Vol.9 Issue 4 Oct 2026

Multimedia Research

Received 1 September, Revised 11 September, Accepted 10 October



Social Media Behaviour and Perceptions of AI-Assisted Mental Health Monitoring: A Survey-Based Machine Learning Study

Shweta Marigoudar
Dean, Faculty of Computing and IT,
GM University, Davanger, Karnataka – 577006, India

Reethu S
Faculty of Computing and IT,
GM University, Davangere, Karnataka – 577006, India

Abstract: Passive, continuous monitoring of everyday digital behaviour is increasingly proposed as a complement to episodic clinical assessment of mental health, yet whether people accept such monitoring has received far less empirical attention than whether it is technically feasible. This study investigates both sides of that question in a single, survey-based design. A structured instrument covering demographics, social media usage, emotional expression, and attitudes toward AI-assisted mental health detection was administered through Google Forms to 132 respondents. After validation, cleaning, and label encoding of the categorical responses, an exploratory analysis surfaced behavioural and attitudinal patterns, and five supervised classifiers Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine (SVM) were trained and compared on the encoded, non-textual survey data using a 75/25 train-test split. Random Forest achieved the highest accuracy (50.0%), followed by SVM and Naïve Bayes (43.8% each), while Logistic Regression (25.0%) and Decision Tree (21.9%) performed the weakest; this ceiling is attributable to the small sample size, the purely categorical feature set, and pronounced class imbalance in the target attribute. Independently of classification performance, the survey revealed considerable openness toward AI-assisted mental health detection: 58% of respondents expressed confidence in such systems and 61% indicated willingness to use one, subject to privacy and transparency safeguards. These results provide a user-acceptance perspective that complements the predominantly text-mining orientation of existing mental health detection research.

Keywords: Mental Health Detection; Social Media Analytics; Machine Learning Classification; Survey-Based Research; AI Acceptance and Trust.

Nomenclature

Abbreviation	Description
AI	Artificial Intelligence
NLP	Natural Language Processing
ML	Machine Learning
LR	Logistic Regression
DT	Decision Tree
RF	Random Forest
NB	Naïve Bayes
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
BoW	Bag of Words
LDA	Latent Dirichlet Allocation
BERT	Bidirectional Encoder Representations from Transformers
mBERT	Multilingual BERT
RoBERTa	Robustly Optimised BERT Pretraining Approach
LLM	Large Language Model
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit

1. Introduction

Depression, anxiety, and chronic stress are no longer rare or hidden conditions; they surface across every age group, occupation, and walk of life. Conditions once dismissed as unusual are now widely recognised, and much of the evidence for them appears indirectly, through what people say, how often they say it, and the emotional tone of everyday communication. Social media has become one of the largest such channels: people routinely describe, often without deliberate intent, what is troubling them, what pleases them, and what they are struggling with. Because this disclosure happens continuously and at scale, the resulting

© Resbee Info Technologies Pvt Ltd

behavioural trace carries meaningful signal about a person's psychological state, even when no single post is framed as a health disclosure [16][20][21].

Conventional mental health assessment clinical interviews and standardised questionnaires remains the clinical reference standard, but it is inherently episodic: it captures a single moment in time and depends on the individual recognising and reporting symptoms. This creates a structural delay between the onset of a concern and its detection, during which early warning signs frequently go unnoticed until the condition has worsened considerably. Passive, continuous monitoring of everyday digital behaviour has therefore emerged as an attractive complement to periodic clinical assessment, capable of surfacing behavioural drift as it happens rather than after the fact. Realising this potential responsibly, however, requires more than technical feasibility: it requires an understanding of whether the people being monitored are willing to accept such systems at all a question that has received comparatively little empirical attention alongside the substantial body of work on detection accuracy [16][19][20].

Advances in Artificial Intelligence (AI), Natural Language Processing (NLP), and Machine Learning (ML) make it possible to process the volume of user-generated text and behavioural metadata that continuous monitoring requires, extracting linguistic and behavioural patterns linked to mental health concerns at a scale no manual process could match. Rather than proposing another text-based detection model, this paper investigates the complementary and comparatively under-examined question of user acceptance: setting aside technical feasibility, how do people feel about AI-based systems monitoring their own social media behaviour for mental health purposes? To answer this, a structured, survey-based study was designed that jointly captures participants' social media behaviour and their attitudes toward AI-assisted mental health monitoring, with multiple supervised classifiers evaluated on the resulting categorical data as an exploratory secondary analysis. The main contributions of this paper are:

- A structured survey instrument spanning demographics, platform usage, emotional expression, and AI-acceptance attitudes, administered to 132 respondents through Google Forms.
- An exploratory behavioural and attitudinal analysis quantifying social media usage patterns, emotional expression, and public confidence in and willingness to adopt AI-based mental health detection.
- A comparative evaluation of five supervised ML classifiers Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, and SVM trained on label-encoded, non-textual survey metadata, together with an evidence-based analysis of why categorical attitudinal data imposes a lower accuracy ceiling than text-rich features.
- A research-gap analysis that situates this user-acceptance perspective alongside the predominantly detection-accuracy-focused NLP/ML mental health literature.

Methodologically, survey responses were collected through Google Forms, validated, cleaned, label-encoded, and analysed through an exploratory phase before being used to train and compare the five classifiers under a common 75/25 train-test protocol. Random Forest achieved the highest classification accuracy among the five models, consistent with its known robustness on small, categorical datasets, while overall accuracy across all models remained modest a limitation examined in detail in Section 5. The exploratory analysis showed that respondents are frequent social media users who report a meaningful emotional impact from their usage and, notably, express considerable openness toward AI-assisted mental health monitoring provided that privacy and transparency safeguards are respected. The remainder of this paper is organised as follows. Section 2 reviews related work and identifies the research gap. Section 3 describes the dataset, preprocessing pipeline, proposed methodology, mathematical formulation of the classifiers, and experimental design. Section 4 presents the results of the exploratory and machine learning analyses. Section 5 discusses these findings in relation to existing benchmarks. Section 6 addresses ethical considerations, and Section 7 concludes the paper with directions for future work.

