Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

CHGAF-Net: A Foundational Framework for Hidden Distress Detection Through Confidence-Aware Multimodal Fusion

Thalla Anusha

Department of Artificial Intelligence, Aurora scientific and technological institute, Ghatkesar, Telangana, India.

Abstract

In recent years, women's safety remains a critical societal concern where individuals experiencing distress often unable to explicitly communicate their emotional state during emergency situations. Existing models advanced in speech emotion recognition and Natural Language Processing (NLP) enabled automates distress analysis. However, these existing approaches predominantly rely on unimodal learning or static multimodal fusion strategies which limited their ability to effectively capture complementary emotional and semantic cues while adapting to varying modality reliability. To address these challenges, a Confidence-Aware Hierarchical Gated Attention Fusion Network (CHGAF-Net) is proposed for hidden distress detection. The speech-derived features such as pitch, energy, spectral characteristics and Mel-Frequency Cepstral Coefficients (MFCCs) are extracted. Consequently, the context embeddings are extracted with Distilled Bi-directional Encoder Representations from Transformer (DistilBERT). Then, these features are fused with a confidence-aware gated attention fusion mechanism. Further, a custom multimodal distress dataset comprising of 48 samples were categorized as safe, warning or danger classes using a fully connected classifier. From the results, the proposed CHGAF-Net attained results in terms of accuracy 70%, precision 74%, recall 70% and F1-score 70% respectively. Qualitative and quantitative analyses further confirm the capability of the proposed framework to capture meaningful distress patterns across multiple modalities.

Keywords: Confidence-Aware Multimodal Fusion, Custom Dataset, Distress Detection, Foundational Framework, Women Safety.

1. Introduction

In recent years, Women's safety has become a vital societal concern due to the increasing prevalence of harassment, abuse and distress-related incidents in both public and private environments [1]. In many situations, the victims may be unable to explicitly communicate their emotional state or seek immediate help from other, therefore, it is needed for the automatic detection of hidden distress as an important research challenge [2][3]. Recent advancements in Artificial Intelligence (AI), Speech Emotion Recognition (SER) and Natural Language Processing (NLP) have enabled intelligent systems to analyze human emotions via speech and text modalities [4][5][6]. The existing studies have primarily focused on

speech-based emotion recognition, sentiment analysis and contextual emotional monitoring using either acoustic or textual information [7]. Unfortunately, it is not uncommon for disruptive situations such as verbal altercations or physical assaults to occur on public or public environments [8]. These events create a discomfort and more significantly poses a safety risk. Therefore, it is crucial to develop the methods for detecting and preventing them. This problem with a strong social impact has not yet attracted the attention it deserves [9]. Moreover, the existing authors who have addressed it so far mostly framed it as the detection of scream or the shouted speech. Even though, the existing models attained promising performance, most rely on single modality and failed to capture the complementary emotions and semantic information present across different data sources [10]. Furthermore, the conventional multimodal systems often employed simple feature concatenation or static fusion strategies which treats all modalities equally and overlooks their varying reliability under different circumstances [11]. Such limitations reduce the effectiveness of detecting subtle and hidden distress cues, particularly in women safety applications where emotional expressions may be implicit, context-dependent and highly dynamic [12]. Motivated by these challenges, this work explores the integration of speech-derived emotional characteristics and contextual textual representations to improve distress detection. By jointly analyzing the acoustic patterns and the sematic content, the proposed framework aims to provide a more comprehensive understanding of distress-related behavior and contribute toward the development of intelligent women safety monitoring systems. The key novelty of this work lies in the proposed Confidence-Aware Hierarchical Gated Attention Fusion Network (CHGAF-Net) which introduces modality-specific attention learning and confidence-aware gating mechanism to dynamically evaluate the reliability of audio and textual information before the fusion. Unlike the conventional fusion methods, the proposed CHGAF-Net adaptively prioritizes the most informative modality by enabling the robust detection of hidden distress cues in women safety contexts. Moreover, the detailed explanation of existing methods for the related works are explained as follows: Eleonora Mancini et al. [13] proposed a disruptive situation detection framework using CNN and SVM by formulating public safety monitoring as a Speech Emotion Recognition (SER) task rather than relying solely on CCTV-based surveillance. The author investigates multiple acoustic features, ML classifiers, data augmentation techniques to achieve the speaker-independent and noise-robust detection. Even though, the study examined only the audio modality and overlooked semantic distress cues present in the textual content. This gap limits their ability to capture hidden or context-dependent distress patterns. Diego Resende Faria et al. [14] proposed a multimodal affective communication framework which combines the speech emotion with sentiment analysis to improve the emotional understanding. The author used acoustic features extracted from the speech and sentiment features derived from speech-to-text conversion. Furthermore, the author proposed a Two Layer-Dynamic Bayesian Mixture Model (2L-DBMM) for multimodal fusion and second-level emotion classification. Nevertheless, the fusion strategy primarily focused on emotional state classification and did not incorporate the adaptive confidence-based modality weighting. This limitation potentially limited its effectiveness in detecting subtle and hidden distress cues. Vithya Ganesan et al. [15] demonstrated a Contextual Emotional Classifier (CEC) framework on interconnected edge devices such as smartphones, smartwatches and voice assistants to monitor and analyze women's emotional state. However, the model primarily focused on emotional wellness and contextual mood analysis rather than detecting hidden distress and risk-related situations. Jessica Ison et al. [16] investigated the experiences of women and gender-diverse individuals facing sexual violence and harassment in public transportation. The study reveals that engage in extensive safety work including behavioral adaptation, route planning and

