KEC Conference 2026

ISSN: 3021-940X (Print)

DOI: 10.3126/injet.v3i2.95516

Nepali Speech Emotion Detection Using Deep Learning

Uttam Pandeya[1], Basanta Joshi[2*]

[1]*Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Lalitpur, Nepal, 080msice020.uttam@pcampus.edu.np*

[2] *Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Lalitpur, Nepal, basanta@ioe.edu.np*

Abstract

Emotionally intelligent human-computer interaction solutions depend on Speech Emotion Recognition (SER), which attempts to recognize emotional states from speech. There is still little research on SER for languages with limited resources, like Nepali. In this work, a one-dimensional Convolutional Neural Network (1D-CNN) and Mel-Frequency Cepstral Coefficients(MFCCs) are used in a deep learning-based Nepali speech emotion detection system. 1,810 audio samples of 632 happy, 560 neutral, and 618 sad utterances were gathered from studio recordings, mobile recordings, podcasts, and broadcast sources to create a specific Nepali emotional speech dataset. Every audio sample underwent preprocessing, resampling to 16 kHz, and conversion to mono. A 1D-CNN model was fed MFCC features that had been retrieved. The suggested model yields an overall accuracy of 88% on the Nepali dataset, according to experimental results. With a precision of 0.96, recall of 0.92, and F1-score of 0.94, the Sad emotion class performed the best. The Neutral class received a precision of 0.89 and an F1-score of 0.81, but the Happy class received a recall of 0.98 and an F1-score of 0.89. Strong discrimination was shown by ROC analysis, with AUC values of 0.97 for neutral and 0.99 for happy and sad.

Keywords: Speech Emotion Recognition, Nepali Language, Deep Learning, MFCC, CNN.

1. Introduction

One of the most organic and efficient ways for people to communicate is through speech. Speech conveys rich emotional information that indicates a speaker's intention, attitude, and mental state in addition to linguistic content. Speech Emotion Recognition (SER), the ability to automatically identify emotions from speech, has drawn a lot of attention lately because of its many uses in intelligent learning systems, virtual assistants, call center analytics, mental health monitoring, and human–computer interaction [1]. SER systems' performance has significantly improved with the development of deep learning and artificial intelligence, especially for high-resource languages like Mandarin, German, and English. However, little is known about how to recognize emotions in low-resource languages like Nepali. Millions of people speak Nepali, which is the official language of Nepal. Emotion recognition is difficult because of its distinctive linguistic features, which include a variety of dialects, phonetic diversity, and distinctive prosodic patterns. The creation of reliable SER models for Nepali is further hampered by the absence of sizable, thoroughly annotated emotional speech datasets. With these potent tools, end-to-end systems that accurately classify emotions like happy, sadness and neutrality in Nepali speech may now be built [2].

Conventional SER methods mostly used conventional machine learning classifiers in conjunction with manually created acoustic variables including pitch, energy, and Mel- Frequency Cepstral Coefficients (MFCCs). Although these techniques produced baseline performance, they frequently had trouble generalizing to different speakers and recording environments. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), two recent deep learning approaches, have shown exceptional ability to automatically acquire discriminative emotional aspects from speech representations.

The goal of this project is to create a framework for Nepali speech emotion detection based on deep learning. The study uses a convolutional neural network architecture and MFCC-based feature extraction to solve language-specific issues and data shortages. The results of this work support the creation of

* *Corresponding author*

emotionally intelligent systems that are inclusive of under-represented linguistic communities and advance effective computing research for low-resource languages.

2. Literature Review

The absence of substantial, realistic, and balanced resources for Speech Emotion Recognition (SER) in the Urdu language led to the creation of the UrduSER dataset. UrduSER offers a more natural and all-encompassing resource than the current Urdu SER datasets, which are small, studio-controlled, and have limited emotional diversity.

Ten professional actors, five male and five female, between the ages of 20 and 65, participated in emotionally charged conversation in Pakistani dramas, telefilms, and TV series. With 3,500 audio samples, it covers seven emotions: anger, fear, boredom, disgust, happiness, neutrality, and sadness. (500 for each emotion), evenly distributed among actors and genders.

