Depression Severity Assessment and Personalized Recommendations Using Structured Data and Random Forest Algorithm

S. Varun
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
Chennai, India
varunsowmyanarayan@gmail.com

Varun K
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
Chennai, India
varunketheboyna@gmail.com

Prince Mary S
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
Chennai, India
princemary.cse@sathyabama.ac.in

***Abstract*—Access to mental health support remains limited for a large section of the population, particularly in regions such as India where psychiatric resources are scarce. Existing screening tools like the PHQ-9 questionnaire, although clinically validated, often reduce a patient's condition to a single aggregate score, failing to capture symptom-level differences in presentation.**

This paper presents MindTrack, a browser-based screening system designed to go beyond score-based evaluation by combining PHQ-9 responses with demographic and lifestyle features such as sleep patterns, physical activity, and self-reported stress. A supervised learning model, based on Random Forest, is used to classify depression severity into four categories: Minimal, Mild, Moderate, and Severe. In addition, the system includes a symptom-level recommendation engine that generates personalized guidance based on individual response patterns rather than aggregate scores.

A safety override mechanism is incorporated to ensure immediate intervention whenever any level of self-harm ideation is reported. The model was trained on a synthetic dataset of 500 samples and evaluated using 5-fold cross-validation, achieving an accuracy of 92.4 percentage and an F1-score of 0.91. While the use of synthetic data limits real-world generalization, the system demonstrates the feasibility of integrating interpretable machine learning with symptom-specific decision support.

The results highlight the potential of combining structured screening tools with machine learning to deliver more personalized and actionable mental health insights.

***Index Terms*—Machine Learning, Random Forest, PHQ-9, Depression Severity, Mental Health Screening, Personalized Recommendation, Streamlit, KNN Imputation.**

I. Introduction

Depression ranks among the most prevalent and disabling health conditions worldwide, yet the gap between its prevalence and the care people actually receive remains staggeringly wide. The World Health Organization estimates that more than 264 million people live with depression worldwide [15]. In India, the scale of unmet needs is particularly acute: around 150 million people require mental health support, while fewer than 30 million seek it [15]. The barriers responsible for this gap are well-documented. Social stigma discourages many from disclosing symptoms or seeking help. A chronic shortage of trained professionals — India has approximately 0.3 psychiatrists per 100,000 population against the WHO-recommended baseline of 3 [15] means that those who reach out may wait months for an appointment. For people in rural or semi-urban areas, a single consultation may require half a day of travel in each direction.

Within the clinical system itself, the tools most widely used for screening carry their own limitation. The PHQ-9 is the established standard, and its validity as a severity measure is well-established [10]. It produces, however, a single numerical total. Two patients who both score 16 can present with completely different clinical pictures, one dominated by persistent sleep disruption, another by recurrent self-harm ideation. A raw sum treats both identically, but their needs and the appropriate next steps diverge significantly. What is absent from most screening workflows is any mechanism for looking at the pattern of responses, not just their aggregate, and translating that pattern into actionable, symptom-specific guidance.

MindTrack was built to address precisely this gap, identifying which PHQ-9 items drive the score and generating symptom-specific recommendations alongside a hard-coded self-harm safety protocol, geolocation facility mapping, and session mood tracking.

The primary contributions of this work are: (1) a Random Forest classifier achieving 92.4% accuracy across four severity categories; (2) a two-stage Recommendation Engine that combines severity-class care plans with item-level symptom trigger identification; (3) an unconditional safety override for any reported self-harm ideation, firing at both submission and results stages; (4) a fully deployed four-page Streamlit

application with geolocation-based facility mapping and session mood trend tracking; and (5) a candid account of the limitations inherent in a synthetic-data prototype, along with a concrete roadmap for clinical translation.

II. Literature Survey

The global burden of depression is well-established. Kessler et al. [7] showed through large-scale surveys that depression accounts for a disproportionate share of years lived with disability and that treatment gaps persist across income levels. Penninx et al. [11] further documented that untreated depression produces measurable physiological changes , disrupted sleep, metabolic dysfunction, and immune dysregulation which correspond directly to the lifestyle features used in MindTrack, providing a clinical rather than merely statistical rationale for their inclusion.

