

## Semi-automatic voice comparison approach using spiking neural network for forensics

Kruthika Siddanakatte Gopalaiah<sup>1,2</sup>, Trisiladevi Chandrakant Nagavi<sup>1</sup>, Parashivamurthy Mahesha<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, S. J. College of Engineering, JSS Science and Technology University, Mysore, India

<sup>2</sup>Department of Artificial Intelligence and Machine Learning, Vidyavardhaka College of Engineering, Mysore, India.

### Article Info

#### Article history:

Received Jul 16, 2024

Revised Mar 30, 2025

Accepted Jun 8, 2025

#### Keywords:

Artificial neural network

Digital forensics

Forensic voice comparison

Free lossless audio codec

Spiking neural networks

### ABSTRACT

This paper explores the application of a semi-automatic technique using spiking neural network (SNN) approach for forensic voice comparison (FVC), addressing the limitations of traditional methods that are time-consuming and subjective. By integrating machine learning with human expertise, the SNN, which mimics the brain's processing of temporal information, is applied to analyze Australian English voice data in .flac format. The model leverages synaptic connection strengths modified by spike timing, allowing for flexible voice feature representation. Performance metrics, including confusion matrices and receiver operating characteristic (ROC) analysis, indicate the model's accuracy of 94.21%, highlighting the effectiveness of the SNN-based approach for FVC.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



### Corresponding Author:

Trisiladevi Chandrakant Nagavi

Department of Computer Science and Engineering, S. J. College of Engineering

JSS Science and Technology University

Mysore, Karnataka 570006, India

Email: trisiladevi@sjce.ac.in

## 1. INTRODUCTION

Digital forensics is a crucial branch of forensic science that focuses on the recovery, investigation, examination and analysis of evidence found in digital devices. This field plays a crucial part in solving crimes by uncovering evidence from various digital sources, such as computers, smartphones and networks. It involves a meticulous process to assure the authenticity and reliability of the data collected, which can be used in legal proceedings. In the field of digital forensics, forensic voice comparison (FVC) is a specialized branch focused on examining the different voice recordings. The primary objective of FVC is to identify a suspect by comparing a trace voice sample with known samples, providing an evidence-based assessment of whether the recordings come from the same speaker for legal or investigative purposes [1]–[4]. Traditionally, FVC has relied heavily on manual analysis, in which trained experts listen to and compare voice samples. Although this method can be effective, it is often time-consuming and prone to human error and bias. To address these limitations, there has been a shift towards semi-automatic approaches that integrate human expertise with machine learning and deep learning techniques. These approaches aim to enhance both the efficiency and accuracy of voice comparisons, reducing reliance on subjective judgment while improving reliability in forensic investigations [5]–[7].

The most challenging task in FVC is analyzing audio samples from both trace and known sources to determine the similarity or dissimilarity in the suspect's voice, which is crucial for identifying any past criminal activity. While earlier studies have explored the impact of semi automatic approach artificial neural

network (ANN) a computational model trained to identify and classify speakers based on patterns in audio features extracted from speech samples. An ANN is made up of networked nodes or neurons, that process information similar to biological neural networks, they have not explicitly address well. Likewise, its influence on the proposed study utilizes spiking neural network (SNN), it offers an innovative method for voice pattern recognition. There is lot of ongoing research in artificial intelligence, machine learning, and neuroscience related to SNN. In order to comprehend brain function and create innovative computational models, a large number of academic institutions and research labs investigate SNN. The field of FVC has limited use of SNN in the existing works. This research aims to address this gap by proposing a SNN model that assist investigative agencies in the effective identification of the suspects. The SNN inspired by the neural architecture of the biological brain, processes information through discrete spikes, mimicking the way neurons communicate. The time-varying processing makes SNN particularly well-suited with adjustable threshold for analyzing complex voice recordings for FVC. The motivation of adopting the SNN for FVC is designed for tasks involving similarity or dissimilarity estimation [8]–[12]. The key contributions of this research are outlined as follows:

- To pre-process input data: remove background noise for a clear speech recording.
- To extract and compare features: use an adjustable threshold SNN to compare voice samples and analyze similarities or differences.
- To conduct performance analysis: evaluate the system using a confusion matrix and its extended metrics for comprehensive assessment.
- To perform comparative analysis: compare the results of the proposed research work with existing studies to highlight improvements and contributions.

The objective of this work is to identify suspects in FVC using SNN. Input voice samples are collected and analyzed to detect similarities and dissimilarities between voices. These samples undergo pre-processing with a stationary noise reduction algorithm to enhance clarity. The pre-processed voices are then converted into discrete spikes, mimicking neuronal communication. This time-dependent processing makes SNNs particularly well-suited for analyzing complex voice recordings in FVC, with an adjustable threshold to improve accuracy. The accuracy of the system is evaluated using a confusion matrix and extended performance metrics. Finally, the proposed approach is compared with existing studies to demonstrate its effectiveness.

The structure of the paper is organized as follows: the review of literature is given in section 2. Section 3 details data collection and experimental setup. Section 4 discusses FVC using a SNN. Additionally, section 5 describes the result analysis and discussion obtained from SNN. Furthermore, the section 6 presents a comparison with existing work. The paper is concluded in section 7.

## 2. LITERATURE REVIEW

Research in FVC has evolved over the decades, introducing various methodologies for authenticating and verifying speech. With the rise of digital technology, voice comparison plays a crucial role in forensics. This review examines key FVC methods, focusing on a semi-automatic approach using SNNs known for analyzing complex speech patterns. It highlights the need for enhanced voice pattern recognition, driving the proposed SNN-based research to improve suspect identification through voice similarity comparisons [13]–[18].

Several studies [19]–[25] have demonstrated the effectiveness of SNNs for spatiotemporal pattern classification. Morales *et al.* [25] developed a multilayer SNN on the SpiNNaker platform, using leaky integrate-and-fire neurons and firing rate-based algorithms to train inter-layer connections. The network achieved over 85% hit rate per class with a signal-to-noise ratio (SNR) above 3 dB, demonstrating its effective configuration and training method. Similarly, Wu *et al.* [19] proposed the self-organizing map (SOM)-SNN, a biologically inspired artificial spiking circuit (ASC) framework combining an unsupervised SOM with an event-based SNN to classify spatiotemporal patterns. On the real world computing partnership (RWCP) and TIDIGITS datasets, SOM-SNN showed robustness to noise and early decision-making, achieving 97.40% and 99.60% accuracy, respectively.