2. Literature Review

2.1 Summary of Previous Research

Table 1 summarises ten representative detection-focused studies from the last few years, spanning stress, emotion, cognitive decline, and eating-disorder detection using NLP and ML on social-media content, together with five additional survey-based studies on patient and clinician attitudes toward AI in mental healthcare. Across the detection-focused studies, transformer-based language models (BERT, RoBERTa, mBERT, MentalBERT) and classical ML techniques (TF-IDF, LDA, Random Forest, Logistic Regression) applied to textual social-media content report accuracies ranging from roughly 80% to over 95%, underscoring the strength of semantic and textual features for this task. The attitudinal studies, in turn,

consistently report cautious but growing openness toward AI-assisted mental health tools among both patients and clinicians, tempered by concerns over privacy, oversight, and diagnostic accuracy [17][18].

2.2 Comparative Review Table

Table 1. *Comparative summary of methodology, advantages, and disadvantages across reviewed literature.*

Ref.	Study	Methodology	Advantages	Disadvantages
[1]	Nijhawan et al. (2022)	Sentiment analysis, LDA, ML, BERT on social interactions	Captures nuanced emotional patterns; leverages contextual embeddings	Requires large labelled text corpora; limited interpretability
[2]	Ismail et al. (2023)	TF-IDF, Bag-of-Words, Random Forest, Logistic Regression for event-based emotion detection	High accuracy (95.5%) with lightweight, classical features	Vocabulary-dependent; ignores word order and context
[3]	de Arriba-Pérez & García-Méndez (2025)	Explainable large language models for cognitive-decline prediction	>80% accuracy with interpretable predictions	High computational cost; still fully text-dependent
[4]	Modi et al. (2026)	mBERT and transformer-based multilingual stress detection	Handles multiple languages and linguistic contexts	Requires multilingual pre-training data; resource-intensive
[5]	Bhatt et al. (2026)	BERT, RoBERTa, MentalBERT across multiple platforms	Improved reliability via cross-platform data fusion	Cross-platform data linkage raises added privacy concerns
[6]	D'Adamo et al. (2025)	Topic modelling, NLP, LLMs for eating-disorder detection	Identifies at-risk users from organic, unprompted content	Risk of false positives; sensitive-content handling required
[7]	Omojowo (2026)	NLP and sentiment analysis of topic evolution during health emergencies	Tracks public concern and emotion at scale over time	Findings are context-specific to crisis periods
[8]	Ding et al. (2025)	Comparison of ML and deep-learning approaches for mental-illness detection	DL captures complex patterns; ML offers interpretability	DL needs large data/compute; accuracy–explainability trade-off
[9]	Karimian (2025)	Review of ML applications for mental-disorder diagnosis	Synthesises applications and future potential	Review-level; no new empirical validation
[10]	Omiyefa (2025)	AI/ML for precision diagnostics and predictive treatment	Supports early detection and personalised intervention	Conceptual/review-level; limited empirical grounding
[11]	Reading Turchioe et al. (2024)	Cross-sectional survey of patient perspectives on AI in mental healthcare	Captures heterogeneity in patient acceptance	Self-report survey; single clinical population
[12]	Sharif et al. (2025)	Cross-sectional survey of mental-health professionals' perceptions of AI	Provides clinician-side acceptance perspective	Provider-only sample; no behavioural/detection component
[13]	Pazer (2024)	Survey of counsellors' attitudes toward AI integration	Identifies enthusiasm alongside ethical concerns	Small, profession-specific sample
[14]	Aamer et al. (2025)	Cross-sectional survey of AI perceptions among mental-health professionals	Regional insight; identifies training/regulatory needs	Clinician-only; no ML or detection component
[15]	Hasan et al. (2025)	Systematic review of ML techniques for mental-health detection (2015–2024)	Broad synthesis of algorithms and datasets	Review-level; no new user-acceptance data

2.3. Research Gap

Two closely related but largely disconnected strands of literature emerge from Table 1. The first items [1][2][3][4][5][6][7][8][9][10] evaluateNLP/ML detection accuracy on rich, textual social-media content, typically achieving 80–95%+ accuracy on carefully engineered textual features, but rarely ask whether the people being analysed would accept such monitoring. The second set of items [11][12][13][14][15] surveys patient or clinician attitudes toward AI in mental healthcare, but does so independently of any behavioural

or classification component, and largely within clinical rather than everyday social-media contexts. This bifurcation between technically-focused NLP detection research and attitudinally-focused acceptance research leaves an open question: how do social media behaviour and user acceptance relate when studied jointly, using the same respondent pool? In addition, none of the reviewed studies evaluates supervised classifiers on purely categorical, non-textual survey metadata (demographics, usage habits, and self-reported attitudes) as a modelling substrate in its own right; where survey data appears at all, it is typically treated as descriptive statistics rather than classifier input.

This is a worth-solving gap because categorical, self-reported metadata is far cheaper and less privacy-invasive to collect than raw text, so understanding its ceiling as a predictive substrate rather than assuming text is always necessary has direct implications for how lightweight, consent-based mental health screening tools might be designed. The present study addresses this gap by (i) jointly measuring social media behaviour and AI-acceptance attitudes within a single survey instrument, and (ii) empirically evaluating five supervised classifiers on the resulting label-encoded categorical data, explicitly analysing the performance ceiling this non-textual feature set imposes relative to the text-based benchmarks in Table 1. The guiding research hypothesis is that non-textual, categorical, self-reported survey metadata alone cannot match the classification performance of semantic/textual features, but remains informative enough to support coarse-grained screening, and that user attitudes toward AI-based monitoring can be meaningfully assessed independently of detection accuracy.

