西南交通大学学报
JOURNAL OF SOUTHWEST JIAOTONG UNIVERSITY
ISSN 02582724
Register Login

Optimizing Speech Emotion Recognition Using Convolutional Neural Network Technique

Silviana Widya Lestari[1*], Rabab Alayham Abbas[2] ,Trismayanti Dwi P[3], Aisha Rahmayanti[4]

[1,2,4]School of Graduate Studies, Management and Science University, Malaysia
[3]Department of Information Technology, Politeknik Negeri Jember, Indonesia

DOI:https://doie.org/10.0226/Jsju.2025422437

Abstract: This study investigates the successful recognition and interpretation of emotional states through the use of technology, specifically speech-based emotional analysis. Since people with mental health disorders frequently have trouble regulating their emotions, accurate emotion recognition is essential to delivering individualized care. Counsellors and mental health specialists can provide individualized interventions and gain a deeper understanding of people's emotional states by utilizing technology. Convolutional Neural Networks (CNNs) for Speech Emotion Recognition (SER) are the main topic of this study, which focuses on deep learning methods. Notwithstanding notable progress, issues like lowering computational complexity and attaining high accuracy still exist in this field. The paper assesses SER utilizing well-known datasets, such as RAVDESS, CREMA, SAVEE, and TESS, in order to overcome these issues. Root Mean Square Error (RMSE) is the most successful feature extraction technique, according to the methodology, which looks at the performance of five distinct feature extraction strategies. The CNN-based model attains remarkable accuracy scores of 93.11% and 96.07% using RMSE. Conv1D layers are used in the model to apply the best feature extraction technique, allowing for real-time emotion recognition. In order to improve SER accuracy, this study emphasizes the significance of making thoughtful decisions about feature extraction techniques and dataset selection. The study significantly advances SER technology by using CNN models and carefully optimizing feature extraction. In order to progress the area and fully solve its enduring issues, it also highlights the necessity of enhanced modelling techniques and further investigation of various datasets.

Keywords: Deep Learning; Feature Extraction; Datasets; Convolutional Neural Network; Speech Emotion Recognition

I. Introduction

According to Blanke, E. S. et al., [1], emotions have a big impact on our well-being and how we connect with other people, thus it's important to understand how someone is feeling in our day-to-day lives. Mental health disorders are complex problems that can take many different forms. In 2021, the World Health Organization (WHO) defined mental illnesses as clinically substantial disruptions in a person's behaviour or emotional regulation that are typically linked to distress or impairment in key areas of life. With using technology, such as emotional analysis in speech and facial expressions, it is possible to detect emotional indicators that humans might find challenging to detect. Counsellors can more precisely and successfully identify their patients' emotions with the use of the human emotion detection system. The counsellors can create a more.

individualized and successful treatment plan that meets the patient's unique needs using the data the system provides. According to recent studies, including the one by Jain et al. [2], one of the most intense areas of research being done on human emotions these days is voice recognition.

Speech is an effective means of communicating emotions and ideas. Determining the emotional content of a speech signal and detecting the emotions from the speech utterance are important challenges for the researcher, according to Swain et al., [3]. Among the various emotional categories are depression, anxiety, boredom, annoyance, fear, joy, neutral, panic, sadness, tension, surprise, shock, and concern. As stated by Varol and Ibrahim [4]. A speaker can be identified using the information provided by speech waves. The literature on speech emotion recognition (SER) has employed a variety of techniques to extract emotions from signals, including some well-known speech analysis and classification methods. Research by Khalil et al. [5] claims that deep learning methods have lately been put out as a substitute for conventional methods in Speech Emotion Recognition.

1.1 Research Background

Understanding and managing human emotions is crucial for social relationships and mental health. The detrimental effects that poorly controlled emotions can have on mental health emphasize the need for effective tools for emotion detection and recognition. Madanian et al. [6] claim that technology, particularly emotional analysis using facial and speech recognition, has become a viable method for more accurately identifying and understanding emotions. The growing intensity of research on human emotions, particularly in the context of recognizing emotions through speech, has led to the emergence of voice recognition as an emotion identification technique that capitalizes on the unique characteristics of human speech.

J. Zhao and colleagues [7] define SER as "the process of automatically identifying the emotional state of a speaker from their spoken words, using various computational methods and techniques." In the framework of current research, this is one definition of SER. They note that SER is applicable to virtual agents, speech-based emotion detection systems, and conversational interfaces. It employs a range of deep learning and signal processing techniques, including feature extraction, classification, and regression.

This study aims to enhance deep learning techniques by identifying voice emotions using Convolutional Neural Networks (CNN). One of the primary objectives is to compare several deep learning techniques for Speech Emotion Recognition (SER) in order to assess their efficacy. Additionally, the project intends to enhance speech emotion recognition by utilizing CNN, focusing on raising the accuracy of emotion detection. The final objective is to validate the effectiveness of the optimization process in speech emotion detection in order to verify that the proposed CNN-based approach contributes to better and more reliable results when it comes to identifying emotional states from speech signals.

1.2 Literature Review

Through literature reviews and suggested techniques, a number of earlier studies have investigated different ways for speech emotion recognition (SER). For example, a study by Hashem et al. [8] conducted a systematic literature analysis on speech emotion identification techniques, covering the most recent discoveries and techniques from 2012 to 2022.

Two noteworthy journal articles significantly advance the field of voice emotion identification systems. First, Zhang et al.'s research [9] examines the challenges in the field and presents a deep learning model that integrates attention mechanisms with recurrent neural networks. Our model demonstrates enhanced accuracy and robustness in comparison to traditional methods, underscoring the potential of deep learning to enhance emotion identification technologies. Second, a paper by Anvarjon et al. [10] proposes a novel Convolutional Neural Network (CNN) technique that captures deep frequency characteristics. Performance examination on the IEMOCAP and EMO-DB speech datasets yields recognition values of 77.01% and
92.02%, respectively.

A study by Kwon [11] claims that those investigations highlight the many challenges in enhancing the accuracy and reducing the computational complexity of Speech Emotion Recognition (SER) models. The proposed model uses stride CNN architectures, which use certain stride values, to down-sample feature maps instead of relying on pooling layers. The proposed method achieves noteworthy accuracy scores of 81.75% on IEMOCAP and 79.5% on RAVDESS. In their research, Amjad et al. [12] employed convolutional neural networks (CNNs) to categorize seven types of emotions using a number of publicly accessible databases. Their approach yields speaker-dependent (SD) identification accuracy rates of 92.02%, 88.77%, 93.61%, and 77.23% for Emo-DB, SAVEE, RAVDESS, and IEMOCAP while employing the feature selection method.

Convolutional neural networks (CNNs) have been used in numerous research to examine the accuracy of voice emotion recognition; the results have shown variation based on the CNN architecture and the dataset used. The Emo-DB database had an accuracy of 82.32%, the CASIA and FAU databases had an accuracy of

48.5%, and the SAVEE database had an accuracy of 53.60%, according to Mountzouris et al. [13]. In a similar vein, research by Ullah et al. [14] showed that CNN performed between 72.19% to 72.88% on the IEMOCAP database and 66.1% better than a feature-based SVM. Additionally, to improve recognition accuracy, Y. Zhao and Shu [15] suggested a unique method that combines CNN with error-correcting output codes (ECOC) based on gamma classifiers. Their approach demonstrated its efficacy in detecting emotional indicators in speech with average accuracies of 93.33% and 85.73% on the Berlin and ShEMO datasets, respectively. These results highlight the role that dataset selection and model architecture play in improving speech emotion identification accuracy.

Numerous approaches have been put out to address the complexity of speech emotion recognition (SER), but improving its accuracy is still a difficult issue. The accuracy of Convolutional Neural Networks (CNNs), which have been frequently used in SER, varies based on the particular architecture and dataset used. The impact of these characteristics is reflected in the reported accuracies for CNN-based SER models, which vary from 48.5% to 93.33% across various databases. But maintaining low computing complexity while increasing accuracy remains a major problem. The subjective and complex character of emotions makes it difficult for robots to reliably identify emotional content in human speech, according to Xu et al. [16].

Detecting emotions in spoken language is the focus of the quickly developing discipline of speech emotion recognition (SER). Scholars have investigated a number of approaches, pointing out possible directions for development, including data augmentation, hybrid models, cross-corpus recognition, and benchmarking against other systems. The importance of feature extraction in recognizing the emotional components of speech was highlighted by Alluhaidan et al. [17]. There is still much need for improvement in terms of accuracy, computational efficiency, and consistency across various architectures and datasets, even if existing SER approaches have produced encouraging results.

Numerous Deep Learning feature extraction strategies have been investigated in recent research, and a range of machine learning and deep learning models are being built to meet signal processing difficulties. Not all extracted audio features are trustworthy, as some may be inaccurate or contain redundancy, as noted by Lieskovská et al. [18]. Speech Emotion Recognition (SER) is essential to Human-Computer Interaction (HCI), according to de Lope & Graña [19]. SER is a focus of development in the present decade as researchers are actively working on novel ways to improve its efficiency and dependability for real-time applications.

II. Methodology

This study aims to investigate the CNN (Convolutional Neural Network) model of feature theory. Journal datasets including RAVDESS, CREMA, SAVEE, and TESS will be analyzed using five feature extraction techniques: RMSE, Energy, Entropy of Energy, Entropy, and Zero Crossing Rate. The best of these five comparative feature extraction methods will be selected and implemented in a real-time system using an author- generated dataset.

2.1 Dataset

Convolutional Neural Networks (CNNs) are used in this study phase to categorize audio data according to emotions. The study combines a freshly created dataset created especially for the goal of voice emotion detection with well-known datasets such as RAVDESS, CREMA, SAVEE, and TESS. Audio files in the ".wav" format are included in the database. A number of simulated emotion datasets, including EMO-DB (German), IEMOCAP, DES (Danish), RAVDESS, TESS, SAVEE, and CREMA-D, were cited by Alzubaidi et al. [20]. Together with the specially created dataset for this work, the chosen datasets—RAVDESS, CREMA, SAVEE, and TESS—form the basis for training and assessing the CNN models.

2.1.1 RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)

Anvarjon et al. [21] created the RAVDESS dataset, which is frequently used to identify emotions in English- language speech and music. Eight different emotions— fear, wrath, surprise, disgust, tranquility, happiness, sadness, and neutrality—are captured on tape by 24 actors (12 men and 12 women). The dataset consists of 1440 WAV files with a sample rate of 48,000 Hz.

2.1.2 CREMA (Crowd Sourced Emotional Multimodal Actors Dataset)

A study by Zielonka et al. [22] introduced the CREMA dataset, which was developed by 91 performers who generated 7443 audio samples in total. Crowdsourcing was used to validate the dataset's dependability, with 2443 raters evaluating the category emotion categories. A variety of emotions, including neutrality, happiness, anger, disgust, fear, and sadness, are included in CREMA-D.

2.1.3 SAVEE (Surrey Audio Visual Expressed Emotion)

SAVEE, according to Mittal et al. [23], is a spoken emotion database made up of recordings from four male performers that correspond to seven different emotional categories. Anger, contempt, fear, happiness, neutrality, sadness, and surprise are the seven emotion classes into which the dataset's vocal emotions are divided.

2.1.4 TESS (Toronto Emotional Speech Set)

200 phrases uttered by two women, ages 26 and 64, respectively, are recorded in the "TESS" dataset, which was first presented by Chatterjee et al. [24]. Seven repetitions of each sentence were made, each expressing a distinct emotional tone.

2.1.5 Own Dataset

In addition to the previously given dataset, the author collected sound in three seconds based on seven distinct types of emotions: anger, fear, disgust, neutral, sadness, joy, and surprise. The dataset contains 700 sound recordings for these 7 emotions, with 100 sound recordings for each emotion. WAV files are used for the audio files.

2.2 Optimizing the Convolutional Neural Network (CNN) for Speech Emotion Recognition (SER)

Artificial neural networks with several hidden layers are used in deep learning to automatically learn features and generate high-level representations from unprocessed input. According to Nanni et al. [25], deep learning models are able to examine big datasets, such spoken audio samples, and find minute features and complex patterns that are frequently difficult for conventional machine learning techniques to pick up on. Nie et al. [26] assessed the performance of the suggested SER model.

An 80-20 split technique, in which 80% of the data was used for CNN model training and the remaining 20% for testing, was used to assess its performance in comparison to current SER baseline methods. Audio data can be transformed into spectrogram images, which show frequency spectra over time, by using Convolutional Neural Networks (CNNs) to extract features. According to Chu et al. [27], these spectrograms are subsequently fed into a CNN for categorization. The CNN for Speech Emotion Recognition optimization procedure used in this study is shown in Figure 1.

.

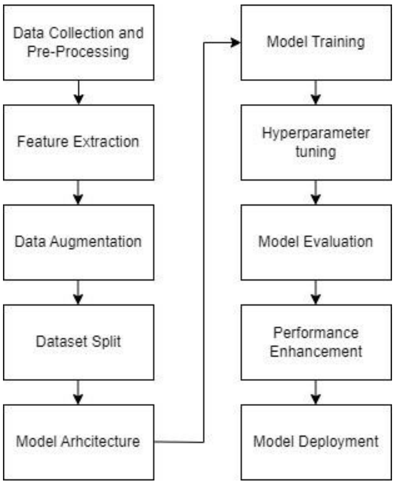

Figure. 1. Figure Flow of Optimizing Convolutional Neural Network (CNN) for Speech Emotion Recognition

2.2.1 Data Collection and Pre-processing

To create a labeled dataset, collect voice samples with matching emotion labels. The RAVDESS, CREMA, SAVEE, and TESS datasets as well as its own dataset— which includes the previously mentioned emotions of anger, contempt, fear, happiness, neutrality, sorrow, and surprise—will be used in this study. Following that, pre- processing methods like pitch correction, time-stretching, shifting, and noise addition will be applied to the RAVDESS, CREMA, SAVEE, and TESS audio datasets.

Wijayasingha & Stankovic [28] point out that by mimicking real-world situations where background noise impacts the audio, adding noise to audio data improves the model. As a result, the algorithm is better able to identify emotions even in noisy settings. In a similar vein, the stretching strategy in CNN optimization for Speech Emotion Recognition seeks to broaden and diversify the dataset, as explained by Ottoni et al. [29]. This technique improves the model's capacity to recognize speech changes that express various emotions by altering the temporal characteristics of audio, allowing the model to learn stable emotional representations.

Using shifting approaches to optimize CNN for Speech Emotion Recognition enhances the model's ability to identify emotions by collecting speech fluctuations that convey various emotional states, according to Parthasarathy & Tashev [30]. By altering the audio's temporal characteristics, this technique enables the model to pick up consistent representations of emotional information.

According to Pan & Wu [31], pitch shifting—which entails changing the audio's pitch—is useful for recording emotional pitch changes. This method uses pitch alterations to improve the model's recognition of emotional content. By recognizing speech variants that transmit a range of emotional states, shifting strategies in CNN optimization for Speech Emotion Recognition further improve the model's ability to detect emotions, as highlighted by Parthasarathy & Tashev [30]. This method allows the model to learn stable representations of emotional information by changing the audio's temporal properties.

2.2.2 Extracting Features

2.2.2.1 Root Mean Square Error (RMSE)

Utilize the audio data that has already been processed to extract pertinent elements. One often used statistic for speech emotion recognition is Root Mean Square Error (RMSE). According to Baird et al. [32], RMSE is a statistical metric that assesses how well a model predicts outcomes. It is also frequently used in feature extraction, especially in signal processing and audio analysis. Here, the root-mean-square (RMS) energy of each audio frame is calculated using RMSE, which offers information on the energy fluctuations of the signal over time.

Because of this, RMSE is a useful characteristic for a variety of audio processing tasks, such as speech recognition. As seen in the graphic below, this study's studies with the RMSE extraction feature on the RAVDESS, CREMA, SAVEE, and TESS datasets yielded a confusion matrix with an accuracy of 93.11%.

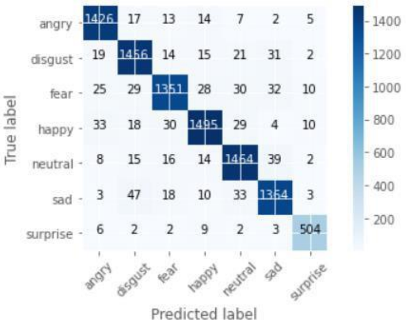

Figure. 6. Figure confusion matrix for combinations datasets processed with RMSE feature extraction.

2.2.2.2 Energy

"Energy" is a characteristic extracted from speech signals that helps identify different emotions in deep learning for speech emotion recognition. A deep learning-based system for identifying voice emotions must incorporate the energy characteristic and its associated parameters, claim Atmaja et al. [33]. Furthermore, the energy component has been effectively used with statistical techniques and machine learning models to find patterns associated with emotional expressions in speech, as shown by Deka & Mittal [34]. The confusion matrix and the 28.68% accuracy attained in this study's experiments employing the Energy extraction feature on the RAVDESS, CREMA, SAVEE, and TESS datasets are shown in the image below.

.

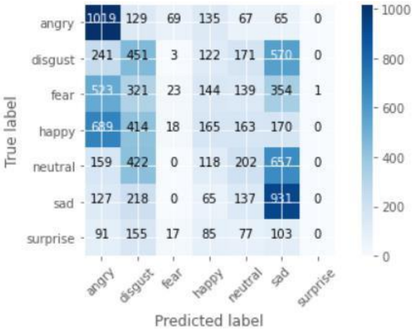

Figure. 7. Figure confusion matrix for combinations datasets processed with Energy feature extraction

2.2.2.3 Entropy of Energy

Deep learning for voice emotion recognition can be aided by the extraction of the "energy entropy" characteristic from speech signals. Despite its potential, this characteristic is rarely utilized in current research. Relevant features are usually extracted from voice signals during the feature extraction step to help categorize various emotions. Aouani & Ayed [35] state that a deep learning algorithm for voice emotion recognition is then fed these extracted features. Experiments employing the Energy Entropy extraction feature on the RAVDESS, CREMA, SAVEE, and TESS datasets yielded aconfusion matrix with an accuracy of 22.89%, as seen in the graphic below.

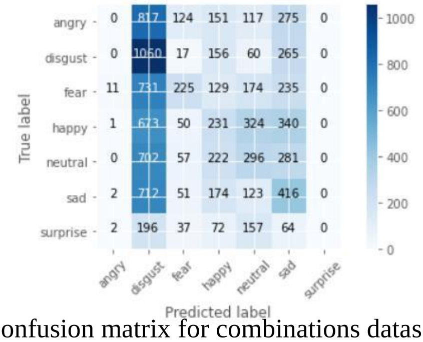

Figure. 8. Figure confusion matrix for combinations datasets processed with Entropy of Energy feature extraction.

2.2.2.4 Entropy

In feature extraction for deep learning-based voice emotion identification, entropy is essential. Sample entropy and approximation entropy are two examples of entropy-related metrics that help identify emotional states by capturing the general characteristics of speech signals. Y. Zhao & Shu [36] claim that entropy characteristics improve speech feature extraction and emotion representation in audio signals when used with deep learning techniques like convolutional neural networks (CNNs). This study's studies on the RAVDESS, CREMA, SAVEE, and TESS datasets utilizing the entropy extraction feature yielded a confusion matrix with an accuracy of 23.83%, as seen in the image below.

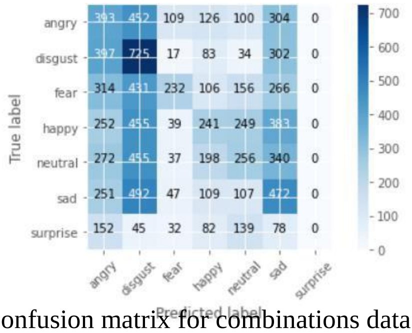

Figure. 9. Figure confusion matrix for combinations datasets processed with Entropy feature extraction.

2.2.2.5 Zero Crossing Rate (ZCR)

A popular feature in deep learning-based speech emotion recognition (SER) is zero-crossing rate (ZCR). It calculates the frequency with which a speech signal's amplitude crosses the zero-value threshold during a certain period of time. Because it aids in the characterization of speech signals and the identification of emotional states in spoken language, Mashhadi & Osei-Bonsu [37] have emphasized its critical importance in feature extraction for SER. Using the Zero Crossing Rate feature on the RAVDESS, CREMA, SAVEE, and TESS datasets, this study's experiments yielded an

accuracy of 61.96%; the related confusion matrix is displayed in the graphic below.

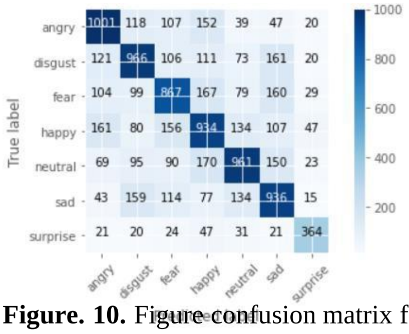

Figure. 10. Figure confusion matrix for combinations datasets processed with Entropy feature extraction.

