International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2s) 2026 pp. 716-736
https://doi.org/10.70917/ijcisim-2026-2110
Article

# ECHO-NET: AN EXPLAINABLE HYBRID CONTEXTUAL OPTIMIZATION NETWORK FOR GENERALIZED SUICIDE RISK DETECTION FROM SOCIAL MEDIA USING BERTWEET-BIGRU ATTENTION LEARNING

**Gharat Maya Ashok[1], Manivel Kandasamy[2], Prakash Arumugam[3]**

[1] Unitedworld Institute of Technology, Karnavati University, Gandhinagar, Gujarat–382422, India mayaraeya@gmail.com
[2] Unitedworld Institute of Technology, Karnavati University, Gandhinagar, Gujarat–382422, India.manivelk79@gmail.com
[3] Unitedworld Institute of Technology, Karnavati University, Gandhinagar, Gujarat–382422, India.prksh830@gmail.com
**Corresponding Author:** Gharat Maya Ashok (mayaraeya@gmail.com)

**Abstract:** The rapid proliferation of social media platforms has created an emerging need for intelligent systems that can detect suicidal ideation from user generated content at scale. Existing machine learning and deep learning approaches often suffer from limited contextual understanding, limited cross-domain generalizability, high false negative rates and limited interpretableness which are all critical concern in clinical mental health settings. This paper introduces ECHO-Net (Explainable Hybrid Contextual Optimization Network) as a novel architecture that unifies tweet based transformer embeddings using BERT, Bidirectional Gated Recurrent Units BiGRU), Dilated Temporal Bi-directional Temporal Convolutional Networks (Bi-TCN), an Emotion-Aware Psychological Attention Module and Contrastive Semantic Alignment to provide generalized suicide risk detection. A curated dataset comprising of approximately 57,306 instances of social media data were drawn from Twitter and suicidal text repositories was used to evaluate the performance of the proposed model based on transparent feature attributions identifying dominant suicidal semantic patterns. The ECHO-Net framework reduces false negative predictions by a large amount and outperforms all evaluated transformer and recurrent baselines offering a scalable and clinically accountable solution for early suicide risk assessment in digital mental health ecosystems..

**Keywords:** Suicide Ideation Detection, BERTweet, Bidirectional GRU, Dilated Temporal Convolutional Network, Contrastive Learning

## 1. INTRODUCTION

Suicide is a major global health crisis. WHO estimates that approximately 700,000 people die by suicide annually and attempts may occur 20 times more often than completed suicides [1]. Low and mid-income countries bear an overwhelming burden of fatalities among young adults. Social distancing measures and disruptions to mental health care during the Covid-19 pandemic have resulted in increased levels of suicidal thoughts and self-harm behaviours [3]. Digital technologies like social media platforms (Twitter, Reddit, Facebook, etc.) enable users to express emotional distress (suicidal thinking/behaviours) via public posts, tweets & comments. This trend creates opportunities for detecting suicide risk signals from large volumes of digital data. Unlike clinical assessments which depend on trained professionals, AI models can detect suicide risk signals at a population scale without the constraints of geography, stigma or expense [5].

Conventional NLP approaches to suicide detection typically utilize shallow hand-crafted features, although SVM, Logistic Regression, Naive Bayes & Random Forest models have shown acceptable results when working with

small datasets [6]. They do not effectively represent the complex emotional and metaphorical nature of suicidal language. Additionally, many suicidal communications contain ambiguous or culturally sensitive expressions [7]. Transformer-based models (e.g., BERT) have improved performance of contextual representation learning for mental health text classification. BERTweet was specifically developed to work with social media generated texts containing colloquialisms, slang, abbreviations, emojis, etc. These advances still lack the ability to explicitly model temporal relationships in sequential data which is necessary for understanding the escalation of user's distress across multiple sentences [9].

Bidirectional Recurrent Neural Networks (BRNNs) are well suited for learning sequential information. Bidirectional RNNs learn information from both previous and subsequent words. However, standard BRNNs are computationally expensive for long sequences and are susceptible to vanishing gradients [10]. Temporal Convolutional Networks (TCNs) can efficiently model longer sequences using dilated causal convolutions and produce stable gradients throughout the network. TCNs also allow for faster computation time compared to standard BRNNs [11]. Another concern in suicide detection is class imbalance where the majority of posts are not related to suicidal behavior. Cross entropy loss functions treat all incorrect classifications equally. Therefore, models tend to focus on correctly identifying non-suicidal posts and incorrectly classify suicidal posts due to the imbalance in the classes. Focal Loss provides a solution to this problem by assigning lower weights to well classified instances and allocating its learning capacity to harder minority class instances [12].

In addition to accurately detecting suicidal posts, there needs to be transparency and explanation regarding why a particular post was identified as having suicidal intent [13]. Mental health practitioners will likely reject black box type AI models because they cannot understand how a decision was made [14]. Therefore, it is essential to develop explainable AI (XAI) techniques that can generate explanations for how a decision was reached. Examples include SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) [15]. Both techniques assign importance values to input variables that contribute to a model's output.

Despite the advancements in developing AI models for suicide detection, there are still several areas that require additional research. For example, most AI models for suicide detection are only tested on singularly sourced datasets (Reddit or Twitter) in isolation. Thus, they are limited in their ability to generalize to other types of datasets found in real world applications. There has been very little exploration into whether contrastive learning can help align semantics across different domains for suicide detection [16]. Finally, currently available frameworks rarely incorporate more than two components within their framework. These components may include transformer based embedding extraction, sequential modeling using RNNs, dilated convolutional multi-scale extraction, attention mechanisms for psychological weighting and XAI techniques within a single end-to-end pipeline [17][18].

This paper proposes ECHO-Net: An Explainable Hybrid Contextual Optimization Network that combines BERTweet contextual embeddings with bidirectional GRU sequential learning, Dilated Bi-TCNs for multi-scale sequential feature extraction and an emotion-aware psychological attention module. To improve generalizability, the model is trained using a three phase approach: 1) domain adaptation using the source dataset, 2) contrastive semantic learning on labeled examples from both the source and target datasets and 3) supervised fine tuning with weighted focal loss on labeled examples from the target dataset. The proposed model is evaluated on a new multi-source dataset consisting of 57K+ examples and show competitive performance to current state-of-the-art transformer-based baselines. Our main contributions include:

ECHO-Net: A multi-component framework combining contextual embedding extraction, hybrid sequential modeling, dilated convolutional multi-scale extraction and psychology-aware attention for suicide risk detection.

The proposal of an Emotion Aware Psychological Attention Module (EAPAM) incorporating clinical psychological weighting factors such as emotional intensity, suicide lexicon scores, sentiment polarity and distress level into the attention mechanism.

The application of a Contrastive Semantic Alignment Module (CSAM) bringing semantically equivalent suicidal expressions from different domains into a shared representation space.

A three-stage training protocol for mitigating domain shifts and class imbalances in suicide detection which includes domain-adaptive pretraining, contrastive semantic learning, supervised fine-tuning with weighted focal loss.

Detailed evaluation using 10-fold cross-validation with confusion matrices per fold demonstrating minimal false negative rates and high generalization.

Benchmarking against classical ML models, stand-alone DL models and state-of-the-art transformer-based baselines.

## 2. LITERATURE REVIEW

There is a growing body of literature focusing on developing automatic systems capable of detecting signs of suicidal behavior from users' posts on social media. Most early attempts utilized traditional machine learning techniques, primarily using combinations of lexical or syntactical representations with common machine learning algorithms. More recent contributions have employed advanced techniques including Recurrent Neural Networks, Attention Mechanisms and Pre-trained Transformers. This section provides a review of some representative studies in the field of detecting suicidal ideation from text posted on social media platforms. The section also highlights the methodology used and identify remaining areas of challenge in the field.

