INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Deep Neural Spectral Subtraction Centroid Mel Frequency Powered Secure Multimodal Biometric Authentication

S. Preethi[1]*, **P. Joseph Charles**[2]*

1 Research Scholar, Department of computer science, Affiliated to Bharathidasan University, Tiruchirappalli - 620002
2 Assistant Professor, Department of Computer Science, Affiliated Bharathidasan University, Tiruchirappalli - 620002

OPEN ACCESS

Received: 02/09/2025
Accepted: 21/09/2025
Published: 16/11/2025

Citation: Preethi S, Joseph Charles P (2025) Deep Neural Spectral Subtraction Centroid Mel Frequency Powered Secure Multimodal Biometric Authentication. Indian Journal of Science and Technology 18(37): 3005-3023. https://doi.org/10.17485/IJST/v18i37.1537

* **Corresponding authors**.

preethidiwakarmphil@gmail.com

hedrpjcharles@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To propose a Deep Neural Spectral Subtraction Centroid Mel Frequency Spatial-Temporal (DN-SCMFST) framework for robust multimodal biometric authentication (face and voice). **Method:** The system uses the Speaking Faces dataset (142 subjects, 13,000 samples). Preprocessing is performed with a Kushner–Stratonovich filter and spectral subtraction. Features are extracted via spectral centroid and spatio-temporal descriptors, followed by score-level fusion. **Findings:** DN-SCMFST achieved accuracy improvements of 5–18%, PSNR gains of 18–28%, FAR reduction of 20–37%, FRR reduction of 18–28%, and recognition time reduction of 15–27% compared with CNN+RNN and DL-MBA models. **Novelty:** Integrates deep neural spectral subtraction with centroid-based mel frequency and temporal-spatial fusion, offering resilience against noise and improved real-time recognition efficiency.

Keywords: Multimodal biometric authentication; Deep neural network; Spectral subtraction; Spectral centroid; Mel frequency; Score-level fusion; Kushner–Stratonovich

1 Introduction

In the emergence of a progressively digital and associated world, both the personal and organization data security has jumped up to the cutting edge of technological revolution and research. In light of this, biometric authentication systems, which grip unique characteristics for identity verification, have made an appearance as a demanding element in boosting security infrastructures. In spite of their heightening ubiquity and significance, conventional biometric systems frequently struggle with issues concerning privacy, accuracy and hence found to be vulnerable to numerous attacks.

In [(1)], a novel multi-modal biometric authentication system that combines facial, vocal, and signature data to boost security measures using a fusion of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) was proposed. The method framework specifically integrated dual shared layers for extensive extraction of feature. Also, the method underwent training employing a joint loss function with

the intent of optimizing for accuracy across different biometric inputs. Moreover, feature-level fusion employing Principal Component Analysis (PCA) and classification employing Gradient Boosting Machines (GBM) were also used with the objective of refining the authentication process. This in turn resulted in significant improvements of authentication accuracy and minimizing processing time and resource utilization. Nevertheless, the method cannot entirely make certain to meet out with privacy as it necessitates cross-data learning that affects security. To address these aspects in this work, Deep Neural-powered Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction has been applied in addition to Dual Aggregated Score Level Fusion that aids in ensuring security, therefore reducing false acceptance and false rejection rate.

The notable growth in online activities on account of contemporary global events has emphasized the significance of biometric person authentication on digital platforms. Even though several biometric devices may be utilized for accuracy biometric authentication, obtaining necessary technology can be insanely high-prices and logistically demanding. To address these aspects, a deep learning powered multimodal biometric authentication to craft a robust multi-classification method was proposed in [2] resulting in exceptional performance improvements in biometric authentication.

The method also reinforced the prospective of combining both static and dynamic biometric features, therefore resulting in high precision and accuracy. However, the recognition time and accuracy were not focused. To concentrate on these two aspects, Deep Neural-powered Kushner–Stratonovich and Spectral Subtraction Non-linear filter-based Pre-processing is applied that aids in not only reduction time but also by preserving edges helps in improving overall recognition accuracy.

Technology Acceptance Model (TAM) was introduced in [3] to focus on two distinct aspects namely, privacy and perceived trust to examine the user acceptance of biometric identification in FinTech application. Here, analytic hierarchy process (AHP) was used to define the objects and also discovered biometric identification techniques that symbolize the objects and criteria. The face and voice recognition were most preferred identification techniques in FinTech applications. However, the biometrics failed to adjust the weights with corresponding conditions. The findings did not allow the readers to understand how users view biometrics in FinTech applications.

A spiking neural network (SNN) approach was introduced in [4] for predicting emotional states based on facial expression and speech data. The designed approach was found to be robust to noise where it attained higher accuracy for facial expression recognition (FER). But the designed approach was not found to be used for continuous multi-modal emotion recognition and also did meet the accuracy at required level.

A new multimodal emotion recognition approach was introduced in [5] to increase the performance of BERT model for emotion recognition by combining heterogeneous features based on language, audio, and visual modalities. Due to the involvement of heterogeneous features of audio and visual modalities, it resulted in improved accuracy. Here two different modules employing Self-Multi-Attention Fusion module and Multi-Attention fusion module were integrated using transformer architecture with the objective of joining fine-grained representation of audio and visual features for fine-tuning. However, the designed approach increased the computation cost because of the trainable weights and hyperparameters.

An automated integrated speech signal and facial image analysis system was designed in [6] for human emotion identification. Statistical and model multi-classification analysis was carried out to choose the features to distinguish human emotions. The selected features were combined with age and gender to construct learning based classifiers for increasing accuracy. In addition, iris together with speech and face information was investigated to join additional information and integrate information from brain activity with improved accuracy.

Over the recent few years, image recognition technique employing deep learning (DL) has improved notably using biometric information like, iris scans, face recognition and so on. To be specific, user authentication methods that use face recognition have been researched over the last few decades. In [7], a biometric authentication system using chest wall displacement recorded by a wearable electromagnetic near-field sensor (NFS) was proposed using [8] dataset. This mechanism was designed with the objective of selecting an optimal classification mechanism for performing authentication on the basis of user states. Despite accuracy factor being focused, the average speed involved was not included. A real time verification system was proposed in [9] with stored user data features with improved face detection accuracy and minimal average speed, therefore ensuring smooth and robust authentication. In spite of making certain that robust authentication is ensured, however the error factor was not analyzed. Neural fuzzy extractor was employed in [10] based on trigonometric correlations that in turn minimized error probability to a greater extent.

Verification of identity poses a pivotal role in contemporary society, with applications ranging from online services to security systems. With the requirement for robust automatic authentication increasing, several software, hardware and biometrics methods have been developed. Amidst these, biometric modalities have received notable awareness owing to high accuracy and minimal falsification. In [11], electrocardiogram (ECG) signals were used for verification identity with improved accuracy. Despite several biometric devices that may be utilized for accurate authentication, acquiring the necessary technology are both found to be expensive and logistically demanding. An innovative approach by combining static and dynamic signature data

with facial data was proposed in [12] with a high-performance recognition framework. Human activity was recorded in [13] for user authentication employing deep features fusion with improved performance.

An improved technology involving authentication and securing files using face recognition, voice and image grid pattern selection elements were proposed in [14] ensuring multi-model security. Despite improvement in security, however, the false rejection rate was not focused, therefore compromising security. Yet another method to ensure security employing deep residual neural network incorporating both frequency and time characteristics with exceptional classification accuracy was proposed in [15].

The success of deep learning motivates us to use deep neural network in the user authentication process. In this paper, we report a new method using Deep Neural-powered Spectral Centroid Mel Frequency Spatio-Temporal (DN-SCMFST). The entire procedure works around deep neural network encompassing one input layer, three hidden layers and one output layer. First, raw Speaking Faces Dataset is analyzed and validated in the input layer. Next, in the first hidden layer, pre-processing employing Kushner–Stratonovich and Spectral Subtraction Non-linear filter is modeled for further processing. In the second hidden layer, feature extraction using Spectral Centroid Mel Frequency and Spatial Temporal function are used for extracting relevant features. Following which Dual Aggregated Score Level Fusion is applied in the third hidden layer for comparison and finally the decision making results are generated as output in the output layer. This experiment was carried out on the Speaking Faces Dataset. The experimental work discloses that the proposed method surpasses previously published results.

2 Methodology

In this section, the proposed Deep Neural-powered Spectral Centroid Mel Frequency Spatio-Temporal (DN-SCMFST) method is detailed. The steps listed below are implemented in our work to secure multimodal biometric-based user authentication. The secure multimodal biometric-based user authentication pipeline using Deep Neural-powered Spectral Centroid Mel Frequency Spatio-Temporal (DN-SCMFST) method has four different stages, namely, acquisition, pre-processing, extractor and comparison with decision making. Figure 1 shows the architecture of Deep Neural-powered Spectral Centroid Mel Frequency Spatio-Temporal (DN-SCMFST) method to perform multimodal biometric-based user authentication.

