NeuroAI: A Multimodal Framework for Text and Speech Emotion Recognition, Suicide/Self-Harm Risk Screening, and PHQ-9-Based Mental Health Assessment

Atharva Naringrekar
Department of Information Technology
Vidyalankar Institute of Technology
Mumbai, India

Aryan Sawant
Department of Information Technology
Vidyalankar Institute of Technology
Mumbai, India

Vedant Gupta
Department of Information Technology
Vidyalankar Institute of Technology
Mumbai, India

Adi Sheth
Department of Information Technology
Vidyalankar Institute of Technology
Mumbai, India

Dr. Neha Kudu
Deprartment of Information Technology
Vidyalankar Institute of Technology
Mumbai, India

Abstract—Mental-health-related information is communicated through several channels, including written language, speech, and structured self-report questionnaires. Systems that analyse only one channel may overlook contextual or behavioural information, while emotion recognition alone cannot establish a mental-health diagnosis or determine suicide risk. This paper presents NeuroAI, a research-oriented multimodal framework that integrates text emotion recognition, text-based suicide/self-harm risk screening, speech emotion recognition, and deterministic Patient Health Questionnaire-9 (PHQ-9) scoring within a unified application.

The text emotion component uses a RoBERTa-based classifier for 28 categories derived from GoEmotions. A separate RoBERTabased binary classifier detects language patterns associated with potential suicide/self-harm risk. The speech pipeline converts audio to mono at 22,050 Hz, extracts 715 acoustic features, and uses a PyTorch VoiceNet model to classify six emotion categories. PHQ-9 responses are summed deterministically to obtain a score between 0 and 27. A transparent weighted fusion mechanism combines available outputs using configured weights of 35% for text, 35% for voice, 20% for PHQ-9, and 10% for the safety-risk signal. The framework also supports missing-modality weight renormalization, rule-based safety overrides, historical summaries, alerts, and clinician-oriented workflows. Reported component results are 56.50% accuracy and 55.92% weighted F1-score for text emotion recognition, 97.38% accuracy and F1-score for the binary safety classifier, and 76.28% held-out validation accuracy for speech emotion recognition. These figures describe separate component tasks and do not represent endto-end clinical performance. The proposed contribution is an integrated and inspectable software workflow; clinical validity and overall fusion accuracy remain to be established through independent evaluation.

Index Terms—Multimodal artificial intelligence, emotion recognition, RoBERTa, speech emotion recognition, suicide/selfharm risk screening, PHQ-9, explainable AI, mental-health decision support.

I. INTRODUCTION

Mental health is influenced by a broad range of psychological, social, and environmental factors. People communicate their feelings and experiences through written messages, spoken language, and structured self-report questionnaires. Each channel captures different information. Text may express semantic content and context, speech can carry prosodic and acoustic cues, and questionnaires provide standardized responses to specific symptom questions. Analysing these sources together may provide a more organized view of available information than treating each source as an isolated result.

Recent advances in natural language processing and machine learning have made emotion classification possible at scale. Transformer architectures learn contextual representations from text, while acoustic models can identify patterns in pitch, energy, spectral structure, and speech rhythm. However, model predictions are not equivalent to psychological truth. A message classified as sad does not necessarily indicate depression, and a calm voice cannot

establish that a person is free from distress. Likewise, language that resembles selfharm-related content requires careful contextual interpretation rather than an automatic clinical conclusion.

A further challenge is that emotion recognition, suicide/selfharm risk screening, and depression symptom measurement are distinct tasks. A single sentiment score cannot appropriately replace a dedicated safety model or a validated questionnaire. The Patient Health Questionnaire-9 (PHQ-9) is a standardized self-report instrument that asks about the frequency of depressive symptoms over a defined period [7]. Its total score is calculated from nine item responses and should be interpreted in context by an appropriately qualified professional.

