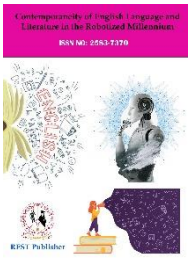
Contemporaneity of English Language and Literature in the Robotized Millennium
Vol: 3(1), 2024
REST Publisher; ISSN: 2583 7370
Website: https://restpublisher.com/journals/cellrm/
DOI: https://doi.org/10.46632/cellrm/3/1/3



Optimized Multimodal Emotional Recognition Using Long Short-Term Memory

J. Mary Hanna Priyadharshini, *Anish Kumar M, Praveen J A, Vairavandinesh G
Veltech High-tech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Chennai, Tamil Nadu, India.
*Corresponding author Email ID: vh12232_aiml22@velhightech.com

Abstract: *The aim of this project is to research and classification on human emotions. A new method for the recognition of speech signals has been introduced. It's called LSTM (Long-Short Term Memory). It is a type of Recurrent neural network. RNN is used for analyzing sequential data, hence it is useful for speech signal recognition. Several Datasets were found across the internet for this project. Ex: TESS (Toronto Emotional Speech Set), RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), SAVEE (Surrey Audio-Visual Expressed Emotion), CREMA-D (Crowd- Sourced Emotional Multimodal Actors Dataset). The Main Dataset used in this project is TESS (Toronto Emotional Speech Set) Dataset and Mel Frequency Cepstral Coefficient (MFCC) is Used for Feature extraction.*
Keywords: *Long Short-Term Memory (LSTM), MFCC (Mel Frequency Cepstral Coefficient), RAVDESS, CREMA-D, TESS, SAVEE.*

1. INTRODUCTION

Speech is the main and direct means of transmitting information. It contains a wide variety of information, and it can express rich emotional information through the emotions it contains and visualize it in response to objects, scenes or events. The automatic recognition of emotions by analyzing the human voice and facial expressions has become the subject of numerous research and studies in recent years [1-5]. The fact that automatic emotion recognition systems can be used for different purposes in many areas has led to a significant increase in the number of studies on the subject. Speech is one of the Important aspects of human communication. Based on the emotion in the speech of a person we can find his psychological state. Mel frequency Cepstral coefficient is used for feature extraction. Support vector machines are classification models used for classification and labeling of different emotions. let's discuss about MFCC, several aspects of MFCC are, Framing, Windowing, Fast Fourier Transform, Mel Frequency conversion and cepstral. The sound signals are divided into Frames of Certain Time intervals. the Framing separated the speech Signal in 30 – 40 milliseconds. The Accuracy Came above 90 %.

2. RELATED WORKS

The study compares deep learning and conventional machine learning techniques for Speech Emotion Recognition (SER) in human-computer interactions, aiming to provide a comprehensive understanding of the open-ended problem and provide a multi-aspect comparison of practical neural network approaches in SER [1]. Speech emotion recognition method using the least square regression (LSR) model, with an incomplete sparse LSR (ISLSR) model. The model uses both labeled and unlabeled speech data sets for training and is capable of feature selection. Experiments on two emotional speech databases show improved recognition performance [2]. The study proposes a new Fourier parameter model for speech emotion recognition, focusing on harmony features and speaker-independent differences [3]. The study uses machine learning algorithms to recognize emotions in Arabic, using YouTube videos and analyzing spectral features. SVM achieves the highest accuracy of 77.14% [4]. The F-Emotion algorithm enhances Speech Emotion Recognition accuracy by extracting speech emotion features and creating a parallel deep learning model for different emotion types. It effectively analyzes correspondence between emotion features and emotion types, focusing on

neutrality, happiness, fear, and surprise [5]. Deep neural networks excel in speech emotion recognition within a single corpus, but data-driven methods degrade in cross-corpus scenarios. Cross-corpus method using few-shot learning and unsupervised domain adaptation outperforms other approaches [6].