All audio was standardized to WAV format (44.1 kHz, 32-bit, 2–6 s), and preprocessing included music separation, noise suppression, and silence shortening. Comprehensive metadata (actor identity, gender, emotion, length, and original Urdu script) and an organized file-naming strategy were supplied. UrduSER is a reliable, realistic, and demographically varied benchmark for Urdu speech emotion recognition (SER) since label reliability was guaranteed through multi-expert annotation and validation by 100 native Urdu speakers, resulting in 94% agreement. To get beyond the drawbacks of MFCC-centric approaches, SER was carried out concurrently by fusing statistical wavelet characteristics with traditional time-frequency features. Experiments using silence removal, framing, and windowing on the RAVDESS dataset (1,440 recordings from 24 actors across 8 emotions) produced 428 characteristics per sample, including wavelet-based statistics generated from DWT coefficients, spectral and time-domain descriptors, and MFCCs.

Optuna was used to optimize Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) models, which were then assessed using 5-fold cross-validation. The findings demonstrate that, in comparison to conventional methods, wavelet enhanced features enhance emotional speech representation and improve Speech Emotion Recognition (SER) performance [3]. E-Speech is an organized and useful resource for furthering SER research since the dataset is recorded in WAV format (44.1 kHz, mono) with metadata describing file name, speaker information, emotion label, and duration [4]. Overall, the results show that raw waveform-based deep learning models, particularly CNN-LSTM, outperform conventional feature engineering techniques and generalize better across datasets, moving SER closer to more reliable and automated systems [5].

A dual feature extraction methodology is suggested to enhance Speech Emotion Recognition (SER). It combines a spectrogram-based CNN encoder (AlexNet variation) with a semantic feature encoder that employs MFCCs and Speech2Vec embedding. The fused features are classified using an LSTM network, which achieves 94.8% accuracy on RAVDESS and 94.0% on EMO-DB, indicating superior robustness than standard single-feature approaches [6]. Experiments with the TESS and RAVDESS datasets demonstrate that the CNN model attained the maximum accuracy of 97.1%, outperforming LSTM 92%, GRU 93%, Random Forest 96%, and SVM 86%. The findings show that CNNs are good at recognizing emotions in speech [7].

This study offers an ensemble-based Speech Emotion Recognition (SER) framework that combines CNN, LSTM, and MFCC features together with an attention mechanism. Tested on the RAVDESS, TESS, SAVEE, and CREMA-D datasets, the model obtained 87.08% accuracy, beating individual CNN and LSTM models and enhancing performance for real-time emotion recognition [8].

Some preliminary datasets, including one with over 3,000 samples spanning multiple emotions, have surfaced on platforms like Kaggle, despite the fact that Nepali Speech Emotion Recognition (SER) is still understudied. Nevertheless, these resources typically lack benchmark evaluation methodologies, standardized collection and annotation methods, and peer-reviewed documentation. They lack the

validation, speaker diversity, and reproducibility of well-known datasets like RAVDESS and EMO-DB. As a result, there is still a glaring need for Nepali SER datasets that are benchmark-quality and well-structured.

3. Methodology

3.1 Block Diagram for Nepali SER

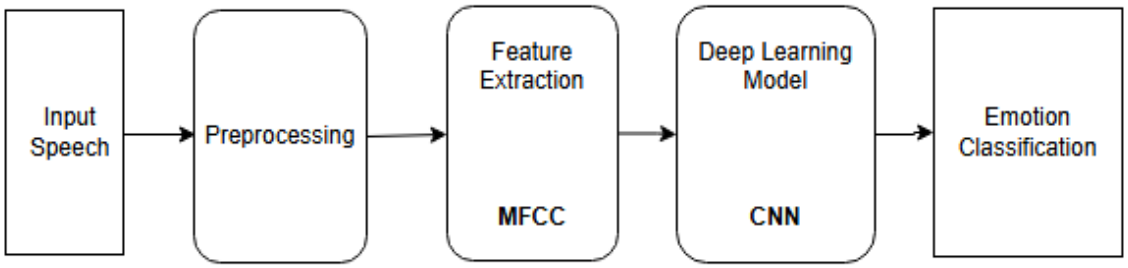

Figure 1. Block diagram of the proposed Nepali Speech Emotion Recognition (SER) system showing preprocessing, feature extraction, CNN model, and classification stages.