situational awareness to reduce potential risks. Even though, the study provides valuable insights into the challenges faced by vulnerable commuters, it is primarily qualitative and lacks automated mechanism for identifying the distress or potential risk situations. Chaitanya Singla et al. [17] proposed a speech emotion recognition system with Convolutional Neural Networks (CNNs) for Punjabi language speakers. The CNN model was trained to learn discriminative emotional patterns directly from spectrogram images for enabling the effective emotion classification. Although, the framework demonstrated the capability in recognizing speech emotions from regional language datasets, it is constrained with relying on acoustic information and does not incorporate semantic contextual cues from textual data which limits its ability to identify subtle or hidden distress expressions.

The contributions of this research are listed as follows:

- The proposed CHGAF-Net dynamically integrates the acoustic and semantic representations while enabling the adaptive modality weighting for improved hidden distress detection in women safety applications.
- By fusing speech-derived emotional features with Distilled Bi-directional Encoder Representations from Transformer (DistilBERT)-based contextual embeddings facilitates comprehensive analysis of both vocal and semantic distress indicators.
- A custom multimodal distress dataset comprising of safe, warning and danger categories is constructed and evaluated to demonstrate the feasibility of confidence-aware fusion mechanism for detecting the hidden distress under limited-data conditions.

The organization of this research is structured as follows: Section 2 represents the methodology along with architecture, section 3 specifies the experimental results with corresponding confusion matrix, section 4 demonstrates the discussion and section 5 illustrates the conclusion.

2. Methodology

In this section, a multimodal learning approach namely CHGAF-Net is introduced for hidden distress detection by integrating audio and textual information. Initially, speech recordings are preprocessed and transformed into acoustic representations using MFCCs, pitch, RMS energy, spectral centroid, spectral bandwidth and Zero-Crossing Rate (ZCR) features. Simultaneously, corresponding transcripts are tokenized and encoded using DistilBERT to obtain contextual semantic embeddings. Consequently, the extracted audio and text features are then processed through modality-specific attention mechanisms and confidence gates. Finally, the proposed CHGAF-Net adaptively fuses both modalities and a fully connected classifier categorizes each sample into safe, danger or warning classes. The architecture of proposed CHGAF-Net is presented in below Figure 1.