This paper presents NeuroAI, a multimodal research prototype that integrates four analytical paths: text emotion recognition, text-based suicide/self-harm risk screening, speech emotion recognition, and PHQ-9 scoring. A fusion layer combines available outputs with configurable weights and rule-based safety handling. The application also provides history, alerts, summaries, and a doctor-facing workspace to support review of previously collected information.

The system is designed around the principle that automation should organize information without replacing human judgement. Component-level results are kept visible so that the combined output does not conceal which model or questionnaire generated a signal. Missing input is handled explicitly, and the system is framed as non-diagnostic decision support. The contributions of this work are:

1) Integration of text emotion, separate suicide/self-harm risk screening, speech emotion, and deterministic PHQ9 scoring in one application workflow.
2) A speech emotion pipeline using mono audio resampled to 22,050 Hz, a 715-feature acoustic representation, and a PyTorch VoiceNet classifier.
3) A transparent weighted fusion method with renormalization for unavailable modalities and rule-based safety overrides.
4) An application architecture supporting historical analysis, alerts, summaries, and clinician-oriented review.
5) A discussion of evaluation boundaries, responsible use, and the validation required before real-world clinical deployment.

II. Related Work

A. Fine-Grained Emotion Recognition in Text

Emotion recognition extends beyond conventional sentiment analysis, which typically identifies positive, negative, or neutral polarity. Fine-grained emotion classification attempts to distinguish states such as sadness, anger, fear, gratitude, confusion, and amusement. The GoEmotions dataset introduced by Demszky *et al.* contains approximately 58,000 Reddit comments annotated with 27 emotion categories and a neutral category [1]. Its label inventory makes it useful for studying subtle affective distinctions, although social-media language may differ from clinical dialogue in style, vocabulary, and context.

Bidirectional Encoder Representations from Transformers (BERT) established contextual language representations through pre-training on large text corpora [2]. RoBERTa improved the pre-training procedure and demonstrated strong language-understanding performance [3]. NeuroAI uses a RoBERTa-based model to map text to 28 emotion categories. The current inference design selects a dominant class using the highest predicted probability. This simplifies presentation but can discard co-occurring emotions, a limitation worth addressing in future multi-label evaluation.

B. Speech Emotion Recognition

Speech conveys information through both words and acoustic delivery. Features such as fundamental frequency, energy, spectral centroid, chroma, zero-crossing rate, and Melfrequency cepstral coefficients (MFCCs) have been used in speech processing. Davis and Mermelstein's work established MFCCs as a useful representation for speech recognition [4]. The CREMA-D dataset contains 7,442 audio-visual clips from 91 actors and supports six emotion categories: anger, disgust, fear, happiness, neutral, and sadness [5].

Speech emotion recognition is challenging because vocal expression varies with speaker identity, accent, language, microphone quality, background noise, and the difference between acted and spontaneous speech. A model trained on acted recordings may not generalize to natural conversations. NeuroAI's implementation extracts 715 acoustic features and uses a PyTorch VoiceNet classifier. The feature count describes the implementation, not a guarantee of robustness across environments.

C. Multimodal Learning

Multimodal learning seeks to combine information from multiple input types. Tsai *et al.* proposed the Multimodal Transformer for unaligned multimodal language sequences, modelling interactions across language, audio, and visual streams [6]. Learned fusion can capture complex relationships but may require large datasets, careful alignment, and substantial computation. It can also be difficult to explain to end users.

NeuroAI adopts a transparent weighted fusion strategy rather than a learned cross-modal architecture. This choice supports auditability because the configured weights and available modalities can be inspected. However, a manually weighted system may fail to capture nonlinear relationships and cannot be assumed to outperform more sophisticated fusion methods without controlled evaluation.

D. Risk Screening and Mental-Health Assessment

Emotion classification and suicide/self-harm risk screening are related but not interchangeable. Emotion models characterize expressed affect, whereas a safety classifier looks for text patterns associated with possible risk. Using sadness as a direct proxy for suicide risk would conflate two different prediction tasks. NeuroAI therefore includes a separate binary text classifier and presents its result as a screening signal rather than a clinical risk determination.