A sequential pipeline comprising input speech, preprocessing, feature extraction, deep learning, and emotion classification is used by the suggested Nepali Speech Emotion Detection system. To enhance signal quality, Nepali voice samples are first gathered and preprocessed using segmentation, noise reduction, normalization, and silence removal. After that, significant spectral and perceptual aspects of speech are captured by extracting Mel-Frequency Cepstral Coefficients (MFCCs). A Convolutional Neural Network (CNN) receives these MFCC information and automatically learns discriminative patterns. Lastly, CNN uses accuracy and loss metrics to assess the system's performance after classifying speech into emotion categories like happy, sad, angry, and neutral.

3.2 Audio Data Collection

Due to the dearth of public resources, a new dataset had to be created in order to collect audio data for Nepali speech emotion identification. Samples of native speakers expressing joyful, sad, and neutral emotions were recorded in a studio setting and meticulously labelled. YouTube was used to gather more natural emotional speech, which needed to be normalized and noise-removed. Only publicly available data was used for research, all audio was verified for quality, and ethical concerns were guaranteed. Consent was acquired for recordings. The dataset is made up of 1,810 utterances that were gathered for the purpose of identifying emotions. Of them, 560 samples (31%) are classified as neutral, 632 samples (34.9%) as happy, and 618 samples (34.1%) as sad. There isn't a particular category that is noticeably over-represented in the sample because the distribution of the three emotion classes is fairly balanced. The total length of recordings for each emotion class was examined in addition to the quantity of samples. There are roughly 44 minutes and 19 seconds of audio in the happy class, 46 minutes and 59 seconds in the neutral class, and 45 minutes and 43 seconds in the sad class. A balanced dataset in terms of both sample count and temporal representation is indicated by the comparatively similar durations across classes.

3.3 Data Preprocessing

Following the collection of happy, sad, and neutral speech Nepali audio datasets, all files were converted from sources such as.m4a,.mp3, and.mp4 to a common.wav format. The audio was standardized to mono, resampled at 16 kHz, and saved as 16-bit PCM WAV files. This sampling rate is frequently employed in speech processing and benchmark datasets for tasks like spoofing detection because it strikes a balance between computing efficiency and audio quality.

3.4 Feature Extraction

For the purpose of identifying emotions, the feature extraction module transforms unprocessed Nepali speech into concise numerical representations. To capture the timbral, tonal, and perceptual aspects of speech, high-quality resampling is used first, and then MFCCs, chroma, Mel-spectrogram, spectral contrast,

and tonnetz features are extracted. A robust representation is created by averaging these properties throughout time.

Because they capture tonal, intensity, and temporal fluctuations that are essential for identifying emotional states like sadness and anger, MFCCs, in conjunction with delta and delta-delta coefficients, are particularly useful for Nepali speech.

