I.J. Information Technology and Computer Science, **2026, 1, 140-162**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2026.01.08

Measuring Cognitive Distortions: A KPI-based Approach to Understanding Faulty Information Processing

Laxmi Jayannavar*
Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bengaluru, Karnataka, India (Affiliated to VTU)
E-mail: laxmijayannavar@gmail.com
ORCID iD: https://orcid.org/0009-0000-5412-4568
*Corresponding author

T. N. R. Kumar
Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bengaluru, Karnataka, India (Affiliated to VTU)
E-mail: tnrkumar@msrit.edu
ORCID iD: https://orcid.org/0000-0002-1670-2336

Shreekant Jere
Accenture AI, Bangalore, India
E-mail: shrikantjere@gmail.com

Received: 07 July 2025; Revised: 08 August 2025; Accepted: 03 October 2025; Published: 08 February 2026

Abstract: Cognitive distortion refers to the patterns of negative thinking which can distort a person's perception of reality. These distorted thoughts lead to unhealthy behaviors, emotional distress, and mental health issues, like depression and anxiety. In order to detect cognitive distortion, Deep Learning (DL) techniques are employed; however, these approaches lead to a high error rate and poor performance. This is mainly because they fail to understand the hierarchical semantics, subtle emotional tones, and long-range dependencies within the text. Hence, a new model termed Hierarchical Attention Neural Harmonic Fusion Network (HAN-HFNet) is exploited for cognitive distortion detection from text. Initially, the input sentence is passed to Bidirectional Encoder Representations from Transformers (BERT) tokenization, which generates context-aware embeddings capable of capturing subtle emotional nuances, long-range dependencies, and hierarchical semantics critical for identifying cognitive distortions in text. Next, various Key Performance Indicators (KPIs), like Severity of Cognitive Distortions (SCD), Frequency of Cognitive Distortion (FCD), Correlation Between Cognitive Distortions and Depression Severity, Cognitive Behavioral Therapy (CBT), self-reports of cognitive distortions from individuals, Long-Term Monitoring of Cognitive Distortions (LT-MCD), and impact on daily functioning is considered. Lastly, the cognitive distortion is detected utilizing HAN-HFNet, which is obtained by integrating Hierarchical Deep Learning for Text classification (HDLTex) and Deep High-order Attention neural Network (DHA-Net) using harmonic analysis. This fusion enables the model to learn both coarse and fine-grained features, enhancing contextual understanding and reducing error. Moreover, the performance of the HAN-HFNet is evaluated using the Faulty Information Processing Dataset (FIPD), and it computed a minimum classification error of 0.072, and maximum recall, accuracy, precision, and F1-score of 94.756%, 92.754%, 91.866%, and 93.289%. Furthermore, the model is suitable for integration into real-world mental health support systems, offering scalability and potential deployment in online therapy platforms, clinical decision-making tools, and cognitive behavioral assessment frameworks.

Index Terms*:* Faulty Information, Cognitive Distortion Detection, Text Data, Deep High-order Attention Neural Network, Hierarchical Deep Learning for Text Classification.

1. Introduction

Cognitive distortions refer to the dysfunctional core beliefs and misconceptions that people may hold that control

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

their feelings toward themselves and the world around them [1]. Negatively biased thought errors, also known as cognitive distortions, are thought to make people more vulnerable to depression. People react to events with automatic thoughts, which produce emotional and behavioral reactions. Automatic thought content usually aligns with a person's fundamental beliefs about significant aspects of the world, other people, and themselves. Maladaptive behaviors and negative affect can be influenced by negative, neutral, or even positive events when negative core beliefs arise and negative automatic thoughts are aroused. Over time, this pattern of emotions, thoughts, and behaviors may contribute to or maintain depressive symptoms [2]. One major factor contributing to the global burden of disability is depression [3] [4]. Depression-related disability has increased over the past few decades, especially among young people [5][6]. Cognitive distortions are distinct and identifiable maladaptive thought patterns that are connected to depression. These patterns cause individuals to think inaccurately and excessively negatively about the world, themselves, and the future. For instance, when people label themselves in negative, absolutist terms (e.g., “I am a loser”), they are connecting in a cognitive distortion that is prevalent in depression [7]. Suicidal ideation is predicted to be directly correlated with cognitive deficits and distortions. Moreover, a stronger correlation between suicidal ideation and cognitive distortions than cognitive deficits is predicted. Furthermore, it is hypothesized that cognitive distortions and deficits will exhibit reciprocal relationships, meaning that improvements in one will be connected to improvements in the other [8].

Social media has become a major platform for news consumption today. As it is free to use and easily accessible, it allows for rapid sharing of posts [9]. People are gradually seeking out and consuming news from social media platforms instead of traditional news organizations as they spend more and more of their lives interacting with others online [10].

Consequently, social media serves as a great platform for people to share and/or consume information. People are spending more and more time on social media. For example, according to Pew Research Center studies, 68% of Americans in 2018 derived some of their news from social media, a percentage that has gradually increased since 2016. The quality of news articles spread on social media is generally inferior to that of traditional news sources because there is no regulatory body in place to regulate social media. Alternatively, fake news is also made possible by social media [10]. Fake news is defined as intentionally disseminating incorrect information to deceive people. Both individuals and society at large are affected by fake news. Initially, fake news has the potential to disturb the news ecosystem's authenticity balance. Second, consumers are more likely to believe biased or false stories from fake news [11]. Some people and organizations use social media to disseminate false information in an effort to advance their political and financial gains. Additionally, it has been reported that the 2016 US presidential election was affected by fake news. Lastly, fake news has the potential for important effects on actual events. For instance, a Reddit fake news article termed "Pizzagate" caused a real shooting. Consequently, detecting fake news is a vital issue that requires attention [9].

Earlier, natural language classifiers are essentially developed using human-designed features (such as knowledge bases, special tree kernels, dictionaries, etc.) [12]. With the advancement of statistical learning approaches, several statistical learning techniques, such as Support Vector Machine (SVM) and Bag of Words (BoW), have been utilized to create automatic depression detection tasks [13] [14]. In recent times, Deep Learning (DL) methods have become popular for tasks involving the classification of natural language. The local features of sentences can be automatically captured by Convolutional Neural Networks (CNN) by exploiting layers with convolving filters. Consequently, CNN has demonstrated high efficacy in natural language classification as well as other Natural Language Processing (NLP) tasks. In terms of word representation, conventional NLP methods always depict words as indices in a vocabulary, excluding the relationships among words [15]. Prior DL techniques, such as CNN and Recurrent Neural Networks (RNN), have been shown to be superior in detecting depression [16] [17]. To boost performance, pre-trained frameworks like Bidirectional Encoder Representations from Transformers (BERT) have been devised [18]. Nevertheless, the majority of researchers currently focus on linguistic statistical features to determine whether or not a user is depressed. They ignore the long-running psychological research on depression in their attempt to model the data employing different statistical learning approaches. In contrast, the psychological understanding of depression is not contemplated. However, professionals will not assist from a binary assessment of depression only for the purposes of further treatment and intervention [19].

The purpose of this research is to detect cognitive distortion effectively using the proposed HAN-HFNet technique from textual messages. At first, the input sentence gathered for the dataset is tokenized using the BERT tokenizer. After this, various KPI indicators, like FCD, SCD, CBT, correlation among depression severity and cognitive distortion, impact on daily functioning, self-report of cognitive distortion, and LT-MCD are determined. Finally, the cognitive distortion detection is performed using HAN-HFNet, which is engineered by the combination of DHA-Net and HDLTex.

The key contribution of this research is as follows

- *Proposed HAN-HFNet for cognitive distortion detection:* In order to detect cognitive distortion from textual messages, a new approach known as HAN-HFNet is developed in this research. Furthermore, the HAN-HFNet technique is formulated by the amalgamation of DHA-Net and HDLTex. Moreover, various KPIs, like SCD, FCD, Correlation Between Cognitive Distortions and Depression Severity, CBT, self-reports of cognitive distortions from individuals, LT-MCD, and impact on daily functioning is considered.

The residual sections of this paper are systematized as follows: The motivation of the work, with the pros and cons of existing methods, is portrayed in Section 2. The detailed description of HAN-HFNet is elucidated in Section 3. The investigational results of HAN-HFNet are illustrated in Section 4, and the conclusion is depicted in Section 5.

2. Motivation

Cognitive distortion detection is the process of recognizing invalid or biased thought patterns in texts, including personalization, overgeneralization, and catastrophizing. Psychological support can be provided by identifying these distortions, which can have a harmful effect on mental health. Conventional methods for identifying cognitive distortions include rule-based systems, supervised machine learning, lexicon-based methods, and DL frameworks, such as transformers. Nonetheless, these techniques face difficulties, like a lack of sufficient labeled data, uncertainty in identifying distortions, the complication of language, and the inability to capture contextual nuances. Hence, in order to detect cognitive distortion more effectively, this research devises an effective technique termed HAN-HFNet.

2.1. Literature Survey

Wang, B., *et al*. [20] developed an Explainable depression detection method for detecting cognitive distortion. This approach achieved maximal accuracy, high reliability, and rapid computation. Unfortunately, this method did not integrate the inevitable combination of psychology and computing methods, particularly Artificial Intelligence (AI) research for more accurate depression detection and treatment. The Suicidal Ideation Support Model (SIdSM) was introduced by Benjachairat, P., *et al.* [19] to categorize the intensity of suicidal ideation in Twitter messages. This technique was more accurate and less expensive. Nevertheless, this approach did not improve the categorization or Cognitive Behavioral Therapy (CBT) for better suicide risk identification and management. Lalk, C., *et al*. [21] devised ML+Explainable AI to predict patient depression symptoms. This approach recognized the distortions with the greatest impact on features and provided an explanation for the predictions of individual transcripts. However, the outcomes of this method were not repeated on other databases to assure their accuracy. Li, F., *et al*. [22] introduced a Support Vector Machine (SVM) for predicting cognitive distortion. This approach was more scalable and was especially helpful for identifying subtle patterns in complex patient-therapist interactions, like cognitive distortions. Still, the performance of ML algorithms was affected because patients who finished the treatment trajectory in less than ten sessions were not included.

Mirtaheri, S.L., *et al*. [23] presented self-attention-based long short-term memory (LSTM)-Bidirectional Temporal Convolutional Networks (AL-BTCN) for suicidal ideation detection from social media posts. The AL-BTCN model attained great promise for suicide prevention programs in a variety of sectors and was able to precisely identify suicidal thoughts from social media posts. Nonetheless, this model did not examine various attention approaches or experiment with various combinations of learning algorithms for improved performance of the technique. Bathina, K.C., *et al.* [24] developed a theory-driven approach for the diagnosis of depression on social media. This method was suitable for real-time situations and showed excellent sensitivity and generalization. But this technique did not manage to reach a broader group of individuals who are familiar with public data availability and the significance of improving transparency through a well-managed consent process. Thekkekara, J.P., *et al*. [25] established an attention mechanism on CNN-Bidirectional long short-term memory (BiLSTM) (CBA) for depression detection on social media text. This method eradicated the high frequency of biased results while attaining high accuracy. Nonetheless, the CBA technique did not fully utilize the attention mechanism to understand words in a way that accounts for social and affective processes, going beyond just personal pronouns or singular terms in both the textual and visual context. Yang, K., *et al*. [26] introduced Knowledge–aware and Contrastive Network (KC-Net) for early stress and depression detection on social media. In practical applications, this method has been demonstrated to be very effective and robust. Nonetheless, this technique was not suitable for real-time applications for enhancing performance.

2.2. Major Challenges

The various approaches for detecting cognitive distortion in DL face the following difficulties:

- The Explainable depression detection [20] was devised for detecting cognitive distortions in depressed individuals. This model demonstrated the potential for AI to uncover complicated psychological patterns. But this method did not assist in determining how pervasive the cognitive distortions are in a person's thought patterns.
- The ML+Explainable AI in [21] demonstrated how language-based metrics can identify relevant processes that predict depression symptoms. However, this approach failed to self-report, giving individuals a chance to reflect on their thoughts, raising awareness of distorted thinking patterns, and assisting them in starting the process of change.
- SVM predicted continuous treatment results in [22] based on early symptom changes and pre-treatment predictors. Nevertheless, this strategy did not enhance the performance, and a comparatively small number of potential predictors had an effect on performance.
- The AL-BTCN [23] produced the superior performance rate in terms of error metrics and accuracy. However, the method was not enhanced by taking into account several words embedding techniques and providing the model with a variety of behavioral and social network profile features.
- Cognitive distortions are frequently subtle, interconnected with ordinary thoughts, and may be difficult to identify. It can be challenging to identify and precisely classify some distortions since they may be implicit and not immediately apparent.