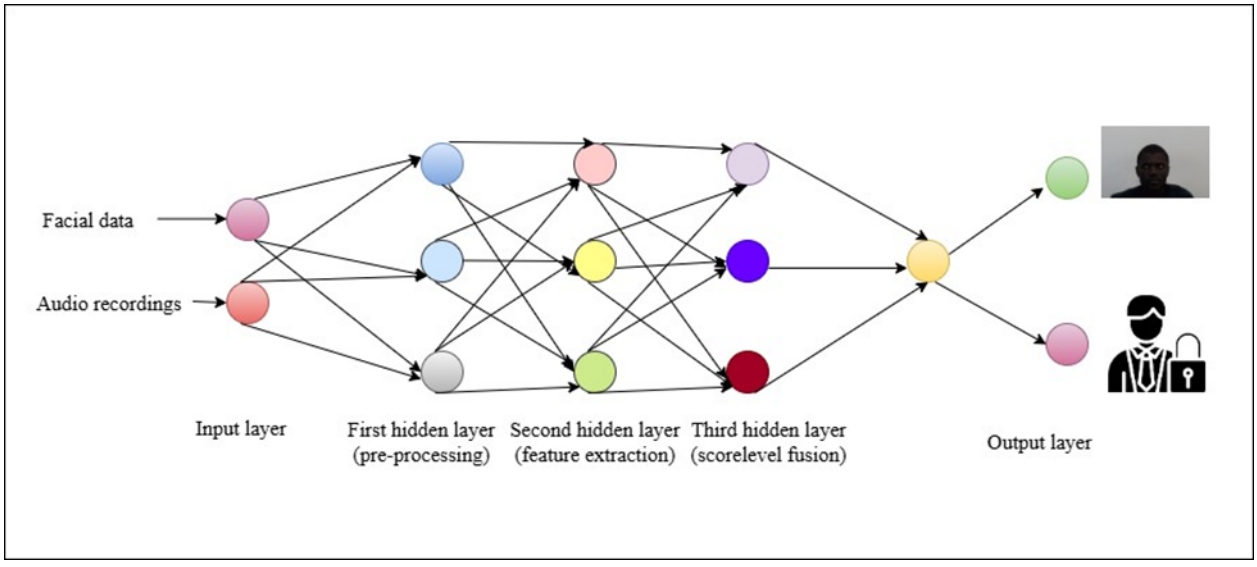

Fig 1. Architecture of DN-SCMFST

As illustrated in figure 1, the deep neural-powered multimodal biometric based user authentication consists of five layers, namely, one input layer obtaining facial data and audio recordings, first hidden layer performing pre-processing, second hidden layer performing feature extraction, third hidden layer doing score level fusion and finally the output layer doing decision making process. The detailed modeling is illustrated in figure 2. Figure 2 illustrates the key elements of the proposed secure multi-modal biometric-based user authentication method using DN-SCMFST method.

As shown in figure 2, the initial step requires providing input and, in our work, SpeakingFaces [30], a large-scale multimodal dataset is employed which is then acquired by launching an audio recorder in the input layer. The audio recorder converts the

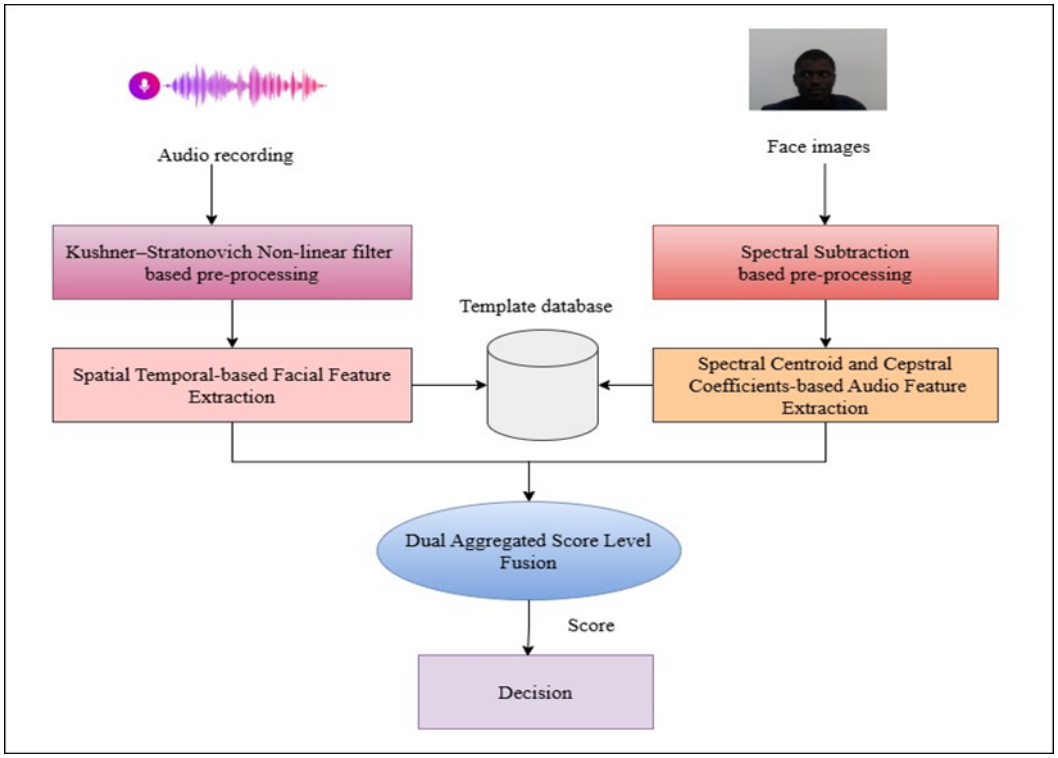

Fig 2. Block diagram of Deep Neural-powered Spectral Centroid Mel Frequency Spatio-Temporal (DN-SCMFST) method

captured input that is then later used for pre-processing. Next, the biometric features frequently contain noise and redundant data that are not required for secure multi-modal biometric-based user authentication and may even be deceptive, or misleading has to be discarded. In our work, Kushner–Stratonovich and Spectral Subtraction Non-linear filter-based Pre-processing is used and performed in the first hidden layer. For example, voice recording may have background noise and facial photos may contain non-face background information.

The purpose of applying pre-processing is to rationalize the raw biometric data (i.e. face and audio recording) for effortless user recognition. The pre-processed face and audio recording data is then organized for the next stage. Although the pre-processing stage cleans up the biometric face and voice recordings data by removing the noise and redundant data, the final dataset is usually immense and may still contain duplicated information. The extraction stage concentrates on extracting features indispensable for authenticating a person using Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction performed in the second hidden layer. Finally, in this stage, users are compared using Dual Aggregated Score Level Fusion in the third hidden layer and split into two groups on the basis of whether they accept or reject the authentication conclusion in the output layer.

2.1 Data Collection and Data Acquisition

Data collection and data acquisition are two distinct processes utilized to gather and prepare Speaking Faces Dataset for analyzing multimodal biometric based user authentication that is being provided as input for deep neural network via input layer. While data acquisition concentrates on the initial steps of encapsulated raw data from two different sources, i.e., face and audio recordings obtained from Speaking Faces Dataset extracted through https://huggingface.co/datasets/issai/Speaking_Faces. On the other hand, data collection confines the comprehensive procedure of gathering, cleaning and preparing data for use in multimodal biometric based user authentication analysis.

SpeakingFaces consists of 142 subjects, gender-balanced and each subject is recorded in close proximity from nine different angles resulting in 13,000 sample instances. It also includes more than 45 hour of video sequences taken from Stanford University. The video setup involved FLIR T540 thermal camera with an attached visual spectrum camera, a Logitech C920 Pro HDweb-camera and a built in dual stereo microphone of 44.1 kHz. Also, with the purpose of maximizing and aligning

frame rates for both cameras, the original web-camera resolution was minimized with the intent of maximizing and aligning the frame rates for both cameras, while preserving the region-of-interest (RoI)–that is, the face.

The data acquisition code initiated by launching an audio recorder and then proceeded with iterative capture of images using both cameras, at a fixed frequency of 28 frames per second (fps). Upon successful capturing of the calculated number of frames, the audio recorder stopped. Subsequently, the camera operator proceeded manually via a series of nine positions to cover face from all major angles with the duration of data collection for each position set to 900 frames. Hence, with the given data collection rate be 28 fps for both cameras, approximately found to be 32 second of video, giving in average 4.5 min of total video per subject. The overall dataset description is provided in table 1.

Table 1. Speaking Faces Dataset repository dataset

S. No	Features	Ranges	Description
1	subID	$\{1, 2, \ldots., 142\}$	Denotes subject number
2	trailID	$\{1, 2\}$	Denotes trial ID
3	sessionID	$\{1\ (or) 2\}$	1 – involve utterance, 0 – otherwise
4	posID	$\{1, 2, \ldots., 9\}$	Denote camera position
5	commandID	$\{1, 2, \ldots., 1297\}$	Denotes command number
6	frameID	$\{1, 2, \ldots., 900\}$	Number of frames in a sequence
7	streamID	$\{1\ (or) 2\ (or) 3\}$	1 – thermal images, 2 – visual images, 3 – aligned version
8	micID	$\{1\ (or) 2\}$	1 – left micro phone, 2 – right micro phone

From the above reposition, for example, the file named 90_1_2_9_12_167_2.png belongs to Trial 1 of Subject 90 and represented as given below.

Table 2. Trial 1 of Subject 90

90	**1**	**2**	**9**	**12**	**167**	**2**
↓	↓	↓	↓	↓	↓	↓
subID	trailID	sessionID	posID	commandID	frameID	streamID

Similarly, the subject ID along with number of trials for corresponding RGB and threshold image both for left and right microphone is listed below for subject 90 in below given table 3.

Table 3. Sample details of subject 90

Subject Id	Number of Trails	rgb_image_cmd	thr_image_cmd	mic1_audio_cmd_trim (left microphone)	mic2_audio_cmd_trim (right microphone)
sub_90_ia	Trial 1	5238	5238	44	44
	Trial II	5553	5553	43	43

With the above detailing, the proposed method design of multimodal based secure authentication system is narrated in the following sub-sections.