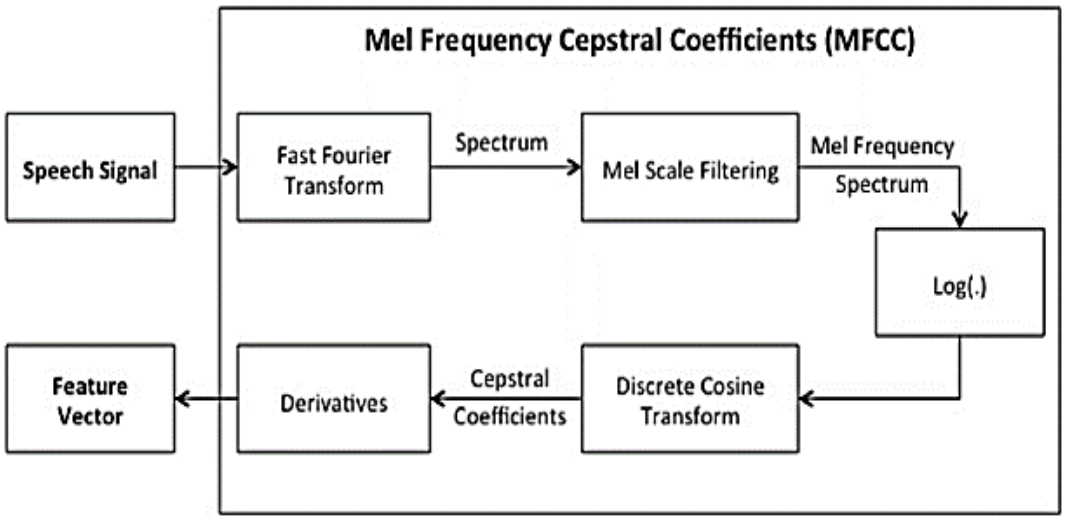

Figure 2. MFCC feature extraction process illustrating the transformation of raw Nepali speech signals into spectral coefficients.

3.4.1 Audio Duration Analysis

Nepali speech recordings ranging in duration from two to fourteen seconds are included in the dataset. The majority of samples range in duration from 4 to 8 seconds, with 6 seconds being the most typical. Shorter clips are more common than longer ones due to the distribution's small right skew. In order to preserve consistency and efficiency while preparing data for speech and emotion recognition models, this duration pattern is crucial for preprocessing operations like segmentation, padding, and cutting.

3.4.2 MFCC Feature Analysis

The average Mel-Frequency Cepstral Coefficients (MFCCs) for every emotion class in the dataset are displayed in the MFCC Mean Heatmap. Because MFCCs capture significant spectrum characteristics of speech signals, they are frequently employed in speech and emotion recognition. The three emotion classes happy, neutral, and sad—are shown on the y-axis, while the 13 MFCC coefficients are shown on the x-axis. The magnitude of the coefficients is reflected in color intensity, where larger values are represented by warmer colors and lower values by cooler colors. The apparent variations in MFCC patterns among emotions imply that every emotional state has unique spectral properties. This illustrates how well MFCC features discriminate in speech emotion identification tasks.

3.4.3 Intensity Distribution Analysis Using Violin Plots

The distribution and fluctuation of intensity levels for various emotion classes are displayed in the charts. With a wider, more symmetrical distribution and a higher average intensity (between 0.02 and 0.20), the Happy class exhibits more expressive speech. The distribution of the Neutral class is narrower and more peaked, with intensity primarily falling between 0.05 and 0.15, showing consistent voice patterns.

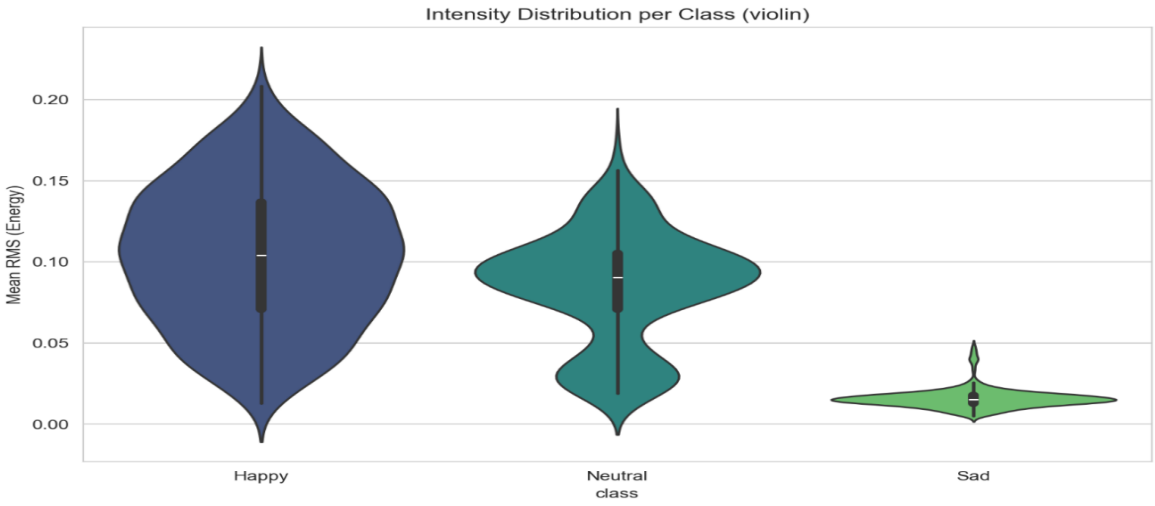

Figure 3. Violin plot showing intensity distribution variations among happy, neutral, and sad speech.