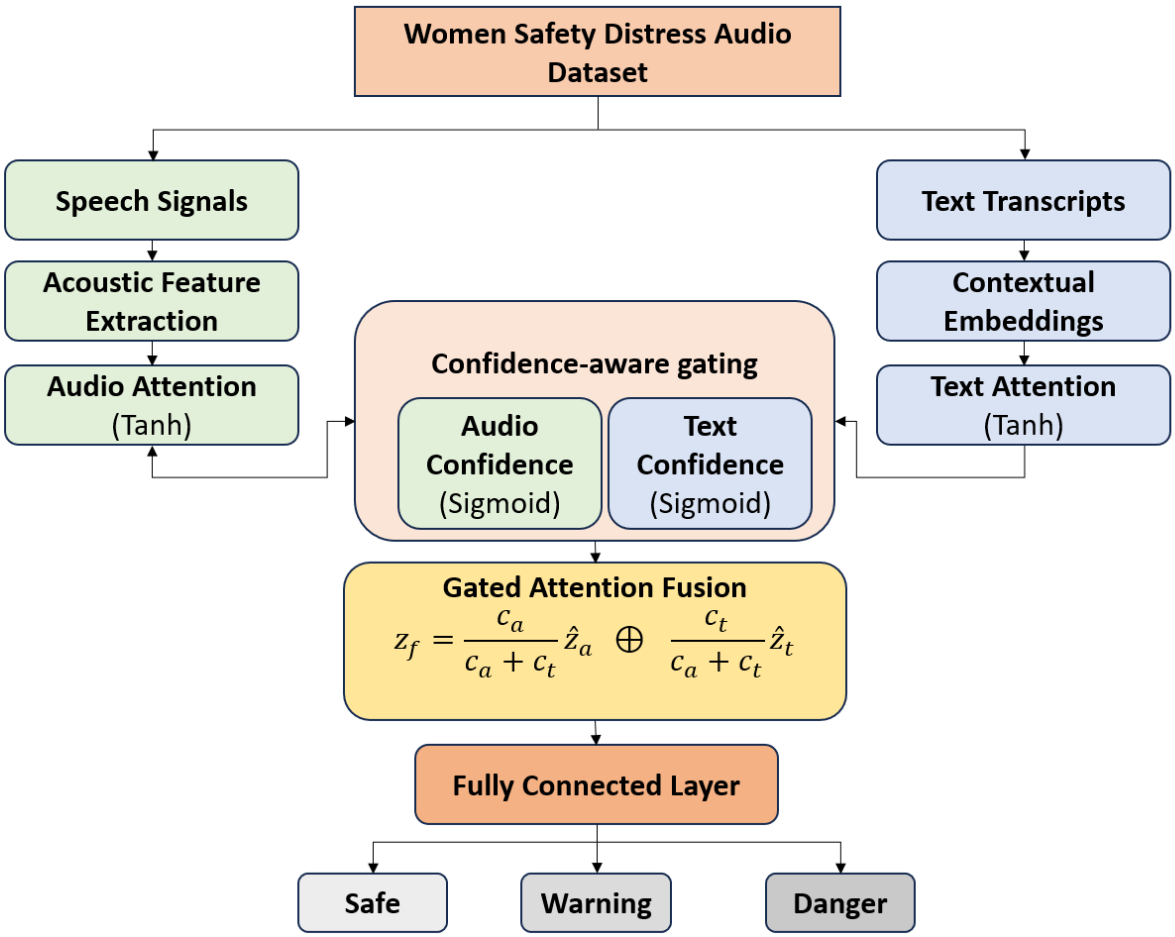

Figure 1. The architecture of proposed CHGAF-Net for hidden distress detection

2.1 Dataset Description

In this study, a custom multimodal dataset was developed for capturing hidden distress cues in women safety scenarios through both speech and textual modalities [18]. The dataset consists of 48 samples which are categorized into three classes including safe, warning and danger. Each samples contains an audio recording and its corresponding transcript for enabling the joint analysis of emotional speech patterns and semantic distress indicators. Due to the limited availability of authentic distress-related recordings and the exploratory nature of this research, a curated dataset of 48 samples was utilized. The dataset was divided into 80% training data and 20% testing data using stratified sampling to preserve the class distribution across both subsets.

