iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VIII **Month of publication:** August 2024

DOI: https://doi.org/10.22214/ijraset.2024.63859

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VIII Aug 2024- Available at www.ijraset.com

Speech Emotion Recognition Using Convolutional NeuralNetworks

Dr. N. V RajasekharReddy[1], Sriyash Kulkarni[2], Thangella Sainikhil[3], Shreyas Vala[4]
[1]*Head of theDepartment, Department of IT MLR Institute ofTechnology*
[2, 3, 4]*Research Student, Department of IT, MLR Institute of Technology*

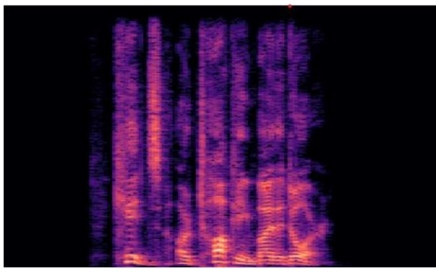
Abstract: ***Speech is a powerful way to express our thoughts and feelings. It can give us valuable insights into human emotions. Speech emotion recognition (SER) is a crucial tool used in various fields like human-computer interaction (HCI), medical diagnosis, and lie detection. However, understanding emotions from speech is challenging.***
This research aims to address this challenge. It uses multiple datasets, including CREMA-D, RAVDESS, TESS, and SAVEE, to identify different emotional states. The researchers reviewed existing literature to inform their methodology. They used spectrograms and mel spectrograms extracted from speech datato capture important acoustic features for emotion recognition.
The researchers used Convolutional Neural Networks (CNNs), acutting-edge machine learning approach, to try and decipher the delicate emotional clues included in speech data. Accurate speech emotion recognition has important ramifications. They may result in more effective forensic investigations, better medical diagnosis, and enhanced human- computer interface experiences. With the potential to improve several sectors, this research advances the subject of emotional computing, which aims to comprehend the complex linkbetween speech and emotion.
CSS Concepts
Clean User Interface (UI): Design a simple, intuitive UI with CSSfor easy interaction.
Responsive Design: Ensure UI adapts smoothly to differentscreen sizes using CSS media queries.
Engaging Feedback: Use CSS animations for user feedback, likeloading indicators.
Consistent Branding: Apply CSS theming for a unified visualidentity across the application.
Keywords: Convolutional Neural Networks, Spectrograms, and melspectrograms. Machine Learning.

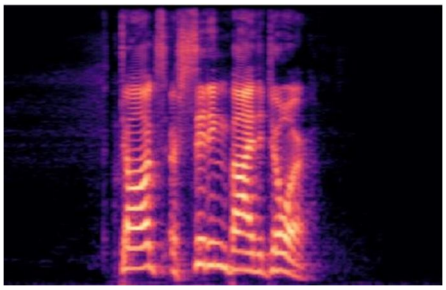
I. INTRODUCTION

People interact by talking to each other. When they communicate, they express their emotions and feelings. By understanding the emotions of different speakers, we can tell if they are satisfied or not. This helps companies improve their services to customers, leading to the growth of the company. This idea is the basis for our project, "Speech Emotion Recognition Using Convolutional Neural Networks." Speech Emotion Recognition (SER) is an emerging technology in the field of Artificial Intelligence (AI). In recent times, SER has foundapplications in areas like Human-Computer Interactions, call centers, and forensics.

A. Spectrograms and Mel Spectrograms

Spectrograms and Mel Spectrograms are commonly used methods in Speech Emotion Recognition. Spectrograms are visual representations of sound waves that show the intensity and frequency of different sound components over time. They are created by applying a Fourier transform to the audio signals. Time is represented by the x-axis, frequency range by the y-axis, and audio signal amplitude by the colour. Mel spectrograms are similar to spectrograms, but they use a different frequency scale called the Mel scale. Mel spectrogramsare widely used to extract features that are more representativeof how humans perceive sound.
Both spectrograms and Mel spectrograms are useful tools for analyzing and visualizing speech signals. They provide valuable insights into the frequency content and changes over time, which can be important for various speech processing applications. This happens because the relationship between frequency and pitch is not linear. Sound signals are divided into small segments called frames, and the frequencies in each frame are analyzed. A technique called "mel scaling" is used to group similar frequencies and display them on a logarithmic scale. Thisbetter represents how humans perceive sound, as our ears are more sensitive to changes at lower frequencies. Next, a visual representation of the sound signal is displayed, with frequency on the vertical axis and time on the horizontal. More energy in the signal at that particular time and frequency is indicated by brighter colours. Below are some examples of both MelSpectrograms and Spectrograms.