Wu *et al.* [20] investigated SNNs for acoustic modeling in large-vocabulary automatic speech recognition (ASR), achieving competitive accuracies with only 10-time steps and 0.68 times the synaptic operations per audio frame. This combination of energy-efficient neuromorphic hardware and deep SNNs shows potential for ASR on mobile and embedded devices, with reported accuracies of 18.7% and 36.9%.

Augé *et al.* [21] explored SNNs for energy-efficient edge devices, emphasizing small-scale neuromorphic implementations. By integrating resonating neurons as the SNN input layer for end-to-end online audio classification, they enabled low-power continuous audio stream analysis. The approach, evaluated using a keyword spotting benchmark, demonstrated strong accuracy using mel-frequency spectral features.

Further, Mukhopadhyay *et al.* [22] studied human footstep sound classification in natural environments using a wireless sensor network (WSN) for security surveillance. By employing an SNN with simple time-domain features, they aimed to create energy-efficient, cost-effective sensor nodes. Simulations showed significant power savings with analog SNNs, despite minor accuracy loss mitigated by redundancy and majority voting. Future research may focus on low-power feature extraction for surveillance systems [23].

Earlier, Yamazaki *et al.* [23] highlighted the limitations of deep neural networks, such as high computational costs and energy consumption in drones and self-driving vehicles. They proposed SNNs as efficient alternatives, mimicking biological neurons through sparsity and temporal coding. The paper reviews biological neuron theories, spike-based neuron models, SNN training methods, and applications in computer vision and robotics, offering future research insights.

Kholkin *et al.* [24] discussed the rising interest in SNNs despite challenges with von Neumann architectures, noting that hardware advancements now enable practical SNN applications. Their comparison of SNN and ANN reservoir computing architectures using the RCNet library showed SNNs had longer run times but superior classification, particularly for complex datasets like industrial sensor faults. In ball bearing diagnosis, SNNs outperformed ANNs, which achieved only 61% accuracy. Table 1 shows reviews of SNN techniques for speech analysis, highlighting their limited application in voice analysis and absence in FVC. This gap motivates our research to integrate SNNs for enhanced voice analysis, with the goal of transforming FVC in legal investigations.

Table 1. Literature review of SNN methods

Citation	Dataset	Method	Overview	Results in (%)
Morales <i>et al.</i> [25]	Pure tone samples	SNN, SpiNNaker	Robustness, efficiency in the neuromorphic field	85
Wu <i>et al.</i> [19]	RWCP & TIDIGITS	SOM-SNN	SOM for frequency representation	SNN for spatiotemporal pattern & 97.40, 99.60
Wu <i>et al.</i> [20]	Disagree TIMIT, Librispeech, FAME	SNN, MFCC, FBANK, FMLLR	Large vocabulary recognition	36.9, 18.7
Mukhopadhyay <i>et al.</i> [22]	Human footstep sounds	SNN, WSN	energy efficiency	Time domain for acoustic classification
Kholkin <i>et al.</i> [24]	Accelerometer data	RCNet, ANN, SNN	Ball bearing diagnosis	SNN =100, ANN =61
Yamazaki <i>et al.</i> [23]	Robotics domains	SNN, ANN	SNN vs deep networks & energy efficient applications	Audio classification

### 3. DATA COLLECTION AND EXPERIMENTAL SETUP

For the FVC study, known speech samples and trace data were collected from the University of New South Wales Faculty of Electrical Engineering and Telecommunications in Sydney, Australia. The benchmark dataset used for evaluation consists of Australian English recordings from over 3899 speakers, featuring various styles such as casual telephone conversations, information exchange tasks, and pseudo-police interviews. This dataset was divided into training and testing data, with access granted upon obtaining permission from the relevant authorities. The datasets used in this FVC experiment were sourced from the FVC data repository [26]. The focus on Australian English allows for precise analysis of speech patterns unique to Australian speakers, capturing variations in accent, pronunciation, and other linguistic features essential for reliable voice comparisons in forensic contexts. The data is provided in free lossless audio codec (.flac) file format, and a summary of the experimental data collection is presented in Table 2.

Table 2. Summary of experimental data collection

Dataset name	Number of samples	Training	Testing	Gender	Audio format
Australian English	3899	2729	170	Female & Male	.flac

### 4. THE METHOD - FORENSIC VOICE COMPARISON USING SPIKING NEURAL NETWORK

Figure 1 illustrates the proposed architecture for FVC using SNN. The proposed method in this study tended to have an inordinately higher proportion of the experimental approach employs an SNN model, where the input layer receives raw speech samples in the form of audio files (.flac). These samples are pre-processed to remove noise and other interferences. After pre-processing, feature extraction and classification are performed using an SNN. Once the pre-processed data is fed into the SNN model, features such as time and frequency differences of spikes are extracted and encoded into a format suitable for the SNN. The encoded data is then passed through the SNN, which consists of multiple layers of interconnected neurons,

allowing for detailed analysis and comparison of voice patterns. Each neuron in the SNN integrates incoming spikes from other neurons and emits its own spike when its membrane potential reaches a specified threshold. The timing of these spikes encodes information about the input speech. Finally, the output layer of the SNN generates a set of spikes representing the recognized speech, which is then decoded back into a conventional format, such as a similarity score between the input speech and a reference sample. The performance is evaluated through adjustable threshold spike timing and assessed using a confusion matrix, which includes metrics such as accuracy, precision, recall, F1 score, and F2 score. This section outlines the proposed research methodology for FVC using SNNs. A detailed description of each subsection, including pre-processing and the SNNs model, is provided in the subsequent sections.