The choice of Random Forest is grounded in consistent evidence across structured health data. Martins et al. [1] demonstrated its robustness on EMG signal classification; Wang [4] applied it to groundwater prediction — a structurally analogous incomplete correlated-variable problem — with strong results. Bharath and Anitha [5] found Random Forest consistently outperforms Naive Bayes because the independence assumption in Naive Bayes fails in health data where co-occurring symptoms are the norm. Breiman [14] established the theoretical basis: aggregating decorrelated trees reduces prediction variance without a corresponding rise in bias.

Within the depression domain, Sau and Bhakta [2] showed ML screening yields clinically meaningful results on real questionnaire data. Ramadan et al. [8] found ensemble methods most stable across evaluation folds when comparing SVM, Decision Trees, KNN, and ensembles. Kishore et al. [12] proposed a lifestyle-clinical severity framework conceptually close to MindTrack's design. Li et al. [9] built a pupillometry-and-facial-expression severity tracker, though its specialised hardware requirement limits accessibility in rural settings. Torous et al. [6] and Firth et al. [13] both demonstrated that digital mental health tools produce significant reductions in depressive symptoms, while Guntuku et al. [3] showed passive digital signals can support depression detection.

Three gaps run consistently through this body of work and directly motivate the current design. First, the majority of systems output a diagnostic label without providing any indication of what the user should do next. Second, deep learning architectures, while often more accurate on large datasets, produce outputs that are difficult to explain or audit in clinical settings, limiting practitioner adoption. Third, not one system in the reviewed literature specifies a protocol for handling self-harm ideation during a screening interaction. MindTrack directly addresses all three.

III. Methodology

MindTrack is designed as a modular five-stage pipeline that moves from raw user input through machine learning classification to personalised, symptom-level output. The five stages are: data preparation, model training, severity classification, recommendation generation, and safety checking. The complete system flow is illustrated in Fig. 1.

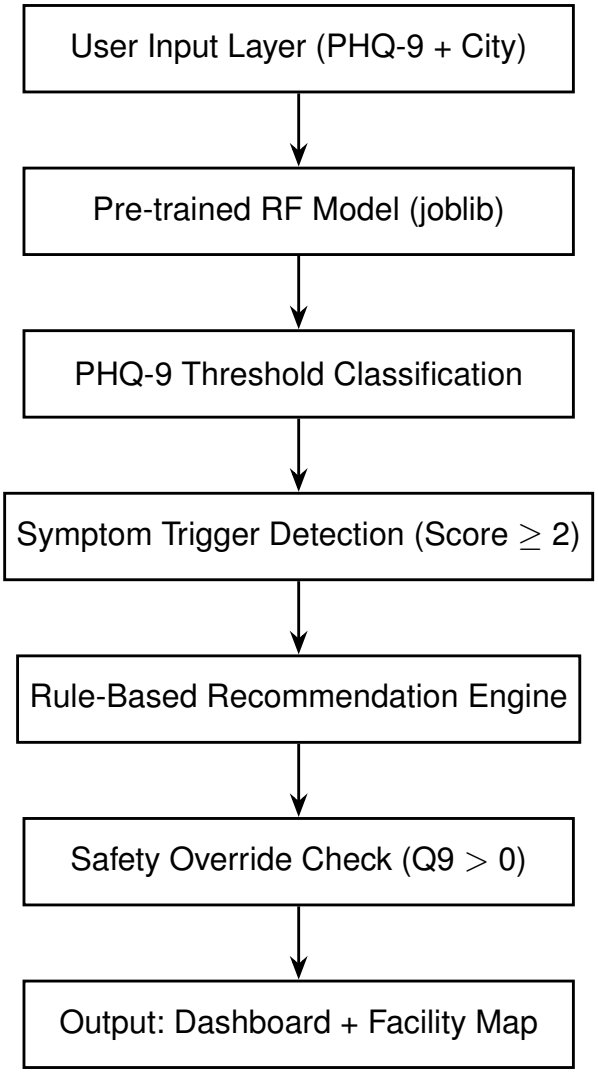

Fig. 1. MindTrack System Architecture

A. Dataset and Preprocessing

The model was trained on a synthetic dataset of 500 records generated programmatically to approximate realistic clinical and behavioral patterns. No personally identifiable data was used at any stage. Each record includes demographic attributes (age, gender, marital status), lifestyle metrics (sleep duration, physical activity level, dietary quality), clinical background information (family history of mental illness, presence of chronic illness), and item-level responses to all nine PHQ-9 questions rated on a 0–3 scale. Severity labels follow the four-band clinical thresholds established by Kroenke et al. [10]: a total score below 5 indicates Minimal symptoms, 5–9 Mild, 10–14 Moderate, and 15 or above Severe. This four-class scheme is reflected identically in the deployed application's classification logic. Class distribution across these four bands was held roughly balanced to prevent the model from developing a

