IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304127 |

Lightweight Deep Learning Framework for Speech Emotion Recognition

Mrs. C.Radha, A.V.Kishore, M.Muthamizhl, M.Manivel, M.Praveen Kumar

Associate Professor, Department of MCA, Muthayammal Engineering College, Namakkal, Tamilnadu, India

Final Year Student, Department of MCA, Muthayammal Engineering College, Namakkal, Tamilnadu, India

ABSTRACT: Artificial intelligence (AI) and machine learning (ML) are employed to make systems smarter. Today, the speech emotion recognition (SER) system evaluates the emotional state of the speaker by investigating his/her speech signal. Emotion recognition is a challenging task for a machine. In addition, making it smarter so that the emotions are efficiently recognized by AI is equally challenging. The speech signal is quite hard to examine using signal processing methods because it consists of different frequencies and features that vary according to emotions, such as anger, fear, sadness, happiness, boredom, disgust, and surprise. Even though different algorithms are being developed for the SER, the success rates are very low according to the languages, the emotions, and the databases. In this paper, we propose a new lightweight effective SER model that has a low computational complexity and a high recognition accuracy. The suggested method uses the convolutional neural network (CNN) approach to learn the deep frequency features by using a plain rectangular filter with a modified pooling strategy that have more discriminative power for the SER. The proposed CNN model was trained on the extracted frequency features from the speech data and was then tested to predict the emotions. The proposed SER model was evaluated over two benchmarks, which included the interactive emotional dyadic motion capture (IEMOCAP) and the berlin emotional speech database (EMO-DB) speech datasets, and it obtained 77.01% and 92.02% recognition results. The experimental results demonstrated that the proposed CNN-based SER system can achieve a better recognition performance than the state-of-the-art SER systems.

KEYWORDS: artificial intelligence; deep learning; deep frequency features extraction; speech emotion recognition; speech spectrograms

I. INTRODUCTION

The affective content analysis of speech signals is an active area of investigation in this era. Speech is the greatest prevailing way to exchange information among human beings, and it is worth paying attention to human-computer interaction (HCI). On the other hand, social media, written text, and other semantic information methods are also very helpful and play a crucial role in HCI. The most significant factor in human speech is emotions, which can analyze for judgments about human expressions, paralanguages, and others. Hence, the speech signal is an efficient way for the fastest communication among HCI, which efficiently recognized human behavior. Today, emotion recognition in a speech signal is one of the fastest emerging research field, where researchers have developed methods to naturally detect emotions from a speech signal [1,2]. The theory of speech emotion recognition (SER) is beneficial for education and health, and it will be widely used in these fields once they are proposed [3]. The SER plays an important role in the HCI, and researchers are making a variety of techniques in the current decade to make it efficient and robust for real-time applications [4]. In the past decade, it has been a challenging task to recognize the emotional facts and the expressive cues from the speech of an individual due to the lack of techniques and technologies [5,6]. In each era, researchers have worked to develop an efficient SER system, and they have developed several methods for preprocessing, features extraction, and classification [7]. Some researchers have concatenated the low-level descriptor (LLD) with high-level statistical functions (HSFs) for emotion recognition tasks, and they have obtained high accuracy as compared to the other features in INTERSPEECH2013 [8]. In addition, many other researchers have developed a lot of new techniques for feature extraction. Turgut et al. [9] presented an idea for feature selection based on the definition set and the emotional state.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304127 |

