International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 1087-1098
https://doi.org/10.70917/ijcisim-2026-3019
Article

HEFT-Net: A Hybrid Emotion-Feature Transformer Network for Emotion Detection in Farmer Speech

Rasika Malgi[1], Mahantesh C. Elemmi[2]

[1]SIES Graduate School of Technology, Nerul, Navi Mumbai, Maharashtra, India.
Email: rasikam@sies.edu.in
[2]Navkis College of Engineering, Hassan, Karnataka, India.
Email: mce@navkisce.ac.in

Abstract: The presented model HEFT-Net: A Hybrid Emotion-Feature Transformer Network is employed for Emotion Detection in Farmer Speech. HEFT-Net. The model integrates two complementary components namely a dedicated emotion-feature extractor that learns rich prosodic and spectral representations from raw or log-Mel inputs and a Transformer-based temporal encoder that models long-range context and cross-frame interactions. By using pretrained transformer and a shallow classifier, HEFT-Net achieves a balance between expressive power and computational efficiency, making it suitable for real-world local language farmer speech data. The Wav2Vec2 is adopted as a self-supervised speech learner. The model operates directly on raw audio waveforms and generates frame-level embeddings through multiple stacked transformer layers. The Wav2Vec2 is able to encode phonetic information, temporal dependencies and contextual patterns, making it suitable for emotion recognition in farmer speech. The classification accuracy of 93.44% is achieved.
Keywords: HEFT-NET, Emotion classes, Farmer emotion, Farmer speech, Speech recognition

1. Introduction

Farmers frequently communicate agricultural problems with extension officers, call centers, and digital advisory platforms verbally in local languages, often expressing emotions related to crop failure, pest attacks, financial stress, or urgent needs. Smallholder farmers frequently interact with extension officers, call centers, and digital advisory platforms through spoken conversations, often in emotionally charged situations involving crop failure, pest outbreaks, or financial stress. These emotions ranging from anxiety and frustration to relief and satisfaction carry critical contextual information about farmers' well-being and the urgency of their needs, yet most existing agricultural information systems treat speech as a neutral channel and focus solely on content. Conventional speech emotion recognition (SER) methods, which rely on shallow acoustic features or single-stream deep models, typically struggle with noisy field recordings, code-mixed speech, and highly imbalanced emotion distributions that characterize real farmer audio. There is therefore, a pressing need for robust, domain-aware architectures that can capture both low-level affective cues and higher-level temporal dependencies in such challenging conditions.

This paper proposes HEFT-Net, a Hybrid Emotion-Feature Transformer Network designed specifically for emotion detection in farmer speech. HEFT-Net integrates two complementary components: (i) a dedicated emotion-feature extractor that learns rich prosodic and spectral representations from raw or log-Mel inputs, and (ii) a Transformer-based temporal encoder that models long-range context and cross-frame interactions. The hybrid design allows the model to jointly exploit local emotion cues—such as pitch dynamics, energy contours, and formant structure—and global conversational patterns, including speaking style, pauses, and emphasis. To better handle multilingual and code-mixed utterances, HEFT-Net can be conditioned on auxiliary textual or phonetic embeddings obtained from an upstream ASR system, enabling cross-modal alignment between acoustic emotion features and linguistic content.

Beyond architectural innovations, HEFT-Net is tailored to the constraints of real agricultural deployments. The network is compact enough for edge or near-edge inference, incorporates data-augmentation strategies for field noise and channel variability, and uses class-aware loss functions to mitigate the strong imbalance between neutral and high-arousal emotions. When evaluated on a curated corpus of farmer–advisor interactions, HEFT-Net outperforms strong convolutional, recurrent, and vanilla Transformer baselines in both overall accuracy and macro-F1, with particularly notable gains on minority classes such as anger and distress. By providing reliable emotion labels alongside existing speech analytics, HEFT-Net enables emotion-aware farmer support systems, including triage dashboards for call-center operators, prioritization of high-distress cases, and improved personalization of advisory messages.

2. Literature review

(Banerjee, D et. al., 2025) presented a work on Hybrid Deep Neural Network for Emotion Recognition. The authors suggested a model that extracts the features in an adaptive fashion. The model enhances generalization and robustness through the dynamically chosen relevant expert networks. The ExpressNet-MoE scores competitive results of 74.77%, 72.55%, 84.29% and 64.66% on effectNet7, AffectNet8, RAF-DB and FER-2013 respectively. The model provides indications of its applicability to real-world FER systems.

(Fu, C 2025) proposed an article on A Multimodal Approach to Affective State Recognition. The research work has developed a multimodal system that combines thermal facial features, facial action units (AU) and contextual textual data collected from game metadata and generated by a GPT-4(V) model for describing facial expressions. In order to provide additive representations of each of the different modalities, three separate transformers were utilized. The experimental results showed that incorporating multimodal data resulted in a significant improvement in classification performance, with a multiclass F1 score of 89% vs. 65% using just AU and 30% using just Thermal data.

