ISSN 2085-4579

Depression Risk Classification Based on Self-Reported Psychosocial and Behavioral Survey Data Using Machine Learning

Marcelinus Jonathan Salim[1], Tegar Anugrah Firdaus[2], Carens Chanda Claudhyta Hasan[3], Yuri Pamungkas[4*]

[1-4]Department of Medical Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
yuri@its.ac.id

Accepted on March 27th, 2026
Approved on June 29th, 2026

***Abstract*—Mental health disorders, particularly depression, remain a major public health concern, while early risk identification outside clinical settings continues to pose significant challenges. The increasing availability of survey-based mental health data provides opportunities for data-driven risk screening, yet effective modeling strategies and evaluation practices remain underexplored. This study presents a comparative evaluation of multiple machine learning algorithms for depression risk classification using a publicly available mental health survey dataset. Rather than predicting clinical depression, the target variable is formulated as a risk proxy derived from social weakness indicators to support screening-oriented analysis. A quantitative experimental framework is employed to compare Logistic Regression, Random Forest, Support Vector Machine, and Extreme Gradient Boosting under consistent preprocessing and data partitioning conditions. Model performance is evaluated using complementary metrics, including accuracy, recall for High-risk cases, and the area under the receiver operating characteristic curve (ROC-AUC). Threshold optimization based on ROC analysis is applied to align model outputs with screening objectives that prioritize sensitivity. The results demonstrate that Logistic Regression and Support Vector Machine consistently achieve superior or comparable performance across all evaluation dimensions, including high overall accuracy, near-perfect sensitivity for High-risk detection, and strong discriminative capability. In contrast, more complex ensemble and distance-based models show mixed outcomes, indicating diminishing performance gains from increased algorithmic complexity. These findings highlight that simple and interpretable models can effectively support depression risk screening using survey-based data, offering a practical balance between predictive performance, transparency, and computational efficiency.**

***Index Terms*—Depression Risk Classification; Machine Learning; Mental Health Screening; Risk Proxy; Survey-based Data**

I. INTRODUCTION

Mental health disorders, particularly depression and closely related conditions, remain a significant global health challenge due to their persistent effects on emotional well-being, social functioning, and daily productivity [1],[2],[3]. Depression is associated not only with clinical impairment but also with broader social and economic consequences that often extend beyond the healthcare system [4],[5],[6]. Despite increased awareness, a substantial proportion of individuals experiencing depressive symptoms remain undetected, especially outside formal clinical environments [7],[8]. In recent years, large-scale survey-based mental health datasets have become increasingly available, enabling population-level analysis of psychological well-being and related risk factors [9],[10]. Such datasets provide an opportunity to explore early indicators of depression risk using self-reported behavioral and psychosocial attributes, particularly in contexts where clinical data are limited or unavailable [11],[12]. However, survey-based data are inherently heterogeneous and subjective, requiring analytical approaches capable of handling noisy and high-dimensional feature spaces [13].

Machine learning techniques have been widely adopted to address these challenges due to their ability to model complex, non-linear relationships among multiple variables [14],[15]. Compared to conventional statistical approaches, machine learning models can accommodate categorical attributes, capture interaction effects, and scale efficiently to larger datasets [16],[17]. Nevertheless, these models are not intended to replace clinical diagnosis, but rather to support exploratory analysis and early-stage risk screening within data-driven mental health research [18],[19]. Although numerous studies have applied machine learning to depression-related datasets, many focus on a single algorithm or rely primarily on accuracy as the sole evaluation metric. In imbalanced or perception-based datasets, such practices may obscure important performance trade-offs, as misclassification of high-risk individuals carries greater implications than overall correctness [20],[21]. Therefore, a more comprehensive evaluation framework that incorporates sensitivity-oriented and discrimination-based metrics is required to

properly assess model behavior in mental health applications.

Different machine learning algorithms exhibit distinct learning mechanisms and inductive biases, which may lead to varying predictive performance when applied to the same dataset [9],[10]. Linear models such as Logistic Regression offer interpretability and stability, while kernel-based methods like Support Vector Machines are effective in high-dimensional feature spaces [22],[23],[24]. Ensemble approaches, including Random Forest and Extreme Gradient Boosting (XGBoost), aim to improve robustness and generalization by aggregating multiple weak learners [22],[23]. In this study, a classification framework is developed using a secondary mental health survey dataset obtained via Mendeley Data [25]. Rather than predicting clinical depression, the target variable is formulated as a risk proxy derived from responses related to social weakness, which has been associated with psychosocial vulnerability and elevated depressive tendencies in population-based studies [17]. This proxy-based formulation is adopted to support methodological comparison and early risk screening, rather than clinical or diagnostic decision-making [18].

Model performance is evaluated using multiple complementary metrics to capture different aspects of predictive behavior [19]. Overall accuracy is reported to reflect general classification correctness, while recall is emphasized to assess sensitivity in identifying potential risk cases, particularly those belonging to the High-risk group. In addition, the area under the receiver operating characteristic curve (ROC-AUC) is employed to evaluate discriminative capability across varying decision thresholds [20]. To further align model outputs with screening-oriented objectives, threshold optimization based on ROC analysis is applied, prioritizing the reduction of missed High-risk cases [21]. The primary objective of this study is to conduct a systematic performance comparison of Logistic Regression, Random Forest, Support Vector Machine, and XGBoost under a consistent experimental framework [9]. By providing a transparent and reproducible comparative analysis, this study aims to highlight performance trade-offs among widely used machine learning algorithms and to contribute methodological insights for future research in mental health risk modeling [10].