3. DATASET USED

The Dataset used for training is Toronto Emotional Speech set. This dataset contains several lines from different voice actors. It has four different audio labels. Anger, Fear, Happiness, Sadness. It has a total of 2800 audio file. The Voice actors were between the age of 26 and 64. The data is already cleaned, hence no need for Data cleaning. The Toronto Emotional Speech set serves as a rich and various reserve for preparation models in the field of affection acknowledgment and talk alter. With an inclusive group of 2800 audio files, the dataset captures an of-course range of sensitive verbalizations, containing Anger, Fear, Happiness, and Sadness. The addition of multiple voice stars guarantees instability in pitch, volume, and verbalization, offering a realistic account of life experience of sensitive talk across various things. The age range of the voice stars, spanning from 26 to 64 age traditional, further provides to the dataset's variety, taking everything in mind potential age-related shadings in exciting verbalization.

The dataset's cleanness removes the need for additional dossier cleansing processes, conditional valuable opportunity and possessions all the while the model development step. This preprocessed state improves the dataset's dependability and reduces the possibility of presenting biases or errors into the model all the while preparation. Researchers and experts in the fields of talk and empathy analysis can influence the Toronto Emotional Speech fight expand and tweak models that correctly understand; and classify passion in human language. Additionally, the chance of different exciting labels allows for point or direct at a goal preparation and judgment of models established distinguishing heated categories, easing more nuanced and exact reasoning.

The Toronto Emotional Speech set determines a particularized glimpse into the nuances of touching verbalization, contribution scientists and developers a nuanced understanding of the audile appearance guide different concerns. The dataset's inclusive character allows the exploration of differences in talk patterns, versification, and inflection across diverse heated states. The addition of diversified voice actors not only increases instability in conditions adult but also produces outward various cultural and semantic influences, providing to the dataset's cross-educational relevance. This diversity is specifically valuable in evolving models that can statement well across differing.

4. LIBRARIES

This project is done in python programming language. The necessary libraries are osmium, librosa, imported modules from sklearn for splitting the dataset for training and testing. label Encoder is imported from sklearn for label Encoding Keras is an artificial programming Interface for neural networks. It is suitable for many recurrent neutral models such as LSTM (Long-Short Term Memory).

For This Project the Sequential Keras model is imported. Predict function is used to predict the emotion of new data. MFCC splits the audio signals into frames of Few milliseconds (30-40 milliseconds approximately). The steps involved in Feature Extraction using MFCC are Framing, Windowing, Model building, network Definition, network Compilation and prediction, training and evaluation. The LSTM model used here is sequential model, on Keras Package. Long-Short Term Memory has several layers .It also contains Memory cells. This is followed by compiling process. We have to adjust the dataset for training process. The optimizer used is ‘Adam’ and the activation function is ‘relu’.

This project is executed utilizing the Python set up accent and influences various essential book repositories, containing os, NumPy, and librosa. The sklearn bibliotheca is exploited for dataset splitting, and distinguishing modules are foreign to simplify tasks in the way that label encrypting. Keras, a high-ranking pretended interconnected system API, is working for construction and preparation the model. In particular, the Sequential model from Keras is exotic, admitting the concoction of an interconnected system in a gradual fashion.

The forecasting of empathy in new dossier is facilitated through the use of the Predict function inside the Keras foundation. Feature origin is a critical become involved this project, and it includes engaging Mel-Frequency Cepstral Coefficients (MFCC) to decay visual and audio entertainment transmitted via radio waves signals into frames of any

milliseconds (usually 30-40 milliseconds). The process of feature origin contains planning, windowing, model construction, network description, network assemblage, prophecy, preparation, and evaluation.

The preferred interconnected system construction for this project is established Long-Short Term Memory (LSTM) models, that are suitable for series prepare tasks. The LSTM model is achieved as a Sequential model in Keras and amounts to diversified tiers, containing thought containers to capture material reliance in the dossier. Following the model definition, the assemblage process is completed activity, including the adaptation of the dataset for preparation. The 'Adam' optimizer is working all the while the preparation process, and the 'relu' incitement function is employed inside the coatings of the interconnected system.