The PHQ-9 is a nine-item self-report measure used to assess depressive symptom frequency [7]. Each item is scored from zero to three, resulting in a total from zero to 27. NeuroAI calculates the sum deterministically. This distinction is important because questionnaire scores represent a user's reported answers, not a machine's inference from their writing or voice.

E. Explainability and Responsible Use

Explainability methods can help make model behaviour more inspectable. In a multimodal mental-health application, interpretability includes showing the separate text, voice, and questionnaire outputs, indicating unavailable inputs, and documenting how the fusion result was produced. Nonetheless, transparent rules do not guarantee that a system is accurate or fair. Responsible use also requires privacy safeguards, appropriate consent, uncertainty communication, and human oversight.

III. Research Gap and Motivation

Existing tools commonly focus on one part of the workflow: emotion classification, speech analytics, symptom questionnaires, or clinical record management. Integrating these outputs presents several practical challenges. First, model outputs may be difficult to compare because they use different label spaces and scales. Second, an absent modality should not be treated as evidence of a negative result. Third, a safety signal should not be suppressed merely because another component reports a positive or neutral emotion. Finally, historical information must be presented without implying that a sequence of model predictions constitutes a diagnosis.

NeuroAI is motivated by the need for a unified interface that preserves task boundaries while allowing users to inspect multiple signals together. The design combines four core capabilities: (1) text emotion recognition, (2) independent suicide/selfharm risk screening, (3) speech emotion recognition, and (4) PHQ-9 scoring. The fusion layer provides an aggregate summary, while application services make component-level history available for authorized review.

The research gap addressed here is primarily an integration and workflow gap. The paper does not claim that the proposed combination is a new clinical instrument or that multimodal fusion necessarily improves diagnostic performance. Rather, it describes a modular implementation and the component results currently available, and identifies the experiments needed to assess the integrated system.

IV. Proposed Contributions

The proposed NeuroAI framework has the following contributions.

1) Task-separated analysis: emotion recognition, safetyrisk screening, speech emotion, and PHQ-9 scoring are handled by separate services.
2) Heterogeneous input support: text, audio, and structured questionnaire responses can be processed through a common backend workflow.
3) Configurable fusion: available outputs are combined using explicit weights and documented handling for missing modalities.
4) Safety-aware interpretation: rule-based overrides can surface a concerning safety signal independently of the aggregate emotion result.
5) Historical review: authorized users can inspect past analyses, trends, alerts, and session summaries.
6) Modular implementation: the frontend, backend, model services, fusion engine, and database can be maintained and evaluated independently.

V. System Architecture

NeuroAI follows a modular, layered architecture. Its main layers are input acquisition, preprocessing, model inference, PHQ-9 scoring, fusion and interpretation, persistence, and application presentation. The frontend is implemented using React, Vite, and Tailwind CSS. FastAPI provides backend endpoints and coordinates application services. Model-specific services perform text and audio inference, while the database stores authorized user and analysis records. The project report describes SQLite with SQLAlchemy 2.0 Async for application data storage.

A. Input Layer

The system accepts three types of input. Text input is used by the emotion classifier and the separate safety classifier. Audio input is passed through the speech preprocessing and feature extraction pipeline. PHQ-9 input consists of nine structured responses, each constrained to the valid range of zero to three. Input validation should reject malformed audio, invalid questionnaire values, and empty or unsupported requests before inference begins.

B. Model Inference Layer

The text services use RoBERTa-based classifiers. One predicts a dominant emotion from 28 categories, while the other predicts a binary potential-risk label. The voice service extracts acoustic features and passes the standardized feature vector to VoiceNet. Each service returns a result in its own task-

specific format, including the predicted class and any confidence value supported by the model.

C. Fusion Layer

The fusion layer receives whichever component outputs are available. Since the outputs do not naturally share a common scale, the implementation must map them to the normalized representation used by its rules before combining them. The configured weights are explicit system settings, not learned clinical coefficients. If a modality is unavailable, its weight is removed and the remaining eligible weights are renormalized.