2.3 Convolutional Neural Network Model

With an accuracy of 93.11% with RMSE, the suggested SER model uses a CNN architecture for feature extraction. After the Conv1D levels, the model has MaxPool1D and Batch Normalization layers. ReLU activation, 512 filters, stride of 1, "same" padding, and kernel size of 5 are all features of the initial Conv1D layer. Batch Normalization and MaxPool1D layers with a pool size of five and stride two come next. With varying filter sizes (512, 256, 128) and kernel sizes (5, 5, 3), this pattern recurs. Two dense layers—one with 512 units and the other with 7 units—are added after the convolutional layers. Softmax activation is used for classification. The output is converted into a one-dimensional array via a flatten layer. The Adam optimizer, 100 epochs, a batch size of 64, and a 0.001 learning rate are used to train the model. The accuracy metric, categorical cross-entropy loss, and RMSprop optimizer are used in its compilation. Table 1 describes the setup in full.

Table 1. Layer structure of the lightweight CNN model

learning rate

III. Result and Discussion

Five Convolutional Neural Network (CNN) extraction features are employed in this study to get the optimal modeling accuracy. A comparison of the experiments' performance accuracy in this study is shown in Table 2 below:

Table 2. The performance accuracy of feature extraction

Feature Extraction	**Datasets**	**Accuracy**
Root Mean Square Error (RMSE)	RAVDESS, CREMA, SAVEE, TESS	93.11%
Energy	RAVDESS, CREMA, SAVEE, TESS	28.68%
Entropy of Energy	RAVDESS, CREMA, SAVEE, TESS	22.89%
Entropy	RAVDESS, CREMA, SAVEE, TESS	23.83%
Zero Crossing Rate (ZCR)	RAVDESS, CREMA, SAVEE, TESS	61.96%

According to the performance table above, modeling with a combination of the RAVDESS, CREMA, SAVEE, and TESS datasets using the Root Mean Square extraction feature produced the best experimental results, with an accuracy of 93.11%. Thus, an author-created dataset was used to test the RMSE extraction function. An accuracy of 96.07% was obtained from trials using this dataset and the RMSE extraction function. Examining these graphs is also essential for determining the model's overall capabilities and performance. Fig. 13 below displays the confusion plot derived from the confusion matrix.

CNN Configuration
conv1d_5 (Conv1D), batch_normalization_6 (BatchNormalization), max_pooling1d_5 (MaxPooling 1D), conv1d_6 (Conv1D), batch_normalization_7 (BatchNormalization), max_pooling1d_6 (MaxPooling1D), conv1d_7 (Conv1D), batch_normalization_8 (BatchNormalization), max_pooling1d_7 (MaxPooling1D), conv1d_8 (Conv1D), batch_normalization_9 (BatchNormalization), max_pooling1d_8 (MaxPooling1D), conv1d_9 (Conv1D), batch_normalization_10 (BatchNormalization), max_pooling1d_9 (MaxPooling1D),flatten_1 (Flatten), dense_2 (Dense), batch_normalization_11 (BatchNormalization), dense_3 (Dense)
Parameters: Adam Optimizer, 100 epochs, 0.0010

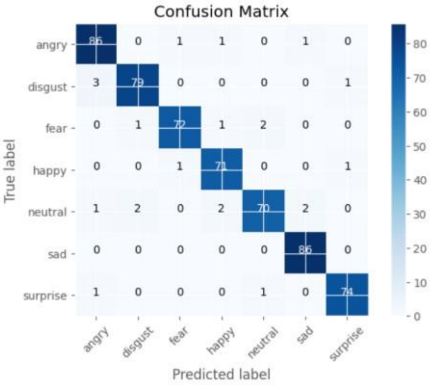

Figure. 13. Figure Confusion Matrix of proposed model

Table 3 below shows the accuracy performance comparison table that was produced based on the outcomes of the tests and experiments that conducted.

Table 3 The comparative performance accuracy

Feature Extraction	**Datasets**	**Accuracy**
Root Mean Square Error (RMSE)	RAVDESS, CREMA, SAVEE, TESS	93.11%
Root Mean Square Error (RMSE)	Proposed author dataset	96.07%

IV.Conclusions

In conclusion, the importance of a carefully considered combination is highlighted by the thorough examination of numerous datasets and extraction features. With an accuracy of 93.11%, the Root Mean Square extraction feature in particular produced remarkable results when paired with the RAVDESS, CREMA, SAVEE, and TESS datasets. This achievement served as the catalyst for additional testing on the author's unique dataset using the Root Mean Square Error extraction function.

