Speech emotion recognition using CNN.

Submitted in partial fulfilment of the requirements

of the degree of

(B.Tech)

by

**Arnav Mishra**

**Aditya Kumar Barawal**

Supervisor:

**Dr. V. Rama**

Assistant Professor

Department of Electronics and Communication Engineering



DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL 2022-2023

**Declaration**

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/ data/ fact/ source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

xxxx

xxxxx

xxxxx

Date: xxx

**CERTIFICATE**

This is to certify that the project entitled “**Speech emotion recognition using CNN**” has been submitted to the Department of Electronics and Communication Engineering, National Institute of Technology, Warangal during the academic year 2024-2025 for the fulfilment of the requirement for Bachelor of Technology degree in “Electrical and Electronics Engineering” by the following students of the B.Tech program.

**Student's name:**

Arnav Mishra (22EEB0A27)

Aditya Kumar Barawal (22EEB0A48)

**Dr. V. Rama**

**Asst. Professor**, Dept. of ECE, NITW

**(Project Guide)**

**ACKNOWLEDGEMENT**

We like to express our sincere gratitude to individuals and organizations. It would not have been possible without their kind support and help.

We would like to thank our Project guide Dr. V. Rama ma’am for providing her guidance and necessary information in partial completion of our project.

We like to express special gratitude to our Department of Electronics and Communication Engineering for giving such a wonderful opportunity to engage ourselves with this project, which is useful to build our theoretical knowledge with industry requirements.

**Contents**

[Abstract 7](#_Toc135033745)

[Chapter 1 – Introductn 8](#_Toc135033746)

[1.1 Overview of Speech Emotion Recognition 8](#_Toc135033747)

[1.2 Types of Emotions 9](#_Toc135033748)

[1.2.1 Basic Emotions 9](#_Toc135033749)

[1.2.2 Complex Emotions 10](#_Toc135033750)

[1.3 Importance of Emotion Detection in Speech 11](#_Toc135033751)

[1.4 Existing Methods of Speech Emotion Recognition and their Demerits 12](#_Toc135033752)

1.5 Motivation.......................................................................................................................13

[1.6 Problem Statement 13](#_Toc135033753)

[Chapter 2: Literature Survey 14](#_Toc135033754)

Chapter 3: Work Flow...............................................................................................................16

Chapter 4: Deep Learning.........................................................................................................17

4.1 Machine Learning............................................................................................................17

4.2 Machine Learning vs Deep Learning vs Neural Networks..............................................17

4.3 Convolutional Neural Network .......................................................................................19

4.4 Transfer Learning.............................................................................................................20

Chapter 5: Data Set, Framework and Loss Functions Used.....................................................24

[5.2 Model Used 24](#_Toc135033763)

[5.3 Loss Functions 27](#_Toc135033764)

[5.3.1 Sparse Categorical Cross Entropy 27](#_Toc135033765)

[5.3.2 Regularisation Losses (L1 and L2 Regularizations) 28](#_Toc135033766)

Chapter 6: Results.....................................................................................................................32

Chapter 7 : Conclusion……………………………………………………………………….34

Chapter 7: References............................................................................................................. **35**

# Abstract

There is great potential for emotion recognition from speech in a number of fields, such as customer service, mental health assessment, and human-computer interaction. Robust emotion recognition from speech has the potential to improve user experiences and make automated systems more sensitive and compassionate.

This study uses the RAVDESS dataset to propose a 1-D Convolutional Neural Network (CNN) based model for emotion recognition from voice. To categorize speech into distinct emotion groups, the model captures and processes audio data such as Chroma, Mel Frequency Cepstral Coefficients (MFCC), and Mel Spectrogram Frequency.

The primary goal of this effort is to increase the precision and dependability of emotion identification systems by utilizing developments in deep learning and audio processing. Machine-human interactions can be revolutionized by being able to recognize and respond to human emotions, which will make them more efficient and natural.

Important techniques are feature extraction (MFCC, Chroma, and Mel Spectrogram), model training, and data augmentation (adding noise and adjusting the audio stream). The augmented dataset is used to train a CNN with several convolutional layers, max pooling layers, and dense layers using the Adam optimizer and sparse categorical cross-entropy loss function along side regularization functions to prevent over-fitting of the model on the test data . Metrics like accuracy, precision, recall, and F1-score are used to assess the model on the test set as well as the entire dataset. And the satisfactory results were obtained with the accuracy being 89.6296% and the F1 score as 0.896529 . The precision and recall being 90.07% and 0.8963 respectively .

Training and assessment are conducted using the RAVDESS dataset. Librosa for audio processing, TensorFlow and Keras for CNN model construction and training, Scikit-learn for data preprocessing and assessment, and Matplotlib and Seaborn for result visualization are some of the tools and libraries available.

The literature review highlights notable studies in the field, including:

• Wang, An, and Li (2020): Achieved 73.74% accuracy with SVM on the EMODB dataset using wavelet packet techniques.

• Roy et al. (2019): Achieved accuracies of 73.67% with SVM, 77.71% with Gaussian Naive Bayes (GNB), and 69.41% with K-Nearest Neighbour (KNN) on the RAVDESS dataset using a feature set incorporating Discrete Wavelet Transform (DWT).

• Van Zwol et al. (2018): Achieved 71% accuracy on the EMODB dataset using Deep Convolutional Neural Networks (DCNN) enhanced with Fast Continuous Wavelet Transform (fCWT).

• Danai (2021): Demonstrated the effectiveness of using multi-modal datasets like IEMOCAP for emotion recognition in an affective service but no knowledge about the results was provided.