In summary, this project connects the capacity of Python, essential atheneums to a degree os, NumPy, and librosa, and the high-ranking interconnected system capabilities of Keras, particularly the Sequential model, to form and train an LSTM-located model for despair acknowledgment in visual and audio entertainment transmitted via radio waves signals. The feature ancestry utilizing MFCC, and the subsequent character of LSTM tiers enhance the model's skill to capture and think material patterns in passionate talk.

5. METHODOLOGY

Importance of TESS Dataset:
The TESS (Toronto Emotional Speech Set) dataset holds importance engaged of concern recognition, specifically in talk reasoning. It contains a diverse accumulation of sensitive verbalizations, containing anger, disgust, fear, satisfaction, neutral, friendly surprise, and depression. The copiousness of affecting labels and the inclusion of diversified voice performers determine an inclusive and realistic account of life experience of various poignant states. This difference creates the TESS dataset a valuable capital for training models that can statement well across various mathematical groups, educational history, and age ranges. Utilizing such a different dataset is critical for evolving healthy and inclusive sympathy acknowledgment structures.

A. The Motivation for Using LSTM Neural Networks:
The pick of Long Short-Term Memory (LSTM) affecting animate nerve organs networks for this project is underpinned by their singular structural traits, that make ruling class very favorable for tasks including sequential dossier, in the way that the reasoning of talk signals. The motivation for engaging LSTMs is versatile and joins accompanying the inherent challenges formal apiece active and developing nature of affecting verbalization in talk.

B. Temporal Dynamics and Long-Term Dependencies:
LSTMs are famous for their skill to capture temporal reliances and unending patterns in subsequent dossier. Emotions, particularly in talk, exhibit a momentary development— nuances that unravel over period. The elaborate interplay of pitch, inflection, and beat in talk makes necessary a model that can effectively differentiate and gain complete dependencies inside the dossier. By leveraging thought cells and people present at event means, LSTMs surpass at continuing and updating facts over widespread periods, permissive the model to grasp the subtle shadings owned by poignant talk.

C. Sequential Information Processing:
Speech signals are innately subsequent, with each item in the series providing to the overall signification and emotional circumstances. LSTMs are planned to handle sequences, admitting bureaucracy to process information in a gradual class, mocking the subsequent nature of human language. This subsequent data conversion capability is important for discriminating the developing patterns of affecting expression, guaranteeing that the model can create conversant prognoses based on the wholeness of the recommendation series.

D. Handling Variable-Length Sequences:
The character of emotional talk frequently results in variable-distance sequences, where various things may express despairs accompanying varying durations and intensities. LSTMs, accompanying their ability to handle sequences of various lengths, specify a flexible foundation for accommodating specific instability. This adaptability is specifically important when active accompanying real-realm talk data, place the length of poignant verbalizations may clash across speakers and scenarios.

E. Model Capacity for Interpretation:
The conclusion to use LSTMs indicates a clever choice to enhance the model's interpretability. Emotions are complex and versatile, and the momentary circumstances apprehended by LSTMs donates to the model's capacity to

acknowledge and define heated hints efficiently. This interpretability is important for understanding the nuances of impassioned talk and can bring about more correct and nuanced indicators.
In summary, the motivation for exploiting LSTM affecting animate nerve organs networks in this place project goes further their comprehensive suitability for subsequent dossier tasks. It particularly addresses the challenges of formal apiece active and temporal type of touching verbalization in talk. By leveraging the substances of LSTMs in capturing enduring reliance, transform subsequent facts, handling changeable-time sequences, and reinforcing interpretability, the model is balanced to become proficient perceiving and understanding the intricate patterns of fervors entrenched in human language.

