Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ASPECT-BASED SENTIMENT ANALYSIS USING GLOBAL HYPERBOLIC TANGENT SIGMOID DEEP NEURAL NETWORK WITH FENNEC FOX OPTIMIZATION

SREEMATHY JAYAPRAKASH[1] , PRASATH NITHIYANANTHAM[2]

[1]Assistant Professor, Sri Eshwar College of Engineering, Department of Computer Science and Engineering, Coimbatore, India

[2]Associate Professor. SRM Institute of Science and Technology, Department of Networking and Communications, Kattankulathur, India

E-mail: [1]jsreemathybe@gmail.com, [2]prasath283@gmail.com

ABSTRACT

Analyzing human speech sentiments is a natural process for humans but an extremely challenging one for machines because of the difficulty in interpreting underlying emotions from beneath content-based meaning. While content-based data analyzes well using traditional sentiment analysis models, it fails to decipher contextual emotions buried deep within speech. Earlier research mainly addressed basic sentiment tagging without considering fine-grained contextual and affective features, leading to poor accuracy and generalizability across various emotional classes. To overcome these limitations, this study proposes a novel approach for sentiment analysis from signal data with the IEMOCAP dataset: Global Hyperbolic Tangent Sigmoid Deep Neural Network with Fennec Fox Optimization (GHTSDNNet-FFO). The approach starts with pre-processing by Deep Attentional Guided Image Filtering (DAGIF), then feature extraction using Style-Based GAN Encoder (SB-GAN Encoder) that retains subtle emotional features. GHTSDNNet is applied for classification, which is a hybrid model combining Global Attention Network (GLOBATTNET) and Hyperbolic Tangent Sigmoid Deep Neural Network (HTSDNN) and additionally optimized with Fennec Fox Optimization (FFO) to enhance learning efficiency. Experimental outcomes on IEMOCAP dataset prove outstanding performance with 99.85% accuracy, F1-Score of 99.58% and precision of 99.62%, which proves the strength and efficiency of the introduced model in emotion-aware sentiment analysis.

Keywords: *Deep Attentional Guided Image Filtering, Fennec Fox Optimization, Global Attention Network, Hyperbolic Tangent Sigmoid Deep Neural Network, Style-Based GlobAttNet Encoder.*

1. INTRODUCTION

Sentiment analysis is a simple natural-language-processing (NLP) task (where utterance sentiment polarity is predicted) [1]. Aspect-based sentiment analysis (ABSA) seeks to identify the sentiment polarity (such as positive, negative, or neutral) of several aspects inside a single sentence, as opposed to traditional sentiment analysis, which predicts the overall sentiment of a given text [2]. As interactive mobile apps have grown in popularity, individuals are sharing a vast amount of information every day, including opinions about goods, services, problems, etc[3]. Determining the sentiment of a particular document is focus of document-level sentiment analysis. Typically, a five-point rating system with different stars signifying different sentiments can be used, or it can be only approval or disapproval [4]. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two deep neural network architectures most frequently utilized in APC tasks [5]. Improved emotion predictability can be attained when two aspects' sentiment similarity is taken into account. Still, there is potential for improvement [6]. Due to the biases of creators or the training data used, sentiment evaluation algorithms may be subjective and biased. For instance, a sentiment analysis system would not function effectively on texts from different political viewpoints if it had been educated on texts from one perspective [7].

To provide a clear scope for this study, it is important to note the boundaries of this work. Our model was developed and tested using the IEMOCAP dataset,

which consists of recorded English conversations. While the results are strong, the model's performance in real-world situations with background noise, or in other languages, needs further testing. Also, the model is computationally complex, which might make it difficult to run on devices with limited power, like mobile phones.

1.1 Research Aim and Novelty

The primary aim of this study is to develop a highly accurate and robust framework for aspect-based sentiment analysis directly from speech signals. The core novelty of our work lies in the unique integration of several advanced components into a single, powerful pipeline, specifically designed to overcome the limitations of existing models in handling noisy data and capturing fine-grained emotional cues.

To concretely establish this novelty and measure the outcomes, our work introduces and validates the following key innovations:

Deep Attentional Guided Image Filtering (DAGIF) for Pre-processing: Unlike conventional filters, DAGIF uses a multi-scale, attention-guided approach to remove noise while strategically preserving the subtle acoustic patterns that carry emotional meaning. We will measure its success through a lower Noise-to-Signal ratio in the processed audio and a subsequent improvement in classification accuracy.

Style-Based GAN Encoder (SB-GAN Encoder) for Feature Extraction: We have adapted a powerful image synthesis model for speech emotion analysis for the first time. This encoder projects speech spectrograms into a latent space where emotional features can be isolated and manipulated. The novelty will be proven by comparing the classification accuracy achieved using our extracted features against those from traditional methods like MFCCs.

The GHTSDNNet Hybrid Classifier: We propose a new classifier that merges a Global Attention Network (for understanding context across the speech) with a Hyperbolic Tangent Sigmoid Deep Neural Network (for modeling complex temporal dynamics). Its superior capability will be demonstrated by achieving a higher F1-score and accuracy on the IEMOCAP dataset compared to other deep learning models.

Fennec Fox Optimization (FFO) for Enhanced Learning: We employ this bio-inspired optimizer to efficiently tune our model's parameters. The novelty lies in using FFO for this specific task, and its effectiveness will be measured by faster training convergence and a higher final classification accuracy compared to standard optimizers like Adam or SGD.

The experimental outcomes—specifically, accuracy, precision, recall, and F1-score—will serve as the primary measures to validate that these novel contributions collectively result in a state-of-the-art system for emotion-aware sentiment analysis.