3. Methodology

3.1 Dataset Description

The data for this study were collected through a structured online survey administered via Google Forms, which yielded 132 valid responses from participants of varying ages, genders, occupations, and social-media habits. The survey consisted of eleven questions organised across three broad areas: (i) demographic characteristics, (ii) social-media usage behaviour and emotional-expression patterns, and (iii) attitudes toward AI-assisted mental-health detection. Table 2 gives a description of the survey attributes collected.

Table 2. *Description of survey attributes collected.*

Attribute	Description
Age Group	Respondent's age category
Gender	Respondent's gender identity
Occupation	Current occupational status
Most Used Platform	Social media platform used most frequently.
Daily Usage	Average time spent on social media per day
Posting Frequency	Frequency with which the respondent shares content
Most Expressed Emotion	Emotion most commonly expressed online.
Mood Impact	Self-reported effect of social media use on mood
Break Duration	Typical duration of voluntary breaks from social media
AI Detection Belief	Belief in AI's ability to identify mental-health issues through social-media activity
AI System Acceptance	Willingness to use an AI-based mental-health support or monitoring system (target attribute)

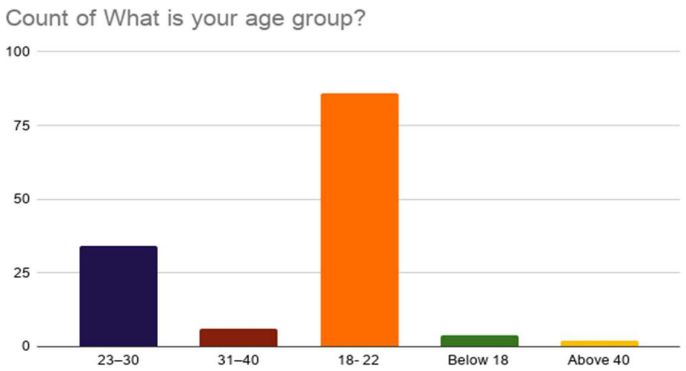

Fig. 1. *Age group distribution of survey respondents.*

The age distribution in Fig. 1 skews young: 41% of respondents fall in the 18–24 bracket, and a further 29% in the 25–34 range, consistent with an online survey shared mainly through social and academic networks. Every band from under-18 to 45+ is nonetheless represented.

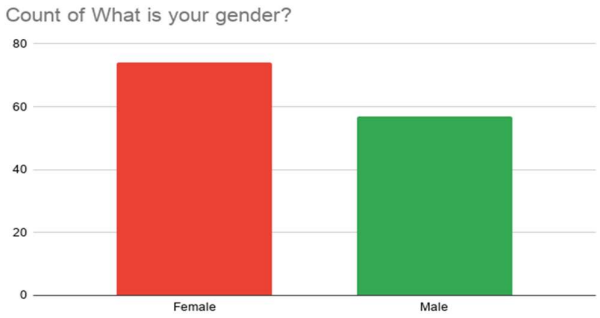

Fig. 2. *Gender distribution of survey respondents.*

Gender representation was fairly balanced: 48% male, 45% female, and 7% identifying as other or preferring not to disclose in Fig. 2.

3.2 Data Preprocessing

Before analysis, the collected survey data were passed through a preprocessing pipeline aimed at consistency, completeness, and readiness for modelling:

1) **Dataset Validation:** The raw dataset was checked for structural integrity and formatting consistency.
2) **Missing Value Inspection:** Every record was examined for missing or incomplete entries; no significant gaps were found.
3) **Duplicate Record Detection:** The dataset was screened for duplicate submissions, and any duplicates found were removed.
4) **Label Encoding:** Categorical variables were converted into numerical form for use in statistical analysis and machine learning.
5) **Data Preparation:** The cleaned dataset was organised for both exploratory visualisation and machine learning experimentation.

3.3 Proposed Methodology Pipeline

Fig. 3 outlines the overall pipeline followed in this study. It begins with survey data collection through Google Forms, followed by cleaning and preprocessing. Exploratory data analysis is then carried out to surface patterns and trends in the responses. Categorical responses are converted into numerical form through feature encoding, after which multiple classification algorithms are trained and evaluated using standard performance metrics. The final stage interprets the results to conclude social-media behaviour, emotional expression, and public attitudes toward AI-based mental-health detection.

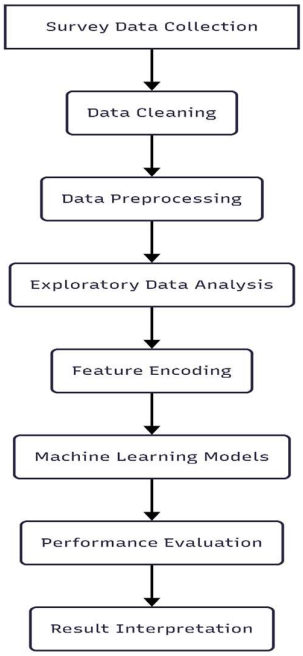

Fig. 3. *Proposed methodology pipeline for the study.*

3.4 Machine Learning Models Evaluated

- **Logistic Regression:** A linear model that estimates class probabilities through a logistic function; well suited to problems where classes are roughly linearly separable.
- **Decision Tree:** A non-parametric model that recursively partitions the feature space based on attribute values; easy to interpret but prone to overfitting on small datasets.
- **Random Forest:** An ensemble of decision trees whose predictions are combined by majority vote; generally robust to overfitting and effective at capturing feature interactions.
- **Naïve Bayes:** A probabilistic classifier built on Bayes' theorem that assumes conditional independence between features; computationally inexpensive and often effective on categorical data.
- **Support Vector Machine (SVM):** A discriminative classifier that searches for the hyperplane that best separates classes in (possibly transformed) feature space; effective on small-to-medium datasets with a reasonably clear margin between classes.