II. Material & Methods

This methods section describes the study workflow for depression risk classification using machine learning, focusing on how the dataset was defined and analyzed within the chosen study design, how data were cleaned and transformed into model-ready representations, and how the train–test split strategy was applied to ensure fair and reproducible comparisons. It then details the machine learning models evaluated in this performance study, including their training procedures and key configurations. Finally, the evaluation metrics are specified to quantify and compare predictive performance in a consistent manner across all models.

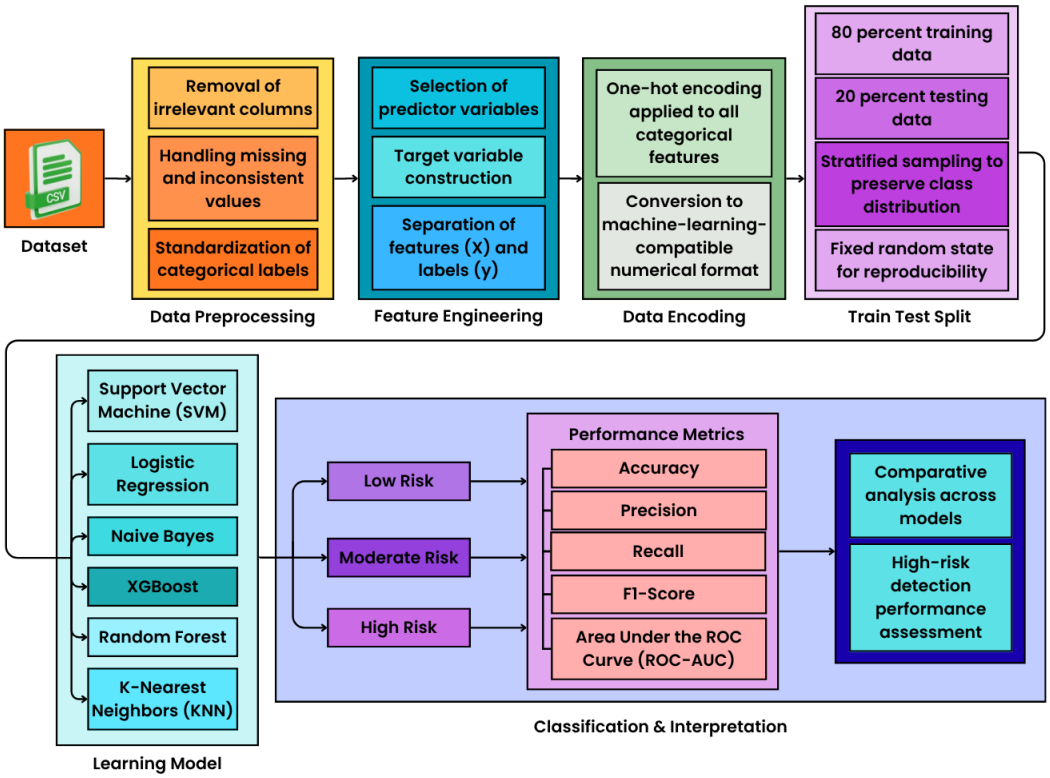

Fig. 1. Methodology of the research.

A. Dataset & Study Design

This study adopts a quantitative experimental design using a publicly available secondary dataset, originally titled "The RHMCD-20 Datasets for Depression and Mental Health Data Analysis with Machine Learning," published by Salehin and Amin via Mendeley Data [25]. The dataset consists of 824 structured survey responses, comprising 12 categorical predictor variables and one target variable (Social_Weakness). The predictor variables represent demographic, psychosocial, behavioral, and lifestyle-related factors associated with mental health conditions, including Age, Gender, Occupation, Days Indoors, Growing Stress, Quarantine Frustrations, Changes in Habits, Mental Health History, Weight Change, Mood Swings, Coping Struggles, and Work Interest. The target variable (Social_Weakness) consists of three response categories (Yes, Maybe, and No), with a distribution of 259 (31.4%), 287 (34.8%), and 278 (33.7%) samples, respectively. Since all predictor variables are categorical, one-hot encoding was applied during preprocessing to convert categorical responses into machine-readable binary representations for model development. Rather than directly predicting clinical depression, this study formulates a three-class depression risk proxy for classification based on the target variable, as described in the subsequent preprocessing stage.

B. Data Preprocessing and Representation

Prior to model training, the dataset undergoes a preprocessing stage to ensure compatibility with machine learning algorithms and consistency across experimental procedures. Because this study utilizes survey-based data and aims to support early risk screening rather than formal clinical diagnosis, the target variable is defined as a depression risk proxy rather than a definitive medical condition. This proxy was constructed directly from the respondents' answers to the Social Weakness attribute, which serves as the target variable in this study. The Social Weakness attribute was selected because it reflects respondent's perceived psychosocial vulnerability and provides an interpretable surrogate outcome for comparative machine learning analysis when clinical diagnostic labels are unavailable. The original response categories were mapped into three risk levels: Low-risk (No), Moderate-risk (Maybe), and High-risk (Yes). This proxy-based categorization enables comparative evaluation of machine learning models for early depression risk screening while avoiding interpretation as a clinical diagnosis.

Regarding the predictor variables, given the survey-based nature of the data, they are treated as categorical attributes reflecting respondents' discrete choices and self-reported conditions [26],[27]. These categorical features are transformed using one-hot encoding, a common and effective approach for representing nominal variables in machine learning models [28],[29],[30]. This technique converts each category into a binary indicator, allowing models to process categorical information without imposing artificial ordinal relationships that may bias learning outcomes [31],[32],[33]. Such a representation is particularly appropriate for survey-based datasets, where categorical responses do not imply inherent ordering. For instance, critical survey attributes such as Growing_Stress consist of responses ('Yes', 'No', 'Maybe'), Coping_Struggles consist of ('Yes', 'No'), and Work_Interest consist of ('Yes', 'No', 'Maybe'). After transforming all the categorical variables using one-hot encoding, the feature space expanded to a total of 39 independent features. Finally, all preprocessing steps are implemented within a unified processing pipeline to ensure identical transformations are applied to both training and testing data, thereby reducing the risk of data leakage and supporting reproducibility. No feature scaling or dimensionality reduction is applied, as the resulting binary-valued features are suitable for direct use in the selected classification algorithms.