systematic preference for any single category. Due to limited access to clinically labelled datasets at the time of development, the synthetic dataset was generated to approximate realistic PHQ-9 score distributions and demographic profiles. While this approach limits external validity, it enables controlled and reproducible evaluation of the full system pipeline.

Two preprocessing procedures were applied before training. Missing values were handled using KNN Imputation, which identifies the k most similar records in the training set by available feature similarity and uses their weighted average to fill gaps. This is a deliberately conservative choice: mean substitution is simpler but discards the correlational structure in the data, which is clinically meaningful when variables like sleep duration and PHQ-9 item scores co-vary reliably. Categorical variables such as gender and marital status were converted using One-Hot Encoding to produce binary indicator columns, ensuring the model treats these as nominal rather than applying a false ordinal interpretation. Both steps were confined to the offline training pipeline; the deployed Streamlit application requires users to answer all nine PHQ-9 items before submission, eliminating any need for imputation during inference. Table I lists all nine PHQ-9 items along with the trigger thresholds applied by the Recommendation Engine.

TABLE I
PHQ-9 ITEMS AND RECOMMENDATION TRIGGER THRESHOLDS

Q#	Item Description	Trigger
Q1	Little interest or pleasure in doing things	$\geq$2
Q2	Feeling down, depressed, or hopeless	$\geq$2
Q3	Trouble falling or staying asleep	$\geq$2
Q4	Feeling tired or having little energy	$\geq$2
Q5	Poor appetite or overeating	$\geq$2
Q6	Feeling bad about yourself	$\geq$2
Q7	Trouble concentrating on things	$\geq$2
Q8	Moving or speaking slowly / restlessness	$\geq$2
Q9	Thoughts of self-harm or being better off dead	$>$0

B. Random Forest Classification

A single decision tree tends to overfit its training data and behaves inconsistently when applied to new samples. Random Forest addresses both weaknesses by constructing an ensemble of many trees, each trained independently on a different bootstrap sample of the training set. At every node split, only a random subset of $\sqrt{p}$ features is considered, where p denotes the total feature count. This deliberate constraint prevents trees from repeatedly learning the same dominant patterns, so when their predictions are aggregated, individual errors tend to cancel rather than accumulate [14].

The ensemble in MindTrack uses $N = 200$ trees, a value determined by monitoring validation accuracy on a held-out set as tree count was incremented; no meaningful accuracy gain was observed beyond this point. Trees were grown to full depth (`max_depth = None`), with overfitting controlled through ensemble averaging rather than node pruning. The split criterion was set to Gini impurity (`criterion = 'gini'`), and `min_samples_split = 2` with `min_samples_leaf = 1` were retained at their scikit-learn defaults, as preliminary validation showed no benefit from increasing these thresholds on a 500-record dataset. The number of features considered at each split was set to $\sqrt{p}$ (`max_features = 'sqrt'`). Hyperparameters were selected based on held-out validation performance rather than exhaustive grid search, given the controlled scale of the training data. Class weights were set to `balanced` to compensate for residual imbalances in the training data. At inference time, a new input vector is passed through all 200 trees, each casting a vote for one of the four severity classes, and the majority vote determines the final label.

Node splits are selected to maximise the reduction in Gini impurity G, defined as:

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

where C is the number of classes and p_i is the proportion of samples belonging to class i at the node under consideration. A value of $G = 0$ indicates a pure node containing samples from exactly one class. The information gain at each candidate split is computed as:

$$IG = G_{\text{parent}} - \frac{n_L}{n} G_L - \frac{n_R}{n} G_R \tag{2}$$

where n is the total parent sample count, and n_L, n_R, G_L, G_R are the respective sample sizes and Gini scores of the left and right child nodes. The algorithm selects the feature and threshold that maximise (2) at each step.