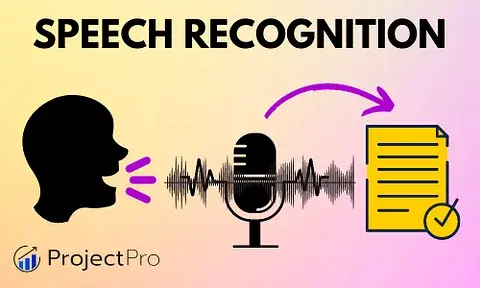
This work presents a robust method for speech emotion recognition using a CNN model. The combination of data augmentation, feature extraction, and deep learning techniques resulted in a highly accurate emotion recognition system, demonstrating potential practical applications in various domains.

# Chapter 1 – Introduction

## 1.1 Overview of Speech Emotion Recognition

Speech Emotion Recognition (SER) is a field that focuses on identifying and classifying emotions from spoken language, leveraging the interplay between computer science, signal processing, and psychology. The ability to detect emotions in speech is essential for enhancing human-computer interaction, improving customer service, and supporting mental health monitoring.

Convolutional Neural Networks (CNNs) present a viable substitute by automatically identifying characteristics from unprocessed input. CNNs can be used in SER to create spectrograms, which are graphic depictions of speech frequency spectrum changes over time. This enhances the accuracy and resilience of emotion detection by enabling CNNs to capture the temporal and spectral patterns linked to various emotions.  
  
CNNs are used in SER because of their higher performance over conventional approaches, capacity to handle big datasets, and ability to automatically extract pertinent features. Notwithstanding obstacles such as fluctuations in speech patterns and ambient noise, CNN-based speech recognition systems exhibit considerable promise for creating dependable emotion identification applications in practical settings.



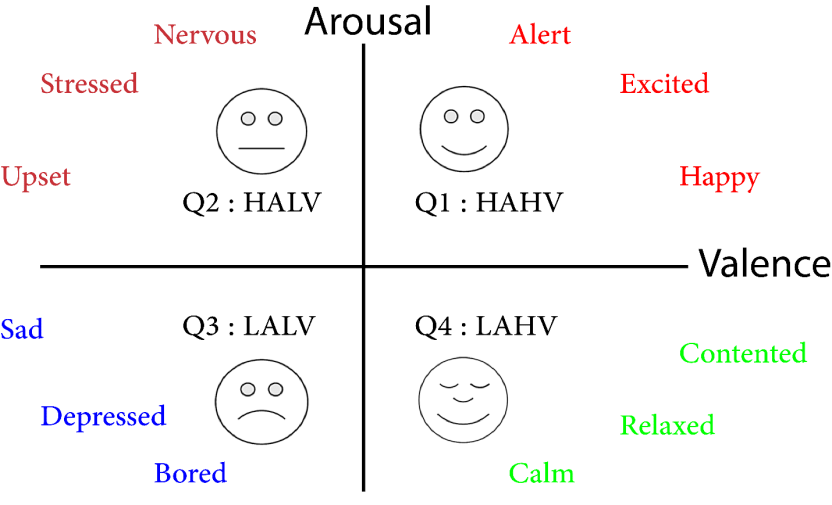
## 1.2 Types of Emotions

Basic emotions and Complex emotions are the two primary categories into which spoken emotions can be generally divided. It is essential to comprehend these categories in order to create Speech Emotion Recognition (SER) systems that work well.

### 1.2.1 Basic Emotions

Basic emotions are universal and are typically recognized easily due to their distinct acoustic features. These emotions are fundamental and are considered to be biologically hardwired in all humans. The primary basic emotions include:

* **Happiness:** Characterized by a higher pitch, increased energy, and a faster speech rate. Happy speech often has a melodious and rhythmic quality.
* **Sadness:** Marked by a lower pitch, reduced energy, and a slower speech rate. Sad speech tends to be softer and more monotonic.
* **Anger:** Exhibits a higher pitch, increased energy, and a faster, more abrupt speech rate. Angry speech is often loud and sharp.
* **Fear:** Features a higher pitch, variable energy, and a faster speech rate. Fearful speech can sound tense and breathy.
* **Surprise:** Involves sudden changes in pitch and energy, often with a rapid speech rate. Surprised speech can be marked by an abrupt onset.
* **Disgust:** Characterized by a lower pitch and a slower, more uneven speech rate. Disgusted speech can sound harsh and guttural.



### 1.2.2 Complex Emotions

Complex emotions are combinations or variations of basic emotions and are more nuanced and context-dependent. These emotions can be harder to recognize due to their subtle and multifaceted nature. Examples of complex emotions include:

* **Frustration:** A mix of anger and sadness, often with a tense and strained speech pattern.
* **Excitement:** A blend of happiness and surprise, marked by a high pitch, rapid speech rate, and heightened energy.
* **Sarcasm:** Involves a contradictory tone where the speaker's intent is opposite to the literal meaning of the words, often requiring contextual understanding for accurate recognition.
* **Embarrassment:** Combines elements of fear and shame, with hesitations, pauses, and a quavering voice.

## 1.3 Importance of Emotion detection in Speech

Speech emotion recognition is important for many applications, including increasing human-machine interaction, boosting mental health programs, and providing better services. The following are some crucial domains where speech emotion detection is very crucial:

* **Enhanced Human-Computer Interaction**

Virtual assistants, chatbots, and other interactive systems can comprehend and react to users' emotional states thanks to emotion detection, which improves the naturalness and empathy of interactions. For instance, a virtual assistant can increase user satisfaction and engagement by providing more help or a more encouraging response if it hears annoyance in the user's voice.