D. Application Layer

The application includes patient-facing and doctor-facing workflows. Users can submit inputs and review available results. Authorized clinicians can review historical analyses, alerts, appointments, session information, and generated summaries. SOAP note assistance is intended to organize information for professional review; any generated note requires verification and editing by the responsible clinician.

E. Data and Security Layer

The application data model described in the project includes users, analysis history, emotion reports, voice reports, PHQ reports, doctor alerts, appointments, session notes, audit logs, and settings. Because mental-health-related information is sensitive, access should be restricted by role and need. Authentication, password hashing, input validation, rate limiting, and audit logging are useful controls, but a production deployment would still require a threat model, security testing, retention policies, incident response procedures, and legal and ethical review.

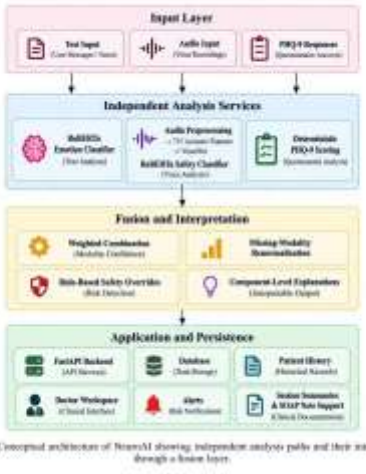


Fig. 1. Conceptual architecture of NeuroAI showing independent analysis paths and their integration through a fusion layer.

VI. METHODOLOGY

A. Text Emotion Classification

Let x represent an input text and let $p_k(x)$ be the probability assigned by the classifier to emotion category k. The predicted category is

$$\hat{y}_t = \arg \max_{k \in \{1,\ldots,28\}} p_k(x). \quad (1)$$

The output is the dominant predicted emotion. The confidence value indicates the model's relative score, not a calibrated probability of psychological truth. The current single-label approach is straightforward to display but can miss cases in which multiple emotions are expressed simultaneously.

The text model is based on RoBERTa, a transformer language model that uses contextual representations. The GoEmotions label space provides 27 emotion categories and neutral. Text normalization, tokenization, truncation, and modelspecific preprocessing must be consistent between training and inference. The final implementation should document the maximum sequence length, tokenizer version, training split, optimizer, learning rate, number of epochs, and model checkpoint used.

B. Suicide/Self-Harm Risk Screening

The safety service is a separate binary RoBERTa-based classifier. Let $q(x)$ be the model score for the positive riskrelated class and let τ denote the configured threshold. The binary output can be represented as

$$\hat{y}_r = \begin{cases} 1, & q(x) \geq \tau, \\ 0, & q(x) < \tau. \end{cases} \quad (2)$$

The available project report does not establish a clinically validated threshold, so $\hat{y}_r$ must be treated as a model-generated screening signal. It should not be used alone to determine whether a person is safe or unsafe. Context, professional assessment, and appropriate safeguarding procedures remain necessary.

C. Audio Preprocessing

The audio pipeline converts recordings to mono and resamples them to 22,050 Hz. Mono conversion provides a consistent channel representation, while resampling standardizes the sampling rate expected by feature extraction. The system then calculates 715 acoustic features, including measures associated with energy, spectral properties, MFCCs, chroma, pitch, and zero-crossing rate. The feature vector is standardized using the same transformation used during model training.

Let the feature vector be $\mathbf{z} \in \mathbb{R}^{715}$. The VoiceNet classifier maps the vector to a distribution over six emotion classes:

$$\mathbf{p}_v = f_\theta(\mathbf{z}), \quad (3)$$

where f_θ is the trained PyTorch model. The selected label is the class with the largest output score. Audio preprocessing

cannot fully remove the effects of background noise, speaker identity, accent, speaking style, or differences between acted and spontaneous speech.