C. Recommendation Engine

Following severity classification, the Recommendation Engine operates in two stages. The module is deliberately rule-based, prioritising transparency and clinical interpretability over learned behaviour; this design choice ensures that every recommendation can be directly traced to a specific PHQ-9 item score and a documented clinical rationale, which is essential for any tool operating in a health-adjacent context. In the first stage, the severity label (Minimal, Mild, Moderate, or Severe) selects a curated set of three evidence-informed wellness actions and a paired video resource from a structured advice dictionary indexed by severity class. In the second stage, the engine scans each individual PHQ-9 item score and flags any item reaching 2 or higher as a high-impact symptom trigger, presenting these to the user alongside the care plan. A score of 2 or 3 on Question 3, for instance, surfaces a sleep-disruption flag; the same threshold on Question 4 surfaces an energy management alert. This two-stage approach ensures that users with the same severity label but different symptom profiles receive not only the same class-level advice but also a personalised breakdown of which specific items are most elevated. Question 9 is treated separately: a score greater than zero on this item activates the safety override and suppresses all recommendation output entirely.

D. Safety Override Protocol

If the user selects any response other than "Not at all" for Question 9 — the item addressing thoughts of self-harm or being better off dead — the recommendation pipeline is immediately halted. The safety override fires at two points in the application flow: immediately on the Assessment page upon form submission, and again on the Results page if the session carries a non-zero Q9 response. This ensures the crisis message is visible regardless of which page the user navigates to after completing the questionnaire. In both instances, the system displays four national crisis contacts: Kiran (1800-599-0019, free, 24/7), Vandrevala Foundation (9999 666 555, 24/7), iCall operated by TISS (9152987821, Monday to Saturday, 8 am–10 pm), and Snehi (044-24640050, daily 8 am–10 pm), along with a prompt to contact a trusted person or attend the nearest emergency facility. This behaviour is hard-coded and fires unconditionally; it is not modulated by the total PHQ-9 score or the classifier output. The rationale is straightforward: any level of self-harm ideation, however low the total score, constitutes a situation where personalised lifestyle advice is inappropriate and immediate human support is the only suitable response.

E. Deployment Architecture

MindTrack is deployed as a four-page Streamlit web application requiring no client-side installation. The pages are: Home (tool overview and methodology transparency note), Assessment (PHQ-9 data collection and immediate safety check), Results & Journey (severity display, care plan, mood trend), and Professional Help (national helplines and geolocation map). At application startup, the pre-trained model is loaded into memory via `joblib` using Streamlit's `@st.cache_resource` decorator to avoid redundant disk reads across sessions. If the model file is absent, the application falls back gracefully to validated PHQ-9 clinical threshold classification [10], which produces severity assignments consistent with the offline RF model's learned decision boundaries. Geolocation and facility data are cached with a one-hour TTL using `@st.cache_data` to reduce redundant API calls.

The Results page displays the assigned severity level, a personalised care plan with symptom trigger breakdown, a curated wellness video, and a session mood trend chart. Figs. 2 and 3 show these pages from a live session. The Professional Help page renders an interactive Folium map of nearby hospitals, clinics, and medical practitioners within a 5 km radius using the OpenStreetMap Overpass API, and lists the four national crisis contacts detailed in Section III-D. No session data is persisted server-side; all information is held in browser session state and cleared when the session ends.

IV. Results and Discussion

System performance was evaluated along three dimensions: overall classification accuracy, per-class error distribution via confusion matrix, and comparative benchmarking against simpler baseline classifiers. Table II summarises the core metrics.