2.2 Data Preprocessing

A preprocessing was applied to collected dataset to ensure the consistency across both modalities. For the audio modality, all recordings were standardized by converting them into WAV format with a fixed sampling rate of 16 kHz. Further, the noise and format inconsistencies were minimized to facilitate reliable feature extraction. On the other hand, for the textual modality, the corresponding transcripts were cleaned by removing unnecessary symbols, special characters and redundant spaces [18]. Subsequently, the textual data were tokenized using the DistilBERT tokenizer to generate contextual representations which are suitable for hidden distress analysis. Finally, the class labels representing safe, warning and danger

categories were encoded into numerical labels and the dataset was partitioned into training and testing subsets using stratified sampling to maintain balanced class distributions.

2.3 Feature Extraction

Followed by preprocessing, the feature extraction is introduced to extract the reliable features from both text and audio modalities. The audio modality was analyzed to capture acoustic and emotional speech characteristics associated with hidden distress. Here, each standardized speech signal was processed to extract Mel-Frequency Cepstral Coefficients (MFCCs), pitch, root mean square (RMS) energy, spectral centroid, spectral bandwidth, and zero-crossing rate. All these features provide a complementary information regarding the speech articulation and emotional intensity with vocal stability and frequency distribution. Particularly, the MFCCs represents the spectral characteristics of speech based on human auditory perception and the mel-scale transformation is mathematically defined in below Equation (1)

$$\text{Mel}(f) = 2595\log_{10}\left(1+\frac{f}{700}\right) \quad (1)$$

Where, f denotes the frequency in hertz and the average MFCC feature is computed using Equation (2)

$$\text{MFCC}_i = \frac{1}{N}\sum_{n=1}^{N} c_i\,(n) \quad (2)$$

Here, $c_i(n)$ denotes the i^{th}cepstral coefficient at frame n and N denotes the total number of frames consequently, the pitch extraction provides the valuable cues corresponding to stress, fear and emotional stability. Therefore, the mean pitch value is calculated using below Equation (3)

$$\text{Pitch}_{\text{mean}} = \frac{1}{M}\sum_{k=1}^{M} p_k \quad (3)$$

Here, p_k refers the detected pitch value and M represents the number of valid pitch observations. Furthermore, the RMSE is applied to reflect the vocal intensity and emotional activation levels. The RMSE energy is calculated by using blow Equation (4)

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \quad (4)$$

Where x_n specifies the speech amplitude and N demonstrates the total number of samples. Consequently, the spectral features contain the spectral centroid which indicates the center of mass for the frequency spectrum and it is computed with Equation (5)

$$SC = \frac{\sum_{k=1}^{K} f_k X(k)}{\sum_{k=1}^{K} X(k)} \quad (5)$$

Here, f_k denotes the frequency bin and X(k) denotes the spectral magnitude. In addition, the spectral bandwidth is explained in Equation (6)

$$SB = \sqrt{\frac{\sum_{k=1}^{K} (f_k - SC)^2 X(k)}{\sum_{k=1}^{K} X(k)}} \quad (6)$$

Here, this spectral bandwidth measures the spread of frequencies around the spectral centroid. Particularly, higher bandwidth often indicates stressed, excited or emotionally intense speech patterns that makes it valuable for distress detection. Another important mechanism is Zero-Crossing Rate (ZCR) which measures the variability and it is computed using Equation (7)

$$ZCR = \frac{1}{N-1} \sum_{n=1}^{N-1} I(x_n x_{n-1} < 0) \quad (7)$$

The final sample with all these features is denoted as z_{audio}, where $I(\cdot)$ is the indicator function and this mechanism mainly measures how frequently an audio signal changes from positive to negative amplitude. The higher ZCR indicates increased speech activity or the emotion intensity by making it useful for distinguishing the calm speech from distress-related vocal patterns.