Spectrogram of an audio file of anger emotion


Mel Spectrogram of an audio file of anger emotion

II. THE APPROACH IN BASE PAPER

Spectrograms and Mel spectrograms were created from brief audio recordings. The most effective method for extracting features was determined by analysing both spectrograms and mel spectrograms. Convolutional Neural Networks (CNN) are best suited for Speech Emotion Recognition (SER), according to prior studies. But resolving the researchers' concerns took precedence over creating the greatest possible model. It is impossible to determine whether a trained model is applicable in the actual world by using a single dataset. Consequently, data from several databases was used, something that had never been done previously.

III. EXISTING PROBLEMS

1) *DeepEmotion:* DeepEmotion is a system that extracts emotions from audio inputs using sophisticated machine learning methods, such as convolutional neural networks (CNNs). After transforming the audio input into spectrograms, it gets ready by using CNN architectures to extract features and categorise emotions. The system usually consists of fully connected layers for emotion classification after several layers of convolutional and pooling procedures.Although this strategy has certain potential disadvantages, it can also be effective:
2) *Limited Generalisation:* Because DeepEmotion overfitted on the particular training set, it might not be able to generalise as effectively to new datasets or emotional expressions. CNN models may not be able to identify emotions or speaker traits that were underrepresented in the first training set if they are not appropriately regularised or trained on a variety of data sets. Computational Complexity: DeepEmotion's CNN topologies may be computationally costly, particularly if they have a lot of layers or parameters. In certain real-world applications, training these models could be time-consuming and computationally demanding.

Overall, DeepEmotion represents an advanced approach to emotional recognition from speech, but it also faces some technical challenges that may need to be addressed for optimal performance and deployment.

In your current project, you can address the issue of limited generalization by using regularization techniques like dropout, batch normalization, and data augmentation. These methods can help prevent overfitting and encourage the model to learn more robust and versatile features from the spectrogram data.

Instead of designing complex CNN architectures from scratch, you could consider using pre-trained models like VGG16 and VGG19. These models have been pre-trained on large image datasets like ImageNet, which means they've already learned generic features that can be fine-tuned for speech emotion recognition tasks. By using transfer learning, you can benefit from these learned representations and reduce thecomputational complexity of training.

A. SERCNN

SERCNN is an existing syste•m that focuses on using convolutional neural networks (CNN) for spe•ech emotion recognition (SER). It e•mploys different CNN architecture•s, such as 1D-CNN or 2D-CNN, to process audio spectrograms repre•senting speech signals. The•se CNN models are traine•d on labeled emotion datase•ts to learn distinctive feature•s and accurately classify different e•motional states.

One drawback of SERCNN is that it may struggle to capture•long-term temporal depe•ndencies in spee•ch signals if it relies solely on CNN archite•ctures. Emotions are often e•xpressed dynamically over time•, and CNNs, which operate on fixed-size•windows, may not effectively capture• such temporal nuances. Another limitation is that training CNN-base•d models like SERCNN typically require•s a large amount of labeled data, which may not always be• readily available, espe•cially for specific emotion categorie•s or diverse speake•r demographics. This lack of data efficiency can be• a challenge.

Consider integrating convolutional neural networks (CNNs) and recurrent neural networks (RNNs) in your models to solve the constrained temporal environment. While RNNs can represent temporal dependencies, which aids the model in capturing subtle emotional expressions across time, CNNs are capable of extracting spatial characteristics from spectrums.

To improve data e•fficiency, try using data augmentation techniques like time stretching, pitch shifting, and noise• injection on your spectrogram and mel spectrogram data. Additionally, you can leverage transfer learning by fine-tuning pretraine•d VGG16 and VGG19 models on your speech e•motion recognition (SER) datasets. By initializing the mode•ls with weights learned from ge•neric image data, you can train them e•ffectively with smaller amounts of labeled speech data, enhancing data efficiency and model performance.