3. Proposed Hierarchical Attention Neural Harmonic Network for Cognitive Distortion Detection

The goal of cognitive distortion detection is to identify and categorize the distorted thinking patterns in the textual data that are mostly seen in several mental health illnesses, namely stress, depression, and anxiety. Various prevailing DL techniques that are exploited to detect these distortions attain high false negatives, which leads to low precision. Therefore, a new DL method termed HAN-HFNet is devised for detecting the cognitive distortion from textual messages. Here, the input sentence is passed to the tokenization phase, where the BERT [27] is employed to tokenize the input. Following this, many KPI [21] [28], [29] indicators, including FCD, SCD, CBT, correlation among depression severity and cognitive distortion, impact on daily functioning, self-report of cognitive distortion, and LT-MCD, are contemplated. Following this, cognitive distortion detection is carried out by exploiting a Hybrid HAN-HFNet model, which is formulated by incorporating DHA-Net [30] and HDLTex [31]. The block diagram of HAN-HFNet for cognitive distortion detection is shown in Fig.1.

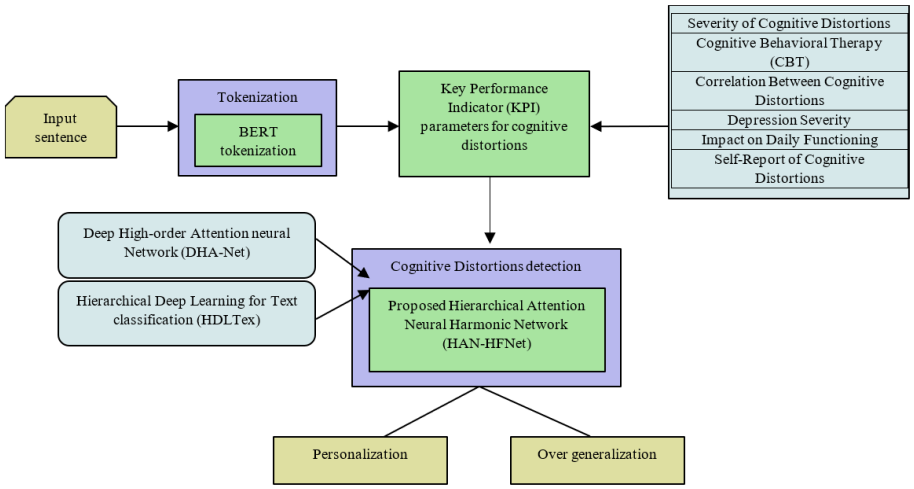


Fig.1. Block diagram of the HAN-HFNet for Cognitive Distortion detection

3.1. Data Acquisition

Consider the textual messages are taken from the Faulty Information Processing Dataset (FIPD) [32] for detecting the cognitive disorder. This database includes 2265 textual messages, with 310 messages taken from mental health blogs and 1955 messages from Twitter. Further, the FIPD database σ is specified by,

$$\sigma = \{\sigma_1, \sigma_2, ..., \sigma_d, ..., \sigma_z\} \tag{1}$$

where, σ_d indicates the d^{th} sentence and z signifies the total number of sentences in the database.

3.2. BERT Tokenization

Tokenization is the process of breaking the text into small units, which are termed tokens. Here, the tokens may be sentences, characters, subwords, or words, which are generated based on the level of tokenization. This process is employed in various NLP tasks, namely speech recognition, machine translation, and text classification. By converting the text into smaller words, the model's ability to understand and process the meaning and structure of the language more effectively is enhanced. Moreover, the BERT [27] is utilized here for tokenization, and this process is performed by considering σ_d as input. Furthermore, a specific type of tokenization termed WordPiece tokenization is employed by BERT for converting the text into tokens and then the tokens into indices. The classification token $[CLS]$ is added at the beginning of the input sequence, which is exploited for classification tasks, such as sentiment analysis. Besides, the separator token $[SEP]$ is utilized for separating the different segments or sentences in the input. Once the text is broken into tokens then the BERT tokenizer adds $[SEP]$ is at the end and $[CLS]$ at the beginning of the sequence. Furthermore, the outcome attained in the tokenization process is denoted as α_q.

3.3. KPI Parameters of Cognitive Distortions

KPI parameters [21,28,29] are utilized for measuring how the overall functioning and mental health of an individual effect thinking patterns. Moreover, these KPI patterns are used for tracking the progress in personal or therapy development when employed to manage or reduce cognitive distortions. Here, the tokenized output α_q is considered as input to the extraction of KPI parameters, such as SCD, FCD, Correlation Between Cognitive Distortions and Depression

Severity, CBT, self-reports of cognitive distortions from individuals, LT-MCD, and impact on daily functioning. Furthermore, the description of KPI parameters is explicated below.

A. FCD

The amount of cognitive distortion that is identified in a specified time frame is termed as the frequency of cognitive distortion. A high frequency represents severe depressive symptoms, and FCD is computed by,

$$W_1 = \frac{\gamma}{R} \tag{2}$$

Here, γ signifies the amount of cognitive disorder identified, W_1 designates the FCD indicator, and R implies the time period measured in months, weeks, or days.

B. SCD

The identified cognitive distortions' impact or intensity on the behavior and mood of an individual is termed SCD. The daily functioning of an individual is significantly affected by severe distortions, which are associated with higher depression severity. The formula given below is used to measure SCD.

$$W_2 = \sum_{h=1}^{\gamma} \left(U_h * Z_h \right) \tag{3}$$

wherein, Z_h typifies the behavioral or psychological impact of h^{th} cognitive distortion, and is measured on a scale, for example $1-5$ or $1-10$, W_2 characterizes the SCD indicator, U_h specifies the intensity score of h^{th} cognitive distortion, which is measured on a scale, for instance $1-5$ or $1-10$.

C. CBT

The CBT parameter refers to the degree of reduction in distortions, suggesting improvement in depressive symptoms after effective therapy. Furthermore, the CBT feature is specified as W_3.

D. Correlation between Cognitive Distortion and Depression Severity

This parameter is defined as the strength of the relation between the level of depressive symptoms and the amount of severity of cognitive distortion, and is formulated as,

$$W_4 = \frac{\sum \left(\xi_h - \overline{\xi} \right) \left(D_x - \overline{D} \right)}{\sqrt{\sum \left(\xi_h - \overline{\xi} \right)^2 \left(D_x - \overline{D} \right)^2}} \tag{4}$$

wherein, $\overline{D}$ exemplifies the mean of depression severity data, which is represented as weak $W_4 < 0.3$ or no correlation $W_4 \approx 0.3$, $\overline{\xi}$ stipulates the mean of cognitive distortion data, D_x implies the individual data points for depression severity, W_4 embodies the correlation among depression severity and cognitive distortion, ξ_h signifies the individual data points for cognitive distortion.

E. Impact on Daily Functioning

The amount to which the cognitive distortion weakens the quality of life and daily activities of an individual is referred to as the impact on daily functioning, and it is designated as,

$$W_5 = \sum \left(G_h - M_h \right) / v \tag{5}$$

Here, v symbolizes the entire amount of functional area evaluated, W_5 represents the impact on daily functioning, M_h designates the severity of h^{th} cognitive distortion in every area, namely the frequency of emotional disturbance or distorted thinking, and is measured on a scale 1 to 5, here 1 denotes low severity and 5 exemplifies high severity, G_h denotes the level of impairment in h^{th} area, like self-care, social interactions, or work, and is measured on a scale 1 to 5, where 5 indicates the extreme impairment and 1 represents no impairment.

F. Self-report of Cognitive Distortion

The severity and amount of cognitive distortion that are reported by the patient themselves is defined as self-report of cognitive disorder, and is represented by,

$$W_6 = \frac{\sum\left(W_1 \times M_h\right)}{v} \tag{6}$$

where, W_1 designates the FCD, which is measured on a scale 0 to 4, here 0 implies never, 1 epitomizes rarely, 2 represents sometimes and 4 symbolizes always. W_6 characterizes the self-report of cognitive disorder.

G. LT-MCD

The LT-MCD parameter is employed to track the reduction or persistence of cognitive disorder over an extended period, and it is expressed as,

$$W_7 = \frac{\sum\left((W_1)^*_{h,R} \times M_{h,R} \times \rho_{h,R}\right)}{v} \tag{7}$$

wherein, the LT-MCD parameter is typified as W_7, W_1 typifies the FCD, for instance number of times the distortion occurs per week or month, $\rho_{h,R}$ represents the impact of h^{th} cognitive distortion at R^{th} time, and is measured on a scale 1 to 5, where 5 characterizes the extreme impact and 1 signifies no impact. In addition, the KPI parameters that are specified above are combined to form a KPI indicator W, and is signified as,

$$W = \{W_1, W_2, W_3, ..., W_7\} \tag{8}$$

3.4. Cognitive Distortion Detection Using Hierarchical Attention Neural Harmonic Network

Cognitive distortion is considered an irrational or biased pattern of thinking that significantly affects the emotional behavior and well-being of an individual. Moreover, this distortion leads to several mental health challenges, such as stress, depression, anxiety, and so on. Moreover, various cognitive distortions can be detected automatically from behavioral data, speech, or text. Numerous prevailing DL models are exploited for detecting the cognitive disorder, but they lead to inaccurate outcomes with high error rates. Therefore, a new approach known as HAN-HFNet is developed to detect cognitive distortion. Moreover, HAN-HFNet is formulated by integrating HDLTex [31] and DHA-Net [30]. Further, the HAN-HFNet comprises three modules, like the DHA-Net model, the HAN-HFNet layer, and the HDLTex model.

Additionally, the process performed to detect cognitive distortion is as follows: First, the multiplication of the KPI indicator W with the weight ϖ_1 is performed, and the resultant output is then normalized. Further, the KPI indicator W is subjected to HDLTex, and the generated outcome is multiplied by the normalized output $\sum \varpi_1 W$ to attain an output η_1. Alternatively, the KPI indicator W is forwarded to DHA-Net, and the resultant output is then multiplied by the weight ϖ_2 and the HDLTex output η_1 for producing η_2. At last, the attained outcomes η_1, and η_2 are unified by employing Harmonic analysis [33] in the HAN-HFNet layer to generate the final output η_3, which is the detected cognitive distortion. Fig. 2 shows the structural model of HAN-HFNet for cognitive distortion detection.

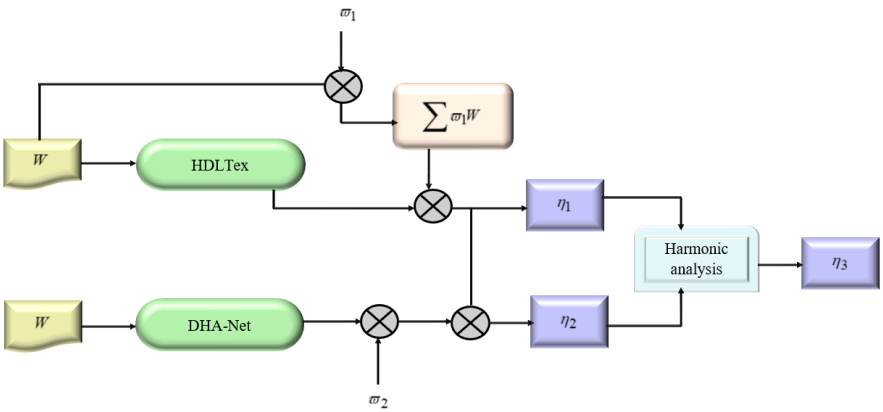


Fig.2. Structural model of HAN-HFNet for cognitive distortion detection

A. HDLTex Model

HDLTex [31] is a progressive DL approach, which is designed for processing and classifying text data by contemplating the hierarchical structure and exploiting the attention mechanism. Moreover, this model intends to detect and classify the cognitive distortions in textual data. Cognitive distortions, like overgeneralization and personalization, manifest in subtle ways in text, and HDLTex provides a sophisticated model for detecting the patterns. Further, the attention mechanism allows HDLTex to concentrate on the most significant features of text, which is essential to detect cognitive distortions. This approach has high scalability and is more effective, which makes it applicable for applications that are required to analyze large volumes of text. Here, the KPI indicator W is fed as input to HDLTex. HDLTex includes three parts, such as (i) a Multi-Layer Perceptron (MLP) classifier which is used for making the prediction of classes at each level on the basis of dynamically generated level masking and document representation, (ii) an attention module, which is used for generating the dynamic document representation in diverse level of classification, (iii) bidirectional Long Short Term Memory (LSTM) encoder is used for transforming every word into vector representation in terms of their context. Besides, the hierarchical classification approach is regarded as a sequence-to-sequence method, where a sequence of hierarchical class labels is generated by employing a sequence of word embeddings. Further, a modified attention module from the conventional attention mechanism is utilized rather than computing attention weights that are trained on the decoder's hidden state at every m^{th} time step, conditioned on the parent category embedding E_{s-1}. Assume a document with u tokens $L=(\tau_1,\tau_2,...,\tau_u)$ and category labels of y levels $A=(E_1,...,E_y)$, $E_s \in \{E_1^{\vartheta_s},...,E_{\ell_s}^{\vartheta_s}\}$. Here, the number of classes in a level s^3 is typified as ℓ_s, and s^{th} level of class taxonomy is represented as ϑ_s. Initially, bidirectional LSTM is exploited for mining the features in the document, and is signified as,

$$\overrightarrow{P_j}=\overrightarrow{LSTM}\left(\tau_j,\overrightarrow{P_{j-1}}\right) \tag{9}$$

$$\overleftarrow{P_j}=\overleftarrow{LSTM}\left(\tau_j,\overleftarrow{P_{j-1}}\right) \tag{10}$$

Here, τ designates the input word embedding at j^{th} time step, the hidden state of the encoder $V=(P_1,...,P_u)$ is generated by concatenating $\left(\overrightarrow{P_j}\right)$ and $\left(\overleftarrow{P_j}\right)$ as $P_m=\left[\overrightarrow{P_j},\overleftarrow{P_j}\right]$, where u represents the tokens.

