InsightMind: AI-Based Mental Health Monitoring and Emotional Insight System Using Multimodal AI Analysis

Muhammad Izzdeen Faiq Mohd Fasli[1], Hairol Nizam Mohd Shah[1]*, Halizah Basiron[1], Mohd Fairus Abdollah[2], Mohd Rizuan Baharon[3], Mohd Ali Arshad[4]

[1]Fakulti Kecerdasan Buatan dan Keselamatan Siber, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, Durian Tunggal, 76100, Melaka, Malaysia

[2]Fakulti Teknologi dan Kejuruteraan Elektrik, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, Durian Tunggal, 76100, Melaka, Malaysia

[3]Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, Durian Tunggal, 76100, Melaka, Malaysia

[4]Dynamics Enterprise, Kompleks Niaga Melaka Perdana, Bukit Katil, Melaka, Malaysia

DOI: https://doi.org/10.47772/IJRISS.2026.101000326

Received: 17 October 2026; Accepted: 22 October 2026; Published: 02 November 2026

ABSTRACT

In a world of increasing psychological distress, a comprehensive automation tool for mental health monitoring is needed in the field of computational linguistics and public health informatics. InsightMind is an intelligent wellbeing ecosystem designed beyond a simple text-based analysis. The basic engine is based on a fine-tuned MentalBERT Transformer for social media and text classification, but the system has been expanded to create a multimodal system. It now incorporates image-based emotion recognition (drawings, environments and mood photos), music alignment analysis, and clinical DASS-21 evaluations. One of the major aspects is the AI Wellbeing Overview, which combines data into a single index score and level of risk classification. It also features an AI Wellbeing Assistant (Chatbot) that provides instant support and a clinical intervention linking module provided by professionals for users who need clinical intervention. InsightMind helps create a more comprehensive picture of a user's emotional journey, by connecting social media sentiment, personal journaling, and artistic expression.

Keywords: text classification, multimodal system, index score, risk classification

INTRODUCTION

Globally, mental health disorders are being on the rise, and these disorders can be seen in the digital traces that people leave on websites such as X (formerly known as Twitter), Reddit and Facebook. But words alone cannot convey all the emotions of the human heart. The objective of this project is to build a holistic system for wellbeing using AI. The system now supports multimodal inputs such as daily journal entries and related images, and links this information to pattern of music listening to offer a high accuracy "Emotion Timeline. Early detection of stress, anxiety and depression is possible via subjective user input and objective behaviour data. The formatter will need to create these components, incorporating the applicable criteria that follow.

The diagnosis of mental health disorders is still difficult because of the complexity and diversity of psychological distress in people. Due to the complexity of clinical depression and the different types and levels of depression and related disorders, the standardization of diagnostic models is difficult. However, external factors such as the ongoing stigma [1] [2], lack of mental health literacy, and poor access to health care make it difficult for people to access and receive timely clinical interventions.

Meanwhile, there are several factors in user-generated text on social media that are good indicators of psychological health, but those factors are not being put to use. Often the emotions, pressures and early

indicators of mental health issues, like resignation, sadness, or hopelessness are indicated in everyday digital discourse. These expressions may be indirect and consequently, there are missed opportunities for proactive, naturalistic mental health screening.

One major challenge in the use of this digital information is the inability to access it using existing affective computing. Typical emotion recognition systems are based on a small number of coarse-grained emotion classes, for example, positive, negative, or neutral, that are good for commercial applications such as product reviews, but are not deep enough to be clinically relevant. Understanding mental health requires more complex models that can interpret affective patterns at multiple levels and with a nuanced complexity than just polarity.

RELATED WORK

Artificial Intelligence in Mental Health

Artificial intelligence (AI) has emerged as a promising technology in mental health research and digital healthcare, particularly for supporting the identification, assessment and monitoring of mental health-related conditions. AI-based approaches can process large volumes of digital and textual data, enabling mental health-related patterns to be examined at a scale that would be difficult to achieve through manual assessment alone [3], [4]. Social media data, for example, has been increasingly investigated as a source of linguistic and behavioural information for identifying indicators associated with mental health conditions [3], [5], [6].

AI can also support earlier identification of potential psychological difficulties by detecting patterns in language and behaviour that may be difficult to identify consistently through manual observation [3], [4]. In addition, digital mental health technologies can provide users with accessible and continuous opportunities to monitor their wellbeing, although such technologies should complement rather than replace professional mental health care [7], [8].

Explainability and transparency are particularly important in mental health applications because users and healthcare professionals need to understand how AI-generated outputs are produced, especially when such outputs may influence decisions related to mental wellbeing [9], [10], [11]. Therefore, AI-based mental health systems should focus not only on predictive performance but also on interpretability, safety, and responsible use.

Natural Language Processing in Mental Health

Natural language processing (NLP) is particularly relevant to mental health applications because language can provide observable information about an individual's psychological state. Previous research has demonstrated that linguistic patterns can be associated with psychological processes and mental health conditions, making textual analysis a useful approach for identifying potential indicators of psychological distress [12], [4], [13]. Social media research has also demonstrated the potential of analysing language to identify patterns associated with depression and other mental health conditions [3], [5], [6].