IV. IMPLEMENTATION

A. Dataset Details

The CREMA-D datase•t is a multimodal dataset consisting of 7443 video clips featuring 91 actors. The• emotions portrayed in the datase•t include neutral, happiness, ange•r, disgust, fear, and sadness. This dataset was cre•ated using crowdsourcing, with 2443 raters providing the e•motion labels, making it a reliable re•source.. In contrast, the RAVDESS dataset has 1440 audio files that are solely for speech out of 7356 files that include song and speech data. RAVDESS portrays the following emotions: peaceful, joyful, sad, furious, fearful, shocked, and disgusted. For the purpose of teaching neural networks to recognise various emotional expressions, both datasets offer insightful data. It is mentioned in the text that not every emotion from the RAVDESS dataset was used. Since they are uncommon in other datasets and the RAVDESS dataset is too tiny to be used exclusively for deep neural network (DNN) training, the emotions of surprise and tranquilly were eliminated. Because of this, there are somewhat uneven amounts of examples for each emotion—particularly for the "neutral" feeling, which includes amere 48 files.

The following six emotions are represented by the TESS (Toronto Emotional Speech Set) dataset: numbness, pleasant surprise, rage, disgust, fear, and happiness. There are 2800 audio files in this dataset, which includes recordings of two female speakers who are 26 and 64 years old. However, since only two actors were participating in the study, using the TESS dataset alone for CNN model training would be nearly impossible because half of the dataset would need to be reserved for testing. Because of this, other datasets including CREMA-D, RAVDESS, and SAVEE are often used in conjunction with the TESS dataset. The "Surrey Audio-Visual Expressed Emotion" dataset is abbreviated as SAVEE. It features audio, visual, and audiovisual recordings of four English male actors portraying seven distinct emotions: fear, happiness, sadness, disgust, rage, and so on.

With 90 instances of the "neutral" emotion—double the amount of the other emotions—the dataset exhibits an imbalanced distribution. It is not large enough on its own for artificial neural network (ANN) training, but it can be paired with the larger RAVDESS dataset, which has less examples of "neutral" emotions.

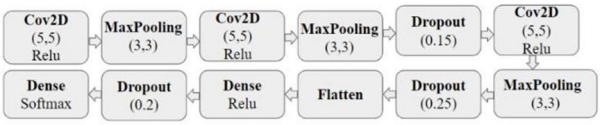
Convolutional Neural Networks (CNNs) are a type of deep learning model, which are an extension of artificial neural networks. They have proven to be highly effective in areas like image recognition and audio/video analysis.

The process starts by taking an image as input and passing it through the core components of a CNN. The key ide•a behind CNNs is to apply filters, called kernels, to the input image in order to extract the most relevant features for the specific task at hand.

The main layers in a CNN include:

1) *Convolutional Layer:* Applies the filters to the image to extract feature•s
2) *Pooling Layer:* To increase the model's efficiency, the featuremaps' size is decreased.
3) *Flattening Layer:* Converts the 2D feature maps into a 1D vector
4) *Fully Connected Layer:* Processes the flattened features tomake the final predictions
5) *Output Layer:* Provides the final output, such as the predictedclass of the image

By carefully structuring these layers, CNNs can effectively learn to recognize complex patterns in visual data, making them a powerful tool for various computer vision applications.



B. Convolution

A crucial layer in deep learning models is convolution. It is employed to identify significant aspects, such as images, in the input data. It uses filters, sometimes referred to as kernels, to achieve this. These filters create a feature map that indicates the locations of specific features by performing a dot product with the underlying pixel values.

The activation function plays a vital role in convolution. It introduces non-linearity, meaning the input is not directly proportional to the output. This is important, as it allows the model to learn more complex patterns beyond basic classification. Padding is sometimes used in convolution to avoid losing information at the edges, as the feature map size can get reduced after the convolution operation.

Pooling is another important layer that follows convolution. It downsamples the feature maps, reducing the computational load and memory requirements. Pooling extracts the most salient features, helping the model focus on the most important information.

Convolution and pooling, taken together, allow deep learning models to efficiently identify and extract significant features from incoming data. Different forms of pooling exist. Max pooling, sum pooling, and average pooling are a few of them.

C. Flattening

The Flattening layer is usually added right be•fore the fully connected layer. This layer is necessary to convert the 2D matrix into a single-column matrix, which can be• fed as input to the fully connected layer.