Furthermore, the contextual word features $\overline{V_s}$ are formed during the classification of class labels at s^{th} level by concatenating the predicted category embedding E_{s-1} with the outcomes of the encoder $V=(P_1,..,P_u)$, and are modeled by,

$$\overline{V_s}=V\oplus E_{s-1} \tag{11}$$

Additionally, the u vectors in $\overline{V_s}$ are transformed into u attention scores via a sequence of non-linear and linear transformations, which is designated as,

$$B_s=soft\max\left[\tau_{\ell 2}\tanh\left(Q_{\ell 1}\overline{V}_s^T\right)\right] \tag{12}$$

Here, weight is exemplified as Q. Since a single attention distribution focuses on a specific module of the semantics in the document, y hops of attention are performed, and then a multi-head attention matrix $F_s(y\times u)$ is formed. Further, the Frobenius norm penalty $\lambda=\left\|F_sF_s^T-I\right\|_\varepsilon^2$ is exploited for encouraging diversity over the attention distribution's multiple hops, and for forcing the attention hops to concentrate on several features of the semantics. Here, the Frobenius norm of a matrix is indicated as $\left\|F_sF_s^T-I\right\|_\varepsilon^2$.

At the s^{th} level, the document representation is attained by multiplying the contextual word features and the multi-head attention matrix, which is represented as,

$$W_s=Q_{\ell 3}F_s\overline{V_s} \tag{13}$$

At last, at s^{th} level, a two-layer MLP is utilized for classifying the category, and the output of the two-layer MLP is written as,

$$n_s = \mathrm{R}\, ELU\left(Q_L\left[W, n_s - 1\right]\right) \tag{14}$$

where, the output generated in HDLTex is indicated as n_s.

Moreover, the resultant outcome is multiplied by normalized output $\sum \varpi_1 W$ to produce η_1, which is signified by,

$$\eta_1 = \mathrm{R}\, ELU\left(Q_L\left[W, n_s - 1\right]\right) * \sum \varpi_1 W \tag{15}$$

wherein, η_1 stipulates the output of the HDLTex module. The HDLTex architecture is displayed in Fig. 3.

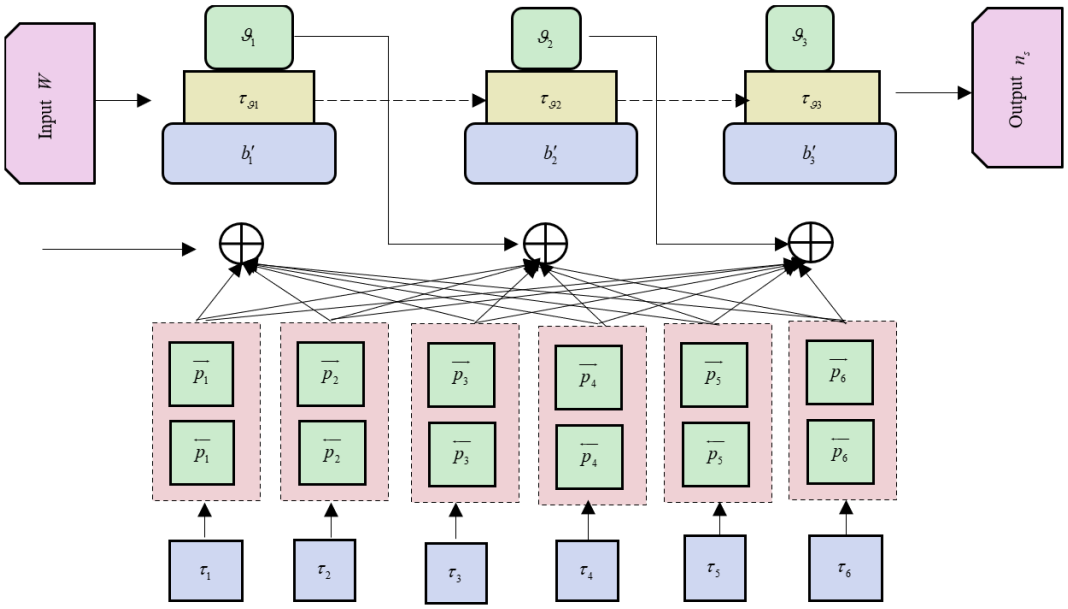


Fig.3. HDLTex architecture

B. DHA-Net Model

DHANet [30] is an advanced neural network framework, which is devised for handling complex tasks, like modeling and detection of cognitive distortions. Moreover, cognitive distortion refers to a biased pattern of thinking, which negatively affects the mental health of individuals. Here, the attention mechanism is leveraged by DHA-Net for capturing the high-order relation in the KPI indicator W, which is vital to understanding and predicting complex cognitive patterns. In addition, a high-order attention mechanism is used to concentrate on the significant aspects in the KPI indicator W, which enhances the accuracy of detecting cognitive distortions.

The DHA-Net approach comprises two components, namely high-order pooling, and Effective Attention Module (EAM). The structure of the ResNet-18 model is utilized by DHA-Net. Besides, at every residual block, the EAM attention module is inserted, which aids with weight revisions through channel attention learning. Moreover, the high-order pooling layer is connected before the last fully connected layer for capturing complex higher-order statistical information, which maximizes the accuracy rate.

a. EAM

In several vision and language-related applications, the attention mechanism is used mostly for attaining performance improvements. EAM-Net is developed by considering an interactive attention approach, called Squeeze-and-Excitation (SENet) for cross-channel attention, and it does not reduce the dimensionality. After applying the global average pooling layer on the convolutional (Conv) features, EAM is applied. The Conv kernel's size is computed by EAM, and then a single-dimensional Conv operation is executed. After this, the sigmoid function is employed for squeezing the values into a range of (0,1). Besides, the reason for using the attention module is explained below.

- ***No Reduction in Dimension:*** Consider the output from the Conv block as $T \in \square^{\varsigma \times p \times l}$, here the number of channels is indicated as ς, features map's height and width is implied as l and p. Further, the weight of the channel Ω is computed by,

$$\Omega = a\left[\mu_{\{\iota_1, \iota_2\}}\left(\nu(i)\right)\right] \tag{16}$$

Here, the sigmoid function is represented as a, $\mu_{\{t_1,t_2\}}$ specifies the function that includes the weight matrices t_1 and t_2, Global Average Pooling (GAP) is implied as $\nu(i)$, and is specified by,

$$\nu(i) = \frac{1}{pl} \sum_{e=1,r=1}^{p,l} T_{e,r} \tag{17}$$

Furthermore, consider $W = \nu(i)$, then $\mu_{\{t_1,t_2\}}$ is computed by,

$$\mu_{\{t_1,t_2\}}(W) = p_2 \operatorname{Re} LU\left(p_{1W}\right) \tag{18}$$

wherein, weights are indicated as p_1 and p_2. The channel does not have any direct link with the weight, as it minimizes the complexity. For example, the weight of every channel in the model is calculated by a fully connected layer. Further, the channel features are projected into a low-dimensional space, and then it is mapped to make an indirect correspondence between the weight and the channel. Moreover, the single-dimensional convolution of size X is performed by EAM on the features gathered from GAP to make a direct correspondence between the weight and channel. Here, X is computed from local channel interaction, and it is explicated beneath.

- *Local cross-channel interaction:* The local interactions in each channel are significant as they contribute to the efficacy and validity of the method. The cross-channel interactions are captured easily by a 1D convolution with kernel dimension X on the input W, and it is designated by,

$$\Omega = a\left[ODC_X(W)\right] \tag{19}$$

Here, ODC epitomizes 1D convolution.

b. High-order Pooling Module

This module is very efficient in contrast to first-order pooling, as it aids in obtaining high statistical information concerning a group of features. Recently, high-order covariance pooling that is applied to features is derived from the first-order pooling, and then it is normalized using matrix power. In order to attain second-order statistical feature information, covariance pooling is exploited. Furthermore, the mathematical process of this module is described as follows.

- *High-order Normalization:* The input feature is considered as $Y \in \square^{\varsigma \times \omega \times f}$, where the height and width of the feature map of the input are indicated as f and ω. Moreover, the overall feature map is written as,

$$g = \phi\left(Y; \omega, c\right) \tag{20}$$

Here, deviation is specified as c, CNN feature mapping is denoted as ϕ. Besides, the g^{th} feature map $g \in \square^{\varsigma \times \omega \times f}$ is converted into a dimensional feature matrix $g \in \square^{o \times b}$, and here $b = f \times \omega$ features. The covariance matrix specified as H is designated as,

$$H = gIg^T \tag{21}$$

wherein, the transpose of the matrix is represented as T, $b \times b$ identity matrix is symbolized as I. The covariance matrix H after eigenvalue decomposition is applied, and is represented by,

$$H \to \left(G', Z'\right), H = G'Z'G'^T \tag{22}$$

Here, the diagonal matrix is characterized as $Z' = diag\left(\zeta_1, ..., \zeta_o\right)$, and G' symbolizes the orthogonal matrix.

- *Covariance Normalization:* In the calculation of the high-order pooling layer, the eigenvalue decomposition plays a significant role. Further, Newton-Schultz iteration (NSI) is utilized to accelerate the computation of each calculation present in the high-order normalization. When applying NSI to expression (22), we get

$$S_w = \frac{1}{2} S_{w-1} \left(3I - N_{w-1} S_{w-1}\right) \tag{23}$$

$$N_w = \frac{1}{2} N_{w-1} \left(3I - N_{w-1} S_{w-1}\right) \tag{24}$$

Expressions (23) and (24) are applicable for GPU computation, which is computed using normal matrix multiplication. NSI is applied to further accelerate the solution, and is specified by the expression given below.

$$\hat{H} = \frac{1}{tr\left(H\right)} H \tag{25}$$

wherein, $tr\left(H\right)$ exemplifies the trace of the matrix. Moreover, the effect of expression (25) is balanced by,

$$H = \sqrt{tr\left(H\right)} S_w \tag{26}$$

The outcome attained from DHA-Net is signified by the following expression.

$$A' = \mu\left(p * W + d'\right) \tag{27}$$

Here, A' denotes the DHA-Net model's outcome, d' implies the bias, p represents weight, and μ denotes the activation function.

Further, the outcome generated by the DHA-Net is first multiplied by a weight ϖ_2 and then with η_1 to obtain η_2, which is represented as,

$$\eta_2 = \mu\left(p * W + d'\right) * \varpi_2 * \left[\mathrm{R}\, ELU\left(Q_L\left[W, n_s - 1\right]\right) * \sum \varpi_1 W\right] \tag{28}$$

wherein, η_2 designates the outcome when multiplied by weight ϖ_2, and then with η_1. Fig. 4 displays the DHA-Net architecture.