* **Mental Health Monitoring**

Identifying emotions in speech might be a useful method for keeping an eye on mental health. Through analysis of alterations in speech patterns, it can assist in identifying symptoms of emotional discomfort, depression, or anxiety in patients. This non-invasive technique offers ongoing, real-time monitoring, supports patients in professional settings, and helps with early mental health issue intervention.

* **Automotive Industry**

By keeping an eye on drivers' emotional states, emotion detection in the automotive sector can improve driving safety. Drivers who exhibit indicators of stress or weariness can be alerted by systems, which could potentially prevent accidents caused by driving while intoxicated. This technology enhances overall road safety and aids in the development of advanced driver assistance systems (ADAS).

* **Social Robotics**

Emotion detection in social robotics allows robots to recognize and react to human emotional cues, hence improving human-robot interaction. This feature is especially useful in environments such as senior care, where robots can enhance people's quality of life by offering support and companionship.

## 1.4 Existing methods of speech emotion recognition and its demerits

Existing methods of speech emotion recognition (SER) generally fall into several categories, each with its own merits and demerits:

* **Feature-Based Methods:**
* **Merits:** These methods extract acoustic features such as pitch, intensity, and spectral features from speech signals. They are computationally efficient and can achieve moderate accuracy in emotion classification tasks.
* **Demerits:** They may struggle with capturing complex emotional nuances that are not well-defined by traditional acoustic features. They also often require careful feature selection and preprocessing to be effective across diverse datasets and languages.
* **Machine Learning Approaches:**
* **Merits:** Machine learning techniques, such as Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and Neural Networks, have been widely used in SER. They can learn complex patterns in speech data and achieve high accuracy with appropriate training.
* **Demerits:** These methods heavily rely on the quality and quantity of labeled training data. They may overfit to specific datasets or struggle with generalization to new speakers or emotional contexts if the training data is not sufficiently diverse. Additionally, they can be computationally intensive, especially deep learning models, requiring substantial computational resources for training and inference.
* **Deep Learning Models:**
* **Merits:** Deep learning, particularly Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and their variants (e.g., LSTM, GRU), have shown state-of-the-art performance in SER. They can automatically learn hierarchical representations of speech features and capture temporal dependencies.
* **Demerits:** Deep learning models often require large amounts of labeled data for training and can be prone to overfitting without proper regularization techniques. They are also computationally expensive both in terms of training time and inference, limiting their deployment in real-time applications without optimized hardware or architectures.
* **Challenges and Limitations:**
* **Lack of Standardization:** There is a lack of standardized datasets and evaluation metrics across SER studies, making it difficult to compare results from different approaches.
* **Cross-Cultural Variability:** Emotional expression varies across cultures and languages, posing challenges for models trained on specific datasets to generalize universally.
* **Contextual Ambiguity:** Speech can exhibit ambiguity in emotional expression, where the same acoustic cues may convey different emotions depending on context or speaker variability.
* **Real-World Deployment:** Despite advances, real-world deployment of SER systems faces challenges such as robustness to environmental noise, speaker variability, and the need for continuous adaptation to evolving emotional contexts.

## 1.4 Motivation

Speech emotion recognition (SER) systems now in use range from sophisticated deep learning models to conventional feature-based approaches. However, these approaches frequently encounter difficulties that restrict their accuracy and practical use, such as the incapacity to capture nuanced emotional nuances, reliance on large amounts of labeled data, and computational complexity. Traditional feature-based techniques, for example, could have trouble processing complex emotional expressions, while deep learning models might overfit if they don't have strong regularization and need a lot of processing power. Accurate detection is further complicated by the diversity of emotional expression across cultures and settings.

## 1.5 Problem Statement

The purpose of this work is to use convolutional neural networks (CNNs) to recognize spoken emotions more accurately and efficiently. The goal is to create a reliable CNN-based model that can precisely identify emotions from voice signals by investigating different data augmentation methods and optimizing loss functions using RAVDESS Dataset.

# Chapter 2: Literature Survey

The work done by various researchers in the classification of emotions and the methodologies they employed are surveyed in this chapter. The possibility of improvement in work and remarks has been stated. A brief description of some of the methodologies which are used by these researchers is also stated.

The description of various literature referred is as follows:

* The application of a Deep Convolutional Neural Network (DCNN) architecture based on the AlexNet model for Speech Emotion Recognition (SER) was investigated by Bjorn E. Van Zwol and colleagues. To improve its suitability for emotion recognition tasks, the AlexNet model's last layer was adjusted to match emotion labels. The Fast Continuous Wavelet Transform (fCWT) was employed by the researchers to enhance the DCNN's functionality and facilitate enhanced feature extraction from voice signals. Using the EMODB dataset, the study's maximum categorization accuracy was 71%. The lack of data, however, was a major drawback that might have caused instability in the deep learning models that the fCWT induced. This volatility affects the regularity and dependability of the findings, emphasizing the necessity for more thorough and diverse datasets in future research.
* Wang Kunxia, An Ning, and Lian Li concentrated their research on improving wavelet packet approaches for speech emotion identification. To extract the subtle emotional cues from voice sounds, their theoretical approach makes use of wavelet packet coefficient models. Through the application of these cutting-edge methods, the study used Support Vector Machine (SVM) on the EMODB dataset to reach the maximum classification accuracy of 73.74%. The study was limited by its short-term investigation, which might not have adequately captured the non-stationary characteristics of voice transmissions, even with these encouraging results. This implies that although the wavelet packet approaches work well, they might need to be improved upon in order to handle speech's dynamic nature over longer time spans and in a wider range of situations.
* Satyakama, Roy Tanmoy, Marwala Tshilidzi, and Chakraverty Snehashish Paul developed a unique method for identifying emotions in speech by using the Discrete Wavelet Transform (DWT) to decode speech signals and examine emotional discrepancies. Three main classification methods were used in the study: Gaussian Naive Bayes (GNB), K-Nearest Neighbor (KNN), and Support Vector Classification (SVC) with Radial Basis Function (RBF) Kernel. Using the RAVDESS dataset, their studies produced classification accuracies of 69.41% for KNN, 77.71% for GNB, and 73.67% for SVM. This study's significant flaw is its reliance on a single dataset, which could limit how broadly the results can be applied. Testing these techniques on various datasets could help future research by confirming their resilience and suitability for real-world situations.
* Styliani Danai's uses Speech Emotion Recognition techniques in conjunction with Multi-Modal Emotion Recognition. The study's methodology trains and evaluates models for emotion identification tasks using the IEMOCAP dataset, which is well-known for its Interactive Emotional Dyadic Motion Capture Database. Although the research shows that IEMOCAP can be effectively used for both model creation and evaluation, it does not go into great detail about the particular difficulties or limitations that come with integrating speech emotion recognition with multi-modal emotion recognition. Future research into the difficulties and practical constraints of incorporating these technologies into practical applications is left open by this approach.