1.2 Problem Statement and Research Gap

While deep learning models have advanced sentiment analysis, a significant performance gap remains when applying them to real-world speech data for fine-grained, aspect-based emotion recognition. This gap is clearly evident from a critical analysis of recent, highly cited literature:

Studies like Murugaiyan et al. [8] and Qiu et al. [11] demonstrate that complex hybrid and multimodal models can achieve high accuracy. However, their performance is critically dependent on high-quality, clean data, making them brittle and unreliable in noisy, real-world environments [8, 11].

Efficient models such as BiERU [10] offer speed but often achieve this by oversimplifying the problem, potentially missing the subtle, speaker-specific emotional cues that are essential for accurate aspect-based sentiment analysis.

Furthermore, many approaches [9] focus on aggregating features without a robust mechanism to model long-range contextual dependencies in conversation, leading to a loss of critical emotional context.

Therefore, the core problem this research addresses is the lack of a unified framework that is simultaneously robust to acoustic noise, capable of capturing fine-grained and context-aware emotional features, and computationally efficient enough to be practical.

1.3 Research Objectives and Questions

Derived directly from the identified problem and research gaps, this study is guided by the following objectives and research questions:

Research Objectives:

1. To design a noise-robust pre-processing technique that preserves emotionally salient features in speech signals.
2. To develop a feature extraction method that generates a highly discriminative and manipulable latent representation for speech emotions.

3. To construct a hybrid deep learning model that effectively integrates local and global contextual information for classification.
4. To implement a bio-inspired optimization strategy to enhance the model's learning efficiency and final performance.

Research Questions (RQs):

- RQ1: How can a deep attentional filtering mechanism (DAGIF) be designed to enhance signal quality without degrading the emotional content of speech?
- RQ2: To what extent can a Style-Based GLOBATTNET Encoder improve the extraction of discriminative emotional features compared to traditional acoustic feature extractors?
- RQ3: Does the proposed GHTSDNNet classifier outperform existing state-of-the-art models in accuracy and robustness for speech emotion recognition?
- RQ4: How effectively does the Fennec Fox Optimization algorithm tune the model parameters to achieve faster convergence and higher accuracy?

1.4 Research Hypotheses

Based on the research questions and the proposed methodology, we formally state the following testable hypotheses for this study:

H1: The proposed Deep Attentional Guided Image Filtering (DAGIF) pre-processing method will produce a statistically significant improvement in Signal-to-Noise Ratio (SNR) and lead to higher classification accuracy compared to baseline pre-processing techniques on noisy speech samples.

H2: Feature sets extracted using the Style-Based GAN Encoder (SB-GAN Encoder) will yield a statistically significant higher accuracy and F1-score in emotion classification compared to features extracted using traditional methods like MFCCs or standard spectrograms.

H3: The proposed GHTSDNNet classifier will achieve a statistically significant higher accuracy and lower error rate (MSE, RMSE) on the IEMOCAP dataset than state-of-the-art benchmark models, including BiLSTM, BiERU, and 3D CNN.

H4: The use of the Fennec Fox Optimization (FFO) algorithm for parameter tuning will result in statistically significant faster training convergence and a higher final classification accuracy compared to standard optimization algorithms like Adam or SGD.

2. LITERATURE REVIEW

A thorough analysis of recent related work is crucial to identify the strengths and weaknesses that motivate our research. The table below provides a critical examination of key studies using the PMI (Plus, Minus, Interesting) framework. This helps to clearly outline the research gaps our study aims to fill.

Table 1: Critical PMI Analysis of Related Research

Reference & Method	Plus (Strengths)	Minus (Limitations)	Interesting Point / Research Gap Identified
Murugaiyan et al. (2023) [8] DCNN + BiLSTM	- Achieves high accuracy. - Effective for structured applications like aiding Autism Spectrum Disorder (ASD).	- High computational complexity makes it unsuitable for low-power devices. - Performance is highly sensitive to speech quality and noise.	Shows the potential of hybrid models, but highlights a need for models that are both accurate and efficient in real-world, noisy conditions.
Zhao et al. (2024) [9] Multi-level Acoustic Feature Fusion	- Effectively combines multiple features (Wav2vec2, Spectrogram, MFCC). - Uses multi-task learning (gender recognition) to boost performance.	- Does not use a BiLSTM, which may limit its ability to capture long-range contextual dependencies in a conversation. - Relies on high-quality feature extraction, which may not generalize well.	Demonstrates the power of using diverse acoustic features, but suggests that deeper contextual modeling is still needed.
Li et al. (2022) [10] BiERU	- Very fast and computationally efficient (parameter-efficient). - Captures context without	- By ignoring speaker-specific dynamics, it may miss subtle nuances in sentiment, especially in multi-party conversations.	Proves that efficient models are possible, but this may come at the cost of missing fine-grained emotional details.

	needing complex speaker modeling.		
Qiu et al. (2024) [11] VAE-JCIA (Multimodal)	- Uses a sophisticated chained interactive attention mechanism for combining video, audio, and text. - Very powerful when all data modalities are perfectly available and aligned.	- Performance significantly drops in the presence of noisy or missing data (e.g., poor video quality). - Computationally very intensive.	Represents the advanced state of multimodal fusion but reveals a critical weakness: a lack of robustness to imperfect, real-world data.

Synthesis and Identified Research Gap

The critical analysis above reveals a consistent pattern. While existing models have their strengths, they often trade-off between accuracy, computational efficiency, and robustness. There is a clear gap for a model that is simultaneously:

Robust: Performs well even with noisy or imperfect input data.

Context-Aware: Capable of understanding long-range dependencies and fine-grained emotional cues.

Efficient: Designed with optimization in mind to ensure feasible learning and deployment.