2.2 Kushner–Stratonovich and Spectral Subtraction Non-linear filter-based Pre-processing

Face images and audio recordings preprocessing involves preparing raw samples obtained from repository dataset for model training. As far as audio recording is concerned, the pre-processing here consists of noise reduction, normalization and so on and on the other hand, in case of facial recordings, it involves normalizing the images, removing noise and so on. By applying appropriate pre-processing technique makes certain cleaner, more consistent data, therefore enhancing the subsequent processing performance. In this work to improve on the peak signal-to-noise ratio while preserving the edges, Kushner–Stratonovich and Spectral Subtraction Non-linear filter-based Pre-processing model is used and performed in the deep neural-powered first hidden layer. Figure 3 shows the block diagram of Kushner–Stratonovich and Spectral Subtraction Non-linear filter-based Pre-processing model.

As shown figure 3, to start with the Kushner–Stratonovich Non-linear filter is applied to the sample face images to eliminate unwanted noise and differentiate between actual and measured values by pixels in the neighborhood in accordance to time instances. This in turn significantly improves the peak signal-to-noise ratio while preserving the edges of sample images.

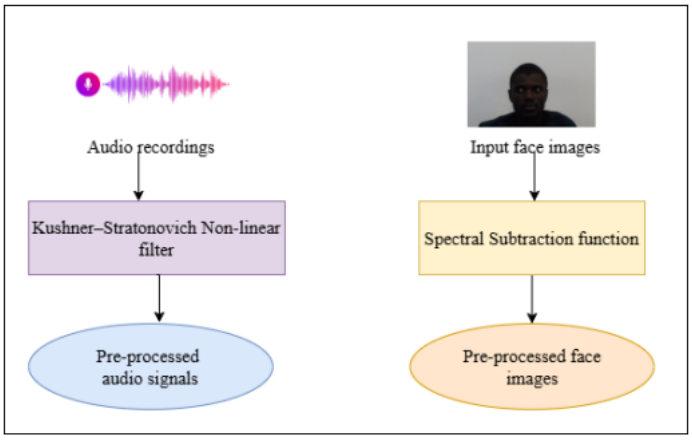

Fig 3. Kushner–Stratonovich and Spectral Subtraction Non-linear filter-based Pre-processing model

Following which, Spectral Subtraction function is employed in audio recording enhancement to eliminate unwanted noise. The working process revolves by evaluating the noise spectrum and subtracting it from noisy signal's spectrum to retrieve the clean speech spectrum.

To start with the face images 'F' from samples 'S' are formulated according to state of system with respect to time. This is mathematically stated as given below.

$$dF = f(T,t)dt + \sigma dw \tag{1}$$

Following which the noisy measurement of face images 'F' from samples 'S' is evaluated accordingly as given below.

$$dzh = h(F,t) + \eta dv \tag{2}$$

From the above **equations (1) and (2)**, 'w'and 'v' denotes the Weiner processes and by applying the above formulates provides a framework for fine tuning the probability distribution of face image given new observations. It assists in incorporating inherent randomness in facial dynamics. Following which the non-linear probability distribution function of face images 'F' from samples 'S' is modeled as given below.

$$PF = dProb(F,t) = L\left[Prob(F,t)\right]dt + Prob(F,t)\left(h(F,t) - Exp_t h(F,T)\right)\left(dzh - Exp_t h(F,T)dt\right) \tag{3}$$

From the above **equation (3)**, the conditional probability density is obtained '$Prob(F,t)$' by taking into considerations the forward operator '$L[Prob]$' and the difference between noisy measurement 'dzh' and expected measurement '$Exp_t h(F,T)dt$' respectively. This notably aids in improving peak signal to noise ratio by also preserving the edges for further processing.

Eliminating unwanted background noise from audio recordings is pivotal for accurate analysis. In this work, Spectral Subtraction function is used that works by measuring the noise spectrum during silence and obtaining the difference between it and the overall noisy signal spectrum. This procedure in turn strives to locate the required signal by eliminating the noisy component. Let us consider a noisy signal that comprises of raw audio and video files degraded by statistically independent additive noise as given below.

$$y[N] = p[N] + q[N] \tag{4}$$

From the above **equation (4)** '$y[N]$', '$p[N]$' and '$q[N]$' represents the sampled noisy audio and video files, clean audio and video files and additive noise, respectively. Owing to the reason that the sampled noisy audio and video files are non-stationery and time differing, it is processed on a frame-by-frame, and their representation is as given below.

$$Y(\alpha, FN) = P(\alpha, FN) + Q(\alpha, FN) \tag{5}$$

From the above **equation (5)**, 'FN' represents the frame number because of the reason that sampled noisy audio and video files are segmented into frames and the simplified representation is as given below.

$$|Y(\alpha)^2| = |P(\alpha)^2| + |Q(\alpha)^2| \tag{6}$$

Following which audio recordings are measured by subtracting noise estimate from received sample and are formulated as given below.

$$|P'(\alpha)^2| = |Y(\alpha)^2| - |Q'(\alpha)^2| \tag{7}$$

Finally, the evaluation of noise '$|P'(\alpha)^2|$' is then generated by aggregating speech pause frames 'SP' as given below.

$$PA = |P'(\alpha)^2| = \frac{1}{M}\sum_{k=1}^{M} |Y_{SP_k}(\alpha)|^2 \tag{8}$$

From the above equation the pre-processed unwanted noise eliminated audio recordings are obtained for further processing. The pseudo code representation of Kushner Spectral Subtraction Non-linear filter-based pre-processing is given below.

Algorithm 1: Kushner Spectral Subtraction Non-linear filter-based pre-processing
Input: Dataset 'DS', Samples '$S = \{S_1, S_2, \ldots, S_N, N \in F, A\}$', Face Images '$F = \{F_1, F_2, \ldots, F_m\}$', Audio Recordings '$A = \{A_1, A_2, \ldots, A_n\}$'
Output: Noise eliminated pre-processed face and audio recordings
1: **Initialize** '$N = 10000$', '$m = 142$', '$n = 142$'
2: **Begin**
3: **For** each Dataset 'DS' with Face Images 'F' and Audio Recordings 'A'
//Pre-processing – First hidden layer
4: Formulate face images from samples according to state of system with respect to time according to (1)
5: Obtain noisy measurement of face images from samples according to (2)
6: Evaluate non-linear probability distribution function according to (3)
7: **Return** pre-processed face images 'PF'
8: **End for**
9: **For** each Dataset 'DS' with Audio Recordings 'A'
10: Formulate noisy signal of raw audio and video files according to (4)
11: Process it on a frame-by-frame according to (5) and (6)
12: Measure audio recordings by subtracting noise estimate from received sample according to (7) and (8)
13: **Return** pre-processed audio recordings 'PA'
14: **End for**
15: **End**

As given in the above algorithm the overall Kushner Spectral Subtraction Non-linear filter-based pre-processing is split into two parts. The first part performs the pre-processing for face images using Kushner–Stratonovich Non-linear filter whereas the second part performs the pre-processing for audio recordings using Spectral Subtraction function. By applying the Kushner–Stratonovich Non-linear filter to face images when evaluating the system state on the basis of noisy measurements aims the preserve edges. Next, by using Spectral Subtraction function to the audio records by obtaining the difference between noise spectrum and noisy signal spectrum in turn improves peak signal to noise ratio.

2.3 Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction

With pre-processed face and audio recording samples as input are subjected to feature extraction process with the intent of extracting specific features from both the facial and voice data. For faces, this might include facial landmarks, texture analysis, or deep learning-based features. For voice, in our work Mel-frequency cepstral coefficients (MFCCs) are analyzed to efficiently capture essential audio signal features. Following which the facial features in our work are extracted using Spatial Temporal-based Facial Feature Extraction performed in the deep neural-powered second hidden layer. The basic proposition behind Spatial Temporal-based Facial Feature Extraction is to utilize the postulate to extract temporal and spatial information of face

recognition. Each pre-processed face input image has identically sampled frames and those frames are used in extracting spatial-temporal information. Figure 4 shows the block diagram of Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction model.

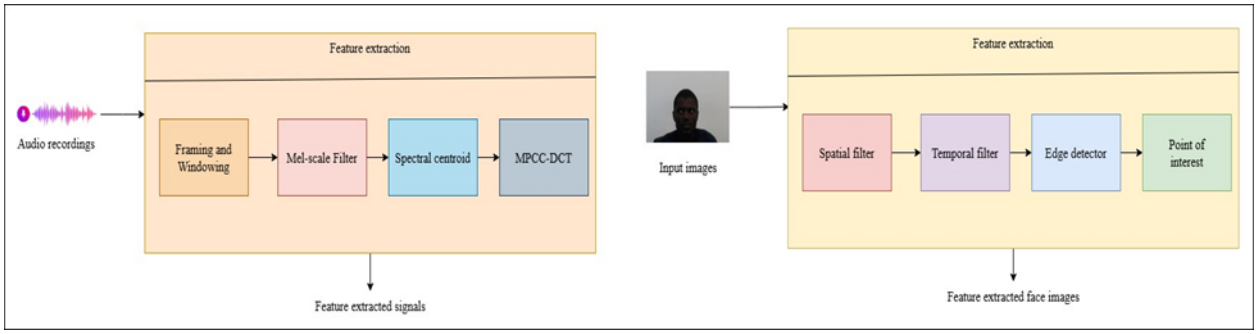

Fig 4. Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction

There are various phases involved in the extraction of features for an audio sample with MFCC. To start with the pre-processed audio signal is divided into frames. The duration for each position was set to 900 frames. With the given data collection rate of 28 fps for both cameras, approximately 32 s of video, yielding on average 4.5 min of total video per subject is utilized. Following which to reduce leakage, windowing is applied to each frame. Next, according to human auditory perception, each frame's power spectrum is filtered via a sequence of triangle filters arranged in relation to the Mel scale. Following which with the intent of creating a set of MFCCs, logarithm of filter bank output is measured, and Discrete Cosine Transform (DCT) is analyzed. These values in turn effectively extract key features of audio stream for user authentication. Next, to extract facial relevant features, interest point that experience fluctuation over space and time is focused. Here, spatial feature, i.e. edges and temporal feature, i.e. data changes over time are used for extracting relevant facial features.

