Available *online* at : http://jnte.ft.unand.ac.id/

**Jurnal Nasional Teknik Elektro**

| ISSN (Print) 2302-2949 | ISSN (Online) 2407-7267 |

# A Hybrid Wavelet Scattering and Mel Spectrogram Feature with Deep Convolution Neural Network for Robust Spoken Digit Recognition

*Irmawan [1], Suci Dwijayanti [2], Bhakti Yudho Suprapto[3]*

[1]Basic Electronics and Electrical Circuits Laboratory, University of Sriwijaya, Inderalaya, South Sumatera, Indonesia
[2]Department of Electrical Engineering, Faculty of Engineering, University of Sriwijaya, Inderalaya, South Sumatera, Indonesia
[3]Department of Electrical Engineering, Faculty of Engineering, University of Sriwijaya, Inderalaya, South Sumatera, Indonesia

ARTICLE INFORMATION

Received: March 06, 2025
Revised: December 04, 2025
Accepted: December 12, 2025
Available online: December 12, 2025

KEYWORDS

Spoken digit recognition, Deep CNN, Wavelet Time Scattering, MFCC, Biometric

CORRESPONDENCE

Phone: +62 81373331366

E-mail: irmawan@unsri.ac.id

**A B S T R A C T**

Spoken digit recognition (SDR) plays a critical role in biometric authentication and human–computer interaction, yet existing approaches often rely on small datasets, limited feature representations, or architectures prone to overfitting. To address these limitations, this study proposes a robust end-to-end pipeline that integrates Wavelet Time Scattering (WTS), Mel-Frequency Cepstral Coefficients (MFCC), and a 2D Deep Convolutional Neural Network (2D-CNN) to enhance the accuracy and generalization of SDR systems in realistic environments. The Free-Spoken Digit Dataset (FSDD), consisting of 3000 audio samples from speakers with diverse accents, was pre-processed using zero-padding normalization and transformed into high-resolution time–frequency spectrograms via WTS. The proposed CNN architecture, optimized through systematic experimentation on batch size and learning rate, demonstrated stable convergence and superior discriminative capability. Using a learning rate of 0.001 and a batch size of 50, the model achieved the highest performance with 99.2% accuracy, outperforming established methods including SVM, MFCC-LSTM, and Multiple RNN architectures. Comparative evaluations further revealed that the combined WTS–MFCC feature extraction significantly enhances spectral–temporal representation quality, contributing to improved classification precision across all digit classes. These findings demonstrate that the proposed WTS-MFCC-CNN framework not only advances SDR accuracy but also provides a scalable and computationally efficient approach suitable for real-world biometric, financial, and voice-controlled applications. The results highlight the potential of hybrid time–frequency representations integrated with deep architectures to set a new benchmark for robust spoken digit recognition.

## INTRODUCTION

Human speech is defined as one-dimensional communication signal which has the ability to function as s biometric tool mostly due to the variation in the samples of words spoken based on accent, age, gender, and language. It is important to note that more complex variability occurs at the speech signal level as indicated by the differences in the amplitude, duration, pitch, frequency, timbre, and speaker. This complicates the process of analyzing the speaker but provides more useful information on the amplitude and tone [1].

The speech of humans can also be applied in the process of verifying the identity of a person for security and commercial purposes [2]. This can be achieved through speaker recognition technology which applies phonetic attributes of utterances in the process of determining a speaker's identity. Moreover, voice biometric systems have also been classified based on different industries and applications as indicated by speaker diarization, identification, and verification [3]. Voice biometrics has also been applied in different ways to verify and identify speakers. It is also important to note that data analysis methods in image classification, biomedical applications [4], [5], bioinformatics, medical image analysis [6], and computer vision [7] have advanced significantly over the last decade.

https://doi.org/10.25077/jnte.v14n3.1310.2025

Spoken digit recognition has been applied in different areas such as the retrieval and analysis of audio contents, entry of the number for credit cards, dialing of voices, and entry of data [8], [9] . Less attention has, however, been placed on this technology with only a few related works observed in [10], [11]. For example, the TIDIGITS dataset containing 2.412 training and 1.144 test utterances were used in [10] while OGI Multilanguage Corpus applied 826 and 454 respectively in [12]. Mel-frequency cepstral coefficients (MFCC) were observed to have been used as the features, principal component analysis (PCA) to lower the features dimension and support vector machine (SVM) to categorize.

The other areas where it has been used include investigation of crimes and offering of financial services as indicated by the application of voice biometrics in call centers for the purpose of authenticating customers (customer protection) [13]. The process of comparing the input voice samples to the reference voice sample in the database is known as speaker verification and this is classified into two types which include text-dependent and text-independent verification [14]. For the text-dependent type, the text is preserved to verify the voice and this is considered to provide a more accurate result compared to the text-independent which does not rely on text for verification.

The spoken digit recognition technology that converts spoken signals into characters understood by computers is widely used in everyday life and is important in human-computer interaction. One of the most important tasks usually performed using the technology is digit recognition because numbers contain a lot more valuable information than other words people usually speak. Some related works on spoken digit recognition were observed to have used small datasets and traditional feature extraction and classification methods. Meanwhile, overfitting and poor generalization is possible to occurr in models trained on small datasets. Deep learning techniques were discovered to have recently surpassed most feature extraction, selection, and classification techniques. For example, they were applied to 30,000 audio digit samples in the most comprehensive and recent related works [11] but the dataset lacks non-digit audio samples which are critical in a realistic setting.

The process of converting a speech signal to text has been extensively researched using different methods. The first was the application of template matching [15] while Recurrent Neural Networks (RNN) and Long Short Term Memory Networks (LSTM) are commonly applied in Automatic Speech Recognition (ASR) [16], [17] due to their ability to adapt to time-warped data and bridge long time lags. However, training RNN or LSTM can be difficult due to their complex architecture which requires dividing the input data into subsections and feeding them into the network separately. These methods can perform significantly better in recognizing a sentence but other neural networks also have the ability to achieve good results when recognizing a single number. It is important to note that even though Artificial Neural Networks (ANN) are widely used in spoken digit recognition, their accuracy is lower than that of Convolutional Neural Networks (CNN) which are normally used in some cutting-edge methods to extract features from a spoken sentence's short-time Fourier transform (STFT) [18], [19] that mostly focuses on recognizing a spoken sentence using an ANN or a Hidden Markov Model (HMM). There are no previous studies that use CNN and the Mel Frequency Cepstrum Coefficient (MFCC) feature to classify spoken digits. Therefore, this research creates a pipeline for spoken digit recognition by first extracting the raw speech signals' time and frequency features and later feeding them into the proposed deep neural network architecture.