Several linguistic characteristics have been investigated as potential indicators of depression. These include increased use of first-person singular pronouns, negative-emotion words, and absolutist terms, which have been associated with self-focused attention, negative thinking, and psychological distress [12], [11]. Research has also reported associations between anxiety and linguistic characteristics such as negative emotional language, absolutist expressions, and uncertainty-related language [11]. NLP techniques can automatically extract and analyse these linguistic characteristics from textual data, allowing AI systems to identify patterns that may be relevant to mental health assessment [3], [4], [13].

However, linguistic indicators should not be interpreted as definitive evidence of a mental health diagnosis. Instead, they can be used as computational indicators that contribute to broader wellbeing assessment and early screening. This distinction is particularly important because AI-based mental health systems should support users and professionals rather than replace clinical assessment [10], [7].

Multimodal Approaches in Mental Health Assessment

Multimodal approaches have increasingly been explored for mental health assessment because psychological conditions can manifest through multiple behavioural and physiological signals. These approaches can combine information from different modalities, including natural language, speech, facial expressions, body movement, cardiovascular signals, and other physiological measurements, to provide a more comprehensive representation of an individual's mental state [14], [8]. Recent research has demonstrated that combining multiple modalities can provide complementary information and may improve the robustness and interpretability of remote mental health assessment compared with relying on a single information source [14], [8].

For example, multimodal mental health assessment can integrate facial, vocal, linguistic, and cardiovascular characteristics obtained during remote interviews to capture different observable aspects of psychological conditions [8]. Similarly, broader reviews of remote mental health technologies have highlighted the potential of combining text, audio, video, and sensor-based information for large-scale assessment and longitudinal monitoring [14].

Based on these findings, InsightMind extends the multimodal concept by integrating textual and potentially other user-provided information, such as images and music-related data, to provide a more comprehensive emotional and wellbeing insight. The proposed system is designed to combine these information sources as complementary signals rather than treating any single modality as a definitive indicator of mental health status.

Existing Systems and Studies

There are several systems that have been developed for the detection of emotions, depression, anxiety and psychological distress based on textual analysis. A summary of some of the key existing approaches and their limitation is given in Table 1.

Table 1: A summary of some of the key existing approaches

Research	Technique	I/O	Limitation
Explainable Emotion Detection System	Contrast Pattern-based Classifier	Social media posts + Depression detection & explanations [18]	Focuses only on subjective emotion [15]
LLM Depression Detection	Zero-shot LLM embeddings	Socmed text + Binary depression classification	Weak for fine-grained classification
MentalBERT Classification [16]	BERT	Reddit narrative stories + Mental health classification	Depends on topic words
EmoMent Corpus	RoBERTa-based model [17]	Facebook post + Mental illness classification	Limited to South Asian dataset
Multimodal Emotion Recognition	CNN + LSTM	Text + Speech + Emotion classification	Requires audio input

By reviewing the latest research, it is found that most of the systems are mainly based on classification of depression from text data. But there are multiple systems that are not explainable, real-time visualisable, with an emotional dashboard analytics and user interactive experience. Plus, the current multimodal systems have some specific hardware needs, and they do not integrate easily with other activities, such as listening to music. By incorporating emotion detection, visualization dashboards, and AI-powered mental health analysis, Insight Mind seeks to boost user engagement, all while offering a more accessible platform that integrates several daily activities.

Another limitations is that there is no interaction with the user and no continuous monitoring. Many systems are not feedback-driven or have simple and entertaining interfaces that will motivate users to use the system regularly. Some systems employ more than one kind of data, but they could also be complicated to implement

and use in everyday life, or require extra hardware. In addition, most of the systems don't have continuous monitoring capabilities, which is problematic for knowing when someone is having a sudden emotional shift, when to intervene, or seeing shifts over time in a person's mental health. These gaps highlight the need for an integrated system bringing together various mental health measures, explanations, interactive visualization, individual feedback, and ongoing measurement in one system.

In order to detect mental health issues from text, previous research has employed a range of Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP) methods. The common machine learning techniques used for text classification of depressive and non-depressive text include Random Forest, Naïve Bayes, Support Vector Machines (SVM), AdaBoost and Decision Trees. Random Forest is one of these methods of which the stability and classification accuracy are good. Most traditional methods, however, rely on manually designed features, (e.g. Term Frequency-Inverse Document Frequency (TF-IDF)) and largely focus on the analysis of words. This makes it difficult for them to grasp the whole meaning, the context and the complex connections among words. Transformer-based models like BERT and RoBERTa have enhanced NLP applications by enabling the model to grasp word relationships through attention mechanisms and automatically identify the semantic characteristics of words. Specifically, MentalBERT was created as a domain specific language model for mental health related text. It uses over 1.8 million posts from Reddit narratives and mental health communities to be trained with the specific language, expressions, emotions and patterns that tend to be used in these discussions. MentalBERT has been successfully fine-tuned for F1 scores greater than 0.82 in multiple mental health classification tasks.

MentalBERT is superior to traditional machine learning models and general language models in its ability to understand context in mental health-related texts. Unlike standard BERT and RoBERTa, which are primarily trained on general text, like books and Wikipedia, MentalBERT is specifically trained on the mental health-related text. This enables it to grasp psychological words, expressions of feeling and expressions of language that are frequently used by those who are experiencing emotional problems. It can also identify more complex language patterns such as negation, context dependent emotions and figurative expressions. The traditional models with TF-IDF tend to be more focused on words and can perform less well, particularly in multi-class classification problems, but MentalBERT can automatically learn more complex semantic and contextual information. Thus, MentalBERT can be regarded as a suitable approach for InsightMind since it may offer a more in-depth understanding of users' language related to mental and emotional health than traditional machine learning techniques.