C. Train Test Split Strategy

The analytical workflow began with data ingestion and risk-proxy target formulation, followed by one-hot encoding of all categorical features. Regarding the target variable, the risk proxy formulation yielded a relatively balanced class distribution: Moderate-risk (287 samples), Low-risk (278 samples), and High-risk (259 samples), resulting in a calculated class imbalance ratio of 1.11. To evaluate model performance under consistent and reproducible conditions, the dataset is partitioned into training and testing subsets using an 80:20 split ratio. This proportion is commonly adopted in supervised learning studies to provide sufficient data for model training while retaining an independent subset for unbiased performance evaluation [34],[35],[36]. Given the presence of class imbalance in the target variable, stratified sampling is applied during the splitting process to preserve the original class distribution across both subsets. This strategy helps prevent biased evaluation results that may arise when minority classes are underrepresented in either the training or testing data [37],[38],[39]. A fixed random seed is used to ensure reproducibility, and the same data partition is maintained across all evaluated models. By applying an identical train–test split for each algorithm, the comparative analysis reflects differences in model behavior rather than variations in data allocation [39].

D. Machine Learning Models

This study evaluates six supervised machine learning algorithms representing complementary classification paradigms. All models were trained and evaluated under identical preprocessing and data partitioning conditions to ensure a fair and consistent comparison. Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Naïve Bayes (NB), and K-Nearest Neighbors (KNN) were selected to represent linear, margin-based, ensemble, probabilistic, and distance-based learning approaches that are widely applied in structured tabular data analysis. This

selection encompasses linear and margin-based models (LR, SVM) known for stability in high dimensions, tree-based and boosting ensembles (Random Forest, XGBoost) recognized for handling complex interactions, as well as probabilistic and distance-based baselines (Naïve Bayes, KNN).

To ensure reproducibility, all experiments were implemented using Python 3.10 and the Scikit-learn library (version 1.2.2). A fixed random seed (random_state = 42) was applied consistently across all data partitioning and model initialization steps. Furthermore, to prevent data leakage and overfitting during hyperparameter tuning, a Stratified 5-Fold Cross-Validation approach was utilized on the training set prior to the final evaluation on the test set. The final hyperparameter settings for each machine learning algorithm were determined through the tuning process and subsequently used for the comparative evaluation. A summary of the selected hyperparameters is presented in Table I.

TABLE I. HYPERPARAMETER CONFIGURATION OF THE EVALUATED MACHINE LEARNING MODELS

Algoritm	Hyperparameter Configuration
Logistic Regression	solver = liblinear, C = 1.0, class_weight = balanced
Support Vector Machine	kernel = RBF, C = 1.0, probability = True, class_weight = balanced
Random Forest	n_estimators = 100, max_depth = None, class_weight = balanced
XGBoost	learning_rate = 0.1, max_depth = 6, objective = multi-softmax
K-Nearest Neighbors	n_neighbors = 5, weights = uniform, metric = minkowski
Naïve Bayes	alpha = 1.0, binarize = 0.0, fit_prior = True

1. *Logistic Regression (LR)*

Logistic Regression is employed as a linear baseline classifier due to its interpretability and suitability for binary risk modeling. The model estimates the probability of the positive class using a logistic function applied to a linear combination of input features [40],[41]:

$$P(y = 1 \mid x) = \frac{1}{1 + \exp\left(-(w^{\top}x + b)\right)} \quad (1)$$

where x denotes the feature vector, w is the weight vector, and b is the bias term. To mitigate class imbalance, balanced class weights are applied during training.

2. *Support Vector Machine (SVM)*

Support Vector Machine is a margin-based classifier that seeks an optimal hyperplane maximizing the separation between classes [23],[42]. The optimization objective can be expressed as:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \text{ subject to } y_i(w^{\top}x_i + b) \geq 1 \quad (2)$$

In this study, SVM is configured to produce probabilistic outputs to support ROC-based evaluation, and class imbalance is handled using class weighting.

3. *Random Forest*

Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and aggregates their predictions to improve robustness and generalization [43],[44]. The final prediction is obtained through majority voting:

$$\hat{y} = mode\{h_1(x), h_2(x), \ldots, h_T(x)\} \quad (3)$$

where $h_T(x)$ denotes the prediction of the ttt-th decision tree and TTT represents the total number of trees. Balanced class weighting is employed to address class imbalance.

4. *Extreme Gradient Boosting*

XGBoost is a gradient boosting ensemble method that sequentially builds weak learners to minimize a regularized objective function, making it effective for structured tabular data [43],[45]. The objective function is defined as:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k}^{\Omega}(f_k) \quad (4)$$

with the regularization term:

$$\Omega(f) = \gamma^T + \frac{1}{2}\lambda||w||^2 \quad (5)$$

where $\ell(\cdot)$ denotes the loss function and $\Omega(\cdot)$ controls model complexity. Class imbalance is handled using the *scale_pos_weight* parameter derived from class proportions.