3.3.1. Text feature extraction with DistilBERT

The DistilBERT model is employed as a lightweight transformer-based language model to capture the hidden distress cues embedded within the textual conversations. The DistilBERT model preserves most of the BERTs' language understanding capabilities while reducing the computational complexity by making it suitable for multimodal distress analysis. For instance, the input transcript is considered as $T = \{w_1, w_2, \dots, w_n\}$ where w_i denotes the i^{th} word, the DistilBERT tokenizer converts the sentence into sub-word tokens which is presented in Equation (8)

$$X = \{t_1, t_2, \dots, t_m\} \quad (8)$$

Here, m denotes the token sequence length and each token is mapped into a dense semantic embedding space as presented in Equation (9)

$$E = \{e_1, e_2, \dots, e_m\} \quad (9)$$

Here, $e_i \in \mathbb{R}^{768}$ and 768 demonstrates the hidden embedding dimension of DistilBERT model. Then, the self-attention mechanism is introduced to compute attention scores for the distress cues. For each token embedding query Q, key K and value V metrices are generated as $Q = XW_Q$, $K = XW_K$ and $V = XW_V$. Where W_Q, W_K, and $W_V$denote learnable projection matrices and the scaled dot-product attention mechanism is expressed as Equation (10)

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

where $d_k$denotes the dimensionality of the key vectors. To attain the contextual features, the DistilBERT model employs an encoder for generating contextualized representations that is mathematically expressed in Equation (11)

$$H = \{h_1, h_2, \dots, h_m\} \quad (11)$$

where each hidden representation captures the semantic dependencies among all the tokens and the final text feature vector is obtained from the contextual embedding of the first token that is expressed through Equation (12)

$$z_{text} = h_{CLS} \quad (12)$$

Where z_{text} denotes the semantic feature vector used for multimodal fusion and this representation

captures contextual meaning, emotional intent, urgency indicators and latent distress patterns present in textual conversations.

2.4 Fusion Model

As the features from both the modalities are extracted separately, the proposed a Confidence-Aware Hierarchical Gated Attention Fusion (CHGAF) is introduces to effectively integrate the information. Unlike conventional concatenation-based fusion strategies, the proposed CHGAF model dynamically estimates the importance of each modality through attention-guided feature refinement and confidence-aware gating. It enables the model to highlight the most informative modality for a given sample while suppressing noisy or the irrelevant representations. The $z_{audio} \in \mathbb{R}^{d_a}$ and $z_{text} \in \mathbb{R}^{d_t}$ denotes the extracted audio and text feature vectors. The modality-specific attention-enhanced representations are computed using Equation (13).

$$\hat{z}_a = z_a \odot \text{Softmax}(W_a z_a + b_a), \quad \hat{z}_t = z_t \odot \text{Softmax}(W_t z_t + b_t) \tag{13}$$

where $\hat{z}_a$and $\hat{z}_t$ represent the attended audio and text features, $W_a$and W_t denote trainable projection matrices, b_a and b_t specifies the bias vectors and $\odot$ represents element-wise multiplication. Moreover, to estimate the reliability of each modality, confidence scores are generated using below Equation (14)

$$c_a = \sigma(w_a^T \hat{z}_a), c_t = \sigma(w_t^T \hat{z}_t) \tag{14}$$

Where c_a and c_t specifies the confidence values of the audio and text modalities, w_a and w_t defined as the learnable weight vectors. In addition, $\sigma(\cdot)$ represents the sigmoid activation function. The confidence values are normalized to obtain adaptive modality weights and subsequently used to construct the fused representation. The fused representation construction is mathematically explained in Equation (15)

$$z_f = \frac{c_a}{c_a + c_t} \hat{z}_a \ \oplus \ \frac{c_t}{c_a + c_t} \hat{z}_t \tag{15}$$

Here, z_f denotes the final multimodal feature representation and $\oplus$ indicates vector concatenation. The fused representation is subsequently forwarded to the fully connected layer for categorizing the input sample into Safe, Warning, or Danger classes. The fully connected layer transforms the fused multimodal feature vector into class-specific scores through learned weights and biases. In the proposed CHGAF-Net, the confidence weighted audio and text features are first fused and fed into the hidden fully connected layer where the discriminative patterns associated with classes are learned. Finally, the FC layer generates class logits which are converted into probability distributions using the SoftMax function and it is given in below Equation (16)

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{16}$$

where z_i represents the output score for class i and C denotes the total number of classes. The class with the highest probability is selected as the final prediction.