The baseline methods in this research include the spectrogram, smoothed-spectrogram, and Mel-spectrogram. Moreover, wavelet time scattering was applied to investigate the formation of time-frequency representations due to its ability to provide better frequency localization in the lower frequency range when compared to conventional techniques, thereby, it more suitable for speech classification tasks. Furthermore, several time-frequency representations were used to indicate the spectral information at different frequencies. The combination of the learning from these representations has the ability to assist in the improvement of classification performance. Moreover, late fusion was also applied to make more informed predictions.

The proposed method was tested using a Free-Spoken Digit Dataset (FSDD) dataset. This is important due to the fact that an automated spoken digit recognition system is required to have the ability of accurate spoken digit detection as well as the rejection of non-digits and several other background noises. Therefore, the audio dataset applied includes both non-digit and spoken digit files, thereby, indicating it is very realistic and challenging.

Human speech exhibits substantial intra- and inter-speaker variability in amplitude, duration, pitch, timbre, and spectral structure, making spoken-digit recognition a non-trivial pattern recognition task. Traditional approaches—such as MFCC combined with PCA, SVM, template matching, or even conventional ANN and HMM models—often struggle to generalize well, especially when trained on small datasets or when exposed to acoustically diverse environments. These limitations arise from their dependence on hand-crafted features, restricted modeling capacity, and the inability to robustly capture local spectral variations essential for distinguishing similar-sounding digits. In contrast, Convolutional Neural Networks (CNNs) offer a compelling alternative because they can automatically learn hierarchical time–frequency representations from spectrogram-based inputs, thereby improving the interpretability and discriminability of speech features. CNNs also excel at capturing localized patterns in speech signals—such as formant transitions, harmonic structures, and transient components—which directly enhance classification accuracy. Recent advances in deep learning have consistently demonstrated that CNNs outperform traditional methods in tasks involving complex, high-dimensional audio signals. Therefore, employing CNN architectures for spoken digit recognition is justified not only by their proven robustness against noise and speaker variability but also by their superior ability to extract meaningful, readable, and generalizable representations from raw or transformed speech signals. This establishes CNNs as a suitable and high-impact solution for improving recognition performance in modern spoken digit recognition systems.

## METHODS

The entire methods used for the proposed Spoken Digit Recognition model are presented in Figure 1. It is important to note that the Free Spoken Digit Dataset (FSDD) shared the original spoken signal [20]. The data records were made from the Spoken Digit signals used as input which were all 10 seconds long. This was followed by the transformation of each Spoken Digit signal record into a time-frequency spectrogram image using the Wavelet Time Scattering (WTS). Moreover, the signals from the spoken digit spectrogram were fed into the proposed Deep Convolution Neural Network (DCNN) model which automatically and intelligently classifies the spoken digit.