# Chapter 3: Work Flow

#### 1. ****Introduction****

* **Objective**: The objective of this research is to develop a machine learning model capable of recognizing emotions from audio signals using a 1D Convolutional Neural Network (CNN).
* **Dataset**: The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset is used, which contains audio files with different emotional expressions.

#### 2. ****Data Preprocessing****

* **Loading Data**:
  + Load audio files from the dataset directory.
  + Each file is associated with a specific emotion based on its filename.
* **Feature Extraction**:
  + **MFCC (Mel Frequency Cepstral Coefficients)**: Capture the power spectrum of the audio signal.
  + **Chroma Features**: Represent the 12 different pitch classes.
  + **Mel Spectrogram**: Shows the power of the spectrum of frequencies.
* **Data Augmentation**:
  + **Adding Noise**: Random white noise is added to the audio to increase diversity.
  + **Shifting**: The audio signal is shifted in time to create variations.
* **Storing Features**:
  + Extracted features are stored in arrays and optionally saved to disk.

#### 3. ****Data Preparation****

* **Label Encoding**:
  + Emotion labels are converted to integers using Label Encoder.
* **Train-Test Split**:
  + Split the dataset into training and testing sets (75% training, 25% testing).

#### 4. ****Model Architecture****

* **Convolutional Layers**
* **Pooling Layer**
* **Flatten Layer**
* **Fully Connected Layer**

5. **Model Compilation**

* **Optimizer**:
  + Adam optimizer is used with a learning rate of 0.001.
* **Loss Function**:
  + Sparse Categorical Crossentropy is used as the primary loss function.
* **Metrics**:
  + Accuracy is used as the evaluation metric.

#### 6. ****Model Training****

* **Training**:
  + The model is trained for 100 epochs with a batch size of 64.
  + The training process involves fitting the model on the training data and validating it on the test data.
* **Plotting Loss**:
  + The loss and validation loss are plotted to visualize the training process.

#### 7. ****Model Evaluation****

* **Confusion Matrix**:
  + A confusion matrix is generated to evaluate the performance of the model in classifying each emotion.
  + A heatmap is plotted to visualize the confusion matrix.
* **Performance Metrics**:
  + **F1 Score**: Weighted F1 score is calculated to measure the balance between precision and recall.
  + **Accuracy**: The accuracy of the model on the test set is evaluated.
  + **Precision and Recall**: Precision and recall are calculated to further assess the model's performance.

#### 8. ****Prediction on Single Audio File****

* **Loading Single Audio File**:
  + A single audio file is loaded and features are extracted.
* **Prediction**:
  + The pre-trained model is used to predict the emotion of the audio file.
  + The predicted class is printed.

#### 9. ****Combining and Evaluating Full Dataset****

* **Combining Training and Test Sets**:
  + The training and test sets are combined for a comprehensive evaluation.
* **Evaluating on Full Dataset**:
  + The model is evaluated on the full dataset.
  + Performance metrics such as loss, accuracy, precision, recall, and F1-score are calculated.

# Chapter 4: Deep Learning

## 4.1 Machine Learning

Machine learning is a subfield of artificial intelligence (AI) and computer science that focuses on using data and algorithms to simulate how human beings learn, gradually increasing the accuracy of the system.

The rapidly expanding discipline of data science includes machine learning as a key element. Algorithms are trained to make classifications or predictions using statistical techniques, revealing important insights in data mining projects. The decisions made as a result of these insights influence key growth metrics in applications and businesses, ideally.