The Sad class, on the other hand, has a more unequal distribution and a lower average intensity (between 0.02 and 0.04), indicating softer speaking. These variations in intensity patterns show the value of intensity features in speech emotion identification and aid in the differentiation of emotional states.

3.4.4 Audio Intensity Analysis Using Boxplots

For every emotion class, the boxplots display the median and distribution of intensity values. The Happy class's speech is more animated and dynamic, as seen by its higher median intensity and greater variety.

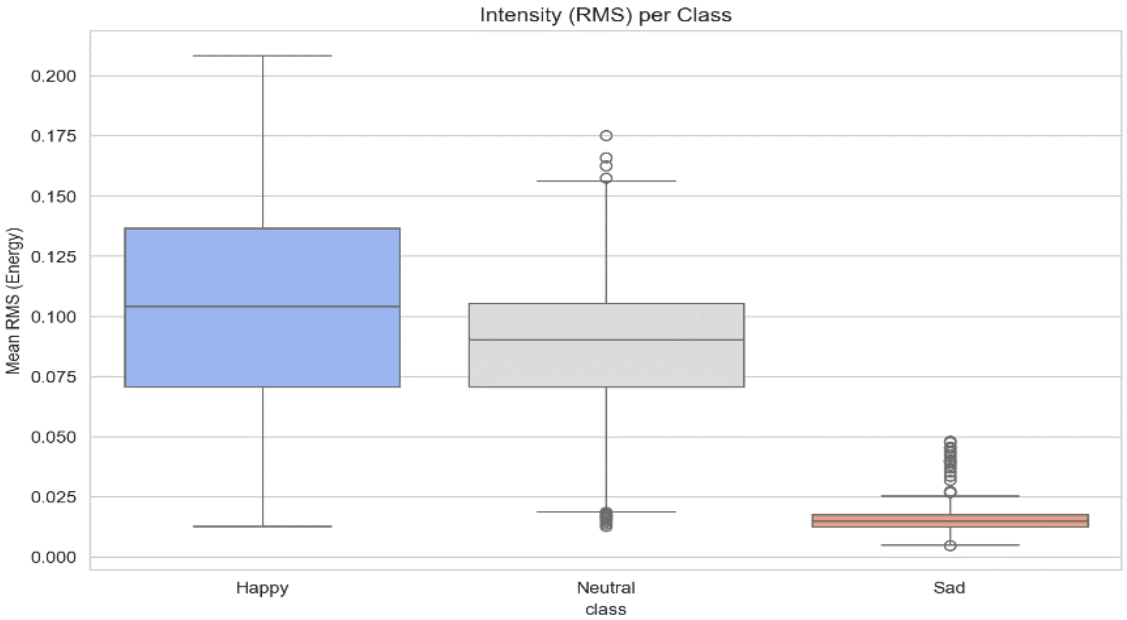

Figure 4. Boxplot illustrating the spread and median of audio intensity across emotion classes.

The Neutral class exhibits consistent speech patterns with a moderate median and less variability. Softer and more muted speech is represented by the Sad class, which has the lowest median intensity and the least variation. These variations demonstrate RMS energy's discriminative function in identifying emotional states and bolster its applicability in feature extraction for speech emotion detection systems.

3.4.5 Pitch Distribution Analysis Using Violin Plots

There are distinct disparities between happy, neutral, and sad speech, according to the acoustic study. While neutral speech has moderate intensity and steady pitch patterns, happy speech has more intensity and more variance, signifying an animated and expressive delivery. Lower intensity and gentler vocal energy, as well as discernible pitch change, are characteristics of sad speech. The high discriminative power of acoustic characteristics for efficient emotion recognition in speech systems is demonstrated by these variations in pitch and intensity.