D. PHQ-9 Scoring

Let $r_i \in \{0,1,2,3\}$ represent the response to PHQ-9 item i. The total score is

$$S_{\text{PHQ9}} = \sum_{i=1}^{9} r_i, \qquad 0 \le S_{\text{PHQ9}} \le 27. \tag{4}$$

The system performs deterministic addition rather than machine-learning inference. It should validate that all nine responses are in the permitted range and clearly communicate if any response is missing. The score represents self-reported symptom frequency and is not, by itself, a diagnosis. Any severity interpretation should follow the appropriate PHQ-9 guidance and professional context.

E. Weighted Fusion

NeuroAI uses configured weights of 35% for text emotion, 35% for speech emotion, 20% for PHQ-9, and 10% for the safety-risk signal. Let w_i denote the configured weight for component i, s_i its normalized contribution, and m_i an availability indicator equal to one if the component is available and zero otherwise. The renormalized weight is

$$\tilde{w}_i = \frac{m_i w_i}{\sum_j m_j w_j}, \tag{5}$$

provided that at least one eligible component is available. The aggregate representation is

$$F = \sum_i \tilde{w}_i s_i. \tag{6}$$

Equations 5 and 6 express the general fusion principle. Exact normalization of individual outputs and thresholds depend on the implemented configuration and should be documented in the released code.

The weights are design settings, not statistically optimized or clinically validated coefficients. In particular, the four outputs measure different concepts and may not be directly commensurate without an explicit normalization scheme. The aggregate score should therefore be interpreted as a software summary rather than a probability of mental-health status.

TABLE I
CONFIGURED FUSION WEIGHTS

Component	Weight
Text emotion	35%
Speech emotion	35%
PHQ-9	20%
Safety-risk signal	10%
Total	100%

F. Missing-Modality Handling

Real usage may include text without audio, audio without a completed questionnaire, or questionnaire responses without text. If an input is absent, its availability indicator is set to zero and its weight is excluded from the denominator in Eq. 5. This avoids treating a missing component as a zerovalued negative signal. However, the resulting aggregate uses a different evidence set, so results from different modality combinations may not be directly comparable. The interface should explicitly list which inputs contributed to each result.

G. Rule-Based Safety Overrides

A weighted aggregate can obscure a concerning signal if other components produce benign-looking outputs. NeuroAI therefore includes rule-based override logic intended to surface safety-related predictions or elevated questionnaire results. The rules must be explicit, testable, and version-controlled. Their thresholds should not be described as clinically validated unless this is supported by a suitable study. An override should trigger review or display guidance rather than automatically produce a diagnosis or a definitive risk label.

H. End-to-End Workflow

The workflow consists of input validation, independent inference, result persistence, fusion, safety-rule evaluation, and presentation. Each component should return a structured status so that a failed model is not confused with a negative prediction. The backend should preserve timestamps, model versions, and the set of available modalities to support reproducibility. Historical analysis should display trends cautiously because changes in model versions, input quality, or available modalities can affect the outputs.

VII. EXPERIMENTAL SETUP

A. Datasets and Tasks

The project report identifies GoEmotions for text emotion classification and CREMA-D for speech emotion recognition. A separate binary dataset was used for suicide/self-harm risk screening, but the material available for this manuscript does not fully specify its source, composition, licensing, class balance, or split strategy. These details should be added from the experiment records before publication.

For reproducibility, a final version of this paper should document exact dataset versions, download dates, licensing

TABLE II
DATASETS AND COMPONENT TASKS

Component	Data / labels		Task
Text emotion	GoEmotions; emotions plus neutral	27	Predict one of 28 emotion labels from text.

Safety screening	Separate binary text dataset	Identify potential riskrelated language patterns.
Speech emotion	CREMA-D; six classes	Predict emotion from acoustic features.
PHQ-9	Nine responses, each 0–3	Calculate a deterministic total score.