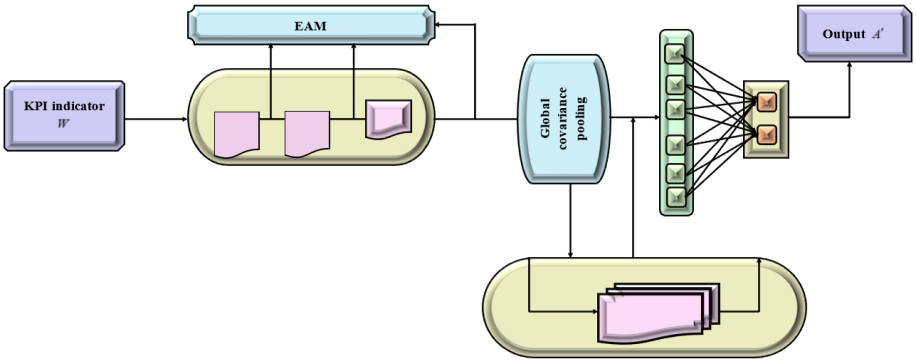


Fig.4. DHA-Net architecture

C. HAN-HFNet Layer

In this layer, the outputs η_1 and η_2 from HDLTex and DHA-Net are combined to attain the outcome η_3. Further, Harmonic analysis [33] is utilized for integrating the outcomes from HDLTex and DHA-Net modules. Here, Harmonic analysis is employed for simplifying the complex issues by allowing effective computation and boosting the performance. This approach intends to provide complex functions to be classified as the amalgamation of modest periodic functions. Here, the time series model is used to assess the unknown deterministic components for fusion methods. In addition, the periodical structure for time series is specified as,

$$C_{(1)}, C_{(2)}, ..., C_{(k)}, ..., C_{(s')} \tag{29}$$

Here, s' postulates the length, the k^{th} time series data is embodied as $C_{(k)}$, and is modeled as,

$$C_{(k)} = D_0' + \sum_{u'=1}^{q'} \left[D_{u'}' \cos\left(2\pi u' k / s'\right) \right] + \left[M_{u'}' \sin\left(2\pi u' k / s'\right) \right] \tag{30}$$

where, the preselected constants are epitomized as D' and M', u' specifies the number of cycles. Assume $q' = 1$ and $s' = 2$, hence equation (30) becomes

$$C_{(k)} = D_0' + D_1' \cos\left(2\pi k / 2\right) + M_1' \sin\left(2\pi k / 2\right) \tag{31}$$

$$C_{(k)} = D_0' + D_1' \cos\left(\pi k\right) + M_1' \sin\left(\pi k\right) \tag{32}$$

Here,

$$D_0' = \frac{1}{s'} \sum_{k=1}^{s'} C_k \tag{33}$$

$$D_0' = \frac{1}{2} \left[C_{(1)} + C_{(2)} \right] \tag{34}$$

$$D_{u'}' = \frac{2}{s'} \sum_{k=1}^{s'} C_k \cos\left(2\pi u' k / s'\right) \tag{35}$$

$$D_1' = C_{(1)}(-1) + C_{(2)}(1) \tag{36}$$

$$M_{u'}' = \frac{2}{s'} \sum_{k=1}^{s'} C_{(k)} \sin\left(2\pi u' k / s'\right) \tag{37}$$

$$M_1' = C_{(1)}(0) + C_{(2)}(0) \tag{38}$$

Additionally, the time series is regarded as $C(k-1), C(k)$ and $C(k+1)$, and hence we attain,

$$\left.\begin{aligned} C_{(1)} &= C(k-1) \\ C_{(2)} &= C(k) \\ C_{(k)} &= C(k+1) \end{aligned}\right\} \tag{39}$$

By applying equations (34), (36), (38), and (39) in equation (32), we get,

$$C(k+1) = \frac{1}{2} \left[C(k-1) + C(k) \right] - \left[C(k-1) - C(k) \right] \cos(\pi k) + 0 \tag{40}$$

$$C(k+1) = C(k-1) \left[\frac{1 - 2\cos(\pi k)}{2} \right] + C(k) \left[\frac{1 + 2\cos(\pi k)}{2} \right] \tag{41}$$

Assume $C(k+1) = \eta_3, C(k-1) = \eta_1$, and $C(k) = \eta_2$, we get

$$\eta_3 = \eta_1 \left[\frac{1 - 2\cos(\pi k)}{2} \right] + \eta_2 \left[\frac{1 + 2\cos(\pi k)}{2} \right] \tag{42}$$

$$\eta_3 = \left(\begin{aligned} &\left(\mathrm{R}\, ELU\left(Q_L \left[W, n_s - 1 \right] \right) * \sum \varpi_1 W \right) \left[\frac{1 - 2\cos(\pi k)}{2} \right] + \\ &\left(\mu\left(p * W + d' \right) * \varpi_2 * \left[\mathrm{R}\, ELU\left(Q_L \left[W, n_s - 1 \right] \right) * \sum \varpi_1 W \right] \right) \left[\frac{1 + 2\cos(\pi k)}{2} \right] \end{aligned} \right) \tag{43}$$

Here, the output attained in the HAN-HFNet layer is exemplified as η_3, which indicates the presence or absence of cognitive distortion.

The pseudocode of the HAN-HFNet technique for cognitive distortion detection is portrayed in Algorithm 1.

Algorithm 1. Pseudocode of HAN-HFNet

Input

Training: Training feature W_{train}, weight of HDLTex ϖ_1, Weight of DHA-Net ϖ_2 and testing data labels

Testing: Testing feature W_{test}, weight of HDLTex ϖ_1, Weight of DHA-Net ϖ_2 and testing data labels

Output

Bias and Weight of HAN-HFNet

Predicted labels of HAN-HFNet η_3

Required parameters

Maximum Epoch U'_{epo}

Learning rate g'

Batch size

Begin

Initialization

Bias and Weight of hybrid HAN-HFNet arbitrarily

Initialize the batch size, learning rate, and epoch

//Training phase

while $t < U'_{epo}$

for all tuning feature W_{train}

Execute HDLTex by multiplying the weight ϖ_1 with the KPI indicator W and the attained outcome to be $\Sigma\varpi_1$

Multiply the outcome $\Sigma\varpi_1$ with the outputs from HDLTex after applying KPI indicator W as input, the final outcome generated from HDLTex after multiplication is η_1.

$$\eta_1 = \mathrm{R}\,ELU\left(Q_L\left[W, n_s - 1\right]\right) * \sum \varpi_1 W$$

Multiply the weight ϖ_2 with the outputs from DHA-Net after applying KPI indicator W as input, and the resultant output is to be again multiplied with η_1 to get the final outcome η_2.

$$\eta_2 = \mu\left(p * W + d'\right) * \varpi_2 * \left[\mathrm{R}\,ELU\left(Q_L\left[W, n_s - 1\right]\right) * \sum \varpi_1 W\right]$$

Apply harmonic analysis to get the output η_3 by integrating η_1 and η_2.

$$\eta_3 = \begin{pmatrix} \left(\mathrm{R}\,ELU\left(Q_L\left[W, n_s - 1\right]\right) * \sum \varpi_1 W\right)\left[\frac{1 - 2\cos(\pi k)}{2}\right] + \\ \left(\mu\left(p * W + d'\right) * \varpi_2 * \left[\mathrm{R}\,ELU\left(Q_L\left[W, n_s - 1\right]\right) * \sum \varpi_1 W\right]\right)\left[\frac{1 + 2\cos(\pi k)}{2}\right] \end{pmatrix}$$

end for

4 Compute the error $\hbar'$ by considering η_3 and tuning features

5 Update the bias and weight of HAN-HFNet based on the Adam optimizer

$t++$

End while

For testing phase

For all testing feature W_{test}

Compute η_1, η_2 and η_3 using tuned HAN-HFNet

Find the predicted label η_3

End for

End

4. Results and Discussion

The outcomes and discussion of the newly formulated HAN-HFNet approach, in contrast to various prevailing models for detecting cognitive distortion, are explicated in this segment.

4.1. Experimental Setup

The implementation of the HAN-HFNet technique for cognitive distortion detection is performed in Python version 3.9.11. The hardware requirements of this analysis are Windows 10 OS, 8GB RAM, and 1.7 GHz CPU. Moreover, the simulation parameters of HAN-HFNet are specified in Table 1.

Table 1. Simulation parameter of HAN-HFNet

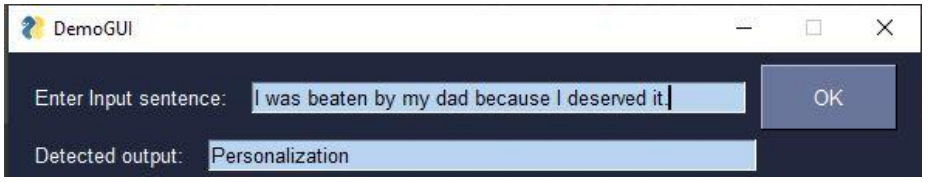
Parameter	Values
Loss	Binary cross entropy
Beta 2	0.9999
Epoch	80
Activation	Sigmoid
Learning rate	0.01
Beta 2	0.9999
Batch size	32
Optimizer	Adam
Epsilon	0.0001

4.2. Dataset Description

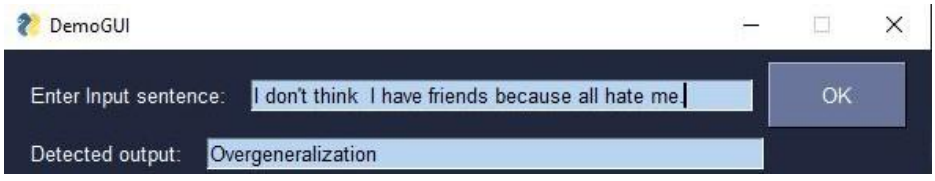
The FIPD dataset [32] is exploited for identifying two cognitive distortions, namely Overgeneralization and Personalization, which play an essential role in diagnosing and predicting depression. This database comprises of 2,265 textual messages, where 310 are from mental health blogs and 1,955 from Twitter. Moreover, FIPD is developed to support the understanding of Faulty Information Processing in clinical chatbot-generated interactions.

4.3. Experimental Outcomes

The experimental outcomes of the newly presented HAN-HFNet model in detecting cognitive distortion are represented in Fig. 5. The experimental result of HAN-HFNet for personalization is displayed in Fig. 5(a). In Fig. 5(b), the experimental result of HAN-HFNet based on overgeneralization is depicted.



(a)



(b)

Fig.5. Experimental outcome of HAN-HFNet for (a) Personalization and (b) Overgeneralization

4.4. Evaluation Metrics

The performance measures, such as recall, classification error, precision, accuracy, and F1-score are used for evaluating the effectiveness of the HAN-HFNet technique. Besides, the expression of these metrics is elucidated below.

- *Accuracy:* The ratio of the number of accurate predictions to the total number of instances predicted by the HAN-HFNet approach is termed accuracy, and is designated as,

$$Accuracy = \frac{H_1' + H_2'}{H_1' + H_2' + H_3' + H_4'} \tag{44}$$

wherein, H_3' and H_2' indicate the false positive and true negative, H_4' and H_1' denote the false negative and true positive.