To start with the pre-processed audio recordings 'PA' obtained as input are subjected to the process of framing and windowing. Here, initially the audio recordings 'PA' streams are split into frames and then Hamming window is applied to each frame as given below.

$$X = win(l) = 0.54 - 0.46\cos\left(\frac{2\pi N}{l-1}\right) \tag{9}$$

From the above **equation (9)**, 'l' represents the frame length with regards to sample 'N' audio recordings 'PA' streams to generate windowing results '$win(l)$'. By applying the above Hamming window to each frame in turn reduces the leakage to a greater extent. Following which a spectral centroid is evaluated denoting the center of mass of spectrum with the intent of improving the recognition accuracy. This is mathematically formulated as given below.

$$SC_i = \frac{\sum_n^0 kX_i k}{\sum_n^0 X_i k} \tag{10}$$

From the above **equation (10)** the spectral centroid of '$i-th$' frame is measured 'SC_i' by measuring the mean value of spectral centroid across all frames. This in turn aids in improving recognition accuracy. Following which each frame's power spectrum is filtered in relation to the Mel scale as given below.

$$Mf = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{11}$$

From the above **equation (11)** Mel scale filter 'Mf' results are generated according to the physical frequency of the sound wave 'f' satisfying human auditory perception. Finally, with the aid of Discrete Cosine Transform (DCT) sequence of log-mel filter bank is converted into concise representation of cepstral coefficients as given below.

$$FE_A = C_n = \sum_{n=1}^{N} \log En_n \cos\left[\frac{\pi N(mf - 0.5)}{MF}\right] \tag{12}$$

From the results of **equation (12)** the spectral information is converted into a domain where the most paramount audio features are extracted 'FE_A' therefore reducing redundancy.

Next, to extract spatial (i.e. edges) temporal (i.e. time) features the linear dimension characterization '$E\left(a, Res, \sigma_{pf}^2, \tau_{pf}^2\right)$' for an input pre-processed face image '$PF(a,b,s)$' is stated by convolving a Gaussian Kernel '$H\left(a,b,\sigma_{pf}^2, \tau_{pf}^2\right)$' with '$PF$' as given below.

$$E\left(a, Res, \sigma_{pf}^2, \tau_{pf}^2\right) = PF(a,b,s) * H\left(a,b,\sigma_{pf}^2, \tau_{pf}^2\right) \tag{13}$$

From the above **equation (13)**, 'σ_{pf}^2' and 'τ_{pf}^2' denotes the spatial and temporal deviations convolved by an operator '$*$' via second-moment matrix 's'. Then the second-moment matrix 's' is convolved with a Gaussian function '$H\left(a,b,\sigma_{pf}^2, \tau_{pf}^2\right)$' to ascertain these positions as given below.

$$T = H\left(a,b,\sigma_{pf}^2, \tau_{pf}^2\right) * \begin{bmatrix} E_a^2 & E_a E_b & E_a E_b \\ E_a E_b & E_b^2 & E_b E_s \\ E_a E_s & E_b E_s & E_s^2 \end{bmatrix} \tag{14}$$

$$E_a^2 = \frac{\partial^2 E}{\partial a^2}, E_b^2 = \frac{\partial^2 E}{\partial b^2}, E_s^2 = \frac{\partial^2 E}{\partial s^2}; E_a E_b = \frac{\partial}{\partial a}\left(\frac{\partial E}{\partial b}\right); E_b E_s = \frac{\partial}{\partial b}\left(\frac{\partial E}{\partial s}\right); E_a E_s = \frac{\partial}{\partial a}\left(\frac{\partial E}{\partial s}\right) \tag{15}$$

With the above obtained results from **equations (14) and (15)** the lambda values 'λ' denoting the eigen values from image gradients '(a,b)' are used to represent the point of interest (i.e. relevant face feature extracted results). This is represented as given below.

$$FE_F = \lambda_1 \lambda_2 - N(\lambda_1 + \lambda_2)^2 \tag{16}$$

From the above **equation (16)** results, the facial features are extracted 'FE_F' by understanding the local image structure and identifying edges, therefore minimizing redundancy to a greater extent.

2.4 Dual Aggregated Score Level Fusion

Finally score level fusion is evaluated by combining the data from face and audio template database is performed in the deep neural-powered third hidden layer. Owing to the difference in match scores, it is mandatory to first convert these match scores in '$[0,1]$' prior to modeling score level fusion. This is performed using tanh indicator as given below.

$$SF' = \frac{1}{2}\left[\tanh\left(0.01\left(\frac{SF-\mu}{\sigma}\right)\right)+1\right] \tag{17}$$

From the above **equation (17)**, 'SF'' represents the fine-tuned score and 'SF', 'μ', 'σ' represent the input score, mean and standard deviation of specific user respectively. Finally, these standardized scores are aggregated Dual Aggregated Score Level Fusion. This is represented as given below.

$$F = F(S_1, S_2, \ldots, S_N) = \prod_{i=1}^{N} F_N;\ A = A(S_1, S_2, \ldots, S_N) = \prod_{i=1}^{N} A_N \tag{18}$$

$$DA(S_1, S_2, \ldots, S_N) = \frac{\prod_{i=1}^{N} F_N}{\prod_{i=1}^{N} F_N + \prod_{i=1}^{N} A_N} \tag{19}$$

From the results of above **equation (18) and (19)**, the dual aggregated sample results 'DA' combining face 'F_N' and audio recordings 'A_N' is obtained with which matching is performed for user authentication. Finally in the output layer the actual results of authenticated or non-authentication are provided. The pseudo code representation of Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction with score level fusion is given below.

Algorithm 2. Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction with score level fusion
Input: Dataset 'DS', Samples '$S = \{S_1, S_2, \ldots, S_N, N \in F, A\}$', Face Images '$F = \{F_1, F_2, \ldots, F_m\}$', Audio Recordings '$A = \{A_1, A_2, \ldots, A_n\}$'
Output: Robust user recognition
1: **Initialize** pre-processed face images 'PF', pre-processed audio recordings 'PA', Threshold 'Th'
2: **Begin**
3: **For** each Dataset 'DS' with pre-processed face images 'PF' and pre-processed audio recordings 'PA'
//Audio recording feature extraction – second hidden layer
4: **For** each Dataset 'DS' with pre-processed audio recordings 'PA'
//Framing and Windowing
5: Apply framing and windowing process according to (9)
//Mel-scale Filter
6: Evaluate spectral centroid for each frame according to (10)
7: Apply Mel-scale Filter to satisfy human auditory perception according to (11)
//MPCC-DCT
8: Obtain cepstral coefficients according to (12)
9: **Return** audio feature extracted results 'FE_A'
10: **End for**
//Face images feature extraction – second hidden layer
11: **For** each Dataset 'DS' with pre-processed face images 'PF'
12: Convolve Gaussian kernel with pre-processed face images according to (13)
13: Evaluate second-moment matrix 's' according to (14) and (15)
14: **Return** facial feature extracted results 'FE_F' according to (16)
15: **End for**
//Dual Aggregated Score Level Fusion – third hidden layer
16: Convert these match scores in '$[0, 1]$' to design score level fusion according to (17)
17: Evaluate Dual Aggregated Score Level Fusion according to (18) and (19)
//Output layer
18: **If** '$DA(S_1, S_2, \ldots, S_N)\ exceed\ Th$'
19: **Then** user is authenticated
20: **End if**
21: **If** '$DA(S_1, S_2, \ldots, S_N)\ does\ not\ exceed\ Th$'
22: **Then** user is not authenticated
23: **End if**
24: **End for**
25: **End**

As given in the above algorithm the overall process of Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction with score level fusion is split into three sections, namely relevant audio feature extraction, face feature extraction and finally performing score level fusion. First, the pre-processed audio recordings are subjected to Spectral Centroid and Cepstral Coefficients-based Audio Feature Extraction for extracting relevant audio signals. Here, by applying the Spectral Centroid function to Cepstral Coefficients in turn aids in minimizing the false acceptance rate. Second, the pre-processed face images are subjected to Spatial Temporal function for extracting relevant point of interest. By applying both spatial and temporal features in turn aids in extracting edges at different time instances, therefore minimizing the false rejection rate. Finally, the relevant audio signals and face images extracted are subjected to Dual Aggregated Score Level Fusion

2.5 Case study and inferences

In this section case analysis of robust multimodal based authentication using deep neural-powered using Speaking Faces Dataset are simulated by applying Deep Neural-powered Spectral Centroid Mel Frequency Spatio-Temporal (DN-SCMFST). Figure 5 given below provides the sample RGB input images **(a) – (d)** with the pre-processed resultant images in **(e) – (h)**.