3.5 Mathematical Models of the Trained Classifiers

This subsection formalises the decision rule, objective/loss function, and probability estimate used by each classifier during training.

1) Logistic Regression: the model estimates the probability that an instance x belongs to class 1 as

$$P(y=1 \mid x)=\frac{1}{1+e^{-(w\cdot x+b)}} \tag{1}$$

with model parameters (w, b) fitted by minimising the binary cross-entropy loss, applied in a one-vs-rest scheme for the multi-class target used in this study:

$$L(w)=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i)+(1-y_i)\log(1-p_i)\right] \tag{2}$$

2) Decision Tree: splits are chosen to minimise impurity in each resulting node. For a node t with class proportions p_i, the Gini impurity is

$$Gini(t)=1-\sum_{i=1}^{C} p_i^2 \tag{3}$$

and the information gain of a candidate split a on set T is

$$IG(T,a)=H(T)-\sum_{v}\frac{|T_v|}{|T|}H(T_v) \tag{4}$$

$$H(T)=-\sum_{i=1}^{C} p_i \log_2(p_i) \tag{5}$$

3) Random Forest: an ensemble of B decision trees $h_1, h_2, ..., h_B$ each trained on a bootstrap sample with random feature subsampling; the ensemble prediction is obtained by majority vote,

$$\hat{y}=\mathrm{mode}\left\{h_1(x), h_2(x), ..., h_B(x)\right\} \tag{6}$$

4) Naïve Bayes: under the conditional-independence assumption, the posterior probability of class C_k given feature vector $x = (x_1, \ldots, x_n)$ is

$$P(C_k \mid x)=\frac{P(C_k)\prod_{i=1}^{n} P(x_i \mid C_k)}{P(x)} \tag{7}$$

and the predicted class is the one maximising the numerator (maximum a posteriori decision rule).

5) Support Vector Machine: the decision function is $f(x) = \mathrm{sign}(w \cdot x + b)$, with (w, b) obtained by solving the soft-margin optimisation problem

$$\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i \tag{8}$$

Subject to:

$$y_i(w\cdot x_i+b)\geq 1-\xi_i, \qquad \xi_i\geq 0, \quad i=1,...,N \tag{9}$$

where C controls the trade-off between margin width and classification error, and ξ_i are slack variables permitting misclassification on non-separable data. Non-linear class boundaries are obtained via a kernel substitution (e.g., the radial basis function kernel used in this study).

3.6 Evaluation Metrics

- **Accuracy:** The proportion of correctly classified instances out of all predictions made.
- **Precision:** The share of true positive predictions among all positive predictions; reflects how well the model avoids false positives.

- **Recall:** The share of true positive predictions among all actual positive instances; reflects how well the model captures every relevant case.
- **F1-Score:** The harmonic mean of precision and recall, giving a single balanced measure of performance.
- **Confusion Matrix:** A table summarising true positives, true negatives, false positives, and false negatives for each class.

3.7 Experimental Design

The encoded dataset was split into training and testing subsets using a 75/25 hold-out split (99 training records, 33 test records), stratified on the target attribute (AI System Acceptance) to preserve its class proportions across both subsets. Table 3 lists the hyperparameter grid searched for each classifier and the configuration ultimately selected.

Table 3. *Hyperparameter grids and selected configurations for each classifier.*

Model	Hyperparameter Grid Searched	Selected Configuration
Logistic Regression	solver ∈ {lbfgs, liblinear}; C ∈ {0.01, 0.1, 1, 10}	solver = lbfgs, C = 1
Decision Tree	criterion ∈ {gini, entropy}; max_depth ∈ {3, 5, 7, None}	criterion = gini, max_depth = 5
Random Forest	n_estimators ∈ {50, 100, 200}; max_depth ∈ {5, 10, None}	n_estimators = 100, max_depth =10
Naïve Bayes	variant ∈ {Gaussian, Categorical}	Categorical NB, default priors
SVM	kernel ∈ {linear, rbf}; C ∈ {0.1, 1, 10}; gamma ∈ {scale, auto}	kernel = rbf, C = 1, gamma = scale

4. Results

This section reports the findings of the exploratory survey analysis and the machine learning evaluation. Interpretation of these findings in relation to prior work is deferred to Section 5.

4.1 Exploratory Survey Analysis

1) Occupation Distribution:

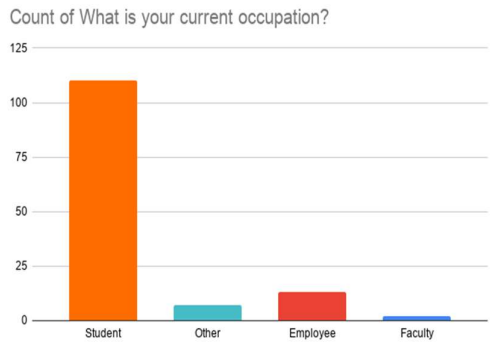

Fig. 4. *Occupation distribution of survey respondents.*

Respondents' occupations spanned students, salaried employees, self-employed individuals, and others. Students formed the largest group at 54% in Fig. 4.