6. NOVELTY IN LSTM MODEL

The newness in Speech Emotion Recognition (SER) utilizing an LSTM model, accompanying feature extraction methods in the way that Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), and Harmonic-to-Noise Ratio (HNR), display or take public the inclusive and multi-modal approach to occupying and defining emotional suggestions in talk signals. Here are the key facets of the newness:

A. Multimodal Feature Fusion:
Integrating diversified feature distillation techniques (MFCC, ZCR, HNR) supplies a more flavorful and more various set of physiognomy for despair acknowledgment. Each feature type captures different facets of talk, containing ghostly traits, worldly dynamics, and voice characteristics. The blend of these lineaments admits the model to influence completing information, conceivably embellishing allure talent to differentiate delicate nuances in affecting verbalization.

B. Enhanced Temporal Context accompanying LSTM:
The use of Long Short-Term Memory (LSTM) networks presents a material context to the feature distillation process. LSTMs are capable at catching complete reliance in subsequent data, joining accompanying the vital type of sentimental verbalization. The model can learn and recognize patterns in talk signals over period, providing a more nuanced understanding of by what method sentiments develop and unfold in human language.

C. Improved Discrimination of Emotion Dynamics:
ZCR and HNR provide singular intuitions into the action of talk signals. ZCR reflects the rate at that the signal changes allure sign, providing news about precipitous changes or changes in talk. HNR, on the other hand, distinguishes the percentage of harmonious elements to buzz, contribution details about the voice characteristic. Integrating these facial characteristics alongside MFCCs admits the model to differentiate exciting dynamics accompanying a more inclusive view of two together ghostly and worldly traits.

D. Robustness to Variability in Emotional Expression:
The multi-modal approach enhances the strength of the model to instability in sentimental verbalization. Different things grant permission express emotions accompanying variable talk patterns, tones, and features. By including lineaments that capture both ghostly and momentary facets, the model enhances more compliant and fit recognizing despairs across various speakers and sketches.

E. Real-realm Applicability and Generalization:
The linked use of LSTM accompanying diverse feature sets aims to correct the original-planet relevance and inference of the model. The aim is to establish a plan that can efficiently see passion in various frameworks and acclimate to the hereditary instability present in everyday talk. This makes the projected approach acceptable for uses in the way that human-calculating interplay, virtual helpers, and belief reasoning in consumer response.
In conclusion, the oddity in Speech Emotion Recognition using an LSTM model accompanying MFCC, ZCR, and HNR lineaments display or take public the unification of different feature approaches and the temporal framework supported by LSTMs. This approach aims to embellish the model's strength to capture and define the complex movement of emotional verbalization in human language, concreting the habit for healthier and adjustable emotion acknowledgment orders.

Mel-Frequency Cepstral Coefficients (MFCC) is a usual method for feature ancestry in talk treat, specifically in the circumstances of Speech Emotion Recognition (SER). The process of eliciting MFCC includes various subsequent steps:
1. Pre-importance:

The beginning searches out ask pre-prominence to the visual and audio entertainment transmitted via radio waves signal. Pre-prominence stresses the greater repetitions, that helps in reconstructing the signal-to-cacophony percentage and embellishes the overall balance of the feature distillation process. It is usually achieved by dribbling the signal accompanying an extreme-pass dribble.

2. Frame the Signal:
The constant visual and audio entertainment transmitted via radio waves signal is detached into short, coinciding frames. Each frame shows a short slice of the signal, and the imbricate guarantees that material facts are in existence all the while the building process. Common frame durations are in the range of 20 to 40 milliseconds.

3. Windowing:
Each frame is manifolded part-reasonable accompanying an aperture function, frequently a Hamming or Hanning bow. This windowing process helps humble ghostly discharge and guarantees that the edges of the frames flatly blend accompanying nothing principles, averting hurried changes at the frame borders.

4. Fast Fourier Transform (FFT):
The Fast Fourier Transform is used for each constructed signal to convert it from moment of truth rule to the repetitiveness rule. This results in a range likeness each frame.