3. Experimental Results

The proposed CHGAF-Net model was evaluated using Python with PyTorch, DistilBERT and Librosa libraries. The software and hardware configurations including windows 11 with intel core i5 processor and 16 GB RAM were considered. The experiments were conducted on a custom multimodal dataset containing 48 samples, split it into 80% training and 20% testing sets. Moreover, the evaluation metrics such as accuracy, precision, recall and F1-score are considered and the mathematical representations are explained in Equation (17) – (20):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

$$Recall = \frac{TP}{TP+FN} \quad (19)$$

$$F1-score = \frac{2\times precision\times recall}{precision+recall} \quad (20)$$

Here, TP denotes the true positive, TN specifies true negative, FP specifies false positive and FN denotes the false negative. Moreover, the hyperparameter setting for the proposed CHGAF-Net model is presented in below Table 1.

Table 1. Hyperparameter setting for proposed CHGAF-Net model

Hyperparameter	**Value**
Audio Feature Dimension	18
Text Embedding Model	DistilBERT Base Uncased
Text Embedding Dimension	768
Maximum Sequence Length	64
Hidden Layer Dimension	128
Number of Classes	3 (Safe, Warning, Danger)
Attention Activation	Tanh
Confidence Gate Activation	Sigmoid
Classifier Activation	ReLU
Dropout Rate	0.3
Loss Function	Cross-Entropy Loss
Optimizer	Adam
Learning Rate	0.001
Batch Size	Full Batch Training
Number of Epochs	10
Train-Test Split	80:20
Random State	42

3.1 Qualitative Analysis

The proposed CHGAF-Net model is further evaluated on a representative test sample and the objective was to examine the model's ability to distinguish among safe, warning and danger classes beyond

performance metrics. Therefore Table 2 presents the subset of actual and predicted labels generated by the model.

Sample	Actual Label	Predicted Label	Outcome
S1	Danger	Danger	Correct
S2	Warning	Warning	Correct
S3	Safe	Safe	Correct
S4	Danger	Danger	Correct
S5	Danger	Danger	Correct
S6	Warning	Warning	Correct
S7	Safe	Safe	Correct
S8	Safe	Danger	Misclassified
S9	Warning	Danger	Misclassified

From Table 2, the qualitative results indicate that the proposed CHGAF-Net model effectively captures explicit distress patterns by achieving correct classification for multiple Safe, Warning, and Danger classes. Most misclassifications occurred between the Warning and Danger categories where it suggests that the presence of overlapping emotional and semantic characteristics within the dataset. Moreover, to validate the cross-attention mechanism, the evaluation was carried out and it is explained in below Table 3.

Sample	Audio Confidence	Text Confidence	Dominant Modality
S1	0.72	0.28	Audio
S2	0.34	0.66	Text
S3	0.51	0.49	Balanced

Table 3 shows that the confidence-aware gating mechanism dynamically adjusted modality contributions according to input characteristics. For the emotionally expressive speech and audio received higher weights whereas semantically explicit distress statements were predominantly influenced by textual representations.

3.2 Quantitative Analysis

Further, the quantitative results of proposed CHGAF-Net model are explained in below Table 4.

Metric	Value (%)	Interpretation
Accuracy	70.00	7 out of 10 test samples classified correctly
Precision	74.00	Predicted classes are relatively reliable
Recall	70.00	Majority of actual distress categories identified

F1-Score	70.67	Good balance between precision and recall

From Table 4, the proposed CHGAF-Net model achieved an overall classification accuracy of 70.00%, with precision, recall, and F1-score values of 74.00%, 70.00%, and 70.67%, respectively. The obtained results demonstrate the effectiveness of integrating acoustic and semantic representations through confidence-aware gated attention fusion. In particular, the precision score of 74.00% indicates that the model generates reliable class predictions while the balanced recall and F1-score values demonstrates its ability to identify distress-related patterns across all categories. Moreover, the F1-score of 70.67% and accuracy of 70.00% demonstrates the effectiveness of confidence-aware multimodal fusion for hidden distress detection under limited-data conditions. Furthermore, the confusion matrix for the proposed CHGAF-Net is presented in Figure 2.

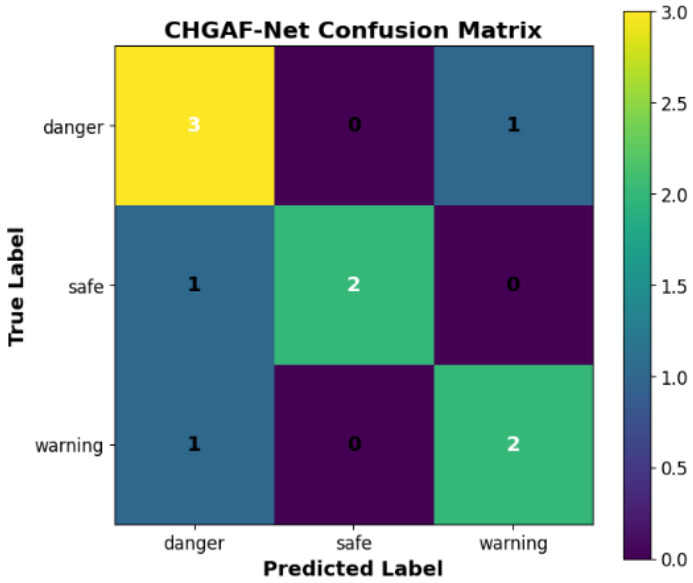

Figure 2. Confusion matrix of proposed CHGAF-Net for distress detection

From Figure 2, the confusion matrix specifies that the model correctly classified three out of four Danger samples, two out of three Safe samples, and two out of three Warning samples. The results indicate that the proposed confidence-aware multimodal fusion strategy effectively distinguishes high-risk distress situations from non-distress conditions. Notably, no direct confusion was observed between Danger and Safe classes, demonstrating the model's capability to learn meaningful distress representations.

4. Discussion

In this section, the effectiveness of experimental findings for distress detection using proposed CHGAF-Net in women safety applications are discussed. The proposed CHGAF-Net effectively captures complementary information from both speech and text modalities by dynamically integrating acoustic and semantic representations which resulted in improved distress understanding compared to unimodal approaches. The confidence-aware gating mechanism adaptively adjusts the modality contributions to the reliability of the input signals while enabling more robust feature fusion and reduced the influence of less

informative representations. Consequently, the integration of contextual embeddings and emotional features facilitates the comprehensive analysis of both vocal expressions and semantic content that allows the model to identify distress cues. The obtained quantitative results including accuracy of 70% validates the feasibility of proposed multimodal framework. Additionally, the custom dataset comprising of safe, warning, and danger categories provides a proof-of-concept environment for evaluating the confidence-aware fusion mechanisms in hidden distress detection tasks. Despite these outcomes, the study has several limitations including the dataset contains only 48 samples which restricts the generalization capability of the model. In addition, it may not adequately represent the real-world distress variations.

5. Conclusion

This study exploited a CHGAF-Net model for hidden distress detection in women safety applications. Existing models predominantly rely on unimodal analysis or static multimodal fusion strategies often struggled to effectively capture the complementary emotional and semantic information. This limitation restricted their adaptively and failed to account for varying modality reliability. Therefore, the proposed CHGAF-Net model dynamically integrates acoustic features extracted from speech signals and contextual semantic embeddings generated using DistilBERT through a confidence-aware gated attention fusion mechanism. The adaptive weighting strategy enables the model to prioritize the most informative modality for each input instance by improving the hidden distress cues representation. The experimental evaluation was conducted on a custom multimodal dataset consisting of safe, warning and danger categories demonstrated the feasibility of the proposed CHGAF-Net model. The results attained in terms of accuracy of 70%, precision of 74%, recall of 70%, and F1-score of 70% while confirming its ability to distinguish distress-related patterns across multiple categories.