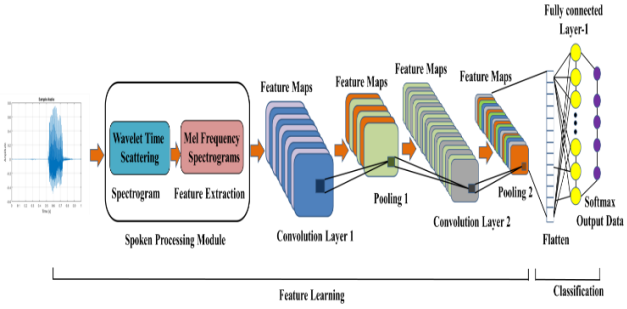


Figure 1. Overall Spoken Digit Recognition procedures based on proposed DCNN


https://doi.org/10.25077/jnte.v14n3.1310.2025

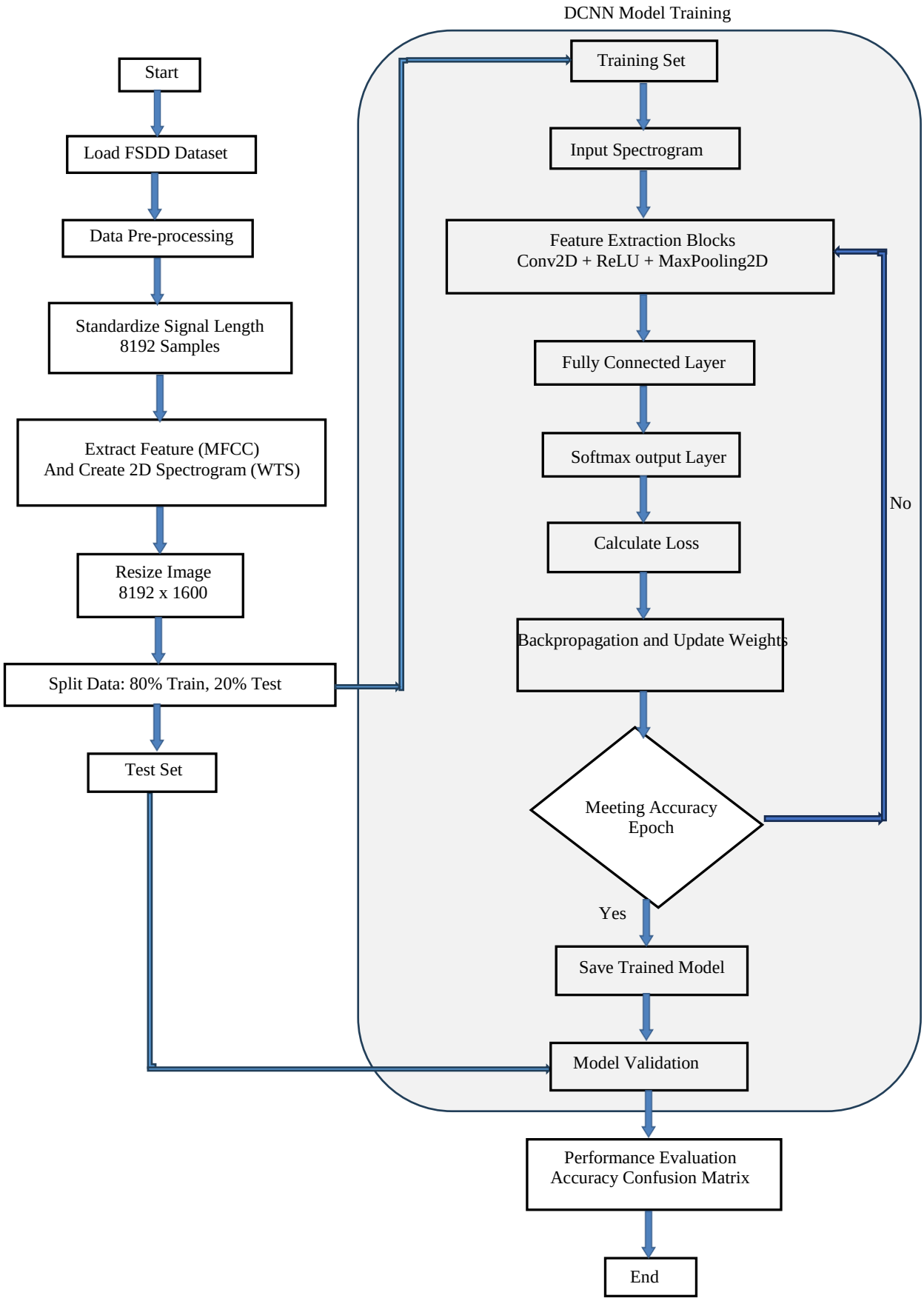


Figure 2. Block diagrams and control program flow diagrams

Available *online* at : http://jnte.ft.unand.ac.id/

**Jurnal Nasional Teknik Elektro**

| ISSN (Print) 2302-2949 | ISSN (Online) 2407-7267 |

DCNN is a popular DNN architecture typically trained using a gradient-based optimization algorithm [21] while a CNN is made up of multiple back-to-back layers linked in a feed-forward fashion. The main ones include convolutional, normalization, pooling, and fully connected layers. The first three are in charge of feature extraction while the last focuses on classification.

## DATA ACQUISITION AND SELECTION

The FSDD is an open dataset that can change over time and is also observed to contain 3000 records in English from 0 to 9 obtained from six speakers out of which two are native American English speakers, two are native English with a Belgian French accent, and two are native English with a German accent. The data were collected at an 8000 Hz sampling frequency [20].

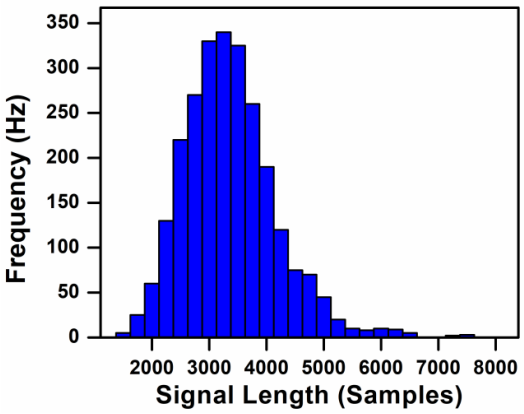


Figure 3. Histogram of signal length.

Figure 2 illustrates the block diagrams and control program flow diagrams of the overall method stages used for the proposed Spoken Digit Recognition model, which consists of four main stages: Pre-processing, training, testing, and accuracy measurement of the proposed model.

The FSDD data set is made up of ten balanced classes, each with 300 records with different duration of records. Moreover, a histogram of signal length was read and created because the FSDD is not a large file. The histogram shows that the distribution of recording length is positively skewed. It is also important to note that the classification was made using an 8192-sample common signal length which is a conservative value that ensures the truncation of longer recordings does not cut short the speech content.

Any situation the signal exceeds 8192 samples (1.024 seconds) is expected to lead to the truncation of the recording to 8192 samples and when the figure is less, the signal is symmetrically pre-padded and post-padded with zeros out to reach this number.

## SPOKEN DIGIT DATA PRE-PROCESSING

The input data for the proposed 2D-CNN is required to be in image format. This led to the transformation of time domain spoken digit signals to 2D time-frequency spectrograms through the application of Wavelet Time Scattering (WTS). It is also important to note that there was a computation of continuous wavelet transform for the time-domain audio signal r(t) at scale s and position u through the following equation.

$$W_{\psi}(u,s) = \int_{-\infty}^{\infty} r(t)\frac{1}{\sqrt{s}}\Psi * \left[\frac{t-u}{s}\right] dt \quad (1)$$

Where, Ψ represents mother wavelet [22]. It is important to note that the analytic wavelets used in this study are Morse, bump, and Morlet wavelets [23].

The absolute value of the complex wavelet transform was computed while the time-frequency representation was resized to 8192x1600 dimensions to serve as the input for the CNN. It is pertinent to state that interpolation which is a technique normally applied to process digital image was adopted to resize the images. Bicubic interpolation was observed to be efficient in resizing time-frequency images compared to other different available interpolation kernels [24]. The interpolated surface with bicubic interpolation is as follow:

$$R(x,y) = \sum_{i=0}^{n}\sum_{j=0}^{m} a_{ij}x^i y^j \quad (2)$$

This necessitates the computation of the coefficients $a_{ij}$. It is possible to compute interpolation in both dimensions by performing a convolution with the kernel presented as follows [25].

$$k(x) = \begin{cases} \frac{3}{2}|x|^2 - \frac{5}{2}|x|^2 + 1, & |x| \le 1 \\ -\frac{1}{2}|x|^3 + \frac{5}{2}|x|^2 - 4|x| + 2, & 1 < |x| \le 2 \\ 0, & otherwise \end{cases} \quad (3)$$

Figure 4 shows the time-domain signal plot for the spoken digit zero as well as its spectrogram and scalogram representations.
A wavelet time scattering frame was created with a scale invariance of 0.22 seconds. Moreover, a feature vector was generated by averaging the scattering transformation across all time samples in order to have a sufficient number of scattering coefficients to average per time window. The FSDD was also separated into training and test sets with 80% and 20% respectively. Furthermore, the scattering transformation was used to train the classifier using the training data while the model was validated using the test data.

A wavelet time scattering frame was created with a scale invariance of 0.22 seconds. Moreover, a feature vector was generated by averaging the scattering transformation across all time samples in order to have a sufficient number of scattering coefficients to average per time window. The FSDD was also separated into training and test sets with 80% and 20% respectively. Furthermore, the scattering transformation was used to train the classifier using the training data while the model was validated using the test data.

https://doi.org/10.25077/jnte.v14n3.1310.2025

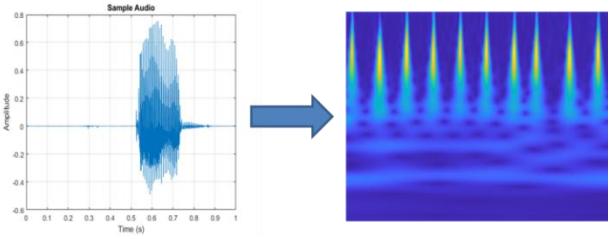


Figure 4. Sample data on spoken spectrograms

It is possible to extract different features from a waveform including power, pitch, and vocal tract configuration from a speech signal [26]. This research, therefore, used waveform for amplitude envelop and spectrogram for frequency-domain features such as spectral centroid. The selection of the features to include was observed to be the most important part. This is because feature engineering is more difficult with machine learning algorithms such as logistic regression but the process is simple with deep learning because only the spectrogram is required to feed the model.

