Hybrid Machine Learning Model for Risk Prediction and Action Recommendation Based on Artificial Mental Systems

Hadi Asnal[1*], Khusaeri Andesa[2], Fitry Erlin[3], Junadhi[4]

[1,2,4]Universitas Sains dan Teknologi Indonesia, Pekanbaru, Riau, Indonesia
[3]Institut Kesehatan Payung Negeri Pekanbaru, Pekanbaru, Riau, Indonesia
E-mail: [1*]hadiasnal@usti.ac.id [2]khusaeriandesa@usti.ac.id, [3]fitryerlin@gmail.com, [4]junadhi@usti.ac.id

(Received: 17 Aug 2025, revised: 13 Sep 2025, accepted: 16 Sep 2025)

Abstract

Mental health problems are increasingly prevalent among the younger generation, particularly those active on social media, yet early detection efforts often remain limited. Previous studies have explored text-based approaches for identifying mental health issues, but many are constrained by low accuracy in differentiating multiple psychological states or lack integration into accessible tools for end-users. This study addresses these gaps by proposing a hybrid machine learning model for early detection of mental health conditions through social media text analysis. Five algorithms were evaluated, and a soft voting ensemble combining Logistic Regression and Support Vector Machine (SVM) was developed to improve classification across five mental states (Anxiety, Depression, Stress, Emotional Exhaustion, and Healthy) and three risk levels (Low, Medium, High). To ensure practical utility, the model was deployed in an Android-based application, SmartRisk, which allows users to input free text and receive automated assessments. The findings show that the proposed hybrid approach significantly improves detection performance, particularly in identifying depression and high-risk cases, while maintaining high usability in real-world application. The novelty of this study lies in combining hybrid ensemble learning with mobile deployment for practical, text-based early detection of mental health, offering both methodological advancement and societal impact.

Keywords: Mental Health, Machine Learning, Risk Classification, Hybrid Model, Mobile Application

I. INTRODUCTION

Mental health is increasingly recognized as a major public health concern, particularly among younger populations who are active users of social media. In Indonesia, more than 9% of adults are reported to suffer from emotional disorders, with a trend that continues to rise each year. Despite its growing importance, access to professional services remains severely limited, especially outside urban areas. The shortage of psychologists and psychiatrists, coupled with social stigma surrounding mental health consultation, creates a substantial barrier for individuals in need of timely support [1], [2]. Global initiatives, including the World Health Organization (WHO) and the Sustainable Development Goals (SDG 3), emphasize mental health as a central element of global well-being. However, translating these commitments into practical solutions is particularly challenging in low-resource settings such as Indonesia. National health policies have recognized the urgency of this issue, but systemic constraints mean that millions of people remain without access to adequate care. This gap highlights the urgent need for scalable, technology-driven approaches that can complement existing healthcare infrastructure [3].

In recent years, machine learning (ML) has shown promise in detecting mental health conditions from digital traces, such as text, survey responses, or smartphone usage patterns. Prior studies have successfully applied algorithms like Support Vector Machine (SVM), Random Forest, and Logistic Regression to predict symptoms of depression, anxiety, and stress with varying degrees of accuracy [4], [5]. These approaches demonstrate that ML can provide fast and continuous monitoring of psychological states, which could play a vital role in prevention and early intervention. Despite these advances, existing works share several limitations. First, the majority of studies focus on predicting a single condition—such as depression—without considering the complex, overlapping nature of mental health disorders. Second, very few approaches have attempted to combine classification of psychological states with stratification of risk levels, even though such dual-layer detection is essential for prioritizing early intervention. Third, many studies remain confined to experimental settings and are rarely deployed in real-world applications accessible to non-clinical users.

DOI: 10.34148/teknika.v14i13.1357

Finally, although the need for intervention has been widely acknowledged, most systems stop at prediction, offering little or no actionable feedback for individuals who may require urgent support [6], [7].

These limitations are particularly problematic in the Indonesian context. With fewer than one psychiatrist per 100,000 people and significant geographic disparities in access to care, mental health services remain inaccessible for much of the population. At the same time, smartphone penetration is high, and digital platforms have become integral to everyday communication, particularly among young people. This combination of unmet clinical need and widespread mobile access creates an opportunity to design AI-driven systems that are not only accurate but also practical, user-friendly, and contextually relevant for Indonesia [8], [9], [10].

To address these gaps, this study introduces SmartRisk, an AI-powered early detection system that leverages a hybrid ensemble model to classify both mental health states (Anxiety, Depression, Stress, Emotional Exhaustion, Healthy) and corresponding risk levels (Low, Medium, High) based on text data. By integrating multiple machine learning methods through a soft-voting mechanism, the system enhances robustness and improves accuracy compared to single-model approaches. More importantly, SmartRisk is deployed as an Android-based mobile application, allowing users to enter free-text reflections of their emotions and receive automatic assessments in real time. Such deployment aligns with findings that mobile apps incorporating ML for mental health support are feasible and show promise, though many existing studies lack usability testing and risk stratification [11], [12]. Unlike prior studies that focus exclusively on prediction, SmartRisk emphasizes practical deployment and end-user validation. Usability testing with the System Usability Scale (SUS) was conducted to ensure that the application is not only technically sound but also accessible and engaging for its intended audience. This approach positions the system as a bridge between academic research and real-world impact, particularly in low-resource healthcare environments. Findings from [13] suggest that for ML applications in mental health to be effective, human-centered design and usability evaluation are critical.