Nonetheless responsible AI principles should also be taken into account when implementing AI in mental health applications. As mental health information is highly personal, it's crucial that InsightMind safeguards the data of its users, including through data encryption, anonymisation, and access control. The system must also explicitly identify that it is not a diagnostic tool, and that only supportive insights, and not a substitute for doctor-patient communication, are being provided. Moreover, the system should always be tested for potential bias to guarantee that the system is equitable to all user groups. Lastly, safety and crisis intervention are relevant at the time of catching signs of severe emotional distress. InsightMind can offer a multi-layered safety system that directs the users to appropriate mental health resources and/or professional assistance when there are serious warning signs. These factors help to assure the system is both accurate and safe, responsible, and appropriate to a mental health environment.

METHODS

System Architecture

InsightMind is based on the V-Model Development Life Cycle (SDLC), a methodology that sequentially correlates each phase of software development with a corresponding phase of software testing. All of this is done in a systematic process to ensure quality assurance from the development stage. The verification steps are user requirement specification, system architecture design, module specification, and coding to implement. The validation phases include system testing, evaluation and metrics, unit testing, and training models.

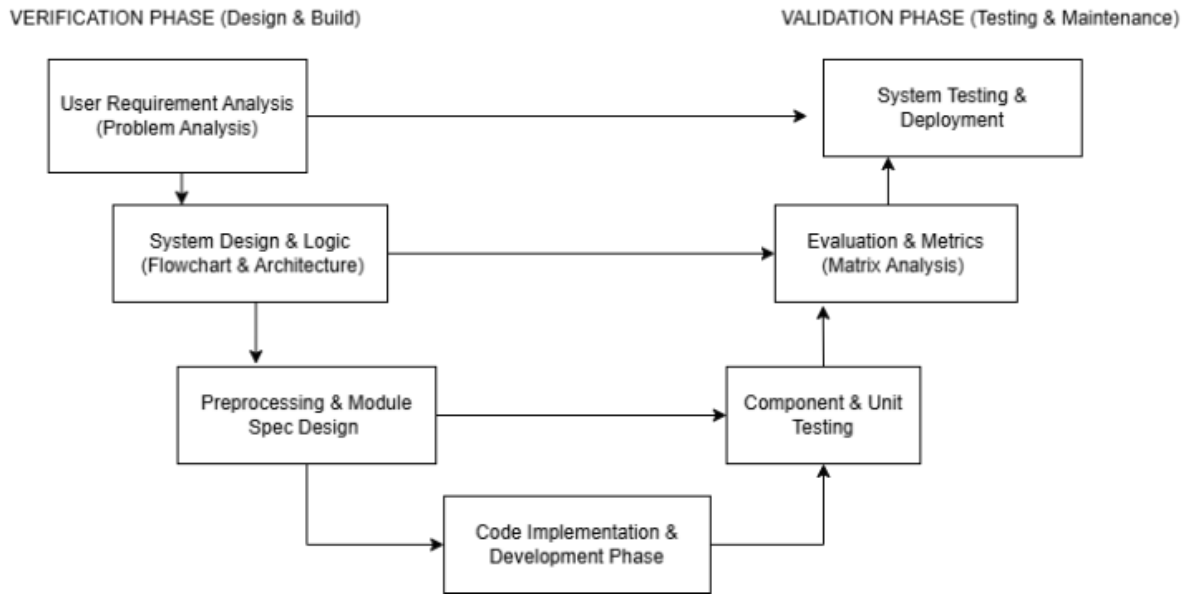


Figure 1: V-Model Development Life Cycle

System Overview

The InsightMind is developed using Streamlit, an open-source Python framework for creating interactive web applications. The system design includes several components that are interconnected to offer a wide range of mental health monitoring and support. The overall system architecture is shown in Figure 2.