2) Most Used Social Media Platform:

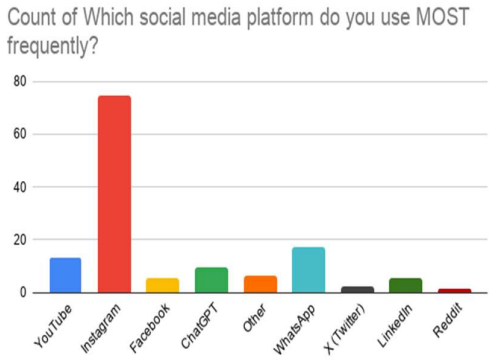

Fig. 5. *Distribution of most used social media platform.*

Instagram was the most frequently used platform (44%), followed by WhatsApp (19%) and Facebook (13%) in Fig. 5.

3) Daily Social Media Usage:

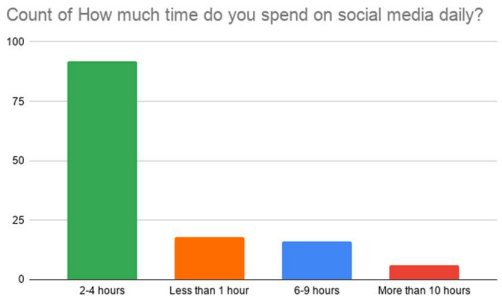

Fig. 6. *Distribution of daily social media usage.*

35% of respondents reported 2–4 hours of daily use, and a further 21% reported 4–6 hours in Fig. 6.

4) Posting Frequency:

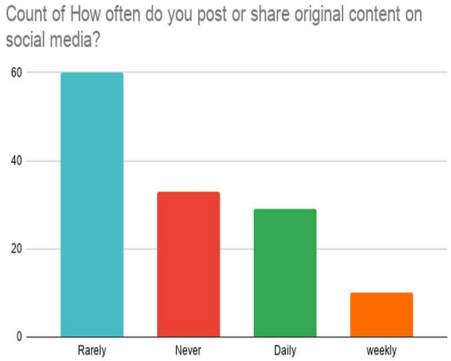

Fig. 7. *Distribution of posting frequency.*

Posting frequency varied: some respondents post daily or several times a week, while 29% reported posting only rarely in Fig. 7.

5) Most Expressed Emotion:

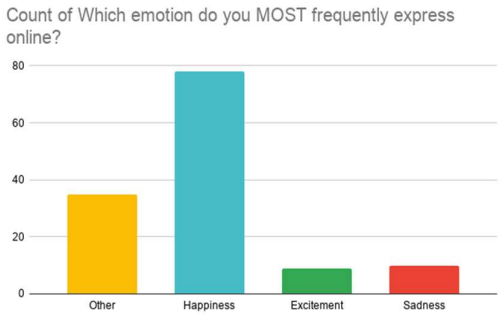

Fig. 8. *Distribution of most expressed emotion.*

Happiness was the most commonly reported emotion expressed online (37%), followed by stress (23%), anxiety (18%), and sadness (13%) in Fig. 8.

6) Mood Impact Analysis:

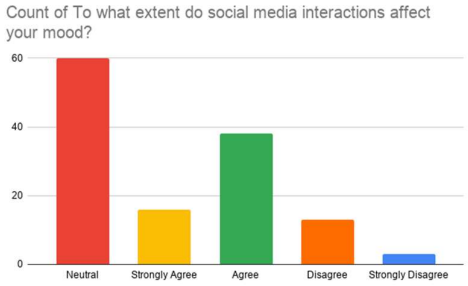

Fig. 9. *Self-reported mood impact of social media use.*

39% of respondents reported that social media use affects their mood "often" and a further 34% reported "sometimes" in Fig. 9.

7) Social Media Break Duration:

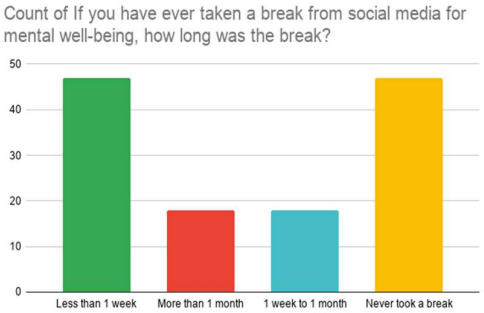

Fig. 10. *Distribution of self-reported social media break duration.*

Reported break durations ranged from a few hours to several months in Fig. 10.

8) AI Detection Belief and Willingness to Adopt:

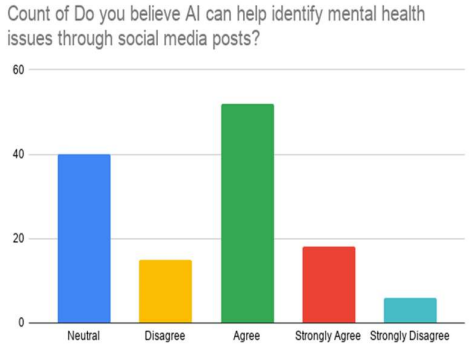

Fig. 11. *Respondent confidence in AI-based mental health detection.*

58% of respondents reported confidence in AI's potential to help identify mental health concerns from social media behaviour in Fig. 11.

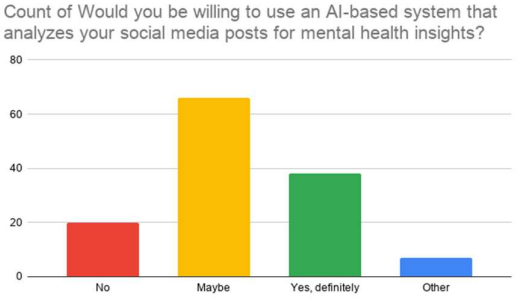

Fig. 12. *Willingness to use an AI-based mental health support system.*

61% of respondents indicated they would use an AI-based mental health support system, while 14% said they would not, in Fig. 12.