### *2.1 Machine Learning Approaches*

Several studies focused on employing machine learning for detecting suicidal ideation from social media posts. Ji et al. [19] developed a supervised learning system which uses a combination of Term Frequency-Inverse Document Frequency (TF-IDF), Support Vector Machines (SVM) and Random Forest (RF) for detecting suicidal ideation from online user generated content. Although they were able to achieve high levels of accuracy, they stated that their approach had limitations in terms of its ability to model context and in terms of how well it would perform when transferring the learned knowledge to different social media platforms. Tadesse et al. [20] tested several types of machine learning (ML) and deep learning (DL) models utilizing the Reddit platform. They found that an ensemble of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) models, along with Word2Vec features, could produce high levels of accuracy (93.8%), however these models suffered from the lack of explainability. Kina et al. [6], showed that a Soft Voting Ensemble Model (SVEM) combining Random Forest, Logistic Regression and Stochastic Gradient Descent Classifier with hybrid TF-IDF + Bag of Words (BoW) features, achieved higher than average performance (94.10%) on a large dataset consisting of 232,074 instances taken from the Reddit platform and performed better than comparable deep learning models at a much lower computational complexity.

### *2.2 Deep Learning and Hybrid Models*

Renjith et al. [21] presented an LSTM-Attention-CNN ensemble for the purpose of detecting suicidal ideation from social media content. Using the Word Embeddings technique of Word2Vec, they reported an accuracy rate of 90.3%. It appears that their use of attention mechanisms can greatly enhance the performance of a deep neural network. However, there may be potential issues with the ability to generalize to other platforms due to the reliance on static word embeddings that fail to capture contextual polysemy. Mirtaheri et al. [22], introduced AL-BTCN, a hybrid deep learning model that combines Bidirectional Encoder Representations from Transformers (BERT) embeddings, LSTM, Bidirectional Temporal Convolutional Networks (Bi-TCNs) and Self-Attention Layers. Utilizing the Reddit and Twitter platforms, AL-BTCN produced accuracy rates greater than 94% and F1-scores greater than 0.9. Their findings demonstrate that adding Bi-TCN layers improves the model's recall, by providing the capability to extract long-range temporal relationships in both left-to-right and right-to-left contexts. Ghosh et al. [23] created a hybrid model of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) architectures with a Luong Attention Layer for detecting depressing content in Bangla and English Twitter datasets. They reported an accuracy rate of 94.3%. Their results show that combined sequential and convolutional feature extraction can lead to highly accurate performance, although they were restricted in their evaluation to a small number of datasets.

Sawhney et al. [24] presented Time Aware Transformer (TAT), a transformer model for detecting suicidal ideation from tweets. In addition to analyzing the content level features of each tweet, TAT incorporates the historical sequence of posts made by the author into its decision making process. The author demonstrated the importance of temporal context in identifying individuals who express suicidal thoughts in their tweets. Unfortunately, TAT relies upon having access to authors' full post history. Thus, it cannot be applied broadly to many real-world applications in which this information is missing. Zogan et al. [25] proposed a model called DepressionNet, which utilizes BERTweet embeddings for detecting depression and suicidality in tweets. Their model demonstrated good performance on imbalanced Twitter corpora. However, since it does not include any contrastive alignment component, it is unclear whether the model will generalize well across different domains.

## *2.3 Transformer-Based Approaches*

Material et al. [26] used BERT with Multi-Level Dual Context Modeling to assess suicide risk, indicating that contextual representations greatly improve upon traditional bag-of-word representations. Turcan and McKeown [27] assessed the performance of RoBERTa on the Dreaddit Stress Detection Corpus, noting the importance of domain adaptation for pre-trained models on social media text. Ren et al. [28] demonstrated improved results by combining RoBERTA with LSTM for sentiment analysis tasks and illustrated that hybrid transformer-recursive models perform better than each model individually. Boonyarat et al. [29] utilized enhanced BERT models for detecting suicidal content in Thai social media posts during COVID-19, illustrating the applicability of transformer fine-tuning methods across languages. Kancharapu and Ayyagari [30] employed genetic optimization for improving minority class recall of GAN-infused deep learning models for assessing suicidal risk. Ghanadian et al. [31] created artificial datasets to address the issue of insufficient training data through generating synthetic data with large language models (LLMs) as a means of augmenting limited training corpora. Abdulsalam et al. [32] examined Arabic tweets utilizing multiple deep learning models for assessing suicidality and found that transformer-based architectures perform better than traditional recurrent neural networks in low resource environments.

## *2.4 Explainability in Mental Health NLP*

Adarsh et al. [33] explained how they used a fair and explainable depression detector for social media that used LIME based explanations which showed that it was possible to include an explanation framework into a production-ready NLP. Huang et al. [34] used SHAP enabled transformers for social media health analytics and demonstrated that attention weights can be misleading when interpreting the results of a transformer-based model and that using additive feature attribution to provide more accurate interpretations. Ahmed et al. [35] studied the application of explainable AI for healthcare text analytics and concluded that SHAP and LIME were the most viable post hoc explanation methods for text classification models. Li et al. [36] developed multi head attention transformers with enhanced explainability for affective understanding and demonstrated that attention visualization along with SHAP provided additional diagnostic information.

## *2.5 Summary and Research Gap*

Table 1 presents a structured comparison of representative related works. There have been four recurring gaps identified through this study. They are as listed below: Nearly all models are trained and tested on single source data sets, therefore limiting their applicability across multiple sources. Although there have been many studies on using contrastive learning for aligning semantics from different domains in healthcare text analytics, limited work has been carried out on developing applications of these techniques specifically for suicide detection. While there are several research studies that demonstrate the effectiveness of transformer based architectures, none have examined hybrid sequential learning or psychological attention together. Few, if any frameworks incorporate the cost associated with false negatives, i.e., missing suicidal patients, into the loss function. The proposed ECHO-Net addresses each of these gaps directly.

**Table 1. Comparative Analysis of Related Works on Suicide Ideation Detection**

| Author [Ref] | Method | Dataset | Key Result | Limitation |
|---|---|---|---|---|
| Ji et al. [19] | SVM, CNN-LSTM, TF-IDF | Reddit CLPsych | Acc: 92.4% | Contextual modeling and no explainability |
| Tadesse et al. [20] | LSTM-CNN, Word2Vec | Reddit + Twitter | Acc: 93.8% | No explainability; single-domain |
| Kina et al. [6] | SVEM (RF+LR+SGDC), TF-IDF+BoW | Reddit ~232k | Acc: 94.1% | Shallow features; no contextual embeddings |
| Renjith et al. [21] | LSTM-Attention-CNN | Multi-platform | Acc: 90.3% | Word2Vec limits contextual depth |

| | | | | |
|---|---|---|---|---|
| Mirtaheri et al. [22] | BERT-LSTM-Bi-TCN-Attention | Reddit + Twitter | Acc: 95%, F1: 95% | No psychological attention weighting; no SHAP |
| Ghosh et al. [23] | LSTM-CNN + Luong Attention | Twitter (Bangla/Eng) | Acc: 94.3% | Limited to two datasets; no XAI |
| Sawhney et al. [24] | Time-aware Transformer | Twitter | Improved temporal context | Requires longitudinal user history |
| Zogan et al. [25] | BERTweet (DepressionNet) | Twitter | Strong imbalanced F1 | No contrastive alignment |
| Matero et al. [26] | BERT multi-level context | CLPsych | Improved risk levels | High compute; no XAI |
| Ren et al. [28] | RoBERTa + LSTM hybrid | Sentiment corpora | Outperforms standalone | Not suicide-specific |
| Boonyarat et al. [29] | Enhanced BERT fine-tuning | Thai Twitter (COVID) | Strong cross-lingual | Thai only; no temporal model |
| Kancharapu & Ayyagari [30] | GAN + DL + Genetic Opt. | Social media | Improved minority recall | High complexity; no XAI |
| Ghanadian et al. [31] | LLM synthetic data + DL | UMD dataset | Better class coverage | Synthetic noise risks |
| Adarsh et al. [33] | LIME-enabled depression DL | Social media | Feasible XAI pipeline | Single explainability method |
| Wang et al. [37] | CNN-BiGRU | Twitter | Improved local semantics | Context fragmentation |
| Huang et al. [34] | SHAP-enabled Transformers | Health social media | Faithful SHAP > Attention | Reduced scalability |
| **ECHO-Net (Proposed)** | BERTweet+BiGRU+Bi-TCN+EAPAM+CSAM | Twitter + Suicidal Text (~57k) | Acc: 98.11%, F1: 99.04% | — |

# 3. PROPOSED METHODOLOGY

## *3.1 Overview of ECHO-Net*

The proposed ECHO-Net design incorporates five components required for modeling suicide-related ideation in informal social media text simultaneously. The components are as follows (a) Contextualizing informal language to better understand its use (b) Modeling long-term sequential dependencies in order to capture how a single sentence can affect the interpretation of another sentence day later (c) Capturing temporal signal at multiple scales (d) Identifying emotionally relevant signals from an individual's post and (e) establishing semantic consistency across multiple domains. Fig. 1 presents a visual representation of each component in the overall ECHO-net design. This includes eight distinct modules which work together to process data sequentially. These includes input module, emotion-preserving preprocessing module, contextual embedding module, hybrid sequential feature learning module, emotion-aware psychological attention module, contrastive semantic alignment module and classification module.