The novelty of this study lies in three interrelated contributions. First, it combines hybrid ensemble learning with multi-label detection, capturing both psychological states and risk levels simultaneously [14]. Second, it validates the system through usability testing, demonstrating practical feasibility beyond algorithmic performance [6]. Third, it is explicitly designed for the Indonesian context, where resource constraints and social stigma make traditional interventions insufficient, but mobile-based AI systems can provide scalable alternatives [8]. Together, these contributions distinguish SmartRisk from previous works and underline its potential as a benchmark for AI-driven digital mental health solutions in Indonesia and other resource-limited settings.

II. RESEARCH METHOD

The research stages in the development of the SMART-RISK model are systematically structured to ensure that each process from data collection to model evaluation and the formulation of intervention recommendations can be carried out in a measurable and replicable manner. This process flow is visually represented in Figure 1, which illustrates the key steps undertaken throughout this study.

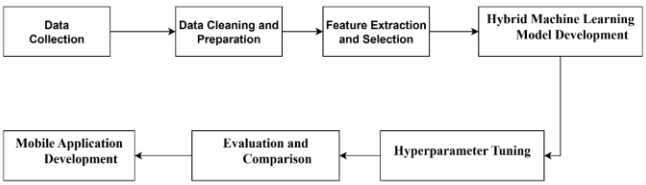

Figure 1. Research Architecture Development

2.1 Data Collection

The data for this study were collected from the social media platform X.com, as it is considered to reflect users' spontaneous expressions related to their psychological states. Data collection was conducted using web scraping techniques with Selenium, due to the limited access to X.com's official API. The data retrieval process involved formulating a set of keyword queries that are semantically related to both negative mental states (e.g., "I'm stressed," "mentally exhausted," "want to give up") and positive ones (e.g., "I'm happy," "I'm grateful") to ensure balanced labeling. Through this approach, a total of 12,687 public status posts were collected, including IDs, timestamps, content, and other metadata, and were stored in CSV format. All data were then manually labeled based on two main attributes.

1. Mental condition categories: Anxiety, Depression, Stress, Emotional Exhaustion, and Healthy.
2. Risk levels: 0 (no risk), 1 (moderate risk), and 2 (high risk).

While X.com provides a rich and openly accessible dataset frequently used in mental health research [4], [5], reliance on a single platform inevitably limits representativeness, as its users are predominantly younger and more digitally active. Thus, findings from this study may not fully generalize to other demographic groups or communication contexts. To address this limitation, future research should extend data sources to other platforms (e.g., Instagram, TikTok, forums) or complement social media text with survey and clinical datasets.

Labeling was conducted through manual annotation by three annotators (two licensed clinical psychologists and one graduate researcher in psychology) following predefined guidelines. The dataset was labeled into five psychological states (Anxiety, Depression, Stress, Emotional Exhaustion, Healthy) and three risk levels (Low, Medium, High). To ensure consistency, Fleiss' Kappa was calculated, yielding 0.78 (substantial agreement), consistent with standards for inter-rater reliability [15]. Disagreements were resolved through consensus discussion, a practice recommended in qualitative content analysis [13].

2.2 Data Cleaning and Preparation

At this stage, the first step involved removing duplicate entries based on text content. These duplicates typically originated from identical status updates resulting from retweets or viral posts that appeared repeatedly in the scraping results. Subsequently, irrelevant data were filtered out, including very short statuses (fewer than three words) and posts containing spam, advertisements, or links without clear emotional narratives. Next, all text underwent a preprocessing phase to generate a cleaned text column. This process included the removal of links, user mentions (@), hashtags (#), numbers, and punctuation marks. The text was then converted to lowercase for standardization, followed by stemming using the Sastrawi library to reduce words to their root forms in accordance with the structure of the Indonesian language. This step aimed to simplify the data and minimize unnecessary word variation [16].

2.3 Feature Extraction and Selection

Once the data were cleaned, the next step was to transform the text into numerical representations that could be processed by machine learning algorithms. For this purpose, the Term Frequency–Inverse Document Frequency (TF-IDF) method was employed, which assigns higher weights to terms that are unique within a specific document but infrequent across the entire corpus [17]. This approach is effective in capturing contextually relevant terms for classification tasks. The TF-IDF configuration in this study included ngram_range = (1,2) to account for both unigrams and bigrams, max_features = 5000 to limit feature dimensionality for computational efficiency, and stop_words = 'indonesian' to remove common words that do not contribute meaningful classification value. The output of this stage was a numerical feature vector matrix, which served as input for training classification models for mental health status and risk level in the subsequent phase [18].

2.4 Hybrid Machine Learning Model Development

At this stage, model training was conducted using a hybrid machine learning approach by implementing five popular classification algorithms: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and Decision Tree [19], [20]. The selection of these five algorithms was based on their proven ability to handle textual data and deliver strong classification performance in various prior studies, particularly in the domain of text-based psychological condition detection. The cleaned dataset, represented as TF-IDF vectors, was then divided into two parts: 80% for training and 20% for testing. While this split is a widely adopted practice, model optimization was further strengthened by employing 5-fold cross-validation during hyperparameter tuning to reduce variance and minimize overfitting. Each algorithm was trained separately to address two classification scenarios:

1. Classification of mental health categories (y_label), which includes labels such as Anxiety, Depression, Stress, Emotional Exhaustion, and Healthy.
2. Classification of mental health risk levels (y_risk), categorized into three levels: No Risk (0), Moderate Risk (1), and High Risk (2).