Fig 1: Speech Emotion Recognition using Machine Learning

With the development of skills and technologies, researchers have adopted artificial intelligence (AI) and deep learning (DL) approaches to enhance the way of the HCI, which includes emotion recognition in speech signals. Today, researchers have developed new techniques for efficient SER using AI and DL in this domain [6]. Many researchers have achieved great success in this era in order to make an efficient and robust SER system using certain DL applications, such as a deep belief network [11] (DBN), CNNs [12], and long short-term memory (LSTM) [13,14]. Most DL-based CNNs models work on 2D data, so researchers utilize deep spectrum approaches for the SER in order to extract high-level discriminative features from the speech spectrograms using the CNN model. All the CNNs models are extremely good working in the spatial domain to extract the salient features from the data [15]. Recently, fully convolutional networks (FCNs) have been introduced as a variant of the CNNs, which handle inputs of variable sizes based on their properties, and they have achieved state-of-the-art accuracy in time-series problems [16]. However, the FCNs model is not able to learn temporal features in this regard. The recurrent neural network (RNN) and the LSTM show good performances to model temporal dependency among the sequences [14,17]. The RNN-LSTM network is suitable to learn long term contextual dependencies, and it is widely used in the SER domain [18]. However, different CNNs models are proposed in the SER domain to improve the performance of the baseline SER system. The approach proposed herein aims to reduce the computational cost and improve the performance by using the lightweight CNN model.

II. LITERATURE REVIEW

In the literature, the SER has been an active research area during this decade, and many techniques are proposed in this field for an efficient SER system that utilizes new technologies. Today, deep learning approaches are rapidly used in the SER due to the increase in data and high computational power. Most researchers are showing interest in deep learning-based methods over the traditional methods, because the deep learning method is the most dominant approach [2,15]. Hence, many researchers have used the CNN model to extract the high-level features in order to show the importance and the power of deep learning, which learn high-level discriminative features. The CNN model achieved a lot of success in the visual recognition task, so researchers have therefore widely used CNNs for the feature representations in many speech analysis problems. For example, in Reference, the authors developed an SER system based on the CNN features using speech spectrograms. In addition, Reference similarly used CNN to learn the salient features and recognized the emotions in spectrograms. Today, deep learning approaches have become a recent trend to extract the CNN features from the speech spectrum and the spectrograms. Moreover, the researchers take benefits from the deep spectrum and the spectrograms by employing transfer learning techniques to trained end-to-end SER models utilizing Alex Net. The deep spectrum and the spectrograms are suitable for 2D representations of speech signals and show better performances to produce salient high-level features in the SER task, which can easily achieve the state-of-the-art results. Furthermore, for temporal information, the researchers have used sequence learning methods, such as RNNs and LSTM networks to recognize the contextual dependencies. In this regard, these networks are frequently adopted in the SER fields.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304127 |

Through the improvement of technologies, artificial intelligence and CNNs are the most popular sources that have achieved excessive success in many fields, such as handwriting recognition, object recognition, natural language processing, and. The convolutional neural networks addressed the scalability issues of the traditional neural networks by allowing them to share similar weights for multiple regions of the inputs. Usually, the CNN model consists of three main building blocks that first include the convolution layers, second the pooling layers, and finally the fully connected layers. The feature map of the input is computed by the convolution operation, and down sampling the dimension of the computed feature map is performed by the pooling strategy. Finally, the features are passed from the fully connected layers to be transformed into more reliable forms for the target predictions. Recently, CNNs has been used widely in the SER to learn the high-level discriminative features using the speech data and classifying them accordingly. The combination of the CNNs with the RNNs-LSTM is presented in improves the performance of the existence SER system.

Currently, researchers have used deep learning approaches for the proposed a new technique for the emotion recognition based on the DCNN. This method extracts features by using the Alex Net model and a trained conventional classifier, which is a support vector machine (SVM), to predict the emotions [38]. A CNN model extracts features from the whole utterance and feeds them to the LSTM or the RNNs to extract long term contextual dependencies in the speech signals. presented a method for the SER using the DBN and the SVM where the high-level features are extracted by the DBN and then classified by the SVM. Similarly, used the DBN model to represent the salient features and fed to the SVM for classification using the Chinese dataset built by Chinese Academy of Sciences' Institute of Automation (CASIA). In, the authors proposed a novel SER system based on a stride CNN to extract the salient discriminative features and trained them in an end-to-end manner. Furthermore,presented a SER based on the key segments to reduce the cost and increase the accuracy utilizing three challenging speech emotion datasets. However, in this research, we proposed a novel CNN architecture for the SER based on the deep frequency pixel features to recognize the emotional features in the speech spectrograms. We used plain rectangular shape filters in the convolution layer with a modified pooling strategy to extract the hidden deep frequency features. In this research, we proposed a simple and lightweight CNN model for the SER using fewer convolutional layers in order to reduce the cost complexity and improve the performance of the baseline methods. A detailed explanation about the proposed SER is illustrated in the subsequent section.