References

1. Yang, X., Zhang, J., Liu, W., Jing, J., Zheng, H. and Xu, W., 2024. Automation in road distress detection, diagnosis and treatment. Journal of Road Engineering, 4(1), pp.1-26.
2. Yi, C., Liu, J., Huang, T., Xiao, H. and Guan, H., 2023. An efficient method of pavement distress detection based on improved YOLOv7. Measurement Science and Technology, 34(11), p.115402.
3. Li, Y., Liu, C., Yue, G., Gao, Q. and Du, Y., 2022. Deep learning-based pavement subsurface distress detection via ground penetrating radar data. Automation in Construction, 142, p.104516.
4. Ibragimov, E., Lee, H.J., Lee, J.J. and Kim, N., 2022. Automated pavement distress detection using region based convolutional neural networks. International Journal of Pavement Engineering, 23(6), pp.1981-1992.
5. Zhang, C., Nateghinia, E., Miranda-Moreno, L.F. and Sun, L., 2022. Pavement distress detection using convolutional neural network (CNN): A case study in Montreal, Canada. International Journal of Transportation Science and Technology, 11(2), pp.298-309.
6. Hu, Y., Chen, N., Hou, Y., Lin, X., Jing, B. and Liu, P., 2025. Lightweight deep learning for real-time road distress detection on mobile devices. Nature Communications, 16(1), p.4212.

7. Liu, Y., Liu, F., Huang, Y., Hu, J., Zhang, W. and Hou, Y., 2025. The real-time pavement distress detection system based on edge-cloud collaborative computing. IEEE Transactions on Intelligent Transportation Systems.
8. Cicmanec, L., Kubica, A. and Maslan, J., 2025. LSTM and TCN application for airport surface distress detection. Results in Engineering, 27, p.105708.
9. Wang, J., Li, X., Xu, Y., Zhou, Z., Wu, W. and Li, Z., 2025. Optimizing CNN for pavement distress detection via edge-enhanced multi-scale feature fusion. PloS one, 20(4), p.e0319299.
10. Lv, Z., Hao, Z., Zhu, Y. and Lu, C., 2025. A review on automated detection and identification algorithms for highway pavement distress. Applied Sciences, 15(11), p.6112.
11. Taiwo, O. and Al-Bander, B., 2025. Emotion-aware psychological first aid: Integrating BERT-based emotional distress detection with Psychological First Aid-Generative Pre-Trained Transformer chatbot for mental health support. Cognitive Computation and Systems, 7(1), p.e12116.
12. Chauhan, A.S., Fatima, S., Sharma, S., Srivastava, A.K., Srivastava, A. and Kumar, M., 2025, November. Multimodal Approaches for Automated Distress Detection. In 2025 5th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU) (pp. 1-6). IEEE.
13. Mancini, E., Galassi, A., Ruggeri, F. and Torroni, P., 2024. Disruptive situation detection on public transport through speech emotion recognition. Intelligent Systems with Applications, 21, p.200305.
14. Resende Faria, D., Weinberg, A.I. and Ayrosa, P.P., 2024. Multimodal affective communication analysis: Fusing speech emotion and text sentiment using machine learning. Applied Sciences, 14(15), p.6631.
15. Ganesan, V., Ramasamy, V., Manoj, C. and Tejaswi, T., 2023. Contextual emotional classifier: An advanced AI-powered emotional health ecosystem for women utilizing edge devices. Traitement du Signal, 40(6), p.2481.
16. Ison, J., Forsdike, K., Henry, N., Hooker, L. and Taft, A., 2025. “I’ll try and make myself as small as possible”: women and gender-diverse people's safety work on public transport. Violence against women, 31(11), pp.2830-2852.
17. Singla, C., Singh, S., Sharma, P., Mittal, N. and Gared, F., 2024. Emotion recognition for human–computer interaction using high-level descriptors. Scientific reports, 14(1), p.12122.
18. Women Safety Distress Audio Dataset: https://www.kaggle.com/datasets/thallaanusha/women-safety-distress-audio-dataset (Accessed on June 2026).