CAR at FAR 0.01	AUC	CAR at FAR 0.1	CAR at FAR 0.01	AUC
Face Only + Euclidean Distance	0.932	0.102	0.948	0.917	0.027	0.951
Face Only + Decision Tree	0.945	0.000	0.939	0.952	0.000	0.951
Face Only + SVM	0.997	0.997	0.998	0.992	0.992	0.996
Face Only + Random Forest	0.985	0.985	0.991	0.985	0.985	0.992
Face Only + XGBOOST	0.990	0.990	0.993	0.985	0.985	0.992
Face Only + ANN	0.997	0.997	0.999	0.997	0.997	0.999

In biometric voice, the experimental results are shown in Table 3. The best model was achieved using ANN. On the validation data, the model resulted in a CAR of 0.9275 at FAR 0.1, a CAR value of 0.83 at FAR 0.01, and an AUC value of 0.949. The model resulted in a CAR value of 0.9175 at FAR 0.1 and an AUC value of 0.952 on the testing data.

Table 4: Experiment Result Voice Only

Model	Validation			Testing		
	CAR at FAR 0.1	CAR at FAR 0.01	AUC	CAR at FAR 0.1	CAR at FAR 0.01	AUC
Voice Only + Euclidean Distance	0.980	0.020	0.938	0.795	0.020	0.898
Voice Only + Decision Tree	0.840	0.000	0.840	0.887	0.000	0.856
Voice Only + SVM	0.872	0.872	0.930	0.917	0.000	0.839
Voice Only + Random Forest	0.865	0.865	0.926	0.897	0.922	0.938
Voice Only + XGBOOST	0.890	0.890	0.936	0.910	0.000	0.932
Voice Only + ANN	0.927	0.833	0.949	0.917	0.000	0.952

Meanwhile, in multimodal biometric using feature-level fusion, the best model was achieved using XGBOOST, as shown in Table 4. On the validation data, the model resulted in a CAR of 0.9925 at FAR 0.1 and FAR 0.01, and an AUC value of 0.996. On the testing data, the model resulted in a CAR of 0.995 at FAR 0.1 and FAR 001, and an AUC value of 0.996.

Table 5: Experiment Result Feature Level Fusion

Model	Validation			Testing		
	CAR at FAR 0.1	CAR at FAR 0.01	AUC	CAR at FAR 0.1	CAR at FAR 0.01	AUC
Feature Level + Euclidean Distance	0.980	0.020	0.938	0.795	0.020	0.895
Feature Level + Decision Tree	0.940	0.000	0.941	0.955	0.000	0.945
Feature Level + SVM	0.905	0.905	0.944	0.905	0.000	0.941
Feature Level + Random Forest	0.987	0.987	0.993	0.980	0.980	0.989
Feature Level + XGBOOST	0.992	0.992	0.996	0.995	0.995	0.996
Feature Level + ANN	0.990	0.977	0.993	0.990	0.912	0.992

In multimodal biometric using score-level fusion use 2 scenario. Table 5 shows that the FLD model provides the best results, achieving a CAR value of 0.997 at FAR 0.1 and FAR 0.01, as well as an AUC value of 0.998, on the validation data. Similarly, on the testing data, it achieves a CAR value of 0.992 at FAR 0.1 and FAR 0.01, with an AUC value of 0.996.

Table 6: Experiment Result Score Level Fusion

Model	Validation			Testing		
	CAR at FAR 0.1	CAR at FAR 0.01	AUC	CAR at FAR 0.1	CAR at FAR 0.01	AUC
Score Level +Average	0.967	0.967	0.983	0.972	0.972	0.985
Score Level + FLD	0.997	0.997	0.998	0.992	0.992	0.996

This research provides several interesting insights. Feature-level fusion is a fusion technique that allows various fusion methods such as merging, averaging, weighted combinations, etc., to effectively integrate information from different modalities. In this research, the performance of feature-level fusion is 0.003 better than score-level fusion for CAR at FAR 0.1 and CAR at FAR 0.01. Score-level fusion involves combining scores from each modality, and the results depend on the classifier model. Therefore, in score-level fusion, multiple classifier experiments are needed to obtain the best-performing model. In contrast, feature-level fusion, is more model-independent, as it directly combines feature vectors without intermediate classification steps. Among the five classifiers used, XGBoost achieved the best result with an AUC value of 0.996.