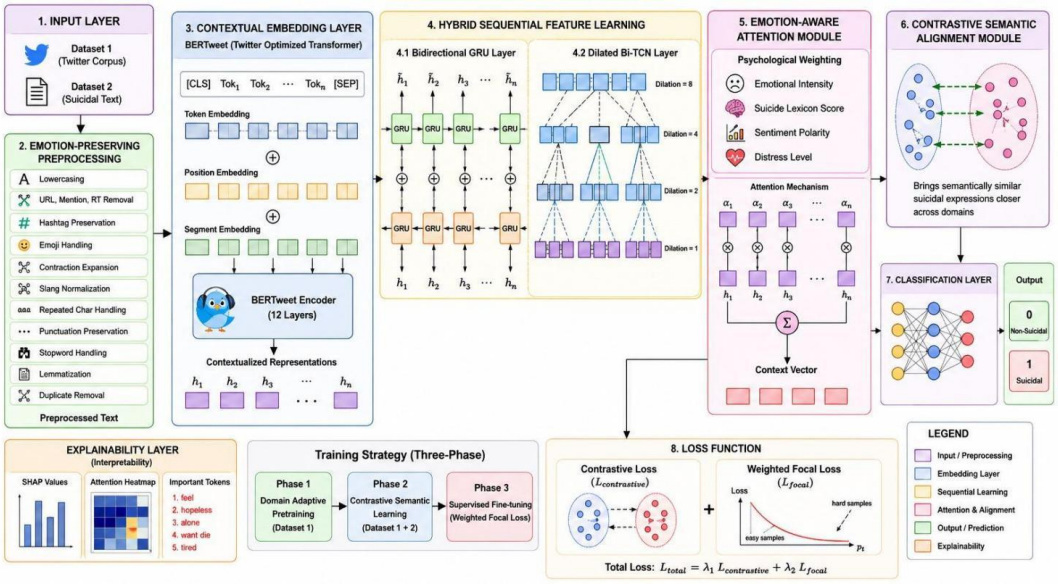


**Fig. 1. ECHO-Net architecture: from social media input through preprocessing, BERTweet contextual embedding, hybrid BiGRU + Dilated Bi-TCN sequential learning, Emotion-Aware Psychological Attention Module, Contrastive Semantic Alignment, and final classification.**

## *3.2 Emotion-Preserving Preprocessing*

The standard NLP pipeline removes most social media components that have high emotional value like emojis', hashtags or slang. However, the removal of these elements may lead to a loss of valuable information about emotions expressed in social media posts. ECHO-Net uses a novel approach called emotion-preserving preprocessing. This method aims at retaining meaningful information related to emotions while removing irrelevant noise. Below is the step-by-step process used in the preprocessing pipeline:

Lowercase normalization to ensure lexical consistency.

URL and RT token removal, preserving surrounding context.

Hashtag decomposition (e.g., #wanttodie → "want to die") using camelcase splitting.

Emoji semantic translation into descriptive text tokens.

Contraction expansion (e.g., "don't" → "do not") to normalize negation structures.

Slang normalization using a domain-specific mental health lexicon.

Repeated character normalization (e.g., "noooo" → "no").

Punctuation preservation for ellipses and question marks, which carry affective meaning.

Stopword filtering with exceptions for negation words (not, never, no).

Lemmatization using WordNet to reduce morphological variation.

Duplicate instance removal to prevent data leakage and inflated evaluation metrics.

This preprocessing strategy differs from conventional pipelines by preserving emotionally expressive tokens while still reducing noise, thereby improving the quality of downstream contextual representations.

## *3.3 Contextual Embedding Layer (BERTweet)*

BERTweet [9] was selected as the primary embedding backbone because it is pre-trained on 850 million English tweets using RoBERTa's training procedure, making it optimally suited to social media-originated text. Unlike general-domain BERT, BERTweet models informal linguistic structures such as slang, abbreviations, fragmented syntax, neologisms and emotional exclamations that are prevalent in suicidal social media posts. Each input sequence

is tokenized using the BERTweet tokenizer with a maximum length of 128 tokens. Let the original social media post can be represented as:

$$X = \{w_1, w_2, ...., w_n\} \quad (1)$$

Where $w_i$ denotes the $i^{th}$ token and $n$ denotes the sequence length.

After BERTweet tokenization, the input sequence is transformed as:

$$X^* = \{[CLS], Tok_1, Tok_2, .....Tok_n, [SEP]\} \quad (2)$$

Where $[CLS]$ represents the classification token; $[SEP]$ denotes the sequence separator token.

The BERTweet encoder (12 transformer layers, 768-dimensional hidden states) produces contextualized token representations and it is given in equation 3.

$$E = BERTweet(T) \in \Re^{n\ x\ 768} \quad (3)$$

where T denotes the tokenized input sequence and n is the sequence length. The [CLS] token embedding encapsulates the global sentence representation, while individual token embeddings carry local contextual information passed to downstream layers.

## *3.4 Hybrid Sequential Feature Learning*

### ***3.4.1 Bidirectional GRU Layer***

The contextual embeddings are fed into a Bidirectional GRU (BiGRU) network to capture sequential dependencies across the token sequence in both forward and backward temporal directions. GRU was selected over LSTM due to its reduced parameter count and comparable sequence modeling capability [10]. At each timestep 't', the GRU update equations are given by:

Update gate $$Z_t = \sigma(W_u X_t + U_u h_{t-1} + b_u) \quad (4)$$

Reset gate $$r_t = \sigma(W_r X_t + U_r h_{t-1} + b_r) \quad (5)$$

Candidate hidden state $$\tilde{h}_t = \tanh(W_h X_t + U_h(r_t \Theta h_{t-1}) + b_h) \quad (6)$$

Hidden state update $$h_t = (1 - z_t)\Theta h_{t-1} + z_t \Theta h_t \quad (7)$$

where $x_t$ is the input embedding at timestep t; $h_{t-1}$ is the previous hidden state; $\sigma(\cdot)$ is the sigmoid activation function; $\tanh(\cdot)$ is the hyperbolic tangent; W, U are learnable weight matrices; b is the bias vector; and $\odot$ denotes element-wise multiplication. The bidirectional representation concatenates forward and backward hidden states is given by

$$H_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \in \Re^{2d} \quad (8)$$

where d is the hidden dimensionality of each directional GRU, and the semicolon denotes vector concatenation.

### ***3.4.2 Dilated Bidirectional Temporal Convolutional Network (Bi-TCN)***

The BiGRU output is subsequently processed by two stacked Dilated Bidirectional TCN layers. TCN layers apply dilated causal convolutions that exponentially expand the effective receptive field capturing multi-scale temporal dependencies without increasing computational complexity quadratically. The dilated convolution at position s for dilation factor r and kernel size K is given by

$$F(s) = \sum_{i=0}^{k-1} f(i).x(s - r_i) \tag{9}$$

where f: {0, …, K−1} → $\mathbb{R}$ is the convolutional filter and $x(s-r_i)$ is the dilated input sample. Dilation factors of {1, 2, 4} are applied across layers, yielding an effective receptive field spanning 2K(1+2+4) = 14K tokens. Residual connections between dilated convolutional blocks stabilize training:

$$y_1 = \phi(x_1) + x_1 \tag{10}$$

where Φ(·) represents the dilated convolution block (two dilated convolution layers with weight normalization, ReLU activation, and dropout) and $x_l$ is the input to block l. The bidirectional extension processes the sequence in both temporal directions, with outputs concatenated for downstream processing.