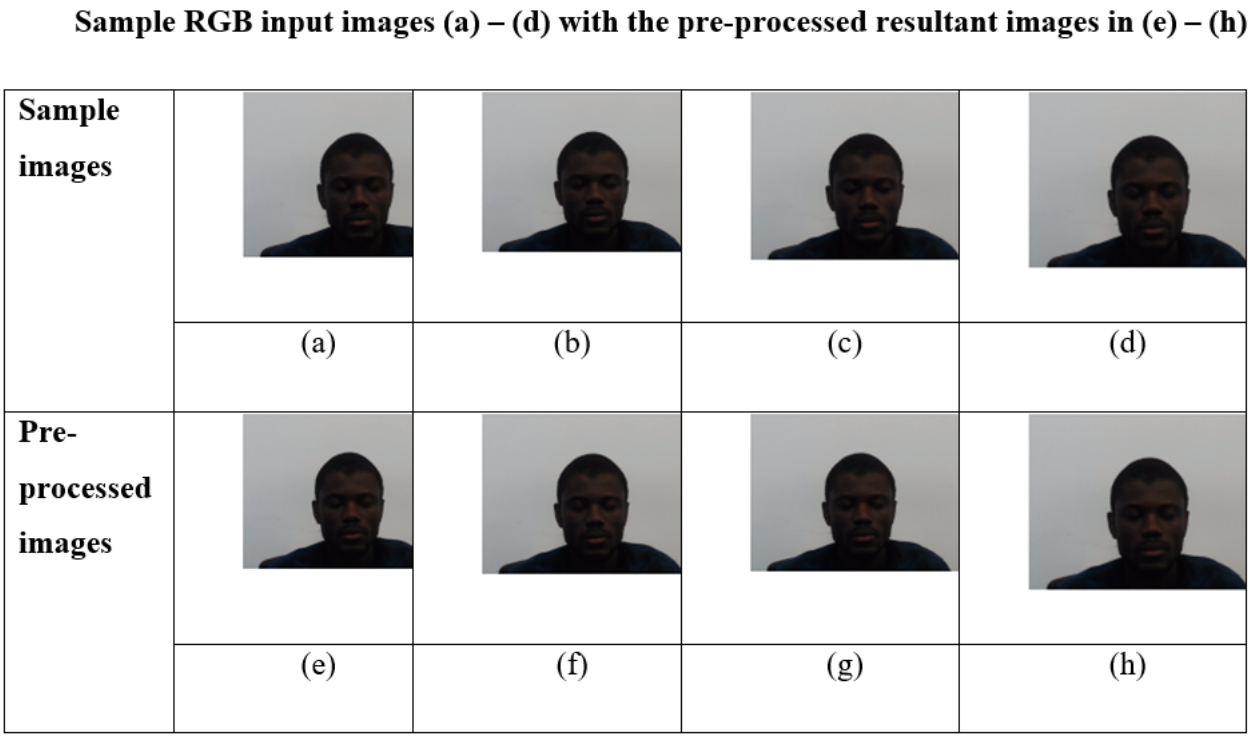

Fig 5. Kushner–Stratonovich and Spectral Subtraction Non-linear filter-based Pre-processing

As shown in figure 5 the results of facial data with corresponding input images, 90_1_2_1_200_100_2, 90_1_2_1_200_101_2, 90_1_2_1_200_102_2, 90_1_2_1_200_103_2 are obtained as input with equivalent values as given in table 4.

Table 4. Sample values of RGB input images with subsequent ID

	Rgb input images id						
S.No	**subID**	**trailID**	**sessionID**	**posID**	**commandID**	**frameID**	**streamID**
1	90	1	2	1	200	100	2
2	90	1	2	1	200	101	2
3	90	1	2	1	200	102	2
4	90	1	2	1	200	103	2

With the sample audio recordings obtained from the dataset, pre-processing is performed using Spectral Subtraction Non-linear filter and its results are shown in figure 6. Figure 6 given below provides the sample audio recordings in **(a)** and pre-processed results in **(b)** after applying Spectral Subtraction Non-linear filter.

With the pre-processed facial data, audio recordings, feature extraction is performed separately using Spatial Temporal function and Spectral Centroid Mel Frequency respectively. Their results are shown in figure 7.

Finally, the results of features extraction for audio recordings are given below.

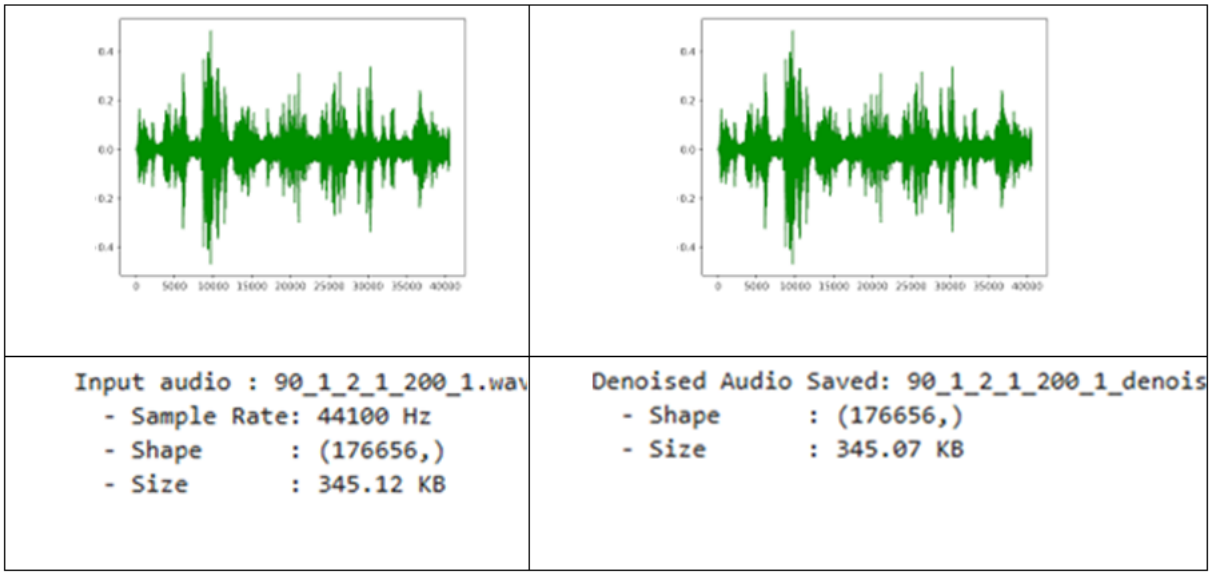

Fig 6. Spectral Subtraction Non-linear filter based pre-processing (a) Sample input audio (b) pre-processed audio

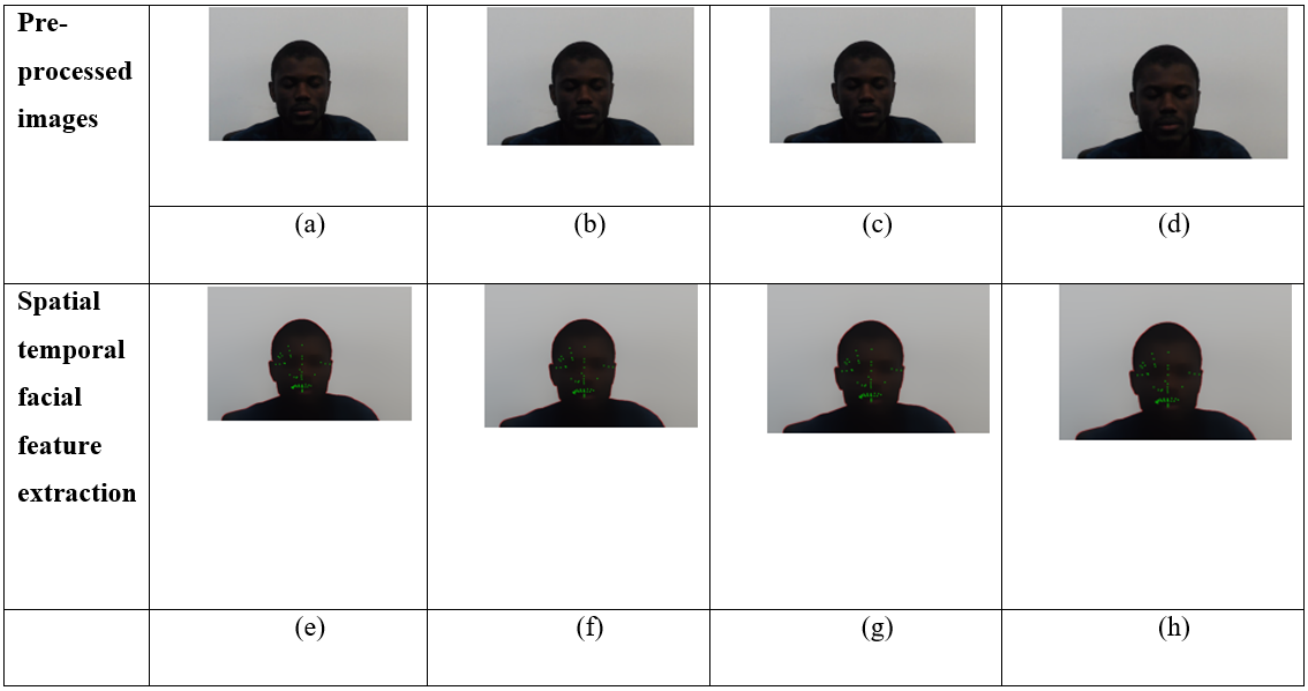

Fig 7. Facial feature extraction results

```
Input audio : 90_1_2_1_200_1_denoised.wav
 - Sample Rate: 44100 Hz
 - Shape      : (176656,)
 - Size       : 345.07 KB

Feature Extracted Audio Saved: feature_extracted_audio.wav
 - Shape      : (176640,)
 - Size       : 345.04 KB
```

From the above case study, it is inferred that by applying the proposed DN-SCMFST method for user authentication security is said to be ensured. Upon completion of feature extraction process in the second hidden layer of deep neural, finally score level fusion were applied for comparison in the third hidden layer. Only upon successful comparison users were provided with access. The quantitative analysis of the DN-SCMFST method is provided in the following sub-sections.