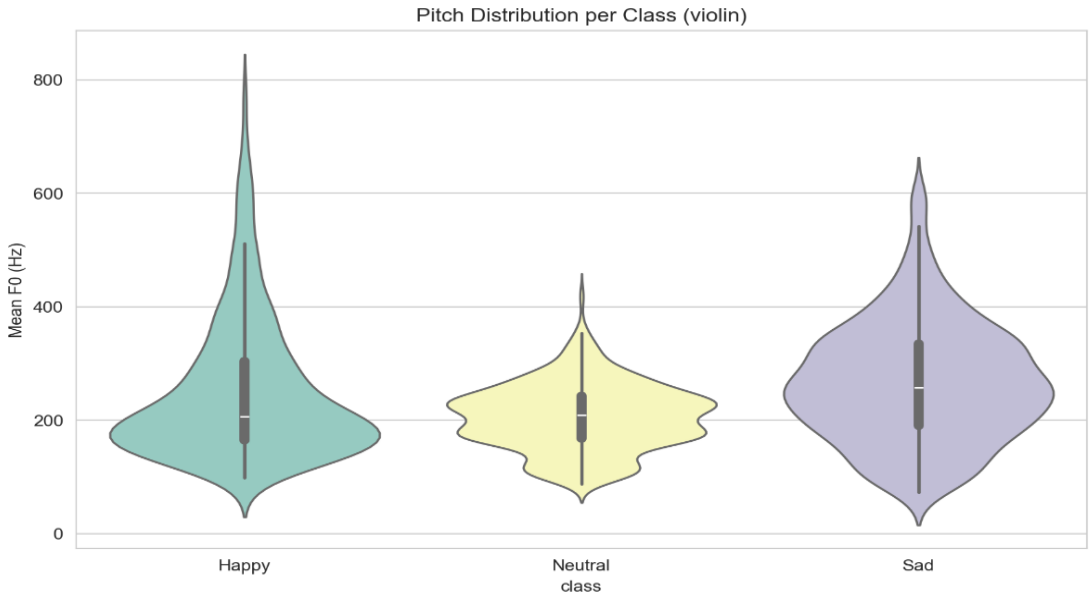

Figure 5. Violin plot representing pitch distribution differences among emotional speech categories.

3.5 1D CNN Model

There are distinct disparities between happy, neutral, and sad speech, according to the acoustic study. While neutral speech has moderate intensity and steady pitch patterns, happy speech has more intensity and more variance, signifying an animated and expressive delivery. Lower intensity and gentler vocal energy, as well as discernible pitch change, are characteristics of sad speech. The high discriminative power of acoustic characteristics for efficient emotion recognition in speech systems is demonstrated by these variations in pitch and intensity.

The suggested model classifies emotions using a one-dimensional Convolutional Neural Network (1D-CNN) architecture. A max-pooling layer with a pool size of two for dimensionality reduction, a dropout layer with a rate of 0.2 to prevent overfitting, batch normalization to stabilize learning, and a Conv1D layer with 128 filters and a kernel size of five using ReLU activation and the same padding make up each of the network's two convolutional blocks. After being flattened, the collected feature maps are sent to a fully connected dense layer with 128 neurons and ReLU activation. A dropout layer with a rate of 0.3 comes after that. Multi-class emotion categorization is carried out by the final output layer using a softmax activation function. A dataset of 1,810 samples, divided into 70% training (1,267 samples), 10% validation (181 samples), and 20% testing (362 samples), was used to train the model. A batch size of 32 was used for training across 200 epochs. The model was trained using the Adam optimizer with a default learning rate of0.001 and optimized using the categorical cross-entropy loss function. A ReduceLROnPlateau callback with a patience of five epochs was used to dynamically change the learning rate and maintain the top-performing model in order to improve training efficiency and avoid overfitting. With a test accuracy of 88.40% and a validation accuracy of 87.85%, the suggested model showed good generalization performance on unknown data.