5. *Additional Baseline Models*

To provide a broader comparative context, K-Nearest Neighbors (KNN) and Bernoulli Naïve Bayes classifiers are included as supplementary baselines. These models are evaluated using the same preprocessing pipeline and train-test split to maintain comparability. Specifically, the K-Nearest Neighbors (KNN) algorithm was configured with k=5 neighbors (n_neighbors=5) using the standard Minkowski distance metric and uniform neighbor weighting. Meanwhile, the Bernoulli Naïve Bayes model was implemented with additive Laplace smoothing (α = 1.0) and binarization thresholds matching the one-hot encoded input feature space. The detailed parameter specifications for all evaluated models are summarized in Table I.

The overall experimental workflow consisted of four main stages: dataset preparation and risk proxy construction, categorical feature encoding through one-hot encoding, stratified train-test splitting, and comparative evaluation of six supervised machine learning algorithms. All models were trained under

identical preprocessing and data partitioning conditions and evaluated using accuracy, macro recall, macro F1-score, High-risk recall, and ROC-AUC to ensure a fair and consistent comparison.

III. Results and Discussions

This chapter reports the classification results obtained from the evaluated machine learning models and discusses their performance using accuracy, recall, and ROC-AUC metrics. The analysis focuses on comparative performance, sensitivity to high-risk cases, and the relationship between model complexity and predictive effectiveness.

A. Overall Classification Performance

The overall classification results indicate that the evaluated models exhibit distinct performance profiles when applied to survey-based depression risk data. Figure 2 provides a visual summary of overall classification performance across the selected models using accuracy, macro-averaged recall, and macro F1-score. Logistic Regression and Support Vector Machine demonstrate the most consistent performance across all evaluation metrics, achieving the highest accuracy values (0.95) alongside strong macro-averaged recall and F1-scores. Despite their relatively simple decision boundaries, both models effectively capture the dominant patterns embedded in the one-hot encoded categorical features. This suggests that the underlying structure of the dataset allows linear or near-linear classifiers to perform competitively.

Random Forest achieves slightly lower overall performance, with an accuracy of 0.91 and reduced macro-averaged recall compared to the linear models. Although ensemble learning improves robustness by aggregating multiple decision trees, the observed performance gain remains limited in this context. The radar chart in Figure 1 highlights this trade-off, where Random Forest exhibits a smaller performance area despite its higher model complexity. Overall, the results show that increasing algorithmic complexity does not necessarily translate into superior classification performance. Instead, simpler and more interpretable models, such as Logistic Regression and Support Vector Machine, provide a favorable balance between predictive accuracy and model stability for depression risk classification based on survey data.

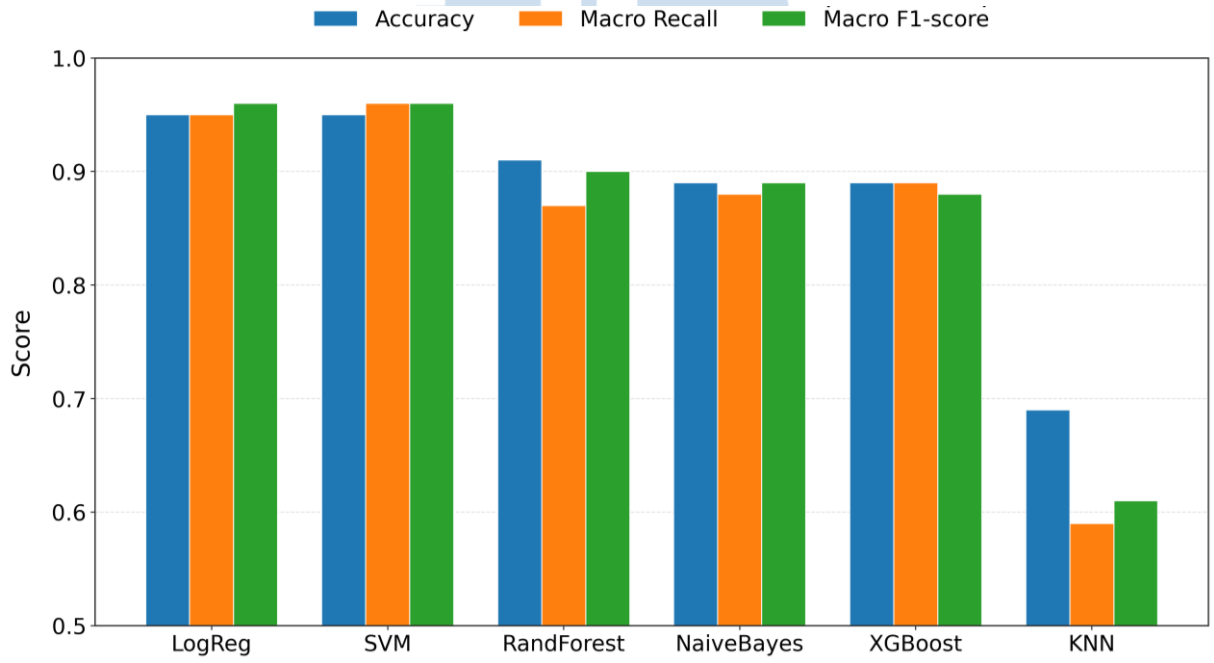

Fig. 2. Overall classification performance including accuracy, macro recall, and macro F1-score for all evaluated models.

B. Sensitivity to High Risk Cases (Recall)

Sensitivity to high-risk cases is a critical evaluation aspect in depression risk classification, as false negatives may delay timely intervention. Therefore, this subsection focuses on recall for the High-risk class to assess each model's ability to correctly identify individuals with elevated risk. The results provided in Figure 3 indicate that Support Vector Machine achieves perfect recall (1.00) for the High-risk class, followed closely by Logistic Regression (0.96). These findings suggest that both margin-based and linear classifiers are highly effective in capturing discriminative patterns associated with high-risk responses in the survey data. Their strong performance aligns with the objective of risk screening, where prioritizing sensitivity over specificity is often preferred.