2.5 Hyperparameter Tuning

To improve model performance and ensure optimal parameter configurations, each machine learning algorithm underwent a hyperparameter tuning process. This process was carried out using the GridSearchCV technique with a 5-fold cross-validation scheme, which was chosen because it provides a systematic and exhaustive search over parameter combinations, while cross-validation minimizes the risk of overfitting and improves model generalizability [21]. The selection of parameter values was tailored to the characteristics of each algorithm. For Random Forest and Decision Tree, parameters such as max_depth, min_samples_split, and n_estimators were tuned, as these directly influence model complexity and the bias-variance trade-off [22]. For the SVM model, parameters including C, gamma, and kernel were optimized, since they control the margin width, non-linear decision boundaries, and kernel transformations that are critical in handling high-dimensional text data. By explicitly optimizing parameters for each model, the tuning process ensured that the evaluation was both fair across algorithms and aligned with best practices in applied machine learning. The procedure was repeated separately for the two classification tasks (mental state classification and risk level detection) to account for differences in label distribution and task complexity.

1. Hyperparameters for Mental Health Categories (y_label)
   For this target, tuning was performed to obtain the optimal model configuration for distinguishing between categories such as Anxiety, Depression, Stress, Emotional Exhaustion, and Healthy. The tuning process focused on parameters that influence the model's sensitivity to variations in emotional text.
2. Hyperparameters for Mental Health Risk Levels (y_risk)
   For the risk level classification target, tuning was aimed at optimizing the model's ability to distinguish between the ordinal risk levels (0, 1, 2). Given the ordinal nature and the fact that the high-risk class was underrepresented, class weighting was applied during model training to mitigate imbalance, ensuring that errors in minority classes were penalized more heavily [23]. In addition, evaluation emphasized recall and F1-score metrics for the high-risk class, as these measures are more informative than accuracy when dealing with imbalanced datasets. While alternative methods such as SMOTE and oversampling have been widely adopted [24], class weighting was chosen in this study to reduce overfitting risks in a relatively small dataset. Future work may extend to hybrid imbalance solutions (e.g., cost-sensitive learning or oversampling combined with class weighting) to further enhance robustnesss.

2.6 Evaluation and Comparison

Evaluation was conducted on the five classification models to assess their performance across two scenarios:

mental health category classification and mental health risk level classification. The models were tested using the testing dataset, and evaluation metrics included Precision, Recall, F1-Score, and Accuracy [25], [26].

1. Mental Health Categories
   The models were evaluated based on their ability to classify users' mental status into five categories: Anxiety, Depression, Stress, Emotional Exhaustion, and Healthy. The assessment included average scores per class and cross-model comparisons to identify the model with the most balanced performance.
2. Mental Health Risk Levels
   Evaluation was also carried out for classifying risk levels (0 = No Risk, 1 = Moderate Risk, 2 = High Risk). Particular emphasis was placed on the Recall and F1-Score of the High-Risk class, due to the critical importance of identifying potentially crisis-level cases. The best-performing model was selected based on its high sensitivity and stable performance across all classes.

2.7 Mobile Application Development

As the final implementation stage of the SMART-RISK system, a mobile application for Android was developed to serve as the user interface for detecting mental health status and providing recommended actions. The application was designed to be user-friendly and accessible to the general public, with a focus on ease of access, fast response time, and an intuitive user interface. Development was carried out using Android Studio, with the Dart programming language via the Flutter framework, and the application was targeted exclusively for the Android platform [27], [28]. The application is integrated with a machine learning model that has been trained and deployed via a backend service. Users simply input a text message reflecting their feelings or personal status, and the system analyzes the input to display the classification results of the user's mental condition along with the associated risk level. In addition to detection, the application provides personalized intervention recommendations based on the identified risk level. The main features of the application include (1) a text input form, (2) detection of mental health status and risk level, (3) display of recommended actions, and (4) a history log of detection results. The development of this application aims to improve public access to proactive, technology-based early mental health detection services.

III. RESULT AND DISCUSSIONS

The results of the study indicate that the five baseline models—Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and Decision Tree—exhibited varying levels of performance across the classification categories. Evaluation was conducted on two primary targets, namely mental health categories and mental health risk levels, using Accuracy, Precision, Recall, and F1-Score metrics on the testing dataset. These variations highlight that while some algorithms capture explicit emotional expressions effectively (e.g., Logistic Regression and SVM for Healthy and Anxiety classes), others struggle with more subtle or imbalanced categories such as Depression and High Risk. This suggests that differences in linguistic patterns and class distribution strongly influence model performance, underscoring the need for ensemble approaches to balance sensitivity across categories.