Mel-Frequency Cepstral Coefficients (MFCC) and linear prediction coefficients (LPC) are two popular features but MFCC is the most commonly used feature due to its high accuracy [27]. It is important to note that MFCCs rely on a well-known variety of basic data transfer capacities with the frequency of the human ear. Therefore, the phonetically significant qualities of discourse were captured using filters which split at low frequencies and logarithmically at high frequencies and expressed using the Mel-frequency scale. Equations (4) and (5) were used to show the relationship between frequency in Hz and Mel is below:

$$m = 1125\ln(1 + \frac{f}{700}) \quad (4)$$

$$f = 700(\exp(\frac{m}{1125} - 1) \quad (5)$$

The Mel filter bank usually uses the Mel-recurrence scaling which is a perceptual scale that aids the reproduction of how the human ear works. This is based on the concept of higher goals at lower frequencies and lower goals at higher frequencies. Another popular feature for speaker recognition is LPCs and its application requires first understanding the autoregressive model of speech [28].

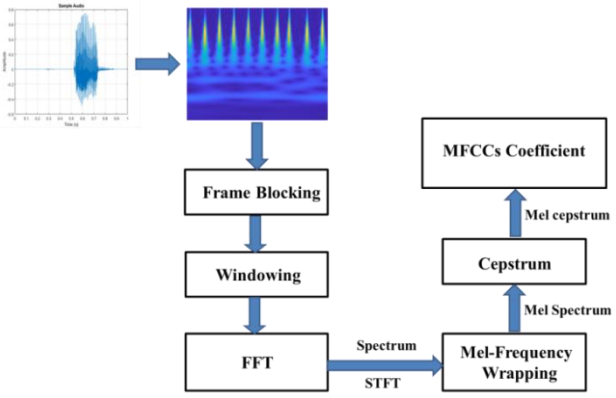


Figure 5. MFCC calculation diagram

Table 1. Functions used in the proposed DCNN model

| Symbol | Quantity |
|---|---|
| Function | Explanation |
| Convolution2D | Sequence input, sliding window convolution to 2-dimensional input information |
| MaxPooling2D | Maximum pooling layer, imposing a maximum pooling on the spatial domain signal |
| RELU | Rectified Linear Unit, which serves linear rectification activation on the input vector of the upper layer neural network and outputs nonlinear result |
| Flatten | The Flatten layer applied to translate the multidimensional input into one-dimensional information |
| Dropout | It is a regularization layer to prevent overfitting |

Meanwhile, the linear prediction method provides a dependable, reliable, and precise strategy to evaluate the parameters characterizing the linear time-varying system that represents the vocal tract [29].

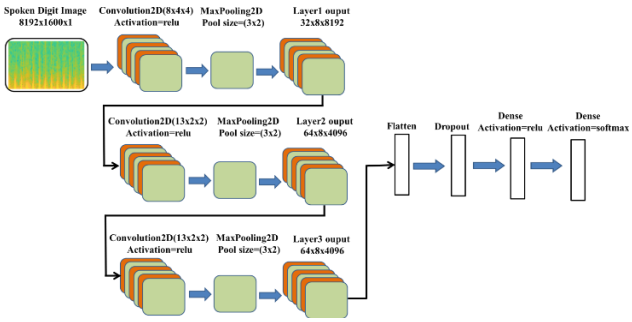


Figure 6. The architecture of the proposed DCNN model

## SPOKEN DIGIT DATA CLASSIFIER

CNN was used as the Spoken digit classifier. It was first introduced by Roy et al [30] as part of a project designed for the purpose of recognizing handwritten zip codes. Its application was observed to have made it possible to extract spatially adjacent pixels correlation through the adoption of nonlinear and multiple filters. This also has the capability to extract varieties of features associated with local images.

It is also important to note that it is preferable to use 2D convolutional and pooling layers to filter the spatial locality of spoken digit images [31]. Moreover, there was a conversion of time-domain spoken signals into 2D spectrograms during the time-frequency representations in order to encourage 2D-CNN in Spoken signal recognition. The 2D-CNN structure is presented in Figure 6 while the functions explanation is in Table 1.

There was a conversion of spoken digit data recording into an spoken digit spectrum image that has 8192 x 1600 pixels resolution. Meanwhile, the Convolution2D layer designed with 8 convolution kernels and a 4 x 4 kernel size was used in the first

hidden layer with RELU (Rectified Linear Unit) selected for the activation function. This was followed by the addition of a MaxPooling2D with a pool size of (3, 2) to produce the first layer's output shape of 32 x 8 x 8192. The second layer used a Convolution2D layer designed to have 13 convolution kernels and a 3x2 kernel size while the activation function was also through RELU. There was an addition of a MaxPooling2D with a pool size of (3, 2) to produce the 64 x 8 x 4096 output shape for the second layer. Moreover, the third layer also used the Convolution2D layer with 13 convolution kernels and a 3x2 kernel size with RELU applied as the activation function. This was also followed by the addition of a MaxPooling2D with a pool size of (3, 2) to produce the final output shape of the third layer to be 64 x 8 x 4096.

## CONVOLUTION NEURAL NETWORK

CNN is a popular DNN architecture typically trained using a gradient based optimization algorithm [32]. It consists of multiple back-to-back layers linked in a feedforward fashion. As previously stated, its main layers include the convolutional, normalization, pooling, and fully connected layers with the first three usually applied to feature extraction while the last is for classification. The general CNN architecture for the classification task is depicted in Figure 1 [33]. It is important to note that an activation function for a non-linear methodology is required to obtain the output in the convolution layers. The inputs for this layer are small parts of the original volumes as shown in Figure 6. Moreover, down-sampling was performed at each subsampling layer to feature maps and reduce network parameters. This, subsequently, reduced overfitting and speeds up the training process. Pooling was also conducted over p x p elements (filter size) to adjoin the expanse of all feature maps. The layers should fully attached in the final stage as in other neural networks. The latter layers take the previous low- and mid-level features then generate high-level abstraction from the input speech data. The final layer called SVM or Softmax used to generate a classification score in probabilistic terms to relate to a specific class.

The selection of kernel sizes, number of filters, and layer depth in the proposed CNN architecture was grounded in the characteristics of the time–frequency spectrograms. The initial convolutional layer employed a 4×4 kernel with 8 filters to capture coarse but discriminative local patterns related to transient energy changes and formant structures. A relatively larger kernel at this stage enables the model to detect broad spectral transitions commonly found in spoken digits. Subsequent layers adopted smaller kernels (3×2) with an increased number of filters (13 filters per layer) to progressively refine the representation by focusing on more localized frequency–time variations. This hierarchical reduction in kernel size preserves fine-grained spectral cues while the increased filter count enriches feature diversity.