TABLE III
REPORTED RESULTS FOR INDIVIDUAL NEUROAI COMPONENTS

Component	Cls.	Acc.	F1-score	Additional Metrics
RoBERTa text emotion model	28	56.50%	55.92% weighted	Precision: 56.21%; Recall: 56.50%
RoBERTa safety classifier	2	97.38%	97.38%	Precision: 97.32%; Recall: 97.44%
PyTorch VoiceNet	6	76.28%	Not reported	Held-out validation accuracy

constraints, preprocessing, duplicate removal, class distributions, and the training/validation/test partition. For speech, the split should be speaker-independent where possible to reduce the risk that the same speaker appears in both training and test partitions. For text, near-duplicate removal and leakage checks are important, particularly when data are gathered from related sources.

B. Evaluation Metrics

Accuracy is the proportion of predictions that match the reference labels:

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All evaluated predictions}} \quad (7)$$

For binary classification, precision and recall are

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \quad (8)$$

where TP, FP, and FN denote true positives, false positives, and false negatives. The F1-score is the harmonic mean of precision and recall:

$$F_1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

For multi-class tasks, macro and weighted averages provide different views of class-level performance. Weighted F1 gives greater influence to classes with more examples, while macro F1 treats each class equally. Both can be useful when label distributions are imbalanced.

C. Reported Component Results

The project report provides the results shown in Table III. These are component-level results; the metrics should not be interpreted as evidence of end-to-end fusion accuracy.

The text emotion classifier achieved 56.50% accuracy and a weighted F1-score of 55.92%. Fine-grained emotion categories can be semantically similar, and text may express multiple emotions at once. Further evaluation should include per-class precision, recall, F1, a confusion matrix, and calibration where suitable.

The binary safety classifier achieved 97.38% accuracy and a 97.38% F1-score, with reported precision of 97.32% and recall of 97.44%. These metrics are encouraging for the reported test setup but are not sufficient to establish real-world safety. The result depends on dataset composition, class prevalence, threshold selection, and possible overlap between training and evaluation data. A safety-sensitive system must particularly examine false negatives and performance on external data.

The VoiceNet speech model achieved 76.28% held-out validation accuracy across six emotion categories. The project report contrasts this with an earlier Keras-based model using 189 acoustic features that achieved 49.43% accuracy. This is a reported comparison between implementations. Without identical data splits, preprocessing, training procedures, and repeated runs, it should not be interpreted as a controlled estimate of the benefit of adding features or changing the architecture.

D. Fusion Evaluation Boundary

No independent, ground-truth-labeled multimodal clinical test set was available in the project report to evaluate the overall fusion engine. Consequently, this paper does not report a fused-system accuracy. The individual model metrics cannot be combined arithmetically to estimate system performance. Future evaluation should define an appropriate target label, collect consented multimodal data under ethical review, preregister the evaluation protocol, and compare fusion against unimodal baselines.

VIII. RESULTS AND DISCUSSION

The reported results show that the implemented components can perform their intended classification or scoring tasks on their respective evaluation data. The text emotion model provides a fine-grained affect label, the separate safety classifier identifies risk-related text patterns, and the voice model assigns one of six speech emotion classes. PHQ-9 scoring is deterministic and can be verified directly from the nine responses. Together, these services provide the building blocks for an integrated workflow.

The most important result at the system level is the integration design rather than a demonstrated improvement in clinical prediction. The architecture keeps each task separate and provides a common point at which results can be inspected. This can be useful for application development because it avoids treating a text emotion label, an audio emotion label, a questionnaire score, and a risk prediction as if they were the same kind of measurement.

The text emotion score indicates that fine-grained affect classification remains challenging. A single-label output is convenient for a dashboard but may be insufficient for ambiguous or mixed emotional language. Multi-label prediction, class-specific analysis, and human review of difficult examples are useful next steps. The model's confidence should not be displayed as certainty, and the interface should make clear that a label refers to the input rather than a definitive assessment of the person.