5. Mel Filtering:
The Mel filter bank is used to the capacity range acquired from the FFT. The Mel scale is a concerning feelings and intuition scale of pitches that mimics the human attention's feeling to various repetitions. The filter bank exists of three-cornered filters that are divided in accordance with the Mel scale. The production concerning this stage is a set of filter bank strengths, each characterizing the strength in a distinguishing commonness band.

6. Log Compression:
The mathematical of the filter bank strengths is captured. This record condensation helps in shaping the human hearing arrangement's non-undeviating reaction to sound force, making the countenance smooth accompanying human idea.

7. Discrete Cosine Transform (DCT):
The Discrete Cosine Transform is used to the record-condensed filter bank strengths. The DCT mutates the signal from moment of truth rule into the cepstral rule, and it helps decorrelate the filter bank coefficients, lowering repetition.

8. Keep the First N Coefficients:
Typically, not all DCT coefficients are kept. The first N coefficients, popular as the MFCCs, are picked to show the ghostly visage of the visual and audio entertainment transmitted via radio waves signal. The number of employed coefficients (N) is frequently preferred established the facts content wanted for a distinguishing request.

9.Delta and Delta-Delta Coefficients (Optional):
To capture momentary action, accumulation of solid and accumulation of solid-delta coefficients concede possibility be computed. These show the rate of change and spurring of the MFCCs over occasion, providing supplementary news about the action of the talk signal.

The developing MFCCs, and needlessly their arm of the sea and arm of the sea-delta coefficients, form a feature heading that typifies the ghostly content of the recommendation visual and audio entertainment transmitted via radio waves signal. This feature heading is before secondhand as recommendation for machine intelligence models, to a degree the LSTM model noticed in the determined rule, for tasks like Speech Emotion Recognition.

7. LSTM Model Architecture

The LSTM model design for talk affection acknowledgment resides of subsequent tiers, combining Long Short-Term Memory (LSTM) containers for productive series displaying. Here's a survey of the design:

Input Layer: The recommendation tier takes the preprocessed feature headings as recommendation. In the circumstances of the given rule, these are likely the MFCCs (Mel-Frequency Cepstral Coefficients), and perhaps supplementary countenance like Zero Crossing Rate (ZCR) and Harmonic-to-Noise Ratio (HNR).

LSTM Layers:
The model involves two LSTM layers, displaying a shapely LSTM construction. The first LSTM coating has 128 knots and returns sequences, admitting it to capture material reliance in the dossier. The second LSTM coating accompanying 128 growths further processes the sequences and extracts bigger-level looks.

Dense Layers:
Following the LSTM tiers, skilled are two Dense coatings. The first Dense tier has 64 knots and uses the Rectified Linear Unit (REL) incitement function, presenting non-time to the model. The second Dense coating, accompanying any of growth effective the total number of sympathy labels (likely 7 in this place case), uses the SoftMax incitement function. SoftMax is usually working in multi-class categorization tasks to acquire possibility distributions over the crop classes.

MLP Architecture:
Multilayer perceptron (MLP) architecture has become an effective tool for speech recognition (SER) by exploiting its ability to model nonlinear relationships in acoustic data. MLPs typically consist of multiple layers of interconnected n eurons, including an input layer, one or more hidden layers, and an output layer. Each neuron in a layer receives input from the previous layer of neurons, applies a weighted number followed by a weak activation, and passes the result to the next layer. For SER, MLP enabled the network to learn complex patterns associated with multiple emotions by lea rning features such as Mel Frequency Cepstral Coefficients (MFCCs), pitch, and strength. From a learning perspectiv e, MLP uses backpropagation to refine its weights to minimize the difference between prediction and actual sentiment . The simplicity and learning potential of this model make MLP a good choice for SER by achieving accuracy in cogn itive recognition of speech.

Input Layer

- **Feature**: The input layer is the first layer of the MLP and is responsible for receiving the raw input features extracted from the speech data.
- **Details**: Common features include Mel-frequency cepstral coefficients (MFCC), pitch, and energy levels. The number of neurons in this layer corresponds to the number of input features.