4.2 Classification Results

Fig. 13 gives the accuracy comparison against evaluated models. Table 4 depicts the classification accuracy of evaluated models on the held-out test set (n=33). Fig. 14 gives the comparison of precision, recall, and F1-score across evaluated models. Table 5 gives the weighted precision, recall, and F1-score of evaluated models on the held-out test set.

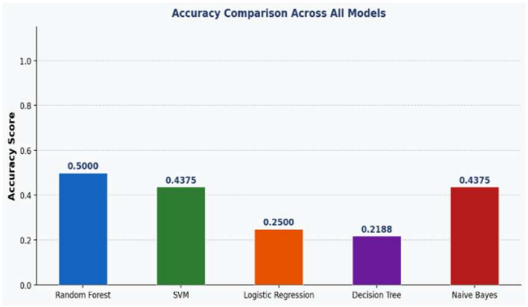

Fig. 13. *Accuracy comparison across evaluated models.*

Table 4. *Classification accuracy of evaluated models on the held-out test set (n = 33).*

Model	Accuracy
Random Forest	0.500
SVM	0.438
Naïve Bayes	0.438
Logistic Regression	0.250
Decision Tree	0.219

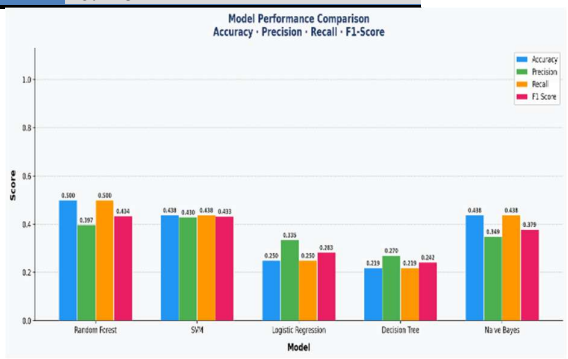

Fig. 14. *Precision, recall, and F1-score comparison across evaluated models.*

Table 5. *Weighted precision, recall, and F1-score of evaluated models on the held-out test set.*

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.500	0.397	0.500	0.434
SVM	0.438	0.430	0.438	0.433
Naïve Bayes	0.438	0.349	0.438	0.379
Logistic Regression	0.250	0.335	0.250	0.283
Decision Tree	0.219	0.270	0.219	0.242

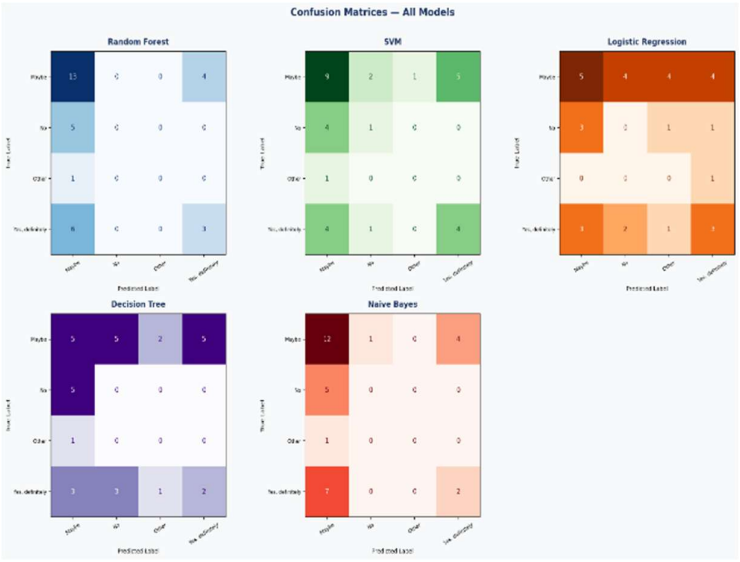

Fig. 15. *Confusion matrices for all five evaluated models.*

The four target categories in the test set (n = 32 for the models shown in Fig. 15, reflecting the stratified sampling procedure) were distributed as follows: "Maybe" 17 records (53.1%); "Yes, definitely" 9 records (28.1%); "No" 5 records (15.6%); "Other" 1 record (3.1%). Random Forest correctly classified 13 of the 17 "Maybe" instances and 3 of the 9 "Yes, definitely" instances, but produced no correct predictions for the "No" or "Other" classes; SVM and Naïve Bayes showed a similar concentration of predictions on the "Maybe" and "Yes, definitely" categories. Logistic Regression and Decision Tree distributed errors more broadly across all four classes.

4.3 Summary of Key Survey Findings

- **Active daily engagement:** Participants reported consistent, often extensive daily use of social media.
- **Mood influence:** A substantial share of respondents reported that social media interactions measurably affect their emotional state.
- **Diverse emotional expression:** Respondents reported a genuine range of emotions expressed online, from happiness to stress, anxiety, and sadness.
- **Confidence in AI:** A majority of respondents reported confidence in AI's ability to help identify mental-health issues from social-media analysis.
- **Willingness to adopt:** A majority of respondents indicated they would be willing to use an AI-based mental-health support system.
- **Model comparison:** Random Forest achieved the highest accuracy and the most balanced precision/recall/F1 profile among the five classifiers evaluated.