This research demonstrates that feature-level fusion outperforms score-level fusion. Feature-level fusion combines raw feature vectors from each modality, preserving all unique characteristics of each modality. This ensures that no information is lost during fusion. Feature-level fusion at the feature level can help in filtering out noise and errors that may occur in one modality but not in another. However, it can also affect the performance of another modality when combined with a modality that performs poorly. Averaging or merging features can reduce noisy data, thus providing more reliable results. To further enhance performance, dimension reduction techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), or Artificial Neural Network (ANN) can be explored. These techniques can reduce the complexity of the fusion process.

6. Conclusion and Future Work

This study demonstrates the potential of deep learning for robust multimodal biometric authentication. Deep learning models FaceNet and wav2vec for face and voice feature extraction. The biometrics used are face and voice, and the fusion method used multimodal is feature level and score level. Classification is done using five machine learning algorithms, including hyperparameter tuning. In addition, Euclidean distance is also used as a comparison. There are 20 models successfully built in this experiment. The proposed feature-level fusion with XGBoost achieves 0.995 CAR at 0.01 FAR, significantly outperforming score-level fusion. The research provides insights into effective fusion strategies for multimodal biometrics.

Research on multimodal biometrics is exciting and extensive. Future research should explore additional modalities with different combinations of biometrics, advanced fusion techniques to achieve

better reliability and accuracy. Additionally, research focusing on performance optimization specific to mobile devices: the speed of verification, memory requirements, and security of biometric data, such as utilizing blockchain for storing biometric data. Overall, this work advances multimodal biometric authentication through novel application of deep learning models and fusion methods.

References

Arshad, S. R., Mujtaba Haider, S., & Mughal, A. B. (2022). An Approach for Identification of Speaker using Deep Learning. In *IJAIMS International Journal of Artificial Intelligence and Mathematical Sciences* (Vol. 01).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460. https://github.com/pytorch/fairseq

Baghelp, S., Sahup, K., Kshitiz Varmap, P., & Assistant, P. (2017). Multimodal Biometric System Advantages over Unimodal Biometric System Authentication Technology. *IJISET-International Journal of Innovative Science, Engineering & Technology*, *4*. www.ijiset.com

Byahatti, P., & Shettar, M. S. (2020). Fusion Strategies for Multimodal Biometric System Using Face and Voice Cues. *IOP Conference Series: Materials Science and Engineering*, *925*(1). https://doi.org/10.1088/1757-899X/925/1/012031

Cahyono, F., Wirawan, W., & Fuad Rachmadi, R. (2020). Face recognition system using facenet algorithm for employee presence. *4th International Conference on Vocational Education and Training, ICOVET 2020*, 57–62. https://doi.org/10.1109/ICOVET50258.2020.9229888

Cherrat, E. M., Alaoui, R., & Bouzahir, H. (2020). A multimodal biometric identification system based on cascade advanced of fingerprint, fingervein and face images. *Indonesian Journal of Electrical Engineering and Computer Science*, *17*(3), 1562–1570. https://doi.org/10.11591/ijeecs.v18.i1.pp1562-1570

Chowdhury, A., Atoum, Y., Tran, L., Liu, X., & Ross, A. (2018). MSU-AVIS dataset: Fusing Face and Voice Modalities for Biometric Recognition in Indoor Surveillance Videos. *Proceedings - International Conference on Pattern Recognition*, *2018-August*, 3567–3573. https://doi.org/10.1109/ICPR.2018.8545260

Dargan, S., & Kumar, M. (2020). A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Systems with Applications*, *143*, 113114. https://doi.org/10.1016/J.ESWA.2019.113114

Dhole, S. A., Patil, V. H., Vidyapeeth COEW Pune, B., & Vidyapeeth, B. (2012). Review of Multimodal Biometric Identification Using Hand Feature and Face. *Buletin Teknik Elektro Dan Informatika (Bulletin of Electrical Engineering and Informatics)*, *1*(3), 179–184.