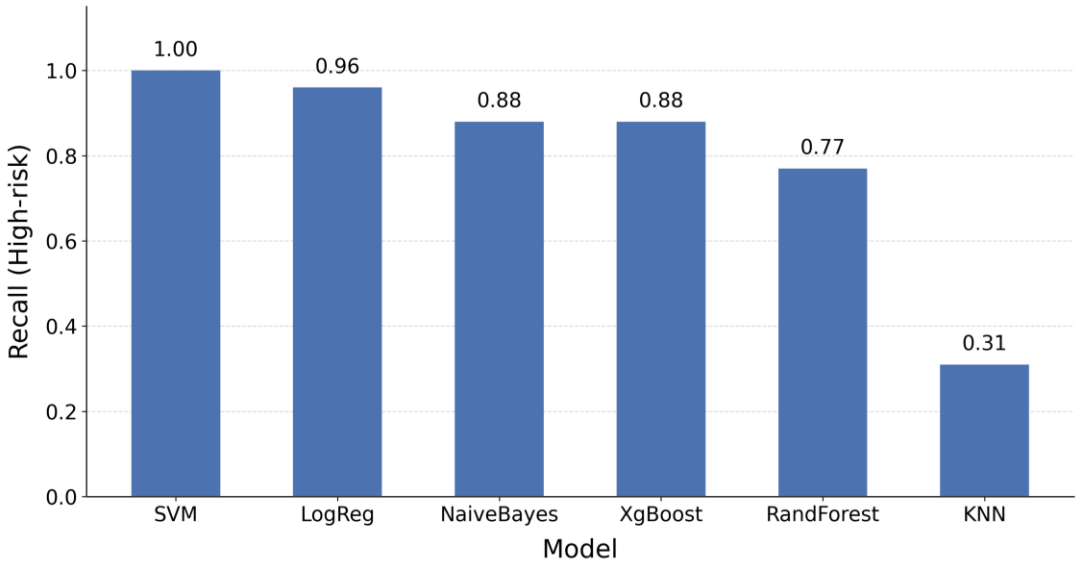

Fig. 3. Sensitivity to high-risk cases for all evaluated models.

Naïve Bayes and Decision Tree demonstrate moderate sensitivity, each achieving a recall of 0.88. While these models correctly identify a substantial proportion of high-risk cases, their performance remains slightly inferior to that of Logistic Regression and SVM. Random Forest shows a lower recall value (0.77), indicating that a portion of high-risk instances is misclassified into adjacent categories, most notably the Moderate-risk class. In contrast, K-Nearest Neighbors exhibits the lowest sensitivity (0.31) for high-risk detection. This substantial drop highlights the limitations of distance-based classifiers in high-dimensional, sparse feature spaces generated by one-hot encoding. Overall, the results emphasize that model selection plays a crucial role in high-risk screening tasks, with simpler linear and margin-based approaches offering superior sensitivity compared to more complex or distance-dependent methods.

C. Discriminative Ability (ROC-AUC)

Discriminative ability was further examined using receiver operating characteristic (ROC) analysis for the High-risk class under a one-vs-rest setting. Unlike accuracy or recall, ROC-AUC provides a threshold-independent view of how well a model separates high-risk instances from the remaining categories across a continuum of decision thresholds. Overall, the ROC profiles indicate that Logistic Regression, Support Vector Machine, and Random Forest achieve near-perfect discrimination, with AUC values approaching unity. This suggests that these models consistently assign higher risk scores to High-risk cases than to Low or Moderate cases, even when the classification threshold is varied. The strong separation shown by these curves supports their suitability for screening-oriented applications where robust ranking performance is desirable.

In comparison, Naïve Bayes and Decision Tree show high but slightly reduced discriminative performance, indicating minor overlap in score distributions between High-risk and non–High-risk observations. The most noticeable performance gap is observed for K-Nearest Neighbors, which yields a substantially lower AUC and a ROC curve closer to the diagonal reference line. This pattern reflects weaker ranking capability in high-dimensional sparse feature spaces generated by one-hot encoding, consistent with its lower sensitivity reported in the previous subsection. Taken together, the ROC analysis reinforces the earlier findings on sensitivity: models with linear or margin-based decision structures provide strong and stable discrimination for High-risk detection, whereas distance-based classification is less reliable in this setting.