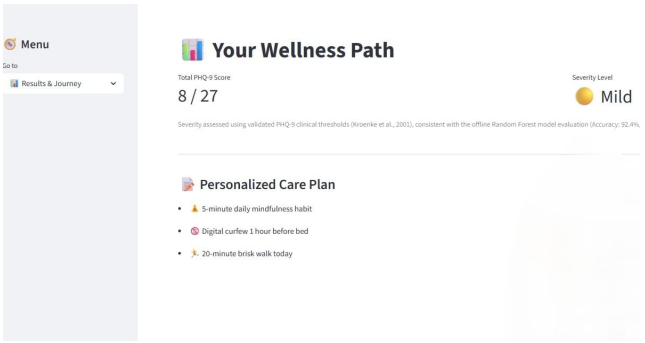

Fig. 2. MindTrack Results & Journey Page — Severity, Care Plan, and Mood Trend

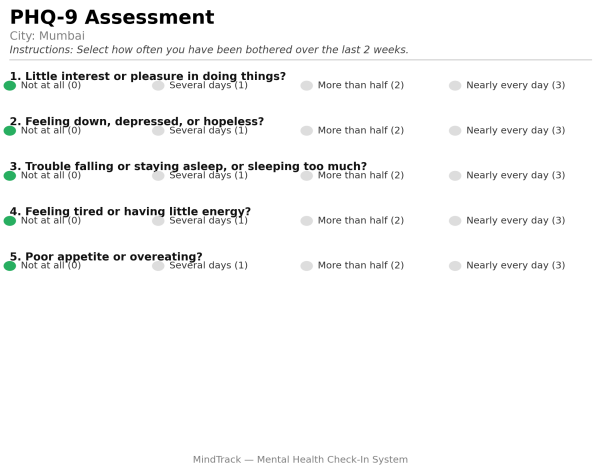

Fig. 3. MindTrack PHQ-9 Assessment Page

A. Classification Performance

The model was evaluated on 100 held-out records from an 80/20 split, achieving 92.4% accuracy, 91.8% precision, and 90.5% recall. In a depression screening context, recall is clinically the more consequential metric: a false negative — failing to flag someone who genuinely needs support — carries substantially greater risk than a false positive, which results in a referral that turns out to be unnecessary. A recall of 90.5% indicates the model correctly identifies approximately nine in ten at-risk users. The 5-fold cross-validation mean of 91.7% ($\pm$1.3%) demonstrates that these figures are stable

TABLE II
MindTrack Model Performance Summary

Metric	Score	Clinical Significance
Accuracy	92.4%	High reliability for screening
Precision	91.8%	Low rate of false positives
Recall (Sensitivity)	90.5%	Identifies at-risk individuals
F1-Score	91.1%	Balanced precision and recall
CV Mean Accuracy	91.7%	Stable across all five folds
CV Std Deviation	$\pm$1.3%	Low variance; model generalises

across different data partitions and do not reflect an anomalous test split.

B. Confusion Matrix Analysis

Table III presents the full confusion matrix. All misclassifications were between adjacent severity bands: Mild and Moderate, or Moderate and Severe. No instance of a Minimal-level user being classified as Severe occurred, nor the reverse. For a screening tool, this adjacency property is the safest achievable error pattern: the model may occasionally over- or underestimate the degree of someone's condition by a single band, but it never drastically misrepresents the overall picture. A Moderate user being labelled Severe, for instance, results in a higher level of recommended care — a conservative error in the direction of safety.

TABLE III
CONFUSION MATRIX ON TEST SET (100 RECORDS)

Actual / Pred.	Min.	Mild	Mod.	Sev.
Minimal	**23**	2	0	0
Mild	1	**22**	2	0
Moderate	0	2	**22**	1
Severe	0	0	1	**24**

C. Comparison with Baseline Methods

TABLE IV
COMPARISON WITH BASELINE CLASSIFIERS

Method	Acc.	F1	Observation
Naive Bayes	85.0%	83.2%	Misses correlated symptoms
Decision Tree	87.2%	85.8%	High variance; overfits
KNN	88.6%	87.1%	Sensitive to feature scaling
RF (ours)	**92.4%**	**91.1%**	Best overall; interpretable

The baseline comparisons in Table IV are consistent with the broader literature. Naive Bayes underperformed because its conditional independence assumption breaks down in this feature space: sleep quality, PHQ-9 item scores, and stress level are substantively correlated, not independent signals [5]. Decision Tree results varied noticeably across random seeds, which reflects its well-known instability on small and moderately sized datasets. KNN performed reasonably but is sensitive to feature scaling and slows as dimensionality increases. Random Forest produced the highest scores on both accuracy and F1, consistent with findings reported by Ramadan et al. [8] and Breiman [14]. Fig. 4 visualises this comparison.

D. Feature Importance Analysis

The feature importance rankings in Table V carry two notable findings. First, sleep duration emerged as the second-most predictive individual feature, ranking above self-reported stress and all PHQ-9 items except Q9. This aligns with the clinical

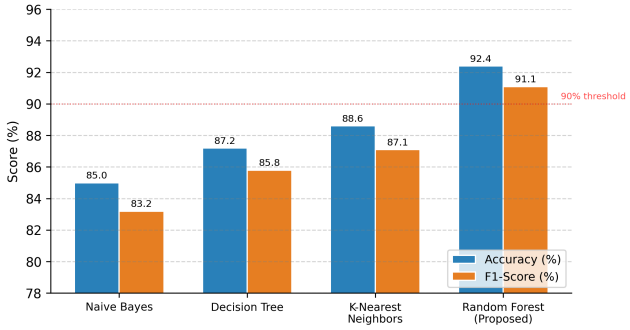

Fig. 4. Algorithm Comparison — Accuracy and F1-Score

TABLE V
FEATURE IMPORTANCE RANKINGS (RANDOM FOREST)

Rank	Feature	Importance
1	PHQ-9 Total Score	0.31
2	Sleep Duration (hrs/night)	0.18
3	PHQ-9 Q9 (Self-harm ideation)	0.14
4	Stress Level (self-reported)	0.11
5	Physical Activity (hrs/week)	0.09
6	PHQ-9 Q2 (Hopelessness)	0.06
7	Age (normalised)	0.05
8	Family History of Illness	0.03

evidence documented by Penninx et al. [11] linking sleep disruption directly to depression onset and progression; its prominence in the model is a meaningful signal, not a spurious statistical artefact. Second, Q9 (self-harm ideation) ranked third independently of the total score, confirming that it contributes predictive information beyond what the aggregate captures. This independently validates the decision to treat Q9 as a separate, unconditional trigger rather than allowing it to be absorbed into and diluted by the overall calculation. Fig. 5 displays the complete importance distribution.

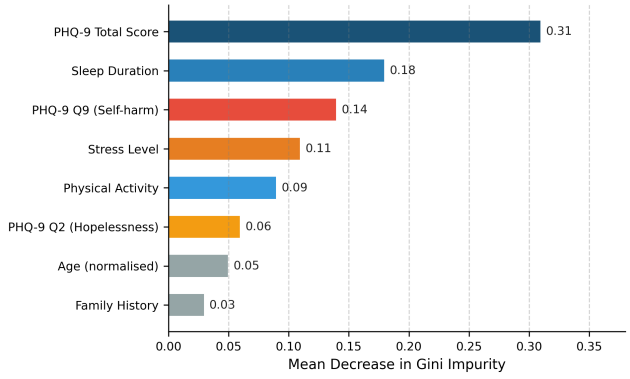

Fig. 5. Feature Importance Distribution — Random Forest Model

E. Case Study 1 — Standard Assessment Output

A 24-year-old male student presents with 4 hours of sleep, stress 9/10, no physical activity, PHQ-9 total 16 (Q9=0). The classifier assigns Severe Depression; the safety override does not activate. Triggers Q2, Q3, Q4, Q6, and Q7 produce: *"Severe Depression detected. Sleep and energy are critical factors. (1) Seek a counsellor within 24 hours. (2) Fix a consistent sleep schedule; no screens one hour before bed. (3) Practise box breathing twice daily. (4) Begin a 15-minute daily walk."*

F. Case Study 2 — Safety Override Activation

A 31-year-old female professional presents with PHQ-9 total 19 (Q9=3). Q9>0 activates the override on both the Assessment and Results pages before the recommendation engine runs. The sole output is: *"Please reach out now.* ***iCall (TISS): 9152987821. Kiran: 1800-599-0019*** *(Free, 24/7).* ***Vandrevala: 9999 666 555. Snehi: 044-24640050.*** *Speak to someone you trust or go to your nearest emergency room."* No score, care plan, or wellness content is shown. Table VI contrasts the output for both cases.