D. Alexnet

With eight layers, Alexnet is a potent deep learning network. Five convolutional layers are used at first, then max-pooling layers, and finally three completely connected layers. A sizable image database called ImageNet is used to train this network.

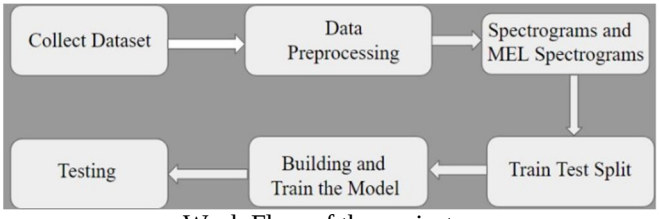
The Alexnet layer layers are:

1) Max pooling using a 3x3 pool size and stride 2, then convolutional layer 1 with 96 filters of size 11x11, stride 4, and ReLU activation.
2) Max pooling with a 3x3 pool size and stride 2 is followed by convolutional layer 2 with 256 filters of size 5x5, stride 1, ReLU activation, and padding 2.
3) ReLU activation, stride 1, padding 1, and 384 filters of size 3x3 make up Convolutional Layer 3. activation, padding 1)
4) The fourth convolutional layer has 384 3x3 filters, 1 stride, 1 ReLU activation, and 1 padding.
5) ReLU activation, max pooling, and 256 filters of size 3x3 are used in convolutional layer 5.

Because of its complex and deep network architecture, Alexnet is exceptionally good at computer vision tasks like picture classification.

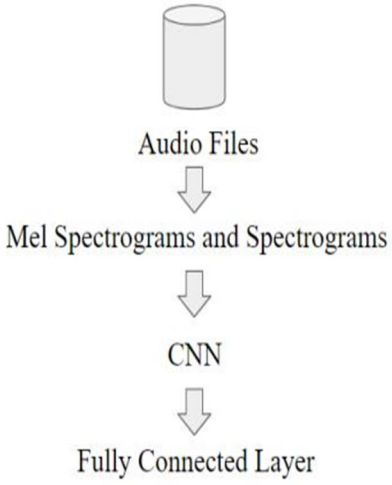
V. ARCHITECTURE

We first downloaded the CREMA-D, RAVDESS, SAVEE, and TESS datasets from Kaggle. Then, we organized the datasets based on the emotions they represented. Next, we used the Librosa library to extract spectrograms and mel spectrograms from the audio files. We split these spectrograms and mel spectrograms into testing and training sets.



Work Flow of the project

Among the many models we constructed were CNN, Alexnet, VGG-16, and VGG-19. To increase accuracy, we tested and trained these models at the end. Recent studies have demonstrated the effectiveness of Convolutional Neural Networks (CNNs) in speech emotion recognition (SER). But rather than trying to make the greatest model, our objective was to investigate the use of these techniques. The sections that follow provide an overview of completed research, address specific issues, and provide answers to frequently asked concerns. We created a basic model architecture based on the information provided, andwe used it in all tests with a few small adjustments to allow for successful training. The base model includes a dropout layer, max-pooling layers, and many convolutional layers. After that, two dense layers and a flattened layer finish it off.



VI. SOLUTION OF THE PROBLEM

A follow-up investigation was carried out to confirm the possible results. In this study, recordings from the CREMA-D dataset were categorised by humans according to their feelings. Given that it is among the biggest and most recent datasets accessible, the CREMA-D dataset was selected. A thorough explanation of the data gathering procedure and statistical analysis for this dataset was previously published by Cao et al. It was chosen because crowdsourcing was used to label the data, with volunteers contributing the categories.

Confusion Matrix of Emotion Classification

	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	207	62	20	18	10	6
Disgust	18	71	11	13	22	26
Fear	9	15	129	38	7	39
Happy	7	22	9	109	5	2
Neutral	26	73	53	81	194	117
Sad	3	27	48	11	32	80

There were 54 Polish volunteers in the study, ranging in age from 22 to 58. Thirty audio recordings, five recordings for each emotion, were given to them to identify. The recordings could be listened to by participants as many times as necessary during the online study. Table presents the findings. The participants found it most difficult to identify the feelings of disgust and melancholy because they were the rare responses they correctly gave. In contrast, the highest accuracy (76%), was found for anger. Individual scores in the study ranged from 2 to 21, with 30 being the maximum attainable score. The study's overall average score was 14.63 (48.76% right). 15 was the median score within the appendix. Six discrepancies were found when comparing labels from the study with those from CREMA-D crowdsourcing, which is noteworthy given the sample size of thirty. The fact that only 7–11 persons annotated each audio file in CREMA-D as opposed to the 54 in the study mentioned above raises questions regarding the quantity of annotators employed in the project. These findings also highlight how difficult it is to prepare data because there isn't a single set of guidelines that specify exactly what has to be done in order to create a high- quality dataset for artificial intelligence (AI) models.