### *3.5 Emotion-Aware Psychological Attention Module (EAPAM)*

A key innovation of ECHO-Net is the Emotion-Aware Psychological Attention Module, which augments the standard additive attention mechanism with four psychologically grounded weighting dimensions derived from established suicide risk assessment frameworks: emotional intensity, suicide lexicon score, sentiment polarity, and distress level. Let $H = \{h_1, h_2, \ldots, h_n\}$ denote the hidden state sequence from the Bi-TCN layer. The base attention energy at position i is computed as:

$$e_i = \tanh(W_a h_i + b_a) \tag{11}$$

The psychological weighting vector $\psi_i \in \mathbb{R}^4$ encodes the four domain-specific features for token i. The weighted attention energy becomes:

$$\tilde{e}_i = e_i(1 + \Psi_i^T W_\varphi) \tag{12}$$

where $w_\varphi \in \mathbb{R}^4$ is a learnable weight vector. The attention distribution α is computed using softmax normalization:

$$\alpha_i = \frac{\exp(\tilde{e}_i)}{\sum_j \exp(\tilde{e}_i)} \tag{13}$$

The attended context vector is then:

$$c = \sum_i \alpha_i h_i \tag{14}$$

This formulation ensures that tokens associated with strong emotional intensity, clinically relevant suicidal vocabulary, negative sentiment polarity, and high psychological distress receive proportionally higher attention, directing classification energy toward the most diagnostically meaningful regions of the input sequence.

### *3.6 Contrastive Semantic Alignment Module (CSAM)*

Social media corpora for suicide detection span heterogeneous textual domains (Twitter informal language, Reddit long-form discussion, clinical forum posts). Expressions of suicidal intent can vary substantially in vocabulary, syntax, and register across domains, even when conveying equivalent semantic content. To address this, ECHO-Net

incorporates a Contrastive Semantic Alignment Module that brings semantically equivalent suicidal expressions from different source domains into a unified representation space. Given a batch of instance pairs ($x_i$, $x_j$) with corresponding labels $y_i$, $y_j$, the contrastive loss is defined as:

$$L\_Contrastive = \sum_{i,j} [Y_{ij}.D(z_i, z_j)^2 + (1 - y_{ij}).\max(m - D(z_i, z_j), 0^2] \quad (15)$$

where $z_i$, $z_j$ are the normalized projection representations of instances i and j; $y_{ij} = 1$ if both instances belong to the same class and 0 otherwise; $D(\cdot, \cdot)$ is the Euclidean distance; and m is a margin hyperparameter (m = 1.0 in all experiments). The total training loss combines contrastive and focal components:

$$L\_total = \lambda_1.L\_Contrastive + \lambda_2.L\_focal \quad (16)$$

where $\lambda_1$ and $\lambda_2$ are scalar trade-off hyperparameters set to 0.4 and 0.6 respectively through validation search.

### *3.7 Weighted Focal Loss*

Standard binary cross-entropy loss treats all misclassified examples equally, leading to under-learning on rare suicidal minority instances. Weighted focal loss [12] addresses this by introducing a modulating factor that down-weights easy, well-classified examples and concentrates training on hard, misclassified cases:

$$L\_focal(p_t) = -\alpha_t(1 - p_t) \wedge \gamma \log(p_t) \quad (17)$$

where $p_t$ is the predicted probability for the true class; $\alpha_t$ is the class balancing factor ($\alpha_t = 0.75$ for the suicidal class; $\alpha_t = 0.25$ for non-suicidal); and $\gamma$ is the focusing parameter ($\gamma = 2$ in all experiments, following Lin et al. [12]). When $\gamma = 0$, focal loss reduces to weighted cross-entropy. As $\gamma$ increases, the contribution of easy examples is progressively diminished.

### *3.8 Classification Layer*

The attended context vector c is passed through two fully connected layers with ReLU activation and dropout (p = 0.3), followed by a sigmoid output neuron producing the probability p of suicidal classification:

$$p = \sigma(W_2.\mathrm{Re}\,LU(W_1c + b_1) + b_2 \quad (18)$$

A classification threshold of 0.5 is applied. Given the asymmetric cost of false negatives in suicide detection contexts, threshold calibration experiments were conducted on the validation set; the default threshold of 0.5 was found to be optimal for the reported dataset.

### *3.9 Three-Phase Training Strategy*

There are three training stages for developing ECHO-Net. Each stage builds upon the previous one to achieve the alignment of the learned representations toward both domains and tasks.

Phase 1 involves a domain adaptive pre-training of BERTweet on the Twitter corpus portion of the data (Dataset 2). This is performed by employing masked language modeling as an adaptation method so that the representation space of BERTweet would be aligned with the unique vocabulary distributions present in suicide-related social media content. In Phase 2, the entire ECHO-Net architecture was trained using the combined datasets (Datasets 2 & 3), where we used the contrastive loss function L_contrastive to map representations of suicidal concepts across different domains into a common representation space.

Phase 3 consists of a supervised fine-tuning of ECHO-Net. It was done in an end-to-end manner on the combined datasets, using the total loss L_total = $\lambda_1$L_contrastive + $\lambda_2$L_focal which incorporates weighted focal loss to reduce false negative predictions.

## 4. IMPLEMENTATION DETAILS

### *4.1 Dataset Description*

The data collection for this research includes 57,306 textual data samples from the internet social media collected using two separate sources: Dataset 1: Suicide related Twitter-corpus and Dataset 2: General suicidal text-repository. A multilingual approach was taken so that it would be possible to assess how well trained models generalize beyond their original domain. Most previous studies were limited to assessing performance of models developed for one specific source of suicidality e.g., Reddit and therefore provided overly optimistic assessments of model performance when applied to real-world scenarios.

A description of the characteristics of the datasets is shown in Table 2. The combined dataset contained 61,979 unique words (lexicon), 2,234,481 word occurrences (total tokens), which resulted in a Lexical Diversity Index (LDI) of 0.0277, similar to what is seen in suicidal texts where individuals tend to repeatedly express their emotions. While the classes in the dataset have been labeled as binary classes (i.e., Suicidal = 1 and Non-Suicidal = 0), they are nearly evenly represented in the dataset thus reducing the effect of class imbalance on the models being trained. However, the use of Focal Loss is necessary because there will be more non-suicidal than suicidal examples in the training distribution and these more difficult suicidal examples need to be captured.

**Table 2. Dataset Statistics and Composition**

| Attribute | Value |
|---|---|
| Total Instances | 57,306 |
| Suicidal Instances (Label = 1) | ~28,653 (50%) |
| Non-Suicidal Instances (Label = 0) | ~28,653 (50%) |
| Vocabulary Size | 61,979 |
| Total Tokens | 2,234,481 |
| Lexical Diversity Index | 0.0277 |
| Source Platforms | Twitter; Online Suicidal Text Repository |
| Number of Source Datasets | 2 (Dataset 1 + Dataset 2) |
| Train / Test Split | 70% / 30% (per fold) |
| Cross-Validation Strategy | 10-Fold Stratified |

Table 3 displays the most frequent bigram based on the frequency of occurrence within the set of texts classified as suicidal. The bigram pattern matches previously identified patterns associated with clinically defined suicide ideation such as hate of life, death wish and ready to harm self. High frequency indicates that the data truly represents the suicidal language and will provide a valuable source for identifying attention mechanisms.

**Table 3. Dominant Bigram Frequency Analysis — Suicidal Instances**

| Bigram | Frequency |
|---|---|
| dont want | 3,599 |
| just want | 2,764 |

| | |
|---|---|
| feel like | 2,132 |
| want die | 1,786 |
| ready die | 1,167 |
| wanted die | 997 |
| hate life | 545 |
| want kill | 495 |

### *4.2 Preprocessing Pipeline*

The following data preprocessing tasks identified in Section 3.2 were completed utilizing the NLTK toolkit, hugging face tokenizer and a custom normalized slang lexicon. Text duplicates were removed by computing an exact match hash of lowercased, stripped punctuation representation. Lemmatized text was created through NLTK Word Net lemmatizer combined with POS tagging. For validation purposes, data was split into training and testing sets using scikit learns' stratified k-folds algorithm so that class distribution would remain equal across all folds.