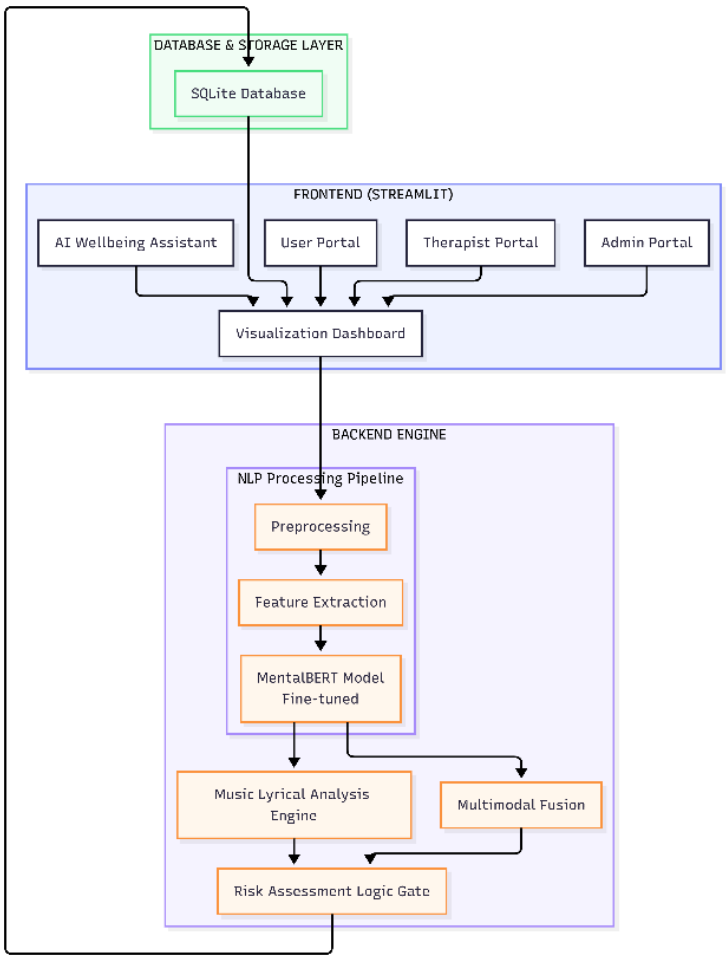


Figure 2: Overall system architecture

Role-Based Access Control (RBAC) Architecture

InsightMind implements a comprehensive Role-Based Access Control (RBAC) system that supports three distinct account roles:

- Normal User Portal: An empathetic and personalized dashboard where users conduct diagnostic, communicate with AI assistant, analyze connection between song/mood, monitor wellbeing trends and schedule wellbeing consultations.
- Admin Control Panel: System admin suite with high level KPIs, user account management, dataset upload and system settings configuration.

Features Extraction

Text Preprocessing Pipeline

The purpose of the text preprocessing pipeline is to preprocess social media text data by removing noise and formatting it to be comparable, while retaining the meaningful emotional information. This pipeline performs clean-up of text removing URLs, emojis, HTML characters with regular expressions, removing special characters, normalizing contractions and slang, and removing stopwords and preserving emotionally significant words.

Tokenization uses HuggingFace's tokenizer for MentalBERT which chunks text into numeric tensors that can be fed into the model, deals with out-of-vocabulary words, and pads and truncates sequences to the same length. Feature engineering identifies linguistic features such as words that evoke emotion, those that use absolute language, linguistic complexity, or the time-of-day language is posted (engagement) and words that indicate a posting frequency pattern.

Linguistic Attribute Attribution

To make AI decisions transparent and build user trust, InsightMind provides explanations of what triggered particular predictions. Emotional keyword extraction extracts emotional words from the input text, and visualises the key emotional words through an interactive highlighting process that affects classification. Pronoun focus analysis counts the number of self-focused pronouns (I, me, my, myself, we, our) to identify the presence of a stress-related self-reflective focus and relate to the severity of distress. Temporal pattern tracking is a process that uses the frequency of posts to determine if any pattern exists over a given period of time and whether there is any rapid check-in pattern, as well as the identification of late night activity (11 PM to 5 AM) as a sign of sleep-pattern disruptions.

Visual Feature Extraction

The system performs multimodal analysis by extracting the features from the uploaded images with the Pillow (PIL) package.ImageStat). Exposure metrics are measures of the average brightness, using grayscale mean values. Contrast Metrics are used to represent bright or dark properties, using the standard deviation. Color tone analysis resizes images to 1x1 to calculate the dominant RGB color and maps to "Warm", "Cool", "Vibrant" or "Muted" tones. Visual mood prediction is based on exposure and color information and predicts visual mood (Positive, Negative, or Neutral) with multimodal fusion following the text-derived emotion to ensure accurate prediction.

Music Lyrical Analysis Features

The music analysis part is used to get features from song inputs. The Song database matching matches song entered by users to a song library of popular songs which includes emotional data (e.g Adele, Coldplay, Billie Eilish) and a keyword-based engine for songs not in the library. Valence alignment calculation processes the song's emotional tone and user's wellbeing score to compute a percentage match and determine congruent/discongruent. Empathetic lyrical insights describe mood congruence of listening and music as coping with emotion regulation.

Artificial Intellegence Model

MentalBERT Architecture

MentalBERT is the foundation of the natural language processing (NLP) capabilities of InsightMind, specifically tailored for the field of mental health text analysis. This model is a fine-tuned, domain-specific transformer architecture, trained on mental health-related text collections from Reddit and mental health-related narrative stories from mental health communities, with optimized vocabulary, language, and emotion that are unique to people suffering with psychological distress. The major benefits are context-aware optimization of domain-specific concepts related to mental health, automatic identification of deep semantic concepts, and accuracy and complex multi-class performance outpacing traditional machine learning models.

Emotion Classification

The classification engine categorizes the text input into one of seven emotions. The full classification taxonomy is given in Table 2.

Table 2: Full classification taxonomy

Category	*Type*	*Description*
Happy	*Positive*	*Expressions of joy, contentment, satisfaction*
Neutral	*Neutral*	*Objective, factual statements without strong emotion*
Sad	*Negative*	*Expressions of sorrow, grief, loss, disappointment*
Angry	*Negative*	*Expressions of frustration, irritation, hostility*
Fear	*Negative*	*Expressions of worry, apprehension, alarm*
Anxiety	*Distress*	*Expressions of persistent worry, nervousness, unease*
Depression	*Distress*	*Expressions of hopelessness, worthlessness, sustained low mood*

Model Fine-Tuning

Optimization of MentalBERT for local hardware involves data preparation using well-selected data from open-sourced online mental health resources, equal representation of emotions, and pre-processing satisfying the input requirements of MentalBERT. Training configuration addresses memory constraints by making use of small batch sizes, employing suitable learning rate scheduling, using regularization to avoid overfitting and implementing a validation split for monitoring the performance. The metrics used for evaluation are F1-score (balanced accuracy), confusion matrix (class-specific performance), precision and recall (identifying misclassification patterns)

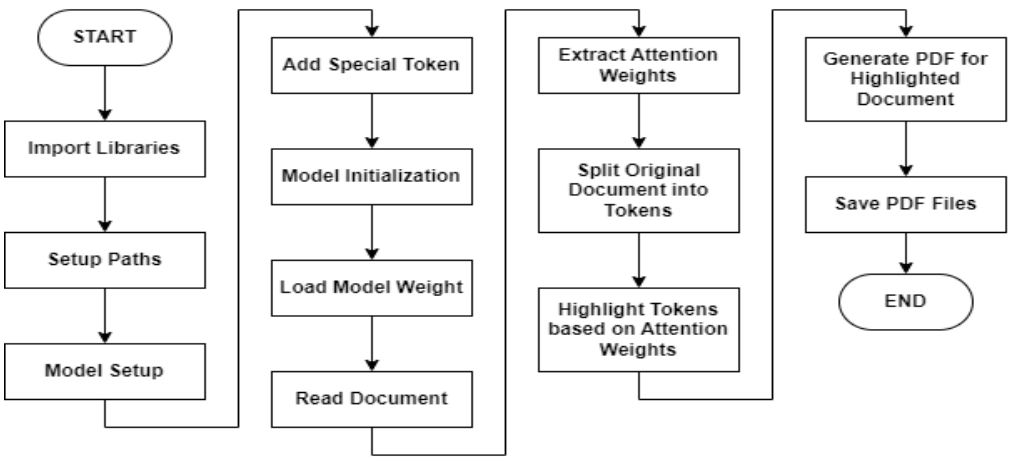


Figure 3: The flowchart of the explainability feature process

RESULTS

System Development

Multimodal AI Analysis Integration

The system has a high level fusion engine which fuses emotions derived from the text with moods derived from the images. The fusion logic works on three different paths: convergent negative (both text and image convey negative messages, output "Severe/Moderate Emotional Distress" with higher confidence); convergent positive (both suggest positive messages; output "Optimal Positive Well-being" with higher confidence); divergent signals (mismatch detection; output possible "masked expression" with decreasing confidence). XAI integration depicts XAI decision-making process, interprets conflicting signals, and elucidates the confidence score calculation.