Fan, Z., Li, M., Zhou, S., & Xu, B. (2020). Exploring wav2vec 2.0 on speaker verification and language identification. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2*, 896–900. https://doi.org/10.21437/Interspeech.2021-1280

Farik, M., Lal, N. A., & Prasad, S. (2016). A Review Of Authentication Methods. *Article in International Journal of Scientific & Technology Research*, *5*(11). www.ijstr.org

Grand View Research. (2023). *Behavioral Biometrics Market Size & Share Report, 2020-2027*. https://www.grandviewresearch.com/industry-analysis/behavioral-biometrics-market#

Gudavalli, M., Viswanadha Raju, S., Vinaya Babu, A., & Srinivasa Kumar, D. (2012). Multimodal biometrics - Sources, architecture and fusion techniques: An overview. *Proceedings - 2012 International Symposium on Biometrics and Security Technologies, ISBAST 2012*, 27–34. https://doi.org/10.1109/ISBAST.2012.24

Nyein, T., & Oo, A. N. (2019). University Classroom Attendance System Using FaceNet and Support Vector Machine. *2019 International Conference on Advanced Information Technologies, ICAIT 2019*, 171–176. https://doi.org/10.1109/AITC.2019.8921316

Oloyede, M. O., & Hancke, G. P. (2016). Unimodal and Multimodal Biometric Sensing Systems: A Review. *IEEE Access*, *4*, 7532–7555. https://doi.org/10.1109/ACCESS.2016.2614720

Saleh, M., & Jouny, I. (2022). Multimodal Person Identification through the Fusion of Face and Voice Biometrics. *2022 17th Annual System of Systems Engineering Conference, SOSE 2022*, 164–169. https://doi.org/10.1109/SOSE55472.2022.9812670

Sanjekar, P. S., & Patil, J. B. (2019). Multimodal biometrics with serial, parallel and hierarchical mode at decision level fusion. *Indonesian Journal of Electrical Engineering and Computer Science*, *16*(3), 1303–1310. https://doi.org/10.11591/ijeecs.v16.i3.pp1303-1310

Sarangi, P. P., Nayak, D. R., Panda, M., & Majhi, B. (2022). A feature-level fusion based improved multimodal biometric recognition system using ear and profile face. *Journal of Ambient Intelligence and Humanized Computing*, *13*(4), 1867–1898. https://doi.org/10.1007/S12652-021-02952-0/METRICS

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June-2015*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Srivastava, N. (2007). Fusion Levels in Multimodal Biometric Systems–A Review. *International Journal of Innovative Research in Science, Engineering and Technology (An ISO*, *3297*(5). https://doi.org/10.15680/IJIRSET.2017.0605266

Sujana, S., & Reddy, V. S. K. (2021). Comparison of levels and fusion approaches for multimodal biometrics. *Indonesian Journal of Electrical Engineering and Computer Science*, *23*(2), 791–801. https://doi.org/10.11591/ijeecs.v23.i2.pp791-801

Walia, G. S., Singh, T., Singh, K., & Verma, N. (2019). Robust multimodal biometric system based on optimal score level fusion model. *Expert Systems with Applications*, *116*, 364–376. https://doi.org/10.1016/J.ESWA.2018.08.036

William, I., Ignatius Moses Setiadi, D. R., Rachmawanto, E. H., Santoso, H. A., & Sari, C. A. (2019). Face Recognition using FaceNet (Survey, Performance Test, and Comparison). *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*. https://doi.org/10.1109/ICIC47613.2019.8985786

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

Zhang, X., Cheng, D., Jia, P., Dai, Y., & Xu, X. (2020). An Efficient Android-Based Multimodal Biometric Authentication System with Face and Voice. *IEEE Access*, *8*, 102757–102772. https://doi.org/10.1109/ACCESS.2020.2999115

Zhang, X., Dai, Y., & Xu, X. (2017). Android-based multimodal biometric identification system using feature level fusion. *2017 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2017 - Proceedings*, *2018-January*, 120–124. https://doi.org/10.1109/ISPACS.2017.8266457

Zhang, X., Xiong, Q., Dai, Y., & Xu, X. (2018). Voice biometric identity authentication system based on android smart phone. *2018 IEEE 4th International Conference on Computer and Communications, ICCC 2018*, 1440–1444. https://doi.org/10.1109/COMPCOMM.2018.8780990