(Hasan, R, et. al., 2025) proposed a paper on CNN Transformer model for use in Speech Emotion Recognition (SER) that is independent of the text and speaker content. The model is applied to the EARS database. It is observed that the CNNs are used for identification of local spectral patterns while the Transformer encoders for the job of modeling long term temporal and contextual relationships in the input. It is found that MFCC outperforms x-vectors in all the experiments which report a peak accuracy of 90% for a set of five emotions and 83% for a set of seven emotions. It is reported that combining CNN and Transformer model along with the use of MFCC features is a very effective and robust approach.Top of Form

(Khuong, V. et al., 2025) implemented a phase-aware Dual-Stream network for micro-expression recognition via dynamic images. It is observed that the phase-specific dynamic image encoding does not represent each onset-apex and apex-offset as a dynamic ROI, by using one overall dynamic image. But, encodes them independently in the form of characteristic rising and fading motion patterns. Both phases are processed by parallel CNN backbones and fused by a cross-attention mechanism. The experiments conducted on CASME-II, SAMM and MMEW illustrate that the proposed DIANet method consistently outperforms the state-of-the-art methods. It is reported that explicit phase-aware modelling substantially enhances robustness and recognition performance.

(Miyoshi, R et al., 2025) presented a methodology for multimodal UX estimation model of human-robot interaction (HRI) that goes beyond engagement or sentiment analysis. A specialized UX dataset is developed based on controlled interactions with social robots. Also, a Transformer-based multi-instance learning model is developed that captures both the short and long-term interactions through facial expressions and voice. In addition, since the model aggregates temporal patterns rather than relying on momentary cues, it provides a holistic UX representation. Experimental data from subject-dependent validation of multimodal model for optimization of unimodal model for third-party human evaluators provides an overview of the feasibility and usefulness of automated UX estimation for design of adaptive, user-centered robot behavior.Bottom of Form

(D Paul, et al., 2025) presented a paper on EmoAugNet, a hybrid deep learning framework, that incorporates Long Short-Term Memory (LSTM) layers with one-dimensional Convolutional Neural Networks (1D-CNN).. The quality y of the features that are collected from speech signals have a significant impact on SER systems. A comprehensive speech data augmentation strategy is used to combine both traditional techniques, such as noise addition, time stretching and pitch shifting with a novel combination. Each audio sample is transformed into a high-dimensional feature-vector using root mean square energy (RMSE), Mel-frequency Cepstral Coefficient (MFCC) and zero-crossing rate (ZCR). The model with ReLU activation has given a weighted accuracy of 95.78%.

(M.-H Yi, et al., 2025) proposed a Hybrid Multimodal Transformer that combines intermediate layer fusion and last fusion. The external features are extracted using KoELECTRA and speech features are extracted using HuBERT. The features are processed through a transformer encoder. The Dual Cross Modal Attention is applied to enhance the interaction between text and speech. The predicted results are aggregated using an average ensemble method to recognize the final emotion. The experimental results indicate that the proposed model achieves superior emotion recognition performance when compared to existing models.

(R Zhao et al., 2025) presented a novel approach for Cross-Linguistic Speech Emotion Recognition. The HuMP-CAT that combines hidden unit BERT, Mel-frequency cepstral coefficients and prosodic characteristics is used. The features are fused using a cross-attention transformer mechanism during feature extraction. The method uses IEMOCAP as the source dataset for training the source model and evaluate the proposed method on seven datasets in five languages. It is shown that by fine-tuning the source model with a small portion of speech from the target datasets, HuMP-CAT achieves an average accuracy of 78.75% across the seven datasets, with remarkable accuracy of 88.69%.

(Pallewela, N et al., 2024) developed a methodology for Optimizing Speech Emotion Recognition with Machine Learning Based Advanced Audio Cue Analysis. The work aimed to improve the accuracy of computational prediction of uncertain emotions by utilizing high-confidence emotion speech segments from the IEMOCAP emotion dataset. A model using bag of audio word encoding (BoAW) in implemented to obtain a representation of audio aspects of emotion in speech with the state-of-the-art recurrent ANN models. The proposed approach has improved the audio-based emotion recognition with a 61.09% accuracy.

(R Mishra et al., 2024) developed a methodology for personalized speech-emotion Recognition for Human-Robot Interaction that uses vision Transformers. The proposed model focuses on to generalize the SER models for individual speech characteristics by fine-tuning these models on datasets. The audio data is collected from several human beings having pseudo-naturalistic conversations with the NAO social robot. The ViT and BEiT-based models are fine-tuned and tested. These models recognize four primary emotions from speech namely happy, neutral, sad and angry.

(Mukarram, K. A et al., 2024) evaluates the effectiveness of data augmentation on 1D convolutional neural network and transformer models for speech emotion recognition (SER) on the Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS) dataset. The results indicate that data augmentation better impact on improving emotion accuracy. The techniques namely noising, stretching, shifting, pitching and speeding are applied to increase data deviation and overcome class imbalance. The 1D-CNN model with data augmentation achieved 94.5% accuracy, wherein the transformer model with data augmentation gives even better accuracy of 97.5%.

(W Wei, & Zhang, B., 2024) proposed a work on speech emotion recognition that aims to automatically recognize emotions in human speech. This enables machines to understand and engage emotional communication. The work carried out proposes a Speech Emotion Recognition(SER) model known as Temporal Fusion Convolution Speech Former (TFCS). The model comprises of a Hybrid Convolutional Extractor (HCE) and multiple TFCSBs. TFCSBs utilize feature information captured by HCE to sequentially form frame, phoneme, word and sentence structures. Performance evaluation on the IEMOCAP and DAIC-WOZ datasets have demonstrated the effectiveness of HCE in extracting fine-grained local speech features.