## 4.2 Machine Learning vs Deep Learning vs Neural Networks

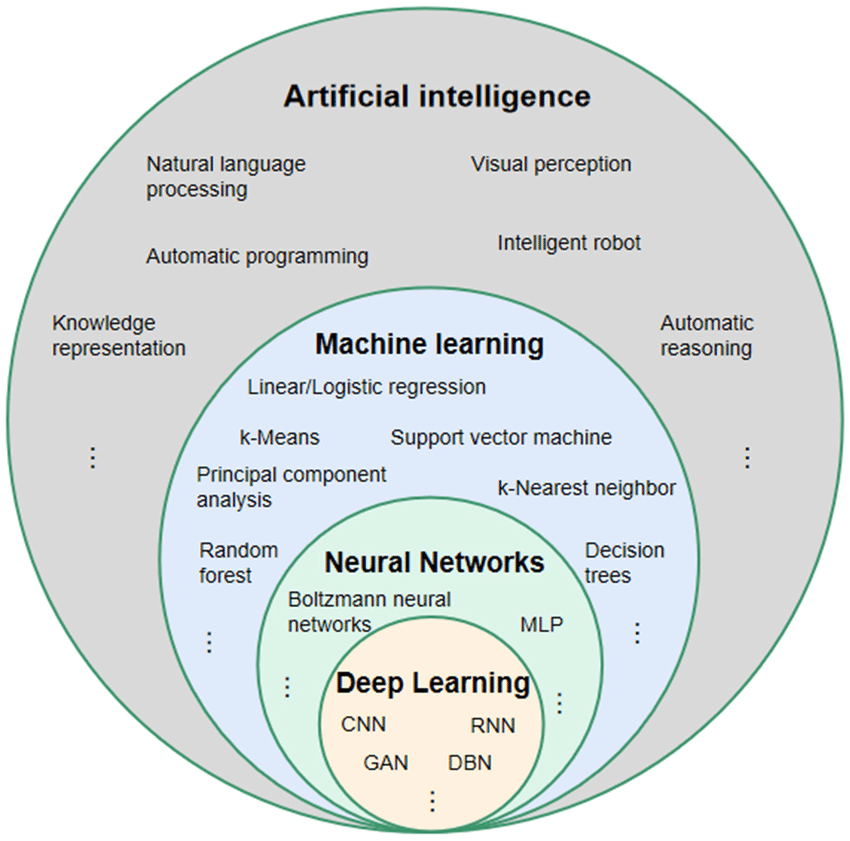
****

Figure 10 Relationship between AI, ML, NN and DL [13]

Given that deep learning and machine learning are frequently used synonymously, it is important to understand their differences. Neural networks, deep learning, and machine learning are all branches of artificial intelligence.

The way each algorithm learns is where deep learning and machine learning diverge. Deep learning significantly reduces the amount of manual human intervention required during the feature extraction phase of the process, allowing for the use of larger data sets.

Neural networks are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, is connected to others and has a weight and threshold that go along with it. Any node whose output exceeds the defined threshold value is activated and begins sending data to the network's uppermost layer. Otherwise, no data is transmitted to the network's next layer. The depth of layers in a neural network is all that is meant by the term "deep learning." A deep learning algorithm or deep neural network is defined as a neural network with more than three layers, inclusive of the inputs and outputs.

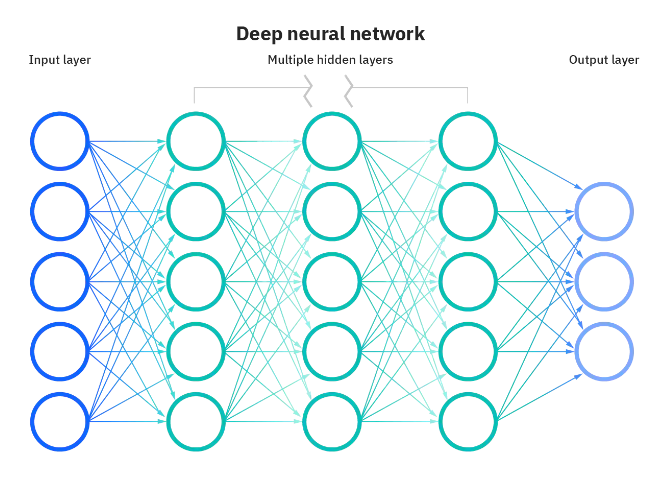


Figure 9 Deep Neural Network [14]

## 4.3 Convolutional Neural Networks

**What is CNN?**

Convolutional neural networks (CNNs) are one kind of neural network that are frequently adopted in deep learning. As they can automatically learn the spatial hierarchies and spatial relationships between pixels in an image, CNNs are especially well suited for image identification and classification applications.

**Layers in CNN**

Convolutional neural networks (CNNs) are a particular kind of neural network that include different layers, each of which processes the input data in a particular way.

A CNN has various kinds of layers, including:

Using a series of trainable filters, the input data is processed in the **convolutional layer**, which is the fundamental component of a CNN, to extract key characteristics.

**Pooling layer** is to reduce the computational complexity and to avoid overfitting, this layer down samples the input data by summing up the data in small patches.

The **activation layer** adds non-linearity to the network by applying a non-linear activation function to the output of the convolutional or pooling layer.

**Fully Connected Layer**, this layer connects every neuron in the past layer to every neuron in the upcoming layer, which enables the network to learn intricate connections between the input data and the output labels.

The network's performance and stability may be enhanced by the **normalisation layer**, which performs normalisation on the activations of the preceding layer.

Together, these various layers process the input data to generate a prediction or classification. The task and the data that a CNN is applied to will determine the specific architecture of a CNN, including the number and kind of layers.

## 4.4 Transfer Learning

A model that has been trained for one task is repurposed for a different or a related task using the machine learning technique known as transfer learning. The model can either be adjusted for the new task or used as a feature extractor to achieve this.

**Fine-tuning**

A model can be "fine-tuned" by changing its parameters to perform better on a new job. To accomplish this, a small collection of instances from the newly created task is used to train the model. The degree to which the new task resembles the task that the model was first trained on determines the amount of fine-tuning that is required.

**Extraction of features**

The technique of extracting features from a dataset is known as feature extraction. The qualities of the data that are pertinent to the task at hand are referred to as features. Once the features have been recovered, they can be utilised to generate predictions for fresh data or to train a new model.

In transfer learning, the following equations are utilised:

* The difference between the predicted labels and the actual labels is calculated using the loss function.
* In order to minimise the loss function, the optimizer is utilised to update the model's parameters.
* The model's performance on a test dataset is gauged by accuracy.