3 Results and Discussion

In this section, elaborate experiments are performed to validate the efficiency of the Multimodal Biometric recognition system using facial data and audio recordings called as Deep Neural-powered Spectral Centroid Mel Frequency Spatio-Temporal (DN-SCMFST). The method is implemented using Python with the aid of the dataset, Speaking Faces Dataset extracted from https://huggingface.co/datasets/issai/Speaking_Faces. Comparison is made using two existing methods, Convolutional Neural Networks and Recurrent Neural Networks (CNNs and RNNs) [1] and Deep learning-powered multimodal biometric authentication (DL- MBA) [2]. To ensure fair comparison, validation is made using the same dataset for all the three methods, DN-SCMFST, CNNs and RNNs [1] and DL- MBA [2]. While performing experiments, data are stored on a computer with an Intel(R) Core(TM) i5-7200 CPU @2.50GHz and 8.00GB of RAM.

3.1 Performance analysis of Peak Signal-to-Noise Ratio

The performance of proposed DN-SCMFST method is validated in terms of image quality metrics like Peak Signal to Noise Ratio (PSNR) for both facial data and audio recordings. The PSNR is mathematically evaluated as given below.

$$PSNR(dB) = 10\log_{10}\frac{255^2}{MSE} \tag{20}$$

From the above **equation (20)**, the resultant value of '$PSNR$' is obtained by taking the mean square error and is mathematically evaluated as given below.

$$MSE = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[Res(i,j) - prob(i,j)]^2 \tag{21}$$

From the above **equation (21)**, the resultant value of mean square error 'MSE' is arrived at by taking into consideration the difference pixel values based on the true resultant 'Res_j' (i.e. facial data and audio recordings) and probability '$prob_j$' for 'j' data point (i.e. facial data and audio recordings) respectively. Table 5 tabulates the validation results of the proposed DN-SCMFST with all other compared methods CNNs and RNNs [1] and DL-MBA [2] for all 10000 samples images respectively.

Figure 8 illustrates the PSNR results of three different methods, DN-SCMFST, CNNs and RNNs [1] and DL-MBA [2] for 10 different tongue sample facial images and audio recordings respectively. According to the sample image sizes for both facial data and audio recording variations in PSNR were observed using all three methods. However, overall PSNR improvement was observed using DN-SCMFST method upon comparison to [1] and [2] for both facial images and audio recordings. The reason was due to the application of Kushner–Stratonovich and Spectral Subtraction Non-linear filter-based Pre-processing model. Here, Kushner–Stratonovich Non-linear filter for facial images and Spectral Subtraction function for audio recordings separately not only aided in removing noise but also eliminated unwanted data. Moreover, by utilizing the Kushner–Stratonovich Non-linear filter to face images by evaluating the system state based on noisy measurements aims at preserving the edges. This in turn aided in improvement of PSNR using DN-SCMFST method by 18% compared to [1] and 28% compared to [2] for facial images and 17% compared to [1] and 26% compared to [2] for audio recordings.

Table 5. Tabulation for PSNR when applied with facial and audio recordings using DN-SCMFST, CNNs and RNNs [1] and DL-MBA [2]

Facial data size (KB)	PSNR (dB) – for facial data			Audio recordings size	PSNR (dB) –for audio recordings		
	DN-SCMFST	**CNNs and RNNs [1]**	**DL-MBA [2]**		**DN-SCMFST**	**CNNs and RNNs [1]**	**DL-MBA [2]**
217	29.14	22.82	19.15	345	35.35	20.25	17.55
218	24.35	18.85	16.35	261	38.15	33.15	30.35
220	31.65	25.65	21.45	275	41.35	36.55	33.45
218	28.55	24.75	22.35	328	35.25	30.25	27.15
218	32.75	26.45	24.15	292	37.35	32.55	29.45
219	31.95	25.75	23.25	344	32.45	27.45	24.35
217	22.65	20.65	18.55	184	30.15	25.35	21.25
220	36.75	30.85	27.45	326	38.35	33.15	30.15
217	30.2	24.85	21.35	344	32.45	27.85	24.35
219	32.35	23.55	21	361	35.15	30.45	27.15

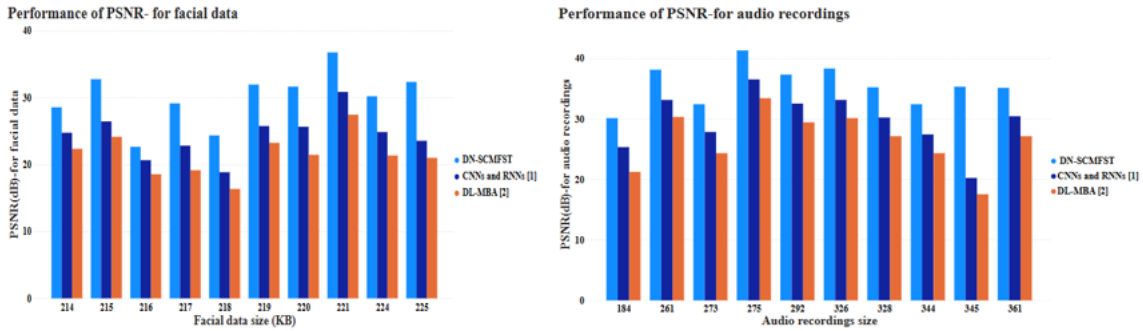

Fig 8. Performance analyses of PSNR (a) using facial data (b) using audio recordings

3.2 Performance analysis of Recognition accuracy

Multimodal biometric systems that combine multiple biometric features for user authentication, specifically achieve higher recognition accuracy upon comparison to unimodal systems. This is due to the reason that they can grip the advantages of distinct modalities to settle accounts for the drawbacks of individual ones, resulting in more robust and reliable identification. The recognition accuracy is mathematically formulated as given below.

$$RA = \sum_{i=1}^{N} \frac{S_{RA}}{S_i} \tag{22}$$

From the above **equation (22)**, recognition accuracy 'RA' is measured by using the samples involved in simulation process 'S_i' and the samples recognized accurately 'S_{RA}'. It is measured in terms of percentage (%). Table 6 lists the validation results of the proposed DN-SCMFST with two existing methods, CNNs and RNNs [1] and DL-MBA [2] in terms of recognition accuracy.

Figure 9 shows the recognition accuracy for proposed DN-SCMFST and two existing methods, CNNs and RNNs [1], DL-MBA [2] with respect to 10000 samples of facial data and audio recordings. From figure 9 three inferences are made. First, increasing the sample size does not have major influence in recognition accuracy when applied with both facial data and audio recordings. Moreover, the recognition accuracy of the proposed DN-SCMFST method was found to be comparatively better than [1] and [2]. Also, the recognition accuracy for facial data using the proposed DN-SCMFST method was comparatively better when applied with audio recordings. The reason was due to the application of Dual Aggregated Score Level Fusion for comparison between users and making decisions accordingly. Here, prior to modeling score level fusion match scores were converted into '$[0, 1]$' prior to modeling score level fusion using tanh indicator. This in turn aided in improvement of recognition accuracy using DN-SCMFST method by 5% compared to [1] and 18% compared to [2] for facial images and 4% compared to [1] and 10% compared to [2] for audio recordings.

Table 6. Tabulation for recognition accuracy for facial and audio recordings using DN-SCMFST, CNNs and RNNs [1] and DL-MBA [2]

Samples	Recognition accuracy (%) – for facial data			Recognition accuracy (%) –for audio recordings		
	DN-SCMFST	**CNNs and RNNs [1]**	**DL-MBA [2]**	**DN-SCMFST**	**CNNs and RNNs [1]**	**DL-MBA [2]**
1000	97.5	96	95	96	95	93
2000	95.25	90.35	87.45	94.35	90.25	84.35
3000	93	88.15	85.25	92.45	88.45	82.45
4000	91.45	86.45	83.55	88.15	84.35	78.25
5000	90	85.35	82.15	88.35	84.55	78.35
6000	88.15	83.15	80.35	85.45	81.35	75.35
7000	91.35	86.25	83.45	88.35	84.25	78.55
8000	93	88.5	85.25	90.25	86.55	80.25
9000	94.55	89.15	86.15	91.35	87.65	81.35
10000	96	91.35	88.54	93	89.25	83.45