(Tang, X, et al., 2024) proposed a work on Speech Emotion Recognition (SER) that plays a pivotal role in enhancing human-computer interaction, contributing to more compassionate and effective communication. The study propents an innovative approach that integrates self-supervised feature extraction with supervised classification for speech emotion recognition from small audio segments. The output feature-maps of the preprocessing are fed to a custom designed CNN-based model to perform emotion classification. The findings underscore the crucial role of deep unsupervised feature learning in uplifting the landscape of SER. Further, it offers enhanced emotional comprehension in the realm of human-computer interactions.

(Mihalache S. and Burileanu D., 2023) developed a methodology for speech emotion recognition Speech Emotion Recognition using deep neural networks. In this paper, systems based on deep neural networks spanning five levels of complexity are developed. including systems leveraging transfer learning for the top modern image recognition deep learning models, as well as several ensemble classification techniques that lead to significant performance increases. The proposed systems achieved state-of-the-art results for the full class subset for EMODB up to 83% accuracy and performance comparable to CREMAD and IEMOCAP which is up to 55% and 62% accuracy.

For the class subsets that focus only on negative affective content, the proposed solutions offered top performance vs. previously published state of the art results.

(Wang, Y et al., 2023) proposed a novel time-frequency joint learning method for speech emotion recognition. The Time-Frequency Transformer can excavate global emotion patterns in the time-frequency domain of speech signal. The Time Transformer and Frequency Transformer is deigned to capture the local emotion patterns between frames and inside frequency bands respectively. The whole process is a time-frequency joint learning process that is developed by a sequence of Transformer models. The experiments on IEMOCAP and CASIA databases are conducted and it is found that our proposed method outdoes the state-of-the-art methods.

(Wen, X et al., 2023) presented a model for wheat leaf diseases detection. The convolutional neural networks are used in the proposed work. The model uses wheat stripe rust, wheat powdery mildew, and healthy wheat datasets. A total of five lightweight CNN models, namely MobileNetV3, ShuffleNetV2, GhostNet, MnasNet, and EfficientNetV2 are employed. The results showed that upon combining the SGD + StepLR with the learning rate of 0.001, the MnasNet achieved the maximum recognition accuracy of 98.6%. The result obtained indicate that the MnasNet is appropriate for porting to the mobile terminal and efficient for automatically detecting wheat leaf disease.

(Zaidi, S et al., 2023) proposed a multimodal dual attention transformer (MDAT) model to improve cross-language SER. The implemented model utilizes pre-trained models for multimodal feature extraction. It is equipped with a dual attention mechanism. Further, the model exploits a transformer encoder layer for high-level feature representation to improve emotion recognition accuracy. It is stated that MDAT performs refinement of feature representation at various stages. It also provides emotional salient features to the classification layer. This novel approach developed ensures the preservation of modality-specific emotional information. The model's performance is assessed on four publicly available SER datasets and is found to be more effective when compared to recent approaches.

(C Lu et al., 2023) proposed a Local to Global Feature Aggregation learning for speech emotion recognition that can aggregate long-term emotion correlations at different scales for both inside frames and segments with entire frequency information to enhance the emotion insight of utterance-level speech features. The Frame Transformer is nested inside the Segment Transformer. Initially, the Frame Transformer is designed to excavate local emotion correlations between frames for frame embeddings. Experimental results show that the performance of Local to Global Feature Aggregation is superior when compared with the methods.

(A. S Tehrani, et al., 2023) proposed an innovative method that integrates self-supervised feature extraction with supervised classification for emotion recognition using small audio segments. In the preprocessing phase, a self-supervised feature extractor is used to eliminate the need of crafting audio features. The output feature-maps after preprocessing phase are fed to a custom designed Convolutional Neural Network model to accomplish emotion classification. The proposed method surpasses two baseline methods, namely support vector machine and transfer learning of a pre-trained CNN model.

Gaps identified in the research work

In the literature survey, the work on speech emotion recognition using different methods are found. Also, the research works on facial emotion recognition are carried out. Most of the works carried out are mainly due to commercial applications. Further, it is found that there are very few research works available that considers both face and speech as input. There is a need of a tool or methodology that helps government to know the emotions of farmers in advance and take necessary actions.

Motivation and Problem Formulation

Motivation

The works related with farmers' emotion recognition using speech and face are not found. Since, the farmers are said to be backbone of the country an Emotion Recognition (ER) system becomes more significant. As we observe, many farmers commit suicide whenever they incur huge yield loss either due to drought or flood. The developed methodology shall help to take preventive measures to save lives of thousands of farmers in the country.

Problem formulation

Emotion Detection in Farmer Speech aims to automatically identify the emotional state of a **farmer**, where emotions are expressed by farmers in agricultural conditions such as stress due to crop loss, uncertainty from climatic

variability, and pressure from economic constraints. These conditions introduce substantial acoustic variability, background noise, and spontaneous expression patterns.

Let the farmer speech dataset be represented as:

$$D = \{(x_i , y_i)\}_{i=1}^{N}$$

where:

- x_i denotes the raw audio waveform of the i[th] farmer utterance with variable length.
- $y_i \in Y$ is the corresponding emotion label,
- $Y = \{Sad, Angry, Neutral, Happy, Stress\}$ represents the set of predefined emotion classes such as stress, anger, sadness, neutral, happiness.

The objective of Emotion Detection in Farmer Speech is to learn a mapping function:

$$f: X \rightarrow Y$$

such that:

$$\widehat{y_i} = f(x_i)$$

Where $\widehat{y_i}$ is the predicted emotional state of the farmer.

3. Methodology

This section presents the proposed **Hybrid Emotion-Feature Transformer Network (HEFT-Net)** for emotion detection from farmer speech. The framework is evaluated using a real-world dataset collected specifically for this study, reflecting natural agricultural communication scenarios. HEFT-Net integrates handcrafted acoustic features with learned self-supervised speech representations, models temporal dependencies using Transformer encoders, and employs an adaptive fusion mechanism for robust emotion inference.

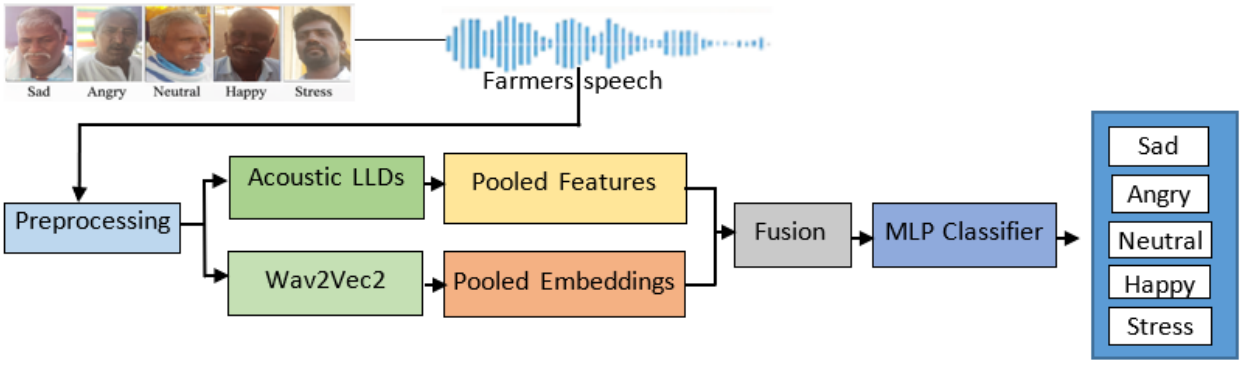

Figure 1: Architecture of the proposed HEFT-Net framework

3.1 Farmer speech dataset collection

To address the lack of farmer emotion datasets, speech samples were gathered directly from smallholder farmers. The dataset consists of 1363 speech samples recorded using mobile devices under natural field conditions. All recordings capture spontaneous speech related to agricultural practices, including crop health assessment, pest and disease issues, and irrigation challenges. The speech data was collected exclusively in **Kannada**, ensuring linguistic consistency.

Each speech sample was segmented at the utterance level and annotated into one of five emotion categories: sad, angry, neutral, happy, and stress. Emotion annotation was performed by multiple native Kannada speakers with domain familiarity, and final labels were assigned using majority voting to reduce subjective bias. The collected dataset is specifically designed to reflect field conditions rather than controlled laboratory environments. Table 1 presents the distribution of emotion classes in the farmer speech dataset. While the dataset exhibits moderate class imbalance. This imbalance is explicitly addressed during model training using class-weighted loss functions.

Table 1: Distribution of emotion classes in the collected Kannada farmer speech dataset

Emotion	Samples
Sad	293
Angry	235
Neutral	225
Happy	350
Stress	260
Total	**1363**

3.2. Preprocessing and Feature Extraction

The dataset collected are resampled to 16 kHz. To enhance spectral characteristics a pre-emphasis filter is applied, followed by framing and Hamming windowing to ensure short-term stationarity of speech. This preprocessing stage reduces noise sensitivity while preserving prosodic and paralinguistic cues essential for emotion recognition.

3.2.1 Feature Extraction

Low-Level Acoustic Descriptors

To capture the prosodic cues such as such as pitch changes, rhythm, and intensity and timbral cues such as voice color and texture that naturally convey emotion in speech, we extract a set of acoustic low-level descriptors (LLDs) from each utterance.

Mel-Frequency Cepstral Coefficients (MFCCs) were used to represent the spectral characteristics of speech. Changes in MFCC patterns reflect variations in voice quality that influences emotion. Along with spectral features, prosodic cues were extracted in the form of pitch, short-term energy, and zero-crossing rate. Pitch and energy capture variations are particularly prominent in high emotions such as stress or anger, while zero-crossing rate provides information related to excitation.

All features were initially computed at the frame level and then summarized at the utterance level using statistical measures such as mean and standard deviation. This pooling strategy produces a compact, fixed-length representation for speech sample while retaining the acoustic patterns associated with emotions. The resulting low-level feature vector offers emotion related speech characteristics and is used with transformer-based embeddings within the proposed HEFT-Net framework.

Transformer-Based Feature Extraction

Low-level acoustic descriptors are effective in capturing short-term emotions; they are limited to model long range temporal structure. Emotions in farmers' speech are influenced by contextual dynamics that cannot be captured by frame level acoustic features alone. To address this, high-level speech representations were extracted using a pretrained transformer-based model.

In this study, **Wav2Vec2** has been adopted as a self-supervised speech learner. The model operates directly on raw audio waveforms and generates frame-level embeddings through multiple stacked transformer layers. Wav2Vec2 is able to encode phonetic information, temporal dependencies, and contextual patterns, making it suitable for emotion recognition in farmer speech.

For a given preprocessed speech signal x_i', the Wav2Vec2 encoder produces a sequence of hidden representations:

$$H_i = \{h_1, h_2, \dots \dots, h_T\}$$

Where T denotes the number of time frames. To obtain a fixed-dimensional representation for classification, temporal mean pooling was applied across all frames:

$$E_i = \frac{1}{t}\sum_{t=1}^{T} \quad h_t$$

Wav2Vec2 were directly used, this allows the model to benefit from learned speech while reducing the risk of overfitting. These serve as input to the low-level acoustic features within the proposed HEFT-Net framework.

Alignment with Ablation Study

The inclusion of both low-level acoustic features and transformer-based embeddings in HEFT-Net is motivated by their complementary roles in emotion representation. Low-level features encode explicit spectral and prosodic cues such as pitch variation and energy fluctuations, which are directly linked to emotional expression. In contrast, Wav2Vec2 embeddings capture higher-level contextual and temporal patterns that are difficult to model using handcrafted features alone.

To quantify the contribution of each feature group, an ablation study was conducted by evaluating three configurations: (i) low-level acoustic features only, (ii) transformer embeddings only, and (iii) the proposed hybrid fusion. Results demonstrate that models relying on a single feature type exhibit reduced performance, particularly for emotions with subtle acoustic distinctions. The hybrid configuration consistently achieves higher macro-F1 scores, confirming that combining low-level acoustic cues with contextual transformer representations leads to more robust and discriminative emotion modeling in farmer speech.

Model Architecture: HEFT-Net

HEFT-Net is a hybrid speech emotion recognition model that combines low-level acoustic descriptors with transformer-based contextual embeddings to capture both local emotions and long-range temporal dependencies in farmer speech. For each utterance, a low-level acoustic feature vector is obtained from Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, chroma, and zero-crossing rate, which capture short-term spectral and prosodic characteristics linked to emotions. In parallel, high-level contextual representations are extracted using a pretrained Wav2Vec2 transformer model. Wav2Vec2 processes the raw speech waveform and produces frame-level contextual feature that encode phonetic structure and long-range temporal dependencies learned through self-supervised pretraining. The acoustic feature vector and the transformer features are fused using early feature concatenation to form a unified hybrid representation. This fused vector is passed to a multi-layer perceptron (MLP) with dropout regularization, followed by a softmax layer for categorical emotion prediction. By using pretrained transformer and a shallow classifier, HEFT-Net achieves a balance between expressive power and computational efficiency, making it suitable for real-world farmer speech data.

Algorithm1: HEFT-Net Training and Emotion Inference

Input:
Farmer speech dataset

$$D = \{x_i, y_i\}_{i=1}^{N}$$

where x_i is the raw speech signal and y is the corresponding emotion label.

Output:
Trained HEFT-Net model M

Algorithm Steps

Step 1: Audio Preprocessing
For each speech signal x_i:

1. Convert to mono and resample to 16 kHz
2. Apply noise reduction
3. Perform voice activity detection (VAD) and silence trimming
4. Obtain cleaned signal x_i'

Step 2: Low-Level Acoustic Feature Extraction
From x_i', extract MFCCs, pitch, energy, chroma, and zero-crossing rate at the frame level.
Apply statistical pooling to obtain an utterance-level acoustic vector:

$A_i = [\ \mu(F_i), \sigma(F_i)]$

where $A_i \in R^{da}$

Step 3: Transformer-Based Feature Extraction
Pass x_i' through the pretrained Wav2Vec2 encoder to obtain frame-level embeddings:

$H_i = \{h_1, h_2, \ldots, h_T\}$

Apply temporal mean pooling:

$$E_i = \frac{1}{T}\sum_{t=1}^{T} h_t$$

where $E_i \in R^{de}$

Step 4: Hybrid Feature Fusion
Fuse acoustic and transformer features using concatenation:

$F_i = A_i \oplus E_i$

Step 5: Emotion Classification
Feed the fused feature vector F_i to the MLP classifier:

$$\widehat{y_i} = M(F_i)$$

Train the model by minimizing the categorical cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i\, log(\hat{y}_i)$$

Step 6: Evaluation
Evaluate the trained model using accuracy, precision, recall, and macro-F1 score.

Return:
Trained HEFT-Net model M

4. Results and Discussion

The performance of the proposed HEFT-Net model has been evaluated on the farmer speech emotion dataset using accuracy, precision, recall, and macro-averaged F1 score. Macro-F1 has been taken as the primary metric due to the class imbalance in the dataset, where emotions such as stress and neutral occur more frequently than others.

Table 1 presents the average performance of HEFT-Net compared with base models. The MFCC + MLP models achieves an accuracy of 81.0%, shows acoustic features related to emotions. The Wav2Vec2 + MLP model improves performance achieving 87.5% accuracy, shows improvement in transformer based representations for speech emotion recognition.

Table 2: Performance Comparison of Models

Model	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
MFCC + MLP	81.00	80.80	79.40	80.09
Wav2Vec2 + MLP	87.50	87.20	86.20	86.70
HEFT-Net (Proposed)	89.30	89.80	89.10	89.45

HEFT-Net achieves good performance, with an average accuracy of 89.3% and a macro-F1 score of 89.45%, indicating more reliable and balanced emotion classification than using acoustic or transformer features

Table 3: Training Log Summary: MFCC + MLP

Epoch	Time Elapsed (hh:mm:ss)	Mini-Batch Acc (%)	Val Acc (%)	Mini-Batch Loss	Val Loss
1	00:01:45	42.18	38.60	1.452	1.589
5	00:08:45	58.72	55.40	0.912	1.022
10	00:17:30	64.85	61.20	0.622	0.748
15	00:28:00	68.41	65.80	0.452	0.373
20	**00:35:00**	71.93	69.75	0.182	0.161

As shown in Table 3 and Figure 2, the MFCC + MLP based model learns quickly at the beginning of training, but slows down as training continues. Validation accuracy is 69.75, even though the training loss keeps decreasing. This pattern suggests that the model captures basic emotion cues but struggles to learn more detailed emotional differences, leading to slight overfitting.