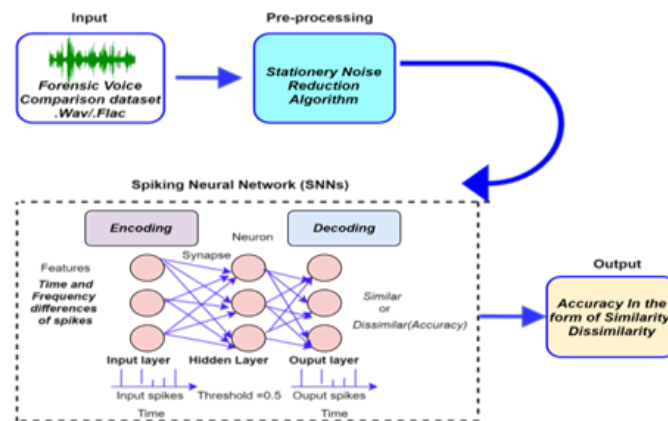


Figure 1. Proposed architecture for FVC using SNN

#### 4.1. Preprocessing

The stationary noise reduction method is employed to eliminate background noise from the forensic voice samples, particularly within the Australian English dataset, which is stored in the .flac file format. These are provided to the model along with a noise sample, encompassing the typical background noise for the sample. This noise sample is combined with a signal clip containing both the noise and the signal that needs to be removed, as illustrated in Figure 2(a) noisy speech input data Figure 2(b) noise reduced speech output data. The following provides an explanation of the stationary noise reduction Algorithm 1

##### Algorithm 1: Stationery noise reduction algorithm

Input: Australian English dataset audio recording samples of voice are used.

Output: Noise Reduced Speech Data.

Step 1: Spectrogram is calculated for the noisy audio clip.

Step 2: In frequency statistics are measured using the noise spectrogram.

Step 3: On the basis of noise statistics a threshold is created.

Step 4: Through the signals spectrogram is calculated.

Step 5: By the signal spectrogram threshold is determined and compared.

Step 6: To smooth the mask over time and frequency the linear filter is used.

Step 7: The mask is applied to the signals spectrogram and inverts the noise signal to produce positive results.

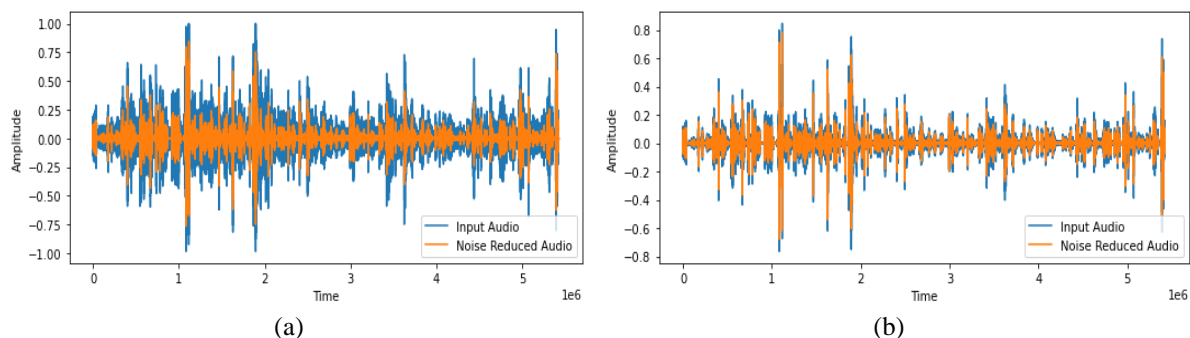


Figure 2. Preprocessing results (a) noisy speech input data (b) noise reduced speech output data

## 4.2. Spiking neural network

The proposed model utilizes SNN inspired by the structure and functionality of biological neural networks found in the human brain. In contrast to conventional neural networks, which depend on signals with continuous values, SNN employ discrete spikes or pulses to communicate information between neurons, resembling the transmission of signals through action potentials in biological systems. Adopting a semi-automatic approach based SNN are considered for FVC in the proposed research model. The process begins with loading and preprocessing the audio data, extracting relevant features and converting continuous audio signals into spike trains using encoding techniques like rate coding or time-to-first-spike coding. Subsequently, synaptic weights of the SNN are initialized either randomly or using pre-trained weights from a neural network. A threshold is applied to determine the similarity or dissimilarity of spike trains. Within the SNN model, neurons communicate through discrete spikes via the membrane potential. Neurons accumulate input over time and emit a spike once the threshold is reached.

Voice patterns are encoded in the timing of spikes, and synaptic weights are updated based on spike timing, commonly through spike-timing-dependent plasticity (STDP) or variants, to identify voice patterns. To decode and predict known and trace voice samples an adjustable threshold is applied to determine the similarity or dissimilarity of spike trains. Several parameters significantly influence SNN behavior and learning, including input spike, membrane potential, spike generation, STDP, and output spike. Input spikes receive information from processed audio samples which then pass through the membrane function. Membrane potential generates spike times and once the spike time surpasses a threshold SNN network identifies voice patterns for classification based on similarity or dissimilarity using STDP. Output spikes generated by neurons aid in evaluating or identifying similar or dissimilar voice samples. The subsections 4.2.1 through 4.2.5 elaborate on these key parameters.

### 4.2.1. Input spikes and membrane potential

The input spikes are derived from pre-processed audio samples, capturing information related to both time and frequency. Consequently, input spikes play a pivotal role in encoding to achieve the desired output through membrane potential. Once the input spike time is generated, the membrane potential indicates the electric potential throughout the voice patterns. It governs the neuron's generation of an action potential spike, with critical factors including its update over time and response to incoming spikes. These factors determine how spikes are generated over time.