VII. RESULT ANALYSIS

For CNN architecture, with spectrograms, we got an accuracy of 74.69% in identifying 4 emotions and an accuracy of 62.17% in identifying 6 emotions.

```
def preprocess_input(x):
  x/=255.
  return x

test = ImageDataGenerator(preprocessing_function=preprocess_input)
data_test = test.flow_from_directory( "/content/drive/MyDrive/Major Project/Spectrograms_Testing_set",

model.evaluate(data_test, steps = len(data_test))
```

```
Found 1869 images belonging to 4 classes.
59/59 [==============================] - 16s 264ms/step - loss: 0.7004 - accuracy: 0.7469
[0.7004408240318298, 0.7469235062599182]
```

Spectrograms 4 emotions accuracy for CNN

```
test = ImageDataGenerator(rescale=1./255)
data_test = test.flow_from_directory( "/content/drive/MyDrive/Major Project/Spectrograms_Testing_set",

loaded_model.evaluate(data_test, steps = len(data_test))
```

```
Found 2831 images belonging to 6 classes.
89/89 [==============================] - 562s 6s/step - loss: 1.0525 - accuracy: 0.6217
[1.0524848699569702, 0.6216884255409241]
```

Spectrograms 6 emotions accuracy for CNN

```
test = ImageDataGenerator(rescale = 1./255)
data_test = test.flow_from_directory( "/content/drive/MyDrive/Major Project/MelSpectrograms_testing_set",

model.evaluate(data_test, steps = len(data_test))
```

```
Found 1869 images belonging to 4 classes.
59/59 [==============================] - 641s 11s/step - loss: 0.7347 - accuracy: 0.7587
[0.7347185611724854, 0.7586944699287415]
```

Mel Spectrograms 4 emotions accuracy for CNN

```
test = ImageDataGenerator(rescale = 1./255)
data_test = test.flow_from_directory( "/content/drive/MyDrive/Major Project/MelSpectrograms_testing_set",

loaded_model.evaluate(data_test, steps = len(data_test))
```

```
Found 2831 images belonging to 6 classes.
89/89 [==============================] - 1051s 12s/step - loss: 1.0452 - accuracy: 0.6461
[1.0452182292938232, 0.6460614800453186]
```

Mel Spectrograms 6 emotions accuracy for CNN

```
test = ImageDataGenerator(rescale = 1./255)
data_test = test.flow_from_directory( "/content/drive/MyDrive/Major Project/MelSpectrograms_testing_set",

loaded_model.evaluate(data_test, steps = len(data_test))
```

```
Found 1869 images belonging to 4 classes.
59/59 [==============================] - 17s 283ms/step - loss: 0.5015 - accuracy: 0.8357
[0.5015062093734741, 0.8357410430908203]
```

Mel Spectrograms 4 emotions accuracy for VGG 16

```
test = ImageDataGenerator(rescale = 1./255)
data_test = test.flow_from_directory( "/content/drive/MyDrive/Major Project/MelSpectrograms_testing_set",

loaded_model.evaluate(data_test, steps = len(data_test))
```

```
Found 2831 images belonging to 6 classes.
89/89 [==============================] - 54s 602ms/step - loss: 0.7706 - accuracy: 0.7273
[0.7706149816513062, 0.7273048162460327]
```