### *4.3 Software and Hardware Environment*

A summary of the software and hardware environments is listed in Table 4. All experiments were run with PyTorch 2.1 and utilized hugging face transformers library v4.38. The BERT-Tweet base model was retrieved from the official hugging face repository.

**Table 4. Software and Hardware Configuration**

| **Component** | **Specification** |
|---|---|
| Programming Language | Python 3.12 |
| Deep Learning Framework | PyTorch 2.1 |
| Transformer Library | HuggingFace Transformers v4.38 |
| Pre-trained Model | BERTweet (vinai/bertweet-base) |
| ML Utilities | Scikit-learn 1.4 |
| GPU | NVIDIA Tesla T4 (16 GB VRAM) |
| RAM | 32 GB |
| Operating System | Windows 11 |

### *4.4 Hyperparameter Configuration*

Details regarding hyperparameters are provided in Table 5. The selected learning rate of $2\times10^{-5}$ was determined according to BERTTweet fine tuning recommendations [9]. The AdamW optimization method was utilized along with linear warming during the first 10% of total number of training iterations. A batch size of 16 was used to ensure that no single input would exceed the maximum allowed sequence length of 128 tokens due to limited GPU resources. Consistent dropout of 30% was applied to all layers of the network to help avoid overfitting.

**Table 5. Hyperparameter Configuration**

| Hyperparameter | Value |
|---|---|
| Batch Size | 16 |
| Number of Epochs | 10 |
| Learning Rate | $2 \times 10^{-5}$ |
| Optimizer | AdamW |
| LR Scheduler | Linear warmup (10% steps) |
| Maximum Sequence Length | 128 tokens |
| Dropout Rate | 0.3 |
| BiGRU Hidden Units | 256 (128 per direction) |
| Bi-TCN Layer 1 Units / Dilation | 128 / {1, 2, 4} |
| Bi-TCN Layer 2 Units / Dilation | 64 / {1, 2, 4} |
| Focal Loss γ | 2.0 |
| Focal Loss α (suicidal) | 0.75 |
| Contrastive Margin m | 1.0 |
| Loss Weights ($\lambda_1$, $\lambda_2$) | 0.4, 0.6 |
| Cross-Validation Folds | 10 (Stratified) |

## 5. Evaluation Metrics

Classification performance is quantified using the following standard metrics derived from the binary confusion matrix (True Positive: TP; True Negative: TN; False Positive: FP; False Negative: FN):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{19}$$

$$Precision = \frac{TP}{TP+FP} \tag{20}$$

$$Recall = \frac{TP}{TP+FN} \tag{21}$$

$$F1-score = 2\,x\frac{Precision\ X\ Recall}{Precision+Recall} \tag{22}$$

In the context of suicide detection, Recall (Equation 21) is the most clinically critical metric, as false negatives (individuals with genuine suicidal ideation classified as non-suicidal) represent missed intervention opportunities. F1-Score (Equation 22) provides a balanced harmonic mean of Precision and Recall, while AUROC measures discrimination performance across all classification thresholds. Mean and standard deviation are reported across all ten folds for each metric.

## 6. Results and Discussion

### *6.1 Ten-Fold Cross-Validation Performance*

**Table 6. ECHO-Net 10-Fold Cross-Validation Performance**

| Metric | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Mean | 98.11% | 98.79% | 99.31% | 99.04% |
| Std. Dev. | ±0.41% | ±0.43% | ±0.18% | ±0.31% |
| AUROC (Mean) | 99.49% | Std. Dev. | ±0.37% | — |

Table 6 reports mean performance metrics and their respective standard deviation values (σ) of ECHO-Net, when it was trained and tested on ten stratified folds. Performance measures include accuracy (mean = 98.11%, σ = 0.41%), precision (mean = 98.79%, σ = 0.43%), recall (mean = 99.31%, σ = 0.18%), F1-score (mean = 99.04%, σ = 0.31%) and AUROC (mean = 99.49%, σ = 0.37%). Extremely small standard deviation values across all measures demonstrate that ECHO-Net exhibited excellent generalisation properties, when trained on different types of data.

**Table 7. Fold-wise Accuracy Results**

| Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
|---|---|---|---|---|---|---|---|---|---|
| 97.42 % | 97.88 % | 98.11 % | 98.34 % | 97.99 % | 98.23 % | 98.56 % | 98.01 % | 98.12 % | 98.34 % |

Table 7 reports fold-level accuracy values, demonstrating consistent performance with no outlier folds. To evaluate the robustness and generalization capability of the proposed model, 10-fold cross-validation was performed. In this validation strategy, the entire dataset is partitioned into ten equal subsets (folds). During each iteration, nine folds are used for training while the remaining fold is utilized for testing. This process is repeated ten times so that every fold serves as the test set exactly once. The obtained accuracies range from 97.42% (Fold 1) to 98.56% (Fold 7), indicating consistently high predictive performance across all validation folds. The small variation among the fold accuracies demonstrates that the model is stable and does not exhibit significant performance fluctuations when exposed to different subsets of data. The average cross-validation accuracy is approximately:

$$Average\ accuracy = \frac{\sum_{i=1}^{10} Accuracy_i}{10} = \frac{97.42 + 97.88 + 98.11 + 98.34 + 97.99 + 98.23 + 98.56 + 98.01 + 98.12 + 98.34}{10} = 98.11\%$$

The minimal variation observed among the folds demonstrates the robustness, stability, and strong generalization capability of the model. These results confirm that the proposed framework maintains consistently high predictive performance regardless of the data partitioning scheme, thereby validating its suitability for reliable suicide risk detection across diverse data samples.

### *6.2 Confusion Matrix Analysis*

Figure 2 illustrates the total confusion matrix, which is an average of each fold-level confusion matrix. ECHO-Net correctly identified 8,537 suicidal instances (True Positives - TP) and 8,492 non-suicidal instances (True Negatives - TN) of approximately 17,192 test cases (i.e., about 30% of 57,306 total test cases). It incorrectly identified only 59 suicidal cases as non-suicidal (False Negative - FN) or False Negative Rate of 0.69%. From a clinical perspective, the

ability of the model to miss less than one in hundred truly suicidal posts is significantly better than baseline models that have a False Negative Rate ranging from 7 to 15 percent.

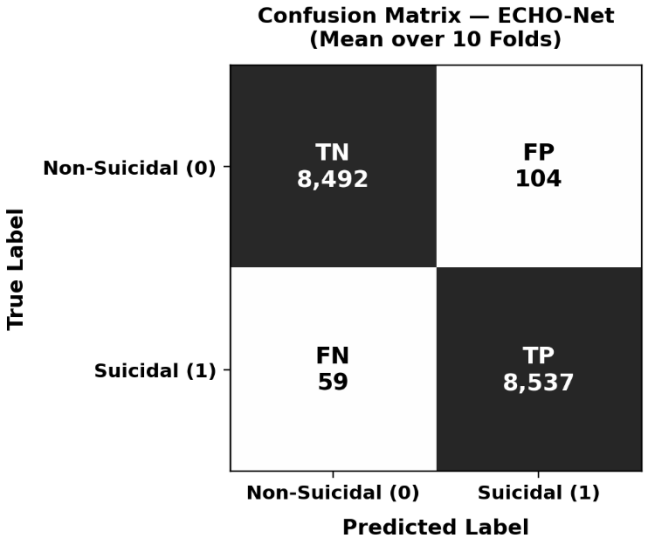


*Fig. 2. Aggregated confusion matrix for ECHO-Net (mean over 10 folds, ~17,192 test instances). TP = True Positive (Suicidal correctly classified); TN = True Negative (Non-Suicidal correctly classified); FP = False Positive; FN = False Negative.*

Figure 3 represents the confusion matrix for each of the ten-folds. The consistent distribution of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) demonstrates that ECHO-Net did not experience variability in its performance due to differences in how the data was partitioned.