III. METHODS

A spectrogram is a two-dimensional representation of the speech signals, and a dimensional setting of the speech data for 2D deep learning models is a challenging task. The dimension conversion of the speech data was necessary for the 2D CNN model to learn the salient discriminative features from the speech signals to make an efficient SER system. A spectrogram is 2D representations of the speech data, which is suitable to show the strength of the speech signals with respect to the frequencies [4].

For a visual representation, we made a spectrogram to plot the speech signals with respect to time by applying a short-term Fourier transformation (STFT) algorithm. The STFT algorithm is used to convert the short-term speech signals or the long-time speech signals into segments, which have similar lengths. Finally, the fast Fourier transformation (FFT) is applied to that segment to compute the Fourier spectrum. Actually, the spectrogram is a combination of the frequencies with respect to time in two dimensions.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304127 |

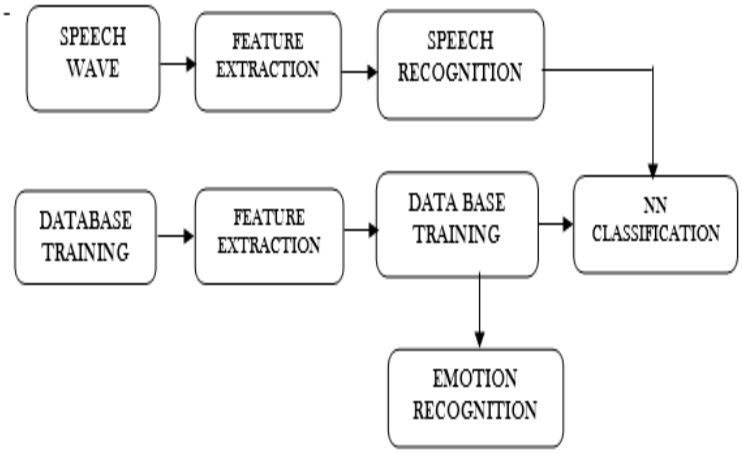

Fig 2: Work Flow

The suggested CNN architecture is illustrated. Most CNN models are designed for different types of tasks that are related to computer visions, such as classification, localization, recognition, tracking, and retrieval in order to achieved high accuracies. We were inspired by their performance and adapted it into the SER domain to make a significant and robust system to recognize emotions in the speech data. Basically, the CNN model has three components, which include the convolution layers, the pooling layers, and the fully connected layers. The convolution layers have many numbers of filters to compute the initial feature map and then feed to the pooling layer in order to reduce the dimensions for fast processing and computations. Similarly, the fully connected layers are used to learn the global features and recognize the hidden cues in the input data and then fed to the SoftMax layer to produce the different probabilities of classes. In this analysis, we utilized the basic structure of the CNN and constructed a simple and lightweight model for the SER using plain rectangular shape filters in the convolution layer with a modified pooling strategy. Our model used the same structure filters in the convolutional layer and in the pooling layer with a small receptive field in order to capture the hidden cues and the important cues. Our model uses the simple rules of CNN architecture, which means that the number of filters must be the same when the same output features map is required. If the size gets cut in half or is reduced up to a half, then the number of filters must be double in the convolution layer to sustain the complexity in each layer.