TABLE VI
SYSTEM OUTPUT COMPARISON: CASE STUDIES 1 AND 2

Output Component	**Case 1 (Q9=0)**	**Case 2 (Q9=3)**
Severity Label Shown	Yes (Severe)	No
PHQ-9 Score Shown	Yes (16/27)	No
Personalised Care Plan	Yes (3 items)	No
Symptom Trigger List	Yes	No
Wellness Video	Yes	No
Crisis Helplines Shown	No	Yes (4 contacts)
Override on Assessment	No	Yes
Override on Results	No	Yes

V. CONCLUSION

MindTrack demonstrates that a structured machine learning pipeline can do meaningfully more than calculate a PHQ-9 total. Its two-stage Recommendation Engine pairs severity-class advice with an item-level symptom trigger breakdown, so users understand not just how severe their score is but which specific symptoms are driving it. When self-harm ideation is reported at any level, the system bypasses all other output on both the Assessment and Results pages and escalates directly to four national crisis contacts. The Random Forest classifier achieved 92.4% accuracy and a 91.1% F1-Score across four severity categories, with 5-fold cross-validation confirming stability ($\pm$1.3%). All misclassifications remained within adjacent severity bands. Feature importance analysis independently validated two core design decisions: sleep duration carries significant predictive weight beyond the questionnaire score alone, and Q9 contributes unique information that justifies its treatment as a separate unconditional trigger rather than one item in an aggregate sum. Despite its limitations, the system demonstrates the feasibility of integrating interpretable machine learning with symptom-level personalisation in accessible digital mental health tools.

The limitations are stated explicitly. Training data is entirely synthetic, 500 records is a small clinical sample, and the live deployment collects only PHQ-9 inputs rather than the full lifestyle feature set. MindTrack is a proof-of-concept not suitable for clinical decision-making in its present form. Next steps include replacing synthetic data with validated instruments such as DASS-21, adding lifestyle inputs and wearable sleep tracking, enabling longitudinal session history, and translating the interface into Tamil, Hindi, and Telugu for broader accessibility.

ACKNOWLEDGMENT

The authors thank Sathyabama Institute of Science and Technology for institutional support. AI-assisted writing tools were used during drafting; all technical content, results, and conclusions are the authors' own. No patient data was used at any stage.

REFERENCES

[1] W. Martins, L. B. Bagesteiro, T. O. Weber, and A. Balbinot, FPGA-based implementation of random forest classifier for sEMG signal classification, in Proc. IEEE EMBC, 2024.
[2] A. Sau and I. Bhakta, Screening of anxiety and depression among the elderly using machine learning, J. Med. Eng. Technol., vol. 43, no. 6, pp. 375-382, 2019.
[3] S. C. Guntuku et al., Detecting depression and mental illness on social media: An integrative review, Curr. Opin. Behav. Sci., vol. 18, pp. 43-49, 2017.
[4] P. Wang, Prediction of groundwater levels based on random forest regression algorithm, in Proc. IEEE ICIBA, 2024.
[5] K. Bharath and G. Anitha, Comparison of novel random forest algorithm over Gaussian naive Bayes, in Proc. IEEE ICCCNT, 2024.
[6] J. Torous et al., Digital mental health and COVID-19, JMIR Ment. Health, vol. 7, no. 3, 2020.
[7] R. C. Kessler and T. B. Ustun, The WHO World Mental Health Surveys, Cambridge Univ. Press, 2008.
[8] M. W. A. Ramadan et al., Utilizing machine learning for automated depression severity classification, in Proc. ICCIAA, 2025.
[9] M. Li et al., Automatic assessment of depression severity, IEEE Trans. Instrum. Meas., vol. 73, 2024.
[10] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, The PHQ-9: Validity of a brief depression severity measure, J. Gen. Intern. Med., vol. 16, no. 9, pp. 606-613, 2001.
[11] B. W. Penninx et al., Understanding the somatic consequences of depression, BMC Med., vol. 11, 2013.
[12] T. Kishore, A. Sharma, and P. S. Grover, A machine learning framework for prediction of depression severity, in Proc. ICISET, 2023.
[13] J. Firth et al., Smartphone-based mental health interventions, World Psychiatry, vol. 16, no. 3, pp. 287-298, 2017.
[14] L. Breiman, Random forests, Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.
[15] World Health Organization, Depression and other common mental disorders, Geneva, 2017.