- *Recall:* This measure is referred to as the ratio of the number of true positives, which is the precisely detected cognitive distortions, to the entire amount of actual cognitive distortions. Here, recall is specified as,

$$\text{Re}\,call = \frac{H_1'}{H_1' + H_4'} \tag{45}$$

- *Classification error:* The classification error metric is the inverse of the accuracy, and it calculates the rate of inaccurate predictions, which is expressed as,

$$Error = 1 - Accuracy \tag{46}$$

- *Precision:* The proportion of true positives which is the precisely detected cognitive distortions to the whole number of instances that HAN-HFNet detected as cognitive distortions is called precision. Further. The precision is modelled as,

$$\text{Pr}\,ecision = \frac{H_1'}{H_1' + H_3'} \tag{47}$$

- *F1-score:* F1-score is used to evaluate the performance of the proposed HAN-HFNet model in accurately detecting cognitive distortions from text data. The formula is given as,

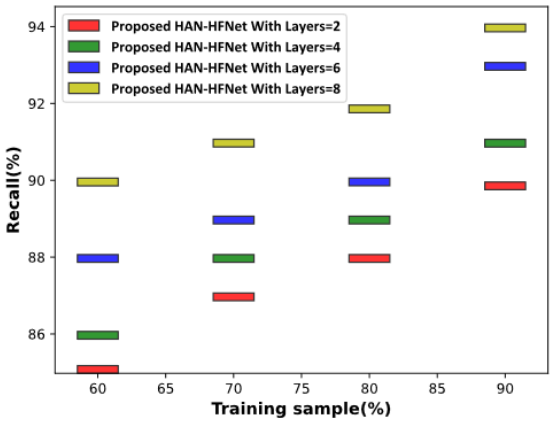
$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{48}$$

4.5. Performance Analysis

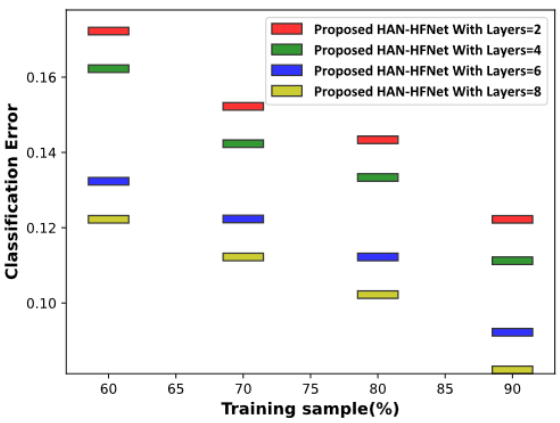
The examination of the HAN-HFNet approach to detect cognitive distortion concerning k-fold and training samples on the basis of several evaluation measures is discussed below.

A. For Training Sample

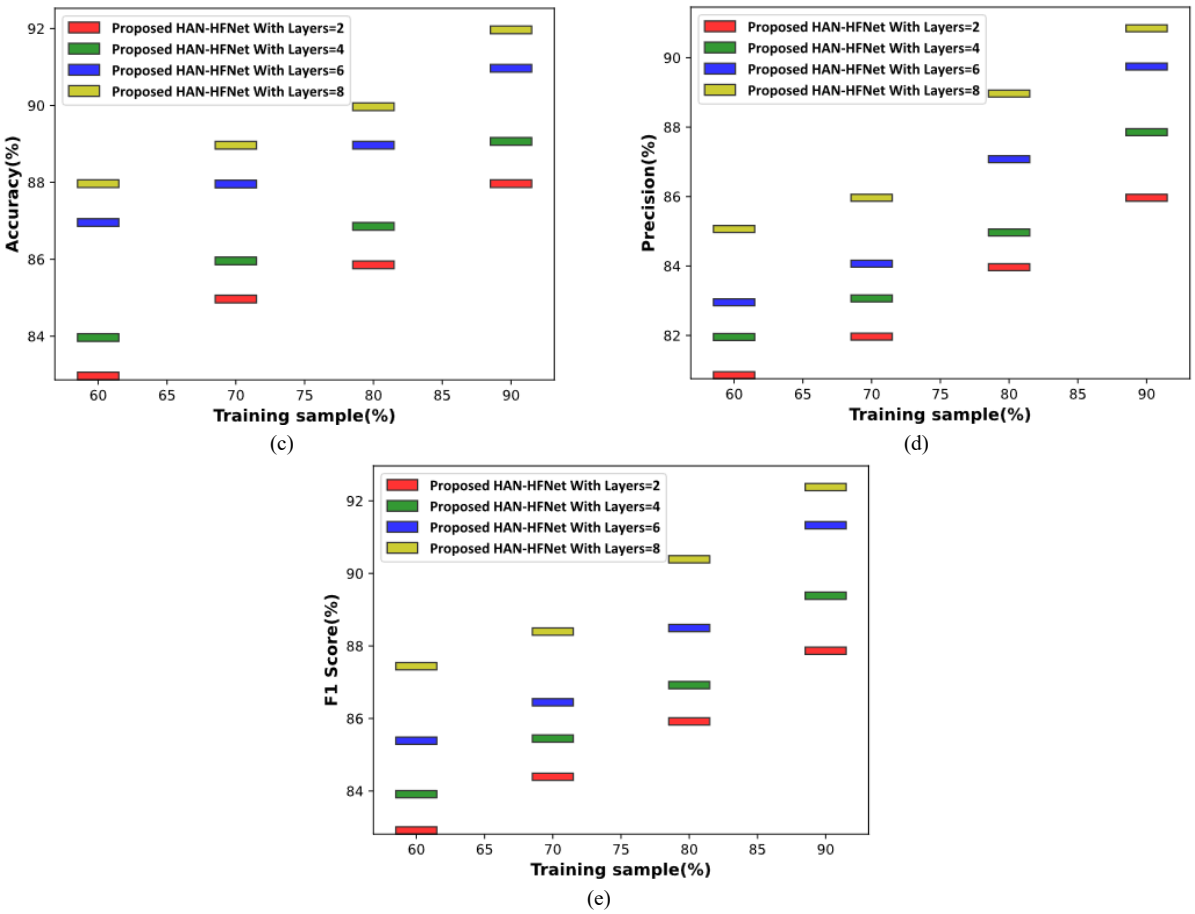
The assessment of the HAN-HFNet technique based on training samples with respect to many metrics is delineated in Fig. 6. The examination of HAN-HFNet regarding recall is specified in Fig. 6(a). At a training sample of 70%, the HAN-HFNet with 2, 4, 6, and 8 layers achieved a recall of 85.876%, 87.876%, 88.876%, and 90.876%. In Fig. 6(b), the classification error-based investigation of HAN-HFNet is signified. The classification error measured by the HAN-HFNet with 2, 4, 6, and 8 layers for a training sample of 80% is 0.131, 0.121, 0.111, and 0.091. The graph drawn among the training sample and accuracy is designated in Fig. 6(c). For the training sample of 90%, the accuracy calculated by HAN-HFNet with 2 layers is 87.877%, 4 layers is 88.876%, 6 layers is 89.876%, and 8 layers is 91.876%. Fig. 6(d) illustrates the valuation of HAN-HFNet concerning precision. When assuming a training sample of 60%, the HAN-HFNet model with 2, 4, 6, and 8 layers recorded a precision of 80.765%, 81.876%, 83.876%, and 85.876%. The valuation of the F1-score is provided in Fig. 6(e). When assuming a training sample of 70%, the HAN-HFNet model with 2, 4, 6, and 8 layers recorded a F1-score of 83.823%, 85.356%, 86.830%, and 89.293%.



(a)



(b)



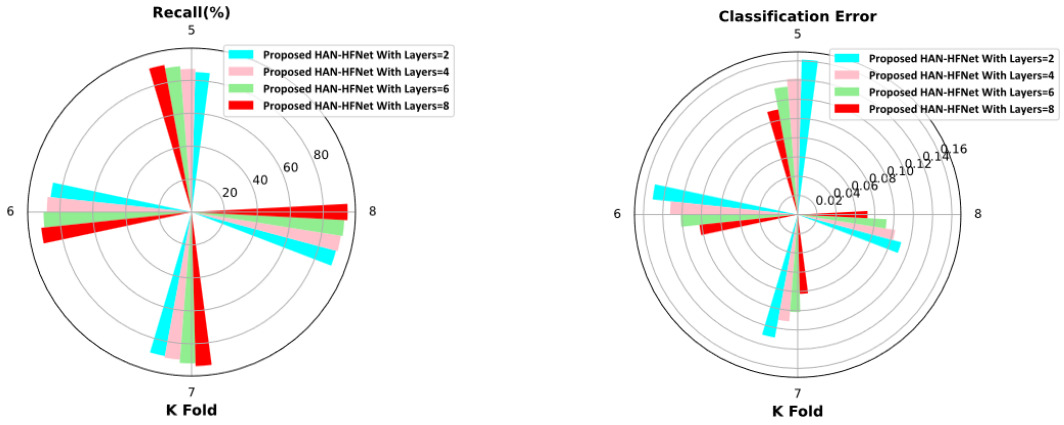
(c) (d)

(e)

Fig.6. Performance valuation of HAN-HFNet for training sample (a) Recall, (b) Classification error, (c) Accuracy, (d) Precision, and (e) F1-score

B. For K-fold

Fig. 7 delineates the assessment of the HAN-HFNet technique for k-fold with respect to many metrics. The graph drawn between k-fold and recall is designated in Fig. 7(a). For a k-fold of 7, the recall calculated by HAN-HFNet with 2 layers is 88.876%, 4 layers is 89.876%, 6 layers is 91.876%, and 8 layers is 93.543%. Fig. 7(b) illustrates the valuation of HAN-HFNet concerning classification error. When assuming a k-fold of 6, the HAN-HFNet model with 2, 4, 6, and 8 layers recorded the classification error of 0.151, 0.132, 0.121, and 0.102. In Fig. 7(c), the accuracy-based investigation of HAN-HFNet is signified. The accuracy measured by the HAN-HFNet with 2, 4, 6, and 8 layers for a k-fold of 8 is 88.876%, 89.765%, 90.766%, and 92.754%. The examination of HAN-HFNet regarding precision is specified in Fig. 7(d). At a k-fold of 5, the HAN-HFNet with 2, 4, 6, and 8 layers achieved a precision of 82.979%, 84.877%, 85.987%, and 87.765%**.** The F1-score examination is provided in Fig. 7(e). At a k-fold of 7, the HAN-HFNet with 2, 4, 6, and 8 layers achieved a F1-score of 87.403%, 88.351%, 90.351%, and 92.186%.



(a) (b)

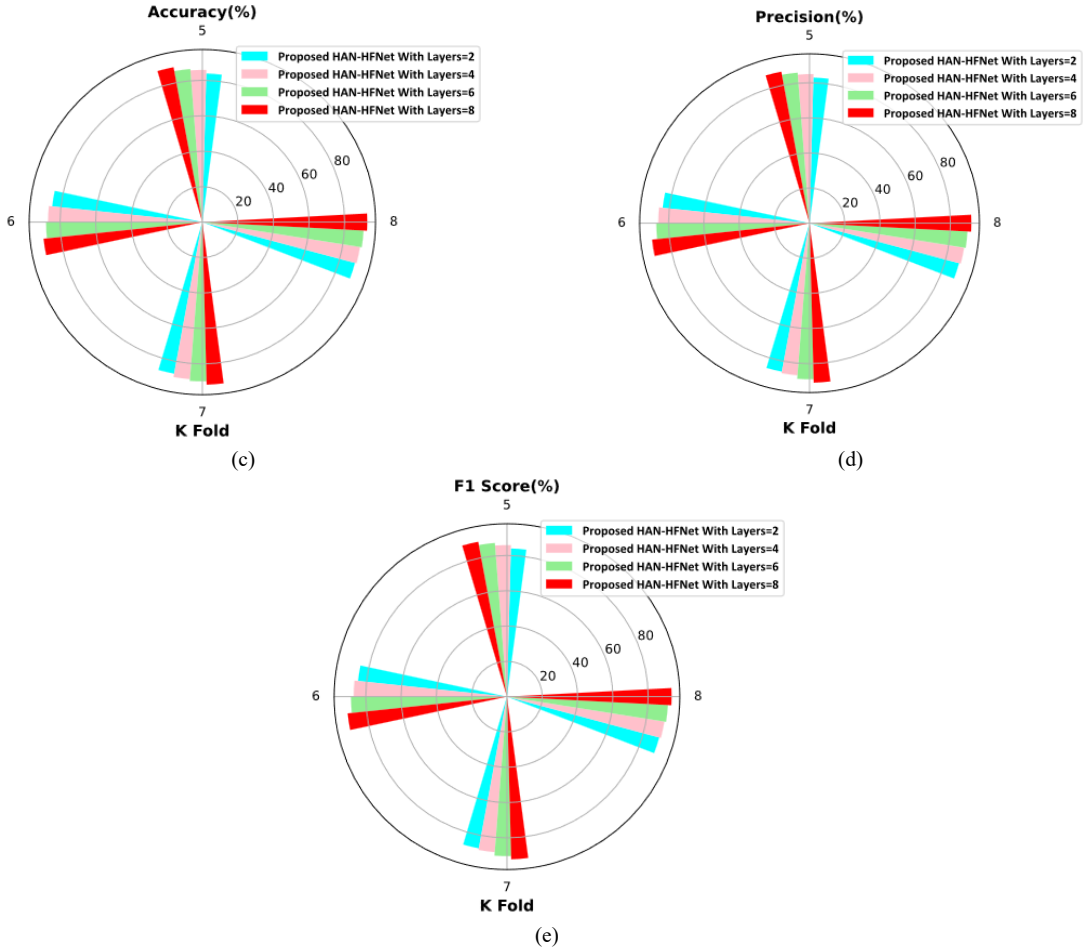


Fig.7. Performance valuation of HAN-HFNet for k-fold (a) Recall, (b) Classification error, (c) Accuracy, (d) Precision, and (e) F1-score

4.6. Comparative Techniques

The conventional approaches, namely Explainable depression detection [20], ML+Explainable AI [21], SVM [22], AL-BTCN [23], CNN [16], and LSTM [34] are correlated with the newly devised HAN-HFNet to examine its efficacy during cognitive distortion detection.