We used a simple and lightweight CNN model to extract the high-level salient features utilizing deep frequency pixel after obtaining the speech spectrograms. Actually, the CNN signifies the feedforward neural networks, which have multiple layers, such as convolution, pooling, and fully connected layers, that exploit the local connectivity and the correlation among the neurons in the connecting layers. Our designed CNN model consists of eight convolution layers with a rectified linear unit (ReLU), three max-pooling layers followed by batch normalization (BN), and two fully connected layers with a SoftMax classifier. The following are detailed explanations of each layer.

The first convolution layer (C1) has ten filters that are a (9 × 1) size to prepare the data or the features map by utilizing ten (10) filters that produce a new tensor deprived of padding. In the second layer, C2 has ten filters, which are size (5 × 1), that evaluate the obtained features map of C1 resulting from the new tensor. The C2 layer focuses on the similarities and measure them accordingly. We used a modified pooling strategy in this model that followed the convolutional layers. We used a more convolutional layer to make deeper networks to extract the high-level salient features. Similarly, the C3 layer has also the same numbers of filters, which are a (3 × 1) size, using no stride setting and padding that obtain a new features map that has the same size as C2. After C3, we used a max-pooling layer of filter size (3 × 1) with no stride setting to reduce the dimensionality [24]. The pooling layers always added on the top of the convolution layer to compute the resolutions and represent the subsampling. The pooling layer followed by the BN layer to rescale or normalize the data to increase the performance of the model and reduced the cost computations.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304127 |

IV. RESULTS ANALYSIS

We practically prove our system in this section, which we tested over two benchmark IEMOCAP, and EMO-DB speech datasets in order to show the significance and the robustness of the model for the SER using speech spectrograms. We checked the performance of our SER system and compared it with other baseline SER systems using the same phenomena. A detailed description of the utilized datasets, the accuracy matrices, and the experimental results are discussed in the upcoming sections.

The IEMOCAP, which is the interactive emotional dyadic capture, is the speech emotion acted dataset consuming the English dialogs. The dataset has five sessions, and each session has two speakers, which include one male and one female to record the different emotions. There is a total of 10 speakers that were used to record the dataset, and all the speakers are professionals or experts. The dataset contains 12 h of audiovisual data that was recorded with a 16 kHz sampling rate. There are different emotions, and all the emotions were scripted

The test set is used to evaluate the model recognition capability on the unknown utterances of data. The error in the model prediction approximates the system generalization error. In this paper, the cross-validation estimation method is utilized to evaluate all the datasets in an effective manner. The data in the database is split into two parts, which include the training data and the testing data. The initial data is separated into k parts. One part is used as a test data, and the other k data is used for training. Each part is then used diagonally for testing. The test operation, which is repeated k-times, in different parts of all the data is called a k-fold cross-validation. We used a well-known 5-fold cross-validation assessment method for the analysis of our method.

We utilized the statistical parameters to investigate the accomplishment of the proposed method, which are computed using the formulas [36]. The True-Positive (TP) which is the number of correctly predictable real positive records, False-Negative (FN), which is the number of wrongly predictable real positive records, True-Negative (TN), which is the number of correctly predictable real negative facts, and the False-Positive (FP), which is the number of wrongly predictable real negative numbers. T describes correct predictions, F describes wrong predictions, and the summation of the TP and the FN is shown as the amount of positive data. The summation of the TN and the FP indicates the amount of negative data in the real condition [50]. The accuracy identifies the overall recognition ratio. Meanwhile, the accuracy alone will not be satisfactory, and additional parameters, such as precision, recall, and the F1-score evaluation criteria are required.