4. EVALUATION METRICS

4.1 Confusion Matrix for Nepali Data

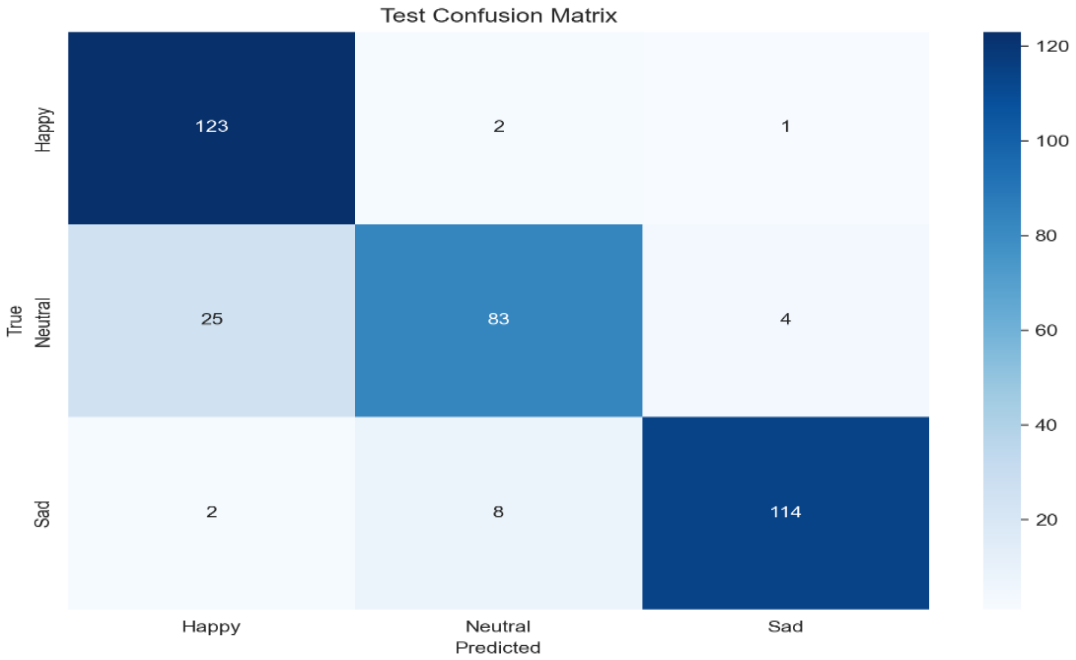

Figure 6. Confusion matrix showing classification performance of the proposed model across emotion classes.

The model's performance in three emotion classes—happy, neutral, and sad—is assessed in the classification report. With an F1-score of 0.89 and a recall of 0.98, the Happy class exhibits strong identification but some false positives because to confusion with Neutral. With a precision of 0.89 and a lower recall of 0.74, the Neutral class gets an F1-score of 0.81, showing problems with misclassification. This implies that neutral emotional states are harder for the model to recognize, maybe because of their less distinguishing auditory characteristics.

Table 1: Classification Report for Nepali Speech Emotion Detection

Class	Prescision	Recall	F1- score	Support
Happy	0.82	0.98	0.89	126
Neutral	0.89	0.74	0.81	112
Sad	0.96	0.92	0.94	124
Accuracy			0.88	362
Macro Avg	0.89	0.88	0.88	362
Weighted Avg	0.89	0.88	0.88	362

4.2 ROC Curve Analysis

Receiver Operating Characteristic (ROC) curves, which plot the True Positive Rate (TPR) against the False Positive Rate (FPR) to assess classification performance, were used to assess three emotion classification models: Happy, Neutral, and Sad.

While the Sad and Happy models both received an AUC of 0.99, showing outstanding performance, the Neutral model received an AUC of 0.97, suggesting high discriminative capacity. The Happy model performed the best overall out of all of them. The results verify that all three models function well, with the Happy model attaining the maximum discriminative power, since ROC curves nearer the upper-left corner and AUC values closer to 1.0 imply greater classification capability.