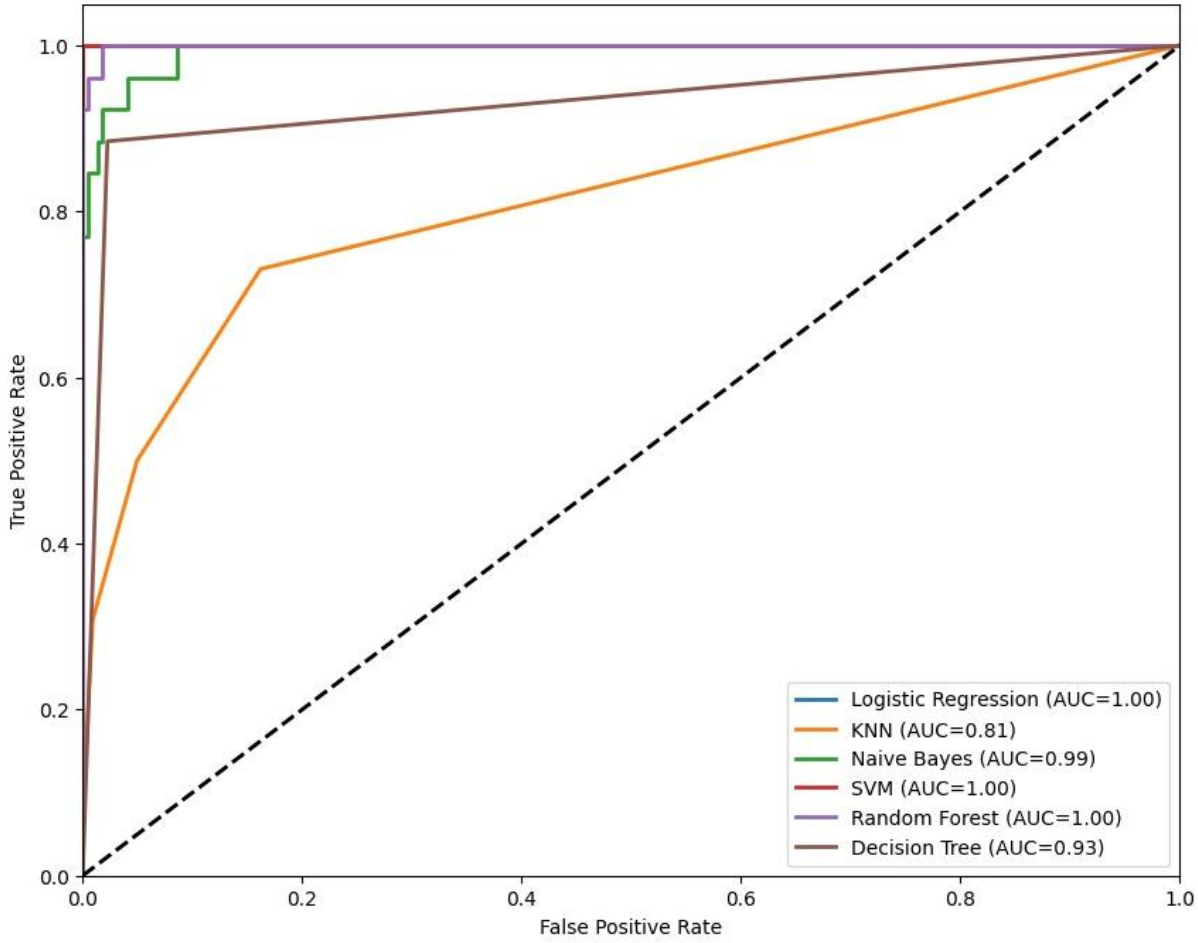

Fig. 4. Combined ROC curves for all evaluated models in detecting the high-risk class.

D. Model Complexity vs Performance

The relationship between model complexity and classification performance is a key consideration in practical depression risk assessment. While more complex algorithms are often expected to yield superior results, the findings of this study reveal that increased model complexity does not necessarily lead to improved performance when applied to survey-based mental health data. Across all evaluation dimensions overall performance, sensitivity to High-risk cases, and discriminative ability Logistic Regression and Support Vector Machine consistently match or outperform more complex models, including Random Forest and XGBoost. In contrast, ensemble-based and distance-based approaches exhibit mixed outcomes, indicating diminishing returns from increased structural complexity.

The superior performance of these simpler, linear, or margin-based models can be directly attributed to the characteristics of the preprocessed dataset. Survey data, which heavily relies on categorical responses, was transformed using one-hot encoding, inherently generating a high-dimensional and highly sparse feature space. In such sparse environments, linear models and SVMs excel because they can efficiently identify a global separating hyperplane without being heavily compromised by the sparsity; their strong performance suggests that the underlying relationships within the dataset are sufficiently linear or margin-separable. Conversely, tree-based ensemble models rely on greedy, recursive data splitting. When features are sparse and binary, individual splits yield very low information gain, making it difficult for tree architectures to construct robust decision paths, which often leads to suboptimal generalization. Similarly, distance-based algorithms like K-Nearest Neighbors show notable performance degradation and suffer significantly from the 'curse of dimensionality' in this one-hot encoded space, as the Euclidean distance between sparse binary vectors becomes uniform and loses its discriminative power. Thus, the inherent linearity of the transformed categorical data makes simpler models structurally better suited for this specific dataset.

From a practical perspective, the results emphasize the advantages of simpler and more interpretable models for depression risk screening. Models such as Logistic Regression and Support Vector Machine not only achieve high predictive accuracy but also offer greater transparency, ease of deployment, and lower computational overhead. These characteristics are

especially important in real-world mental health applications, where interpretability and reliability are often as critical as predictive performance.

IV. CONCLUSION

This study presents a comparative evaluation of multiple machine learning algorithms for depression risk classification using survey-based mental health data. Model performance was assessed in terms of overall accuracy, sensitivity to High-risk cases, discriminative ability, and the relationship between model complexity and predictive performance. The results show that Logistic Regression and Support Vector Machine consistently outperform or match more complex models across all evaluation dimensions. Both approaches achieve high accuracy, strong sensitivity for High-risk detection, and excellent discriminative capability, indicating that linear and margin-based classifiers are sufficient to capture dominant patterns in categorical mental health survey data. In contrast, ensemble and distance-based methods demonstrate mixed performance, with no clear advantage from increased algorithmic complexity. From a practical perspective, these findings suggest that simple and interpretable models provide an effective balance between performance and usability for depression risk screening. Although this study is limited by its reliance on a single self-reported dataset, future work may explore external validation and alternative feature representations to further enhance model generalizability.

ACKNOWLEDGMENT

The authors would like to acknowledge the Department of Medical Technology, Institut Teknologi Sepuluh Nopember, for the facilities and support in this research.

REFERENCES

[1] C. Stecher, S. Cloonan, and M. E. Domino, "The Economics of Treatment for Depression," Annu. Rev. Public Health, vol. 45, no. 1, pp. 527–551, May 2024, doi: 10.1146/annurev-publhealth-061022-040533.