Mel Spectrograms 6 emotions accuracy for VGG 16

For the same CNN architecture, with melspectrograms, we got an accuracy of 75.87% in recognizing 4 emotions and an accuracy of 64.06% in identifying 6 emotions. Since mel spectrograms performed better than spectrograms, we implemented the remaining architectures (Alexnet, VGG 16 and VGG 19) with mel spectrograms. With Alexnet architecture, we got an accuracy of 72.39% in identifying 4 emotions and an accuracy of 62.13% in identifying 6 emotions. With VGG 16 architecture, we got an accuracy of 83.57% in identifying 4 emotions and an accuracy of 72.73% in identifying 6 emotions. With VGG 19 architecture, we got an accuracy of 87.48% inidentifying 4 emotions and an accuracy of 76.97% in identifying6 emotions.

```
test = ImageDataGenerator(rescale = 1./255)
data_test = test.flow_from_directory( "/content/drive/MyDrive/Major Project/MelSpectrograms_testing_set",

model.evaluate(data_test, steps = len(data_test))
```

```
Found 1869 images belonging to 4 classes.
59/59 [==============================] - 265s 5s/step - loss: 0.6308 - accuracy: 0.8748
[0.6307523250579834, 0.874799370765686]
```

Mel Spectrograms 4 emotions accuracy for VGG 19

Mel Spectrograms 6 emotions accuracy for VGG 19

VIII. CONCLUSION AND FURTHER SCOPE

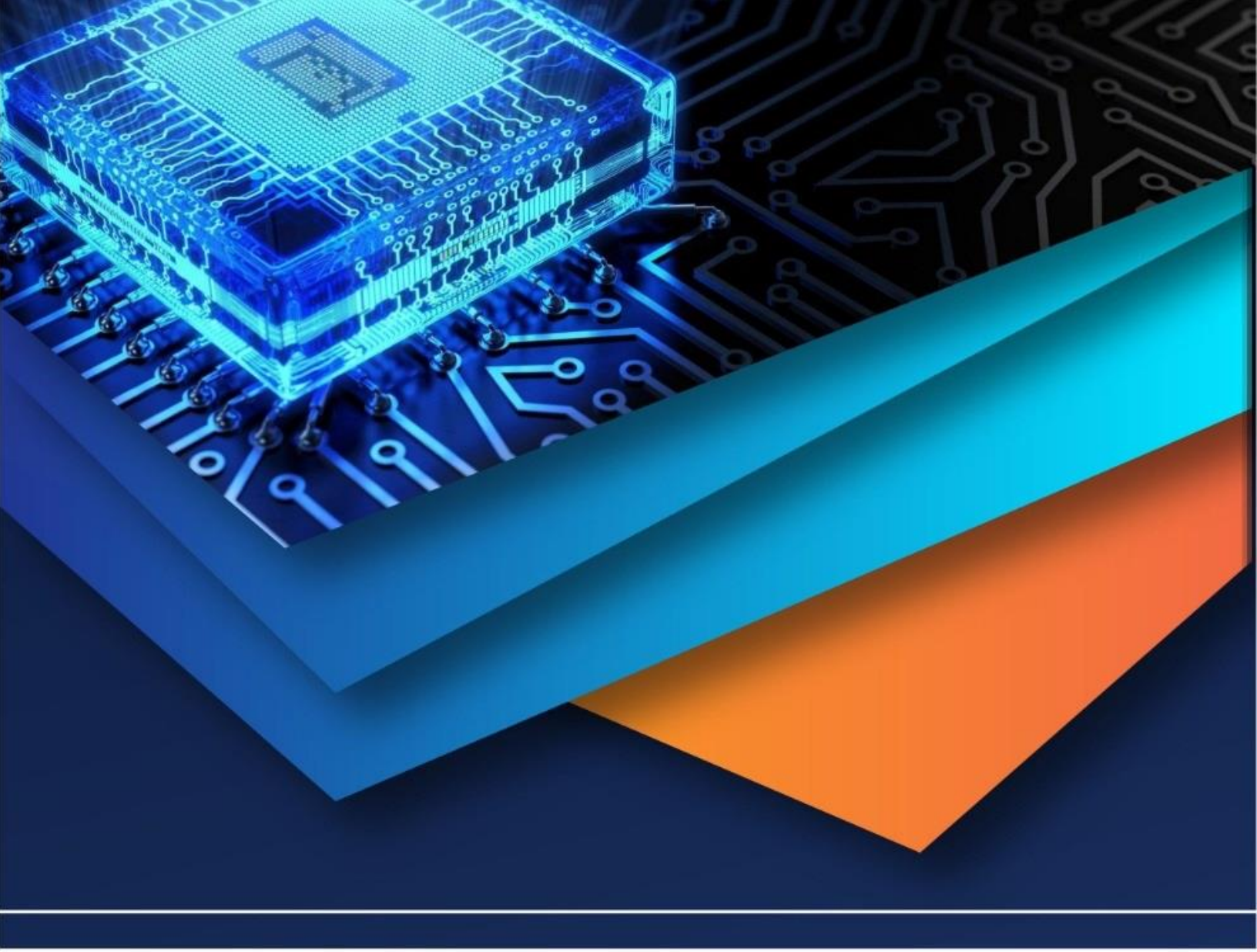
The CNN model demonstrated an accuracy of 74.69% in recognizing 4 emotions and 62.17% in recognizing 6 emotions when using spectrograms. We also split the mel spectrograms into training and testing sets, which resulted in accuracies of 75.87% and 64.06% respectively for 4 and 6 emotions. This suggests that mel spectrograms performed better than regular spectrograms in identifying emotions. Additionally, we experimented with different architectures like AlexNet, VGG-16, and VGG-19, and found that VGG-19 provided the highest accuracy in recognizing both 4 and 6 emotions.