Indeed, the results achieved an exceptional accuracy of 96.07%, surpassing earlier benchmarks. In comparison to the suggested model, this is a 2.96% improvement. These results highlight the significance of strategic decisions in improving accuracy and overall performance by highlighting how modeling techniques may be improved by customizing dataset combinations and extraction characteristics. This demonstrates the suggested Speech Emotion Recognition (SER) method's efficacy and resilience. Comparing feature extraction methods in SER approaches utilizing Deep Learning with various datasets will be one future direction. Furthermore, by efficiently recording high-level acoustic data, the combination of Convolutional Neural Networks and Root Mean Square Error (RMSE) for feature extraction may improve SER accuracy even more.

Acknowledgements

We would like to extend our gratitude to the anonymous reviewers for their valuable feedback and comments on this article. We also appreciate the support from Management & Science University (MSU) in facilitating this research.

References

[1] Blanke, E. S., A. Brose, E. K. Kalokerinos, Y. Erbas, M. Riediger, and P. Kuppens. 2020. "Mix it

[2] Jain, U., K. Nathani, N. Ruban, A. N. J. Raj, Z. Zhuang, and V. G. V. Mahesh. 2019. "Cubic SVM classifier based feature extraction and emotion detection from speech signals." In Proceedings - 2018 International Conference on Sensor Networks and Signal Processing, SNSP 2018, February 2019,

-

386–391. https://doi.org/10.1109/SNSP.2018.00081.

[3] Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, *21*(1), 93–120. https://doi.org/10.1007/s10772-018-9491-z

[4] Ibrahim, H., & Varol, A. (2020). A Study on Automatic Speech Recognition Systems. *8th International Symposium on Digital Forensics and Security*, *ISDFS 2020*, 12–16. https://doi.org/10.1109/ISDFS49300.2020.9116286

[5] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access*, *7*, 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

[6] Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning — A systematic review. *Intelligent Systems with Applications*, *20*(September 2022), 200266. https://doi.org/10.1016/j.iswa.2023.200266

[7] Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035

[8] Hashem, A., Arif, M., & Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. *Speech Communication*, *154*(September), 102974. https://doi.org/10.1016/j.specom.2023.102974

[9] Zhang, Y., Du, J., Wang, Z., Zhang, J., & Tu, Y. (2019). Attention Based Fully Convolutional Network for Speech Emotion Recognition. *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2018 - Proceedings*, 1771–1775. https://doi.org/10.23919/APSIPA.2018.8659587

[10] Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors (Switzerland)*, *20*(18), 1–16.

https://doi.org/10.3390/s20185212

[11] Kwon, S. (2020). A CNN-Assisted Enhanced Audio Signal Processing. *Sensors*.

[12] Amjad, A., Khan, L., & Chang, H. T. (2021). Effect on speech emotion classification of a feature selection approach using a convolutional neural network. *PeerJ Computer Science*, *7*. https://doi.org/10.7717/PEERJ-CS.766

[13] Mountzouris, K., Perikos, I., & Hatzilygeroudis, I. (2023). Speech Emotion Recognition Using Convolutional Neural Networks with Attention Mechanism. *Electronics (Switzerland)*, *12*(20), 1–31. https://doi.org/10.3390/electronics12204376

[14] Ullah, R., Asif, M., Shah, W. A., Anjam, F., Ullah, I., Khurshaid, T., Wuttisittikulkij, L., Shah, S., Ali, S. M., & Alibakhshikenari, M. (2023). Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer. *Sensors*, *23*(13), 1–20. https://doi.org/10.3390/s23136212

[15] Zhao, Y., & Shu, X. (2023). Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC). *Scientific Reports*, *13*(1), 1–18. https://doi.org/10.1038/s41598-023-47118-4

[16] Xu, M., Zhang, F., & Khan, S. U. (2020). Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion. *2020 10th Annual Computing and Communication Workshop and Conference, CCWC 2020*, 1058–1064. https://doi.org/10.1109/CCWC47524.2020.9031207