Compared to building a model from scratch, transfer learning provides the following many benefits like saving time and resources as you do not need to gather a sizable dataset. As the model can start with a good set of parameters that have already been trained on a related task, it can enhance performance. The model may transfer knowledge from a comparable activity with more data, it can be utilised to address problems with little data.

The effectiveness of machine learning models can be increased by using the potent technique of transfer learning. Since it may transfer knowledge from related jobs with more data, it is especially helpful for tasks with less data.

# Chapter 5: Data Set, Framework and Loss Functions Used.

## 5.1 Data Set

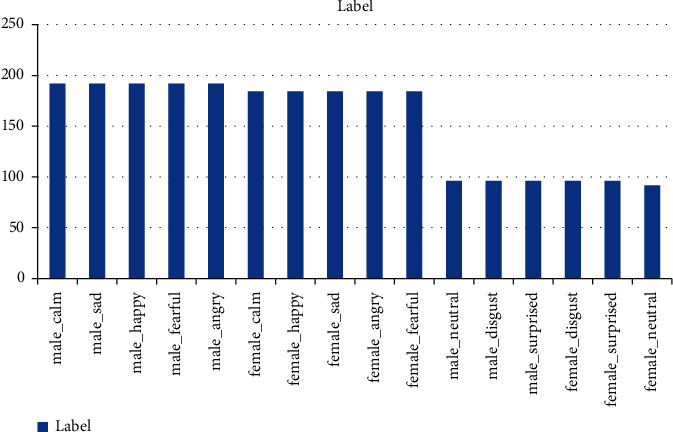
For this machine learning program, the RAVDESS data-set has been incorporated. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7,356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound). Note, there are no song files for Actor\_18.

The High Quality of the RAVDESS dataset are one of the major advantages of using it as it is essential for accurate analysis for training the model. It is widely used as it includes a wide range of emotions(neutral, calm, happy, sad, angry, fearful, disgust, surprised) expressed through speech and song . In-short it is quite a versatile and a balanced dataset .

Its free accessibility and standardization has enabled researchers and students compare their result more easily with others , hence fostering a sense of progress in the field of emotion recognition.

The entire dataset is 24.8GB from 24 actors, but for our project we’ve lowered the sample rate on all the files. In our model, we are taking the emotion (third identifier) in consideration only. The 7356 RAVDESS files are all identified by a different filename. A seven-part number identification makes up the filename (02-01-06-01-02-01-12.mp4).

`



Distribution of data RAVDESS dataset[7]––

## 5.2 Model Used

For our Project we have incorporated a basic 1-D CNN network. The choice of the model being used was this kind of model is designed to work wit sequential or time-series data like audio signals.

CNNs were basically designed to tackle 2-D data(like image processing) , the same principle was extended to 1-D data like sequences and audio signals. CNN was introduced in the early 1980s by Yann LeCun to classify visual patterns which then evolved into LeNet-5 specified for image recognition. In the early 2000s researchers began to adapt CNN beyond the task of image recognition . Its principles were adopted to 1-D data processing. Audio and speech processing was one of the initial fields in which 1D CNNs were used. CNNs can automatically learn feature representations, which makes them useful for applications like speaker identification, emotion recognition from audio data, and speech recognition.

Architecture:

1. Input Shape: The model desires the shape of input to be ‘(180,1)’. This shape represents 180 features(extracted from audio signal) and 1 channel. The data is reshaped to match this input size.
2. First Convolutional Layer: Extracts low level features.

* **Conv1D Layer**: Applies 256 filters, each of size 5, to the input data. The padding value ensures that the output has the same length as the input by padding the input with zeros wherever needed
* **ReLU Activation**: This is the activation function that is incorporated in our project . It introduces non-linearity into the model, allowing it to learn more complex patterns.

1. Second Convolutional Layer: Further processing with regularization.

* Again the 1D layer is applied but this time with 128 filters each of size 5 , with this 2 regularizers L1 and L2 are also added.
* Followed by ReLU activation function.

1. Max Pooling Layer:

* This layer serves to reduce the dimensionality of the feature maps by taking the maximum value over a window size 8. This process helps to reduce the computational complexities and helps heightening the important features.

1. Third Convolutional Layer:

* 1D layer is applied with 128 filters each of size 5 , with this 2 regularizers L1 and L2 are also added.
* ReLU activation to introduce non linearity.

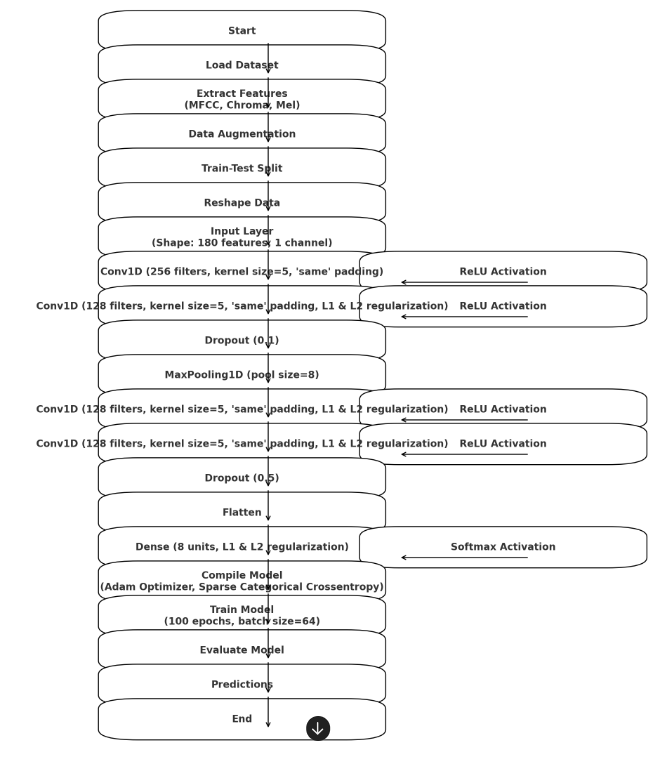
1. Fourth Convolutional Layer:

* 1D layer is applied with 128 filters each of size 5 , with this 2 regularizers L1 and L2 are also added.
* ReLU activation to introduce non linearity.
* Dropout of 50% is added to prevent overfitting of the model during training process.