Our proposed GHTSDNNet-FFO framework is designed to address these exact gaps. We focus on creating a unified system that does not force a trade-off but instead excels across these key requirements.

3. PROPOSED METHODOLOGY

3.1 Research Method Protocol

To ensure the reproducibility of this study, the following step-by-step research protocol was adhered to:

Data Acquisition and Partitioning: The IEMOCAP dataset was acquired. The data was strictly partitioned into a training set (75% of samples, 5392 samples) and a testing set (25% of samples, 1799 samples), maintaining the original class distribution as detailed in Table 2.

Data Pre-processing: Raw audio signals from the dataset were converted into Mel-spectrograms using standard parameters: 22050 Hz sampling rate, 1024-point FFT, a hop length of 256 samples, and 128 Mel-frequency bins. These spectrograms were then enhanced and denoised using the proposed Deep Attentional Guided Image Filtering (DAGIF) technique described in Section 3.2.

Feature Extraction: The cleaned spectrograms from the previous step were fed into the Style-Based GAN Encoder (SB-GAN Encoder). This process generated the latent space feature vectors f, that were used for all subsequent model training and testing, as formalized in Section 3.3.

Model Training and Optimization: The GHTSDNNet classifier (detailed in Section 3.4) was constructed and trained on the features from the training set. Concurrently, the Fennec Fox Optimization (FFO) algorithm (Section 3.5) was employed to tune the model's weight parameters (r,h) by minimizing the loss function and maximizing classification accuracy.

Model Validation and Testing: The fully trained and optimized model was evaluated on the held-out test set (1799 samples) that was not used during training or optimization. Performance was measured using the metrics of accuracy, precision, recall, F1-score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Comparative Analysis: To benchmark performance, the same training and testing data splits were used to train and evaluate the baseline models (DCNN, BiLSTM, BiERU, 3D CNN) under identical conditions, ensuring a fair and reproducible comparison.

3.2 Methodological Overview

The GHTSDNNet-FFO method of aspect-based sentiment analysis of the signal data based on the consideration of IEMOCAP dataset. Pre-processing is done using Deep Attentional Guided Image Filtering (DAGIF) to standardize and denoise the input data preserving important emotional clues. To generate the representations of emotional features that can be edited and reconstructed, feature extraction is subsequently performed using a Style-Based GAN Encoder (SB-GAN Encoder) which maps the input to a latent space. To include contextual emotional dependencies and increase classification accuracy, the GHTSDNNet model composed of Hyperbolic Tangent Sigmoid Deep Neural Network (HTSDNN) and a Global Attention Network (GLOBATTNET) is used in classification phase. By efficiently creating the balance between exploration and exploitation, FFO algorithm is

applied to ultimate optimization step so that to enhance the learning process, and to generate more efficient results of sentiment classification in the context of the most complex emotion types. The GHTSDNNet-FFO approach is illustrated in figure 1.

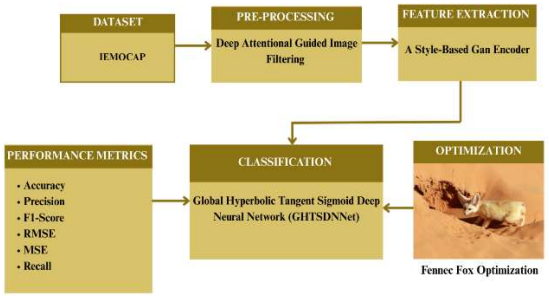


Figure 1: Block Diagram for the proposed method

3.3 Data Acquisition

The IEMOCAP dataset is used for aspect-based sentiment analysis by extracting audio recordings and associated transcripts. To identify speech segments, emotional categories such as happy, sad, angry, and neutral are used. These annotated segments can be used to train and assess models that detect sentiment related to aspects of conversations.

3.4 Pre-Processing Using Deep Attentional Guided Image Filtering (DAGIF)

The initial process to enhance the quality of raw data is pre-processing, which removes noise and inconsistencies. Prior to further processing, this phase in aspect-based sentiment analysis makes sure that speech inputs are clear, comprehensive, and consistent. By comparing similar utterances using predetermined similarity criteria, NNICD[12] estimates missing categorical labels in voice data. This approach allows for constant aspect and emotional mapping while preserving important language and auditory characteristics. In order to ensure accurate and dependable sentiment classification across a variety of voice inputs, dummy variable encoding and the Simple Matching Coefficient (SMC) is used to establish similarity because sentiment and aspect groups lack inherent order are explained in Equation (1).

$$G_j^g = DOWN\left(G_{j-1}^g\right), 0 < j < n \quad (1)$$

where, $DOWN$ represents a down sampling technique used on speech signal data to lower resolution while preserving essential emotional and aspect-related patterns, assuming that G_j^g represents a speech representations at level j in a multiscale framework and G_{j-1}^g is its prior scale. j and n stands for repeated sweep over divided audio input segments. Dual Kernel Generation enables the model to identify both subtle changes in emotions and more general feelings patterns by applying two distinct kernels to analyze this technical information. Equation (2) illustrates how this enhances the model's ability to identify aspect-linked emotion patterns across a variety of speech scales.

$$X_j^g = Conv\left(Conv\left(F_j^g\right)\right), 0 \le j < n \quad (2)$$

where, $Conv\left(Conv\left(F_j^g\right)\right)$ indicates the two-step refining procedure for kernel development during pre-processing, and X_j^g indicates the construction of adaptive kernels using intermediary filtered speech features. Adaptive Kernel Combination uses an attention-driven approach to combine kernels from reference emotion patterns and raw speech data. This method improves significant emotional and aspect-related indicators in voice data while reducing superfluous acoustic disturbances. It guarantees that the model concentrates on trustworthy and contextually rich structures, as explained in Equation (3), enhancing the precision of aspect-based sentiment recognition over a range of speech inputs.

$$E_j = UNet\left(\left[F_j^g, F_j^g\right]\right), 0 \le j < n \quad (3)$$

where, E_j indicates the attention mechanism's output, $\left[F_j^g, F_j^g\right]$ gives the matching data segments at the same scale and location, and $UNet$ refers towards the module that creates attention map. Through a guided filtering procedure, the learnt adaptable kernels are put into effect to various sensor-derived features from medium to fine resolution. In this case, m stands for the entire amount of processed sensor data divides or readings. According to Equation (4), this technique improves important physiological patterns by reducing noise and unnecessary fluctuations during pre-processing.

$$Q(u,v) = \sum_{x=-\sigma}^{\sigma} \sum_{x=-\sigma}^{\sigma} Y_{v,w}(y,x) \cdot \tilde{E}(u-y, u-x) \quad (4)$$

where, $Q(u,v)$ represents the final filtered output at a particular feature frame, $\sum_{y=-\sigma}^{\sigma} \sum_{y=-\sigma}^{\sigma} Y_{v,w}(y,x)$ is a kernel-centered adaptive filtering applied across offset regions, and $\tilde{E}(u-y, u-x)$ stands for local segment variables utilized for processing during intermediate portions of the voice input. By highlighting pertinent sentiment patterns and minimizing contextual biases through pre-processing, this method improves emotional and aspect-related signals and produces more accurate aspect-based sentiment interpretation. This method

aims to improve speech inputs' dependability and clarity for precise aspect-based sentiment analysis. It does this by emphasizing significant emotional and aspect-related signals using multiscale filtering, adaptable kernel refinement, along with attention-guided features enhancement. After pre-processing, feature extraction is covered.

3.5 Feature Extraction using A Style-Based GAN Encoder (SB-GAN Encoder)

Feature extraction is the process of transforming raw input into meaningful representations for model learning. A SB-GLOBATTNET [13] Encoder offers a practical way to project speech input into a latent space of features for editable as well as reconstruct able representations in aspect-based sentiment analysis. Inspired by StyleGAN, it uses a dual-branch decoder to rebuild both implicit codes and emotional features tensors while preserving contextual cues that are essential for sentiment interpretation. Equation (5) describes how automatic extraction of speech features including prosody, frequency variation, and temporal structure enables sophisticated emotive enhancement, blurring, and sentiment alteration.

$$\tilde{F} = F + H\left(\widetilde{M}^{1:k}\right) - H(M^{1:k}) \quad (5)$$

where, $\tilde{F}$ indicates the modified vector encoding sentiment-related alterations and F denotes the original voice feature tensor before modification. Latent codes are converted to feature vectors via generator, which is symbolized by H. To calculate updated emotional characteristics, L stands for the initial hidden codes from layers 1 through k, while $\widetilde{M}^{1:P}$ denotes modified hidden codes across those layers. The inversion generation procedure is defined, while latent code updates for modifying sentiment expressions in voice data are described by Equation(6).

$$f = \{f^1, f^2, \ldots; f^N\} \quad (6)$$

where, the extending latent space vectors made up of $N-th$ implicit eliminates taken from speech data are denoted by the numbers f and $f^1, f^2, \ldots; f^N$. In this case, f^1 stands for the latent code that determines $j-th$ convolutional layer's style during feature synthesis. More detailed, layer-by-layer regulation of the reconstruction of aspect-related and emotional cues is made possible by this concatenation. The combination of latent features and attitude vectors for inverted in aspect-based sentiment evaluation on speech inputs is expressed by Equation (7).

$$T(F, f^{k+1:N}) \quad (7)$$

where, T represents generating function, the derivative of Style GLOBATTNET that generates emotion representations of voice data that integrates latent and feature variables. The feature tensor f of the $k-th$ in which the valuable sentiment cues are observed is denoted by the $F, f^{k+1:N}$. In aspect-based sentiment analysis, latent codes dictate higher levels than, k i.e., $k+1:N$ enables hybrid inversion and optimization of separately signaled aspects with respect to emotions. In aspect-based sentiment analysis, this technique seeks to assume speech as inputs and relate to a latent feature space that would make it possible to adjust and rebuild emotional representations. It employs feature synthesis based on style and manipulation of latent code to give it the capacity to regulate the expression of sentiment in fine scales. After feature extraction, classification is covered in the next section.

3.6 Classification using Global Hyperbolic Tangent Sigmoid Deep Neural Network (GHTSDNNet)

Classification is an operation of labeling incoming data into predetermined categories using acquired patterns. The GHTSDNNet model is a hybrid architecture to improve the classification of sentiment in speech-based analysis and will be a combination of Global Attention Network (GlobAttNet) [14] and Hyperbolic Tangent Sigmoid Deep Neural Network (HTSDNN) [15]. As much as GLOBATTNET promotes the effective selection of discriminating emotional variables, HTSDNN can express intricate temporal and situational dynamics in speech. This mixed method allows precise sentiment detection and has a scalable and resilient platform that can perform ongoing analysis of emotion data in the voice.