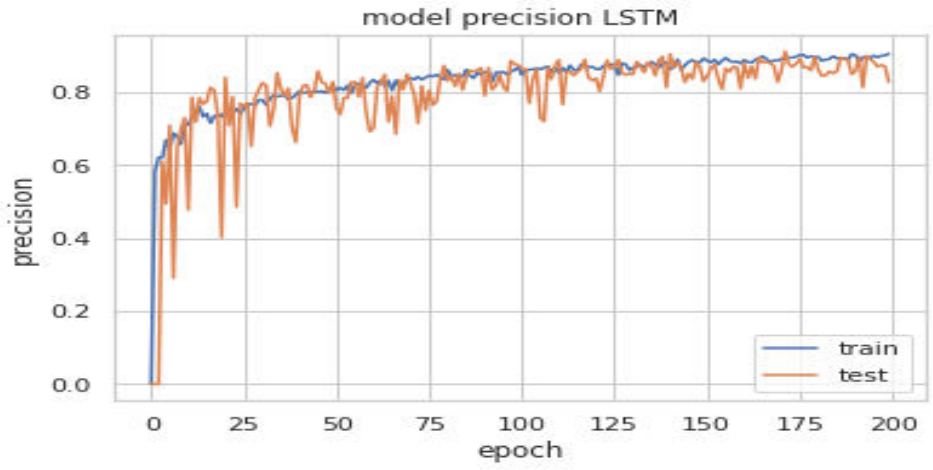

Fig 3: Result Analysis

In this study, the proposed CNN architecture, which is a plain rectangular shape convolutional filter, and a modified pooling strategy are considered as a major contribution for the SER. We used a novel filter shape and recognized the deep frequencies features from the speech spectrogram by applying a modified pooling technique in order to reduce the dimensionality and increase the accuracy. According to the best of our knowledge, the SER domain lacks this type of filter and frequency feature selection to recognize the emotions in the speech data. We deeply investigated the literature

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304127 |

of the SER domain, which has two major issues that include i: the recognition rate and ii: the cost computations or the time complexity. The researchers have developed many techniques in this domain utilizing hand-crafted features and high-level features using the classical models and the CNN models, but the recognition accuracy is still low, and the cost complexity is very high. In this research, we addressed these issues and made a novel CNN model for the SER using a modified filter shape and a pooling scheme in order to recognize the emotions in the speech data. Our model used a speech spectrogram as the input, which is a visual representation of the frequencies. We designed a new filter for the convolutional operation with a plain rectangular shape to extract the deep CNN features from the frequency pixels, which were recognized accordingly. With the usage of the deep frequency features, our model recognized the speech emotions with high accuracy. After resolving this issue, we focused on time complexity, which followed the designed filter and a pooling scheme, and made a simple CNN model that utilized fewer layers. Our proposed model used eight convolutions, three poolings, 2 batch normalizations, and 2 fully connected layers with a SoftMax classifier in order to recognized the emotions in the speech. Due to this simple structure,

V. CONCLUSION

In this paper, a novel and lightweight CNN model was proposed for the SER based on the deep frequency features using speech spectrograms. A plain rectangular shape filter was used to extract the deep frequency features in order to ensure the accuracy. We designed a CNN model that utilized eight convolutional, three max-pooling, and two fully connected layers for the SER. Due to the usage of fewer layers, we reduced the cost and the time computation. The model performance was evaluated over two benchmarks, which include the IEMOCAP and the EMO-DB datasets, in order to show the effectiveness and the significance. The prediction results of the proposed system outperformed the state-of-the-art methods. Our model achieved outstanding 77.01% recognition results for the IEMOCAP dataset and 92.02% for the EMO-DB dataset. The proposed model increased by 5% and 6% over the state-of-the-art accuracy, which is a great achievement in the current era to recognize all the emotions with better accuracy and a reduced model size in order to show computational friendly output. The output of the proposed system showed the strength of the system for a SER that utilizes speech spectrograms.

In the future, we will focus on adopting this type of system in an automatic speaker recognition system and further elaborate it for signals processing tasks. Similarly, it would be advisable to test the strength of the system and the achieved outcomes on diverse databases, which include natural databases. The recognition rates can be increased and even combined with other deep learning methods. The studies can be enriched by both spectrums and spectrograms.