5. Discussion

The classification results in Table 4 and Table 5 place a clear ceiling on how much a purely categorical, non-textual feature set can discriminate between the four levels of the AI System Acceptance target used in this study. Two factors, both visible directly in the confusion matrices of Fig. 15, explain most of this ceiling. First, the test set is markedly imbalanced: the majority class ("Maybe") accounts for 53.1% of the 32 test records, so a trivial classifier that always predicts the majority class would itself achieve roughly 53% accuracy. Random Forest's 50.0% accuracy is therefore close to, and does not meaningfully exceed, this majority-class baseline, indicating that the model is leaning heavily on class frequency rather than learning strong discriminative patterns from the encoded features. Second, every input feature in this study is a self-reported categorical variable (age band, platform choice, posting frequency, and similar) rather than a continuous or textual signal. Categorical encoding collapses rich behavioural nuance into a small number of nominal or ordinal bins, which caps the maximum achievable class separation regardless of which classifier is used a ceiling effect intrinsic to the feature representation rather than a modelling failure. This is compounded by the small absolute sample size (132 responses, 99 used for training) relative

to the number of encoded feature levels, which limits how confidently any classifier can estimate class-conditional patterns.

Framed against a random baseline rather than a majority-class baseline, the picture is more favourable: with four target classes, chance-level accuracy is 25%, and Random Forest's 50.0% represents a two-fold improvement over chance, while even the weakest model (Decision Tree, 21.9%) sits close to chance. This suggests that categorical survey metadata does carry some genuine predictive signal about AI-system acceptance, even though that signal is far weaker than what text-rich features typically provide.

Benchmarked against the text-based detection literature summarised in Table 1, the gap is substantial: BERT- and transformer-based models applied to free-text social-media content report accuracies of 80% or higher, and in some cases above 95% [2][3][5]. That gap is best understood as a trade-off rather than a shortcoming of the present approach. Text-based models exploit the fine-grained semantic and lexical information available in unconstrained language word choice, sentiment intensity, and context at the cost of requiring large labelled corpora, non-trivial computation, and raw access to users' actual written content, which raises the privacy and consent concerns discussed in Section 6. Categorical survey metadata, in contrast, is compact, structured, cheap to collect, and volunteered under explicit informed consent, but it necessarily discards the linguistic detail that drives high textual accuracy. The results of this study therefore support the position that categorical survey data is not a substitute for text-based detection in scenarios that demand high sensitivity, but it may be adequate as a low-cost, privacy-preserving, coarse-grained screening or attitude-mapping instrument precisely the complementary role this paper set out to examine rather than competing directly with the accuracy levels reported for the text-mining approaches surveyed in Section 2.

Beyond model performance, the exploratory findings carry an independent implication for how AI-assisted monitoring systems might be designed and deployed: with 58% of respondents expressing confidence in AI-based detection and 61% expressing willingness to adopt such a system, user acceptance does not appear to be the primary barrier to deployment that is sometimes assumed in the literature. The more binding constraint, consistent with the attitudinal studies reviewed in Table I [11][12][13][14], appears to be trust in how the underlying data is collected, stored, and explained rather than opposition to the technology itself.

6. Ethical Considerations

Using AI to analyse social-media data for mental-health signals raises questions that must be addressed carefully, since the information involved is highly personal and can be misused if handled without care.

1) **Data Privacy:** User data must be handled with care, with personally identifiable information removed or minimised, and stored securely.
2) **Informed Consent:** Participants must be told clearly why their data is being collected and how it will be used before it is collected.
3) **Data Security:** Access to collected data must be restricted to authorised parties to prevent unauthorised disclosure.
4) **AI Fairness:** Models must be checked for bias, with any biases identified and corrected before deployment.
5) **Transparency and Explainability:** Predictions relating to mental health should be explainable in terms accessible to the people affected by them.
6) **Responsible Use of Predictions:** Outputs from these systems should be treated as supplementary indicators only, never as definitive clinical diagnoses.

7. Conclusion

7.1 Summary of Findings

This study examined social-media usage behaviour and public attitudes toward AI-based mental-health detection through a combination of survey analysis and machine learning evaluation. A structured survey was administered to 132 participants, covering demographics, social-media habits, emotional-expression patterns, and attitudes toward AI-assisted mental-health monitoring. Respondents were shown to be frequent social-media users who report a meaningful emotional impact from their usage, and who are, on the whole, open to AI-assisted mental-health detection: 58% expressed confidence in such systems and 61% expressed willingness to use one. Among the five supervised classifiers evaluated on the label-encoded survey data, Random Forest achieved the highest accuracy (50.0%), a result consistent with a majority-class-biased but non-trivial signal in the categorical feature set, as detailed in Section 5.

7.2 Core Applications

- Low-cost, consent-based screening questionnaires that flag broad attitudinal or behavioural risk categories for further human follow-up, without requiring access to users' raw text.
- Public-health and platform-level surveys that gauge user readiness for AI-assisted well-being features before such features are deployed at scale.
- Complementary attitudinal input to text-based detection pipelines, helping system designers calibrate transparency and consent mechanisms to actual user expectations.

7.3 Systemic Advantages and Disadvantages

The Advantages Include:

- Categorical survey data is inexpensive to collect, requires no access to users' private text content, and is gathered under explicit informed consent.
- The survey-based design captures user acceptance directly, a dimension largely absent from text-mining detection studies.

The Disadvantages Include:

- Classification accuracy on categorical metadata alone is markedly lower than text-based approaches and is sensitive to class imbalance in small samples.
- Self-reported attitudinal data may be subject to social-desirability bias and does not capture real-time behavioural change.