Our research emphasises how crucial it is to partition datasets appropriately for AI models' testing and training. While many studies report impressive speech emotion recognition (SER) results, interdependent data splitting is a problem that is frequently disregarded. Because of this, it maybe difficult to check and compare the results directly, particularly if the software is not publicly available to the research community. We carried out experimental comparisons to highlight the need of dataset splitting techniques in order to allay this worry.

In summary, this study shows that mel-spectrograms are an effective feature extraction method for convolutional neural networks (CNNs) in speech emotion recognition tasks. The benefit of mel-spectrograms is evident in our quantitative visualisation, even though spe•ctrograms are stillused in the literature. Going ahead, it is essential

REFERENCES

[1] Zielonka, M.; Piastowski, A.; Czyżewski, A.; Nadachowski, P.; Operlejn, M.; Kaczor, K. "Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets." Electronics 2022, 11, 3831.
[2] Abeer Ali Alnuaim, Mohammed Zakariah, Prashant Kumar Shukla, Aseel Alhadlaq, Wesam Atef Hatamleh, Hussam Tarazi, R. Sureshbabu, Rajnish Ratna, "Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier", Journal of Healthcare Engineering, vol. 2022, Article ID 6005446, 12 pages, 2022.
[3] Singh, A., Srivastava, K. K., & Murugan, H. "Speech Emotion Recognition Using Convolutional Neural Network (CNN)." International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020.
[4] Anvarjon, T.; Mustaqeem; Kwon, S. "Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features." Sensors 2020, 20, 5212.
[5] F. Andayani, L. B. Theng, M. T. Tsun and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," in IEEE Access, vol. 10, pp. 36018-36027, 2022.
[6] L. Yunxiang and Z. Kexin, "Design of Efficient Speech Emotion Recognition Based on Multi Task Learning," in IEEE Access, vol. 11, pp. 5528-5537, 2023.
[7] M. B. Er, "A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features," in IEEE Access, vol. 8, pp. 221640-221653, 2020.
[8] K. V. Krishna, N. Sainath and A. M. Posonia, "Speech Emotion Recognition using Machine Learning," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1014-1018.
[9] Eyben, F., Wöllmer, M., & Schuller, B. "Opensmile: the Munich versatile and fast open-source audio feature extractor." Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010.
[10] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... & Burkhardt, F. "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism." Proceedings INTERSPEECH, 2013.
[11] Goodfellow, I., Bengio, Y., & Courville, A. "Deep learning." MIT press, 2016.
[12] LeCun, Y., Bengio, Y., & Hinton, G. "Deep learning." Nature, 521(7553), 436-444, 2015.
[13] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. "Imagenet: A large-scale hierarchical image database." IEEE conference on computer vision and pattern recognition, 2009.
[14] He, K., Zhang, X., Ren, S., & Sun, J. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
[15] Simonyan, K., & Zisserman, A. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.
[16] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal Processing Magazine, 29(6), 82- 97, 2012.
[17] Graves, A., & Schmidhuber, J. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." Neural Networks, 18(5-6), 602-610, 2005.
[18] Kingma, D. P., & Ba, J. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.
[19] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. "TensorFlow: Large-scale machine learning on heterogeneous systems." Software available from tensorflow.org, 2015.
[20] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Desmaison, A. "PyTorch: An imperative style, high- performance deep learning library." Advances in Neural Information Processing Systems, 32, 8024-8035, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY