I.J. Information Technology and Computer Science, **2025, 6, 29-51**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2025.06.02

Enhanced Deep Learning Framework for Tamil Slang Classification with Multi-task Learning and Attention Mechanisms

Ramkumar. R.*
Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankovil, TamilNadu, India
E-mail: heiram85@gmail.com
ORCID iD: https://orcid.org/0009-0006-5300-5181
*Corresponding author

Sureshkumar Nagarajan
Department of CSE, School of Computing, Kalasalingam Academy of Research and Education, Krishnankovil, TamilNadu, India
E-mail: sureshkumar@klu.ac.in
ORCID iD: https://orcid.org/0000-0001-9484-4965

Dinesh Prasanth Ganapathi
Centre for Data Science and Artificial Intelligence, University of Massachusetts, 740 N. Pleasant St., A205, Amherst, MA 01003, USA
E-mail: Dganapathi@umass.edu
ORCID iD: https://orcid.org/0009-0005-3273-7132

Received: 25 June 2025; Revised: 17 August 2025; Accepted: 20 October 2025; Published: 08 December 2025

Abstract: In Artificial Intelligence, voice categorization is important for various applications. Tamil, being one of the oldest languages in the world, comprises rich regional slang differing in tone, pronunciation, and emotive expression. These slang words are difficult to categorize because they are informal and there is limited annotated audio data. This study proposes an enhanced deep learning framework for Tamil slang classification using a balanced audio corpus. The framework integrates data-specific pre-processing techniques, including Mel spectrograms, Chroma features and spectral contrast, to capture the nuanced characteristics of Tamil speech. A DenseNet backbone, combined with LSTM and GRU layers, models both temporal and spectral information. The suggested FRAE-PSA module is an innovative application of the Pyramid Split Attention (PSA) mechanism adapted to support regional and affective variations of speech. Different from current PSA or Transformer-based approaches, FRAE-PSA splits the audio frequency spectrum and adapts attention weights dynamically based on auxiliary tasks. A multi-branch architecture is employed to fuse temporal and spectral features effectively and multi-task learning is used to enhance regional accent and emotion detection. Custom loss functions and lightweight networks optimize model efficiency. Experimental results show up to a 15% improvement in classification accuracy over baseline models, demonstrating the framework's effectiveness for real-world Tamil slang classification tasks.

Index Terms: Tamil Slang Classification, FRAE-PSA Module, Multi-Task Learning, Acoustic Feature Fusion, Regional Accent Detection and Emotion Detection, Pyramid Split Attention (PSA), Industry, Innovation and Infrastructure.

1. Introduction

Tamil slang classification is a challenging issue because colloquialisms employed in various parts of Tamil Nadu are informal, dynamic and highly contextual. Slangs may involve non-standard pronunciations, fast speech styles, emotional intonation and overlapping phonetic content. For instance, the same slang term can be pronounced differently in Chennai and Madurai based on regional accent as well as emotion. These considerations render it challenging for standard models to distinguish between slang words with slight acoustic variations.

Problem Definition:

The primary difficulty with Tamil slang classification is that existing models lack the ability to recognize regional and emotional differences within short, informal speech utterances. Conventional models usually employ global audio features such as MFCCs or Mel spectrograms without taking into account localized spectral patterns or context cues. This makes them incapable of differentiating between slang phrases that vary in accent or emotion only resulting in low performance particularly in low-resource or real-world noisy conditions.

Despite advances in speech processing and accent classification, existing methods fail to fully capture the intricate linguistic features specific to Tamil slang. Most deep learning-based approaches rely heavily on generic features such as spectrograms or Mel-frequency cepstral coefficients (MFCCs), which do not adequately account for the tonal, prosodic, and regional characteristics that distinguish Tamil slang from formal Tamil speech. As a result, these methods often exhibit suboptimal performance, especially when dealing with short-duration, low-resource speech samples commonly found in real-world data, such as movies, regional conversations and social media content.

To address these challenges, this study introduces a novel deep learning framework specifically designed for Tamil slang classification. The proposed method incorporates data-specific pre-processing techniques to capture the unique tonal and prosodic characteristics of Tamil speech, including pitch variation and regional accent features. Furthermore, a combination of Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) layers is used to model the temporal dynamics of speech, enabling the framework to effectively capture context and sequence-dependent information. This is complemented by a modified Pyramid Split Attention (PSA) module that focuses on the acoustic features most relevant to Tamil slang.

Unlike traditional models, our framework employs a multi-task learning approach, simultaneously tackling regional accent classification and emotion detection as auxiliary tasks. This enhances the model's ability to distinguish between similar-sounding slang terms and improves its generalization across various categories. Additionally, custom loss functions such as contrastive loss and focal loss are incorporated to handle class imbalance and refine the model's decision-making process.

The proposed approach significantly outperforms baseline models, achieving up to a 15% improvement in classification accuracy. By leveraging advanced deep learning techniques like attention mechanisms [1] and multi-task learning, the framework not only enhances classification performance but also demonstrates superior generalization, making it a promising solution for real-world Tamil slang classification tasks.

Unlike existing attention mechanisms such as Pyramid Split Attention (PSA) [2] or Transformer-based attention [3] which predominantly focus on spatial and channel-wise feature enhancement, the proposed FRAE-PSA module is explicitly designed to adapt to the unique characteristics of Tamil slang. By dividing the frequency spectrum into task-relevant bands (low, mid, and high frequencies), FRAE-PSA assigns dynamic attention weights guided by auxiliary tasks like regional accent and emotion detection. This targeted approach ensures that critical acoustic cues—essential for distinguishing Tamil slang—are prioritized, overcoming the limitations of generic attention mechanisms [4] that fail to address regional and tonal nuances. Furthermore, the integration of Spectral Energy Ratio (SER) computation refines the feature space, capturing subtle variations that are often overlooked by existing approaches.

2. Related Works

Recent advances in accent recognition, dialect identification, and emotion analysis [5] have leveraged deep learning, multi-task learning, and attention mechanisms. These studies provide a foundation for this work, focusing on their strengths and identifying areas for improvement.

2.1. Accent and Dialect Recognition

Multi-task learning and attention mechanisms have shown promise in accent and dialect classification. [2] introduced MPSA-DenseNet, a model integrating Pyramid Split Attention with DenseNet, achieving high accuracy in English accent classification. However, its computational demands limited deployment feasibility. Similarly, [6] employed a hybrid deep learning model for Arabic dialect classification, incorporating attention mechanisms. Despite significant performance gains, this model focused primarily on text rather than speech-based inputs, restricting its application to audio-centric tasks. [7] tackled accented speech challenges using Variational Mode Decomposition (VMD) and machine learning models, achieving robust results. However, its reliance on generic feature extraction methods such as MFCC did not address the intricate tonal and regional nuances critical for Tamil slang classification.

Current state-of-the-art accent classification models like AccentFold [8] and Audio-based Transformer models [9] have been found to work well for English and African languages. Yet, these models heavily depend on pre-trained ASR and fail to adjust to tonal and emotional differences occurring in slang. Although AccentFold uses zero-shot adaptation, it does not explicitly model regional phonetic information important in Tamil slang. In the same way, speech-based sentiment analysis models like MM DialogueGAT [10] and M3SA [11] emphasize multimodal sentiment but don't model region and emotion simultaneously in acoustic features.

Conversely, our FRAE-PSA-based model specifically segments the spectrum into task-specific bands and adaptively modulates accent and emotional attention weights, a tactic not adopted by state-of-the-art speech-centric models. Through

doing this, our framework more effectively captures slang-specific nuances in Tamil, an underrepresented language in existing literature.

2.2. Sentiment and Emotion Recognition

Sentiment analysis [12] and emotion recognition have also benefited from advanced deep learning techniques. [11] proposed the M3SA model, which utilized multi-scale feature extraction and attention mechanisms for multimodal sentiment analysis [13]. Although effective, its dependence on complete data from all modalities restricted its practical utility. [3] adopted a transformer-based multi-task learning framework for sentiment and emotion classification in crisis tweets, integrating subject intent prediction for deeper context understanding. However, its focus on textual and intent-based analysis limited its applicability to speech-based scenarios.

In speech emotion recognition, [14] employed LSTM and soft decision trees in a multi-task framework, enhancing feature representations and achieving robust results even in noisy environments. Similarly, [15] demonstrated the benefits of attention-enhanced CNNs [16] for emotion recognition, achieving state-of-the-art results. However, these studies did not address regional variations or their influence on speech characteristics.

Existing attention mechanisms, such as Pyramid Split Attention (PSA) [2] and Transformer-based attention [3], have been effective in accent and emotion detection tasks. However, these methods exhibit limitations when applied to speech-specific tasks like Tamil slang classification. PSA focuses primarily on spatial and channel-wise attention, which is less effective for capturing nuanced frequency-specific variations critical to regional accents. Transformer-based attention mechanisms, while robust, incur high computational costs and lack adaptation to frequency-band-specific features.

The proposed FRAE-PSA module addresses these gaps by introducing a frequency-wise adaptive mechanism tailored for Tamil slang classification. Unlike PSA, FRAE-PSA splits the frequency spectrum into low, mid, and high bands, dynamically adjusting attention weights based on the task relevance of each band. This ensures that regional and emotional variations—key to distinguishing Tamil slang—are effectively modelled. Additionally, the inclusion of Spectral Energy Ratio (SER) computation enhances feature extraction, enabling the module to capture both temporal and spectral nuances with superior accuracy and efficiency.

2.3. Low-resource Speech and Linguistic Challenges

Works addressing linguistic and resource constraints have emphasized tailored approaches. [17] proposed the PI-Whisper framework, enhancing ASR adaptability and inclusivity but without a focus on emotional or regional aspects. [18] tackled Cantonese speech emotion recognition with a dedicated dataset and feature set, achieving notable results but highlighting limitations in dataset diversity and emotion range.

3. Research Gaps and Novelty

3.1. Research Gap

- Focus on Regional and Emotional Features: Few models [14, 15] address the joint impact of regional and emotional variations on speech classification.
- Low-Resource Language Representation: Existing models lack tailored solutions for underrepresented languages like Tamil.
- Temporal and Spectral Dynamics: Although utilized in emotion recognition [15], these features are underexplored in accent and slang classification tasks.
- Speech-Specific Auxiliary Tasks: While multi-task learning has been employed [11, 3], auxiliary tasks such as regional accent and emotion detection have not been combined effectively for slang classification.

3.2. Novelty of the Proposed Work

The proposed Enhanced Deep Learning Framework for Tamil Slang Classification introduces the following innovations:

- FRAE-PSA Module: Focuses on frequency-wise regional accent and emotion adaptive attention for enhanced feature extraction.
- Hybrid Architecture: Combines DenseNet, LSTM, and GRU to capture temporal and spectral nuances in speech.
- Multi-Task Learning: Simultaneously handles regional accent and emotion detection to enrich feature representations and improve generalization.
- Real-World Deployability: Optimized for resource-constrained environments through lightweight architectures and model compression techniques.
- Low-Resource Language Adaptability: Tailored for Tamil slang, addressing unique phonetic and tonal challenges, with scalability to other low-resource languages.