4.7. Comparative Analysis

The comparative examination of the HAN-HFNet approach in detecting cognitive distortion is effectuated with varying k-fold and training samples on the basis of several evaluation measures and is demonstrated below.

A. For Training Sample

In Fig. 8, the examination of HAN-HFNet based on various measures concerning the training sample is exemplified. Fig. 8(a) specifies the investigation of HAN-HFNet with respect to recall. The recall value calculated by HAN-HFNet for the training sample 70% is 90.876%, whereas the traditional models, namely Explainable depression detection, ML+Explainable AI, SVM, AL-BTCN, CNN, and LSTM attained a recall of 82.977%, 84.876%, 86.876%, 88.543%, 89.389%, and 89.777%. The performance of HAN-HFNet is maximized by 8.69%, 6.60%, 4.40%, 2.57%, 1.64%, and 1.21%. The plot of training samples and classification error is represented in Fig. 8(b). At a training sample of 80%, the classification error quantified by Explainable depression detection, ML+Explainable AI, SVM, AL-BTCN, CNN, LSTM, and HAN-HFNet is 0.151, 0.141, 0.131, 0.113, 0.107, 0.101, and 0.091. Fig. 8(c) depicts the assessment of HAN-HFNet based on accuracy. For the training sample of 60%, the accuracy computed by Explainable depression detection is 80.877%, ML+Explainable AI is 82.977%, SVM is 83.876%, AL-BTCN is 85.876%, CNN is 86.456%, LSTM is 86.877%, and HAN-HFNet is 87.877%. The performance enhancement of HAN-HFNet is 7.97%, 5.58%, 4.55%, 2.28%, 1.62%, and 1.14%. In Fig. 8(d), the precision-based valuation of HAN-HFNet is represented. Here, the prevailing models, like Explainable depression detection, ML+Explainable AI, SVM, AL-BTCN, CNN, and LSTM measured precision of 80.868%, 81.876%, 85.866%, 86.765%, 89.369%, and 89.887% for a training sample of 90%, while the proposed HAN-HFNet quantified a precision of 89.654%. The F1-score analysis is provided in Fig. 8(e). Here, the Explainable depression detection, ML+Explainable AI, SVM, AL-BTCN, CNN, and LSTM measured F1-score of 84.877%, 87.876%, 88.876%, 89.766%, 90.367%, and 91.770% for a training sample of 80%, while the HAN-HFNet quantified a F1-score of 92.877%.

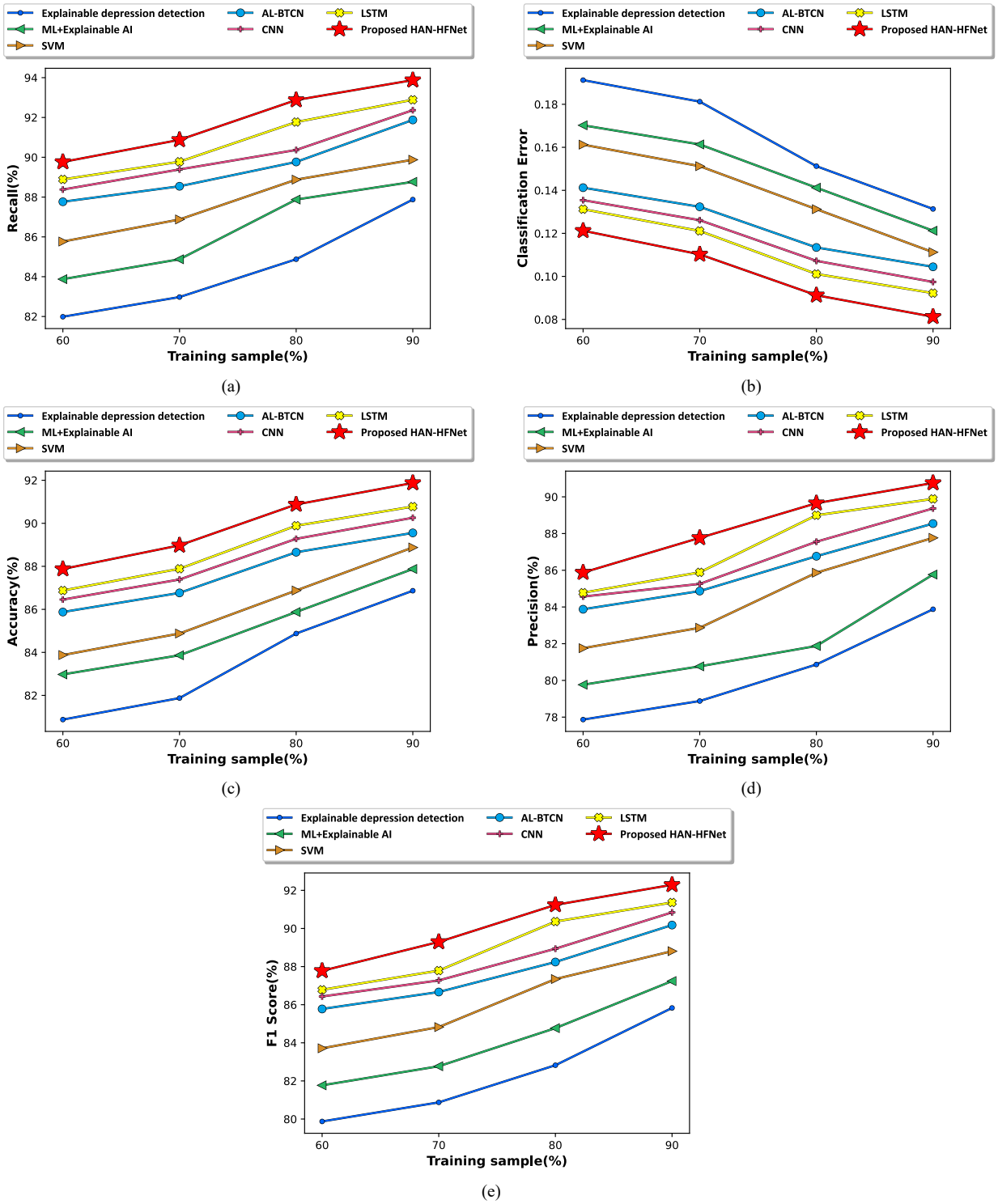


Fig.8. Comparative valuation of HAN-HFNet concerning training sample (a) Recall, (b) Classification error, (c) Accuracy, (d) Precision, and (e) F1-score

B. For K-fold

Fig. 9 designates the examination of HAN-HFNet relating to k-fold. The plot of k-fold and recall is represented in Fig. 9(a). At a k-fold of 8, the recall quantified by Explainable depression detection, ML+Explainable AI, SVM, AL-BTCN, CNN, LSTM and HAN-HFNet is 88.765%, 89.877%, 91.765%, 92.676%, 93.468%, 93.877%, and 94.756%. The performance of HAN-HFNet is maximized by 6.32%, 5.15%, 3.16%, 2.20%, 1.36%, and 0.93%. Fig. 9(b) specifies the investigation of HAN-HFNet with respect to classification error. The classification error value calculated by HAN-HFNet for a k-fold of 5 is 0.111, whereas the traditional models, namely Explainable depression detection, ML+Explainable AI, SVM, AL-BTCN, CNN, and LSTM attained an error of 0.171, 0.160, 0.142, 0.133, 0.126, and 0.122. In Fig. 9(c), the accuracy-based valuation of HAN-HFNet is represented. Here, the prevailing models, like Explainable depression

detection, ML+Explainable AI, SVM, AL-BTCN, CNN, and LSTM measured an accuracy of 84.987%, 86.977%, 88.765%, 89.876%, 90.367%, and 90.777% for a k-fold of 7, while the proposed HAN-HFNet quantified a higher accuracy of 91.754%. Fig. 9(d) depicts the assessment of HAN-HFNet based on precision. For k-fold of 8, the precision computed by Explainable depression detection is 83.768%, ML+Explainable AI is 85.977%, SVM is 88.877%, AL-BTCN is 89.765%, CNN is 90.368%, LSTM is 90.877% and HAN-HFNet is 91.866%. The F1-score assessment is given in Fig. 9(d). For k-fold of 6, the F1-score computed by Explainable depression detection is 82.891%, ML+Explainable AI is 84.829%, SVM is 86.826%, AL-BTCN is 88.190%, CNN is 88.800%, LSTM is 89.762% and HAN-HFNet is 90.806%.

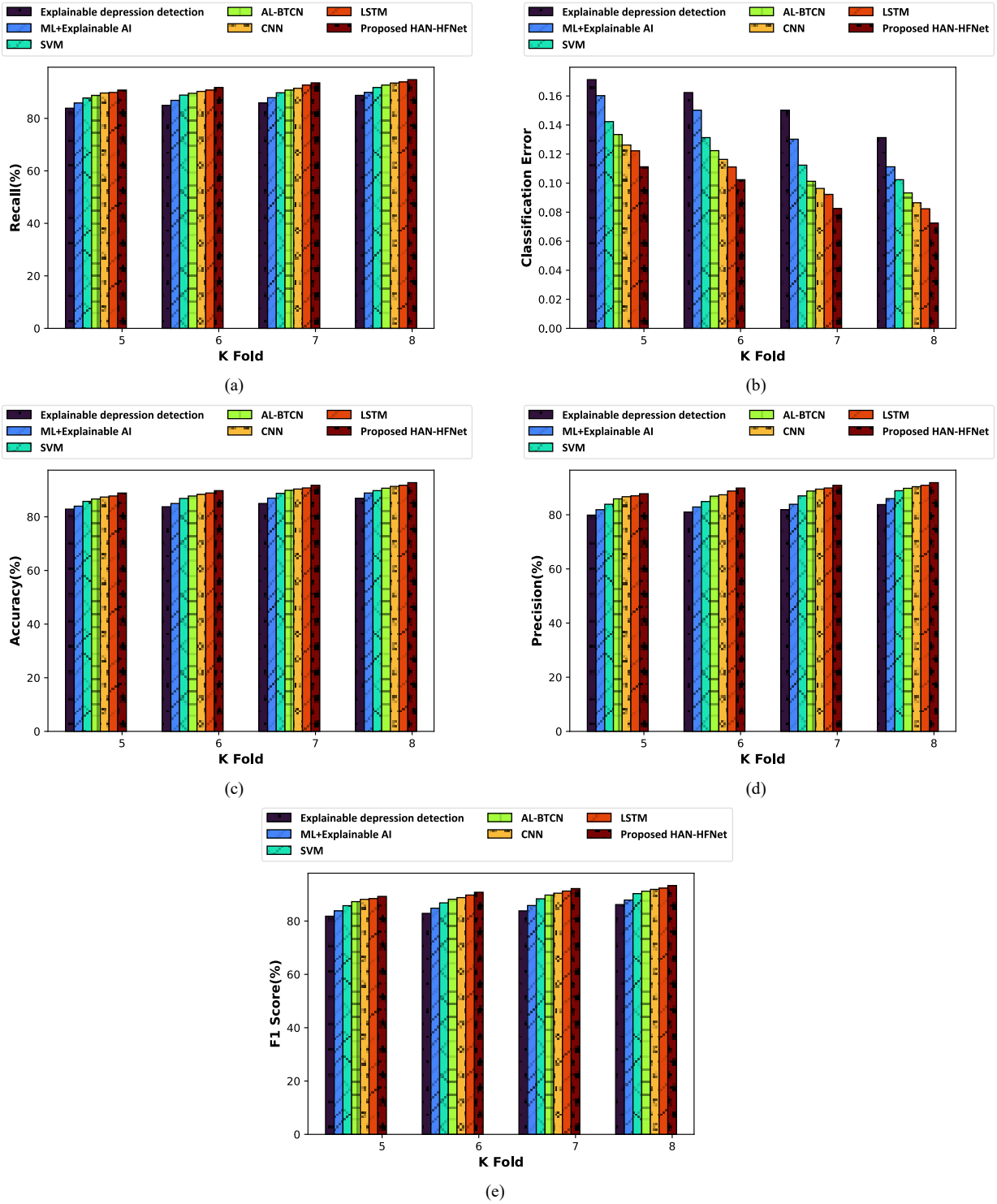


Fig.9. Comparative valuation of HAN-HFNet with respect to k-fold (a) Recall, (b) Classification error, (c) Accuracy, (d) Precision, and (e) F1-score