Update over time: the membrane potential is dynamic and changes over time in response to incoming signals or spikes. The dynamics of this change are typically described by a set of equations that model how the neuron integrates incoming information. The response to incoming spikes: the membrane potential is influenced by the synaptic inputs received from connected neurons, with each incoming spike contributing to the change in the membrane potential. In (1) represents the membrane potential over time.

$$V(t) = \sum_i w_i * s_i(t - t_i) \quad (1)$$

Where:

- The membrane potential function at time  $t$  is denoted by  $V(t)$ .
- $\sum_i$  denotes the summation over the index  $i$ .
- $w_i$  represents the weight associated with each function.
- $s_i(t-t_i)$  is the spike train from the presynaptic neuron  $i$  at time  $t-t_i$

### 4.2.2. Spike generation

After a spike is generated by the membrane potential, it reaches a threshold that allows the neuron to identify both dissimilarity and similarity in voice patterns. The parameters of spike generation are influenced by threshold crossing, action potential, and neuronal response. Threshold crossing refers to the process where neurons possess a specific threshold level of membrane potential. When the membrane potential surpasses this threshold, the neuron generates a spike, also known as an action potential. The action potential is a brief electrical pulse that travels along the neuron's axon, signaling the neuron's activation to other neurons or target cells. This spike represents an all-or-nothing response: if the membrane potential exceeds the threshold, a spike is generated; otherwise, no spike occurs. Mathematically, the neuron generates spikes when its membrane potential crosses a threshold, represented by  $\theta$  in (2).

$$\text{If } V(t) \geq \theta, \text{ then the neuron emits a spike} \quad (2)$$

### 4.2.3. Spike time dependent plasticity

As the spike reaches the threshold, STDP is utilized to adjust the synaptic weights, assessing the strength and weakness of connections in the SNN network to discern similarity and dissimilarity in

recognizing voice patterns. The strength of a connection (synapse) in an SNN network should alter based on the relative timing of spikes between the presynaptic and postsynaptic neurons. If a presynaptic neuron consistently fires before a postsynaptic neuron, the connection between them strengthens. Conversely, if the postsynaptic neuron fires first, the connection weakens. This process enables the network to adapt to patterns in the input data. The synaptic weights undergo plasticity adjustments based on the timing of pre and postsynaptic spikes as represented in (3).

$$\Delta w_i = \eta \cdot s_i(t - t_i) \cdot \text{PostSynapticSpike}(t) \quad (3)$$

Where:

- $\Delta w_i$  represents the change in the synaptic weight  $w_i$ .
- $\eta$  is the learning rate, controlling the magnitude of weight adjustments.
- $s_i(t - t_i)$  is the function associated with the timing of the presynaptic spike at time  $t_i$ .
- $\text{PostSynapticSpike}(t)$  is a function that indicates if a spike occurred in the postsynaptic neuron at time  $t$ .

#### 4.2.4. Output spike

The output spike aids in decoding voice similarity or dissimilarity, with performance evaluated using accuracy, precision, recall, F1 score, and F2 score. Membrane potential updates integrate signals, while spike generation indicates neuron activation. Synaptic plasticity enables adaptive learning, mimicking biological neural systems. These steps are outlined in Algorithm 2.

**Algorithm 2:** To identify the similarity or dissimilarity of voices through the preprocessed data

Input: Preprocessed forensic voice samples.

Output: Prediction for FVC based on the evaluation on Confusion matrix.

Step 1: Initialization

- Synaptic Weights, Membrane Potentials, and Thresholds: Set the initial values for synaptic weights, membrane potentials, and thresholds for all neurons. In this case, they are initialized to 0.5.
- Learning Rate ( $\eta$ ): Choose a learning rate parameter ( $\eta$ ) to control the magnitude of weight adjustments during the learning process. Here, it is set to 0.001.

Step 2: Training

- Adjusting Synaptic Weights: Utilize a learning rule based on spike time to update synaptic weights. The spike timing difference based on the adjustable threshold.

$$\text{spike\_times} = (y > 0.5).\text{nonzero}()[0] * 0.001 \text{ \# if spike amplitude} > 0.5 \quad (4)$$

In the provided code for voice comparison, the exact time of a spike is determined based on the threshold condition.

Where,

- $y$  is the audio signal.
- $(y > 0.5)$  creates a binary mask where the amplitude values greater than 0.5 are marked as True and others as 'False'.
- $\text{.nonzero}()$  returns the indices where the condition is True.
- $*0.001$  scales the indices to represent time in seconds (assuming the audio is sampled at 1,000 Hz).
- The resulting `spike_times` variable contains the times (in seconds) when the amplitude of the audio signal exceeds the threshold of 0.5. These times correspond to the occurrences of spikes in the audio signal, as determined by the chosen threshold. The specific value of 0.5 can be adjusted according to the characteristics of voice samples and the desired sensitivity of spike detection.
- Presenting Training Samples: Introduce training samples to the network. These samples represent patterns or data points that the network will learn to recognize or classify. By using the library functions such as tensor flow and pytorch the SNN network is built. Where the optimization function used is Adam and binary cross entropy is the loss function.

Step 3: Testing

The following pseudocode is utilized to identify and evaluate voice samples between the known and trace. In this context, 0 represents false, and 1 represents true, indicating whether the suspect is identified through the voice samples. This evaluation is performed using accuracy to assess the similarity and dissimilarity in the voice.

```
defevaluate_voice_samples (known_sample, trace_sample):
```

```
    If known_sample == trace_sample:
```

```
        return 1 # True, suspect identified
```

```
    else:
```

```
        return 0 # False, suspect not identified
```

```
    # Example usage
```

```
    known_sample = "voice_sample_1"
```

```

trace_sample = "voice_sample_2"
result =evaluate_voice_samples (known_sample, trace_sample)
print ("Result:", result)
Step 4: Inference of proposed research model
The proposed research model identifies the suspect based on similarity or dissimilarity.

```

## 5. RESULT AND DISCUSSION

The performance evaluation of the proposed framework is conducted using various metrics, with a detailed analysis presented in the form of a confusion matrix and receiver operating characteristic-area under the curve (ROC-AUC) analysis. These performance assessments are systematically discussed in sections 5.1 and 5.2, where the confusion matrix provides insights into classification accuracy by displaying true positives, false positives, true negatives, and false negatives. Meanwhile, the ROC-AUC analysis evaluates the model's discriminative ability, illustrating its effectiveness in distinguishing between different classes.