1. Flattening Layer : This layer serves to convert 3D output from the convolutional layers into a 1D feature vector that is further fed into the dense/fully connected layers.
2. Fully connected Layer:

* Dense Layer: We have used fully connected layers with 8 units , corresponding to the number of emotion classes . further regularizations is applied to prevent overfitting.
* SoftMax Activation : It is used to convert the output to a probability distribution over the 8 classes , with a combined total sum being equal to 1.

The Next page presents with a flowchart that describes the flow of our model working.

**5.3 Loss Functions**

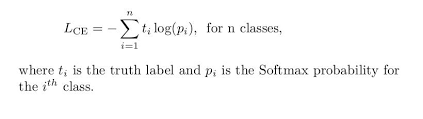
The main loss function used is Sparse Categorical Cross Entropy function along with L1 and L2 added penalty loss functions.

### 

### 5.3.1 Sparse Categorical Cross entropy:

Sparse categorical Crossentropy is used in case where the target labels are integers(sparse) instead of one-hot encoded vectors. Basically this used in multiclass classification models.

The Sparse Categorical Cross Entropy Loss is defined as follows [13]:



This can help improve the performance of models on imbalanced datasets by giving more weight to underrepresented classes and less weight to overrepresented classes. The loss function measures the difference between the true labels and the predicted labels by the model. Sparse categorical cross entropy is specifically useful in multi-class classification problems where the model outputs a probability distribution over multiple class.

**5.3.2 Regularization Losses(L1 and L2 Regularizations)**

By incorporating a penalty into the loss function that is determined by the size of the model coefficients, regularisation losses are employed to stop overfitting. These are described in our model's dense layer and convolutional layers

**L1 Regularization** (Lasso) :

* Adds a penalty equal to the absolute value of the magnitude of coefficients.
* Encourages sparsity, meaning it can drive some coefficients to zero.
* **Formula**: λ | , here λ is a regularization parameter and Wi are the model weights.

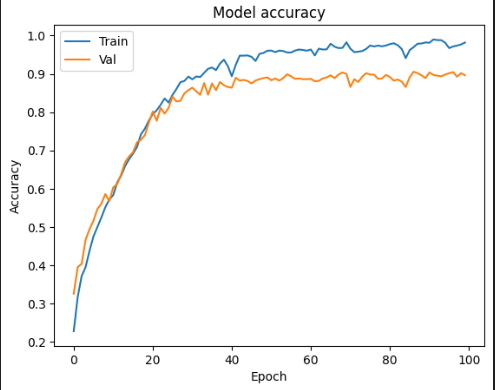
**L2 Regularization(Ridge) :**

* Adds a penalty equal to the square of magnitude of coefficients.
* Improves smaller and more evenly distributed coefficients.
* Formula λ
* By adding these penalties to the loss function we can ensure that the model can not fit the training data too close , hence in better for overall generalization of the model to account for the unseen data. The Adam optimizer was used to reduce the loss function.
* Total Loss = Sparse Categorical Cross entropy Loss + Regularization losses
* These combined loss functions help in effectively training your model while maintaining its ability to generalize well to new, unseen data.

**Chapter 6: Results**

The model was trained for a total of 100 epochs and the following are the Results obtained against the training and validation data…

1. Accuracy:

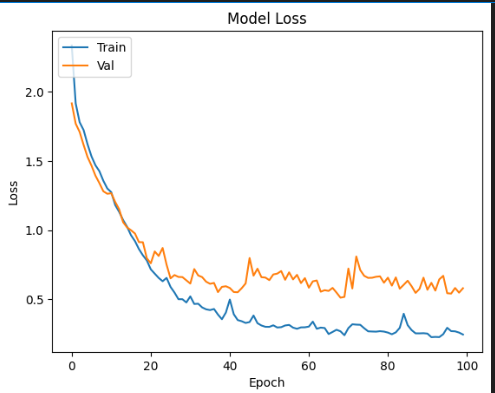


* The total test Accuracy obtained of the training data was close to

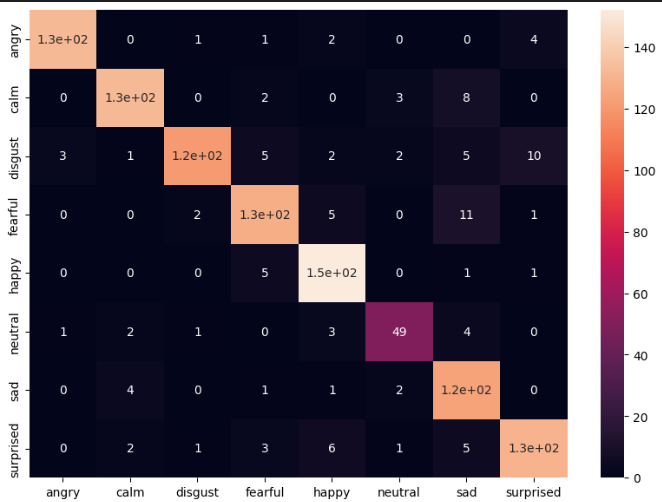
**89.6296 %** , which was better than any of the results obtained.