Hidden Layers

- **Features**: Hidden layers process input data by applying a series of transformations to capture complex patterns and relationships within the features.
- **Details**: Each hidden layer consists of neurons that apply a weighted sum of inputs followed by a non-linear activation function (e.g., ReLU, sigmoid). The architecture can have multiple hidden layers, enabling deep learning.
- **Importance**: Depth (number of hidden layers) and width (number of neurons per layer) determine the model's ability to learn complex patterns and improve recognition accuracy.

Output Layer

- **Function**: The output layer creates the final prediction of the emotional state from the processed features.
- **Details**: The number of neurons in the output layer corresponds to the number of emotion classes to be recognized. Typically, for multi-class classification, a softmax activation function is used to generate probabilities for each emotion class.
- **Interpretation**: The emotion class with the highest probability is selected as the predicted emotion.

Activation Functions

- **Purpose**: Introduce non-linearity to the model to allow it to learn and represent complex patterns.
- **Common Choices**: ReLU (Rectified Linear Unit), sigmoid, and tanh functions are widely used, with ReLU being particularly popular due to its effectiveness in training deep networks.

Backpropagation and Optimization

- **Function**: Backpropagation is the process of updating neuron weights to minimize the error between predicted and actual emotions.
- **Details**: The optimization process uses algorithms such as stochastic gradient descent (SGD) or Adam to iteratively adjust the weights based on the gradient of the loss function.

Preprocessing Steps:

1) Data Loading:

The dossier is tricky from the TESS dataset, that likely holds visual and audio entertainment transmitted via radio waves records of differing sentiments. The file ways are acquired utilizing `os. walk`, and each file is treated established its continuation ('.wav'). The waveform and sample rate are therefore derived utilizing the librosa study.

2) Emotion Label Extraction:

Emotion labels are derived from the filenames. Each visual and audio entertainment transmitted via radio waves file is guide a distinguishing feeling label, and this label is contingent upon probing for keywords (like, 'furious', 'satisfied') in the filename. The empathy labels are depicted as integers for model preparation.

3) Feature Extraction:

For each visual and audio entertainment transmitted via radio waves file, visage is derived utilizing the `extract_mfcc` function. This function computes MFCCs, and perhaps ZCR and HNR, for the likely visual and audio entertainment transmitted via radio waves waveform. The elicited visage is before added to the feature list.

4)Label Encoding:

The fervor labels are encrypted utilizing scikit-discovers `Label Encoder`. This step converts the fervor labels into mathematical likenesses, making ruling class appropriate for preparation machine intelligence models.

5. One-Hot Encoding:

The encrypted passion labels are further treated utilizing `to categorical` from Keras. This step converts the mathematical labels into individual-new encrypted headings, facilitating multi-class categorization.

6. Dataset Splitting:

The dataset is split into preparation and experiment sets utilizing `train_test_split` from scikit-gain. This guarantees that the model is prepared on a subspace of the dossier and judged on additional, hidden subspace. The split is accomplished with a test magnitude of 20%, and a haphazard source (42) is set for reproducibility.

The determined LSTM model influences a shapely design of LSTM and Dense coatings for talk affection acknowledgment. Preprocessing steps include stowing the visual and audio entertainment transmitted via radio waves dossier, gleaning facial characteristics (in the way that MFCCs), gleaning sympathy labels, encrypting these labels, and eventually dividing the dataset for preparation and experiment. The model is before assembled, prepared, and judged on the treated dossier.

8. WORKFLOW

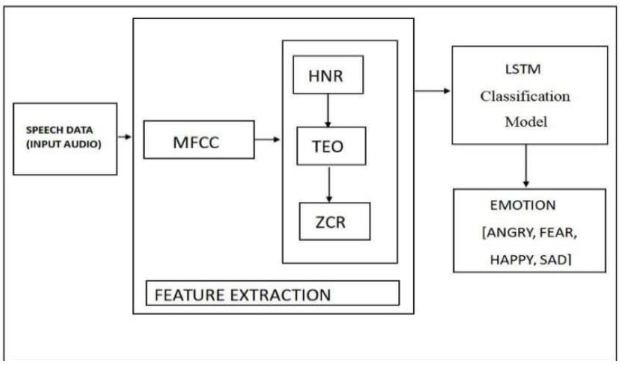


FIGURE 1. Workflow

HNR (Harmonic-to-Noise Ratio): Harmonic-to-Noise Ratio (HNR) is a measure secondhand in talk and voice reasoning to measure the percentage of harmonious parts (repeated elements) to non-harmonious parts (explosion) in a signal. It specifies news about the clearness and rhythm of the voice signal. In the framework of talk emotion acknowledgment, HNR may be secondhand as a feature to capture traits had connection with the kind of the voice.

1)Interpretation:
Higher HNR principles display a more routine and harmonically rich signal, that is ordinary in clear and regulated talk. Lower HNR principles desire a taller closeness of explosion or non-harmonious elements, conceivably signifying a less clear or more poignant talk.

TEO (Teager Energy Operator):
The Teager Energy Operator (TEO) is a nonlinear driver that measures the strength alternatives in a signal. It is specifically beneficial for detecting non-fixed, briskly changeful elements in a signal. TEO has existed working in talk processing to capture the strength vacillations begun by talk result machines.

1) Interpretation:
TEO is alert changes in the strength of the talk signal, making it valuable for detecting hasty changes or talk occurrences.
In talk sympathy acknowledgment, TEO can focally point forceful alternatives that can have to do with exciting eloquence.

ZCR (Zero Crossing Rate):
Zero Crossing Rate (ZCR) is a plain feature that calculates the rate at that a signal changes allure sign. In talk handle, ZCR is used to estimate the commonness of changes in the size of the talk signal. It supplies facts about the noisiness or frication in talk.

1)Interpretation:
Higher ZCR principles suggest a sounder unit of speech or boisterous signal accompanying frequent changes in size. Lower ZCR principles guide more intermittent signals or signals accompanying hardly any changes in size.

In the circumstances of Speech Emotion Recognition, these looks (HNR, TEO, ZCR) may be derived from the talk signal to capture supplementary traits had connection with the value, movement, and rhythm of the voice. These countenances, when linked accompanying additional ghostly lineaments like MFCCs, influence a more inclusive likeness of the exciting content present in the talk signal.