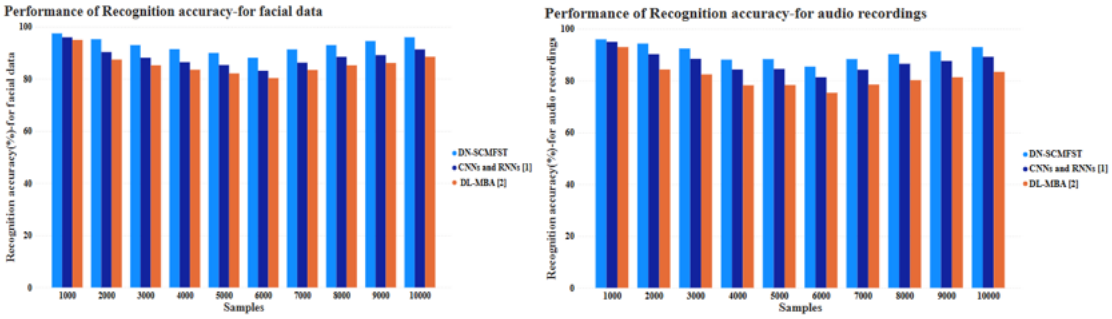

Fig 9. Performance analyses of recognition accuracy (a) using facial data (b) using audio recordings

3.3 Performance analysis of Recognition time

As far as multimodal biometric authentication systems are concerned, recognition time refers to the total time it consumes to verify or identify an individual using multiple biometric modalities, using facial data and audio recordings. This includes the time to capture, process, and compare the biometric data from different sources. The recognition time is measured as given below.

$$RT = \sum_{i=1}^{N} S_i * Time\ (UA) \tag{23}$$

From the above **equation (23)** recognition time 'RT' is referred based on the samples (i.e. facial data and audio recordings) 'S_i' and the time consumed in performing user authentication '$Time\ (UA)$'. It is measured in terms of seconds (sec). Table 7 given below provides the validation results of the proposed DN-SCMFST with two existing methods, CNNs and RNNs [1] and DL-MBA [2] in terms of recognition time by substituting the values in **equation (23)**.

Figure 10 illustrates the results of recognition time for two distinct modals, facial data and audio recording using three methods, DN-SCMFST, CNNs and RNNs [1], DL-MBA [2] for 13000 different samples. From figure 10 increasing the sample size did not have any impact on recognition time for both facial data and audio recording. This is evident from 6000 samples where the recognition time for DN-SCMFST, CNNs and RNNs [1], DL-MBA [2] using facial data was observed to be 90sec, 100sec, 95sec respectively, whereas for 7000 samples the recognition time for DN-SCMFST, CNNs and RNNs [1], DL-MBA [2] using facial data was found to be 83sec, 93sec and 88sec respectively. Also, from the inferences it is observed that the recognition time using facial data for DN-SCMFST method was found to be better than [1] and [2]. The reason was due to the application of deep neural-powered Dual Aggregated Score Level Fusion performed in the third hidden layer. Alos, only relevant audio signals and face images extracted were subjected to Dual Aggregated Score Level Fusion. This in turn reduces the recognition time for facial data using DN-SCMFST method by 15% compared to [1] and 10% compared to [2]. Moreover, the recognition

Table 7. Tabulation for recognition time for facial and audio recordings using DN-SCMFST, CNNs and RNNs [1], DL-MBA [2]

Samples	Recognition time (sec) – for facial data			Recognition time (sec) –for audio recordings		
	DN-SCMFST	**CNNs and RNNs [1]**	**DL-MBA [2]**	**DN-SCMFST**	**CNNs and RNNs [1]**	**DL-MBA [2]**
1000	35	41	45	25	29	35
2000	45	55	50	30	38	45
3000	55	65	60	33	41	48
4000	68	78	73	35	43	50
5000	75	85	80	38	46	53
6000	90	100	95	45	53	60
7000	83	93	88	58	61	73
8000	74	84	79	45	53	60
9000	69	79	74	40	48	55
10000	55	65	60	35	43	50

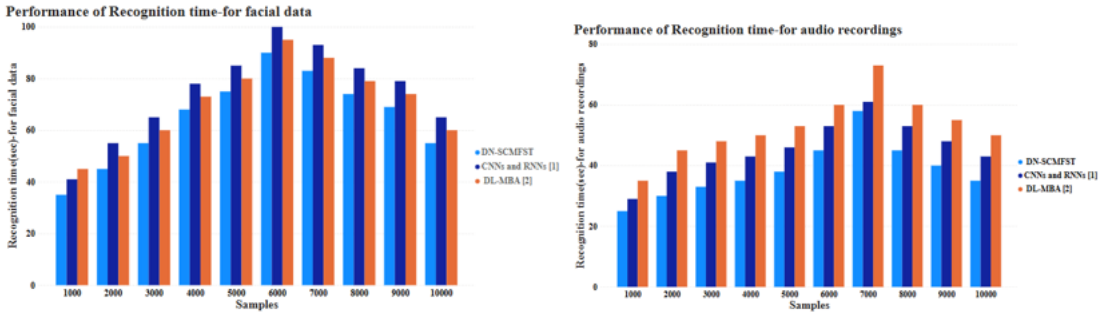

Fig 10. Performance analyses of recognition time (a) using facial data (b) using audio recordings

time for audio recording using DN-SCMFST method was found to be reduced by 19% upon comparison to [1] and 27% upon comparison to [2].

3.4 Performance analysis of False acceptance rate

The performance metric of a multi-modal biometric-based user authentication is measured by its accuracy and security in authentication, frequently quantified using metrics such as the False Acceptance Rate (FAR) and False Rejection Rate (FRR). The percentage of accurately matched authenticated users among all users is referred to as accuracy. The probability of falsely disclosing an unauthenticated user is referred to as the False Acceptance Rate (FAR). It is indicated as the percentage ratio of false acceptances divided by sum of true rejections and false acceptances as given below.

$$FAR = \frac{FP}{FP+TN} \tag{24}$$

From the above **equation (24)** the results of False Acceptance Rate 'FAR' are arrived at based on the false positive rate 'FP' and true negative rate 'TN' respectively. It is measured in terms of percentage (%). Table 8 given below gives the results of the DN-SCMFST, CNNs and RNNs [1], DL-MBA [2] results in terms of FAR by substituting the values in **equation (24)**.

Figure 11 shows the results of false acceptance rate for facial data and audio recordings using three methods, DN-SCMFST, CNNs and RNNs [1], DL-MBA [2] with respect to 10000 samples. From figure 11 the FAR of DN-SCMFST was observed to be comparatively lesser than [1] and [2] for both facial data and audio recordings. The improvement in FAR would be contributed to the application of Spectral Centroid Mel Frequency for audio extraction and spatial temporal data based feature extraction for facial images. With the aid of Discrete Cosine Transform (DCT) sequence of log-mel filter bank the pre-processed audio signals were converted into concise representation of cepstral coefficients that in turn reduced the overall false positive rate. With this the FAR using proposed DN-SCMFST method for audio recordings was found to be comparatively better by 25%

Table 8. Tabulation for false acceptance rate for facial and audio recordings using DN-SCMFST, CNNs and RNNs [1], DL-MBA [2]

Samples	FAR (%) – for facial data			FAR (%) – for audio recordings		
	DN-SCMFST	CNNs and RNNs [1]	DL-MBA [2]	DN-SCMFST	CNNs and RNNs [1]	DL-MBA [2]
1000	13.51	17.54	21.18	9.34	13.51	17.54
2000	11.15	13.35	14.55	10.15	12.45	13.15
3000	8.35	10.55	11.25	11.35	13.15	14.35
4000	7.45	9.55	10.45	12.45	14.55	15.15
5000	5.35	7.25	8.53	13	15	16
6000	8.25	10.55	11.55	13.85	16	17
7000	9.15	11.65	12.15	12	14.35	16.55
8000	10.35	13.55	13.25	11.35	13.15	15.35
9000	11.25	12.35	13.75	11	13	14.85
10000	12	14	15	10	12	13

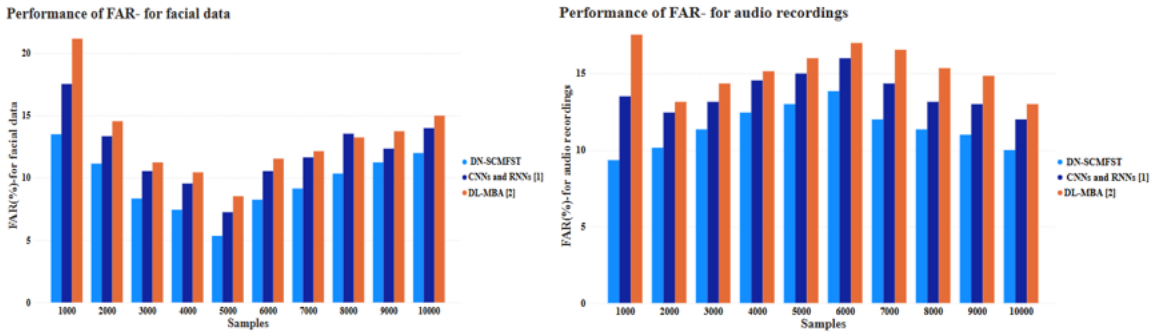

Fig 11. Performance analysis of FAR (a) using facial data (b) using audio recordings

compared to [1] and 37% compared to [2]. Moreover, using the spatial and temporal deviations convolved via second-moment matrix 's' aids in extracting the point of interest of corresponding face images. With this the FAR using proposed DN-SCMFST method for facial data was found to be comparatively better by 20% compared to [1] and 35% compared to [2].

3.5 Performance analysis of False rejection rate

On the other hand, the probability of falsely rejecting an authenticated user is referred to as the False Rejection Rate (FRR). It is indicated as the percentage ratio of false rejections divided by the sum of correct acceptances and incorrect rejections. This is mathematically formulated as given below.

$$FRR = \frac{FN}{FN + TP} \tag{25}$$

From the above **equation (25)** the false rejection rate 'FRR' is arrived at based on the false negative value 'FN' and true positive value 'TP'. It is measured in terms of percentage (%). Table 9 given below lists the FRR values (facial and audio recordings).