### 5.1. Confusion matrix and performance analysis

The Figure 3(a) displays the confusion matrix performance evaluation of the SNN-based approach in classifying the Australian English dataset for FVC. Performance metrics such as accuracy, precision, recall, F1 score, and F2 score are used to evaluate the classifier's effectiveness. The confusion matrix displays the number of correctly identified matching voices as true positives (TP) entries in the bottom right quadrant of matrix indicate accurate match predictions. The true negative (TN) refers to the entries in the upper-left quadrant of the matrix that accurately predict non-matches. Comparably, false positives represented by the entries in the upper-right quadrant of the matrix, which indicate inaccurate match predictions, while false negatives represented by the entries in the bottom-left quadrant of the matrix, which indicate inaccurate predictions of non-matches. In the context of FVC using SNN the Figure 3(a) represented the actual confusion matrix for considering 2,729 samples for training and 1,170 samples for testing in the context of identification of matching or non-matching of the voice. A value of 0 signifies dissimilarity in voice recognition, while a value of 114 signifies similarity, indicating the identification of a suspect using 1,170 testing samples as represented in confusion matrix depicted in the Figure 3(a). In confusion matrix the values in each cell indicate the number of instances. The heatmap visualization with `sns.heatmap` provides a color-coded representation, where darker shades indicate higher counts, helping to quickly identify the performance of the model. The Figure 3(a) is visual represents of values tabulated in Table 3(a).

Compared the current state of art to the existing works [1]-[3], identified the deep learning techniques for our proposed method which demonstrates a potential enhancement in classification accuracy, achieving an impressive 94.21% on the Australian English dataset. Additionally, the model's effectiveness is further supported by its strong performance across multiple evaluation metrics, with a precision of 85.21%, recall of 82.16%, F1 score of 81.11%, and F2 score of 80.10% across varied samples. These results indicate the robustness and reliability of our approach in FVC. The findings are effectively presented and visually depicted in Figure 3(b), accompanied by detailed numerical data in Table 3(b).

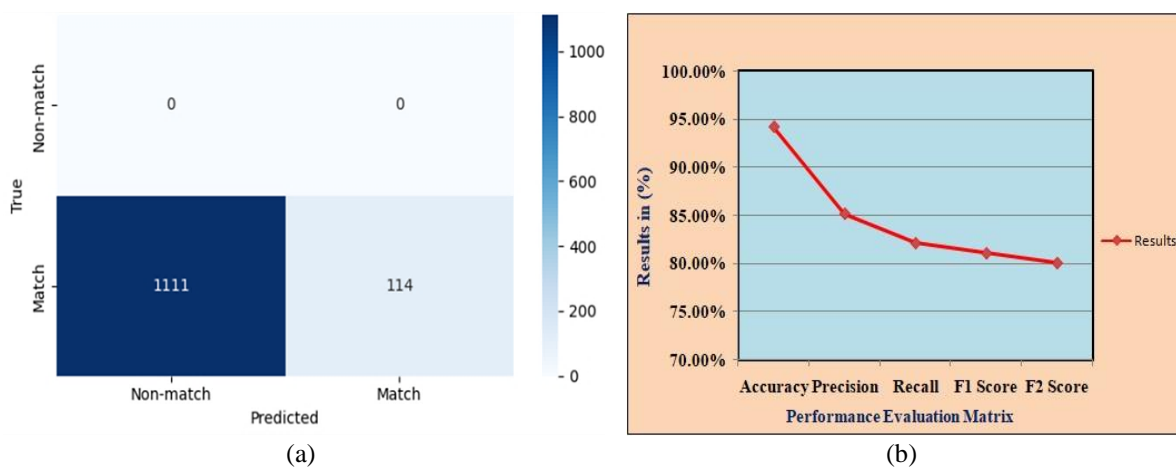


Figure 3. Graphical representation of (a) confusion matrix and (b) performance measures

The SNN are ANN inspired by biological neurons, utilizing spikes for communication rather than weighted sums of inputs in traditional ANN. The SNN performance is assessed using the confusion matrix; encompassing accuracy, precision, recall, F1 score, and F2 score. Moreover, the graph portrays consistent performance across all metrics, indicating robustness and generalization of the SNN model to unseen data. This suggests that the SNN effectively avoids over fitting and maintains good performance on varied datasets. In summary, Figure 3(b) underscores the SNNs potential as promising neural network architecture for proficiently identifying voice patterns in FVC tasks.

Table 3. Tabulation of results: (a) confusion matrix and (b) performance measure

(a)		
	Predicted non match	Predicted match
Actual non match	True negatives	False positives
Actual match	False negatives	rue positives

(b)		
Dataset name	Performance measures	Results (%)
Australian English	Accuracy	94.21
	Precision	85.21
	Recall	82.16
	F1 score	81.11
	F2 score	80.10

## 5.2. Receiver operating characteristic and area under the curve analysis

Furthermore, the ROC curve serves as a tool for assessing the performance of the classification model, specifically the SNN. The degree of separability is indicated by the AUC of the ROC curve. The ROC-AUC curve provides a graphical representation illustrating the classification performance metrics at various thresholds. The ROC curve is a graphical plot that depicts the performance of binary classifier system as its threshold varies. It is extensively utilized in machine learning to evaluate the diagnostic capability of tests, especially in scenarios with imbalanced outcomes. In this context, the binary classifier is represented by the SNN, which aims to differentiate between two classes: true positives and false positives, as represented in Figure 4(a).

The true positive rate (TPR), displayed on the y-axis, signifies the proportion of correctly classified positive cases. Conversely, the false positive rate (FPR), depicted on the x-axis, represents the proportion of incorrectly classified negative cases. The area under the ROC curve (AUC) serves as a measure of the overall performance of the classifier. A perfect discrimination is represented by an AUC of 1, while an AUC of 0.94 indicates the degree of dissimilarity and similarity in the voice patterns. The red dot on the curve denotes the point where the TPR equals the FPR. This point is commonly referred to as the "operating point" of the classifier, where the classifier strikes a balance between true positives and false positives, making it the optimal threshold for classification.