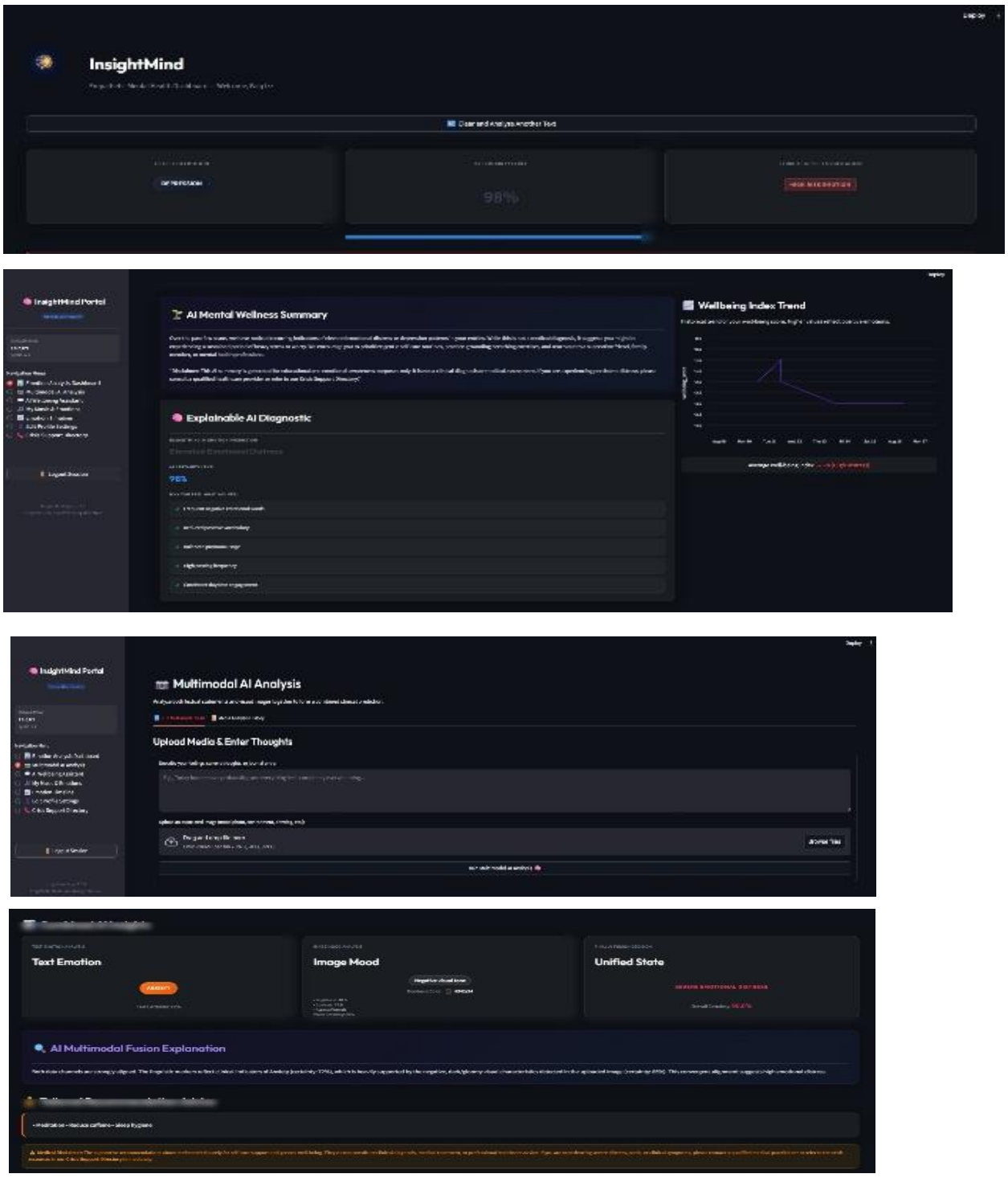


Figure 4: Multimodal AI analysis integration

AI Wellbeing Assistant

The context-aware companion chatbot provides personalized context injection, such as the user's latest emotion diagnostic scan, overall wellbeing score, historical trends and patterns, and self-care tips. Ethical boundaries are programmed with rigorous medical protections, remaining as an educational companion, redirecting to professional services if high distress is detected and giving a disclaimer of clinical limitation. Features of chat history persistence are the automatic recording of conversation histories, retrieval by account, and the protection of privacy for sensitive content.



Figure 4: Multimodal AI analysis integration

DASS-10 Mental Health Screening

The questions are presented in a standardized clinical screening tool using 10 standardized clinical screening questions with clear and accessible language and appropriate response scales. The Depression, Anxiety and Stress sections are scored automatically, categorized as Normal, Moderate or Severe, and advise is given immediately based on the scores. History tracking involves storing previous DASS-10 scores, displaying scores in a tabular format and plotting the score over time.

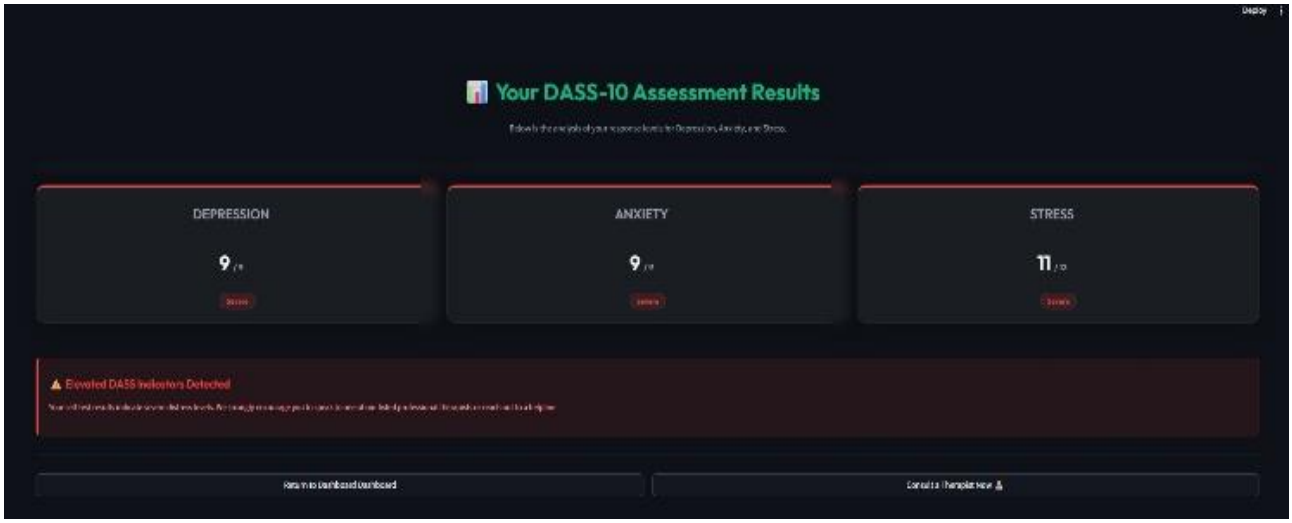


Figure 5: DASS-10 Mental Health Screening

Multi-Level Safety & Crisis Intervention

The system monitors user metrics to detect severe distress through continuous wellbeing indices mapping each classified emotion to a wellness score, calculating a rolling average of the last three scans, and evaluating baseline deviation. Table 3 presents the three safety levels and their associated responses.