The safety classifier's reported accuracy is high, but the value of a safety model cannot be judged from accuracy alone. If the negative class dominates, a high accuracy can coexist with poor recall for the positive class. Evaluation should report the confusion matrix, sensitivity, specificity, precisionrecall curves, calibration, and performance across relevant subgroups. False positives can cause unnecessary concern; false negatives can fail to surface text that merits attention. Thresholds must therefore be selected with domain experts and not solely to maximize overall accuracy.

The speech model result suggests that the acoustic pipeline captures useful patterns in its evaluation data. Nonetheless, the external validity of speech emotion models is a known challenge. A model trained on acted speech may react differently to spontaneous speech, different accents, or low-quality recordings. Robustness tests should vary recording conditions and evaluate on speakers and datasets not seen during training.

The fusion mechanism is transparent, but its weights are hand-configured. Text emotion and voice emotion are not direct measures of the same underlying variable as PHQ-9 or safety risk. A single aggregate may therefore conceal meaningful distinctions. For this reason, NeuroAI should present component-level results alongside any aggregate and describe the aggregate as a configurable summary. The configured weights should be treated as provisional engineering choices until tested in a suitable evaluation.

IX. Application Workflow and Implementation

The application follows a client-server architecture. The React-based frontend collects user inputs and presents the results. FastAPI exposes backend operations and coordinates the model services. The inference services are modular so that one model can be updated without rewriting the entire interface. The database stores user records and analysis history, subject to access control and data-handling requirements.

A. Analysis Request

A request begins when a user submits text, audio, PHQ-9 responses, or a supported combination. The backend validates the request and invokes the relevant service. Text can be passed to the emotion and safety models independently. Audio is preprocessed and transformed into the feature vector expected by VoiceNet. Questionnaire responses are validated and summed. Each service returns a structured response with its own status and output.

B. Result Presentation

The results interface should display each available component separately. For text, it can show the dominant emotion and model score. For voice, it can show the predicted emotion and the recording status. For PHQ-9, it can show the total and the fact that it was calculated from the user's responses. For the safety classifier, it should clearly state that the output is an automated screening signal and not a definitive risk assessment. The fused summary should list which components were included.

C. History and Clinician Workspace

The project includes history and clinician-oriented workflow functions. Historical views can organize prior component outputs by time and allow authorized review. Alerts can identify records requiring attention under configured rules. Session summaries and SOAP note support can reduce administrative effort, but generated content must be verified by a qualified professional. A model-generated summary should never be represented as a confirmed clinical finding unless the responsible professional has reviewed and adopted it.

D. Reproducibility and Versioning

For reliable longitudinal interpretation, the system should retain model version identifiers, preprocessing configuration, timestamps, and available-modality indicators. When a model is updated, historical outputs may not be directly comparable with newly generated results. A version-aware system can communicate this limitation and support reproducible evaluation. Exact software versions, random seeds, hardware, and training configuration should be documented in a public reproducibility package when permitted.

X. Limitations and Threats to Validity

Several limitations affect the conclusions of this study.

1) Dataset representativeness: GoEmotions and CREMAD are useful research datasets, but they may not represent the language, accents, communication patterns, and recording conditions encountered in real clinical settings.
2) Incomplete dataset details: the available project material does not fully document the safety dataset's provenance, licensing, class balance, and split strategy. These details must be supplied for reproducibility.
3) Task-specific metrics: text emotion, safety screening, and speech emotion are different tasks. Their accuracy values are not directly comparable.

4) No end-to-end validation: the fusion engine has not been independently evaluated on a labeled multimodal clinical test set. No overall system accuracy can therefore be claimed.
5) Manually configured weights: the fusion weights and rule thresholds are engineering settings, not clinically optimized parameters.
6) Domain shift: noise, accent, language, sarcasm, indirect phrasing, and spontaneous speech can reduce model reliability.
7) Potential bias: models may perform differently across demographic or linguistic groups. Subgroup evaluation is required before making claims of equitable performance.
8) Privacy and consent: text, voice recordings, and questionnaire responses may contain highly sensitive information. Secure storage and access controls are essential but do not alone establish compliance.
9) No clinical validation: the prototype has not been independently validated for diagnosis, treatment selection, or autonomous safety decisions.