The blue line in the graph represents the ROC for a random classifier. A random classifier typically produces an AUC of 0.5, resulting in a diagonal line on the ROC curve. However, in this case, due to a specific scenario, the random classifier's AUC is mentioned as 0.94. The observation that the SNN ROC curve lies above the ROC curve of the random classifier implies that the SNN exhibits superior performance compared to random chance. Nevertheless, the margin of improvement is relatively modest. Overall, the graph indicates that the SNN performs satisfactorily in voice comparison tasks, with an AUC close to 94.21%. The AUC represents the entire two-dimensional area beneath the ROC curve, signifying the classifier's overall performance. Mathematically, the AUC is determined by calculating the definite integral of the function  $f(x)$  with respect to the vertical boundaries, as described by (5).

$$AUC = \int_a^b f(x)dx = F(b) - F(a) \quad (5)$$

Where:

- $\int_a^b f(x)dx$  denotes the definite integral of the function  $f(x)$  over the interval from  $a$  and  $b$
- $F(x)$  represents the antiderivative of  $f(x)$  often referred to as the cumulative distribution function (CDF).
- $F(b)-F(a)$
- It computes the difference between the antiderivative values at the upper ( $b$ ) and lower ( $a$ ) bounds, representing the accumulated area under the curve within the given interval.



A larger ROC-Area signifies improved accuracy of the classifier in identifying individuals. Figure 4(a) displays several subsequent ROC and AUC graphs generated during the analysis. The graphical histogram representations of the voice comparison in SNN are shown in Figure 4(b). The voice calculated distribution in relation to amplitude and time. Peaks in the histogram indicate the range and distribution of amplitudes in a voice pattern or audio signal. In Figure 4(b), audio A and audio B represent the voice sample. Where the amplitude represents the strength or intensity of the sound wave with higher amplitudes corresponding to louder voice. The voice pattern is encoded in SNN when the frequency of the pattern reaches a certain threshold. Here the frequency of the threshold is 0.5 seconds is taken for the experimentation. During this spike time the voice patterns are identified to know the similarity of the known and trace.

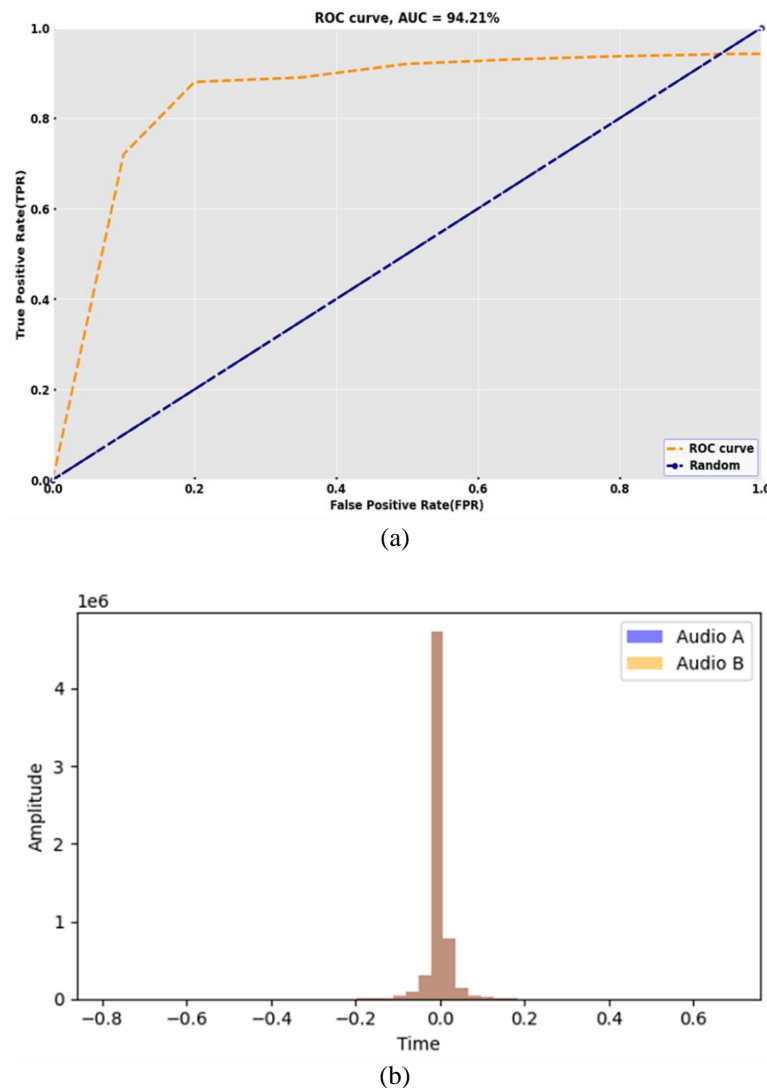


Figure 4. Graphical representation of SNN results (a) ROC-AUC graph and (b) audio signal histogram plot

## 6. COMPARISON WITH EXISTING WORK

The Table 4 presents a comprehensive comparative analysis of existing research studies alongside the proposed SNN approach for the Australian English datasets. The main advantage of SNN for FVC is their ability to process speech data in a way that closely resembles how the human brain processes sound. SNNs are highly efficient at capturing the timing details in speech, such as pauses, pitch, and intonation, which are essential for distinguishing between speakers accurately. They also use less energy than traditional neural networks, making them ideal for real-time applications. Moreover, SNNs are robust in noisy environments, which is beneficial when analyzing low-quality audio recordings often encountered in forensic cases. In our

comparative study, the proposed framework achieved a similarity accuracy of 94.21%, outperforming existing methods. These features make SNNs particularly well-suited for the complex challenges of forensic voice analysis. Overall, a SNN model may suffice, particularly for capturing temporal patterns in spike times for voice pattern recognition. The comparison underscores the versatility of SNN models across different applications, exhibiting varying levels of accuracy, efficiency, and robustness contingent upon the specific dataset and application context.