3.1 Baseline Model Performance

3.1.1. Mental Health Categories

The evaluation of the five baseline models served as a critical step in assessing the performance of each algorithm in classifying mental health status based on textual data from social media. The five target categories Anxiety, Depression, Emotional Exhaustion, Stress, and Healthy represent a spectrum of emotional and psychological states commonly encountered in daily life. Model performance was assessed using the testing set, which had previously been separated from the training data with an 80:20 split ratio. This evaluation aimed to determine how well each model could generalize its classification to new, unseen data.

Based on the evaluation results of the five machine learning algorithms Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and Decision Tree it can be concluded that Logistic Regression achieved the highest overall accuracy of 0.86, followed by Gradient Boosting, Random Forest, and SVM, each with an accuracy of 0.85. In contrast, the Decision Tree algorithm exhibited the lowest performance with an accuracy of 0.79. For the "Anxiety" category, all models demonstrated strong performance, with F1-scores ranging from 0.83 to 0.90. Gradient Boosting and SVM recorded the highest F1-scores (0.90 and 0.89, respectively), indicating good consistency in detecting expressions of anxiety. The "Depression" category proved to be the most challenging, with all models showing relatively lower F1-scores—ranging from 0.66 for Decision Tree to 0.77 for Logistic Regression. This suggests that expressions of depression tend to be more implicit and variable, making them more difficult for models to accurately recognize.

In the "Emotional Exhaustion" category, model performance was relatively stable, with F1-scores ranging from 0.77 to 0.85. Logistic Regression and Gradient Boosting once again outperformed other models in this category. The "Stress" category demonstrated high precision across nearly all models (0.93–0.95), but with comparatively lower recall (0.81), indicating a tendency for the models to accurately identify stress when predicted, yet potentially miss actual stress cases (false negatives). The "Healthy" category exhibited the highest and most consistent F1-scores, reaching up to 0.94 for Logistic Regression, Random Forest, and Gradient Boosting. Notably, Random Forest achieved a recall of 1.00, indicating its exceptional capability in correctly identifying healthy conditions. Overall, Logistic Regression and Gradient Boosting emerged as the most balanced models in detecting various mental health conditions. Meanwhile, although simpler in structure, the Decision Tree model proved

less reliable in handling the complexity of psychological expressions in text. These findings underscore the need for a hybrid modeling approach to leverage the strengths of the best-performing models across different label categories.

However, a closer look at misclassified cases provides further insight into the limitations of the models. For instance, the text “I’m just tired of everything, maybe tomorrow will be better” was labeled by annotators as Depression, yet the model frequently predicted Stress, reflecting the semantic overlap between fatigue-related expressions and stress symptoms. Similarly, short ambiguous posts such as “I’m fine, don’t worry” were often classified as Healthy, while annotators labeled them as Emotional Exhaustion, highlighting the difficulty of detecting irony and implicit negativity. Errors were also common in the risk-level classification: posts like “I just want to disappear” were sometimes predicted as Moderate Risk rather than High Risk, likely due to the small number of high-risk training examples. These errors indicate that misclassification is influenced by the subtle and context-dependent nature of mental health language as well as by class imbalance, especially for critical categories such as Depression and High Risk. This analysis underscores the importance of integrating more context-aware language models and imbalance-handling strategies to improve sensitivity in future work.

3.1.2 Mental Disorder Risk Level Categories

Model evaluation for the classification of mental health risk levels was conducted on five machine learning algorithms: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and Decision Tree. The evaluation was performed using the testing dataset and assessed model performance based on key metrics such as Precision, Recall, F1-score, and overall Accuracy, in order to obtain a comprehensive understanding of the models' predictive accuracy and consistency. This classification task presented an additional significant challenge: class imbalance. The high-risk class (label 2) contained substantially fewer samples than the moderate-risk (1) and low/no-risk (0) classes, which caused models to be biased toward the majority classes. Such imbalance can lead to decreased model performance in detecting high-risk cases—precisely the most critical to identify early. Therefore, the interpretation of metrics such as Recall and F1-score for minority classes is of particular importance, as a high overall accuracy alone is insufficient to ensure the model's effectiveness in real-world, high-risk scenarios. A more in-depth class-wise performance analysis is necessary to evaluate how well each model can identify signs of mental health issues across different risk levels, especially in the context of early detection for high-risk groups that require immediate attention and intervention.

The evaluation of the classification performance of five models Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and Decision Tree in predicting mental health risk levels revealed that SVM achieved the highest overall accuracy at 78%, followed by Logistic Regression at 77%. While accuracy provides a general overview of performance, the F1-score is considered a more representative metric, particularly under conditions of significant class imbalance, such as in the high-risk class (class 2). In the Class 0 category (no risk), all models demonstrated highly stable and strong performance, with F1-scores ranging from 0.90 to 0.93 and consistently high recall values (≥ 0.92). This indicates that all models were highly effective in identifying psychologically stable or healthy conditions. Gradient Boosting and SVM, in particular, showed the most superior performance for this class. Meanwhile, in Class 1 (moderate risk), the models also demonstrated reasonably strong performance, with F1-scores ranging from 0.72 to 0.84. Logistic Regression and SVM once again achieved the best performance, with F1-scores of 0.84 and 0.82, respectively. The consistently high recall values across all models (above 0.72) indicate that the ability to detect individuals with moderate emotional risk is relatively reliable.