* The precision obtained was **90.07 % .**
* Recall score obtained was **0.8963.**
* F1 Score obtained was close to **0.8965**

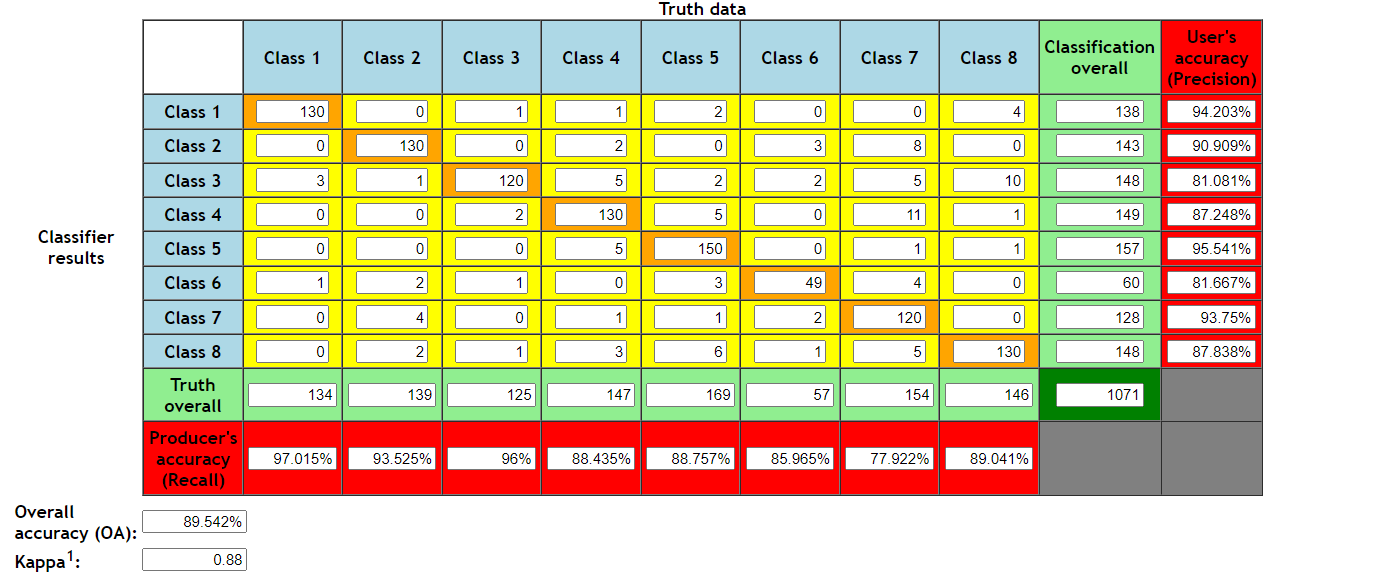
1. Loss : test loss obtained : **58.06%**



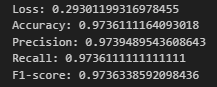
* Confusion Matrix:



* **Accuracy Result obtained by means of confusion matrix:**



* **The final values obtained upon combining the training and test data:**



* **Conclusion:**

Upon concluding our project we were successfully able to get the desired results and improve the accuracy of the model for speech emotion recognition. By means of this research we can summarise that 1D CNN networks are an efficient machine learning model to incorporate speech analysis for emotion detection.

The combination of advanced feature extraction, data augmentation, and a well-designed CNN architecture resulted in a model that performs admirably in classifying emotions. This work lays the foundation for further advancements and applications in emotion recognition technology, paving the way for more empathetic and responsive machines that can better understand and interact with human emotions.

While our current model shows promising results, future research can explore other deep learning architectures, integrate more advanced data augmentation techniques, and leverage larger and more diverse datasets to further enhance performance and robustness and improve and incorporate real time accessibility.

# 

# Chapter 7: References

|  |  |
| --- | --- |
| [1] | S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database  of emotional speech and song (ravdess),” Public Library of Science,  vol. 13, no. 5, pp. 1–35, 05 2018. |
| [2] | J. Lee and I. Tashev, “High-level feature representation using recurrent  neural network for speech emotion recognition,” in Proc. INTER  SPEECH, 2015 |
| [3] | H. M. Fayek, M. Lech, and L. Cavedonb, “Evaluating deep learning  architectures for speech emotion recognition,” Neural Networks, vol. 92,  pp. 60–68, 2017. |
| [4] | F. Albano, T. Longobradi, N. M. Maldonato, E. Moretto, S.D. Orco,  A. D. Sarno, R. Sperandeo, Y. Tedesco “Co-regulation of the voice  between patient and therapist in psychotherapy: machine learning for  enhancing the synchronization of the experience of anger emotion”  9th IEEE International Conference on Cognitive Infocommunications,  2018 |
| [5] | Akcay M. B., Oguz K., “Speech emotion recognition: Emotional  models, databases, features, preprocessing methods, supporting  modalities, and classifiers”. Speech Communication, Volume 116,  2020 |
| [6] | M. Sajjad et al., “Clustering-based speech emotion recognition by in  corporating learned features and deep BILSTM,” IEEE Access, vol. 8,  pp. 79861–79875, 2020. |
| [7] | https://www.researchgate.net/figure/Arousal-Valence-Plane-for-Emotions-With-Some-Basic-Emotions-on-it-The-horizontal-and\_fig1\_334034255 |
| [8] | https://www.projectpro.io/article/speech-emotion-recognition-project-using-machine-learning/573 |
| [9] | The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5955500/ |

[10] https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network

[11] https://www.tensorflow.org/api\_docs/python/tf/keras/losses/SparseCategoricalCrossentropy

[12] https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization#:~:text=The%20differences%20between%20L1%20and,regularization%20solution%20is%20non%2Dsparse.