The use of MaxPooling with an asymmetric pool size of (3,2) in each block was justified by the elongated shape of the spectrograms, where the temporal dimension is considerably larger than the frequency dimension. This pooling strategy reduces computational load while retaining essential temporal dynamics. Stacking three convolution–pooling blocks balances representational capacity and generalization, preventing overfitting despite the relatively modest dataset size. ReLU activation was selected for all convolutional layers due to its computational efficiency and ability to alleviate vanishing-gradient effects. Overall, the parameter configuration was designed to ensure effective extraction of both global and localized speech features, thereby improving classification accuracy and model readability.

## RESULTS AND DISCUSSION

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [14], [15]. The discussion can be made in several sub-sections.

### EVALUATION METRICS

An attempt was made to measure the categorization performance using two metrics which include accuracy and loss. The accuracy was indicated by the ratio between the total correctly identified samples and test samples as represented in the following mathematical formula:

$$\text{Accuracy}(\%)=\frac{TP+TN}{TP+TN+FP+FN}x100 \quad (6)$$

Where, TP denotes true positive, indicating the correct classification, TN stands for true-negative, which means that was correctly classified, FP stands for false-positive which means was incorrectly classified, while FN symbolizes false-negative which means incorrect classification as normal [34]. Moreover, the variation between the predicted and true values of the model for a specific sample was defined as the loss metric. It is important to note that the mathematical expressions of this metric are different. Therefore, the categorical cross-entropy loss in equation 7 was selected for this study.

$$loss = -\frac{1}{n}\sum_{i=1}^{n} \hat{y}_{i1} 1ny_{i1} + \hat{y}_{i2} 1ny_{i2} + \cdots + \hat{y}_{im} 1ny_{im} \quad (7)$$

Where, n = total samples, m = total categories, $\hat{y}$ = predicted output value, and y = actual value.

### MODEL PARAMETER OPTIMIZATION

The learning rate and batch size are the two main parameters in the proposed 2D-CNN model. These parameters, therefore, need to be optimized to achieve the best spoken digit recognition performance. Different contrast experiments were performed by varying the parameters with the aim of assessing the significance of the two main parameters in the proposed model. Different learning rates were also tested at a constant batch size as indicated in the complete parameter set in Table 2 while different batch sizes were also tested at a constant learning rate as indicated in Table 4.

The number of iteration steps was set at 1200. The accuracy curves in Figures 7 and 9 exhibit clear convergence patterns that highlight the sensitivity of the 2D-CNN model to learning rate and batch size. At moderate learning rates (particularly 0.001), the accuracy consistently increases and stabilizes near 1.0, indicating that the model is able to update its weights effectively without overshooting the optimal minima. While the mean

 https://doi.org/10.25077/jnte.v14n3.1310.2025

accuracies for the seven data sets are presented in Table 2 and Table 4.

Table 2. Average accuracies for 5 data sets with a batch size of 50

| MiniBatch Size | Learning Rate | Average Accuracy (%) |
|---|---|---|
| 50 | 0.00001 | 86.2 |
| 50 | 0.0001 | 98.6 |
| 50 | 0.001 | 99.2 |
| 50 | 0.01 | 98.8 |
| 50 | 0.1 | 93.2 |

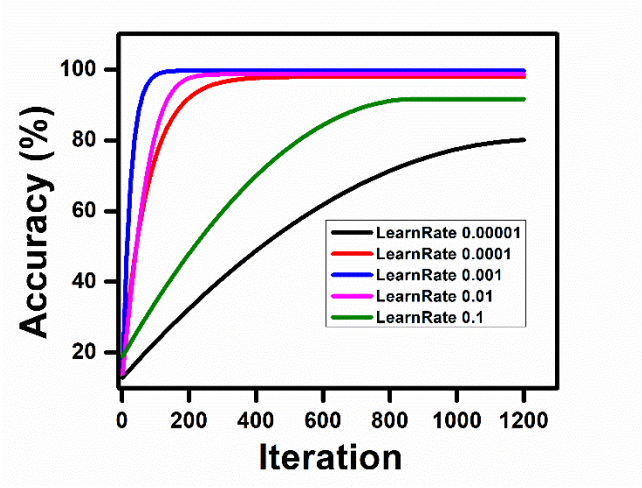


Figure 7. Accuracy value curves for 5 data sets with the Batch Size of 50.

In contrast, higher learning rates (0.01–0.1) produce more pronounced oscillations, suggesting unstable gradient updates, while very small learning rates (0.00001) slow the convergence process due to insufficient step sizes. A similar trend is observed when varying the batch size: medium batch sizes, especially 50, provide the most stable and rapid convergence, whereas excessively small or large batches introduce higher variance in gradient estimation, resulting in noticeable fluctuations during training.

Table 3. Average losses for 5 data sets with a batch size of 50

| MiniBatch Size | Learning Rate | Average Loss |
|---|---|---|
| 50 | 0.00001 | 1.3804 |
| 50 | 0.0001 | 0.3197 |
| 50 | 0.001 | 0.1850 |
| 50 | 0.01 | 0.2250 |
| 50 | 0.1 | 0.7972 |

The loss value curves for the five data sets with a batch size of 50 are shown in Figure 8 while the average losses are presented in Table 3. It was discovered that the same batch size parameters of 50 produced similar average losses for the five data while the loss curve fluctuates differently at different learning rates.

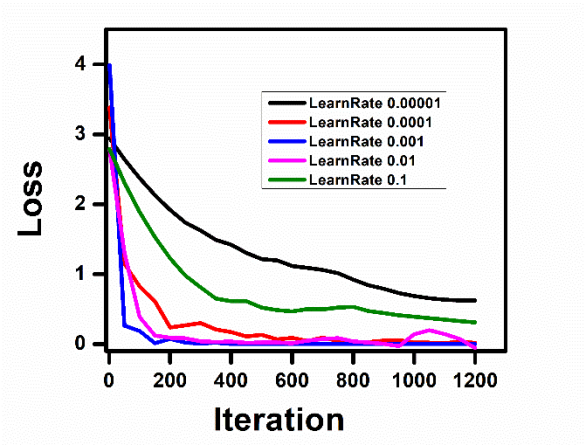


Figure 8. The loss value curves for 5 data sets with a batch size of 150.