Finally figure 12 illustrates the False Rejection Rate (FRR) using three methods, DN-SCMFST, [1] and [2] for 10000 different samples involving both facial data and audio recordings. Though for the initial 6 rounds the FRR was found to be increased in all the methods for facial data and audio recordings, however for the remaining 4 rounds it was found to be decreased. But comparative analysis showed improved results for proposed DN-SCMFST method upon comparison to [1] and [2] for two types of modalities. The reason was due to the application of deep neural-powered Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction with score level fusion algorithm. By applying this algorithm in the second hidden layer features were extracted separately for audio and facial data and accordingly score level fusion were applied in the third hidden layer. This in turn aided in minimizing the false negative rate. As a result, the overall FRR using proposed DN-SCMFST method for facial data was found to be improved by 18% compared to [1] and 28% compared to [2].

Table 9. Tabulation for false rejection rate (facial and audio recordings) using DN-SCMFST, CNNs and RNNs [1], DL-MBA [2]

Samples	FRR (%) – for facial data			FRR (%) –for audio recordings		
	DN-SCMFST	CNNs and RNNs [1]	DL-MBA [2]	DN-SCMFST	CNNs and RNNs [1]	DL-MBA [2]
1000	0.44	0.67	0.79	0.33	0.44	0.67
2000	0.51	0.68	0.81	0.35	0.48	0.68
3000	0.58	0.7	0.73	0.37	0.51	0.7
4000	0.65	0.72	0.84	0.41	0.54	0.71
5000	0.75	0.77	0.85	0.45	0.55	0.72
6000	0.85	0.89	0.88	0.55	0.57	0.73
7000	0.77	0.81	0.83	0.48	0.53	0.54
8000	0.68	0.73	0.75	0.4	0.51	0.52
9000	0.53	0.67	0.7	0.35	0.4	0.42
10000	0.47	0.55	0.58	0.31	0.38	0.39

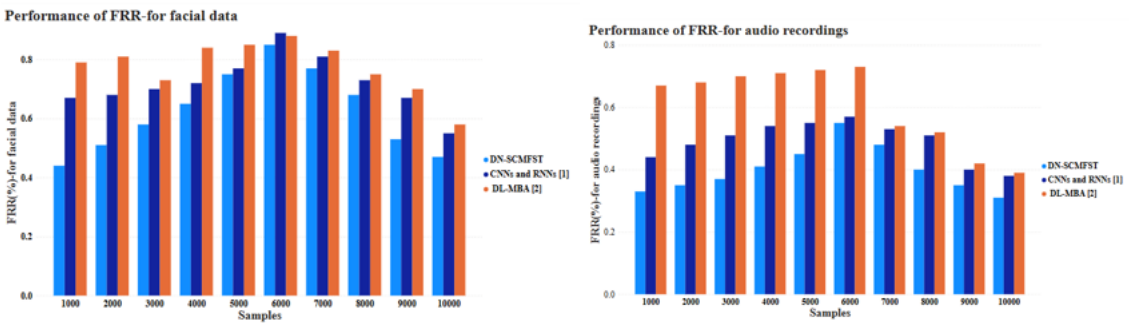

Fig 12. Performance analyses of FRR (a) using facial data (b) using audio recordings

4 Conclusion

This study presented a novel Deep Neural Network powered method for user authentication using multimodal biometrics, like, facial data and audio recordings near the conclusion of the conversation using Deep Neural-powered Spectral Centroid Mel Frequency Spatio-Temporal (DN-SCMFST) method. The proposed method combines Kushner–Stratonovich and Spectral Subtraction Non-linear filter and the Spectral Centroid Mel Frequency and Spatial Temporal-based feature extraction to ensure robust multimodal based biometric authentication. The algorithm is designed to simultaneously optimize four eminent objectives including reducing recognition time, FAR, and FRR with improved recognition accuracy. The proposed DN-SCMFST method is assessed on various samples. The performance of the proposed DN-SCMFST method is compared to the popular user authentication methods such as CNNs and RNNs [1], DL-MBA [2] using different performance metrics. Experimental results prove that the DN-SCMFST method efficiently performs biometric authentication using multimodal with improved recognition accuracy and minimizing recognition time, FAR and FRR.

5 Acknowledgement

The authors thank, DST-FIST, Government of India for funding towards infrastructure facilities at St.Joseph's College (Autonomous), Tiruchirappalli - 620 002. Non-Cited References [16], [17], [18], [19], [20]

References

1) Vatchala S, Yogesh C, Ganesan YGVPA, Raja MK, Arul AV, Ramesh D. Multi-modal biometric authentication: Leveraging shared layer architectures for enhanced security. *IEEE Access*. 2025;13:1–12. https://doi.org/10.1109/ACCESS.2025.3534223.
2) Salturk S, Kahraman N. Deep learning-powered multimodal biometric authentication: Integrating dynamic signatures and facial data for enhanced online security. *Neural Computing and Applications*. 2024. https://doi.org/10.1007/s00521-024-09690-2.

3) Wang JS. Exploring biometric identification in FinTech applications based on the modified TAM. *Financial Innovation*. 2021;7(42):1–24. https://doi.org/10.1186/s40854-021-00268-0.
4) Mansouri-Benssassi E, Ye J. Generalisation and robustness investigation for facial and speech emotion recognition using bio-inspired spiking neural networks. *Soft Computing*. 2021;25:1717–1730. https://doi.org/10.1007/s00500-020-05212-0.
5) Lee S, Han DK, Ko H. Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification. *IEEE Access*. 2021;9:94557–94572. https://doi.org/10.1109/ACCESS.2021.3092364.
6) Loizou CP. An automated integrated speech and face image analysis system for the identification of human emotions. *Speech Communication*. 2021;130:15–26. https://doi.org/10.1016/j.specom.2021.04.001.
7) Hinatsu S, Sakiwada. Biometric authentication using chest wall displacement recorded by VHF-band loop antenna. *IEEE Access*. 2024;12. https://doi.org/10.1109/access.2024.3521013.
8) Abuqaaud K, Nassif AB, Shahin I. FaciaVox: A diverse multimodal biometric dataset of facial images and voice recordings. *Data in Brief*. 2025;60:110123. https://doi.org/10.1016/j.dib.2025.110123.
9) Mun HJ, Lee MH. Design for visitor authentication based on face recognition technology using CCTV. *IEEE Access*. 2022;12:118345–118356. https://doi.org/10.1109/ACCESS.2022.3221234.
10) Sulavko A, Panfilova I, Samotuga A, Inivatov D, Lozhnikov P, Vulfin A. Biometric-based key generation and user authentication using voice password images and neural fuzzy extractor. *Applied System Innovation*. 2025;8(1):3. https://doi.org/10.3390/asi8010003.
11) Lonbar SM, Beigi A, Bagheri N, Peris-Lopez P, Camara C. Deep learning-based biometric authentication system using a high temporal/frequency resolution transform. *Frontiers in Digital Health*. 2024;6:1506782. https://doi.org/10.3389/fdgth.2024.1506782.
12) Salturk S, Kahraman N. Deep learning-powered multimodal biometric authentication: Integrating dynamic signatures and facial data for enhanced online security. *Neural Computing and Applications*. 2024;30:11311–11322. https://doi.org/10.1007/s00521-024-09723-6.
13) Piugie YBW, Charrier C, Manno JD, Rosenberger C. Deep features fusion for user authentication based on human activity. *IET Biometrics*. 2023;12(4):415–427. https://doi.org/10.1049/bme2.12109.
14) Pravallika T, Varma SM, Muthulakshmi A. Secure multi-factor authentication using deep learning. *IOSR Journal of Computer Engineering*. 2025;27(1).
15) Seyfizadeh A, Peach RL, Tovote P, Isaias IU, Volkmann J, Muthuraman M. Enhancing security in brain–computer interface applications with deep learning: Electroencephalogram-based user identification. *Expert Systems with Applications*. 2024;253:124512. https://doi.org/10.1016/j.eswa.2024.124512.
16) Saxena N, Varshney D. Smart home security solutions using facial authentication and speaker recognition through artificial neural networks. *International Journal of Cognitive Computing in Engineering*. 2021;2:57–65. https://doi.org/10.1016/j.ijcce.2021.05.003.
17) Akdeniz F, Becerikli Y. Recurrent neural network and long short-term memory models for audio copy-move forgery detection: A comprehensive study. *The Journal of Supercomputing*. 2024;80(9):8319–8341. https://doi.org/10.1007/s11227-023-05702-4.
18) Ayeswarya S, Johnsingh K. A comprehensive review on secure biometric-based continuous authentication and user profiling. *IEEE Access*. 2024;12:70012–70034. https://doi.org/10.1109/ACCESS.2024.3401123.
19) Qahwaji MAAR, Kyeremeh GK, Ali NT, Abd-Alhameed RA. The evolution of biometric authentication: A deep dive into multi-modal facial recognition – A review case study. *IEEE Access*. 2024;12:145678–145695. http://dx.doi.org/10.1109/ACCESS.2024.3486552.
20) Salturk S, Kahraman N. Deep learning-powered multimodal biometric authentication: Integrating dynamic signatures and facial data for enhanced online security. *Neural Computing and Applications*. 2024;36. https://doi.org/10.1007/s00521-024-09723-6.