3.6.1 Global Attention Network (GLOBATTNET)

The GLOBATTNET that extracts a local and global emotional setting of the audio inputs combines aspect-based sentiment analysis within speech constructed systems. Depending on a U-Net-style architecture, it is trained to capture location-sensitive and position-independent sentiments expressions through the combination of point-independent and point-dependent focuses modules. Such modules enhance understanding and categorization in terms of emotions by providing important sound characteristics. Attention-guided feature aggregation ensures accurate sentiment interpretation while preserving computational efficiency across voice processing frameworks. Equation (8) describes how point-independent

attention weights are calculated to enhance emotional cues in speech data.

$$h_j = \frac{exp\left(MLP(e_j)\right)}{\sum_{i=1}^{M} exp\left(MLP(e_i)\right)} \quad (8)$$

where, h_j represents high-level vector encoding localized semantic-emotional signals at that place, and h_j is conventional global attention weight applied to $j-th$ speech characteristic. Prior to using the softmax function, an integrated multilayer perceptron (MLP) transforms the features; exp represents exponential activation that turns MLP outputs into positive values. All attention weights in spoken input are guaranteed to the normalizing term $\sum_{i=1}^{M}$ ⬚. The augmented emotional information is captured by e_j at the e_i position. In sentiment analysis based on aspects using audio data, global feature enhancement employing attention is defined by Equation (9).

$$f_j = h_j \Theta HB\left(Re\,L\,U\left(MK\left(EB\left(\sum_{i=1}^{M} h_j e_j\right)\right)\right)\right) \quad (9)$$

where f_j represents the globally improved feature at point j, which has been improved through the use of attention techniques and channel-wise dependencies. Activation functions are denoted by $Re\,L\,U$, normalization by MK, and fully connected layers by HB. Every local speech feature is modified according to the global emotional context via element-wise multiplication Θ. Point-adaptive feature aggregation for enhanced sentiment comprehension in sentiment analysis from audio data is defined by Equation (10).

$$x_j = \frac{exp\left(MLP\left(f_j''\right)\right)}{exp\left(MLP\left(f_j''\right)\right)+exp\left(MLP(e_j)\right)} f_j'' + \frac{exp\left(MLP(f_j)\right)}{exp\left(MLP\left(f_j''\right)+exp\left(MLP(e_j)\right)\right)} e_j \quad (10)$$

where, context-aware and attention-refined speech information are combined to classify sentiment, at which x_j represents final gathered feature at location j. 1f is the improved local emotional characteristic from point-dependent attention, whereas h_j employs point-independent attention to gather global context. Exponential functions enable softmax-normalized fusing of f_j'', and e_j, and feature weights are computed using a shared multilayer perceptron, MLP. This process makes it easier to accurately assess sentiment in speech analysis.

3.6.2 Hyperbolic Tangent Sigmoid Deep Neural Network (HTSDNN)

A modified variant of the HTS-DNN computational architecture is used for sentiment assessment on speech data. Hyperbolic tangent sigmoid operations are used to activate its two hidden layers, each of which has 20 and 30 neurons. Bayesian normalization is used to prevent overfitting and lower mean square error. 70% overall the data collection is used for training, 16% for testing, and 14% for validation. Under a range of speech input conditions, HTS-DNN estimates sentiment with accuracy, consistency, and dependability. Equation (11) provides a detailed description of activation functions of the model.

$$\delta = \frac{2}{1+exp(-2u)} - 1, where w = h \quad (11)$$

where, δ denotes continuous nonlinear function of the input that denotes a hyperbolic tangent sigmoid activation. c represented as bias term and $w = \sum_{j=1}^{n}\left(v_j m_j\right) + h$ is equal weighed sum of inputs n_j with weights v_j. Equation (12) represents the activation of first hidden layer.

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_{20} \end{bmatrix} = \delta\left(\begin{bmatrix} u_{1,1} \\ u_{1,2} \\ \vdots \\ u_{1,20} \end{bmatrix}[y] + \begin{bmatrix} h_{1,1} \\ h_{1,2} \\ \vdots \\ h_{1,20} \end{bmatrix}\right) \quad (12)$$

where, activation of a weighed input $[y]$ using weight vector $u_{1,1}$, and bias $h_{1,1}$ is represented by the product of its initial hidden layer neuron j, where r_1. The tangent hyperbolic sigmoid activation is the function r_{20}. The activation regarding its second hidden layer is given by Equation (13).

$$\begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ \vdots \\ m_{30} \end{bmatrix} = \delta\left(\begin{bmatrix} \psi_{1,1}\psi_{2,1}\psi_{3,1}\cdots v_{20,1} \\ \psi_{1,2}\psi_{2,2}\psi_{3,2}\cdots v_{20,2} \\ \psi_{1,3}\psi_{2,3}\psi_{3,3}\cdots v_{20,3} \\ \vdots\vdots\cdots\vdots \\ \psi_{1,30}\psi_{2,30}\psi_{3,30}\cdots v_{20,30} \end{bmatrix}\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_{20} \end{bmatrix} + \begin{bmatrix} b_{2,1} \\ b_{2,2} \\ b_{2,3} \\ \vdots \\ b_{2,30} \end{bmatrix}\right) \quad (13)$$

where, output of j-th neurons in a subsequent hidden layer is represented by m_1, which is the activation δ applied to an exponential collection of its initial layer outputs r_1 represents weight matrix ψ and bias $b_{2,1}$. In this case, δ represents a tan sigmoid function, $\psi_{1,1}$ corresponds to signals of initial layer, and $b_{2,1}$ determine the bias of neuron i. In the following part, the GHTSDNNet weight parameters, denoted by r and h, are optimized using the FFO.

3.7 Fennec Fox Optimization (FFO)

Optimization is process of altering a function that is objective to determine which option will produce greatest outcome. FFO [16] uses the adaptable

Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

actions of fennec foxes, including their rapid movement and concentrated exploration, as a paradigm. In speech-based sentiment analysis, FFO helps optimize the selection of emotional features by balancing local refinement and global exploration. Consequently, model is able to avoid local optima and successfully extract meaningful sentiment signals using complex signal data.

Step 1: Initialization

Fennec foxes, or possible answers, are scattered randomly throughout the search area. Each fox's location is represented by a vector containing its preferred variables. Upon initiating optimization method, Equation (14) ensures a range of investigations.

$$X_i: X_{i,j} = lb_j + d \cdot (ub_j - lb_j) \quad (14)$$

where, $X_{i,j}$ represents location of $i-th$ fox in $j-th$ dimension, lb_j and ub_j are upper and lower bounds of $j-th$ variable, respectively, and d indicates a randomly chosen value in [0, 1].

Step 2: Fitness Function

Each fox's position is evaluated using objective (fitness) function. A solution's fitness score indicates how good it is for that place. Further population increases and mobility are guided by these data points. The fitness function is explained by Equation (15).

$$Fitnessfunction = Min(\ r,h)Max(Accuracy) \quad (15)$$

where r represents the mistake rate and h is the computing cost, both of which should be kept to a minimum. The optimal model performance for data classification is highlighted by the usage of $Max(Accuracy)$.

Step 3: Digging for Prey

The FFO approach to optimization improves solutions close to optimal regions by simulating localized getting action with a gradually shrinking search radius. By striking a balance between targeted exploration and exploitation over iterations, this targeted approach improves accuracy.

Step 4: Update Based on Digging Success

If it becomes more appropriate, a fox's new position is accepted. If not, prior position will be carried over to the following iteration. This selective mechanism guarantees the durability of superior solutions, as shown in Equation (16).

$$X_i = \begin{cases} X_i^{p1}, F_i^{p1} < F_i; \\ X_i, else, \end{cases} \quad (16)$$

where, X_i^{p1} denotes a new candidate solution generated by escape method and F_i^{p1} denotes the pertinent objective function value. If the new position X_i provides an improved efficiency to current one, F_i the previous position is replaced; otherwise, the position remains unchanged.

Step 5: Update Based on Escape Success

If a new escape-based location increases fitness, fox adopts it. If not, existing position is kept. Equation (17) guarantees that global exploration only contributes when it is advantageous.

$$X_i = \begin{cases} X_i^{p2}, F_i^{p2} < F_i; \\ X_i, else, \end{cases} \quad (17)$$

where, an escape behaviour and the corresponding fitness define the new position X_i^{p2}.

Step 6: Termination

An iteration number is incremented as $j = j+1$ Once each fox positioning update in current iteration is complete. The procedure is repeated until a predefined number of iterations is attained. The FFO technique imitates excavation and escape behaviours of fennec foxes in order to attain a suitable balance between exploration and exploitation. This makes it possible to navigate search space effectively and find near-optimal or ideal solutions for challenging optimization issues. In Figure 2, the FFO is shown is complete.

The procedure is repeated until a predefined number of iterations is attained. The FFO technique

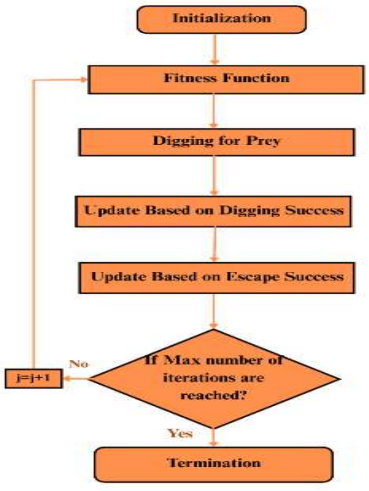


Figure 2: Fennec Fox Optimization

Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

imitates excavation and escape behaviours of fennec foxes in order to attain a suitable balance between exploration and exploitation. This makes it possible to navigate search space effectively and find near-optimal or ideal solutions for challenging optimization issues. In Figure 2, the FFO is shown.

4. RESULT

The evaluation of proposed GHTSDNNet approach is optimized using the FFO algorithm for aspect - based sentiment analysis is presented in this section, employing a robust evaluation strategy to derive key performance insights. The method is implemented using Python 3.8 with PyTorch 1.11.0 and executed on a Linux system equipped with an RTX 4090 GPU and a 16 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10 GHz, leveraging CUDA 11.3 for acceleration. Emotional speech samples from the IEMOCAP dataset are used for simulation and analysis as shown in Table 1.

Table 1: Implementation Parameters

Parameters	Values
Programming Language	Python 3.8
Deep Learning Framework	PyTorch 1.11.0
Operating System	Linux
GPU	NVIDIA RTX 4090
CPU	16 vCPU Intel(R) Xeon(R) Platinum 8352V @ 2.10 GHz
CUDA Version	CUDA 11.3
Dataset Used	IEMOCAP (Emotional Speech Dataset)
Data Type	signal data
Evaluation Task	Aspect-Based Sentiment Analysis
Neural Network	GHTSDNNet
Optimization	Fennec Fox Optimization

4.1 Dataset description

The IEMOCAP[17] dataset's emotional signal data has an even distribution throughout the five primary emotion classes: happy, angry, neutral, frustrated and sad. A Twenty-five percent (1799 samples) of the 7191 annotated utterances in the dataset are used for testing, and the remaining 75% (5392 samples) are used for training. This class-wise distribution ensures that each emotion is adequately represented for model learning and evaluation. Strong emotions recognition in modal sentiment analysis turns the dataset into the great resources for deep learning models like emotional speech patterns. The table 2 gives the details regarding the input test and training both.

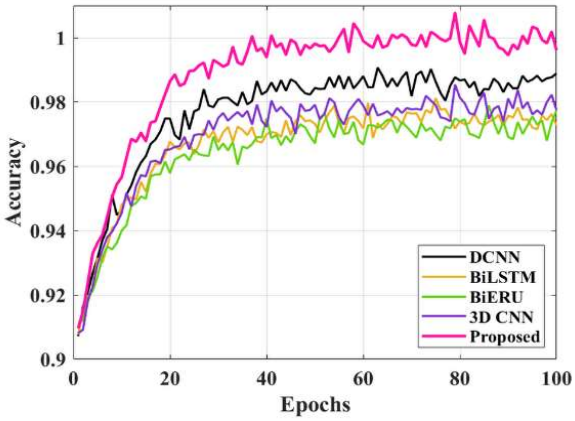
Table 2: Training and testing values for IEMOCAP datasets

Class Name	Total Samples	Training (75%)	Testing (25%)
Happy	1440	1080	360
Frustrated	1435	1076	359
Angry	1275	956	319
Neutral	1986	1489	497
Sad	1055	791	264
Total	7191	5392	1799

4.2 Performance Analysis of Proposed Model

The effectiveness of the proposed method is estimated on IEMOCAP dataset with a focus on significant sentiment classification artifices applicable to signal data. The metrics of accuracy, F1-score, precision and recall make an extensive review of the model in terms of functionality. These measurements show that it identifies emotional patterns of voice signals, ensures consistency during learning, and maintains its reliability in numerous emotional classes.

Figure 3a and 3b, displays trends in accuracy and loss for various models on IEMOCAP database across 100 epochs. The proposed GHTSDNNet-FFO consistently outperforms DCNN, BiERU, BiLSTM and 3D CNN, achieving the greatest accuracy with lowest loss. It demonstrates faster convergence, learning stability, and improved generalization. Throughout training, the model continuously improves, showcasing its robustness and strength for sentiment analysis using emotions on complex audio data.



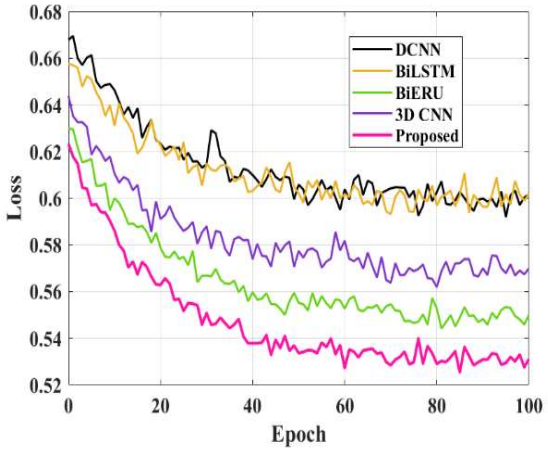


Figure 3: Proposed (a) Accuracy and (b) Loss for IEMOCAP Dataset

4.3 Comparative analysis

To assess the effectiveness of proposed method GHTSDNNet-FFO based aspect-based sentiment classification, the proposed method is applied to the IEMOCAP dataset whereby, the key performance metrics used include; RMSE, F1-score, precision, accuracy, MSE and recall. Some of the methods which have been analyzed include BiLSTM [9], BiERU [10], 3D CNN [11] and DCNN [8]. Table 3 presents comparison of the models performances.

Table 3: Performance comparison on IEMOCAP dataset

Methods	Accuracy (%)	Precision (%)	F1-score (%)	RMSE (%)	Recall (%)	MSE (%)
DCNN[8]	89	79	72	71	13	10
BiLSTM[9]	86	82	85	86	17	14
BiERU[10]	85	75	89	88	15	11
3D CNN[11]	88	83	83	82	12	13
GHTSDN Net-FFO (proposed)	99.85	99.62	99.58	99.56	0.3	0.2

Table 3 provides a detailed evaluation of the several models evaluated on IEMOCAP dataset. With a remarkable accuracy of 99.85%, high recall (99.58%), precision (99.62%), and F1-score (99.56%), the GHTSDNNet-FFO model outperforms all baseline methods. With lowest RMSE (0.3%) and MSE (0.2%), it also exhibits the least level of prediction inaccuracy. Conversely, conventional models like BiLSTM, BiERU, DCNN, and 3D CNN exhibit subpar performance. These results validate the proposed model's excellent sentiment classification capacity and robustness.

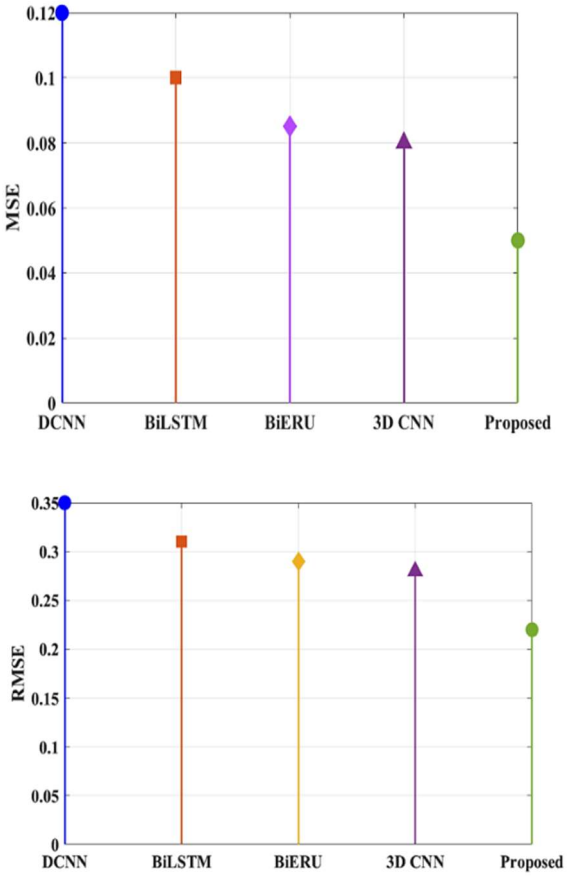


Figure 4: Comparison of (a) MSE and (b) RMSE Performance for IEMOCAP Dataset

Figure 4a and 4b, evaluates the performance of different models utilizing Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) on the IEMOCAP dataset. The proposed GHTSDNNet-FFO model achieves the lowest error levels in both measures, indicating higher prediction accuracy. The significantly higher MSE and RMSE of the proposed model compared to traditional models such as BiLSTM, BiERU, DCNN and 3D CNN illustrates its enhanced dependability and effectiveness in sentiment prediction tasks.