These limitations define the appropriate scope of the current work. The evidence supports describing NeuroAI as an integrated software prototype that brings together separate analytical components and exposes their outputs for supervised review. It does not support describing the framework as a validated clinical assessment system.

XI. Ethical Considerations and Responsible Use

Mental-health-related predictions can influence how users interpret their experiences and how professionals prioritize attention. The system must avoid overconfident or diagnostic language. Outputs should be presented with their limitations, and users should be told that automated results cannot replace professional assessment. A risk classifier result should not be used as the sole basis for determining whether an individual is in danger.

Data minimization should be applied so that the application collects only information necessary for the intended workflow. Users should be informed about how their text, audio, questionnaire responses, and analysis history are stored and used. Access should be role-based, and sensitive records should be protected in transit and at rest. Retention and deletion rules should be explicit, and audit logs should record access to sensitive records without unnecessarily duplicating their contents.

A production deployment would require review by relevant clinical, legal, and institutional stakeholders. Any evaluation involving patients or sensitive mental-health information should follow appropriate ethics approval, consent, and data governance procedures. Human review should remain central to any safety-related response.

XII. Future Work

Future work should prioritize validation, robustness, and transparent reporting. First, the component models should be evaluated on larger and more diverse datasets. Text emotion evaluation should include per-class metrics, multi-label alternatives, and error analysis. The safety classifier should be tested with external data, threshold analysis, calibration, and explicit measurement of false-negative and false-positive rates. Speech evaluation should use speaker-independent partitions and additional datasets with varied recording conditions.

Second, the fusion mechanism should be evaluated using an independently labeled multimodal dataset and a prespecified protocol. Appropriate baselines include each individual modality, simple weighted fusion, and learned fusion. The study should measure performance when modalities are missing, compare different normalization schemes, and assess whether the fused output provides measurable value beyond the strongest single component.

Third, the speech pipeline could be improved through noise augmentation, speaker normalization, self-supervised representations, and tests across accents and languages. Text modelling could explore multi-label prediction and better calibration. However, any added modality, such as facial affect, should be justified by measurable benefit and balanced against additional privacy and bias risks.

Fourth, the application should undergo usability evaluation with intended users and clinicians. Security testing should include access-control verification, penetration testing, retention and deletion checks, and audit-log review. Any clinical study should involve qualified professionals, appropriate ethical review, and a clear process for human assessment of safety alerts. Until these steps are completed, NeuroAI should remain a research prototype.

XIII. Conclusion

This paper presented NeuroAI, a multimodal framework integrating text emotion recognition, separate suicide/selfharm risk screening, speech emotion recognition, deterministic PHQ-9 scoring, weighted fusion, and clinician-oriented application workflows. The reported component results were 56.50% accuracy and 55.92% weighted F1-score for 28-class text emotion recognition, 97.38% accuracy and F1-score for binary safety screening, and 76.28% held-out validation accuracy for six-class speech emotion recognition. The framework uses configurable weights, missing-modality renormalization, and rule-based safety overrides to make the combined workflow more inspectable.

The main contribution is the integration of heterogeneous model outputs and structured questionnaire scoring into one software workflow, rather than a claim of clinical validation. Because the fusion engine has not been evaluated on an

independent labeled multimodal clinical dataset, no overall system accuracy is reported. NeuroAI should be considered a non-diagnostic research and decision-support prototype. Future work must establish external validity, measure fusion performance, assess robustness and fairness, and evaluate safety under qualified human oversight before clinical deployment.

ACKNOWLEDGMENT

The authors may acknowledge project mentors, institutional support, or other contributions here, provided that the information is accurate and permission has been obtained where needed.

REFERENCES

[1] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040–4054.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[3] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.

[4] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[5] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[6] Y.-H. H. Tsai *et al.*, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

[7] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a Brief Depression Severity Measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.