[2] D. Proudman, P. Greenberg, and D. Nellesen, "The Growing Burden of Major Depressive Disorders (MDD): Implications for Researchers and Policy Makers," Pharmacoeconomics, vol. 39, no. 6, pp. 619–625, Jun. 2021, doi: 10.1007/s40273-021-01040-7.

[3] K. S. Chaudhari, M. P. Dhapkas, A. Kumar, and R. G. Ingle, "Mental Disorders – A Serious Global Concern that Needs to Address," International Journal of Pharmaceutical Quality Assurance, vol. 15, no. 02, pp. 973–978, Jun. 2024, doi: 10.25258/ijpqa.15.2.66.

[4] A. Kupferberg and G. Hasler, "The social cost of depression: Investigating the impact of impaired social emotion regulation, social cognition, and interpersonal behavior on social functioning," J. Affect. Disord. Rep., vol. 14, p. 100631, Dec. 2023, doi: 10.1016/j.jadr.2023.100631.

[5] P. Greenberg and others, "The Economic Burden of Adults with Major Depressive Disorder in the United States (2019)," Adv. Ther., vol. 40, no. 10, pp. 4460–4479, Oct. 2023, doi: 10.1007/s12325-023-02622-x.

[6] P. Cuijpers and C. F. Reynolds, "Increasing the Impact of Prevention of Depression—New Opportunities," JAMA Psychiatry, vol. 79, no. 1, p. 11, Jan. 2022, doi: 10.1001/jamapsychiatry.2021.3153.

[7] G. Purebl, K. Schnitzspahn, and É. Zsák, "Overcoming treatment gaps in the management of depression with non-pharmacological adjunctive strategies," Front. Psychiatry, vol. 14, Nov. 2023, doi: 10.3389/fpsyt.2023.1268194.

[8] A. Faisal-Cury, C. Ziebold, D. M. de O. Rodrigues, and A. Matijasevich, "Depression underdiagnosis: Prevalence and associated factors. A population-based study," J. Psychiatr. Res., vol. 151, pp. 157–165, Jul. 2022, doi: 10.1016/j.jpsychires.2022.04.025.

[9] K. L. Mansfield, S. Puntis, E. Soneson, A. Cipriani, G. Geulayov, and M. Fazel, "Study protocol: the OxWell school survey investigating social, emotional and behavioural factors associated with mental health and well-being," BMJ Open, vol. 11, no. 12, p. e052717, Nov. 2021, doi: 10.1136/bmjopen-2021-052717.

[10] J. J. Newson, J. Bala, J. N. Giedd, B. Maxwell, and T. C. Thiagarajan, "Leveraging big data for causal understanding in mental health: a research framework," Front. Psychiatry, vol. 15, Feb. 2024, doi: 10.3389/fpsyt.2024.1337740.

[11] M. Z. Uddin, K. K. Dysthe, A. Følstad, and P. B. Brandtzaeg, "Deep learning for prediction of depressive symptoms in a large textual dataset," Neural Comput. Appl., vol. 34, no. 1, pp. 721–744, Jan. 2022, doi: 10.1007/s00521-021-06426-4.

[12] M. Garg, "Mental Health Analysis in Social Media Posts: A Survey," Archives of Computational Methods in Engineering, vol. 30, no. 3, pp. 1819–1842, Apr. 2023, doi: 10.1007/s11831-022-09863-z.

[13] M. Salama, H. A. Kader, and A. Abdelwahab, "An analytic framework for enhancing the performance of big heterogeneous data analysis," International Journal of Engineering Business Management, vol. 13, Jan. 2021, doi: 10.1177/1847979021990523.

[14] D. Feldner-Busztin and others, "Dealing with dimensionality: the application of machine learning to multi-omics data," Bioinformatics, vol. 39, no. 2, Feb. 2023, doi: 10.1093/bioinformatics/btad021.

[15] L. K. Andersen and B. J. Reading, "A supervised machine learning workflow for the reduction of highly dimensional biological data," Artificial Intelligence in the Life Sciences, vol. 5, p. 100090, Jun. 2024, doi: 10.1016/j.ailsci.2023.100090.

[16] W. Li and K. M. Kockelman, "How does machine learning compare to conventional econometrics for transport data sets? A test of ML versus MLE," Growth Change, vol. 53, no. 1, pp. 342–376, Mar. 2022, doi: 10.1111/grow.12587.

[17] S. K. Dhillon, M. D. Ganggayah, S. Sinnadurai, P. Lio, and N. A. Taib, "Theory and Practice of Integrating Machine Learning and Conventional Statistics in Medical Data Analysis," Diagnostics, vol. 12, no. 10, p. 2526, Oct. 2022, doi: 10.3390/diagnostics12102526.

[18] Md. M. Islam, S. Hassan, S. Akter, F. A. Jibon, and Md. Sahidullah, "A comprehensive review of predictive analytics models for mental illness using machine learning algorithms," Healthcare Analytics, vol. 6, p. 100350, Dec. 2024, doi: 10.1016/j.health.2024.100350.

[19] D. Ostojic, P. A. Lalousis, G. Donohoe, and D. W. Morris, "The challenges of using machine learning models in psychiatric research and clinical practice," European Neuropsychopharmacology, vol. 88, pp. 53–65, Nov. 2024, doi: 10.1016/j.euroneuro.2024.08.005.

[20] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced Big Data," J. Big Data, vol. 10, no. 1, p. 42, Apr. 2023, doi: 10.1186/s40537-023-00724-5.

[21] C. Mosquera, L. Ferrer, D. H. Milone, D. Luna, and E. Ferrante, "Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance," Eur. Radiol., vol. 34, no. 12, pp. 7895–7903, Jun. 2024, doi: 10.1007/s00330-024-10834-0.