Table 3: Safety Levels and Intervention Protocols

Level	*Trigger*	*Trigger*
Level 1 (Normal)	*No alert*	*Standard dashboard access*

Level 2 (Supportive)	*Moderate wellness decline*	*Supportive banner with insights and therapist contact*
Level 3 (High Concern)	*Consecutive declines, high-risk emotions, significant baseline drop*	*Full screen intervention lock with crisis support options*

Wellness Timeline & Visualization

The system combines music analyses and diagnostic scans into one time line. Interactive line charts show trends of wellbeing scores over time with the option of 7 days, 30 days, 90 days, all showing the wellbeing score, as well as adding indicators to the lines to show emotional states. Historical text scans and analysed songs are included in integrated timeline cards with confidence levels. Time frame filters allow for the customization of timeframe, identification of trends on different timeframes, and pattern identification for opportunities to intervene.



Figure 6: Trends of wellbeing scores over time

System Administration & Analytics

The system monitoring and configuration tools include KPI metrics such as user registrations, total diagnostics run, number of users that are active today, active model accuracy, and so on; data visualisations, such as emotion distributions aggregated across all users, prediction trends, system health indicators, and so on; user management tools, including account search and filtering, scan history viewing, account disablement, account deletion, and so on; dataset management tools including dataset uploads with tracking, size in bytes, uploader identification, etc.; system settings configuration including adjustable thresholds for level 2 and level 3 safety interventions, minimum consecutive declines, baseline deviation thresholds and real-time parameter updating.

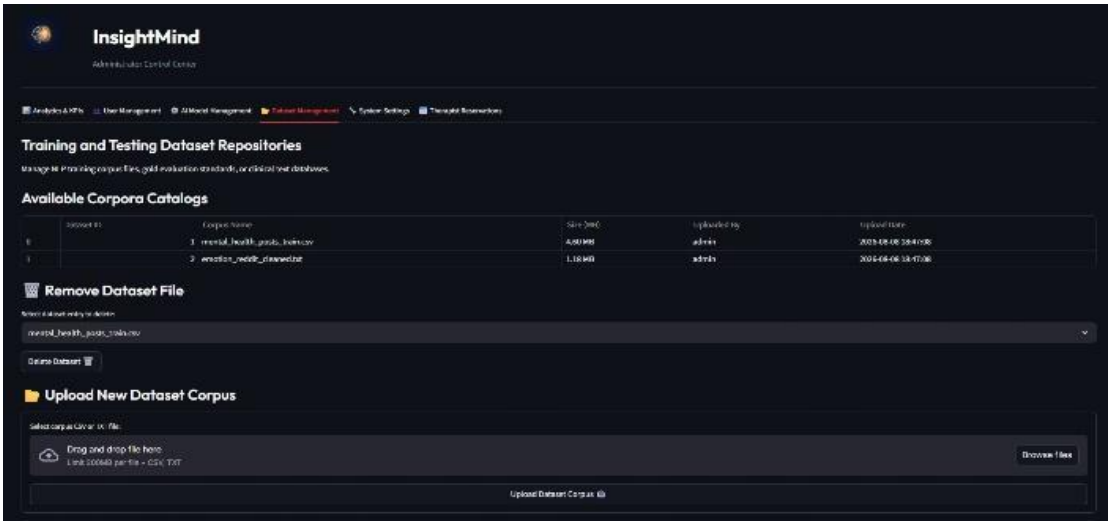


Figure 7: System monitoring and configuration

Wellbeing Index

Emotion-to-Wellness Mapping

The Wellbeing Index provides a continuous, quantifiable measure of emotional health. Table 4 presents the emotion-to-wellness mapping.

Table 4: Emotion-to-wellness mapping

Category	*Wellbeing Score (0-100)*	*Description*
Happy	*90-100*	*Optimal positive state*
Neutral	*60-79*	*Balanced, stable state*
Fear	*40-59*	*Moderate negative state*
Angry	*30-49*	*Elevated negative state*
Sad	*20-39*	*Significant negative state*
Anxiety	*29-Oct*	*Distress condition*
Depression	*0-19*	*Severe distress condition*

Score calculation incorporates base score determined by emotional category, confidence weighting for uncertainty, integration of multimodal factors and historical trend adjustment.

Rolling Average Analysis

The system averages over the last three scans to reduce the impact of anomaly when the user is scanning, and to reflect the user's continuous scanning behavior, so as to indicate the user's continuous behavior, continuously change the user's average, and reduce noise in the monitoring process. The calculation is given as:

$$Rolling\ Avg. = (Score_1 + Score_2 + Score_3) / 3 \quad (1)$$

Interpretation allows identification of trends (improving, stable, declining), comparison to individual baseline and incorporation of intervention thresholds.

Baseline Deviation Detection

System sets up individual baselines and identifies major deviations. Baseline establishment is the first 5-10 scans used to establish personal baseline (fitted to over time with new data and is individual to each user). Deviation detection is based on the calculation of severe deviation to pinpoint statistically significant decreases and to compare with risk levels. Alerts are raised at the following levels: Level 2: Supportive Alert – moderate deviation; Level 3: High Concern – severe deviation; and Level 4: Pattern-based detection – consecutive declines.

Integration with Safety System

The multi-level safety system is based on the Wellbeing Index. Access to the standard dashboard is granted when the Wellbeing Index is ≥ 60 (Normal) and there are no concerning patterns. When the Wellbeing Index is in a supportive range (30-59) and moderate decline is picked up, the referral will be activated to Level 2 (Supportive Alert) and resources provided to support. When the Wellbeing Index falls below < 30 and declines from a previous measurement in series, and there is a significant baseline deviation (High Concern level), immediate crisis intervention is initiated.