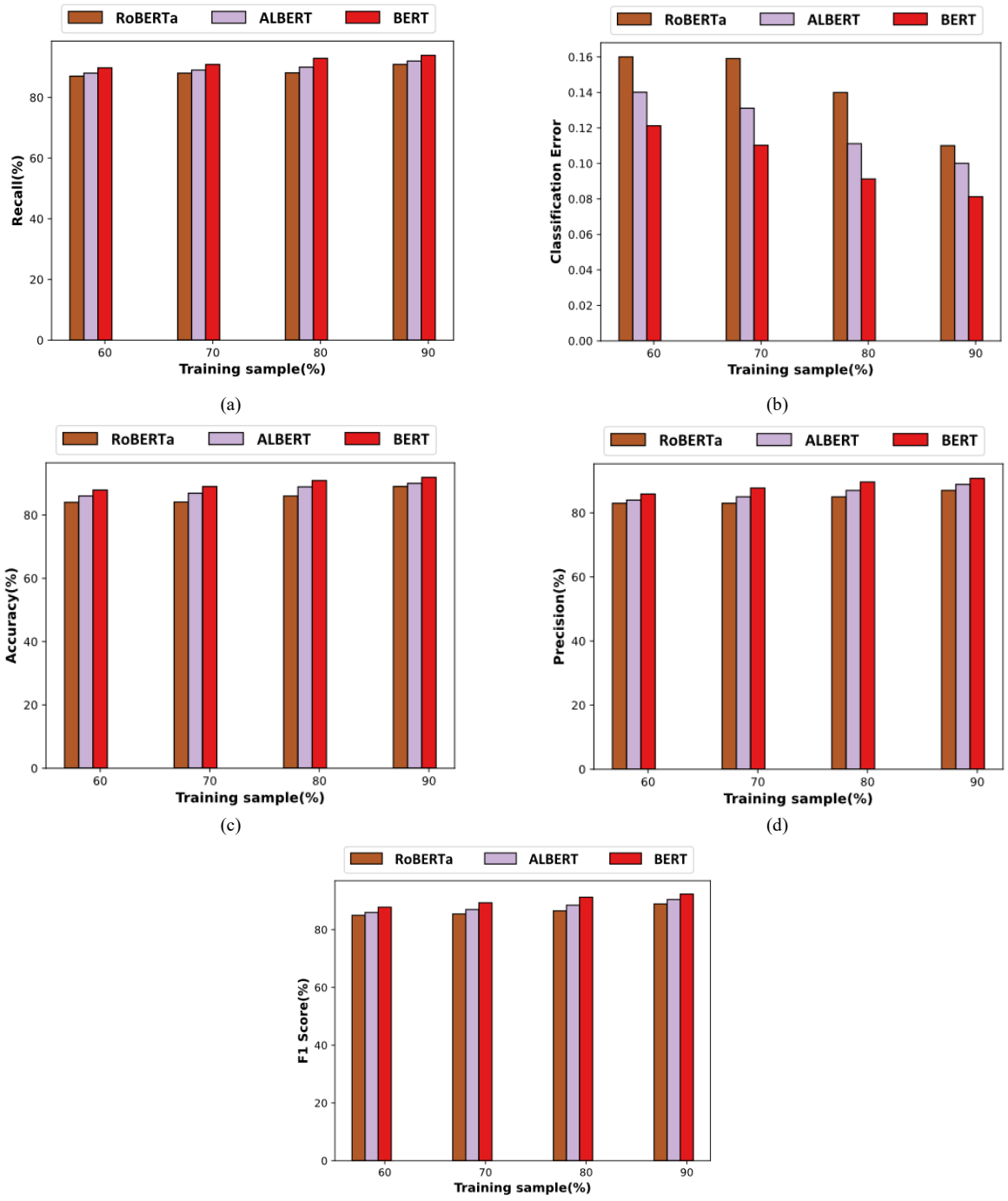


Fig.10. Comparative valuation of different tokenization methods (a) Recall, (b) Classification error, (c) Accuracy, (d) Precision, and (e) F1-score

4.8. Analysis based on Tokenization Methods

Fig. 10 depicts the examination of HAN-HFNet based on various tokenization methods. Here, the tokenization performance of BERT is compared with other tokenization methods, such as Robustly Optimized BERT Approach (RoBERTa), and A Lite BERT (ALBERT). The recall plot is represented in Fig. 10(a). At a training sample of 90%, the recall quantified by the RoBERTa, ALBERT, and BERT methods is 90.89%, 91.99%, and 93.88%. Fig. 10(b) specifies the investigation with respect to classification error. The classification error value calculated by BERT for a training sample of 70% is 0.110, whereas the RoBERTa, and ALBERT attained classification errors of 0.159, and 0.131. In Fig. 10(c), the accuracy-based valuation is represented. Here, the models, like RoBERTa, ALBERT, and BERT measured an accuracy of 86.00%, 88.89%, and 90.88% for a training sample of 80%. Fig. 10(d) depicts the assessment based on precision. For training sample of 60%, the precision computed by RoBERTa is 82.99%, ALBERT is 83.99%, and BERT is 85.88%. The F1-score assessment is given in Fig. 10(d). For the training sample of 90%, the F1-score computed by RoBERTa is 88.89%, ALBERT is 90.41%, and BERT is 92.29%. Overall, across all metrics, BERT consistently

outperformed RoBERTa and ALBERT, making it the most effective tokenizer for cognitive distortion detection in the proposed HAN-HFNet framework.

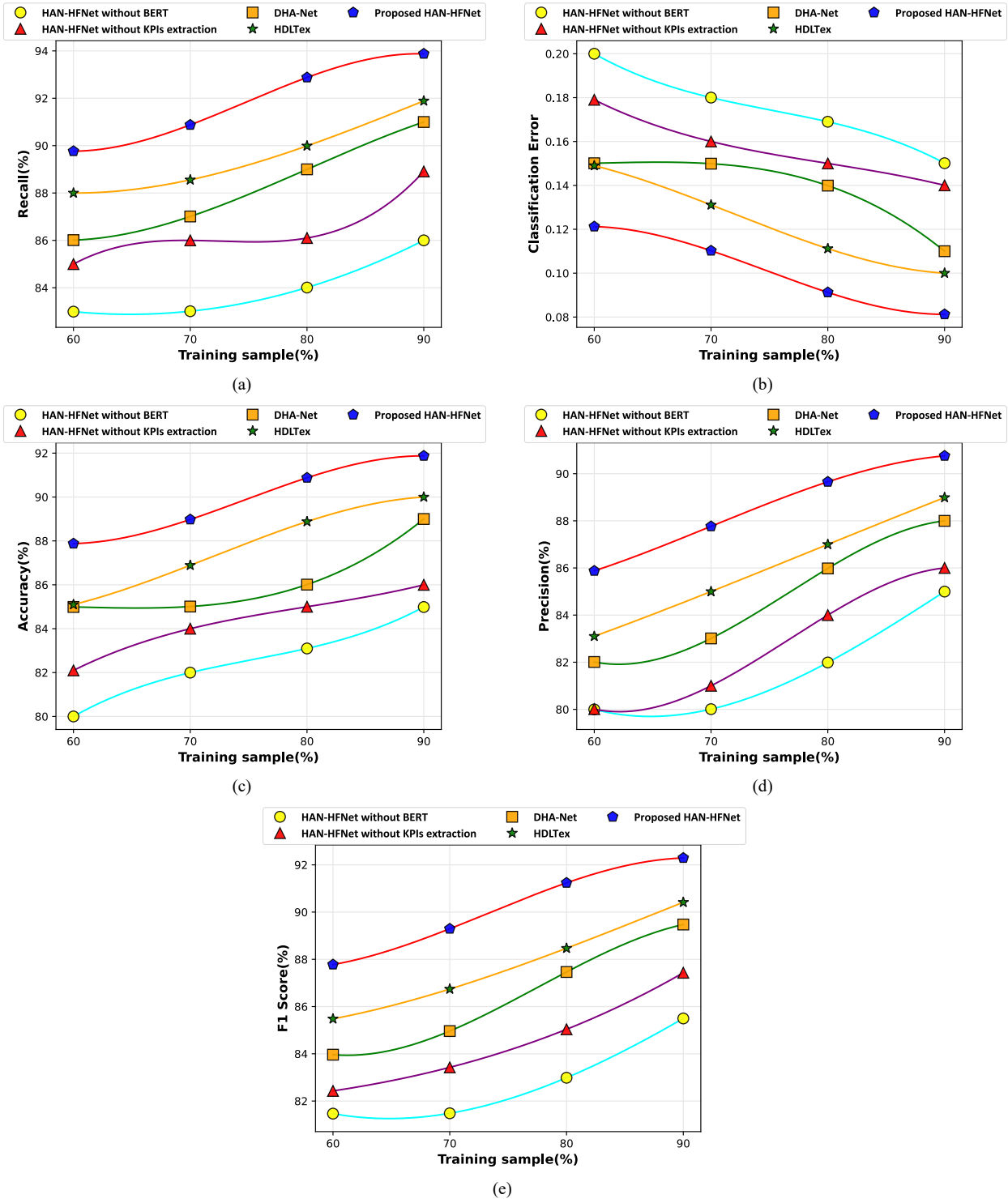


Fig. 11. Ablation study (a) Recall, (b) Classification error, (c) Accuracy, (d) Precision, and (e) F1-score

4.9. Ablation Study

Fig. 11 presents the ablation study conducted to evaluate the individual contribution of each component within the proposed HAN-HFNet framework. The recall plot is given in Fig. 11(a). At a training sample of 60%, the recall quantified by the HAN-HFNet without BERT, HAN-HFNet without KPIs extraction, DHA-Net, HDLTex, and proposed HAN-HFNet are 82.99%, 85.00%, 86.01%, 88.00%, and 89.77%. Fig. 11(b) specifies the investigation with respect to classification error. The classification error value calculated by the proposed HAN-HFNet for a training sample of 70% is 0.110, whereas the HAN-HFNet without BERT, HAN-HFNet without KPIs extraction, DHA-Net, and HDLTex

attained classification errors of 0.180, 0.160, 0.150, and 0.131. In Fig. 11(c), the accuracy-based valuation is represented. Here, the HAN-HFNet without BERT, HAN-HFNet without KPIs extraction, DHA-Net, HDLTex, and proposed HAN-HFNet measured an accuracy of 83.10%, 85.00%, 86.01%, 88.88%, and 90.88% for a training sample of 80%. Fig. 11(d) depicts the assessment based on precision. For the training sample of 90%, the precision computed by the HAN-HFNet without BERT is 85.00%, HAN-HFNet without KPIs extraction is 86.00%, DHA-Net is 88.00%, HDLTex is 88.99%, and the proposed HAN-HFNet is 90.76%. The F1-score assessment is given in Fig. 11(d). For the training sample of 80%, the F1-score computed by the HAN-HFNet without BERT is 82.99%, HAN-HFNet without KPIs extraction is 85.04%, DHA-Net is 87.46%, HDLTex is 88.47%, and the proposed HAN-HFNet is 91.24%. This confirms that integrating all components results in a more balanced and effective classification model.

4.10. Comparative Discussion

The comparative discussion of the HAN-HFNet approach with various evaluation metrics contrasted to different existing methods based on k-fold and training samples is illustrated in Table 2. Here, the newly introduced HAN-HFNet computed a minimum classification error of 0.072, and maximum recall, accuracy, precision, and F1-score of 94.756%, 92.754%, 91.866%, and 93.289% for a k-fold of 8. Besides, the traditional techniques including Explainable depression detection, ML+Explainable AI, SVM, AL-BTCN, CNN, and LSTM figured a recall of 88.765%, 89.877%, 91.765%, 92.676%, 93.468%, and 93.877% while accuracy recorded by these prevailing models are 86.868%, 88.876%, 89.766%, 90.676%, 91.357%, and 91.766%. Moreover, the precision value measured by Explainable depression detection is 83.768%, ML+Explainable AI is 85.977%, SVM is 88.877%, AL-BTCN is 89.765%, CNN is 90.368%, and LSTM is 90.877% whereas the error computed by these models is 0.131, 0.111, 0.102, 0.093, 0.086, and 0.082. The F1-score achieved by the existing models are 86.194%, 87.883%, 90.298%, 91.197%, 91.892%, and 92.353%. Hence, the HAN-HFNet technique measured better results than other traditional models. Moreover, HDLTex is flexible and has attained high scalability, while DHA-Net has more flexibility in focusing on the most relevant features of the text. Thus, the HAN-HFNet increases the model's generalizability and reduces the false negatives by detecting the complex patterns indicative of cognitive distortion.