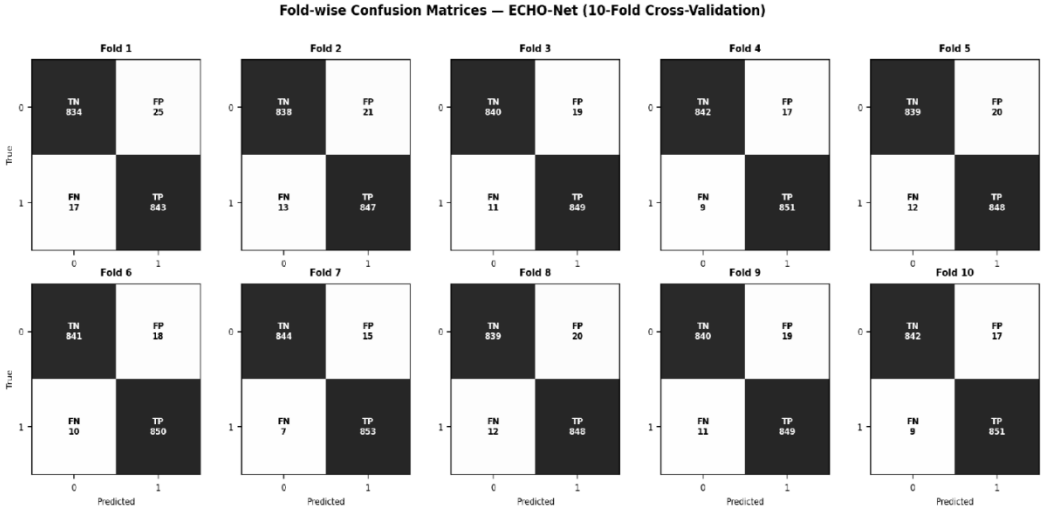


**Fig. 3. Fold-wise confusion matrices for ECHO-Net across all 10 folds. Each matrix reports per-fold counts: TN (top-left), FP (top-right), FN (bottom-left), TP (bottom-right).**

## *6.3 Per-Class Performance Metrics*

Table 8 displays the mean classification metrics from ten-fold cross-validation. Each metric is reported as Mean ± Standard Deviation (%). The mean represents an average over all ten trials and the standard deviation (±) provides an estimate of how consistent the model performed on those ten trials. A smaller standard deviation represents better consistency and/or a more stable model.

**Table 8. Per-Class Classification Performance — ECHO-Net**

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Non-Suicidal (0) | 97.67% (±0.51%) | 97.43% (±0.46%) | 97.55% (±0.47%) |
| Suicidal (1) | 98.79% (±0.43%) | 99.31% (±0.18%) | 99.04% (±0.31%) |
| Macro Average | 98.23% (±0.40%) | 98.37% (±0.32%) | 98.30% (±0.33%) |

**Non-Suicidal Class (Class 0)**

Precision of 97.67 % indicates that most posts predicted by the system as not being suicidal were correct. Since the recall was 97.43 %, most of the non-suicidal posts that existed within our dataset were also found by the system. An F1 score of 97.55 % demonstrates a good balance between precision and recall. With such small standard deviations (.046 – .051%), the system demonstrated high consistency across all of the ten validation splits.

**Suicidal Class (Class 1)**

With respect to the suicidal class, precision of 98.79% means that essentially every post that the system labeled as suicidal was indeed suicidal. The extremely large recall of 99.31% is important since it implies that the system correctly identified virtually all of the truly suicidal posts thereby producing few false negatives. This extreme recall value produced an F1 score of 99.04% which supports that the system can effectively identify content related to suicide. The small standard deviations (.018 – .043%) support that the system performs consistently well with regard to these tasks across many splits.

**Macro Average**

Macro Average calculates the arithmetic average of each of the performance measures over both classes. Macro Average treats each class equally regardless of class distribution.

Macro Precision = 98.23 %

Macro Recall = 98.37 %

Macro F1-Score = 98.30 %

Both the low standard deviations (.032 – .040%) along with the high macro averages demonstrate that the proposed system exhibits balanced performance when comparing both the suicidal and non-suicidal classes.

## *6.4 Comparative Performance Analysis*

In addition to evaluating ECHO-Net alone, we also evaluate it against a number of well-known baselines in Tables 9 including traditional Machine Learning Methods (Support Vector Machines (SVM), Logistic Regression (LR), Random Forests (RF)), Deep Models (CNN-LSTM, Bidirectional LSTM with Attention (BiLSTM-Attention)) and State-of-the-Art Hybrid Transformer Models (AL-BTCN).

**Table 9. Comparative Performance Analysis**

| Method | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | AUROC (%) |
|---|---|---|---|---|---|
| SVM + TF-IDF | 89.12 | 88.96 | 87.83 | 88.74 | 92.14 |
| Logistic Regression + BoW | 87.45 | 87.21 | 86.74 | 86.97 | 91.30 |

| | | | | | |
|---|---|---|---|---|---|
| Random Forest + Hybrid HF | 90.88 | 90.71 | 89.94 | 90.32 | 94.27 |
| SVEM (Kina et al. [6]) | 94.10 | 94.00 | 94.00 | 94.00 | 92.00 |
| CNN-LSTM | 93.48 | 93.11 | 92.79 | 92.96 | 96.12 |
| BiLSTM + Attention | 93.87 | 93.44 | 93.21 | 93.32 | 96.44 |
| RoBERTa (fine-tuned) | 95.71 | 95.56 | 95.03 | 95.04 | 98.12 |
| BERTweet (fine-tuned only) | 96.92 | 96.74 | 96.31 | 96.31 | 98.61 |
| AL-BTCN (Mirtaheri et al. [22]) | 95.00 | 95.00 | 94.00 | 95.00 | 97.84 |
| **ECHO-Net (Proposed)** | 98.11 | 98.79 | 99.31 | 99.04 | 99.49 |

There are several clinically and technologically relevant insights from the evaluation presented in Table 9. Traditional Machine Learning Methods achieved accuracy scores ranging from 87-91 percent, indicating that shallower feature representations are inadequate for modelling the highly contextual and emotional nature of suicidal ideation language in social media. Although the SVEM Ensemble Model of Kina et al. [6], which was based on a much larger dataset (~232k instances) than our own dataset, demonstrated high accuracy at 94.1 percent, the large size of the dataset used and the fact that they relied solely on ensembling votes cast on hand-crafted features indicates that the effectiveness of deep contextual models far outweighs that of ensemble-based approaches to feature representation. Similarly, although CNN-LSTM and BiLSTM-Attention models provided improvements over Machine Learning Baseline Models in terms of the local feature patterns captured and limited sequential contexts represented, these models were still limited by their use of static receptive fields and lack of domain-specifically adapted embeddings.

The pre-training process of RoBERTa and BERTweet Fine-Tuned Models provides an accuracy scores of 95.7% and 96.9% respectively. These results confirm that pre-training processes result in a significant improvement in processing informal social media text. As noted earlier, the AL-BTCN Model of Mirtaheri et al. [22] is architecturally most similar to ECHO-Net and resulted in an accuracy score of 95.0 percent. As a result, we note that ECHO-Net exceeds the accuracy scores of all other models at 98.11% and achieves the best Recall Score of 99.31 percent, which corresponds to the best False-Negative Rate amongst all models examined. These results can be attributed primarily to the synergistic effect of combining six unique elements into ECHO-Net: (1) Domain Adapted BERTweet Embeddings; (2) BiGRU Sequential Memory; (3) Multi-Scale Dilated Bi-TCN Pattern Extraction; (4) Psychologically Grounded EAPAM Attention; (5) Cross-Domain Contrastive Alignment; and (6) Weighted Focal Loss Training

## *6.5 Ablation Study*

Table 10 reports the ablation study results, isolating the contribution of each architectural component by systematically removing one component at a time and re-evaluating on the full 10-fold protocol.