In contrast, Class 2 (high risk) was the most challenging category for all models to classify accurately. F1-scores for this class ranged from 0.40 (Decision Tree) to 0.63 (Logistic Regression), with low recall values (0.44–0.59), indicating a high number of false negatives. This weakness is strongly suspected to result from the imbalanced data distribution, where high-risk samples are underrepresented, as well as from the more implicit and varied linguistic expressions typically associated with this risk category. Among the five models, Logistic Regression and SVM demonstrated the most balanced and stable performance across all risk levels, whereas Decision Tree consistently showed the lowest performance, particularly in detecting high-risk cases. Therefore, it can be concluded that although most models are effective in identifying low to moderate mental health risk conditions, a more adaptive or ensemble-based approach is necessary to improve classification accuracy for high-risk cases.

3.2 Hybrid Model Results

The evaluation results of the five baseline models indicate that no single model consistently outperformed the others across all label categories, whether in classifying mental health status or predicting risk levels. Variations in F1-scores across models and label classes suggest that each model possesses unique strengths in detecting specific categories. Therefore, to optimize the overall system performance, a hybrid modeling approach was employed combining the specific advantages of individual baseline models into a collective classification mechanism.

3.2.1 Model Selection Strategy

The selection of models for the hybrid approach was conducted in a data-driven manner, based on the highest F1-score achieved by each model within individual label classes. The guiding principle applied was "cover the weakness with others’ strength", aiming to compensate for the limitations of one model with the strengths of another that performs better on a specific class. Table 1 presents the hybrid model configuration for predicting mental health status categories (y_label).

DOI: 10.34148/teknika.v14i13.1357
TEKNIKA, Volume 14(3), November 2025, pp. 415-423
ISSN 2549-8037, EISSN 2549-8045

Table 1. The Best Model for Each Mental Health Category Class

Category Labels	Best Model (F1-Score)
Anxiety	Gradient Boosting / SVM (0.90/0.89)
Depression	Logistic Regression (0.77)
Emotional Exhaustion	Logistic Regression (0.85)
Stress	Gradient Boosting / SVM (0.88)
Healthy	All models are of equal height (0.93–0.94)

The hybrid approach in this study employed Logistic Regression and Support Vector Machine (SVM). Logistic Regression was selected based on its superior performance in classifying the Depression and Emotional Exhaustion categories, where it achieved the highest F1-scores compared to other models. On the other hand, SVM was chosen for its consistent and balanced performance across various categories, particularly in dominant classes such as Anxiety, Stress, and Healthy. The combination of these two models was implemented using a soft voting strategy, which aggregates the predicted probabilities from each model and selects the final class based on the highest combined score. This approach offers greater flexibility in addressing class imbalance and enables the system to consider each model's confidence level in its predictions, thereby enhancing the overall accuracy and stability of classification. The results of the hybrid model for predicting mental health risk levels (y_risk) are presented in Table 2 below.

Table 2. The Best Model for Each Risk Level Category

Risk Level	Best Model (F1-Score)
0 (Low Risk)	Gradient Boosting / SVM (0.93)
1 (Medium Risk)	Logistic Regression (0.84)
2 (High Risk)	Logistic Regression (0.63)

The models selected to construct the hybrid classifier for mental health risk level classification were Logistic Regression and Support Vector Machine (SVM). Logistic Regression was chosen due to its strong performance in the moderate-risk (class 1) and high-risk (class 2) categories, particularly in achieving superior F1-scores compared to other models. Meanwhile, SVM proved effective in detecting low-risk cases (class 0), exhibiting high precision and recall, along with stable performance across other classes. This combination was strategically designed to balance predictive strengths across classes, especially in addressing the limitations of baseline models that tended to underperform in identifying high-risk cases due to data imbalance. By integrating the strength of Logistic Regression in handling minority classes with the stability of SVM in recognizing general patterns, the hybrid classifier is expected to enhance both sensitivity and overall accuracy particularly for critical cases that require greater clinical attention.

3.3 Hybrid vs Baseline Evaluation and Comparison

A final evaluation was conducted to compare the performance of the hybrid model with the highest performance of each baseline model. The results showed that the hybrid classifier offered significant stability and performance improvements, especially for labels that were previously difficult to accurately classify. Table 3 presents the test results.

Table 3. Hybrid Accuracy Metrics Mental Health Category Using Test Data

Mental Condition	Precision	Recall	F1-score
Anxiety	0.87	0.91	0.89
Depression	0.76	0.78	0.77
Emotional Exhaustion	0.85	0.85	0.85
Stress	0.95	0.82	0.88
Healthy	0.90	0.97	0.94
Accuracy			**0.86**

Based on these results, the hybrid model combining Logistic Regression (LR) and Support Vector Machine (SVC) demonstrated more consistent and competitive classification performance compared to individual baseline models. The highest F1-scores were achieved in the Healthy (0.94) and Anxiety (0.89) categories, while performance improvements were also observed in the Depression category, where the F1-score increased from 0.73 to 0.77, indicating enhanced sensitivity to minority classes.