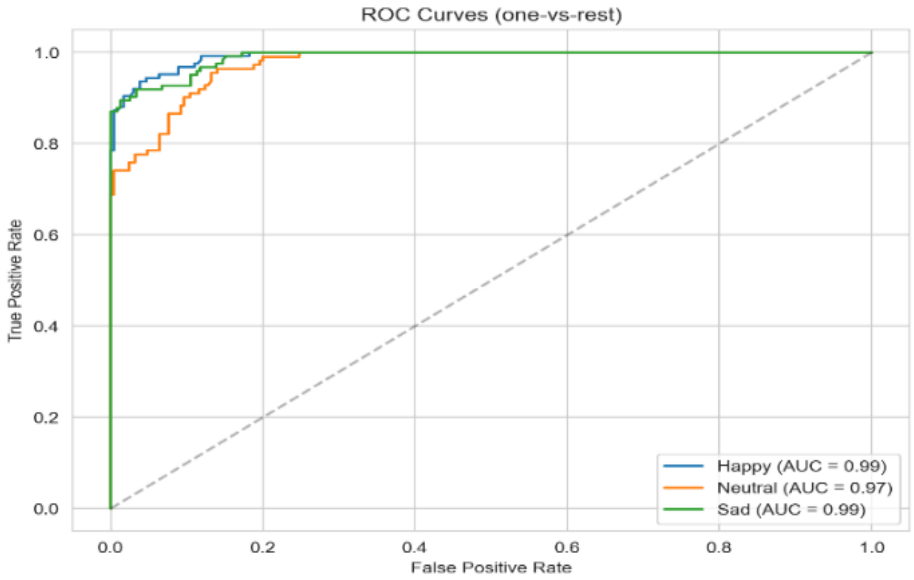

Figure 7. ROC curves showing classification performance of emotion classes.

5. CONCLUSION

Experimental results demonstrated that the proposed 1D-CNN model achieved an overall accuracy of 88%, indicating strong performance in recognizing emotional states from Nepali speech. Among the emotion classes, the Sad category achieved the best performance with a precision of 0.96, recall of 0.92, and F1-score of 0.94, while the Happy class showed the highest recall 0.98. ROC analysis further confirmed the robustness of the model, with AUC values of 0.97 for Neutral and 0.99 for both Happy and Sad emotions.

References

[1] K. Chowdhury, and J. Hassan R. S. Kotecha, "Speech emotion recognition with lightweight deep neural ensemble model using handcrafted features," Scientific Reports, vol. 15, April 2025.

[2] R. Jahangir, Q. Ain, A. Nauman, M. Uddin, and S. Ullah M. Akhtar, "Urduser: A comprehensive dataset for speech emotion recognition in Urdu language," Data in Brief, vol. 60, p. 111627, May 2025.

[3] A. Martínez-Rebollar, H. Estrada-Esquivel, E. Clemente, and O. A. Pliego-Martínez A. A. Colunga-Rodriguez, "Developing a dataset of audio features to classify emotions in speech," Computation, vol. 13, 2025.

[4] J. Shi, S. Zhang, L. Zhou, H. Liu, and Y. Tan W. Liu, "E-Speech: Development of a dataset for speech emotion recognition and analysis," International Journal of Intelligent Systems, 2024.

[5] U. Bayraktar, and A. Kucukmanisa Z. Kilimci, "Evaluating raw waveforms with deep learning frameworks for speech emotion recognition," Multimedia Tools and Applications, pp. 1-31, 2025.

[6] R. Oteniyazov, F. Makhmudov, and Y.-I. Cho I. Pulatov, "Enhancing speech emotion recognition using dual feature extraction encoders," Sensors, vol. 23, no. 14, 2023.

[7] G. Meena, and K. K. Mohbey R. R. Choudhary, "Speech emotion based features to classify emotions in speech," Computation, vol. 13, p. 39, 2025.

[8] P. M. D. R. Vincent, Q. Shambour, S. I. Mohammad, A. Vasudevan, E. E. H. Soon, and M. T. Alshurideh S. A. K. Basha, "Exploring deep learning methods for audio speech emotion detection: An ensemble of MFCCs, CNNs and LSTM," Applied Mathematics and Information

Sciences, vol. 19, pp. 75-85, Jan 2025.