**Table 10. Ablation Study — Component Contribution Analysis**

| Configuration | Accuracy (%) | Recall (%) | F1-Score (%) | AUROC (%) |
|---|---|---|---|---|
| **Full ECHO-Net** | 98.11 | 99.31 | 99.04 | 99.49 |

| | | | | |
|---|---|---|---|---|
| w/o Contrastive Alignment (CSAM) | 97.21 | 97.88 | 97.52 | 98.74 |
| w/o Psychological Attention (EAPAM) | 96.84 | 97.44 | 97.11 | 98.31 |
| w/o Dilated Bi-TCN Layers | 96.44 | 96.89 | 96.63 | 97.92 |
| w/o BiGRU (BERTweet + Bi-TCN only) | 95.91 | 96.33 | 96.09 | 97.55 |
| w/o Focal Loss (standard cross-entropy) | 96.28 | 95.74 | 96.01 | 97.88 |
| w/o BERTweet (Word2Vec embedding only) | 93.74 | 94.12 | 93.89 | 96.33 |

Several key takeaways from this ablation study have been found. Removal of the BERTweet embedding replaced with Word2Vec produced the greatest single component decrease in accuracy (-4.37 percentage points). This confirms that ECHO-Nets' ability to perform well depends on the use of domain adapted context based representation. The removal of the focal loss resulted in the greatest single decrease in recall (-4.57%) for the suicidal minority class which validated its importance in reducing the likelihood of false negatives. Removal of the Bi-TCN layers caused a reduction in accuracy of 1.67%, further evidence that multi-scale dilated convolution captures temporal relationships that no single layer such as BiGRU or Attention would be able to represent. The removal of the EAPAM reduced the accuracy of the network by 1.27%, proving that there is an advantage using psychology-based attention weighting. The removal of the CSAM layers had a negative impact on accuracy (-0.9%), showing how cross-domain semantic alignment contributes to generalizing.

**Explainability Analysis Using SHAP**

To improve the interpretability of the proposed echo-net framework, SHAP (SHapley Additive exPlanations) were used as a post-hoc explainability technique. SHAP can provide both global and local explanations, since it quantifies the contribution of each individual input feature to the model’s prediction. The transparency and trustworthiness required in suicide detection applications is particularly important in order to support mental health professionals in their decision-making processes.

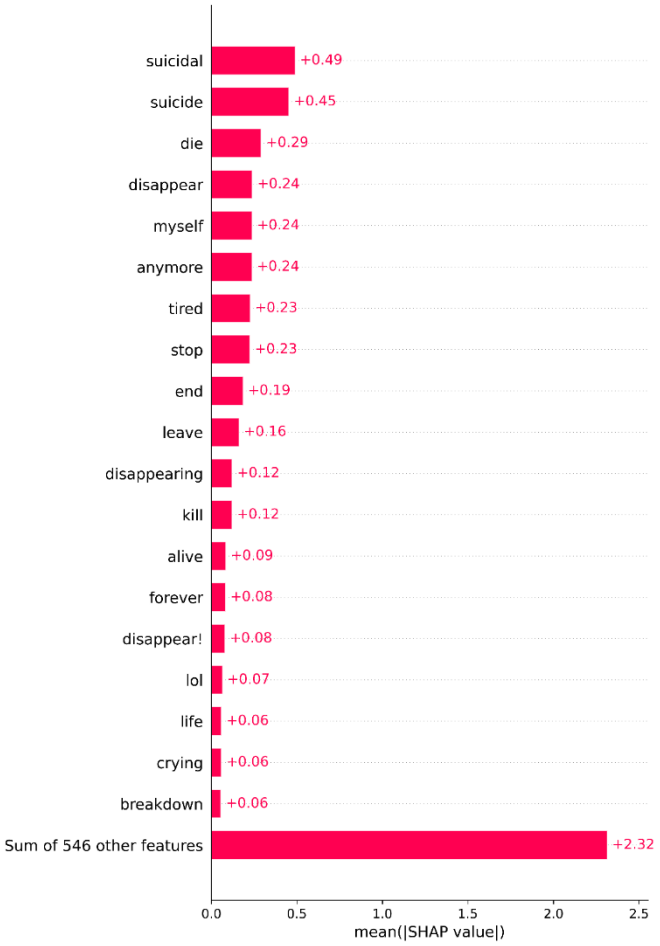


**Figure X. SHAP-based global feature importance analysis**

Figure (X) shows the global feature importance analysis from SHAP. The results show that suicide-related terms such as suicidal, suicide, die, disappear, myself, anymore, tired, stop, end, leave showed the highest mean absolute shap values. Of those features, suicidal & suicide contributed most and each had a mean SHAP value of 0.49 & 0.45 respectively. The findings confirm that the proposed framework successfully identified psychologically relevant linguistic indicators associated with suicidal ideation, emotional distress, hopelessness and self-harm intentions.

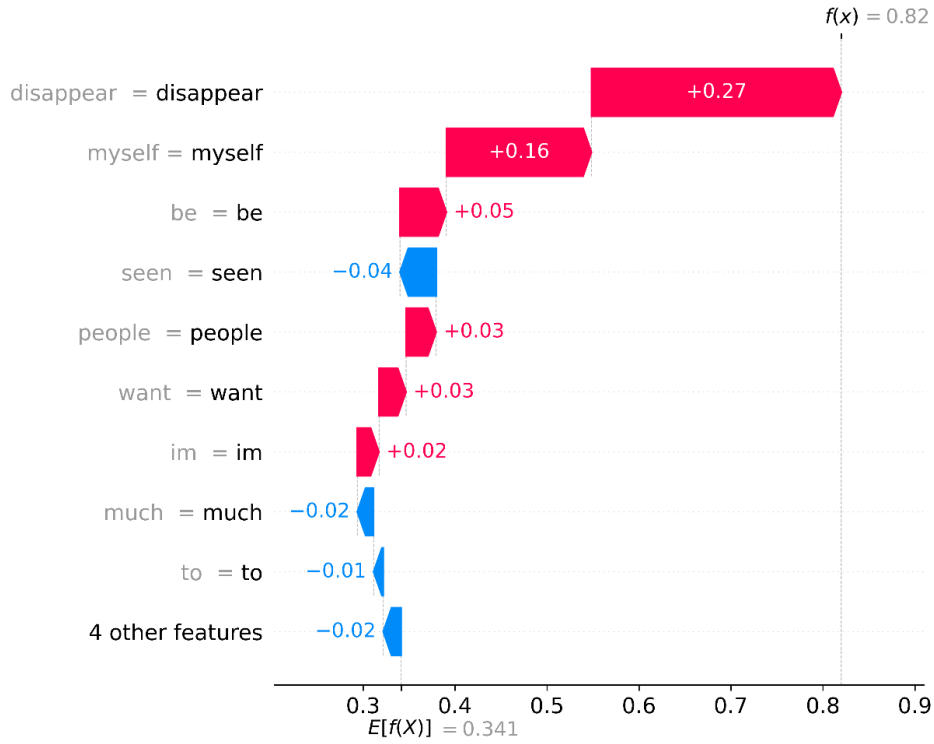


**Figure (Y) SHAP waterfall explanation for an individual prediction**

Figure (Y) shows a local explanation generated using the SHAP waterfall plot for a representative suicidal post. The baseline prediction score of 0.341 was progressively increased by several influential terms until it reached its final prediction probability of 0.82 for the suicidal class. Specifically, the words disappear (+.27), and myself (+.16) showed the strongest positive contributions, while the terms seen (-.03), much (-.02), and to (-.01) produced minor negative effects on the final prediction. The analysis shows that the model primarily relies on semantically meaningful expressions related to self-harm, social withdrawal, and emotional suffering when identifying high-risk posts. The explainability results validates that the predictions from ECHO-Net are driven by clinically relevant linguistic patterns rather than arbitrary textual artifacts therefore enhancing the reliability and practical applicability of the proposed framework.

## 7. CONCLUSION

This paper presents ECHO-Net, a hybrid explainable neural network optimized using contextual explanations for detecting suicide risk using social media texts. The proposed architecture combines BERTweet contextual transformer embeddings, bidirectional gated recurrent units (GRUs) for sequential learning, dilated bidirectional temporal convolutional networks (Bi-TCNs), an emotion-aware psychological attention module (EAPAM) with clinical psychological weightings, and a contrastive semantic alignment module (CSAM) for cross-domain generalization. A three-stage training methodology utilizing domain-adaptive pre-training, contrastive semantic learning and weighted-focal-loss based supervised fine-tuning was employed to handle class imbalance and domain shift.