[17] Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Applied Sciences (Switzerland)*, *13*(8). https://doi.org/10.3390/app13084750

[18] Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics (Switzerland)*, *10*(10). https://doi.org/10.3390/electronics10101163

[19] de Lope, J., & Graña, M. (2023). An ongoing review of speech emotion recognition. *Neurocomputing*, *528*, 1–11. https://doi.org/10.1016/j.neucom.2023.01.002

[20] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future

directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8

[21] Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors (Switzerland)*, *20*(18), 1–16. https://doi.org/10.3390/s20185212

[22] Zielonka, M., Piastowski, A., Czyżewski, A., Nadachowski, P., Operlejn, M., & Kaczor, K. (2022). Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets. *Electronics (Switzerland)*, *11*(22). https://doi.org/10.3390/electronics11223831

[23] Mittal, R., Vart, S., Shokeen, P., & Kumar, M. (2022). Speech Emotion Recognition. *2022 2nd International Conference on Intelligent Technologies, CONIT 2022*. https://doi.org/10.1109/CONIT55038.2022.9848265

[24] Chatterjee, R., Mazumdar, S., Sherratt, R. S., Halder, R., Maitra, T., & Giri, D. (2021). Real-Time Speech Emotion Analysis for Smart Home Assistants. *IEEE Transactions on Consumer Electronics*, *67*(1), 68–76. https://doi.org/10.1109/TCE.2021.3056421

[25] Nanni, L., Costa, Y. M. G., Aguiar, R. L., Mangolin, R. B., Brahnam, S., & Silla, C. N. (2020). Ensemble of convolutional neural networks to improve animal audio classification. *Eurasip Journal on Audio, Speech, and Music Processing*, *2020*(1). https://doi.org/10.1186/s13636-020-00175-3

[27] Chu, H. C., Zhang, Y. L., & Chiang, H. C. (2023). A CNN Sound Classification Mechanism Using Data Augmentation. *Sensors*, *23*(15). https://doi.org/10.3390/s23156972

[28] Wijayasingha, L., & Stankovic, J. A. (2021). Robustness to noise for speech emotion classification using CNNs and attention mechanisms. *Smart Health*, *19*. https://doi.org/10.1016/j.smhl.2020.100165

[29] Ottoni, L. T. C., Ottoni, A. L. C., & Cerqueira, J. de J. F. (2023). A Deep Learning Approach for Speech Emotion Recognition Optimization Using Meta- Learning. *Electronics (Switzerland)*, *12*(23), 1–23. https://doi.org/10.3390/electronics12234859

[30] Parthasarathy, S., & Tashev, I. (2018). Convolutional neural network techniques for speech emotion recognition. *16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018 - Proceedings, September*, 121–125.

https://doi.org/10.1109/IWAENC.2018.8521333

[31] Pan, S. T., & Wu, H. J. (2023). Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with Data Augmentation. *Electronics (Switzerland)*, *12*(11). https://doi.org/10.3390/electronics12112436

[32] Baird, A., Triantafyllopoulos, A., Zänkert, S., Ottl, S., Christ, L., Stappen, L., Konzok, J., Sturmbauer, S., Meßner, E. M., Kudielka, B. M., Rohleder, N., Baumeister, H., & Schuller, B. W. (2021). An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress. *Frontiers in Computer Science*, *3*(December), 1–19. https://doi.org/10.3389/fcomp.2021.750284

[33] Atmaja, B. T., Shirai, K., & Akagi, M. (2019). Speech emotion recognition using speech feature and word embedding. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, 519–523. https://doi.org/10.1109/APSIPAASC47483.2019.9023098

[34] Deka, H., & Mittal, V. (2023). *Deep learning for speech emotion feature extraction and classification: current trends and future directions*. *44*(7), 1580–1600.

[35] Aouani, H., & Ayed, Y. Ben. (2020). Speech Emotion Recognition with deep learning. *Procedia Computer Science*, *176*, 251–260. https://doi.org/10.1016/j.procs.2020.08.027

[36] Zhao, Y., & Shu, X. (2023). Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC). *Scientific Reports*, *13*(1), 1–18. https://doi.org/10.1038/s41598-023-47118-4

[37] Mashhadi, M. M. R., & Osei-Bonsu, K. (2023). Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. *PLoS ONE*, *18*(11 November), 1–13. https://doi.org/10.1371/journal.pone.0291500