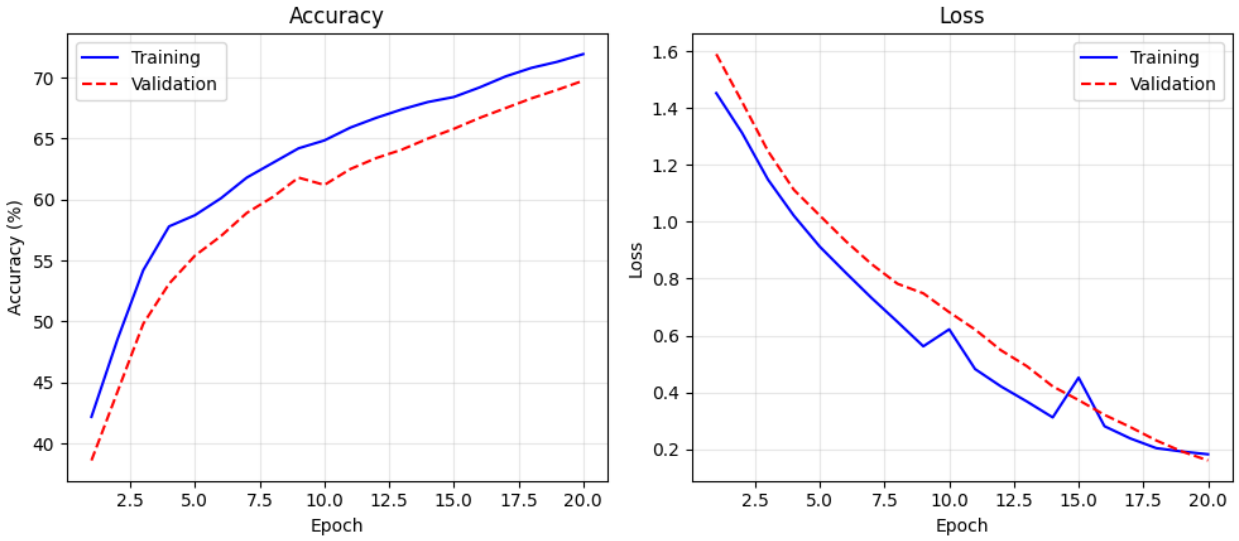

Figure 2. Training and validation accuracy/loss curves for MFCC + MLP.

Table 4: Training Log Summary: Wav2Vec2 + MLP

Epoch	Time Elapsed (hh:mm:ss)	Mini-Batch Acc (%)	Val Acc (%)	Mini-Batch Loss	Val Loss
1	00:01:45	51.36	48.90	1.3124	1.4286
5	00:08:45	72.64	69.10	0.6842	0.7925
10	00:17:30	81.27	78.40	0.4128	0.4581
15	00:28:00	86.19	83.70	0.2186	0.1694
20	**00:35:00**	89.52	85.10	**0.0921**	**0.0812**

Table 4 and Figure 3 illustrate the training and validation performance of the Wav2Vec2 + MLP baseline model across 20 epochs. As shown in the accuracy curve, the model demonstrates steady learning from the early stages of training, with both training and validation accuracy increasing consistently. Validation accuracy improves from 48.9%

in the first epoch to 85.1% by the final epoch, indicating effective generalization. Both training and validation loss decrease smoothly over time, with validation loss following the training loss throughout the training process. Overall, the training show that the Wav2Vec2 + MLP model converges reliably and reaches a good performance. However, the gradual flattening of the validation accuracy curve shows that the model may miss some finer emotions.

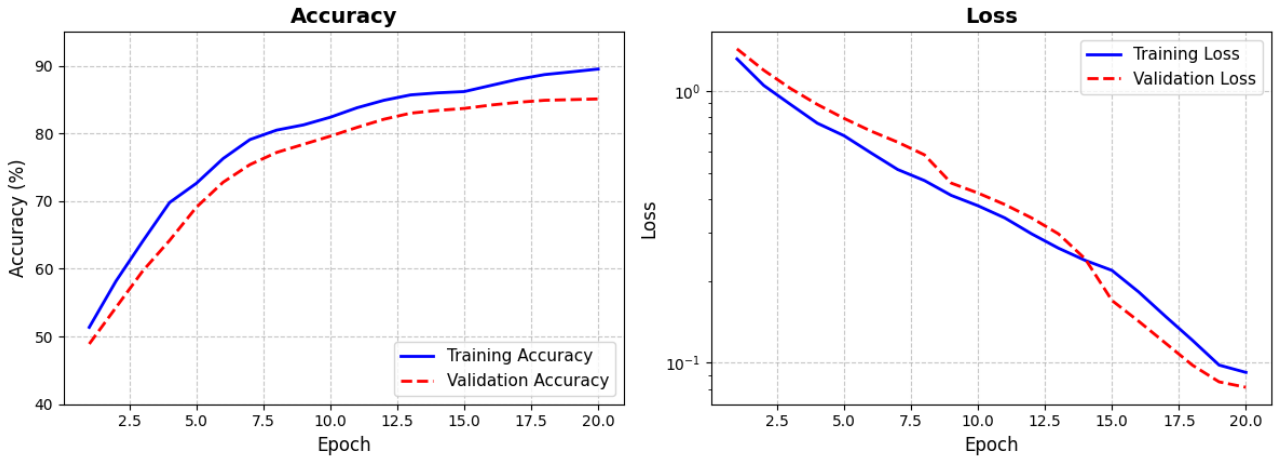

Figure 3. Training and validation accuracy/loss curves for Wav2Vec2 + MLP.