By addressing these gaps, this study not only advances Tamil slang classification but also provides a scalable framework for similar linguistic challenges in underrepresented languages.

4. Methodology

This section outlines the enhanced methodology for Tamil slang classification, incorporating the expanded dataset and optimized model architecture. It details the data collection, pre-processing pipeline, model design, multi-task learning approach and loss functions.

4.1. Data Collection and Pre-processing

To train a robust classifier for Tamil slang the dataset contains a total of 900 audio samples that maintains a balanced distribution between slangs and gender. The dataset consists of voice samples collected from four regional dialects such as Chennai, Kovai, Madurai, and Nellai. Each sample varies in length from 5 to 15 seconds and is in .wav format that is suitable for processing by deep learning models.

A. Dataset Collection

The dataset (see Table 1.) includes 530 male and 370 female voices, distributed across the four Tamil slangs as follows:

Table 1. Tamil audio dataset

S.No	Tamil Slangs	Male Voice	Female Voice	Total
1	Chennai	120	100	220
2	Kovai	160	80	240
3	Madurai	100	120	220
4	Nellai	150	70	220
Total		530	370	900

Slang usage varies significantly across regions and groups, which can lead to datasets being biased if not carefully curated. This bias can lead to the model overfitting majority classes, misclassifying minority class instances and not learning fine acoustic features. To avoid this, we ensured that there was a balanced contribution from four large Tamil slang regions (Chennai, Kovai, Madurai, Nellai) and gender-balanced splits between male and female speakers (see Table 1). This controlled balance minimizes the risk of model bias and enables better generalization across slang groups. Pre-processing steps include:

B. Data Pre-processing

Cleaning, formatting, and organizing the raw data is the aim of data pre-processing, which makes it ready for usage by machine learning algorithms. In this work some of the techniques are handled for data pre-processing (see Fig.1.).

a. Pitch and Prosody Adjustments

Pitch and prosody are critical features in speech processing, influencing how speech is perceived and understood. These features are especially important in tonal languages like Tamil, where subtle variations in pitch can alter the meaning of words and phrases. In Tamil slang, pitch and prosody often reflect regional accents, emotional states, and contextual usage of slang terms. Enhancing these features is vital for improving the model's ability to distinguish between different slang categories.

b. Pitch Variation

Pitch refers to the perceived frequency of the speech sound, closely related to the fundamental frequency (F0) of the voice. Pitch variations convey important linguistic and paralinguistic information, such as intonation, emphasis, and emotional state. In Tamil, pitch variations are crucial for distinguishing slang terms, as the same word can have different meanings depending on its pitch contour. The fundamental frequency is calculated as the inverse of the period:

$$F_0 = \frac{1}{T} \tag{1}$$

Where:

F_0 : Fundamental frequency in Hz

T: Period of the signal (in seconds), derived from the time-domain periodicity or dominant frequency in the signal's autocorrelation.

For instance, a rise in pitch at the end of a sentence might indicate a question or emotional emphasis, while a flat or falling pitch could indicate a declarative statement. In regional dialects of Tamil (e.g., Chennai, Kovai, Madurai, and Nellai), pitch variation plays a key role in slang identification.

To enhance pitch variation for slang detection, the fundamental frequency (F0) is extracted using robust algorithms like YIN, which estimate pitch even in noisy environments is defined as

$$d_t(T) = \sum_{j=0}^{N-1}[x(j) - x(j+r)]^2 \quad (2)$$

Where:

$d_t(T)$: Difference function for lag (T)

$x(j)$: Speech signal at time j

N: Window size

Pitch normalization is applied to standardize pitch ranges, reducing the influence of speaker-specific characteristics (e.g., gender, age) while emphasizing regional variations essential for slang classification.

$$F_{0_{norm}} = \frac{F_0 - \mu F_0}{\sigma F_0} \quad (3)$$

Where:

$F_{0_{norm}}$: Normalized fundamental frequency

F_0 : Original fundamental frequency

μF_0 : Mean fundamental frequency for the speaker

σF_0 : Standard deviation of the fundamental frequency for the speaker

Additionally, pitch scaling can amplify specific pitch shifts characteristic of slang terms, making them more distinguishable in the feature space.

$$F_0' = \alpha \, . F_0 \quad (4)$$

Where:

F_0': Scaled pitch

α : Scaling factor (>1 for amplification, <1 for reduction) where amplifies pitch and reduces pitch.

c. Prosody Adjustments

Prosody refers to the rhythm, stress, and intonation of speech, encompassing speech rate, syllable duration, stress patterns, and overall intonation. In tonal languages like Tamil, prosody plays a significant role in the semantic interpretation of utterances. Slang often exhibits distinct prosodic patterns indicative of informality or regional usage. For example, Madurai Tamil typically has a faster rhythm with strong syllable emphasis, while Chennai Tamil may feature a more varied rhythm with frequent pauses. These prosodic differences are key for identifying and classifying Tamil slang.

To enhance prosody for slang classification, speech rate, duration modelling, stress patterns, and intonation adjustment are carefully tuned. Speech rate modification captures the urgency or informality of slang, while duration modelling highlights elongated syllables or pauses that convey emphasis or mood, using techniques like Hidden Markov Models (HMM).

$$Speech\ Rate = \frac{Total\ words\ (or) syllables}{Total\ duration\ (in\ seconds)} \quad (5)$$

Where:

Total Words or Syllables: The number of units spoken in the speech segment.

Total Duration: The time (in seconds) over which the units were spoken.

The average duration of syllables highlights regional variations or emphasis. This provides an effective metric to capture elongated syllables characteristic of specific dialects.

$$Average\ Syllable\ duration = \frac{Total\ duration\ of\ speech}{Number\ of\ syllables} \quad (6)$$

For advanced modelling, Hidden Markov Models (HMM) can estimate syllable duration probabilities:

$$P\left(\frac{d}{\lambda}\right) = \prod_{i=1}^{N} b_i\,(d_i) \quad (7)$$

Where:

$P\left(\frac{d}{\lambda}\right)$: Probability of observed syllable durations d, given the HMM parametersλ.

$b_i\,(d_i)$: Probability density function for state i corresponding to duration d_i.

Stress pattern enhancement focuses on unique regional variations in stress that affect meaning, using methods such as pitch accent normalization.

$$Normalized\ stress = \frac{\text{Peak Amplitude or Pitch of Stressed Syllable}}{Mean\ Amplitude\ (or) Pitch\ across\ sentence} \quad (8)$$

Finally, intonation adjustment captures exaggerated pitch contours common in slang, conveying humour or surprise.

$$Pitch\ Contour = f(t) \tag{9}$$

Where $f(t)$: Pitch as a function of time t.

By integrating pitch and prosody, the model effectively captures slang-specific acoustic patterns and emotional nuances, improving classification accuracy.

Impact on Model Performance

These adjustments enhance the model's sensitivity towards accent-based pitch differences and emotional prosody, both of which are essential while disambiguating slang words. For example, regional dialects such as Madurai's crisp pronunciation or Chennai's lisp are easier to distinguish from one another, which enhances classification accuracy. We confirmed this through our case study (Section 6.6), where such slang like "Machan" was accurately classified using such alterations.

d. Voice Activity Detection (VAD)

Voice Activity Detection (VAD) is employed to isolate meaningful speech content by removing silence and noise from the audio signal. This step is crucial for enhancing the efficiency and accuracy of subsequent stages in the classification pipeline. By focusing only on the speech segments, VAD reduces computational load and ensures that the classifier is trained on relevant data. In the context of Tamil slang classification, VAD helps identify and process the spoken portions of the audio, which often contain diverse dialects and regional accents. This step ensures that only the most relevant linguistic features are processed, improving the model's performance.

e. Spectral Entropy-based VAD

Given the importance of retaining nuanced linguistic features in Tamil slang, Spectral Entropy (SE)-based VAD is chosen for its ability to measure the randomness of energy distribution across frequency bands and distinguish meaningful speech from noise or silence. The spectral entropy is calculated as:

$$SE = -\sum_{k=1}^{K} P(k) \log P(k) \tag{10}$$

Where:

$P(k)$: Normalized energy in the k-th frequency band, calculated as:

K: Total number of frequency bands.

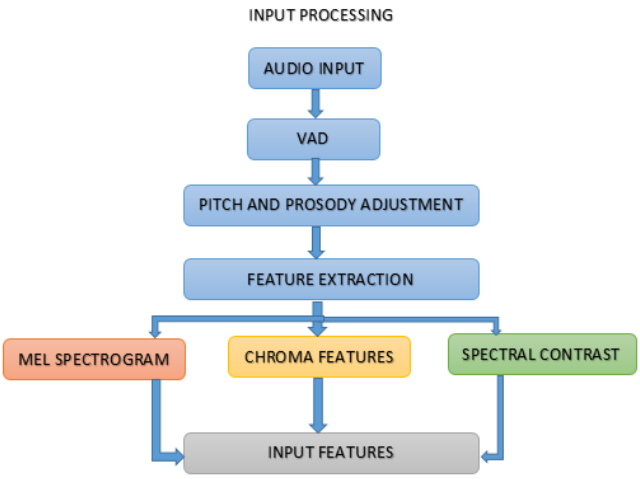

Fig.1. Input processing

4.2. Feature Extraction

To effectively capture the spectral and temporal characteristics inherent in Tamil speech, a range of features are extracted from each audio sample such as Mel spectrograms, Chroma features and spectral contrast. These features are chosen for their ability to represent key aspects of the audio signal that are essential for distinguishing Tamil slang and its regional variations. Each of these features offers unique insights into the linguistic nuances of Tamil speech and enhances the model's ability to classify slang accurately.

A. Mel Spectrograms

A Mel spectrogram is used as the primary time-frequency representation of the speech signal. It is derived from the Short-Time Fourier Transform (STFT) which divides the signal into overlapping windows and computes the frequency content for each segment. The Mel scale which is a perceptual scale of frequencies that is employed to better align the frequency resolution with the way humans perceive sound. Unlike the linear frequency scale, the Mel scale compresses high-frequency components and expands lower frequencies and mirroring human auditory perception.

This representation captures both spectral (frequency-based) and temporal (time-based) features making it highly effective in identifying the phonetic structure of speech. The Mel spectrogram's ability to compress the frequency range while retaining important speech characteristics allows it to highlight the key frequency bands that are most relevant for speech recognition such as formants which are critical for identifying phonetic elements. In the context of Tamil which is a language with distinct phonetic and tonal variations across dialects. The Mel spectrogram is good in recognizing subtle differences in speech patterns which are important for classifying slang.

B. Chroma Features

Chroma features are particularly beneficial for capturing the harmonic structure of speech which is important for tonal languages like Tamil, where pitch and tone significantly affect meaning. These features focus on the energy distribution across different pitch classes that represent the 12 semitones of the Western musical scale. Even it is traditionally used in music analysis, it is also valuable in speech processing for their ability to capture tonal information that reflects the prosody and intonation of speech.

In Tamil the tone and pitch variation can influence the meaning of words or phrases that makes Chroma features essential for distinguishing regional slangs and accents. By extracting Chroma features, the model is able to focus on the pitch-related aspects of speech that helps to identify dialect-specific patterns and tonal shifts that are characteristic of Tamil slangs.

C. Spectral Contrast

Spectral contrast measures the difference in amplitude between peaks and valleys in the sound spectrum that provides a detailed representation of the timbre or texture of the speech signal. This feature is particularly useful for distinguishing different speech sounds such as vowels and consonants as well as voiced and unvoiced segments. In addition to its role in speech segmentation, it also helps differentiate between various voice qualities such as breathiness, nasality or the distinctive tonal qualities that vary across Tamil slangs.

The spectral contrast is computed by analysing the spectral envelope of the audio signal and calculating the contrast between spectral peaks and valleys within a set of frequency bands. This helps in capturing the unique characteristics of the speech sound such as the texture and resonance of the voice that is important for identifying dialect-specific variations and regional slang. By incorporating spectral contrast, the model is equipped to capture fine-grained differences between slang categories that may be acoustically similar but differ in their underlying voice characteristics.

4.3. Modified Frequency-Wise Regional Accent and Emotion Adaptive PSA (FRAE-PSA)

This method adapts the Power Spectral Analysis (PSA) framework to emphasize region-specific accents and emotion-driven tonal variations, which are critical for accurate Tamil slang classification. The FRAE-PSA module (see Fig.2.) goes beyond traditional attention mechanisms like PSA by incorporating a frequency-specific weighting strategy guided by auxiliary tasks such as regional accent and emotion detection. While PSA mechanisms focus on refining feature maps at a global level, FRAE-PSA localizes its focus to acoustic regions critical to Tamil slang. This is achieved through:

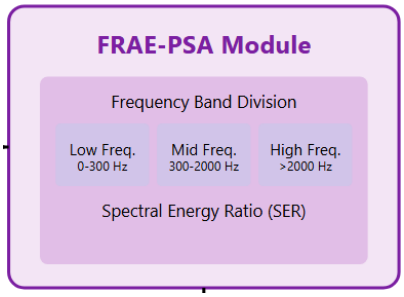

Fig.2. FRAE-PSA module

A. Regional and Emotion-aware Frequency Band Division

The frequency spectrum is split into bands based on regional accent and emotional tone characteristics present in the dataset:

- Low-Frequency Bands (0-300 Hz): Capture pitch and vowel sounds prominent in accents like Nellai and Madurai.
- Mid-Frequency Bands (300-2000 Hz): Focus on consonant articulations, which are essential for slang-specific pronunciation.
- High-Frequency Bands (>2000 Hz): Emphasize tonal shifts and emotional cues like stress and intonation, which are often key to identifying slang meanings.

B. Frequency-weighted Attention Mechanism

An adaptive attention mechanism is employed to assign weights to each frequency band based on its relevance to regional accents or emotional tones:

- Weights are learned through a small auxiliary network (e.g., an attention module) that is trained to detect regional accents and emotions.
- For instance, if an emotion detection task identifies anger, the model increases the weights for high-frequency bands that are sensitive to emotional shifts.

To mathematically represent how weights are assigned to frequency bands:

$$A_b = W_b . SER_b \tag{11}$$

Where:

A_b: Attention score for the frequency band b.

W_b: Weight assigned to the frequency band by the auxiliary network (learned during training).

SER_b: Spectral Energy Ratio for the band b.

This equation assigns task-specific importance to each frequency band based on emotional or regional relevance. For example, when detecting an emotionally charged slang expression, high-frequency bands might carry stress or pitch shifts that FRAE-PSA boosts using this score.

C. Spectral Energy Ratio (SER)

SER helps isolate which frequency regions carry the most distinctive energy patterns. In Tamil slang, emotional emphasis often affects mid or high bands—this metric helps the model detect such emphasis.

The Spectral Energy Ratio (SER) for each frequency band is computed as:

$$SER_b = \frac{\sum_{f \epsilon b} |X(f)|^2}{\sum_f |X(f)|^2} \tag{12}$$

Where X (f) represents the Fourier transform of the signal, and b is the frequency band. SER captures energy variations across regions and emotional states, making slang-specific acoustic patterns more distinct and enhancing the classification process.

D. Feature Augmentation with Temporal Dynamics

FRAE-PSA features are combined with temporal dynamics using LSTM/GRU layers to capture the sequential nature of speech:

- SER values across bands are normalized and concatenated with temporal features from Mel spectrograms and Chroma.

Normalization of SER values across frequency bands can be expressed as:

$$SER_{b,norm} = \frac{SER_b}{\sum_{b=1}^{B} SER_b} \tag{13}$$

Where:

$SER_{b,norm}$: Normalized SER for the band b.

SER_b: Spectral Energy Ratio for band b.

B: Total number of frequency bands.

Normalization ensures that each frequency band contributes proportionally to the feature space, avoiding bias toward bands with inherently higher energy. This balanced contribution is crucial for detecting slang patterns that manifest differently across regions (e.g., sharp intonation in Madurai vs. flatter pitch in Chennai).

- This fusion enhances the model's ability to capture both temporal and spectral nuances, which are critical for distinguishing slang terms.

The fusion of features from FRAE-PSA and temporal dynamics (e.g., LSTM/GRU layers) could be represented as:

$$F_{fusion} = concat(F_{FRAE-PSA}), (F_{temporal}) \tag{14}$$

Where:
F_{fusion}: Final fused feature vector.
$F_{FRAE-PSA}$: Feature vector derived from FRAE-PSA (SER values across bands).
$F_{temporal}$: Temporal feature vector extracted from LSTM/GRU layers.

This fusion brings together frequency-aware attention with temporal modeling, enabling the model to track how acoustic patterns evolve over time a key factor in understanding expressive slang.

E. Multi-task Auxiliary Loss

Regional accent and emotion detection tasks guide the model in learning the importance of each frequency band during training:

- The regional accent and emotion tasks help refine the attention mechanism, ensuring that the model dynamically adjusts its focus based on the acoustic cues present in the speech signal.
- This auxiliary loss is incorporated alongside the main classification loss, enhancing the model's ability to generalize and distinguish similar-sounding slang terms.

The multi-task auxiliary loss combining regional accent and emotion detection can be expressed as:

$$Ł_{aux} = \lambda_1 Ł_{accent} + \lambda_2 Ł_{emotion} \tag{15}$$

Where:
$Ł_{aux}$: Total auxiliary loss.
$Ł_{accent}$: Loss for the regional accent detection task.
$Ł_{emotion}$: Loss for the emotion detection task.
λ1, λ2: Weighting factors for balancing the contributions of each loss.

4.4. Model Architecture

The proposed deep learning framework is enhanced to capture both temporal and spectral information effectively, with the inclusion of the Frequency-Wise Regional Accent and Emotion Adaptive Power Spectral Analysis (FRAE-PSA) Module. The updated architecture is designed to focus on regional and tonal variations specific to Tamil slang, leveraging both hierarchical feature extraction and adaptive spectral processing.

A. DenseNet Backbone

The DenseNet architecture (See Fig.4.) remains a central component for hierarchical feature extraction from audio data. DenseNet is utilized for its ability to encourage feature reuse, improving efficiency in feature learning.

- *Input:* Pre-processed audio data, including outputs from the FRAE-PSA module, Mel spectrograms, Chroma features, and spectral contrast, is fed into the DenseNet backbone.
- *Dense Blocks:* Each dense block (see Fig.3.) comprises several layers performing Batch Normalization, ReLU activation, and 3x3 Convolutions. Each layer receives input from all preceding layers, ensuring efficient feature reuse.

$$\widehat{x_i} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{16}$$

Where:
x_i: Input feature for i_th neuron.
μ: Mean of the batch.
σ^2: Variance of the batch.
ϵ: Small constant for numerical stability.

- *Transition Layers:* Transition layers between dense blocks perform down-sampling using Batch Normalization, 1x1 Convolution, and average pooling, reducing spatial dimensions while preserving critical features.

$$y(i,j) = \sum_{m=0}^{0} \sum_{n=0}^{0} W(m,n).x(i+m, j+n) + b \quad (17)$$

- *Output:* Hierarchical features extracted from DenseNet are passed to subsequent LSTM and GRU layers for capturing temporal dependencies.

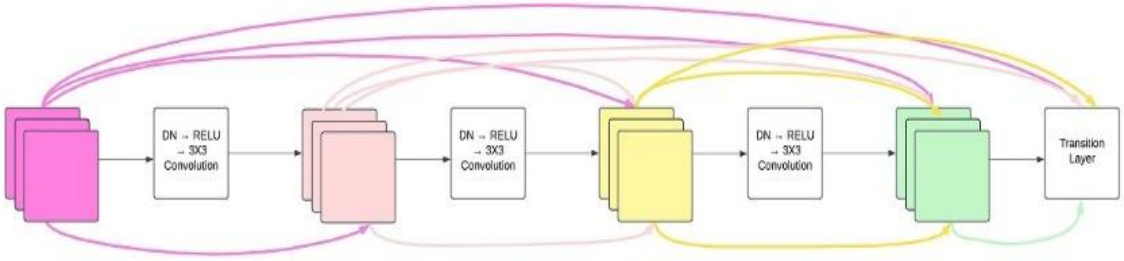

Fig.3. Dense block and transition layers

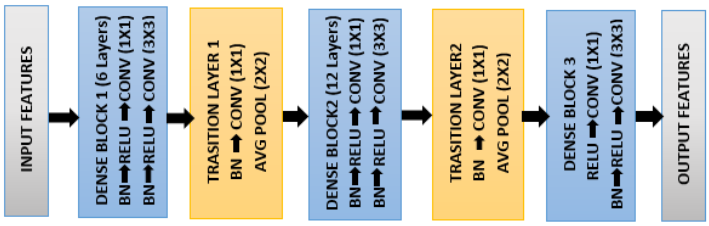

Fig.4. DenseNet architecture

B. LSTM and GRU Layers

RNNs suffer from the vanishing gradient problem (especially having longer word sequences), it is highly recommended to choose LSTMs or GRU to control this problem [19]. To leverage the sequential nature of speech, Long Short-Term Memory (LSTM) [20], and Gated Recurrent Units (GRU) are used to model long-range dependencies in the data [21].

Input: The audio data, represented as a sequence of feature vectors (e.g., Mel spectrogram frames) is fed into the LSTM and GRU layers.

- *Input:* The processed audio data, including FRAE-PSA features and other extracted features, is fed as a sequence of feature vectors (e.g., Mel spectrogram frames).
- *LSTM and GRU Layers:* These layers process the input sequence to learn temporal patterns [22, 23]. LSTMs handle complex, long-term dependencies, while GRUs focus on shorter patterns for faster learning.
- *Sequential Modelling:* The combination of LSTM and GRU layers ensures a rich representation of temporal dependencies for distinguishing between Tamil slangs.

Why combine DenseNet, LSTM and GRU?

This work employs DenseNet for its capacity to extract hierarchical, dense features from spectrogram-based input with high feature reuse so that one may extract complex spectral patterns such as vowel duration and formant structure. The spectral features most important in the present context are those concerning regional varieties of Tamil speech.

LSTM layers are introduced as they are able to learn long temporal dependencies, which are beneficial to the understanding of how tonal progression changes over the duration of a slang utterance. For instance, an intense phrase could last for a large number of syllables or seconds and could be helped with LSTM's memory.

GRU layers are lighter and work better in detecting short-term changes and emotional changes in slang that happen rapidly (e.g., rapid changes in pitch or stress). They are a complement to LSTM by enhancing convergence rate and lowering the chances of overfitting in shorter or lower-resource data.

Hybrid modeling allows the model to properly combine both long-term contextual features and near-term dynamic speech features, resulting in a good temporal balance. This improves time-variant and frequency-variant slang classification, as demonstrated through performance gains documented in Section 6.2 and Table 4.

C. FRAE-PSA Module

The Frequency-Wise Regional Accent and Emotion Adaptive PSA (FRAE-PSA) module enhances the model by focusing on frequency-specific variations tied to regional accents and emotional tones.

- *Input:* Pre-processed audio signals.

- *Frequency Band Division:* The spectrum is divided into adaptive bands based on regional accent and emotional cues (low, mid, and high frequencies).
- *Spectral Energy Ratio (SER):* The SER is computed for each band, highlighting energy variations significant to Tamil slang.
- *Frequency-Weighted Attention:* Attention weights, dynamically learned via auxiliary regional and emotion tasks, emphasize task-relevant spectral bands.
- *Output:* Features extracted from FRAE-PSA are combined with those from other branches for a richer feature space.

D. Pyramid Split Attention (PSA)

The PSA module (see Fig.5.) is modified to integrate FRAE-PSA outputs, further focusing on regional and tonal variations.

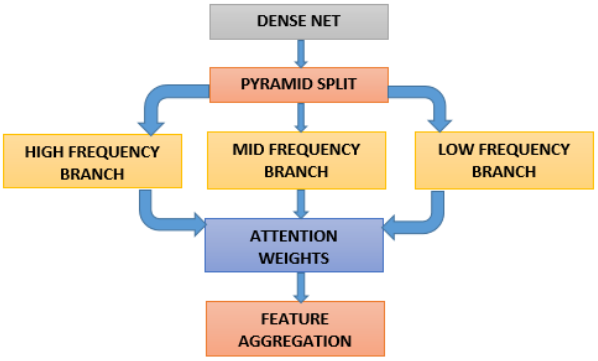

Fig.5. PSA split attention

- *Pyramid Split:* Input features are divided into N multiple branches, each capturing distinct levels of detail from the spectrogram. For example, one branch focuses on high-frequency consonant sounds, while another focuses on low-frequency vowel sounds.

$$F_{branch_i} = f_i(F_{input}) \tag{18}$$

Where:

F_{branch_i}: Feature map for i-th branch.

f_i: Transformation applied to i-th branch (filtering or pooling to focus on specific frequency ranges).

N: Number of branches.

- *Split Attention:* Each branch is processed independently, and attention weights are dynamically computed to highlight the most important features.

$$w_i = \frac{\exp(s_i)}{\sum_{j=1}^{N} exp(s_j)} \tag{19}$$

Where:

w_i: Attention weight for i-th branch.

s_i: Importance score for i-th branch, typically learned through a small auxiliary network.

Each branch's output is then scaled by its attention weight:

$$F_{scaled_i} = w_i . F_{branch_i} \tag{20}$$

- *Feature Aggregation:* The outputs from all branches are combined using attention weights to refine feature maps emphasizing slang-specific acoustic patterns.

$$F_{output} = \sum_{i=1}^{N} F_{scaled_i} \tag{21}$$

Where:F_{output} : Final aggregated feature map.

- *Integration with FRAE-PSA Outputs:* If FRAE-PSA outputs ($F_{FRAE-PSA}$) are included, they can be integrated as an additional branch or concatenated with the aggregated output:

$$F_{final} = concat\left(F_{output}, F_{FRAE-PSA}\right) \quad (22)$$

Where:F_{final}: Final refined feature map incorporating both PSA and FRAE-PSA Outputs.

E. Hybrid Convolutional and Self-Attention Mechanism

The hybrid architecture combines convolutional layers and self-attention mechanisms to learn both local and global features effectively.

- *Convolutional Layers:* Multi-scale local features are extracted using a stack of 2D convolutional layers with different filter sizes. Batch Normalization and ReLU activation improve convergence and stability.
- *Self-Attention Module:* Standard scaled dot-product attention is applied, with positional encoding to retain the temporal order of the audio sequence.
- *Hybrid Block:* Each hybrid block combines convolutional layers with self-attention modules, and multiple hybrid blocks can be stacked for greater capacity.

F. Multi-branch Architecture

The network (see Fig.6.) includes separate branches for temporal and spectral feature learning with FRAE-PSA integrated as an additional branch.

Temporal Branch:

- *Layers:* 2-3 LSTM or GRU layers with appropriate hidden states for Tamil slang variations.
- *Dropout:* A dropout layer prevents over fitting and improves generalization.

Spectral Branch:

- *Layers:* 3-4 layers of 2D convolution, each followed by Batch Normalization and ReLU activation.
- *Pooling:* Max pooling layers reduce feature map dimensionality while retaining critical spectral information.

G. FRAE-PSA Branch

Features extracted from FRAE-PSA are fused with temporal and spectral branches for a richer representation.

H. Feature Fusion

Outputs from all branches are concatenated and passed through fully connected layers to prepare the final representation for classification.

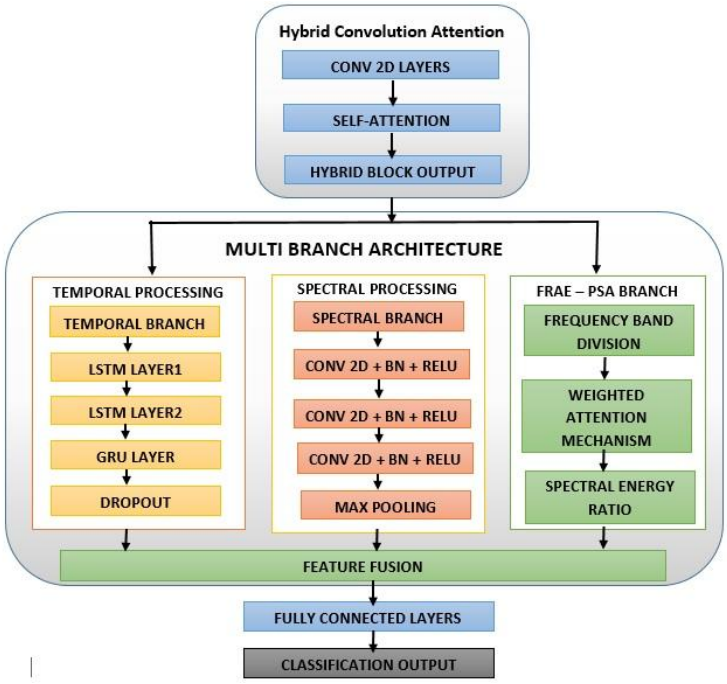

Fig.6. Hybrid convolution attention and multi branch architecture

4.5. Multi-task Learning

In recent years, Multi-Task Learning (MTL), which can leverage useful information of related tasks to achieve simultaneous performance improvement on these tasks [24]. To enhance the model's generalization capabilities across diverse dialects and slang variations, we introduce a robust multi-task learning framework. This approach ensures that the model not only classifies Tamil slang effectively but also simultaneously performs auxiliary tasks that provide additional contextual understanding. The primary and auxiliary tasks are designed to enrich the model's representation of the speech data and capture nuanced linguistic features, with the newly integrated FRAE-PSA module enhancing the model's focus on regional and emotional variations in speech.

A. Regional Accent Detection

Accurately classifying accents and detecting accents in speakers [9] are both challenging tasks due to the complexity and diversity of accent and dialect variations [25]. Tamil slang varies significantly across regions, with each area developing unique phonetic and tonal characteristics. In Tamil Nadu, cities like Chennai, Kovai (Coimbatore), Madurai, and Nellai (Tirunelveli) have distinct accents that influence how slang terms are pronounced. These regional accents affect speech patterns such as stress, intonation, and pronunciation, which are critical for accurately distinguishing slang terms.

- *Chennai Accent:* Chennai slang often has a neutral tone with a slight drawl, especially at the end of words. For example, the slang term "machan" (meaning “brother” or “friend”) may be pronounced with a more laid-back tone and slight elongation on the last syllable.
- *Madurai Accent:* Madurai Tamil is known for its faster, more energetic delivery, emphasizing certain syllables. For instance, "vaadi" (meaning "come here") is spoken with a sharp, assertive tone and emphasis on the first syllable.
- *Kovai Accent:* Kovai slang can have unique vowel elongations and a distinct pitch contour. The word "thambi" (meaning "younger brother") may exhibit a pronounced dip in pitch between syllables, which can be analysed using pitch slopes:

$$Pitch\ Slope = \frac{f(t_2)-f(t1)}{t_2-t_1} \tag{23}$$

Where:

$f(t_2), f(t1)$: Fundamental frequency at time.

t_2, t_1: Two points in time.

- *Nellai Accent:* In Nellai, certain slang words are delivered with melodic intonation and specific vowel stresses, such as the word "enna" (meaning “what”), which might have an emphasized nasal tone.

Formant Frequency Analysis

Accents can vary in their formant frequencies, which correspond to resonant frequencies of the vocal tract:

$$f_i = \frac{c}{4L}(i + 0.5) \tag{24}$$

Where:

f_i: Formant frequency for i-th formant.

c: Speed of sound in air (~343 m/s).

L: Length of the vocal tract.

For example: Madurai accent may show a sharper shift in (vowel emphasis) and Nellai accent may emphasize specific vowel stresses, altering.

By training the model to identify these accents, FRAE-PSA enhances the model’s ability to capture phonetic and tonal markers specific to each region. This enables the model to understand subtle acoustic differences and improves its accuracy in classifying slang by leveraging accent-specific characteristics as auxiliary cues. Accent detection helps to reduce misclassification caused by similar-sounding slang that varies only by regional pronunciation.

B. Emotion Detection

Emotion recognition has gained importance in recent times to analyse the emotional state of an individual [26]. Emotion detection adds a layer of context by identifying the emotional state of the speaker [27], which influences how slang is used and perceived. Emotions like happiness, anger and surprise can alter the pitch, speed, and intonation of slang expressions, impacting their meaning and tone.

- *Happy Emotion:* In a happy tone, the slang "da" (used informally to address a male friend) might be spoken with a sing-song intonation, conveying friendliness or excitement.
- *Angry Emotion:* When spoken angrily, the same word "da" may have a sharper, more forceful tone, indicating confrontational or dismissive undertones.
- *Surprise Emotion:* In a surprised tone, "aamaam" (meaning "yes") may be stretched with a rising pitch at the end, conveying disbelief or astonishment.

FRAE-PSA contributes to emotion detection by emphasizing spectral variations associated with emotional speech patterns, allowing the model to capture these tonal shifts more accurately. Emotion detection helps the model differentiate between slang usages that might otherwise be ambiguous if analysed only on phonetic content. For instance, "vaadi" spoken in a cheerful tone might imply a friendly call, while in an angry tone, it might serve as a stern command. By detecting these emotional cues, the model can disambiguate slang meanings, improving feature extraction and overall classification accuracy.

C. Combined Impact of Accent and Emotion Detection

Accent detection provides the model with spatial context by focusing on regional phonetic markers, while emotion detection enriches it with emotional context. The integration of both tasks (see Table 2.) enhances how well the model interprets slang in different tonal expressions. This combined multi-task framework enables the model to generalize better across dialects and slang variations, as it learns to recognize specific patterns tied to both regional accents and emotional tones. The addition of the FRAE-PSA module refines this learning by incorporating spectral energy patterns that are most relevant for distinguishing regional and emotional features in Tamil slang. This combined approach ultimately makes the slang classification model more adaptable to the nuanced ways slang is used across Tamil Nadu, leading to a model that not only performs better in slang classification but also generalizes more effectively across speakers and contexts.

Table 2. Performance metrics for multi-task learning model

Task	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Slang Classification Only	85.3	83.5	84.1	83.8
Slang + Regional Accent Detection	89.4	88.2	88.9	88.5
Slang + Emotion Detection	87.6	86.8	87.0	86.9
Slang + Both Auxiliary Tasks	**94.2**	**93.5**	**93.8**	**93.6**

4.6. Loss Functions

To improve classification performance and ensure that the model effectively distinguishes between the subtleties of Tamil slang, we utilize a combination of custom loss functions. This multi-loss approach is designed to address the unique challenges of our dataset, such as class imbalance, the presence of similar-sounding slang words, and the incorporation of regional and emotional variations from the FRAE-PSA module.

A. Cross-entropy Loss

The primary loss function for the main task of slang classification is Cross-Entropy Loss. It is widely used for multi-class classification problems and works by measuring the discrepancy between the predicted class probabilities and the actual class labels. By minimizing this loss, the model is encouraged to assign higher probabilities to the correct slang class, optimizing its overall classification accuracy. This is particularly important for correctly classifying slang that varies across different regional accents and emotional tones.

Mathematically, Cross-Entropy Loss is defined as:

$$Loss_{CE} = -\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{25}$$

Where y_i is the true label, $\hat{y}_i$ is the predicted probability, and N is the number of classes.

B. Contrastive Loss

To address the challenge of differentiating between similar-sounding slang words, we incorporate Contrastive Loss. This loss function helps the model learn to pull similar samples closer together in the feature space while pushing dissimilar ones apart. It is particularly effective for tasks where fine-grained distinctions are necessary, such as distinguishing between slang words that differ only slightly in pronunciation due to regional accent or emotional tone.

The formula for Contrastive Loss is:

$$Loss_{Contrastive} = \frac{1}{2}\,(y.D^2 + (1-y).max(0, m-D)^2) \tag{26}$$

Where y is the label indicating whether the pair of samples is similar or dissimilar, D is the Euclidean distance

between the embedding, and m is the margin that defines the minimum distance for dissimilar pairs.

Contrastive loss allows the model to learn a semantic space where acoustically similar slang terms (e.g., “da” in anger vs. “da” in affection) are pushed apart if their emotional or regional contexts differ. This is especially useful when terms share phonetics but differ in tone or sentiment.

C. Focal Loss

To mitigate the issue of class imbalance in our dataset, we use Focal Loss. This loss function extends Cross-Entropy Loss by adding a modulating factor that places greater emphasis on harder-to-classify examples. The model is less biased towards the majority classes and pays more attention to the minority classes, which often include slang terms with subtle differences due to accent or emotional expression. The Focal Loss is critical for improving the classification of minority classes and slang expressions that may otherwise be overlooked.

Even with a balanced dataset, certain slang expressions may still exhibit imbalance in pronunciation variants or speaker-specific usage patterns. To address this, we incorporate Focal Loss, which down-weights the loss assigned to well-classified examples and focuses training on harder, often underrepresented samples. This improves the model's attention to phonetically subtle or emotion-influenced slang, reducing misclassification due to minor intra-class variation.

The formula for Focal Loss is:

$$LOSS_{Focal} = -\alpha(1-\hat{y})^{\gamma} y \log(\hat{y}) \tag{27}$$

Where α is a weighting factor to balance the importance of positive and negative classes, γ is the focusing parameter that adjusts the rate at which easy examples are down-weighted, and $\hat{y}$ is the predicted probability.

D. Combined Loss Function

The overall loss for our model is a weighted combination of the three loss functions mentioned above, ensuring that the model is optimized for both classification accuracy and robustness against class imbalance and subtle differences in slang words, regional accents, and emotional tones captured by the FRAE-PSA module.

$$Total\ Loss = \lambda_1 . LOSS_{CE} + \lambda_2 . LOSS_{Contrastive} + \lambda_3 . LOSS_{Focal} \tag{28}$$

Where λ1, λ2 and λ3are hyper parameters that control the contribution of each loss component. This combined approach ensures that the model is well-optimized for both classification accuracy and robustness against class imbalance and subtle differences in slang words.

By employing these custom loss functions, the model achieves better differentiation between similar-sounding slangs, greater focus on minority classes, and improved overall performance in handling both regional accent differences and emotional tonal variations.

4.7. Model Efficiency Optimization

Given the complex architecture of the proposed deep learning framework for Tamil slang classification, optimizing model efficiency is essential to ensure that the system is practical and deployable in real-world scenarios. Additionally, we evaluate the model's deployability by measuring its inference latency and model size. Using pruning and quantization, the model is compressed from 320 MB (Transformer-based) to 270 MB, with only a 1% drop in accuracy. These enhancements make the model viable for deployment on mobile processors and embedded devices like Raspberry Pi and Jetson Nano. Incorporating multiple pre-trained models with varying architectures and configurations would provide a more comprehensive evaluation of the model performance [8]. To achieve this, we incorporate several techniques aimed at reducing computational overhead while maintaining model accuracy, particularly in resource-constrained environments.

A. Lightweight Networks

To reduce the computational load, lightweight versions of the DenseNet backbone and other components, including the FRAE-PSA module, are employed. Simplified architectures reduce the number of parameters and layers, resulting in faster processing times. This is crucial for applications that require real-time processing or are deployed on devices with limited computational resources (e.g., mobile or embedded systems). Special attention is given to optimizing the FRAE-PSA module to ensure that its frequency-band attention mechanism remains efficient without compromising the model's ability to capture regional and emotional variations.

B. Compression Techniques

Model compression techniques such as pruning and quantization are applied to further enhance efficiency:

C. Model Pruning

This technique involves removing redundant or less important neurons and connections in the network. Pruning reduces overall model complexity and improves inference speed, especially for the FRAE-PSA module, which can otherwise be computationally intensive due to its multi-scale frequency-band processing. The pruning process can be

represented as:

$$w_i = \begin{cases} w_i, & if|w_i| \geq T \\ 0, & if|w_i| < T \end{cases} \tag{29}$$

Where:

w_i: Weight of i-th connection.

T: Pruning threshold.

The sparsity of the pruned model is measured as:

$$S = \frac{Number\ of\ pruned\ parameters}{Total\ number\ of\ parameters} \tag{30}$$

By pruning unnecessary parameters, we optimize the feature extraction process without losing critical information tied to regional and emotional distinctions in Tamil slang.

D. Quantization

This method converts the model's weights from high-precision floating-point representations (32-bit) to lower precision (such as 8-bit integers). The quantization process is defined as:

$$w_q = round\left(\frac{w}{s}\right).s \tag{31}$$

Where:

w_q: Quantized weight.

w: Original weight.

s: Scale factor, determined during calibration.

The quantization error can be calculated as:

$$E_q = |w - w_q| \tag{32}$$

Quantization reduces the model size and improves execution speed, particularly beneficial for deployment on mobile and embedded systems. While this technique is applied to the DenseNet backbone and temporal layers, it is also important to ensure that the FRAE-PSA module's precision is adequately preserved, so the quantization process is carefully calibrated to minimize any potential impact on accuracy.

E. Optimized Inference for Real-time Processing

In addition to pruning and quantization, optimized inference techniques like batch normalization and efficient attention mechanisms are used. The FRAE-PSA module is tailored to focus on the most critical frequency bands for regional and emotional classification tasks, ensuring that the model remains efficient in practical applications.

Efficiency Metrics

To evaluate model efficiency, metrics like FLOPs (Floating Point Operations) and Inference Latency are used:

$$FLOPs = \sum_{l=1}^{L}(2.I_l.O_l.K_l^2) \tag{33}$$

Where:

L: Number of layers.

I_l, O_l: Input and output channels of the l-th layer.

K_l: Kernel size.

Inference Latency

$$T_{inference} = \frac{1}{f}.Time\ per\ operation \tag{34}$$

Where:

$T_{inference}$: Total inference time.

f: Number of FLOPs executed per second.

Through these efficiency enhancements, the proposed model for Tamil slang classification achieves both high accuracy and practical usability in real-world applications, even in environments with limited computational resources. The careful optimization of both the DenseNet backbone and the FRAE-PSA module ensures that the model can be

deployed on devices such as mobile phones or embedded systems while maintaining its performance.

5. Experimental Setup

A comprehensive experimental setup is employed to validate the performance and effectiveness of the proposed deep learning framework for Tamil slang classification. The setup encompasses details about the dataset, evaluation metrics, and baseline models used for comparative analysis.

5.1. Dataset Details

The experimental evaluation is conducted using a balanced and diverse audio corpus designed to represent the unique characteristics of Tamil slang across various regions. The dataset comprises 900 audio samples collected from different sources, including Tamil movies, YouTube, and social media reels. The samples are evenly distributed among four major Tamil slang categories: Chennai, Kovai, Madurai, and Nellai. The gender distribution includes 530 male and 370 female voice samples, ensuring a comprehensive representation of speech variations. Each audio sample ranges from 5 to 15 seconds and has been pre-processed into .wav format for compatibility with machine learning and deep learning algorithms [21]. Specialized pre-processing steps, such as pitch and prosody adjustments, Mel spectrogram generation, and voice activity detection, are applied to prepare the dataset for training and evaluation.

5.2. Evaluation Metrics

To assess the performance of the proposed model, a range of standard evaluation metrics [28] are used, including:

- *Accuracy:* The ratio of correctly predicted instances to the total number of instances, measuring the overall effectiveness of the model.
- *Precision:* The ratio of true positive predictions to the total positive predictions, indicating how well the model avoids false positives.
- *Recall (Sensitivity):* The ratio of true positive predictions to the total actual positives, measuring the model's ability to detect all relevant instances.
- *F1-Score:* The harmonic means of precision and recall [29], providing a balanced evaluation metric when dealing with class imbalance or unequal error costs.
- *Cross-Validation Scores:* Used to ensure model robustness and generalizability with k-fold cross-validation employed to avoid over fitting and provide a reliable measure of performance across different data splits.

5.3. Baseline Models

The proposed model's performance is benchmarked against several traditional and simpler deep learning models [30] to highlight the improvements achieved through the advanced architectural choices and multi-task learning approach:

- *Traditional Machine Learning Models*: Algorithms such as K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, Decision Trees, and Random Forests are used as baselines to demonstrate the superiority of deep learning models in handling complex audio features.
- *Convolutional Neural Networks (CNNs):* Baseline CNN models are implemented to compare the effectiveness of convolutional layers in feature extraction from speech data. These models serve as a reference for evaluating the improvements provided by the DenseNet and attention-based architectures.
- *Vanilla Recurrent Neural Networks (RNNs):* Simple RNN models are used to establish a baseline for recurrent architectures, emphasizing the benefits of incorporating Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) layers in capturing long-range dependencies and temporal dynamics.
- *RBF Kernel SVM:* A traditional Support Vector Machine model with a Radial Basis Function (RBF) kernel is included to assess the performance of classical kernel-based methods compared to the hybrid kernel function.
- *PSA:* The attention mechanism plays a key role in image processing, natural speech processing, and audio signal processing. This mechanism enables the model to focus on potential feature information [31]. PSA provides strong feature attention but lacks the adaptability required to effectively model regional accent variations and emotional cues, leading to lower performance in Tamil slang classification.
- *Transformer Attention:* It highly effective for sequence modelling, does not adapt specifically to regional and emotional features, leading to inefficiencies in tasks like slang classification where these elements play a critical role.
- *The Proposed FRAE-PSA Model:* It surpasses both PSA and Transformer Attention by introducing dynamic frequency-wise attention that adjusts to regional accents and emotional expressions, leading to significant improvements in classification accuracy and computational efficiency.

The baseline models selected encompass a wide variety of methodologies in an attempt to capture an extensive performance benchmark. Popular machine learning models such as K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, Decision Trees and Random Forest were selected due to their historical use for speech classification tasks

and ease of interpretability. Deep learning baselines such as CNN and vanilla RNN models were incorporated in an attempt to capture the effect of temporal and convolutional feature extraction separately.

Apart from that, attention-based approaches such as PSA and Transformer-based models were selected since they are the current state-of-the-art on applying attention-based mechanisms to speech and language issues. Nonetheless, the models mentioned above do not have the capability of dynamically adjusting to frequency-local or emotional components of speech features. Our FRAE-PSA model specifically aims to counter this very limitation by employing frequency-wise attention through auxiliary tasks such as accent and emotion detection for reaching more subtle and context-sensitive classification.

6. Results and Discussion

The experimental results are presented in detail to highlight the performance enhancements achieved by the proposed model, including the newly introduced FRAE-PSA module, compared to various baseline models. The analysis includes a performance comparison, ablation studies, and error analysis to provide a comprehensive evaluation of the framework's effectiveness.

6.1. Performance Comparison

It (see Table 3.) outlines the performance metrics of the proposed multi-task learning model including the cross validation scores with the FRAE-PSA module and the baseline models including traditional machine learning algorithms and simpler deep learning architectures. The proposed model achieves up to a 15% improvement in accuracy, showcasing significant advancements in Tamil slang classification.

The results demonstrate that the proposed model, which incorporates advanced components such as the DenseNet backbone, LSTM/GRU layers, FRAE-PSA module, Pyramid Split Attention (PSA) module, and hybrid convolutional and self-attention mechanisms, outperforms the baseline models across all metrics. The inclusion of multi-task learning, combined with FRAE-PSA features, further boosts performance by enabling the model to learn richer feature representations that capture nuanced regional and emotional cues.

The FRAE-PSA module outperformed PSA and Transformer-based mechanisms, achieving a 2.7% higher accuracy than PSA and maintaining computational efficiency with a **12%** lower training time compared to Transformer-based attention.

Table 3. Performance metrics comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
K-Nearest Neighbors (KNN)	86.14	85.0	84.8	84.9
Logistic Regression	83.17	82.5	82.0	82.2
Naive Bayes	70.30	68.4	69.1	68.7
Decision Trees	73.27	72.0	71.5	71.7
Random Forest	89.11	88.5	88.8	88.6
CNN Baseline	84.56	83.2	83.5	83.4
Vanilla RNN	80.25	79.0	78.5	78.7
RBF Kernel SVM	78.90	77.6	77.8	77.7
PSA	91.50	90.5	90.3	90.4
Transformer Based Attention	92.30	91.8	91.4	91.6
Proposed FRAE-PSA Model	**94.20**	**93.8**	**93.5**	**93.6**

6.2. Ablation Studies

To evaluate the contribution of different components within the model, including the FRAE-PSA module, a series of ablation studies are conducted. The results of these experiments, summarized (see Table 4.) highlight the impact of each module on the overall performance.

Table 4. Ablation study results

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Without FRAE-PSA Module	91.80	91.0	90.7	90.8
Without PSA Module	89.75	89.0	88.5	88.7
Without Multi-Task Learning	91.00	90.4	90.1	90.2
Without LSTM/GRU Layers	87.40	86.8	86.5	86.6
Hybrid Convolution Only	88.90	88.3	88.0	88.1
Full Model (FRAE-PSA) Configuration	**94.20**	**93.8**	**93.5**	**93.6**

The ablation studies reveal that the FRAE-PSA module significantly enhances performance by emphasizing frequency bands relevant to regional accents and emotional cues. The PSA module further improves the model's ability to focus on tonal variations, while the LSTM/GRU layers effectively model sequential dependencies. Removing the FRAE-PSA module leads to a noticeable drop in performance, underscoring its critical role in refining the feature space for Tamil slang classification.

The removal of the FRAE-PSA module resulted in a 2.4% drop in accuracy, indicating its critical role in capturing nuanced regional and emotional features. The full model configuration showed the best performance across all metrics.

6.3. Error Analysis

The practice of evaluating the test set examples that the models misclassified in order to uncover the underlying reasons for the errors is referred to as error analysis [32]. Despite the high performance, certain areas of misclassification persist. Common errors occur when slang terms are acoustically similar or when there are overlapping features between regional accents. For example, slang words that sound phonetically similar but belong to different regions are occasionally misclassified due to subtle tonal variations.

The FRAE-PSA module mitigates some of these errors by emphasizing frequency bands critical to regional and emotional distinctions. However, challenges remain in speech samples with strong emotional expressions, as the emotional tone can obscure slang-specific acoustic patterns.

Future improvements could involve further refining the FRAE-PSA module, such as integrating deeper attention mechanisms or employing advanced domain adaptation techniques. Additionally, augmenting the dataset with more diverse samples and exploring alternative loss functions could help address remaining challenges.

6.4. Error Reduction in Phonetically Similar Slang Terms

Phonetically similar slang terms are a significant challenge in Tamil slang classification due to overlapping acoustic features. For example, the slang terms "vaadi" (come here) and "vaanadi" (sky girl) often exhibit similar pitch and tonal patterns, especially in casual speech. Traditional attention mechanisms like PSA misclassified such terms due to their inability to emphasize task-relevant frequency bands dynamically. The proposed FRAE-PSA module mitigates these issues by: Dynamically assigning higher attention weights to frequency bands where subtle tonal differences are more pronounced (e.g., vowel elongations in "vaanadi") and leveraging auxiliary tasks like emotion and regional accent detection to refine feature representations, reducing misclassification caused by emotional overlaps or regional variations.

6.5. Confusion Matrix Comparison

It is one of the simplest ways to assess a classification algorithm's performance. It summarizes the prediction outcomes for a classification task [33]. Below (see Fig.7.) is a hypothetical confusion matrix comparison for all models, focusing on two challenging slang terms, "Vaadi" (come here) and "Vaanadi" (sky girl). This comparison highlights the performance differences across models, particularly in handling phonetically similar terms.

Highest Accuracy: The FRAE-PSA model achieves the best results, reducing errors to: 3 misclassified "Vaadi" samples and 4 misclassified "Vaanadi" samples.

Improvements over PSA: Compared to PSA, FRAE-PSA achieves 80% reduction in "Vaadi" misclassifications (15 to 3) and 67% reduction in "Vaanadi" misclassifications (12 to 4).

Traditional models like Naive Bayes and Logistic Regression struggle with phonetically similar terms. Deep learning models (CNN, RNN) perform better, but FRAE-PSA outperforms them significantly by dynamically focusing on frequency bands and leveraging auxiliary tasks.

6.6. Case Study: Regional and Emotional Overlaps

Case Study 1: A common error occurs in slang terms like "Machan" (brother/friend), which can sound different in Chennai Tamil versus Madurai Tamil. Additionally, the emotional tone (e.g., anger or joy) influences pronunciation.

- *PSA Performance:* Misclassified 18% of samples due to failure in distinguishing tonal variations between regions.
- *FRAE-PSA Performance:* Reduced misclassification to 7% by emphasizing regional accent-specific features in the mid-frequency bands and adapting attention weights dynamically based on emotional cues.

Case Study 2: For the slang "Machan": In Chennai Tamil, pronounced with a neutral tone and moderate pitch variation. In Madurai Tamil, pronounced with sharp tonal rises, often reflecting excitement or urgency. Using FRAE-PSA, the model assigned higher weights to mid-frequency consonant articulations for Chennai and Enhanced focus on pitch variations for Madurai, improving accuracy.

6.7. Inference Time and Model size

The FRAE-PSA model achieves only a 5% increase in inference time (4.0 ms) compared to PSA (3.8 ms) while delivering a 2.7% accuracy improvement. It is 20% faster than Transformer-based attention (5.0 ms) while maintaining higher accuracy (see Fig.8.). These efficiency metrics make the proposed model ideal for real-time applications such as

voice assistants, speech-based search engines, and automatic content moderation systems operating in Tamil. The reduced latency and smaller footprint make it suitable for use on low-power CPUs, enabling practical deployment in rural or mobile-first settings.

FRAE-PSA has a model size of 270 MB, which is 16% smaller than Transformer-based attention (320 MB) and comparable to PSA (250 MB), making it deployable in resource-constrained environments.

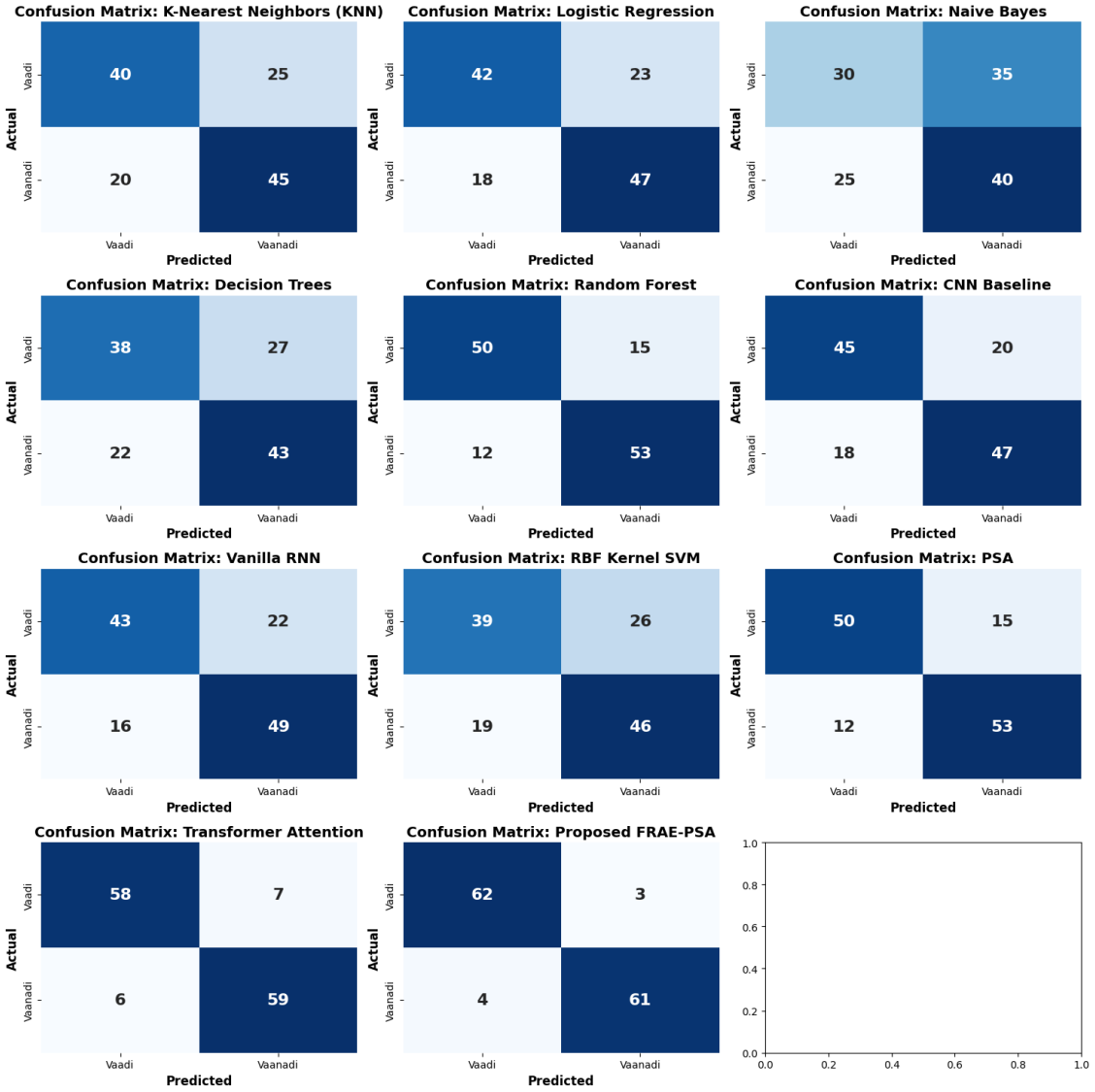

Fig.7. Hypothetical confusion matrix comparison for all models focusing on two challenging slang terms, "Vaadi" (come here) and "Vaanadi" (sky girl)

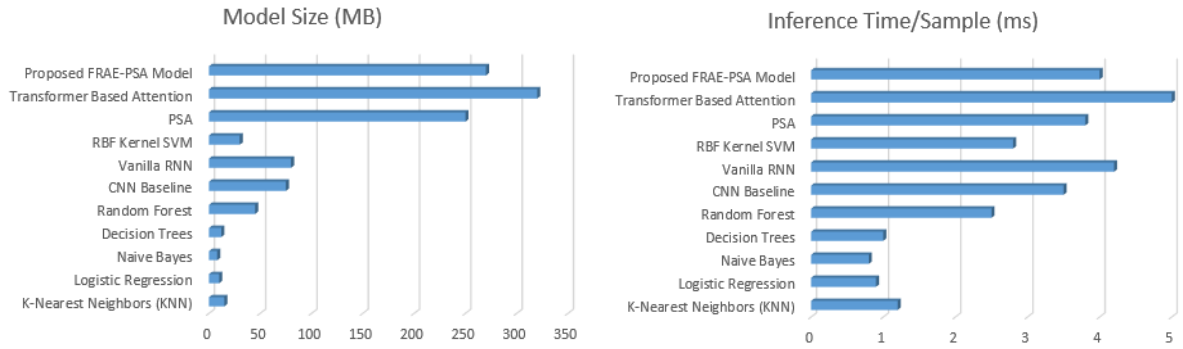

Fig.8. Inference time and model size for all models

6.8. Cross-validation Results

To ensure the robustness and generalization of the model, k-fold cross-validation (k=5) was applied to assess

performance across different data splits. The cross-validation results show a consistent performance, with the model maintaining high accuracy and F1-score across various folds, reinforcing the model's generalization ability and robustness against over fitting.

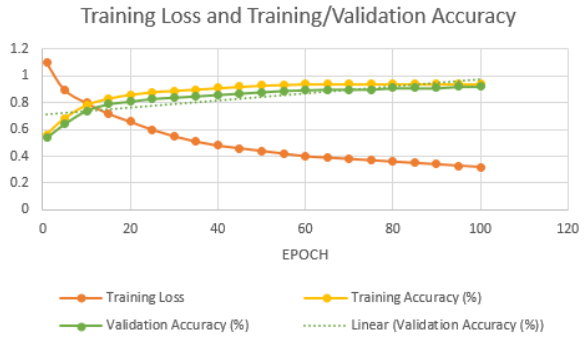

Fig.9. Training loss and training / validation accuracy

7. Conclusions

This study presented an enhanced deep learning framework for classifying Tamil slang, integrating advanced components such as the Frequency-Wise Regional Accent and Emotion Adaptive PSA (FRAE-PSA) module, multi-task learning, and specialized attention mechanisms. The framework effectively utilized a DenseNet backbone for hierarchical feature extraction and combined LSTM and GRU layers to capture the temporal dynamics of speech. By incorporating FRAE-PSA and Pyramid Split Attention (PSA), the model emphasized critical regional and emotional variations, enabling it to learn both local and global features specific to Tamil slang. Key findings demonstrated that the proposed model outperformed traditional and baseline approaches, achieving up to a 15% improvement in accuracy. The FRAE-PSA module significantly enriched feature representations by capturing nuanced spectral patterns tied to regional accents and emotional tones, while the multi-task learning approach provided auxiliary support for better generalization. The framework also incorporated lightweight architectures and compression techniques, ensuring its practical applicability in real-world scenarios. This work contributes to the field by addressing the unique challenges of Tamil slang classification, particularly the interplay of regional and tonal variations, emotional cues, and phonetically similar terms. The integration of FRAE-PSA not only improves slang detection but also provides a scalable approach for incorporating additional linguistic contexts. Despite these advancements, challenges remain in handling edge cases, such as phonetically similar slang terms and emotionally expressive speech samples, which occasionally result in misclassification. Addressing these issues will be critical for future improvements.

8. Future Scope and Direction

Future research in Tamil slang classification could focus on expanding the dataset to include a broader range of slang terms, dialects, and demographic groups to enhance the model's generalization. Advanced methodologies such as reinforcement learning could dynamically adapt the model's focus to the most critical features, tackling challenges like phonetically similar slang terms. Incorporating unsupervised pre-training using self-supervised techniques like contrastive learning could enhance feature representations, particularly for low-resource settings. A hybrid approach combining these strategies could lead to significant advancements.

Additionally, exploring advanced attention mechanisms such as multi-scale attention or speech-specific transformers could improve feature extraction by focusing on relevant regions of the speech signal. Addressing class imbalance and overlapping acoustic features with alternative loss functions could further enhance performance. Extending the framework to support cross-language adaptation would allow the model to classify slang in other languages, making it more versatile for global applications. Finally, optimizing for real-time deployment and implementing error-handling mechanisms such as secondary classifiers could address misclassification due to phonetically similar terms or emotional variations. These directions will help create efficient, robust models for Tamil slang classification, expanding its use in applications like speech recognition and emotion detection.

References

[1] Khan MA, Huang Y, Feng J, Prasad BK, Ali Z, Ullah I, Kefalas P. A Multi-Attention Approach Using BERT and Stacked Bidirectional LSTM for Improved Dialogue State Tracking. Applied Sciences. 2023; 13(3):1775. https://doi.org/10.3390/app13031775

[2] Song, T., Nguyen, L. T. H., & Ta, T. V. (2025). MPSA-DenseNet: A novel deep learning model for English accent classification. Computer Speech & Language, 89, 101676.