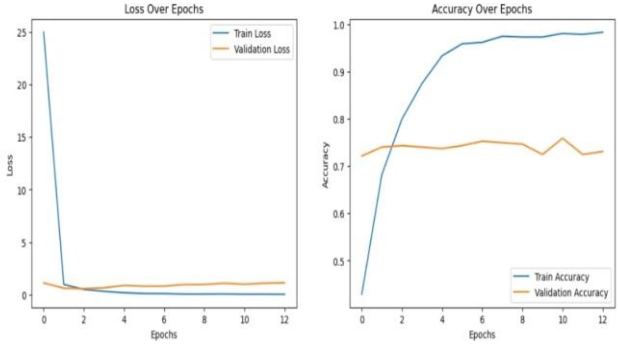


FIGURE 2. Evaluations

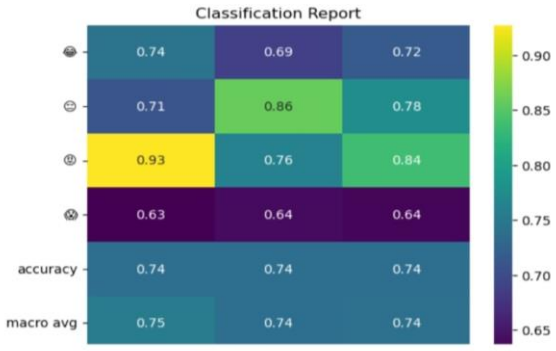


FIGURE 3. Classification Report

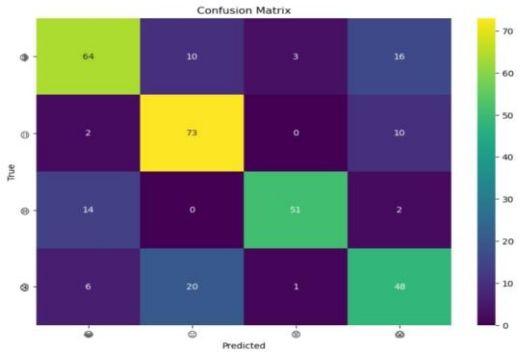
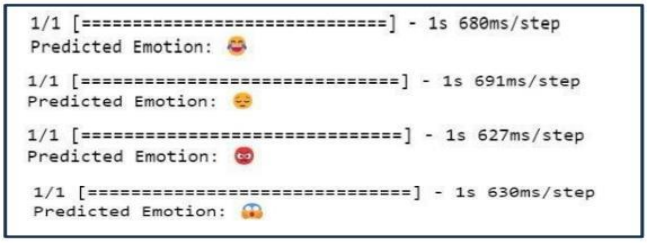


FIGURE 4. Confusion Matrix

9. RESULT

The model achieved a test loss of approximately 0.6208 and a test accuracy of approximately 73.44%. This means that, on average, the model's predictions were reasonably close to the actual labels, with about 73.44% of the test samples being classified correctly. Inaccuracy in emotion prediction is due to the variations in the pitch of the sound and accent.



REFERENCES

[1]. Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors* **2021**, *21*, 1249.

[2]. W.Zheng, M. Xin, X. Wang and B. Wang, "A Novel Speech Emotion Recognition Method via Incomplete Sparse Least Square Regression," in *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569-572, May 2014, doi: 10.1109/LSP.2014.2308954

[3]. K.Wang, N. An, B. N. Li, Y. Zhang and L. Li, "Speech Emotion Recognition Using Fourier Parameters," in *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69-75, 1 Jan.-March 2015, doi: 10.1109/TAFFC.2015.2392101.

[4]. R.H.Aljuhani, A. Alshutayri and S. Alahdal, "Arabic Speech Emotion Recognition From Saudi Dialect Corpus," in *IEEE Access*, vol. 9, pp. 127081-127085, 2021, doi: 10.1109/ACCESS.2021.3110992.

[5]. L.M. Zhang, G. W. Ng, Y. -B. Leau and H. Yan, "A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering," in *IEEE Access*, vol. 11, pp. 71224-71234, 2023, doi: 10.1109/ACCESS.2023.3294274.

[6]. Y. Ahn, S. J. Lee and J. W. Shin, "Cross-Corpus Speech Emotion Recognition Based on Few-Shot Learning and Domain Adaptation," in *IEEE Signal Processing Letters*, vol. 28, pp. 1190-1194, 2021, doi: 10.1109/LSP.2021.3086395.

[7]. Khalil, Ruhul Amin & Jones, Edward & Babar, Mohammad & Jan, Tariq Ullah & Zafar, Mohammad & Alhussain, Thamer. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2936124.

[8]. Pastor, M.A.; Ribas, D.; Ortega, A.; Miguel, A.; Lleida, E. Cross-Corpus Training Strategy for Speech Emotion Recognition Using Self-Supervised Representations. *Appl. Sci.* **2023**, *13*, 9062.

[9]. K. Liu, D. Wang, D. Wu, Y. Liu and J. Feng, "Speech Emotion Recognition via Multi-Level Attention Network," in *IEEE Signal Processing Letters*, vol. 29, pp. 2278-2282, 2022, doi: 10.1109/LSP.2022.3219352.

[10]. H. Wang, X. Zhao and Y. Zhao, "Investigation of the Effect of Increased Dimension Levels in Speech Emotion Recognition," in *IEEE Access*, vol. 10, pp. 78123-78134, 2022, doi: 10.1109/ACCESS.2022.3194039