When compared with prior research, the achieved accuracy of 86% is competitive. For instance, a feature-based Twitter depression detection study reported accuracies between 78% and 89% depending on the classifier used [18] achieved around 91% accuracy for depression symptom detection using Twitter texts (PMC). More recent work by [14] on Indonesian Twitter data using CNN and LSTM models reported accuracies ranging between 83% and 91%, while [14] using BiLSTM achieved about 94% accuracy on Indonesian datasets. These comparisons suggest that while certain deep learning models can achieve slightly higher performance under specific conditions, the proposed hybrid approach remains highly competitive. Its added novelty lies in addressing multi-label classification (mental states and risk levels) and integrating usability testing through a mobile application, features that are often missing in prior studies.

Table 4. Hybrid Accuracy Metrics Risk Level Category Using Test Data

Risk Level	Precision	Recall	F1-score
0 (Low Risk)	0.90	0.95	0.92
1 (Medium Risk)	0.78	0.86	0.82
2 (High Risk)	0.72	0.59	0.64
Accuracy			**0.78**

Based on the evaluation results in Table 4, the hybrid model combining Logistic Regression (LR) and Support Vector Machine (SVM) demonstrated more consistent

classification performance, and in several categories, even outperformed the individual baseline models. The highest F1-scores were achieved in the Healthy (0.94) and Anxiety (0.89) categories, indicating the model's ability to detect common mental states with high precision and sensitivity. Furthermore, performance improvements were observed in the Depression category, which had previously posed a challenge due to its highly variable and implicit linguistic expressions. The F1-score for this class increased from the highest baseline value of 0.73 to 0.77 following the implementation of the hybrid model. A similar enhancement occurred in the Stress category, where the hybrid model exhibited a meaningful performance boost, suggesting that the combination of LR and SVM effectively compensates for each model's individual limitations—particularly in identifying minority classes that are often more difficult to detect accurately.

The strength of the hybrid model lies not only in its ability to accurately detect categories with dominant data, but also in maintaining stable performance across all label categories. This combination proves effective in improving classification for weaker labels without compromising performance on already well-predicted categories. Therefore, this integrative approach demonstrates itself as an efficient and adaptive strategy for handling multi-class classification problems in the domain of text-based mental health analysis. The success of this hybrid method also opens avenues for further exploration into more advanced ensemble techniques such as stacking or blending which may enhance sensitivity to rare or implicit categories within the context of psychological classification based on social data.

3.4 Application Implementation

This section discusses the implementation of the mental health detection system in the form of an Android-based mobile application. The primary objective of this application development is to provide an accessible early detection tool that operates in real-time, utilizing text analysis derived from users' personal posts or status updates. The implementation includes the integration of a pre-trained hybrid model, the design of a user-friendly interface that supports emotional comfort, and the provision of early intervention features in the form of personalized self-help suggestions based on the classification outcomes. This approach is intended to be not only informative but also solution-oriented, while ensuring the privacy of users is maintained. The results of the application implementation are illustrated in Figure 2.

The SmartRisk mobile application was developed with a simple and user-friendly interface where users input daily feelings for analysis. Using a hybrid model (Logistic Regression + SVM), the system classifies both mental status (Anxiety, Depression, Emotional Exhaustion, Stress, Healthy) and risk levels (Low, Moderate, High). Results are displayed empathetically with tailored self-help recommendations, making the app function as both a detection tool and a self-care guide. Beyond technical design, SmartRisk has strong practical relevance. In Indonesia, where mental health professionals are limited, it can serve as a first-line screening tool to support triage—prioritizing high-risk users for referral while guiding others with self-care. Aligned with national digital health strategies and SDG 3, the system demonstrates how AI can be translated into scalable, cost-effective interventions to strengthen public mental health services, particularly in resource-constrained settings.

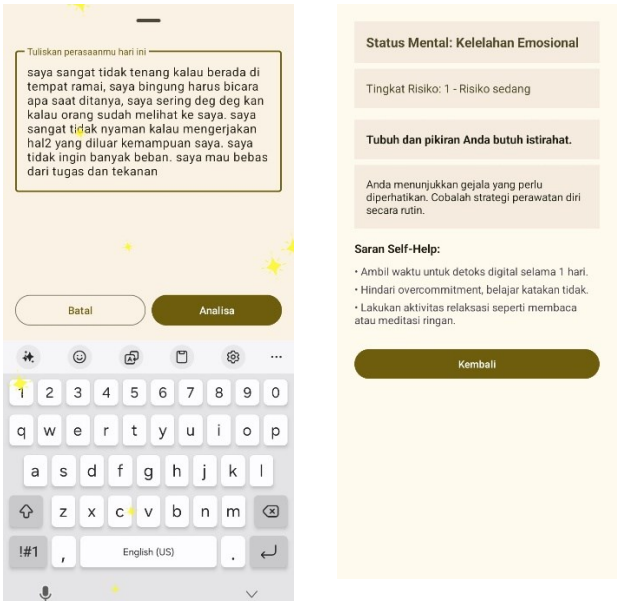

Figure 2. SMART-RISK Application

3.4.1 System Usability Testing

Based on the usability evaluation conducted using the System Usability Scale (SUS) method, the developed application demonstrated a high level of usability [29]. The evaluation involved 10 respondents who directly interacted with the core features of the application, namely submitting emotional expressions through text input and receiving mental health status classifications along with personalized self-help recommendations. From the SUS questionnaire consisting of 10 items rated on a 5-point Likert scale, an average score of 82.5 was obtained. This score falls within the "Good to Excellent" category of usability and corresponds to approximately the 90th percentile, indicating that the application performs above average compared to typical systems [30]. These results suggest that the application is perceived as highly usable, intuitive, and provides a positive user experience. However, it is important to note that the evaluation involved only 10 respondents, which represents a relatively small sample size. While such numbers are often acceptable for exploratory usability testing [14], they are insufficient for drawing robust conclusions about general usability across broader user groups. Moreover, the voluntary nature of participation may have introduced response bias, as the respondents could represent more digitally literate or motivated individuals. Thus, the SUS results should be interpreted as preliminary findings, providing an initial indication of usability rather than a definitive evaluation. Future usability assessments should involve a larger and more diverse respondent pool to enhance the reliability, representativeness, and generalizability of the findings.

IV. CONCLUSION

This study successfully designed and developed an early detection system for mental health conditions based on machine learning, capable of classifying users' psychological states through textual analysis of social media content. The system integrates two types of classification: mental status categories (Anxiety, Depression, Stress, Emotional Exhaustion, and Healthy) and levels of mental health risk (Low, Moderate, and High), using a hybrid approach that combines Logistic Regression and Support Vector Machine (SVM) models. Evaluation results demonstrate that the hybrid model outperforms individual baseline models, achieving an accuracy of 86% for mental status classification and 78% for risk level prediction. The most significant performance improvements were observed in the Depression and High Risk categories, which previously exhibited lower detection rates in baseline models, particularly in terms of F1-score. As a form of implementation, the system was integrated into an Android-based mobile application called SmartRisk, enabling users to receive rapid assessments by submitting narrative text inputs, accompanied by personalized and preventive self-help recommendations. From a user experience perspective, the usability evaluation using the System Usability Scale (SUS) yielded an average score of 82.5, indicating "good to excellent" usability. These findings suggest that the application is not only technically effective but also user-friendly, intuitive, and comfortable for end users. Beyond technical validation, SmartRisk has practical implications for mental health services in Indonesia. In a context where the ratio of mental health professionals remains far below global recommendations, the system can function as a first-line screening tool, supporting early detection and triage so that high-risk individuals can be prioritized for referral while others benefit from guided self-care. By leveraging mobile technology, SmartRisk offers a scalable and cost-effective solution to extend mental health support to underserved communities, aligning with national digital health strategies and the Sustainable Development Goal 3 (Good Health and Well-being). In conclusion, the developed system and application present strong potential not only as a research prototype but also as a practical digital intervention that bridges the gap between academic innovation and real-world mental health services. Future work will focus on validating the system with independent datasets, expanding usability testing to larger populations, and exploring integration with public health programs to maximize impact.

REFERENCES

[1] Rokom, "Kemenkes Beberkan Masalah Permasalahan Kesehatan Jiwa di Indonesia," https://sehatnegeriku.kemkes.go.id/.

[2] Rokom and S. N. Tarmizi, "Menjaga Kesehatan Mental Para Penerus Bangsa," https://sehatnegeriku.kemkes.go.id/. Accessed: Mar. 26, 2025. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20231012/3644025/menjaga-kesehatan-mental-para-penerus-bangsa/

[3] United Nations, "Department of Economic and Social Affairs Sustainable Development." Accessed: Mar. 24, 2025. [Online]. Available: https://sdgs.un.org/goals

[4] D. Liu, X. L. Feng, F. Ahmed, M. Shahid, and J. Guo, "Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review," *JMIR Ment Health*, vol. 9, no. 3, p. e27244, Mar. 2022, doi: 10.2196/27244.

[5] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," *Curr Opin Behav Sci*, vol. 18, pp. 43–49, 2017, doi: https://doi.org/10.1016/j.cobeha.2017.07.005.

[6] J. Koh, G. Y. Q. Tng, and A. Hartanto, "Potential and Pitfalls of Mobile Mental Health Apps in Traditional Treatment: An Umbrella Review," *J Pers Med*, vol. 12, no. 9, 2022, doi: 10.3390/jpm12091376.

[7] Q. Y. Leong *et al.*, "Characteristics of Mobile Health Platforms for Depression and Anxiety: Content Analysis Through a Systematic Review of the Literature and Systematic Search of Two App Stores," *J Med Internet Res*, vol. 24, no. 2, p. e27388, Feb. 2022, doi: 10.2196/27388.

[8] F. Efendi, G. E. Aurizki, A. Yusuf, and L. McKenna, "'Not shifting, but sharing': stakeholders' perspectives on mental health task-shifting in Indonesia.," *BMC Nurs*, vol. 21, no. 1, p. 165, Jun. 2022, doi: 10.1186/s12912-022-00945-8.

[9] R. W. Basrowi *et al.*, "Exploring Mental Health Issues and Priorities in Indonesia Through Qualitative Expert Consensus.," *Clin Pract Epidemiol Ment Health*, vol. 20, p. e17450179331951, 2024, doi: 10.2174/0117450179331951241022175443.

[10] D. I. Yasmine, F. Colombijn, A. J. A. M. van Deursen, and E. van Ingen, "Internet skills, digital engagement, and outcomes among urban youth in Jakarta, Indonesia: incorporating a Global South context in measuring digital inequality," *Inf Commun Soc*, vol. 4462, 2025, doi: 10.1080/1369118X.2025.2479778.