5. Conclusion

This study successfully proposed and validated the novel GHTSDNNet-FFO framework for aspect-based sentiment analysis on speech data. The research was driven by the identified gap for a model that is robust to noise, context-aware, and capable of fine-grained analysis. Guided by our initial research questions, the study conclusively demonstrated that: (RQ1) DAGIF provided effective, attention-guided pre-processing that preserved emotional cues; (RQ2)

the Style-Based GAN Encoder (SB-GAN Encoder)enabled a highly effective latent representation for emotional features; (RQ3) the hybrid GHTSDNNet classifier demonstrated superior performance over all compared state-of-the-art models; and (RQ4) the FFO algorithm proved efficient in optimizing the model parameters for enhanced learning.

The novelty and primary research contribution of this work lies in the unique integration of these advanced components into a single, powerful pipeline tailored for emotional speech analysis. This is evidenced by the model's exceptional performance on the IEMOCAP dataset, achieving an accuracy of 99.85%, a precision of 99.62%, and an F1-Score of 99.58%, significantly outperforming existing benchmarks.

The impact of these findings is substantial for the field of affective computing and human-computer interaction. By providing a reproducible method protocol and a model that addresses key limitations in literature (such as noise robustness and contextual understanding), this work offers a reliable foundation for future applications. These include customer service analytics, mental health monitoring tools, and advanced interactive systems that require a deep understanding of user emotion.

However, this study is not without its limitations. The model's high computational complexity may challenge deployment in real-time or resource-constrained environments. Furthermore, its performance is currently validated only on the IEMOCAP dataset, which may limit its generalizability to more spontaneous, multi-lingual, or noisy speech. For future work, we plan to: (1) explore model compression techniques (e.g., pruning, quantization) to enhance computational efficiency; (2) validate the framework on larger, more diverse, and multi-lingual datasets; and (3) extend its application to real-time streaming audio analysis for live emotion recognition systems.

REFERENCES:

[1] Xu, L. and Wang, W., 2023. Improving aspect-based sentiment analysis with contrastive learning. *Natural Language Processing Journal*, *3*, p.100009.

[2] Liang, B., Su, H., Gui, L., Cambria, E. and Xu, R., 2022. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, *235*, p.107643.

[3] Maroof, A., Wasi, S., Jami, S.I. and Siddiqui, M.S., 2024. Aspect based sentiment analysis for service industry. *IEEE Access*.

[4] Wu, H., Huang, C. and Deng, S., 2023. Improving aspect-based sentiment analysis with Knowledge-aware Dependency Graph Network. *Information Fusion*, *92*, pp.289-299.

[5] Zhao, G., Luo, Y., Chen, Q. and Qian, X., 2023. Aspect-based sentiment analysis via multitask learning for online reviews. *Knowledge-Based Systems*, *264*, p.110326.

[6] Li, P., Li, P. and Xiao, X., 2023. Aspect-pair supervised contrastive learning for aspect-based sentiment analysis. *Knowledge-Based Systems*, *274*, p.110648.

[7] Chifu, A.G. and Fournier, S., 2023. Sentiment difficulty in aspect-based sentiment analysis. *Mathematics*, *11*(22), p.4647.

[8] Murugaiyan, S. and Uyyala, S.R., 2023. Aspect-based sentiment analysis of customer speech data using deep convolutional neural network and bilstm. *Cognitive Computation*, *15*(3), pp.914-931.

[9] Zhao, H., Huang, N. and Chen, H., 2024. Knowledge enhancement for speech emotion recognition via multi-level acoustic feature. *Connection Science*, *36*(1), p.2312103.

[10] Li, W., Shao, W., Ji, S. and Cambria, E., 2022. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, *467*, pp.73-82.

[11] Qiu, K., Zhang, Y., Zhao, J., Zhang, S., Wang, Q. and Chen, F., 2024. A multimodal sentiment analysis approach based on a joint chained interactive attention mechanism. *Electronics*, *13*(10), p.1922.

[12] Faisal, S. and Tutz, G., 2022. Nearest neighbor imputation for categorical data by weighting of attributes. *Information Sciences*, *592*, pp.306-319.

[13] Yao, X., Newson, A., Gousseau, Y. and Hellier, P., 2022, October. A Style-Based GAN Encoder for high fidelity reconstruction of images and videos. In *European conference on computer vision* (pp. 581-597). Cham: Springer Nature Switzerland.

[14] Deng, S. and Dong, Q., 2021. GA-NET: Global attention network for point cloud semantic segmentation. *IEEE Signal Processing Letters*, *28*, pp.1300-1304.

[15] Sabir, Z., Kotob, I.A., Sheikh, L.A. and Saeed, T., 2025. A novel computational approach-based hyperbolic tangent sigmoid deep neural

network for the hepatitis B virus model. *International Journal of Geometric Methods in Modern Physics*, *22*(4), pp.2450315-1945.

[16] Trojovska, E., Dehghani, M. and Trojovský, P., 2022. Fennec fox optimization: a new nature-inspired optimization algorithm. IEEE Access, 10, pp.84417-84443.

[17] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4), 335–359.