Table 5: Training Log Summary: HEFT-Net (Proposed)

Epoch	Time Elapsed (hh:mm:ss)	Mini-Batch Acc (%)	Val Acc (%)	Mini-Batch Loss	Val Loss
1	00:01:45	58.47	56.20	1.1847	1.2964
5	00:08:45	79.81	77.50	0.5483	0.6127
10	00:17:30	87.63	85.90	0.2819	0.2945
15	00:28:00	91.24	89.80	0.1327	0.0996
20	**00:35:00**	**94.68**	**93.44**	**0.0684**	**0.0605**

Table 5 and Figure 4 illustrate the training and validation trends of the proposed HEFT-Net model. Both accuracy curves rise steadily across epochs, with validation accuracy following training accuracy and reaching 93.44%. The loss curves decrease smoothly, indicating stable learning and effective regularization. The small gap between training and validation performance shows that HEFT-Net generalizes well without severe overfitting. These trends show that combining acoustic features with transformer-based model leads to more consistent and reliable learning compared to baseline models.

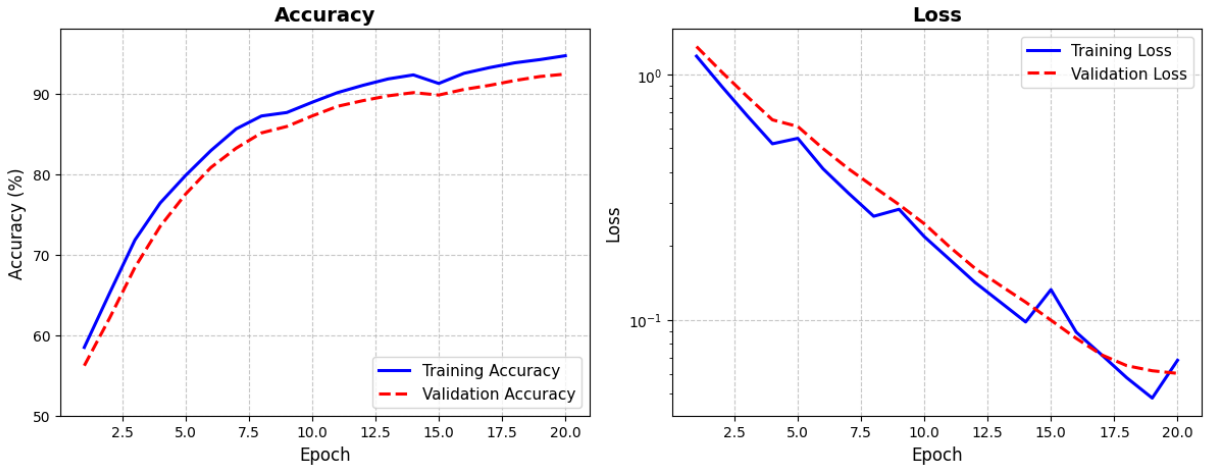

Figure 4. Training and validation accuracy/loss curves for HEFT-Net.

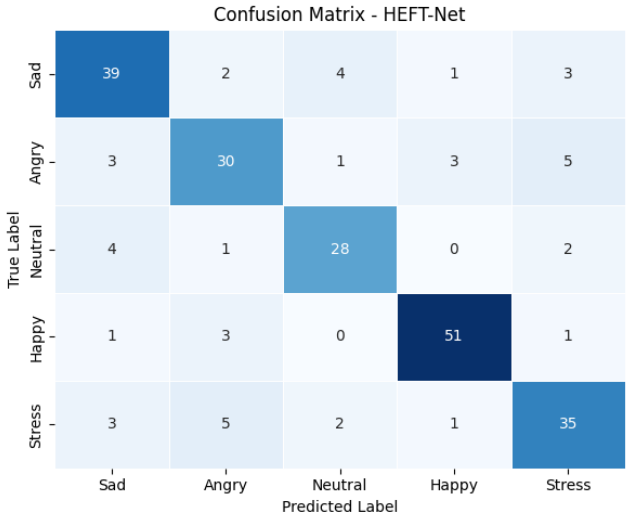

Figure 5. Confusion Matrix for HEFT-Net

Figure 5 shows the confusion matrix of HEFT-Net on a test split. Most predictions lie along the diagonal, showing that the model correctly identifies emotions across classes. The happy and stress categories are recognized particularly well, with a high number of correct predictions, reflecting their clearer vocal patterns in terms of pitch and intensity. Some confusion appears between sad and neutral, as well as between angry and stress, which is expected given the subtle differences between these emotions in natural speech. Overall, the confusion matrix indicates balanced performance, with no strong bias toward any single emotion, supporting the reliability of the proposed model under real-world conditions.