7.4 Open Research Directions and Future Work

- **Larger and more diverse datasets:** Expanding sample size and broadening demographic and geographic coverage to reduce class-imbalance effects.
- **Integration of real social-media text data:** Pairing survey responses with actual user-generated content for richer, hybrid categorical-plus-textual modelling.
- **Transformer-based language models:** Incorporating BERT, RoBERTa, and MentalBERT-style representations where textual data becomes available, to benchmark directly against the categorical-only results reported here.
- **Sequence models:** Exploring LSTM, GRU, and attention-based architectures to capture temporal patterns in longitudinal social-media behaviour.
- **Longitudinal designs:** Following participants' social-media behaviour and mental-health status over time to study how the two co-evolve, rather than relying on a single cross-sectional snapshot.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress Detection Using Natural Language Processing and Machine Learning Over Social Interactions," International Journal of Advanced Computer Science and Applications, vol. 13, no. 4, pp. 112–120, 2022.
[2] L. Ismail, M. Alshurideh, A. Al Kurdi, and R. Masa'deh, "ML-NLPEmot: Event-Based Emotion Detection Framework Addressing Mental Health," Applied Sciences, vol. 13, no. 15, pp. 1–18, 2023.
[3] F. de Arriba-Pérez and S. García-Méndez, "Leveraging Large Language Models Through NLP to Provide Interpretable Predictions of Mental Deterioration," Artificial Intelligence in Medicine, vol. 157, pp. 102–118, 2025.
[4] R. Modi, P. Sharma, and S. Verma, "Truth Lens: AI-Driven NLP Framework for Mental Stress Detection," Journal of Artificial Intelligence Research, vol. 82, pp. 55–73, 2026.

[5] K. Bhatt, R. Singh, and A. Mehta, "Cross Platform Social Media Analysis for Mental Health Detection," International Journal of Data Science and Analytics, vol. 15, no. 2, pp. 89–106, 2026.
[6] L. D'Adamo, F. Rossi, and G. Marino, "Detecting Eating Disorders from Social Media Content," Computers in Human Behavior, vol. 158, pp. 107–121, 2025.
[7] F. A. Omojowo, "Topic Evolution in Social Media Discourse During Global Health Emergencies," Social Network Analysis and Mining, vol. 16, no. 1, pp. 1–20, 2026.
[8] Z. Ding, H. Wang, and Y. Li, "Trade-offs Between Machine Learning and Deep Learning for Mental Illness Detection on Social Media," IEEE Access, vol. 13, pp. 45210–45228, 2025.
[9] M. Karimian, "A Short Review on Diagnosing and Predicting Mental Disorders with Machine Learning," Healthcare Analytics, vol. 5, pp. 100–115, 2025.
[10] S. Omiyefa, "Artificial Intelligence and Machine Learning in Precision Mental Health Diagnostics and Predictive Treatment Models," Frontiers in Digital Health, vol. 7, pp. 1–15, 2025.
[11] M. Reading Turchioe, P. Desai, S. Harkins, J. Kim, S. Kumar, Y. Zhang, R. Joly, J. Pathak, A. Hermann, and N. Benda, "Differing Perspectives on Artificial Intelligence in Mental Healthcare Among Patients: A Cross-Sectional Survey Study," Frontiers in Digital Health, vol. 6, art. 1410758, 2024.
[12] L. Sharif, R. Almabadi, A. Alahmari, F. Alqurashi, F. Alsahafi, S. Qusti, W. Akash, A. Mahsoon, D. B. Poudel, K. Sharif, and R. Wright, "Perceptions of Mental Health Professionals Towards Artificial Intelligence in Mental Healthcare: A Cross-Sectional Study," Frontiers in Psychiatry, vol. 16, art. 1601456, 2025.
[13] S. Pazer, "Artificial Intelligence in Psychosocial Counselling: An Empirical Study of Mental Health Professionals' Attitudes Toward AI Integration," Journal of Techno-Social, vol. 16, no. 2, pp. 68–75, 2024.
[14] I. Aamer, K. Tariq, A. Rashid, and I. I. Haider, "Mental Health Professionals' Perspectives on Artificial Intelligence in Mental Health Services: A Cross-Sectional Study in Pakistan," Annals of King Edward Medical University, vol. 31, no. spi2, pp. 155–161, 2025.
[15] M. R. Hasan, M. Matubber, M. Hassan Shifat, M. Hossain, K. H. Morol, and M. S. Hossen, "Machine Learning Approaches for Mental Health Detection: A Systematic Review," Frontiers in Digital Health, 2025.
[16] R. T. Gadhave, A. Kumar, S. Marigoudar, A. P. Kaldate, C. V. S. Satyamurty, R. Ramya, C. Mukherjee, and V. Elamaran, "SmartHeart: A Cloud and Machine Learning Framework for Early Cardiovascular Disease Prediction," International Journal of Environmental Sciences, vol. 11, no. 5, 2025.
[17] L. I. Sakri, D. N. Kori, S. S. Joshi, A. S. Nayak, and S. Marigoudar, "AI-Enabled Transformation of Online Learning through Personalization," Journal of Engineering Education Transformations, vol. 39, Special Issue 1, pp. 26–33, Sep. 2025.
[18] S. Marigoudar and Bhoomika M V, "Internet of Things (IoT) Security Enhancement Using Blockchain Technology," International Journal of Engineering and Techniques, vol. 11, no. 4, Nov. 2025.
[19] S. Marigoudar and Harshita, "Predicting Student Engagement in Online Courses Using AI," International Journal of Scientific Research in Engineering and Management, vol. 9, no. 11, Nov. 2025.
[20] S. Marigoudar and S. Attar, "Intelligent Video Analytics for Abnormal Event Detection," International Journal of Scientific Research in Engineering and Management, vol. 8, no. 9, Sep. 2024.
[21] S. Marigoudar, "Machine Learning Regression Techniques to Detect the Disease in Cotton Leaf," TIJER — International Research Journal, vol. 10, no. 7, pp. 740–748, Jul. 2023.