Table 2. Comparative discussion of HAN-HFNet

Metrics	Explainable depression detection	ML+Explainable AI	SVM	AL-BTCN	CNN	LSTM	Proposed HAN-HFNet
Training sample 90%							
Recall (%)	87.876	88.766	89.876	91.876	92.367	92.888	93.877
Classification error	0.131	0.121	0.111	0.104	0.097	0.092	0.081
Accuracy (%)	86.866	87.877	88.876	89.554	90.258	90.777	91.876
Precision (%)	83.876	85.766	87.766	88.544	89.369	89.887	90.757
F1-score (%)	85.829	87.240	88.808	90.179	90.843	91.363	92.290
K-fold 8							
Recall (%)	88.765	89.877	91.765	92.676	93.468	93.877	**94.756**
Classification error	0.131	0.111	0.102	0.093	0.086	0.082	**0.072**
Accuracy (%)	86.868	88.876	89.766	90.676	91.357	91.766	**92.754**
Precision (%)	83.768	85.977	88.877	89.765	90.368	90.877	**91.866**
F1-score (%)	86.194	87.883	90.298	91.197	91.892	92.353	**93.289**

5. Conclusions

Cognitive distortion refers to irrational or biased ways of thinking that contribute to negative behavior and emotions. This distortion includes personalization, and overgeneralization, which leads individuals to mental health problems, namely stress, depression, anxiety, and so on. Moreover, the conventional approaches used to detect these distortions attained high error rates with less accuracy and precision. Henceforth, a new model termed the HAN-HFNet technique is devised for cognitive distortion detection. Initially, the input sentence is passed to tokenization, where BERT is utilized for providing a contextual understanding of the text. Next, various KPIs, like SCD, FCD, Correlation Between Cognitive Distortions and Depression Severity, CBT, self-reports of cognitive distortions from individuals, LT-MCD, and impact on daily functioning are contemplated. Lastly, cognitive distortion is detected utilizing HAN-HFNet, which is obtained by integrating HDLTex and DHA-Net. Moreover, the HAN-HFNet computed a minimum classification error of 0.072, and maximum recall, accuracy, precision, and F1-score of 94.756%, 92.754%, 91.866%, and 93.289%. However, the model evaluation is conducted on a single dataset, which may not sufficiently represent the full range of linguistic, cultural, and psychological diversity found in real-world populations. Also, the dataset itself may contain inherent biases, which could influence the model's predictions and raise concerns about fairness and inclusivity. Moreover, the proposed method does not specifically detect things like sarcasm or expressions. Future work will aim to fine-tune the DL components of HAN-HFNet to improve its adaptability across different languages, dialects, and individual expression

styles. Additionally, potential biases in the dataset will be explored and assess fairness and ethical considerations to ensure responsible deployment in sensitive mental health applications in the further extension of this work. Furthermore, advanced tools will be implemented to detect sarcasm, support multiple languages and cultural expressions, and handle situations where a person may have more than one mental health issue.

References

[1] Mostafa, M., El Bolock, A. and Abdennadher, S., "Automatic Detection and Classification of Cognitive Distortions in Journaling Text", In WEBIST, pp. 444-452, 2021.

[2] Rnic, K., Dozois, D.J. and Martin, R.A., "Cognitive distortions, humor styles, and depression", Europe's journal of psychology, vol. 12, no. 3, pp. 348, 2016.

[3] Greenberg, P.E., Fournier, A.A., Sisitsky, T., Simes, M., Berman, R., Koenigsberg, S.H. and Kessler, R.C., "The economic burden of adults with major depressive disorder in the United States (2010 and 2018)", Pharmacoeconomics, vol. 39, no. 6, pp.653-665, 2021.

[4] World Health Organization, "Depression and other common mental disorders: global health estimates", 2017.

[5] Mojtabai, R., Olfson, M. and Han, B., "National trends in the prevalence and treatment of depression in adolescents and young adults", Pediatrics, vol. 138, no. 6, 2016.

[6] Keyes, K.M., Gary, D., O'Malley, P.M., Hamilton, A. and Schulenberg, J., "Recent increases in depressive symptoms among US adolescents: trends from 1991 to 2018", Social psychiatry and psychiatric epidemiology, vol. 54, pp.987-996, 2019.

[7] Bollen, J., Ten Thij, M., Breithaupt, F., Barron, A.T., Rutter, L.A., Lorenzo-Luaces, L. and Scheffer, M., "Historical language records reveal a surge of cognitive distortions in recent decades", Proceedings of the National Academy of Sciences, vol. 118, no. 30, pp. e2102061118, 2021.

[8] Fazakas-DeHoog, L.L., Rnic, K. and Dozois, D.J., "A cognitive distortions and deficits model of suicide ideation", Europe's journal of psychology, vol. 13, no. 2, pp. 178, 2017.

[9] Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H., "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media", Big data, vol. 8, no. 3, pp.171-188, 2020.

[10] Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., "Fake news detection on social media: A data mining perspective", ACM SIGKDD explorations newsletter, vol. 19, no. 1, pp.22-36, 2017.

[11] Gil de Zúñiga, H., González-González, P. and Goyanes, M., "Pathways to political persuasion: Linking online, social media, and fake news with political attitude change through political discussion", American Behavioral Scientist, vol. 69, no. 2, pp.240-261, 2025.

[12] Zhao, X., Xing, Z., Kabir, M.A., Sawada, N., Li, J. and Lin, S.W., "Hdskg: Harvesting domain specific knowledge graph from content of webpages", In Proceedings of 24th international conference on software analysis, evolution and reengineering (saner), IEEE, pp. 56-67, February 2017.

[13] Hiraga, M., "Predicting depression for japanese blog text", In Proceedings of ACL 2017, student research workshop, pp. 107-113, July 2017.

[14] Tadesse, M.M., Lin, H., Xu, B. and Yang, L., "Detection of depression-related posts in reddit social media forum", Ieee Access, vol. 7, pp.44883-44893, 2019.

[15] Xing, Z., Zhao, X. and Miao, C., "Identifying cognitive distortion by convolutional neural network-based text classification", 2017.

[16] Orabi, A.H., Buddhitha, P., Orabi, M.H. and Inkpen, D., "Deep learning for depression detection of twitter users", In Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic, pp. 88-97, 2018, June.

[17] Gopendra Vikram Singh, Sai Vardhan Vemulapalli, Mauajama Firdaus, and Asif Ekbal, "Deciphering Cognitive Distortions in Patient-Doctor Mental Health Conversations: A Multimodal LLM-Based Detection and Reasoning Framework," In the proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 22546–22570, 2024.

[18] An, M., Wang, J., Li, S. and Zhou, G., "Multimodal topic-enriched auxiliary learning for depression detection", In Proceedings of the 28th international conference on computational linguistics, pp. 1078-1089, 2020, December.

[19] Benjachairat, P., Senivongse, T., Taephant, N., Puvapaisankit, J., Maturosjamnan, C. and Kultananawat, T., "Classification of suicidal ideation severity from Twitter messages using machine learning", International Journal of Information Management Data Insights, vol. 4, no. 2, pp.100280, 2024.

[20] Wang, B., Zhao, Y., Lu, X. and Qin, B., "Cognitive distortion based explainable depression detection and analysis technologies for the adolescent internet users on social media", Frontiers in Public Health, vol. 10, pp.1045777, 2023.

[21] Lalk, C., Steinbrenner, T., Pena, J.S., Kania, W., Schaffrath, J., Eberhardt, S., Schwartz, B., Lutz, W. and Rubel, J., "Depression Symptoms are Associated with Frequency of Cognitive Distortions in Psychotherapy Transcripts", Cognitive Therapy and Research, pp.1-13, 2024.

[22] Li, F., Jörg, F., Merkx, M.J. and Feenstra, T., "Early symptom change contributes to the outcome prediction of cognitive behavioral therapy for depression patients: A machine learning approach", Journal of Affective Disorders, vol.334, pp.352-357, 2023.

[23] Mirtaheri, S.L., Greco, S. and Shahbazian, R., "A self-attention TCN-based model for suicidal ideation detection from social media posts", Expert Systems with Applications, vol.255, pp.124855, 2024.

[24] Bathina, K.C., Ten Thij, M., Lorenzo-Luaces, L., Rutter, L.A. and Bollen, J., "Individuals with depression express more distorted thinking on social media", Nature human behaviour, vol. 5, no. 4, pp.458-466, 2021.

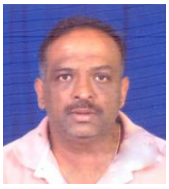
[25] Thekkekara, J.P., Yongchareon, S. and Liesaputra, V., "An attention-based CNN-BiLSTM model for depression detection on social media text", Expert Systems with Applications, vol.249, pp.123834, 2024.

[26] Yang, K., Zhang, T. and Ananiadou, S., "A mental state Knowledge–aware and Contrastive Network for early stress and depression detection on social media", Information Processing & Management, vol. 59, no. 4, pp.102961, 2022.

[27] Grail, Q., Perez, J. and Gaussier, E., "Globalizing BERT-based transformer architectures for long document summarization", In Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume, pp. 1792-1810, April 2021.
[28] Blake, E., Dobson, K.S., Sheptycki, A.R. and Drapeau, M., "The relationship between depression severity and cognitive errors", American Journal of Psychotherapy, vol.70, no.2, pp.203-221, 2016.
[29] Nukhat, A., Salim, A.B.Z. and Shaffie, M.D.F., "COGNITIVE DISTORIONS AND DEPRESSION AMONG OLDER ADULTS: MODERATING ROLE OF RESILIENCE", vol.18, no.2, pp.1-12, 2024.
[30] Waqas, M., Ahmed, A., Maul, T. and Liao, I.Y., "Enhancing breast cancer histopathological image classification using attention-based high order covariance pooling", Neural Computing and Applications, vol.36, no.36, pp.23275-23293, 2024.
[31] Sinha, K., Dong, Y., Cheung, J.C.K. and Ruths, D., "A hierarchical neural attention-based text classifier", In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 817-823, 2018.
[32] Faulty Information Processing Dataset is taken from "https://data.mendeley.com/datasets/npp23hwfnt/1", and accessed on February 2025.
[33] Damsleth, E. and Spjøtvoll, E., "Estimation of trigonometric components in time series", Journal of the American Statistical Association, vol.77, no.378, pp.381-387, 1982.
[34] Vandana Rajput, Amita Jain, and Minni Jain, "An Automatic Approach for Detecting Cognitive Distortion from Spontaneous Thinking," Procedia Computer Science, vol. 260, pp. 768-775, 2025.

Authors' Profiles

Laxmi Jayannavar is currently working as Assistant Professor at the Department of CSE(IOT), Cambridge Institute of Technology, Bangalore, India. She received her M.E in CSE from UVCE, Bangalore, India. She is currently a research candidate at Computer science and Engineering, MSRIT, Bangalore, India. Her research interest secure machine learning. She is a professional member of CSI.



T. N. R. Kumar is currently working as Associate Professor at the Department of CSE, Ramaiah Institute of Technology, Bangalore, India. He has completed his Ph.D from Vels University, Chennai, India. He has been a technology educator. His area of research interest includes image processing, software engineering and computer networks.

Shreekant Jere is currently working as Associate Manager in Accenture AI, Bangalore, India. He has completed his Ph.D from Visvesvaraya Technological University, Belgaum, India. His area of research interest includes Natural Language Processing, Generative AI and Agentic AI.

How to cite this paper: Laxmi Jayannavar, T. N. R. Kumar, Shreekant Jere, "Measuring Cognitive Distortions: A KPI-based Approach to Understanding Faulty Information Processing", International Journal of Information Technology and Computer Science(IJITCS), Vol.18, No.1, pp.140-162, 2026. DOI:10.5815/ijitcs.2026.01.08