International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 4, July – August 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Privacy-Preserving AI Chatbot for Real-Time Stress Detection and Personalized Mental Health Support

Prajakta N. Sardar*, Rashmi Tale**
*(Department of Computer Science and Engineering, College of Engineering and Technology, Akola, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India
Email: prajaktasardar41@gmail.com)
** (Department of Computer Science and Engineering, College of Engineering and Technology, Akola, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India
Email: hodcompsci@sgbau.ac.in)

----------**********----------

Abstract:

Mental health data is among the most sensitive information a user can share, yet most deployed chatbots route user messages through external cloud APIs with no guarantee of confidentiality. The Gentle Sanctuary addresses this directly: all stress classification — keyword matching and transformer inference alike — runs locally on the application server, so no user message ever leaves the system boundary. Catching user stress in real time is one of the most safety-critical jobs a mental wellness chatbot has to do, yet most deployed systems still lean on basic keyword matching — and nobody has properly measured how badly that approach breaks down when users do not type the "right" words. This paper describes how we built, tested, and evaluated a hybrid keyword–transformer stress detection system inside The Gentle Sanctuary, a privacy-preserving FastAPI mental wellness chatbot. The system pairs a fast, rule-based 16-term keyword checker that responds in under a millisecond with a DistilBERT model fine-tuned on the DAIR-AI/Emotion dataset, running in the background so it never slows down the chat. We tested all three classifiers — keyword-only, DistilBERT-only, and the hybrid — on a 120-message dataset that covered six different ways people express distress: direct keyword use, compound stress phrases, implied distress, idioms, code-switched Hindi/English, and cultural metaphors. Two independent raters labelled every message (Cohen's kappa = 0.91). The keyword approach got 70.0% accuracy (binary F1 = 0.714) and never raised a false alarm, but it completely missed implicit, idiomatic, and code-switched messages — 0% detection on those. DistilBERT pushed accuracy to 90.83% (binary F1 = 0.929), catching 66.7% of implied distress, 100% of idioms, and 85.7% of cultural metaphors. The hybrid came out on top at 91.67% accuracy (binary F1 = 0.936) with only 17.2 ms latency per message — a 21.67 percentage point jump over keyword-only, while keeping the immediate-response speed guarantee of the keyword layer. Beyond detection accuracy, the system delivers personalized mental health support through a per-user rolling stress index and an adaptive crisis escalation pathway that surfaces tailored resources only when an individual user's session signals genuine distress. The numbers make a clear case: combining both layers closes the gap that keyword matching leaves wide open, without hurting the real-time performance or the privacy guarantee that production mental health chatbots need.

Keywords — **stress detection, mental health chatbot, hybrid architecture, transformer, DistilBERT; NLP, crisis escalation, privacy-preserving, personalized support, conversational AI, sentiment analysis, keyword classification.**

----------**********----------

I. INTRODUCTION

Mental health chatbots have become a real option for people who cannot easily reach a therapist, especially in areas where mental health professionals are in short supply [1]. The World Health Organization makes the point clearly: digital wellness tools need to be able to spot and respond to serious distress, supporting professional care rather than replacing it [1]. Yet adoption faces a critical barrier: mental health conversations contain deeply personal and sensitive information, and users must trust that their disclosures are handled with care. Most commercial chatbot platforms route user messages through external cloud inference APIs, creating privacy risks that are particularly acute for sensitive mental health data. A privacy-preserving architecture — one in which all classification runs locally, with no user text leaving the system boundary — is not just a technical preference but a clinical and ethical necessity. NLP research has come a long way — fine-tuned BERT models now clear 95% accuracy on mental health sentiment tasks covering anxiety, depression, and stress [2]. Systems like AI-DASA have hit 94% accuracy detecting emotional disorders from student text [3].

A critical safety requirement for any mental health conversational AI is the ability to detect escalating user distress in real time and trigger appropriate crisis escalation interventions. Existing approaches span a broad spectrum. At one end, simple keyword matching provides zero computational overhead and fully deterministic, auditable behaviour [4]. At the other end, transformer-based semantic classifiers—including BERT, DistilBERT, and RoBERTa—achieve superior accuracy on emotionally nuanced text but impose latency penalties that may degrade the conversational user experience [5, 6]. For instance, RoBERTa has been shown to achieve 96.30% accuracy and an F1-score of 98.11% on sentiment classification tasks [5], whereas fine-tuned BERT models reach 91.77% accuracy on domain-specific text classification [7]. However, these models typically require 20–300 ms per inference, compared to sub-millisecond execution for lexicon-based approaches [6].

The literature on emotion and stress detection in text has progressed through three methodological generations. Lexicon-based methods offer high interpretability but fail on negated expressions, figurative language, and non-English inputs [4, 8]. Classical machine learning methods employing TF-IDF and N-gram feature extraction with models such as LinearSVC and Logistic Regression achieve 70–89% accuracy on benchmark emotion datasets [4]. Transformer-based methods represent the current state of the art, with DistilBERT fine-tuned on the DAIR-AI/Emotion dataset achieving 92.7% accuracy on the emotion classification benchmark [16], and large language models demonstrating strong few-shot and zero-shot generalisation across diverse NLP classification tasks [10]. Recent work has also explored hybrid architectures that combine multiple model families; for example, GMM-enhanced N-gram LSTM networks have been proposed to capture fine-grained sentiment patterns [8], and ensemble frameworks combining transformers with stacked classifiers have been developed for code-mixed multilingual text [11].

The integration of emotional intelligence into conversational AI has received growing attention. Research on sentiment and emotion modelling in text-based conversations demonstrates that contemporary language models exhibit a strong capacity for understanding contextual emotional cues [12]. In the specific domain of counselling chatbots, explainable emotion recognition frameworks using Valence–Arousal–Dominance (VAD) calibrated transformer models have improved emotion classification stability, achieving a 52% reduction in prediction flip rate and an Expected Calibration Error of 0.03 [13]. Early deployed systems such as Woebot [14] rely primarily on rule-based and keyword-driven responses for crisis detection—an approach that, while deterministic, is acknowledged to be insufficient for capturing indirect, figurative, or culturally-specific expressions of distress [15]

Research on cultural and linguistic variation in distress expression highlights several failure modes

for English-only keyword approaches: idiomatic expressions ("I'm at my limit"), code-switching by multilingual users, culture-specific metaphors, and indirect expressions such as fatigue complaints [8, 13]. These failure modes carry direct patient safety implications, as false negatives in crisis detection mean that high-stress users receive no escalation intervention. Despite this well-documented gap, no prior study has simultaneously implemented, measured, and compared keyword, transformer, and hybrid classifiers in a production mental health chatbot across linguistically stratified inputs—leaving practitioners without empirical guidance on which architecture best balances latency, accuracy, and clinical safety.

This study addresses three research questions. RQ1: What is the accuracy of keyword-based, DistilBERT-based, and hybrid stress classifiers on a linguistically stratified 120-message test set? RQ2: Which input categories reveal the largest gap between keyword and transformer classifiers, and by how much does the hybrid architecture recover that gap? RQ3: What is the latency–accuracy trade-off of each classifier, and what architecture best satisfies the dual constraints of real-time response and semantic depth in a production chatbot?

The contributions of this work are fourfold: (1) the first empirical comparison of keyword, DistilBERT, and hybrid classifiers on an annotated, linguistically stratified stress detection test set within a production mental health chatbot; (2) a quantified failure mode analysis showing that keyword classifiers achieve 0% detection on implicit, idiomatic, and code-switched distress inputs; (3) a novel synchronous-asynchronous hybrid architecture that achieves 91.67% accuracy and binary F1 = 0.936 at 17.2 ms latency while preserving a public API contract for zero-disruption deployment; and (4) a reproducible experimental framework—including test set, evaluation code, and pre-trained model reference—enabling direct replication and extension.

II. SYSTEM DESCRIPTION

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A. The Gentle Sanctuary — System Context

The stress detection system is deployed as the real-time crisis detection engine within The Gentle Sanctuary, a FastAPI-based mental wellness chatbot backend that additionally integrates Google Gemini as its primary large language model with a Retrieval-Augmented Generation (RAG) layer for verified AI response generation; the present paper focuses specifically on the stress detection and privacy-preserving architecture subsystems. The stress classifier is invoked on every incoming user message within chat_service.send_message(), and the result determines: (a) the sentiment markers attached to the AI response, (b) the stress level stored in the chat_messages.stress_level database field, and (c) whether crisis escalation resources are surfaced to the user.

B. Keyword Classifier

The keyword classifier operates on a curated 16-term English distress lexicon: stress, anxious, anxiety, overwhelmed, panic, worried, tired, exhausted, burnout, depressed, hopeless, crying, pressure, scared, fearful, and tense. Binary classification is performed through case-insensitive substring matching: if any keyword appears in the input text, the message is classified as high_stress; otherwise, it is classified as normal. The classifier is stateless, deterministic, and operates in O(k) time complexity where k = 16, achieving sub-millisecond latency on any hardware.

C. DistilBERT Classifier

The semantic classifier employs the bhadresh-savani/distilbert-base-uncased-emotion model [16], a DistilBERT-base architecture [6] fine-tuned on the DAIR-AI/Emotion dataset comprising 16,000 English text samples across six emotion categories: sadness, joy, love, anger, fear, and surprise. For binary stress classification, the emotion output labels are mapped to the target schema: sadness, anger, and fear are mapped to high_stress; joy, love, and surprise are mapped to normal. This mapping reflects established dimensional models of affect, in which negative high-arousal states such as fear and

anger and negative low-arousal states such as sadness constitute the primary affective components of psychological stress [13]. The model is loaded as a HuggingFace text-classification pipeline with maximum sequence length of 128 tokens.

D. Hybrid Architecture

The hybrid classifier implements a synchronous-asynchronous two-layer decision architecture. In the synchronous layer, the keyword classifier runs immediately upon message receipt, returning a binary classification within sub-millisecond latency for immediate response generation and crisis escalation triggering. In the asynchronous layer, the DistilBERT classifier is invoked and its result is used to update the rolling stress index and trend analysis. The hybrid decision logic applies an OR gate: a message is classified as high_stress if either the keyword layer or the DistilBERT layer returns high_stress. This design maximises recall—the safety-critical metric in crisis detection, where false negatives (missed distress signals) carry greater clinical risk than false positives [1, 13]. The public API of the system (accepting a string input and returning a binary classification label) is identical for all three classifiers, enabling seamless comparison and transparent drop-in deployment.

E. Stress Index and Live Analysis

The get_stress_index() function maps the binary classification to a numeric score via uniform random sampling within class-specific ranges: high_stress maps to [60, 90] and normal maps to [20, 45]. The get_live_analysis() function computes an aggregate stress assessment across the most recent N messages, calculates rolling averages, and detects escalating versus de-escalating trends using a ±5-point threshold on the difference between the first-half and second-half message averages.

F. Crisis Escalation Pathway

When the stress classifier returns high_stress, the system automatically routes the user to crisis resources via the support service, surfacing the National Crisis Hotline (988 Suicide and Crisis Lifeline), Mindfulness-Based Stress Reduction (MBSR) programme information, and CBT Thought Record resources.

III. EXPERIMENTAL DESIGN

A. Test Set Construction

A structured 120-message test set was constructed across six linguistically distinct input categories, specifically designed to expose the failure modes of keyword-based classifiers by including realistic proportions of implicit, idiomatic, and code-switched distress expressions — the categories in which lexicon-based methods are documented to fail [4, 8]. Unlike standard benchmark evaluations that favour explicit vocabulary, this stratified construction enables direct measurement of the accuracy gap that keyword classifiers produce in production deployments serving linguistically diverse populations. Ground truth labels (high_stress or normal) were assigned by two independent annotators with a Cohen's kappa of 0.91, indicating near-perfect inter-annotator agreement [17]. Table I presents the test set composition by category.

TABLE I
TEST SET COMPOSITION BY CATEGORY (N = 120; 81 HIGH_STRESS, 39 NORMAL).

Category	Total	High Stress	Normal
Explicit keyword match	45	30	15
Compound stress expression	20	15	5
Implicit distress	18	12	6
Idiomatic expression	15	10	5
Code-switched non-English	10	7	3
Cultural metaphors	12	7	5
Total	**120**	**81**	**39**

Explicit keyword match examples include direct vocabulary use ("I am so stressed out"). Compound expressions combine multiple stress indicators ("I am exhausted and depressed and don't know what to do"). Implicit distress examples contain no stress vocabulary but convey distress semantically ("I don't see the point in trying anymore"). Idiomatic

examples use figurative language ("I'm hanging on by a thread"). Code-switched examples mix Hindi/Marathi with English ("Main bahut pareshaan hoon aaj kal"). Cultural metaphor examples use imagery-based expressions ("I carry the weight of a mountain every day").

B. Classifier Implementation

All three classifiers were implemented in Python 3.14. The keyword classifier uses built-in string operations with no external dependencies. The DistilBERT classifier uses the HuggingFace Transformers library (version 5.12.1) with the bhadresh-savani/distilbert-base-uncased-emotion pre-trained model [16]. The hybrid classifier combines both using OR-gate logic as described in the System Description. All experiments were run on a Windows 11 workstation with an NVIDIA GPU and 16 GB RAM.

C. Evaluation Metrics

The primary evaluation metrics are: Overall Classification Accuracy, Binary F1-Score (treating high_stress as the positive class), Macro-Average F1-Score, Precision, Recall (True Positive Rate), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), and per-category detection rate. The binary F1-score is the primary metric for clinical evaluation, as it directly captures the classifier's ability to detect distress—the safety-critical task. Average per-message latency (ms) is measured over the full 120-message test set after a 5-message warm-up pass.

IV. RESULTS

A. Test Set Construction

Table II presents the complete accuracy metrics for all three classifiers across the 120-message test set.

TABLE IIIII
OVERALL ACCURACY METRICS FOR ALL THREE CLASSIFIERS (N = 120).

Metric	Keyword	DistilBERT	Hybrid
Overall Accuracy	70.00%	90.83%	91.67%
F1-Score (binary: high_stress)	0.714	0.929	0.936
F1-Score (macro-average)	0.699	0.900	0.908
Precision	100.00%	97.30%	97.33%
Recall (TPR)	55.56%	88.89%	90.12%
True Negative Rate (TNR)	100.00%	94.87%	94.87%
False Positive Rate (FPR)	0.00%	5.13%	5.13%
False Negative Rate (FNR)	44.44%	11.11%	9.88%
Avg. Latency (ms/msg)	<0.01	27.10	17.17

The keyword classifier achieves perfect precision (100%) and zero false positive rate, meaning it never misclassifies a normal message as stressful. However, its recall of 55.56% and FNR of 44.44% reveal that it misses nearly half of all genuine distress signals. The DistilBERT classifier substantially improves recall to 88.89% (FNR = 11.11%) while accepting a modest 5.13% false positive rate. The hybrid classifier achieves the highest recall of 90.12% (FNR = 9.88%), further reducing missed crisis signals while maintaining near-identical precision (97.33%) and a measured latency of 17.17 ms—approximately 2,700 times slower than keyword-only but 37% faster than DistilBERT-only, owing to the short-circuit evaluation where keyword matches avoid DistilBERT inference entirely.

B. Confusion Matrices

Tables III and IV present the confusion matrices for the keyword and hybrid classifiers respectively.

TABLE IVVVI
CONFUSION MATRIX — KEYWORD CLASSIFIER (N = 120).

	Predicted: Normal	Predicted: High Stress
Actual: Normal	39 (TN)	0 (FP)

	Predicted: Normal	Predicted: High Stress
Actual: High Stress	36 (FN)	45 (TP)

TABLE VIIV
CONFUSION MATRIX — HYBRID CLASSIFIER (N = 120).

	Predicted: Normal	Predicted: High Stress
Actual: Normal	37 (TN)	2 (FP)
Actual: High Stress	9 (FN)	72 (TP)

The hybrid classifier reduces false negatives from 36 (keyword) to 8—a 77.8% reduction in missed crisis signals. The 2 false positives introduced are clinically acceptable: incorrectly offering crisis resources to a non-distressed user is a far lower-risk error than failing to respond to a distressed user.

C. Detection Rate by Input Category

Table V presents the per-category detection rates, revealing the failure modes and recovery achieved by the proposed hybrid architecture across different stress categories.

TABLE V
FONT SIZES FOR PAPERS

Category	Keyword	DistilBERT	Hybrid
Explicit keyword match	96.7%	96.7%	100.0%
Compound stress expression	100.0%	100.0%	100.0%
Implicit distress	0.0%	66.7%	66.7%
Idiomatic expression	0.0%	100.0%	100.0%
Code-switched non-English	0.0%	57.1%	57.1%
Cultural metaphors	14.3%	85.7%	85.7%

The keyword classifier achieves 0% detection on three of six categories—implicit distress, idiomatic expressions, and code-switched inputs—because these categories contain no words from the 16-term distress lexicon. DistilBERT eliminates the zero-detection floor entirely, achieving 66.7% on implicit distress, 100% on idiomatic expressions, and 57.1% on code-switched inputs. Cultural metaphor detection improves from 14.3% (keyword) to 85.7% (DistilBERT/Hybrid). The hybrid recovers the additional 3.3% on explicit keyword match over DistilBERT alone (96.7% → 100%), demonstrating complementarity between the two layers.

D. Latency Analysis

The keyword classifier achieves sub-millisecond latency (measured at <0.01 ms per message), consistent with its O(k) string matching complexity. DistilBERT requires 27.10 ms per message on CPU. The hybrid architecture achieves 17.17 ms per message — 37% faster than DistilBERT alone — because 45 of the 120 test messages (37.5%, measured directly from Table I: all 45 explicit keyword-match messages triggered the synchronous keyword layer and bypassed DistilBERT inference through short-circuit evaluation). In production, where explicit-keyword distress messages are common, this short-circuit effect reduces average hybrid latency even further.

V. DISCUSSION

A. Detection Rate by Input Category

The experimental results reveal a more severe limitation of keyword-based stress detection than previous evaluations have suggested. The 70.0% overall accuracy observed in this study—compared to figures of 87–90% sometimes reported in simpler test conditions—reflects the importance of linguistically stratified evaluation. When the test set includes realistic proportions of implicit, idiomatic, and code-switched distress expressions (as experienced in production deployments serving multilingual Indian users), the keyword classifier fails on 44.44% of all genuine distress signals. The three categories with 0% keyword detection—implicit distress, idiomatic expressions, and code-switched inputs—represent exactly the failure modes most commonly observed in real-world chatbot deployments [4, 8, 11]. These results align

with broader findings that feature extraction technique selection significantly impacts classification performance [4] and that lexicon-based approaches are fundamentally unable to capture semantic or pragmatic meaning.

The keyword classifier's perfect precision (100%) and zero FPR are clinically relevant properties: in a conservative screening system where, false positives carry high cost (e.g., unnecessary emergency contacts), keyword matching may be preferred. However, in the mental wellness chatbot context, where false negatives (missed crisis signals) represent the higher-risk failure mode, a 44.44% FNR renders the classifier inadequate as a sole crisis detection mechanism [1].

B. Hybrid Architecture Performance

The hybrid OR-gate architecture achieves the best overall performance across all primary metrics: 91.67% accuracy, binary F1 = 0.936, and FNR = 9.88%. Critically, it reduces missed crisis signals from 36 (keyword-only) to 8—a 77.8% reduction in false negatives—at the cost of introducing only 2 false positives. This asymmetric improvement is clinically desirable, aligning with the safety principle that missed crisis signals are more dangerous than unnecessary escalation [13].

The latency of 17.17 ms is well within the acceptable range for conversational AI systems, where sub-100 ms response times are widely considered acceptable for real-time interaction in production chatbot deployments [6]. The synchronous keyword layer preserves the immediate-response guarantee for real-time crisis escalation, while the asynchronous DistilBERT layer updates the rolling stress index without blocking the response pipeline. This design pattern—combining a fast deterministic layer with a slower semantic layer—generalises to other production AI safety applications where both latency and accuracy are constrained.

The hybrid system is also complementary in a structural sense: the keyword layer contributes unique recall on explicit keyword matches (100% vs 96.7% for DistilBERT alone), while the DistilBERT layer captures idiomatic (100%), metaphorical (85.7%), and implicit (66.7%) distress that the keyword layer misses entirely. This demonstrates that the two classifiers are not redundant but genuinely complementary [8, 11].

C. Remaining Limitations

Three limitations remain. First, implicit distress detection reaches only 66.7% even with DistilBERT, as messages like "I don't see the point in trying anymore" require inference of intent that the emotion classification model does not always capture —a recognised challenge across ML-based mental health detection systems [9]. Second, code-switched detection (57.1%) is limited by the English-only training corpus of the base model; multilingual models such as mBERT or XLM-RoBERTa would improve this category. Third, the test set (n = 120) is structured and controlled; evaluation on naturalistic conversation logs would strengthen external validity. Regarding privacy guarantees: direct inspection of the production database following test conversations confirmed zero plaintext values in the encrypted_message and encrypted_response columns — all stored values are URL-safe base64 Fernet ciphertext, consistent with the authenticated encryption requirement of GDPR Article 32.

VI. COMPARATIVE ANALYSIS OF RESULTS

Situating the hybrid system within the broader literature requires comparing across accuracy, latency, and task scope. Because test sets, emotion taxonomies, and task definitions differ across studies, direct numerical comparison must be interpreted carefully — these differences are highlighted explicitly below.

TABLE VVIII
COMPARISON OF THE PROPOSED HYBRID SYSTEM WITH RELATED PUBLISHED APPROACHES.

Study	*Approach*	*Accuracy / F1*	*Dataset / Task*	*Latency*
Barakaz et al. [5]	RoBERTa fine-tuned	96.30% / F1 = 0.981	Standard sentiment benchmark	~80–150 ms

Study	*Approach*	*Accuracy / F1*	*Dataset / Task*	*Latency*
AI-DASA [3]	Multi-model ensemble	94.00% / —	Student text; multi-class (depression, anxiety, stress)	Not reported
Salih et al. [7]	BERT fine-tuned	91.77% / —	News text; topic classification	Not reported
Loekito et al. [13]	VAD-calibrated IndoBERT	— (52% stability gain)	Indonesian counseling chatbot; emotion	Not reported
Devi et al. [8]	GMM-enhanced N-gram LSTM	~85–89% / —	General sentiment; multi-class	Not reported
Saif et al. [4]	LinearSVC, Logistic Regression	70–89% / —	Social media text; emotion detection	< 1 ms
DistilBERT benchmark [16]	DistilBERT (DAIR-AI/Emotion)	92.70% / —	DAIR-AI/Emotion held-out test set	~27 ms
This work — Keyword-only	Lexicon lookup	70.00% / F1 = 0.714	Stratified stress detection; binary	**< 0.01 ms**
This work — DistilBERT-only	DistilBERT fine-tuned	90.83% / F1 = 0.929	Stratified stress detection; binary	27.10 ms
This work — Hybrid	Keyword + DistilBERT (OR gate)	**91.67% / F1 = 0.936**	Stratified stress detection; binary	**17.17 ms**

Several observations emerge from this comparison. First, the hybrid system's F1 = 0.936 is competitive with the strongest transformer results in the ETASR literature. Barakaz et al. [5] report F1 = 0.981 for RoBERTa, but on standard general-sentiment benchmarks where language is comparatively uniform and the classification task is not safety-critical. Our evaluation set is deliberately harder: it includes implicit distress, code-switching, idioms, and cultural metaphors that standard benchmarks do not contain. On that more challenging six-category set, F1 = 0.936 at sub-20 ms latency represents strong performance.

Second, none of the reviewed works report inference latency in a real-time chatbot context. To our knowledge, this study is the first to benchmark stress detection classifiers against an explicit conversational latency constraint. At 17.17 ms per message, the hybrid comfortably fits inside the responsiveness window required by production chat systems [6]. Pure transformer alternatives — including RoBERTa, which typically exceeds 80 ms on CPU — cannot match this. The asynchronous keyword-first architecture is the key design decision that enables this speed advantage without sacrificing accuracy.

Third, the comparison with the public DistilBERT benchmark [16] is directly informative. The bhadresh-savani/distilbert-base-uncased-emotion model achieves 92.7% on the DAIR-AI/Emotion held-out set. Our DistilBERT-only classifier, using the same model, achieves 90.83% on our custom test set — a gap of 1.87 percentage points. This gap confirms that our evaluation set is harder than the standard benchmark, because it includes message categories not represented in DAIR-AI/Emotion: code-switched inputs (detection rate 57.1%) and culturally metaphorical expressions (detection rate 85.7%), as well as a higher proportion of implied distress. The hybrid's OR-gate design recovers part of this gap, pushing accuracy to 91.67%.

Fourth, domain specificity is critical for safety-sensitive mental health applications. Loekito et al. [13] make this point clearly for Indonesian counseling chatbots, where domain-adapted IndoBERT substantially outperforms general models in prediction stability. Our domain-specific test set — designed around the six categories of distress expression that appear in real mental wellness chatbot conversations — reflects the same principle. The hybrid's 77.8% reduction in missed crisis signals relative to keyword-only is only quantifiable because the evaluation was designed to capture linguistically diverse failure modes, rather than relying on a standard emotion benchmark that does not contain them.

Fifth, the comparison with classical ML approaches [4] is instructive for deployment context. Saif et al. [4] report 70–89% accuracy for feature-engineered classifiers (TF-IDF, N-gram, LinearSVC) on social media text — a range that overlaps with our keyword-only baseline (70.0%). This confirms that simple approaches, while fast, leave a clinically unacceptable accuracy floor for stress detection. The hybrid design bridges this gap, matching transformer-level accuracy while retaining the sub-20 ms speed that keyword-based and classical ML methods provide.

VII. CONCLUSIONS

This study presents the design, implementation, and empirical evaluation of a hybrid keyword–transformer stress detection architecture for a production mental health chatbot. The primary finding is that a simple 16-term keyword classifier achieves only 70.0% accuracy (binary F1 = 0.714) with a 44.44% false negative rate on a linguistically stratified test set—substantially below the performance often assumed for keyword-based systems—while a DistilBERT-based classifier achieves 90.83% accuracy (F1 = 0.929) and the hybrid achieves 91.67% accuracy (F1 = 0.936). The hybrid architecture reduces missed crisis signals by 77.8% compared to keyword-only classification while maintaining 17.17 ms latency and preserving the public API contract of the production system for zero-disruption deployment.

The category-level analysis reveals that keyword classifiers achieve 0% detection on implicit, idiomatic, and code-switched distress inputs—which together represent a substantial proportion of real-world mental health chatbot traffic, particularly in linguistically diverse populations. DistilBERT-based classification eliminates the zero-detection floor on all categories, achieving 66.7–100% detection across previously failed categories. The hybrid OR-gate design is shown to be strictly superior to either component alone, offering higher recall than DistilBERT while recovering explicit keyword matches that DistilBERT occasionally misses.

Compared to recent emotion detection systems in the literature, the hybrid classifier's F1 = 0.936 is competitive with the best transformer-based approaches reported in ETASR [2, 5] while operating at latencies suitable for real-time conversational deployment—far below the 40–300 ms typically required for standalone transformer inference [6]. This positions the hybrid architecture as a practical, production-deployable solution for the critical patient safety requirement of real-time stress detection in mental health conversational AI systems. Future work will extend evaluation to naturalistic conversation logs, explore multilingual transformer models to improve code-switched input detection, and investigate active learning strategies for continuous classifier improvement from production interaction data.

DECLARATION OF COMPETING INTERESTS

The authors declare no conflict of interest.

ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Science and Engineering, College of Engineering and Technology, Akola, Sant Gadge Baba Amravati University, for their institutional support during this research.

DATA AVAILABILITY

The 120-message annotated test set, evaluation code, and experiment results are available from the corresponding author upon reasonable request. The DistilBERT model used in this study is publicly available at https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion [16].

AUTHOR CONTRIBUTIONS

Conceptualization, P.N.S. and R.T.; methodology, P.N.S.; software, P.N.S.; validation, P.N.S.; formal analysis, P.N.S.; writing—original draft, P.N.S.; writing—review and editing, R.T.; supervision, R.T. All authors have read and agreed to the published version.

FUNDING

This research received no external funding.

REFERENCES

[1] World Health Organization, *World Mental Health Report: Transforming Mental Health for All*. Geneva, Switzerland: WHO Press, 2022.

[2] A. Rehman, M. Ahmed, H. U. Khan, A. Bukhari, A. Daud, and H. Dawood, "Mental health sentiment analysis: Exploring an optimized BERT with deep encodings," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 26242–26248, Oct. 2025. https://doi.org/10.48084/etasr.10469

[3] W. Minaya, F. Aramburu, and J. Santisteban, "AI-DASA: AI-based depression, anxiety, and stress assessment," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 28511–28522, Dec. 2025. https://doi.org/10.48084/etasr.13210

[4] W. Q. A. Saif, M. K. Alshammari, B. A. Mohammed, and A. A. Sallam, "Enhancing emotion detection in textual data: A comparative analysis of machine learning models and feature extraction techniques," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, Oct. 2024. https://doi.org/10.48084/etasr.7806

[5] S. M. Barakaz, S. A. Alanazi, M. Alruily, and F. E. Mouatasim, "Advancing sentiment analysis: Evaluating RoBERTa against traditional and deep learning models," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20167–20174, Feb. 2025. https://doi.org/10.48084/etasr.9703

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in *Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2-NeurIPS)*, 2019. https://arxiv.org/abs/1910.01108

[7] M. I. Salih, S. M. Mohammed, A. Kh. Ibrahim, O. M. Ahmed, and L. M. Haji, "Fine-tuning BERT for automated news classification," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, Jun. 2025. https://doi.org/10.48084/etasr.10625

[8] K. Dhana Sree Devi, B. Suvarna, S. Sowmya, M. Dayakar, and V. Ravi Kumar, "A novel approach to sentiment analysis using GMM-enhanced N-gram LSTM networks," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, Jun. 2025. https://doi.org/10.48084/etasr.12586

[9] R. D. Lim, J. Valdez, and M. Poblete, "Detecting mental health conditions from social media text using machine learning: A systematic review," *IEEE Access*, vol. 11, pp. 52667–52687, 2023. https://doi.org/10.1109/ACCESS.2023.3278716

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[11] D. Barua and T. S. Walia, "A deep learning-based framework for dataset creation and sentiment classification of English-Bengali code-mixed texts," *Engineering, Technology & Applied Science Research*, vol. 16, no. 1, Feb. 2026. https://doi.org/10.48084/etasr.15475

[12] P. Mullangi, N. Dimmita, S. S. R. Kollu, V. R. Potnuru, K. R. Kalahasthi, and M. V. B. Kondaiah, "Sentiment and emotion modeling in text-based conversations utilizing ChatGPT," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20042–20048, Feb. 2025. https://doi.org/10.48084/etasr.9508

[13] J. A. Loekito, A. Tjahyanto, and R. Indraswari, "Explainable emotion recognition using a valence arousal dominance calibrated IndoBERT for Indonesian counseling chatbots," *Engineering, Technology & Applied Science Research*, vol. 16, no. 3, pp. 35937–35949, Jun. 2026. https://doi.org/10.48084/etasr.17581

[14] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," *JMIR Mental Health*, vol. 4, no. 2, Art. no. e19, 2017. https://doi.org/10.2196/mental.7785

[15] K. Coghlan, A. Leins, K. Broad, N. Sellis, J. Biderman, and S. Kieseler, "Safe messaging and artificial intelligence: An evaluation of mental health-related responses from AI chatbots," *JMIR Mental Health*, vol. 10, Art. no. e46710, 2023. https://doi.org/10.2196/46710

[16] B. Savani, "distilbert-base-uncased-emotion," HuggingFace Model Hub, 2021. [Online]. Available: https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

[17] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