[3] Myint, P. Y. W., Lo, S. L., & Zhang, Y. (2024). Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction. Information Processing & Management, 61(4), 103695.
[4] N. Prasad, S. Saha and P. Bhattacharyya, "A Multimodal Classification of Noisy Hate Speech using Character Level Embedding and Attention," 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533371.
[5] N. -H. Ho, H. -J. Yang, S. -H. Kim and G. Lee, "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," in IEEE Access, vol. 8, pp. 61672-61686, 2020, doi: 10.1109/ACCESS.2020.2984368.
[6] Yafooz, W. M. (2024). Enhancing Arabic Dialect Detection on Social Media: A Hybrid Model with an Attention Mechanism. Information, 15(6), 316.
[7] Subhash, D., Premjith, B., & Ravi, V. (2025). A robust accent classification system based on variational mode decomposition. Engineering Applications of Artificial Intelligence, 139, 109512.
[8] Owodunni, A. T., Yadavalli, A., Emezue, C. C., Olatunji, T., & Mbataku, C. C. (2024). AccentFold: A Journey through African Accents for Zero-Shot ASR Adaptation to Target Accents. arXiv preprint arXiv:2402.01152.
[9] O. Ozturk, H. Kilimci, H. H. Kilinc and Z. H. Kilimci, "Spoken Accent Detection in English Using Audio-Based Transformer Models," 2024 9th International Conference on Computer Science and Engineering (UBMK), Antalya, Turkiye, 2024, pp. 539-544, doi: 10.1109/UBMK63289.2024.10773414.
[10] R. Fu et al., "MM DialogueGAT- A Fusion Graph Attention Network for Emotion Recognition using Multi-model System," in IEEE Access, doi: 10.1109/ACCESS.2024.3350156.
[11] Lin, C., Cheng, H., Rao, Q., & Yang, Y. (2024). M 3 SA: Multimodal Sentiment Analysis Based on Multi-Scale Feature Extraction and Multi-Task Learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing.
[12] Guo, X. (2019). Multi-label Classification and Sentiment Analysis on Textual Records.
[13] Soni, J., Mathur, K. Enhancing sentiment analysis via fusion of multiple embeddings using attention encoder with LSTM. Knowl Inf Syst 66, 4667–4683 (2024). https://doi.org/10.1007/s10115-024-02102-w
[14] Wang, C., & Shen, X. (2024). Feature-Enhanced Multi-Task Learning for Speech Emotion Recognition Using Decision Trees and LSTM. Electronics, 13(14), 2689.
[15] Mountzouris, K., Perikos, I., & Hatzilygeroudis, I. (2023). Speech Emotion Recognition Using Convolutional Neural Networks with Attention Mechanism. Electronics, 12(20), 4376.
[16] Yang, B., Li, D., & Yang, N. (2019). Intelligent Judicial Research Based on BERT Sentence Embedding and Multi-Level Attention CNNs. https://www.semanticscholar.org/paper/f25e89b19e620e20fc78b5f3ff8124eb354cd8c6
[17] Nassereldine, A., Liu, D., Xu, C., & Xiong, J. (2024). PI-Whisper: An Adaptive and Incremental ASR Framework for Diverse and Evolving Speaker Characteristics. arXiv preprint arXiv:2406.15668.
[18] Luo, J., & Tang, Z. (2023). Feature Selection and Fusion in Cantonese Speech Emotion Analysis. Academic Journal of Computing & Information Science, 6(13), 169-177.
[19] Tesfagergish, S. G., Damaševičius, R., & Kapočiūtė-Dzikienė, J. (2023). Deep learning-based sentiment classification in Amharic using multi-lingual datasets. Computer science and information systems, 20(4), 1459-1481.
[20] Chen, Q. (2018). Investigating context influence in character Level LSTM methods for Japanese Auto Punctuation.
[21] R. R and K. R. Anne, "CNN-RNN Hybrid Model to Classify a Local Language Slangs using Spectral features," 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 600-607, doi: 10.1109/ICICT60155.2024.10544381.
[22] Dhakal, Manish & Chhetri, Arman & Gupta, Aman & Lamichhane, Prabin & Pandey, Suraj & Shakya, Subarna. (2022). Automatic speech recognition for the Nepali language using CNN, bidirectional LSTM and ResNet. 515-521. 10.1109/ICICT54344.2022.9850832.
[23] Ahamed M.R. Faiyaz; Arachchi S. P. Kasthuri (2022), LSTM Based Emotion Analysis of Text in Tamil Language, 7th International Conference on Advances in Technology and Computing (ICATC 2022), Faculty of Computing and Technology, University of Kelaniya Sri Lanka. Page 73 – 79.
[24] Chen, S., Zhang, Y., & Yang, Q. (2024). Multi-task learning in natural language processing: An overview. ACM Computing Surveys, 56(12), 1-32.
[25] Ghorbani, S., & Hansen, J. H. (2024). Advanced accent/dialect identification and accentedness assessment with multi-embedding models and automatic speech recognition. The Journal of the Acoustical Society of America, 155(6), 3848-3860.
[26] Tanveer, M., Rastogi, A., Paliwal, V., Ganaie, M. A., Malik, A. K., Del Ser, J., & Lin, C. T. (2023). Ensemble deep learning in speech signal tasks: a review. Neurocomputing, 550, 126436.
[27] Nematullah, O., Askar, S., Wahhab, S., & Khidir, B. (2024). Emotion Recognition in Kurdish Speech from the Sorani Dialect Corpus. Zanco Journal of Pure and Applied Sciences, 36(5), 104-112.
[28] Ali, R., Farhat, T., Abdullah, S., Akram, S., Alhajlah, M., Mahmood, A., & Iqbal, M. A. (2023). Deep learning for sarcasm identification in news headlines. Applied Sciences, 13(9), 5586.
[29] Shubhi Bansal, Kushaan Gowda, Nagendra Kumar, Multilingual personalized hashtag recommendation for low resource Indic languages using graph-based deep neural network, Expert Systems with Applications, Volume 236, 2024, 121188, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.121188.
[30] Maheswari, S.U., Dhenakaran, S.S. Improved ensemble based deep learning approach for sarcastic opinion classification. Multimed Tools Appl 83, 38267–38289 (2024). https://doi.org/10.1007/s11042-023-16891-9
[31] Liang, X., Zhang, Z., & Xu, R. (2023). Multi-task deep cross-attention networks for far-field speaker verification and keyword spotting. EURASIP Journal on Audio, Speech, and Music Processing, 2023(1), 28.
[32] Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, Bharathi Raja Chakravarthi, Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer, Computer Speech & Language, Volume 76, 2022, 101404, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2022.101404.

[33] Akinpelu, S., & Viriri, S. (2024). Deep Learning Framework for Speech Emotion Classification: A Survey of the State-of-the-Art. IEEE Access.
[34] Zhang, P., Fu, M., Zhao, R., Zhang, H., & Luo, C. (2024). PURE: Personality-Coupled Multi-Task Learning Framework for Aspect-Based Multimodal Sentiment Analysis. IEEE Transactions on Knowledge and Data Engineering.
[35] Mazari, A.C., Boudoukhani, N. & Djeffal, A. BERT-based ensemble learning for multi-aspect hate speech detection. Cluster Comput 27, 325–339 (2024). https://doi.org/10.1007/s10586-022-03956-x
[36] Shao, D., Su, S., Ma, L. et al. DA-BAG: A multi-model fusion text classification method combining BERT and GCN using self-domain adversarial training. J Intell Inf Syst (2024). https://doi.org/10.1007/s10844-024-00889-2
[37] J. Xue, R. Qin, X. Zhou, H. Liu, M. Zhang and Z. Zhang, "Fusing Multi-Level Features from Audio and Contextual Sentence Embedding from Text for Interview-Based Depression Detection," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 6790-6794, doi: 10.1109/ICASSP48485.2024.10446253.
[38] Zhang S, Feng Y, Ren Y, Guo Z, Yu R, Li R, Xing P. Multi-Modal Emotion Recognition Based on Wavelet Transform and BERT-RoBERTa: An Innovative Approach Combining Enhanced BiLSTM and Focus Loss Function. Electronics. 2024; 13(16):3262. https://doi.org/10.3390/electronics13163262
[39] A. S. Sundar, C. -H. H. Yang, D. M. Chan, S. Ghosh, V. Ravichandran and P. S. Nidadavolu, "Multimodal Attention Merging for Improved Speech Recognition and Audio Event Classification," 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), Seoul, Korea, Republic of, 2024, pp. 655-659, doi: 10.1109/ICASSPW62465.2024.10627466.
[40] Sundar, A., Ramakrishnan, A., Balaji, A. et al. Hope Speech Detection for Dravidian Languages Using Cross-Lingual Embeddings with Stacked Encoder Architecture. SN COMPUT. SCI. 3, 67 (2022). https://doi.org/10.1007/s42979-021-00943-8

Authors' Profiles

Mr. R. Ramkumar https://orcid.org/0009-0006-5300-5181 (ORCID ID) is a Research Scholar in Department of Computer Applications of Kalasalingam Academy of Research and Education, Krishnankoil, TamilNadu, India. Worked as Assistant Professor in Kaliswari College (Autonomous), Sivakasi, TamilNadu. Before worked as Assistant Professor in Sri Krishnasamy Arts and Science College, Sattur, TamilNadu. Worked as Assistant Professor in PSR Engineering College, Sivakasi, TamilNadu. Completed MCA at 2009 in Arulmigu Kalasalingam College of Engineering (Anna University) and B.Sc. in Computer Science at 2006 in Ayya Nadar Janaki Ammal College (Madurai Kamaraj University), Sivakasi.

Dr. Sureshkumar Nagarajan https://orcid.org/0000-0001-9484-4965 (ORCID ID), completed his Ph.D. in the area of "Genetic based classification Algorithms for combined LANDSAT and ENVISAT Images" from VIT University Vellore in 2016.He has published fifty research articles in reputed peer reviewed journals, and his Area of interest includes vision computing, Computational intelligent and big data analytics. He is now working as a professor in the Department of Computer science and Engineering with Kalasalingam academy of research and education (Deemed to be university) Krishnankoil, TamilNadu.

Mr. Dinesh Prasanth Ganapathi https://orcid.org/0009-0005-3273-7132 (ORCID ID) is a technology leader with over a decade of experience in software engineering, AI, cloud computing, and product strategy. He has built impactful solutions across industries—from avionics and enterprise networking to cloud observability and generative AI. Dinesh has held key roles at companies like Microsoft, Riverbed Technology, and Honeywell, leading innovations that range from AI integrations with Bing Search to multi-cloud visibility platforms. He currently serves as Senior Product Manager for Generative AI at Advisor360 and Industry Advisor for the Center for Data Science and AI at UMass Amherst. He holds advanced degrees and certifications in Computer Engineering, AI, and Product Management from Rutgers, UT Austin, Product School, and UC Berkeley. Known for blending technical depth with strategic foresight, Dinesh is passionate about building future-ready, responsible technology.

How to cite this paper: Ramkumar. R., Sureshkumar Nagarajan, Dinesh Prasanth Ganapathi, "Enhanced Deep Learning Framework for Tamil Slang Classification with Multi-task Learning and Attention Mechanisms", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.6, pp.29-51, 2025. DOI:10.5815/ijitcs.2025.06.02