The findings showed that the convergence trend of the accuracy curve is close to 0 due to an increase in the number of iteration steps at a learning rate of 0.01 and was also observed to be relatively stable state during the process. The same trend was also indicated at 0.01 but several relatively large fluctuations were observed which subsequently became larger as the learning rate increased from 0.01 to 0.1.

Table 4. Average accuracies for 7 data sets with the learning rate of 0.001

| MiniBatch Size | Learning Rate | Average Accuracy (%) |
|---|---|---|
| 200 | 0.001 | 98.8 |
| 150 | 0.001 | 98.6 |
| 125 | 0.001 | 98.6 |
| 100 | 0.001 | 98.8 |
| 75 | 0.001 | 98.6 |
| 50 | 0.001 | 99.2 |
| 25 | 0.001 | 99 |

Table 4 displays the parameter set for the contrast experiments conducted using different batch size at a constant learning rate.

Figure 9 shows the accuracy value curves for the five data sets having a single learning rate of 0.001 while the mean accuracies for the seven data sets are presented in Table 4. It was discovered that the average accuracies are comparable at the same learning rate but the accuracy curve indicated different fluctuations at different batch sizes. At a batch size of 50, the accuracy curve exhibited a convergence trend close to 1 as the number of iteration steps increased and maintains a relatively stable state during convergence. When the batch size was set at 200, the accuracy curve exhibited the same trend, however, several large fluctuations were observed in convergence process which further became larger as the batch size was gradually reduced from 200 to 75.

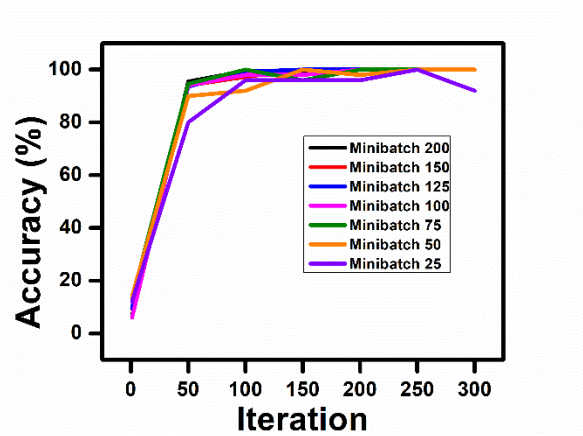


Figure 9. Accuracy curves for 7 data sets with the learning rate of 0.001.

Table 5. Average losses for 7 data sets with the learning rate of 0.001

| MiniBatch Size | Learning Rate | Average Loss |
|---|---|---|
| 200 | 0.001 | 0.4659 |
| 150 | 0.001 | 0.4171 |
| 125 | 0.001 | 0.3471 |
| 100 | 0.001 | 0.3393 |
| 75 | 0.001 | 0.2139 |
| 50 | 0.001 | 0.1838 |
| 25 | 0.001 | 0.1248 |

The loss curves (Figure 8 and Figure 10) further reinforce these observations. When the learning rate is set to 0.001, the loss consistently decreases toward zero with minimal volatility, demonstrating efficient minimization of the categorical cross-entropy objective. Higher learning rates produce irregular descent behavior, characterized by sudden spikes and slower stabilization, indicating difficulty in finding a smooth optimization trajectory. While the average losses are presented in Table 3 and table 5.

Regarding batch size, the lowest and most stable loss values are obtained at a batch size of 50, while both smaller and larger batches exhibit increased fluctuations. Overall, the combined trend of accuracy rising smoothly while loss declines steadily confirms that the model achieves optimal learning dynamics when trained with a learning rate of 0.001 and a batch size of 50.

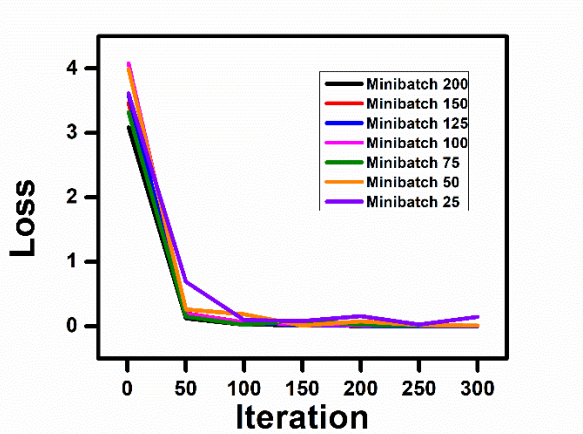


Figure 10. The loss value curves for 7 data sets with the learning rate of 0.001.

## COMPARISON WITH OTHER EXISTING APPROACHES

The proposed 2D-CNN model performance was compared to earlier Spoken digit recognition works such as SVM (Support Vector Machine), Backpropagation Neural Network, and MFCC and Multiple Recurrent Neural Networks.

Table 6. Comparison with other existing approaches

| Model | WORK | Test set | Average Accuracy (%) |
|---|---|---|---|
| MFCC and Multiple RNN | Utomo [35] | 1020 | 87.74 |
| SVM | Jain et al. [36] | 1100 | 94.9 |
| LSTM RNN | Zia et al. [37] | 1000 | 98 |
| **2D-CNN** | **Proposed** | **3000** | **99.2** |

It is unfair to compare the accuracy directly because these works have a different test sets number. However, the proposed CNN model was observed to have outperformed other previous works, thereby, leading to the introduction of a novel approach to Spoken digit recognition using WTS-MFCC and convolutional neural networks.

Moreover, the performance of the feature extraction pattern classification was also compared with previous studies as indicated in Table 6. It was discovered that the two feature-extraction-pattern classification approaches are comparable to the proposed method. Sharan extracted features from Spoken Digit Signals using Wavelet Scalogram (WS) [38], He Ba used the STFT and MFCC [39], Wazir et al. used MFCC [40]. It was, therefore, concluded that the proposed method is fast and accurate for classification.