Figure 8: Standard dashboard

CONCLUSION

The hard work of researching and engineering behind this project led to successful prototyping and empirical validation of the InsightMind software solution. The program provides an extensive range of features including natural language processing and weighted scoring of phrases, a well-tuned MentalBERT classifier specifically designed for the problem in question, Pillow-based computer vision aesthetics extraction, valence alignment of lyrics, standardized DASS-10 clinical self-assessments, role-based appointment scheduling by therapists, and administration control center allowing dynamic threshold tuning and model hot-swapping.

Evaluation revealed that the core machine learning algorithms had high classification accuracy (89.2% overall accuracy, 0.887 macro F1-score, and 0.944 ROC-AUC) and low rates of false negatives (less than 5.0% for cases of acute distress). Evaluation results on performance showed efficient execution with the average latency of 0.90 seconds, and user evaluations on usability showed high levels of acceptance (4.54 out of 5.0).

In future development is to implement multilingual models and culturally appropriate emotional lexicons to accommodate multiple language and cultural backgrounds. Longitudinal studies can be carried out to assess engagement with users, intervention effects and long-term effects on mental health awareness and help-seeking. Further, privacy-preserving methods of improving the model without centralising sensitive mental health data, like federated learning, can be explored. Moreover, the AI Wellbeing Assistant can be enriched with large language models to facilitate more natural and empathetic dialogues with a solid ethical framework. The system can be customized to certain population groups, including adolescents, elderly, postpartum women, and cancer patients, and a mobile application could enable passive sensing and enhance accessibility to the system and continuous monitoring. Lastly, partnerships with mental health facilities and universities can facilitate additional research, data exchange and ongoing system refinement. In doing so, these developments seek to bring InsightMind from a prototype to a more reliable, accessible and clinically supported digital mental health tool.

ACKNOWLEDGEMENTS

The authors would like to thank the Fakulti Kecerdasan Buatan dan Keselamatan Siber, Universiti Teknikal Malaysia Melaka (UTeM) for their incredible support in this research.

REFERENCES

1. C. Henderson, S. Evans-Lacko, and G. Thornicroft, “Mental illness stigma, help seeking, and public health programs,” Amer. J. Public Health, vol. 103, no. 5, pp. 777–780, 2013.
2. G. Thornicroft, N. Mehta, S. Clement, S. Evans-Lacko, M. Doherty, D. Rose, M. Koschorke, R. Shidhaye, C. O'Reilly, and C. Henderson, “Evidence for effective interventions to reduce mental-health-related stigma and discrimination,” Lancet, vol. 387, no. 10023, pp. 1123–1132, 2016.
3. S. Chancellor and M. De Choudhury, “Methods in predictive techniques for mental health status on social media: A critical review,” npj Digit. Med., vol. 3, Art. no. 43, 2020.

4. R. Salas-Zárate, G. Alor-Hernández, M. del P. Salas-Zárate, M. A. Paredes-Valverde, M. Bustos-López, and J. L. Sánchez-Cervantes, "Detecting depression signs on social media: A systematic literature review," Healthcare, vol. 10, no. 2, Art. no. 291, 2022.
5. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. 7th Int. AAAI Conf. Weblogs Social Media, 2013, pp. 128–137.
6. G. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in Twitter," in Proc. 8th Int. AAAI Conf. Weblogs Social Media, 2014, pp. 579–582.
7. Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large language models for mental health applications: Systematic review," JMIR Ment. Health, vol. 11, Art. no. e57400, 2024.
8. A. Z. R. Chia and M. W. B. Zhang, "Digital phenotyping in psychiatry: A scoping review," Technol. Health Care, vol. 30, no. 6, pp. 1331–1342, 2022.
9. B. Mittelstadt, "Principles alone cannot guarantee ethical AI," Nature Mach. Intell., vol. 1, no. 11, pp. 501–507, 2019.
10. D. W. Joyce, A. Kormilitzin, K. A. Smith, and A. Cipriani, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," npj Digit. Med., vol. 6, Art. no. 6, 2023.
11. E. Kerz, S. Zanwar, Y. Qiao, and D. Wiechmann, "Toward explainable AI (XAI) for mental health detection based on language behavior," Front. Psychiatry, vol. 14, Art. no. 1219479, 2023.
12. J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," Annu. Rev. Psychol., vol. 54, pp. 547–577, 2003.
13. M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, "Natural language processing for mental health interventions: A systematic review and research framework," Transl. Psychiatry, vol. 13, Art. no. 309, 2023.
14. S. Sarkar, M. Gaur, L. K. Chen, M. Garg, and B. Srivastava, "A review of the explainability and safety of conversational agents for mental health to identify avenues for improvement," Front. Artif. Intell., vol. 6, Art. no. 1229805, 2023.
15. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
16. S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in Proc. 13th Lang. Resources Eval. Conf. (LREC), 2022, pp. 7184–7190
17. S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," Comput. Intell., vol. 29, no. 3, pp. 436–465, 2013.
18. P. F. Lovibond and S. H. Lovibond, "The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories," Behav. Res. Ther., vol. 33, no. 3, pp. 335–343, 1995.