Existing research studies utilize diverse datasets, ranging from environmental sound datasets like RWCP and spoken digits datasets such as TIDIGITS to more specialized datasets like those from industrial sensor data analysis. For instance, methodologies like the SOM-SNN model achieve impressive accuracies of 97.40% and 99.60% on the TIDIGITS spoken digits and RWCP environmental sound datasets, respectively. Conversely, SNNs employed for acoustic modeling demonstrate lower accuracy rates of 36.9% and 18.7% on the TIMIT Corpus and Librispeech datasets. In a study focused on ball bearing diagnosis, an ANN achieves 61% accuracy. Furthermore, a multilayer SNN designed for audio sample classification using SpiNNaker exhibits robustness and efficiency in neuromorphic engineering, achieving an accuracy of 85%. However, these methodologies often integrate advanced SNN technology with other models. In contrast, the proposed research exclusively employs SNN, which is advantageous considering factors like resource constraints and the unavailability of advanced systems like GPU. The SNN with adjustable threshold effectively determines the similarity or dissimilarity of spike trains in voice samples. Performance evaluation utilizing a confusion matrix with its extended metric values like accuracy 94.21%, precision 85.21%, recall 82.16%, F1 score 81.11%, and F2 score 80.10% are achieved. There are challenges in a proposed framework of SNNs for FVC such as complex and time-consuming training due to spike-based learning mechanisms like STDP. They also require specialized neuromorphic hardware, limiting accessibility.

Table 4. Comparison of the proposed approach with existing work

Citation	Dataset	Method	Results in (%)
Morales <i>et al.</i> [25]	Pure tone samples	SNN, SpiNNaker	85
Wu <i>et al.</i> [19]	RWCP, TIDIGITS Disagree	SOM-SNN	SNN for spatiotemporal pattern & 97.40, 99.60
Wu <i>et al.</i> [20]	TIMIT, Librispeech, FAME	SNN, MFCC, FBANK, FMLLR	36.9, 18.7
Auge <i>et al.</i> [21]	keyword spotting	SNN, MFCC	80
Kholkin <i>et al.</i> [24]	Accelerometer Data	RCNet, ANN, SNN	SNN=100%, ANN=61
Proposed approach model	Australian English	SNN	94.21

## 7. CONCLUSION

The work proposed explores the potential of FVC to enhance suspect identification using forensic speech recordings. It applies an SNN model to analyze an Australian English dataset of 3,899 .flac file recordings, utilizing stationary noise reduction for pre-processing. The SNN model uses a threshold to assess spike train similarities, where neurons communicate via discrete spikes through membrane potentials. Synaptic weights, updated using STDP or its variants, help recognize and decode voice patterns. Our findings provide conclusive evidence that this phenomenon is associated with a SNN model achieves 94.21% accuracy. For future studies may investigate on refining the SNN architecture to enhance real-world forensic applications.

## ACKNOWLEDGEMENTS

The authors gratefully to Rose P., Zhang C., and Geoffrey Stewart Morrison from the FVC Laboratory, UNSW, Sydney, Australia, for providing access to their database for this study.

## FUNDING INFORMATION

This research is supported by the Department of Science and Technology, New Delhi, India, through the fundamental research grant scheme DST WISE-Kiran fellowship for Ph.D. File No: DST/WISE-PhD/ET/2023/4 (G).

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Kruthika Siddanakatte Gopalaiah	✓	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓
Trisiladevi Chandrakant Nagavi		✓		✓	✓		✓			✓	✓	✓		✓
Parashivamurthy Mahesha		✓		✓	✓		✓			✓	✓	✓		✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

**CONFLICT OF INTEREST STATEMENT**

Authors state no conflict of interest with respect to this article.

**INFORMED CONSENT**

This study does not require any informed consent.

**ETHICAL APPROVAL**

Authors state no ethical approval for this study.

**DATA AVAILABILITY**

The datasets utilized in this FVC experiment were sourced from the FVC data repository, available at <http://databases.forensic-voice-comparison.net/>.

**REFERENCES**

[1]

S. Ekhande, U. Patil, and K. V. Kulhalli, "Review on effectiveness of deep learning approach in digital forensics," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 5, pp. 5481–5492, 2022, doi: 10.11591/ijece.v12i5.pp5481-5592.

[2]

T. Sutikno and I. Busthomi, "Capabilities of celebrete universal forensics extraction device in mobile device forensics," *Computer Science and Information Technologies*, vol. 5, no. 3, pp. 254–264, 2024, doi: 10.11591/csit.v5i3.p254-264.

[3]

M. A. Qamhan, H. Altaheri, A. H. Meftah, G. Muhammad, and Y. A. Alotaibi, "Digital audio forensics: microphone and environment classification using deep learning," *IEEE Access*, vol. 9, pp. 62719–62733, 2021, doi: 10.1109/ACCESS.2021.3073786.

[4]

G. Horsman, "Sources of error in digital forensics," *Forensic Science International: Digital Investigation*, vol. 48, 2024, doi: 10.1016/j.fsidi.2024.301693.

[5]

S. M. Geoffrey, E. Ewald, D. Ramos, J. González-Rodríguez, and A. Lozano-Diez, "Statistical models in forensic voice comparison," *Handbook of Forensic Statistics*, pp. 451–497, 2020, doi: 10.1201/9780367527709-20.

[6]

R. Sharma, D. Govind, J. Mishra, A. K. Dubey, K. T. Deepak, and S. R. M. Prasanna, "Milestones in speaker recognition," *Artificial Intelligence Review*, vol. 57, no. 3, 2024, doi: 10.1007/s10462-023-10688-w.

[7]

S. G. Kruthika and T. C. Nagavi, "Speech Processing and Analysis for Forensics and Cybercrime: A Systematic Review," in *Cybercrime in Social Media*, 2023, pp. 191–224, doi: 10.1201/9781003304180-10.

[8]

D. Sztahó and A. Fejes, "Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings," *Journal of Forensic Sciences*, vol. 68, no. 3, pp. 871–883, 2023, doi: 10.1111/1556-4029.15250.

[9]

R. Potapova, V. Potapov, and I. Kuryanova, "Analysis of formant trajectories of a speech signal for the purpose of forensic identification of a foreign speaker," in *Speech and Computer: 25th International Conference, SPECOM 2023*, 2023, pp. 287–300, doi: 10.1007/978-3-031-48309-7\_24.

[10]

J. C. Cavalcanti, R. R. da Silva, A. Eriksson, and P. A. Barbosa, "Exploring the performance of automatic speaker recognition using twin speech and deep learning-based artificial neural networks," *Frontiers in Artificial Intelligence*, vol. 7, 2024, doi: 10.3389/frai.2024.1287877.