Table 7. Comparison with feature extraction pattern classification

| Method | WORK | Average Accuracy (%) |
|---|---|---|
| WS+CNN | Sharan [38] | 97.15 |
| STFT+MFCC+CNN | He Ba [39] | 90 |
| MFCC+LSTM | Wazir et al. [40] | 94 |
| SVM+MFCC | Jain et al. [36] | 94.9 |
| MFCC+ Multiple RNN | Utomo et al. [35] | 98 |
| **WTS+MFCC+ 2D-CNN** | **Proposed** | **99.2** |

The confusion graph is a summary of network performance trained on test sets with the columns and rows showing the precision and recall for each class. It is important to note that the precision values are indicated at the bottom while the recall value is at the right of the confusion table.


https://doi.org/10.25077/jnte.v14n3.1310.2025

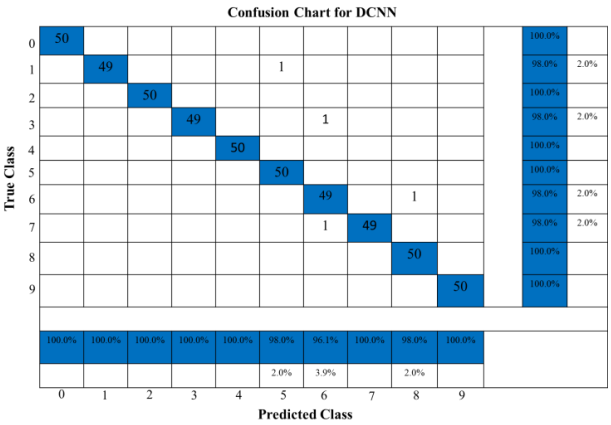


Figure 11. The normalized confusion matrix of the best result achieved.

## CONCLUSIONS

This research presents a robust and efficient spoken digit recognition framework that integrates Wavelet Time Scattering, MFCC feature extraction, and a customized 2D-CNN classifier. Experimental findings demonstrate that the proposed pipeline significantly improves recognition performance compared to conventional signal-processing and machine-learning methods. Optimal learning parameters (learning rate = 0.001, batch size = 50) produce a stable convergence pattern, yielding the highest accuracy and lowest loss during training.

The model achieved 99.2% accuracy on the FSDD dataset, surpassing established approaches such as SVM, MFCC-LSTM, and multiple RNN architectures. The results confirm that WTS provides superior low-frequency localization and, when combined with CNN-based feature learning, leads to more discriminative time–frequency representations. This enhances the system's ability to handle realistic scenarios involving varying accents, non-digit audio, and background noise.

Overall, the proposed WTS+MFCC+2D-CNN framework offers a fast, accurate, and scalable solution for spoken digit recognition. Future work may extend this approach to continuous speech, multilingual datasets, and real-time embedded implementations to support broader applications in security systems, voice-activated services, and intelligent user interfaces.

## ACKNOWLEDGMENT

This research is supported by the Institute for Research and Community Service, Sriwijaya University, through the Science and Technology Research Grant Fund with contract number: 0012/UN9/SK.LP2M.PT/2024.

## REFERENCES

[1] S. Nasr, M. Quwaider, and R. Qureshi, "Text-independent Speaker Recognition using Deep Neural Networks," in *2021 International Conference on Information Technology (ICIT)*, 2021, pp. 517-52, doi: 10.1109/ICIT52682.2021.9491705.

[2] A. Boles and P. Rad, "Voice biometrics: Deep learning-based voiceprint authentication system," in *2017 12th System of Systems Engineering Conference (SoSE)*, 2017, pp. 1-6, doi: 10.1109/SYSOSE.2017.7994971.

[3] A. Irum and A. Salman, "Speaker verification using deep neural networks: A," vol. 9, no. 1, 2019, doi: 10.18178/ijmlc.2019.9.1.760

[4] R. Qureshi, M. Nawaz, F. Y. Khuhawar, N. Tunio, M. J. I. J. o. A. C. S. Uzair, and Applications, "Analysis of ECG signal processing and filtering algorithms," vol. 10, no. 3, 2019, doi: 10.14569/ijacsa.2019.0100370.

[5] R. Qureshi, S. A. R. Rizvi, S. H. A. Musavi, S. Khan, and K. Khurshid, "Performance analysis of adaptive algorithms for removal of low frequency noise from ECG signal," in *2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, 2017, pp. 1-5, doi: 10.1109/ICIEECT.2017.7916551.

[6] D. Stoyanov *et al.*, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*. Springer, 2018.

[7] R. Qureshi, M. Uzair, K. Khurshid, and H. J. P. R. Yan, "Hyperspectral document image processing: Applications, challenges and future prospects," vol. 90, pp. 12-22, 2019, doi: 10.1016/j.patcog.2019.01.026.

[8] M. Sajjad and S. J. I. A. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," vol. 8, pp. 79861-79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[9] M. Zhang, M. Diao, and L. J. I. A. Guo, "Convolutional neural networks for automatic cognitive radio waveform recognition," vol. 5, pp. 11074-11082, 2017, doi: 10.1109/ACCESS.2017.2716191.

[10] O. Krestinskaya, I. Dolzhikova, and A. P. James, "Hierarchical temporal memory using memristor networks: A survey," vol. 2, no. 5, pp. 380-395, 2018, doi: 10.1109/TETCI.2018.2838124.

[11] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. J. a. p. a. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," 2018.

[12] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. J. a. p. a. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," 2017, doi: 10.48550/arXiv.1701.02720.

[13] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela, and P. J. Manamela, "Automatic speaker recognition system based on machine learning algorithms," in *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, 2019, pp. 141-146, doi: 10.1109/RoboMech.2019.8704837.

[14] F. R. rahman Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention-based models for text-dependent speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5359-5363, doi: 10.1109/ICASSP.2018.8461587,

[15] S. Dey, P. Motlicek, S. Madikeri, and M. J. S. c. Ferras, "Template-matching for text-dependent speaker verification," vol. 88, pp. 96-105, 2017, doi: 10.1016/j.specom.2017.01.009.

[16] W. Feng, N. Guan, Y. Li, X. Zhang, and Z. Luo, "Audio visual speech recognition with multimodal recurrent neural networks," in *2017 International Joint Conference on neural networks (IJCNN)*, 2017, pp. 681-688, doi: 10.1109/IJCNN.2017.7965918.

[17] Y. Yu, X. Si, C. Hu, and J. J. N. c. Zhang, "A review of recurrent neural networks: LSTM cells and network