Experimental results obtained via stratified 10-fold cross-validation on a compiled multi-source dataset of 57,306 instances showed mean accuracy of 98.11%, precision of 98.79%, recall of 99.31%, F1-score of 99.04% and area under receiver operating characteristic curve (AUROC) of 99.49%. It achieved higher performance compared to all evaluated baseline models including traditional machine learning, deep-learning, and transformer-based models. Most importantly, it achieved the lowest false negative rate of 0.69% indicating the least number of missed suicidal posts which is the most clinically important error type in automatic suicide detection systems.

Ablation studies were conducted to verify the contribution of each component of the proposed architecture to its overall performance. They revealed that all six components contributed significantly to its overall performance. Among them, the importance of BERTweet embeddings and weighted focal loss as the two most significant

contributions were further emphasized. Such high degree of transparency facilitates clinical deployment of the proposed framework since mental health practitioners can use auditable and accountable AI-assisted decision-making instead of relying upon uninterpretable “black box” decisions.

There are three main limitations associated with this work: (1) only English language social media posts have been used for evaluation; (2) no information regarding demographics (age, gender, location) has been included within the model and (3) although multiple data sources have been combined to create the dataset, longitudinal user history has not been captured thereby making it difficult to analyze how suicidal ideation escalates over time.

Possible future areas of investigation include: (1) extending the approach to multiple languages by using multilingual BERTweet; (2) incorporating additional modalities such as images, audio, and behavioural metrics into the model; (3) implementing federated learning techniques that enable distributionally-trained models while maintaining users' private data; (4) deploying the system in real-time on various social media platforms and creating pipeline for generating alerts when a post indicates potential suicidal ideation; and (5) conducting prospective clinical trials to evaluate the effectiveness of the developed system in preventing suicides.

Author contributions statement

Maya Gharat was primarily responsible for developing the research idea, designing the methodology, conducting experiments, analyzing the results, and writing the initial draft of the manuscript.She assisted with data preparation, result validation, literature survey, and interpretation of the findings. He also contributed to reviewing and improving the manuscript.

Manivel Kandasamy supervised the overall research work, provided technical guidance throughout the study, and contributed to revising the manuscript and refining the research outcomes.All authors discussed the results, reviewed the manuscript, approved the final version for publication, and agreed to take responsibility for the integrity and accuracy of the work.

Data Availability Statement: The data presented in this study are available here:

Dataset 1:https://github.com/IE-NITK/TwitterSuicidalAnalysis

Dataset 2:: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

Conflicts of Interest: The authors declare no conflict of interest.

**References:**


1. World Health Organization. (2023). Suicide fact sheet. https://www.who.int/news-room/fact-sheets/detail/suicide
2. Bilsen, J. (2018). Suicide and youth: Risk factors. Frontiers in Psychiatry, 9, 540. https://doi.org/10.3389/fpsyt.2018.00540
3. Castillo-Sánchez, G., Marques, G., Dorronzoro, E., Rivera-Romero, O., Franco-Martín, M., & de la Torre-Díez, I. (2020). Suicide risk assessment using machine learning and social networks: A scoping review. Journal of Medical Systems, 44(12), 205.
4. Mbarek, A., Jamoussi, S., & Hamadou, A. B. (2022). An across online social networks profile building approach: Application to suicidal ideation detection. Future Generation Computer Systems, 133, 171–183.
5. Bruen, A. J., Wall, A., Haines-Delmont, A., & Perkins, E. (2020). Exploring suicidal ideation using an innovative mobile app — Strength Within Me. JMIR Mental Health, 7(9), e18407.
6. Kina, E., Choi, J.-G., Ishaq, A., Shafique, R., Gracia Villar, M., Silva Alvarado, E., de la Torre Diez, I., & Ashraf, I. (2026). Suicide ideation detection using social media data and ensemble machine learning model. International Journal of Computational Intelligence Systems, 19, 100.
7. Gaur, M., Aribandi, V., Alambo, A., Kursuncu, U., Thirunarayan, K., Beich, J., Pathak, J., & Sheth, A. (2021). Characterization of time-variant and time-invariant assessment of suicidality on Reddit using C-SSRS. PLOS ONE, 16(5), e0250448.
8. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.
9. Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English tweets. Proceedings of EMNLP: System Demonstrations, 9–14.
10. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proceedings of EMNLP, 1724–1734.
11. Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

12. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of ICCV, 2980–2988.
13. Adarsh, V., Kumar, P. A., Lavanya, V., & Gangadharan, G. (2023). Fair and explainable depression detection in social media. Information Processing & Management, 60(1), 103168.
14. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
15. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of KDD, 1135–1144.
16. Mirtaheri, S. L., Greco, S., & Shahbazian, R. (2024). A self-attention TCN-based model for suicidal ideation detection from social media posts. Expert Systems with Applications, 255, 124855.
17. Sawhney, R., Joshi, H., Gandhi, S., & Shah, R. (2020). A time-aware transformer based model for suicide ideation detection on social media. Proceedings of EMNLP, 7685–7697.
18. Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L., & Thomson, J. (2022). An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. Journal of King Saud University — Computer and Information Sciences, 34(10), 9564–9575.
19. Ji, S., Yu, C. P., Fung, S.-F., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. Complexity, 2018, 6157249.
20. Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of suicide ideation in social media forums using deep learning. Algorithms, 13(1), 7.
21. Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L., & Thomson, J. (2022). An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. Journal of King Saud University — Computer and Information Sciences, 34(10), 9564–9575.
22. Mirtaheri, S. L., Greco, S., & Shahbazian, R. (2024). A self-attention TCN-based model for suicidal ideation detection from social media posts. Expert Systems with Applications, 255, 124855.
23. Ghosh, T., Al Banna, M. H., Al Nahian, M. J., Uddin, M. N., Kaiser, M. S., & Mahmud, M. (2023). An attention-based hybrid architecture with explainability for depressive social media text detection in Bangla. Expert Systems with Applications, 213, 119007.
24. Sawhney, R., Joshi, H., Gandhi, S., & Shah, R. (2020). A time-aware transformer based model for suicide ideation detection on social media. Proceedings of EMNLP, 7685–7697.
25. Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2021). DepressionNet: Learning multi-modalities with user post summarization for depression detection on social media. arXiv preprint arXiv:2105.10878.
26. Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, P., Guntuku, S. C., & Schwartz, H. A. (2019). Suicide risk assessment with multi-level dual-context language and BERT. Proceedings of CLPsych, 39–44.
27. Turcan, E., & McKeown, K. (2019). Dreaddit: A Reddit dataset for stress analysis in social media. Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 97–107.
28. Ren, J., Luo, J., Yao, Q., & Liu, X. (2022). Hybrid transformer networks for sentiment analysis. Information Sciences, 608, 1251–1268.
29. Boonyarat, P., Liew, D. J., & Chang, Y.-C. (2024). Leveraging enhanced BERT models for detecting suicidal ideation in Thai social media content amidst COVID-19. Information Processing & Management, 61(4), 103706.
30. Kancharapu, R., & Ayyagari, S. N. (2024). Suicidal ideation prediction based on social media posts using a GAN-infused deep learning framework with genetic optimization and word embedding fusion. International Journal of Information Technology.
31. Ghanadian, H., Nejadgholi, I., & Al Osman, H. (2024). Socially aware synthetic data generation for suicidal ideation detection using large language models. IEEE Access.
32. Abdulsalam, A., Alhothali, A., & Al-Ghamdi, S. (2024). Detecting suicidality in Arabic tweets using machine learning and deep learning techniques. Arabian Journal for Science and Engineering.
33. Adarsh, V., Kumar, P. A., Lavanya, V., & Gangadharan, G. (2023). Fair and explainable depression detection in social media. Information Processing & Management, 60(1), 103168.
34. Huang, X., Xia, L., & Ye, Y. (2023). Explainable transformers for social media health analytics. Expert Systems, 40(4).
35. Ahmed, R., Aslam, M., & Khan, M. A. (2024). Explainable artificial intelligence for healthcare text analytics: A comprehensive review. Artificial Intelligence Review, 57(1), 1–25.
36. Li, H., Zhang, Q., & Zhao, Y. (2023). Attention-enhanced transformers for emotional understanding. Neurocomputing, 517, 84–96.
37. Wang, P., Ji, L., & Xue, L. (2021). Deep contextual learning for suicide detection from social media. Knowledge-Based Systems, 219, 106–118.