[11] M. Milne-Ives, E. Selby, B. Inkster, C. Lam, and E. Meinert, "Artificial intelligence and machine learning in mobile apps for mental health: A scoping review," *PLOS Digital Health*, vol. 1, no. 8, pp. 1–13, 2022, doi: 10.1371/journal.pdig.0000079.

[12] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches," *Sensors*, vol. 24, no. 2, 2024, doi: 10.3390/s24020348.

[13] A. Thieme, D. Belgrave, and G. Doherty, "Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems," *ACM Trans. Comput.-Hum. Interact.*, vol. 27, no. 5, Aug. 2020, doi: 10.1145/3398069.

[14] K. K. Putri and E. B. Setiawan, "Depression Detection in Indonesian X Social Media Text using Convolutional Neural Networks and Long Short-Term Memory with TF- IDF and FastText Methods," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 2, pp. 557–574, 2025, doi: 10.52436/1.jutif.2025.6.2.4206.

[15] M. L. McHugh, "Interrater reliability: the kappa statistic.," *Biochem Med (Zagreb)*, vol. 22, no. 3, pp. 276–282, 2012.

[16] J. Junadhi, A. Agustin, L. Efrizoni, F. Okmayura, H. Rahman, Dedi, and Muslim, "Improving Evaluation Metrics for Text Summarization: A Comparative Study and Proposal of a Novel Metric," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 885–896, May 2025, doi: 10.47738/jads.v6i2.547.

[17] A. Lubis, Irawan Yuda, Junadhi, and Defit Sarjon, "Leveraging K-Nearest Neighbors with SMOTE and Boosting Techniques for Data Imbalance and Accuracy Improvement," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1625–1638, Dec. 2024, doi: 10.47738/jads.v5i4.343.

[18] Erlin, A. Yunianta, L. A. Wulandhari, Y. Desnelita, N. Nasution, and Junadhi, "Enhancing Rice Production Prediction in Indonesia Using Advanced Machine Learning Models," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3478738.

[19] A. Qasim, G. Mehak, N. Hussain, A. Gelbukh, and G. Sidorov, "Detection of Depression Severity in Social Media Text Using Transformer-Based Models," *Information*, vol. 16, no. 2, 2025, doi: 10.3390/info16020114.

[20] M. Kanahuati-Ceballos and L. J. Valdivia, "Detection of depressive comments on social media using RNN, LSTM, and random forest: comparison and optimization," *Soc Netw Anal Min*, vol. 14, no. 1, p. 44, 2024, doi: 10.1007/s13278-024-01206-z.

[21] Junadhi, Agustin, M. Rifqi, and M. K. Anam, "Sentiment Analysis of Online Lectures using K-Nearest Neighbors based on Feature Selection," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 3, pp. 216–225, Dec. 2022, doi: 10.23887/janapati.v11i3.51531.

[22] D. K. Barupal and O. Fiehn, "Generating the blood exposome database using a comprehensive text mining and database fusion approach," *Environ Health Perspect*, vol. 127, no. 9, pp. 2825–2830, 2019, doi: 10.1289/EHP4713.

[23] K. Kusuma *et al.*, "The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: A meta-analysis and systematic review," *J Psychiatr Res*, vol. 155, pp. 579–588, 2022, doi: https://doi.org/10.1016/j.jpsychires.2022.09.050.

[24] H. Ghanadian, I. Nejadgholi, and H. Al Osman, "Improving Suicidal Ideation Detection in Social Media Posts: Topic Modeling and Synthetic Data Augmentation Approach," *JMIR Form Res*, vol. 9, 2025, doi: https://doi.org/10.2196/63272.

[25] M. Conciatori, A. Valletta, and A. Segalini, "Improving the quality evaluation process of machine learning algorithms applied to landslide time series analysis," *Comput Geosci*, vol. 184, p. 105531, 2024, doi: https://doi.org/10.1016/j.cageo.2024.105531.

[26] Dewi Widyawati and Amaliah Faradibah, "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indonesian Journal of Data and Science*, vol. 4, no. 2, pp. 80–89, Jul. 2023, doi: 10.56705/ijodas.v4i2.76.

[27] F. Nurjanah and S. E. Y. Suprihatin, "The development of Android-based learning media using Kodular in making suit patterns subject," *Jurnal Pendidikan Vokasi*, vol. 13, no. 3, pp. 232–245, 2023, doi: 10.21831/jpv.v13i3.54542.

[28] I. P. Laksana, E. D. Wahyuni, and C. S. K. Aditya, "Game Design for Mobile App-Based IoT Introduction Education in STEM Learning," *Jurnal RESTI*, vol. 7, no. 3, pp. 688–696, 2023, doi: 10.29207/resti.v7i3.5007.

[29] J. Riko Kurniawan Putra, Junadhi, Agustin, Herwin, F. Zoromi, and K. Andesa, "Analisis User Experience Sistem Tracer Alumni USTI dengan Metode SUS sebagai Upaya Peningkatan Kualitas Penggunaan," vol. 6, no. 1, 2025.

[30] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean; adding an adjective rating," *J Usability Stud*, vol. 4, no. 3, pp. 114–23, 2009.