[11]

D. R. Edla, A. Bablani, S. Bhattacharyya, R. Dharavath, R. Cheruku, and V. Boddu, "Spatial spiking neural network for classification of EEG signals for concealed information test," *Multimedia Tools and Applications*, vol. 83, pp. 79259–79280, 2024, doi: 10.1007/s11042-024-18698-8.

[12]

P. Leferink, "Transfer learning in spiking neural networks," Doctoral dissertation, University of Groningen, Groningen, Netherlands, 2023.

[13]

X. Chen, Q. Yang, J. Wu, H. Li, and K. C. Tan, "A hybrid neural coding approach for pattern recognition with spiking neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3064–3078, 2024, doi: 10.1109/TPAMI.2023.3339211.

[14]

M. Umair, W. H. Tan, and Y. L. Foo, "Efficient malware classification with spiking neural networks: a case study on N-BaIoT dataset," in *International Conference on Ubiquitous and Future Networks, ICUFN*, 2023, pp. 231–236, doi: 10.1109/ICUFN57995.2023.10200941.





[15]

M. Dong, X. Huang, and B. Xu, "Unsupervised speech recognition through spike-timing-dependent plasticity in a convolutional spiking neural network," *PLoS ONE*, vol. 13, no. 11, 2018, doi: 10.1371/journal.pone.0204596.





- [16] Y. Wang, "Construction and improvement of English vocabulary learning model integrating spiking neural network and convolutional long short-term memory algorithm," *PLoS ONE*, vol. 19, no. 3, 2024, doi: 10.1371/journal.pone.0299425.
- [17] Z. Roozbehi, A. Narayanan, M. Mohaghegh, and S. A. Saeedinia, "Dynamic-structured reservoir spiking neural network in sound localization," *IEEE Access*, vol. 12, pp. 24596–24608, 2024, doi: 10.1109/ACCESS.2024.3360491.
- [18] S. Carmo *et al.*, "Forensic analysis of auditorily similar voices," *Revista CEFAC*, vol. 25, no. 2, 2023, doi: 10.1590/1982-0216/20232524022.
- [19] J. Wu, Y. Chua, M. Zhang, H. Li, and K. C. Tan, "A spiking neural network framework for robust sound classification," *Frontiers in Neuroscience*, vol. 12, 2018, doi: 10.3389/fnins.2018.00836.
- [20] J. Wu, E. Yilmaz, M. Zhang, H. Li, and K. C. Tan, "Deep spiking neural networks for large vocabulary automatic speech recognition," *Frontiers in Neuroscience*, vol. 14, 2020, doi: 10.3389/fnins.2020.00199.
- [21] D. Auge, J. Hille, F. Kreutz, E. Mueller, and A. Knoll, "End-to-end spiking neural network for speech recognition using resonating input neurons," in *Artificial Neural Networks and Machine Learning – ICANN 2021*, 2021, pp. 245–256, doi: 10.1007/978-3-030-86383-8\_20.
- [22] A. K. Mukhopadhyay, M. P. Naligala, D. L. Duggisetty, I. Chakrabarti, and M. Sharad, "Acoustic scene analysis using analog spiking neural network," *Neuromorphic Computing and Engineering*, vol. 2, no. 4, 2022, doi: 10.1088/2634-4386/ac90e5.
- [23] K. Yamazaki, V. K. Vo-Ho, D. Bulsara, and N. Le, "Spiking neural networks and their applications: a review," *Brain Sciences*, vol. 12, no. 7, p. 863, 2022, doi: 10.3390/brainsci12070863.
- [24] V. Kholkin, O. Druzhina, V. Vatrik, M. Kulagin, T. Karimov, and D. Butusov, "Comparing reservoir artificial and spiking neural networks in machine fault detection tasks," *Big Data and Cognitive Computing*, vol. 7, no. 2, 2023, doi: 10.3390/bdcc7020110.
- [25] J. P. D. -Morales *et al.*, "Multilayer spiking neural network for audio samples classification using SpiNNaker," in *Artificial Neural Networks and Machine Learning – ICANN 2016*, 2016, pp. 45–53, doi: 10.1007/978-3-319-44778-0\_6.
- [26] G. S. Morrison *et al.*, "Forensic database of voice recordings of 500+ Australian English speakers," *Forensic Voice Comparison Databases*. 2015. [Online]. Available: <https://forensic-voice-comparison.net/databases/>.

## BIOGRAPHIES OF AUTHORS







**Kruthika Siddanakatte Gopalaiah**     is a Ph.D. scholar in Department of Computer Science and Engineering at JSS STU, Mysuru. She earned her M.Tech. in CSE from GSS College of Engineering, Bangalore, in 2014. Currently, she is a full-time research scholar in the Department of Computer Science and Engineering at SJCE. Awarded the WISE-Kiran for Ph.D. Women Scientist fellowship by the DST, New Delhi, her research interests include digital forensics, speech signal processing, artificial intelligence, and machine learning. She can be contacted at email: [sgkruthi@jssstuniv.in](mailto:sgkruthi@jssstuniv.in).



**Dr. Trisiladevi Chandrakant Nagavi**     is an Associate Professor in the Department of Computer Science and Engineering at SJCE, JSS STU, Mysuru. She holds UG degree from Karnataka University and PG degree from VTU Belgaum. Her expertise encompasses audio, music, speech and image signal processing, digital signal forensics, and machine learning. She is an active IEEE member. She can be contacted at email: [trisiladevi@sjce.ac.in](mailto:trisiladevi@sjce.ac.in).



**Dr. Parashivamurthy Mahesha**     is an Associate Professor in the Department of Computer Science and Engineering. His research interests include speech signal processing, machine learning, data analytics, and digital signal forensics. He has presented and published papers in reputable conferences and journals, serves on international conference committees, and reviews for journals. He holds a BE from the UOM and an M.Tech. and Ph.D. from VTU, Belgaum. He can be contacted at email: [maheshap@sjce.ac.in](mailto:maheshap@sjce.ac.in).