REFERENCES

1. Nardelli, M.; Valenza, G.; Greco, A.; Lanata, A.; Scilingo, E.P. Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Trans. Affect. Comput.* **2015**, *6*, 385–394. [Google Scholar] [CrossRef]
2. Kwon, S. A CNN-Assisted enhanced audio signal processing for speech emotion recognition. *Sensors* **2020**, *20*, 183. [Google Scholar]
3. Swain, M.; Routray, A.; Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: A review. *Int. J. Speech Technol.* **2018**, *21*, 93–120. [Google Scholar] [CrossRef]
4. Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Baik, S.W. Deep features-based speech emotion recognition for smart affective services. *Multimed. Tools Appl.* **2019**, *78*, 5571–5589. [Google Scholar] [CrossRef]
5. Pandey, S.K.; Shekhawat, H.; Prasanna, S. Deep learning techniques for speech emotion recognition: A review. In Proceedings of the 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 16–18 April 2019. [Google Scholar]
6. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* **2019**, *7*, 117327–117345. [Google Scholar] [CrossRef]
7. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [Google Scholar] [CrossRef]
8. Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E.; et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In Proceedings of the INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013. [Google Scholar]

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304127 |

9. Özseven, T. A novel feature selection method for speech emotion recognition. *Appl. Acoust.* **2019**, *146*, 320–326. [Google Scholar] [CrossRef]
10. Jing, S.; Mao, X.; Chen, L. Prominence features: Effective emotional features for speech emotion recognition. *Digit. Signal Process.* **2018**, *72*, 216–231. [Google Scholar] [CrossRef]
11. Zhu, L.; Chen, L.; Zhao, D.; Zhou, J.; Zhang, W. Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. *Sensors* **2017**, *17*, 1694. [Google Scholar] [CrossRef] [PubMed] [Green Version]
12. Liu, X.; van de Weijer, J.; Bagdanov, A.D. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1862–1878. [Google Scholar] [CrossRef] [Green Version]
13. Mustaqeem, M.; Sajjad, M.; Kwon, S. Clustering based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* **2020**, *8*, 79861–79875. [Google Scholar] [CrossRef]
14. Karim, F.; Majumdar, S.; Darabi, H. Insights into LSTM fully convolutional networks for time series classification. *IEEE Access* **2019**, *7*, 67718–67725. [Google Scholar] [CrossRef]
15. Wang, H.; Zhang, Q.; Wu, J.; Pan, S.; Chen, Y. Time series feature learning with labeled and unlabeled data. *Pattern Recognit.* **2019**, *89*, 55–66. [Google Scholar] [CrossRef]
16. Naqvi, R.A.; Arsalan, M.; Rehman, A.; Rehman, A.U.; Loh, W.K.; Paul, A. Deep learning-based drivers emotion classification system in time series data for remote applications. *Remote Sens.* **2020**, *12*, 587. [Google Scholar] [CrossRef] [Green Version]
17. Zeng, M.; Xiao, N. Effective combination of DenseNet and BiLSTM for keyword spotting. *IEEE Access* **2019**, *7*, 10767–10775. [Google Scholar] [CrossRef]
18. Tao, F.; Liu, G. Advanced LSTM: A study about better time dependency modeling in emotion recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018. [Google Scholar]
19. Wang, H.; Wu, J.; Zhang, P.; Chen, Y. Learning shapelet patterns from network-based time series. *IEEE Trans. Ind. Inform.* **2018**, *15*, 3864–3876. [Google Scholar] [CrossRef]
20. Huang, Z.; Dong, M.; Mao, Q.; Zhan, Y. Speech emotion recognition using CNN. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014. [Google Scholar]
21. Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; Zhang, S. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimed.* **2017**, *20*, 1576–1590. [Google Scholar] [CrossRef]

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details