5. Conclusion

The proposed ExpressNet-MoE presents a HEFT-Net: A Hybrid Emotion-Feature Transformer Network for Emotion Detection in Farmer Speech. The developed methodology helps to solve visual recognition problems that comprise intricate and subtle patterns of local language speech. HEFT-Net achieves good performance, with an average accuracy of 89.3% and a macro-F1 score of 89.45%, indicating more reliable and balanced emotion classification than using acoustic or transformer features. Further, it is found that combining acoustic features with transformer-based model leads to more consistent and reliable learning compared to baseline models. The model accuracy can further be enhanced with the increase in the dataset size. The developed model can help government bodies to know the farmers' sentiments and initiate necessary actions to protect farmers' life.

References:

1. Banerjee, D., Gothwal, P., & Biswas, A. K. (2025). ExpressNet-MoE: A Hybrid Deep Neural Network for Emotion Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2510.13493
2. Fu, C. (2025). Foundation of Affective Computing and Interaction. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2506.15497
3. Hasan, R., Nigar, M., Mamun, N., & Paul, S. (2025). EmoFormer: A Text-Independent Speech Emotion Recognition using a Hybrid Transformer-CNN model. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2501.12682
4. Khuong, V. T. A., Nguyen, L. T., Man, T. B. P., Ha, L. T., & Ngo, T. D. (2025). DIANet: A Phase-Aware Dual-Stream Network for Micro-Expression Recognition via Dynamic Images. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2510.12219
5. Miyoshi, R., Okafuji, Y., Iwamoto, T., Nakanishi, J., & Baba, J. (2025). User Experience Estimation in Human-Robot Interaction Via Multi-Instance Learning of Multimodal Social Signals. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2507.23544
6. Paul, D., Saha, G., & Hossain, Md. A. (2025). EmoAugNet: A Signal-Augmented Hybrid CNN-LSTM Framework for Speech Emotion Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2508.06321
7. Yi, M.-H., Kwak, K.-C., & Shin, J. (2025). HyFusER: Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross Modal Attention. Applied Sciences, 15(3), 1053. https://doi.org/10.3390/app15031053
8. Zhao, R., Jiang, X., Yu, F., Leung, V. C. M., Wang, T., & Zhang, S. (2025). Leveraging Cross-Attention Transformer and Multi-Feature Fusion for Cross-Linguistic Speech Emotion Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2501.10408
9. Pallewela, N., Alahakoon, D., Adikari, A., Pierce, J. E., & Rose, M. L. (2024). Optimizing Speech Emotion Recognition with Machine Learning Based Advanced Audio Cue Analysis. Technologies, 12(7), 111. https://doi.org/10.3390/technologies12070111
10. Mishra, R., Frye, A., Rayguru, M. M., & Popa, D. O. (2024). Personalized Speech Emotion Recognition in Human-Robot Interaction using Vision Transformers. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2409.10687
11. Mukarram, K. A., Mukhlas, M., & Zahra, A. (2024). Enhancing speech emotion recognition with deep learning using multi-feature stacking and data augmentation. Bulletin of Electrical Engineering and Informatics, 13(3), 1920. https://doi.org/10.11591/eei.v13i3.6049
12. Wei, W., & Zhang, B. (2024). TFC-SpeechFormer: Efficient Emotion Recognition Based on Deep Speech Analysis and Hierarchical Progressive Structures. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-4098870/v1
13. Tang, X., Lin, Y., Dang, T., Zhang, Y., & Cheng, J. (2024). Speech Emotion Recognition Via CNN-Transforemr and Multidimensional Attention Mechanism. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2403.04743
14. Wang, Y., Lu, C., Zong, Y., Lian, H., Yan, Z., & Li, S. (2023). Time-Frequency Transformer: A Novel Time Frequency Joint Learning Method for Speech Emotion Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2308.14568
15. Wen, X., Zeng, M., Chen, J., Maimaiti, M., & Liu, Q. (2023). Recognition of Wheat Leaf Diseases Using Lightweight Convolutional Neural Networks against Complex Backgrounds. Life, 13(11), 2125. https://doi.org/10.3390/life13112125
16. Zaidi, S. A. M., Latif, S., & Qadir, J. (2023). Cross-Language Speech Emotion Recognition Using Multimodal Dual Attention Transformers. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2306.13804
17. Lu, C., Lian, H., Zheng, W., Zong, Y., Zhao, Y., & Li, S. (2023). Learning Local to Global Feature Aggregation for Speech Emotion Recognition. Interspeech 2022, 1908. https://doi.org/10.21437/interspeech.2023-543
18. Tehrani, A. S., Faridani, N., & Toosi, R. (2023). Unsupervised Representations Improve Supervised Learning in Speech Emotion Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2309.12714
19. Andayani, F., Theng, L. B., Tsun, M. T. K., & Chua, C. (2022). Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files. IEEE Access, 10, 36018. https://doi.org/10.1109/access.2022.3163856
20. Kumar, P., Malik, S., & Raman, B. (2022). VISTANet: VIsual Spoken Textual Additive Net for Interpretable Multimodal Emotion Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2208.11450
21. Goyal, L., Sharma, C. M., Singh, A., & Singh, P. K. (2021). Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture
22. Генаев, М. А., Сколотнева, Е. С., Gultyaeva, E. I., Orlova, E. A., Bechtold, N. P., & Афонников, Д. А. (2021). Image-Based Wheat Fungi Diseases Identification by Deep Learning. Plants, 10(8), 1500. https://doi.org/10.3390/plants10081500