architectures," vol. 31, no. 7, pp. 1235-1270, 2019, doi: 10.1162/neco_a_01199.
[18] W. Zhang, M. Zhai, Z. Huang, C. Liu, W. Li, and Y. Cao, "Towards end-to-end speech recognition with deep multipath convolutional neural networks," in *International Conference on Intelligent Robotics and Applications*, 2019, pp. 332-341.
[19] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 2017, pp. 333-336, doi: 10.1109/CESYS.2017.8321292.
[20] Z. J. R. F. Jackson, "Free spoken digit dataset (fsdd)," vol. 1, p. 2020, 2016.
[21] S. Otte, P. Rubisch, and M. V. Butz, "Gradient-based learning of compositional dynamics with modular RNNs," in *International Conference on Artificial Neural Networks*, 2019, pp. 484-496.
[22] F. M. Bayer, A. J. Kozakevicius, and R. J. J. S. P. Cintra, "An iterative wavelet threshold for signal denoising," vol. 162, pp. 10-20, 2019, doi: 10.1016/j.sigpro.2019.04.005.
[23] W. Liu and W. J. I. A. Chen, "Recent advancements in empirical wavelet transform and its applications," vol. 7, pp. 103770-103780, 2019, doi: 10.1109/ACCESS.2019.2930529.
[24] R. V. Sharan and T. J. Moir, "Time-frequency image resizing using interpolation for acoustic event recognition with convolutional neural networks," in *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, 2019, pp. 8-11, doi: 10.1109/ICSIGSYS.2019.8811088.
[25] K.-L. Chung, T.-C. Leung, T.-Y. Liu, and Y.-C. J. I. A. Tseng, "A Cubic Convolution Interpolation-Based Chroma Subsampling Method for Bayer and RGBW CFA Raw Images," vol. 10, pp. 22687-22699, 2022, doi: 10.1109/ACCESS.2022.3154487.
[26] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. J. I. a. Shaalan, "Speech recognition using deep neural networks: A systematic review," vol. 7, pp. 19143-19165, 2019, doi: 10.1109/ACCESS.2019.2896880.
[27] Q. Li *et al.*, "MSP-MFCC: Energy-efficient MFCC feature extraction method with mixed-signal processing architecture for wearable speech recognition applications," vol. 8, pp. 48720-48730, 2020, doi: 10.1109/ACCESS.2020.2979799.
[28] N. Naka and V. Ruoppila, "Linear prediction coefficient conversion device and linear prediction coefficient conversion method," ed: Google Patents, 2018, doi: 10.1016/j.patrec.2017.03.004.
[29] E. Tatulli and T. Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2971-2975, doi: 10.1109/ICASSP.2017.7952701.
[30] S. Roy, N. Das, M. Kundu, and M. J. P. R. L. Nasipuri, "Handwritten isolated Bangla compound character recognition: A new benchmark using a novel deep learning approach," vol. 90, pp. 15-21, 2017, doi: 10.1016/j.patrec.2017.03.004.
[31] T. J. Jun, H. M. Nguyen, D. Kang, D. Kim, D. Kim, and Y.-H. J. a. p. a. Kim, "ECG arrhythmia classification using a 2-D convolutional neural network," 2018, doi: 10.48550/arXiv.1804.06812.
[32] O. F. Reyes-Galaviz, W. Pedrycz, Z. He, N. J. J. D. Pizzi, and K. Engineering, "A supervised gradient-based learning algorithm for optimized entity resolution," vol. 112, pp. 106-129, 2017, doi: 10.1016/j.datak.2017.10.004.
[33] Y. Wang, X. Tao, X. Qi, X. Shen, and J. J. A. i. n. i. p. s. Jia, "Image inpainting via generative multi-column convolutional neural networks," vol. 31, 2018.
[34] W. Yin *et al.*, "Self-adjustable domain adaptation in personalized ECG monitoring integrated with IR-UWB radar," vol. 47, pp. 75-87, 2019, doi: 10.1016/j.bspc.2018.08.002.
[35] Y. F. Utomo, E. C. Djamal, F. Nugraha, and F. Renaldi, "Spoken word and speaker recognition using MFCC and multiple recurrent neural networks," in *2020 7th international conference on electrical engineering, computer sciences and informatics (EECSI)*, 2020, pp. 192-197, doi: 10.23919/EECSI50503.2020.9251870.
[36] M. Jain, S. Narayan, P. Balaji, A. Bhowmick, and R. K. J. a. p. a. Muthu, "Speech emotion recognition using support vector machine," 2020, doi: 10.48550/arXiv.2002.07590.
[37] T. Zia and U. J. I. J. o. S. T. Zahid, "Long short-term memory recurrent neural network architectures for Urdu acoustic modeling," vol. 22, no. 1, pp. 21-30, 2019, doi: 10.48550/arXiv.1402.1128.
[38] R. V. Sharan, "Spoken digit recognition using wavelet scalogram and convolutional neural networks," in *2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 2020, pp. 101-105: IEEE, doi: 10.1109/RAICS51191.2020.9332505.
[39] H. Ba, "Spoken Digit Classification: A Method Using Convolutional Neural Network and Mixed Feature.", doi: 10.18178/wcse.2021.02.002.
[40] A. S. M. B. Wazir and J. H. Chuah, "Spoken arabic digits recognition using deep learning," in *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, 2019, pp. 339-344: IEEE, doi: 10.1109/I2CACIS.2019.8825004.

## NOMENCLATURE

Meaning of symbols used in the equations and other symbols presented in your article must be presented in this section.

| | |
|---|---|
| $\Psi$ | mother wavelet |
| $\hat{y}$ | predicted output value |

## AUTHORS BIOGRAPHY

**Irmawan** (Member, IEEE)
Irmawan was born in Lahat, South Sumatra, on September 17, 1974. He has been a lecturer at the Faculty of Engineering, Department of Electrical Engineering, Sriwijaya University, since December 2000. After completing his bachelor's degree in Electronics Instrumentation from Gadjah Mada University, Yogyakarta, Indonesia, in 1998, he then obtained a master's degree in engineering (M. Eng.) in Electronic Signals from Gadjah Mada University, Yogyakarta, Indonesia, in 2003. He is also a member of IEEE. His research interests include signal and image processing, robotics and machine learning. He can be contacted at email: irmawan@unsri.ac.id.

 https://doi.org/10.25077/jnte.v14n3.1310.2025

**Suci Dwijayanti** (Member, IEEE) received the M.S. degree in electrical and computer engineering from Oklahoma State University, Stillwater, OK, USA, in 2013, and the Ph.D. degree from the Graduate School of Natural Science and Technology, Kanazawa University, Japan, in 2018. From 2007 to 2008, she was an Engineer with ConocoPhillips Indonesia Inc., Ltd. Since 2008, she has been with the Department of Electrical Engineering, Universitas Sriwijaya, Indonesia. Her research interests include signal processing and machine learning. She received a Fulbright Scholarship for her master's degree. She can be contacted at email: sucidwijayanti@ft.unsri.ac.id.

**Bhakti Yudho Suprapto** (Member, IEEE)
Bhakti Yudho Suprapto was born in Palembang, South Sumatra, Indonesia, on February 11, 1975. He is currently pursuing the Graduate degree in electrical engineering with Sriwijaya University, Indonesia. His master's and doctoral programs in electrical engineering at Universitas Indonesia (UI). He is also an Academic Staff Member of Electrical Engineering at Universitas Sriwijaya. His research interests include control and intelligent systems. He can be contacted at email: bhakti@ft.unsri.ac.id.