[22] L. Wu, R. Huang, I. V Tetko, Z. Xia, J. Xu, and W. Tong, "Trade-off Predictivity and Explainability for Machine-

Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets," Chem. Res. Toxicol., vol. 34, no. 2, pp. 541–549, Feb. 2021, doi: 10.1021/acs.chemrestox.0c00373.

[23] K.-L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions," Mathematics, vol. 12, no. 24, p. 3935, Dec. 2024, doi: 10.3390/math12243935.

[24] M. Gimeno, K. S. del Real, and A. Rubio, "Precision oncology: a review to assess interpretability in several explainable methods," Brief. Bioinform., vol. 24, no. 4, Jul. 2023, doi: 10.1093/bib/bbad200.

[25] I. Salehin and N. Amin, "The RHMCD-20 Datasets for Depression and Mental Health Data Analysis with Machine Learning," 2024. doi: 10.17632/pxjmjyfdh2.

[26] W. L. Ku and H. Min, "Evaluating Machine Learning Stability in Predicting Depression and Anxiety Amidst Subjective Response Errors," Healthcare, vol. 12, no. 6, p. 625, Mar. 2024, doi: 10.3390/healthcare12060625.

[27] A. Dixit, A. K. Gupta, N. Chaplot, and V. Bharti, "Emoji Support Predictive Mental Health Assessment Using Machine Learning: Unveiling Personalized Intervention Avenues," International Journal of Experimental Research and Review, vol. 42, pp. 228–240, Aug. 2024, doi: 10.52756/ijerr.2024.v42.020.

[28] M. R. Rezvan, A. G. Sorkhi, J. Pirgazi, and M. M. P. Kallehbasti, "AdvanceSplice: Integrating N-gram one-hot encoding and ensemble modeling for enhanced accuracy," Biomed. Signal Process. Control, vol. 92, p. 106017, Jun. 2024, doi: 10.1016/j.bspc.2024.106017.

[29] S. Bagui, D. Nandi, S. Bagui, and R. J. White, "Machine Learning and Deep Learning for Phishing Email Classification using One-Hot Encoding," Journal of Computer Science, vol. 17, no. 7, pp. 610–623, Jul. 2021, doi: 10.3844/jcssp.2021.610.623.

[30] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," Comput. Stat., vol. 37, no. 5, pp. 2671–2692, Nov. 2022, doi: 10.1007/s00180-022-01207-6.

[31] W. Yustanti, N. Iriawan, and I. Irhamah, "Categorical encoder based performance comparison in pre-processing imbalanced multiclass classification," Indonesian Journal of Electrical Engineering and Computer Science, vol. 31, no. 3, p. 1705, Sep. 2023, doi: 10.11591/ijeecs.v31.i3.pp1705-1715.

[32] Y. Sun and others, "Modifying the one-hot encoding technique can enhance the adversarial robustness of the visual model for symbol recognition," Expert Syst. Appl., vol. 250, p. 123751, Sep. 2024, doi: 10.1016/j.eswa.2024.123751.

[33] M. Klimo, P. Lukáč, and P. Tarábek, "Deep Neural Networks Classification via Binary Error-Detecting Output Codes," Applied Sciences, vol. 11, no. 8, p. 3563, Apr. 2021, doi: 10.3390/app11083563.

[34] M. Fallahpoor, S. Chakraborty, M. T. Heshejin, H. Chegeni, M. J. Horry, and B. Pradhan, "Generalizability assessment of COVID-19 3D CT data for deep learning-based disease detection," Comput. Biol. Med., vol. 145, p. 105464, Jun. 2022, doi: 10.1016/j.compbiomed.2022.105464.

[35] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," PeerJ Comput. Sci., vol. 10, p. e2245, Sep. 2024, doi: 10.7717/peerj-cs.2245.

[36] S. N. Selamat, N. A. Majid, and A. M. Taib, "A Comparative Assessment of Sampling Ratios Using Artificial Neural Network (ANN) for Landslide Predictive Model in Langat River Basin, Selangor, Malaysia," Sustainability, vol. 15, no. 1, p. 861, Jan. 2023, doi: 10.3390/su15010861.

[37] M. Kim and K.-B. Hwang, "An empirical evaluation of sampling methods for the classification of imbalanced data," PLoS One, vol. 17, no. 7, p. e0271260, Jul. 2022, doi: 10.1371/journal.pone.0271260.

[38] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, and C. Zhang, "Stratified Sampling-Based Deep Learning Approach to Increase Prediction Accuracy of Unbalanced Dataset," Electronics (Basel)., vol. 12, no. 21, p. 4423, Oct. 2023, doi: 10.3390/electronics12214423.

[39] A. Jude and J. Uddin, "Explainable Software Defects Classification Using SMOTE and Machine Learning," Annals of Emerging Technologies in Computing, vol. 8, no. 1, pp. 36–49, Jan. 2024, doi: 10.33166/AETiC.2024.01.004.

[40] S. Kumar and V. Gota, "Logistic regression in cancer research: A narrative review of the concept, analysis, and interpretation," Cancer Research, Statistics, and Treatment, vol. 6, no. 4, pp. 573–578, Oct. 2023, doi: 10.4103/crst.crst_293_23.

[41] A. Zaidi and A. S. M. Al Luhayb, "Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic Regression," Math. Probl. Eng., vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/5525675.

[42] A. Rizwan, N. Iqbal, R. Ahmad, and D.-H. Kim, "WR-SVM Model Based on the Margin Radius Approach for Solving the Minimum Enclosing Ball Problem in Support Vector Machine Classification," Applied Sciences, vol. 11, no. 10, p. 4657, May 2021, doi: 10.3390/app11104657.

[43] M. Azad, T. H. Nehal, and M. Moshkov, "A novel ensemble learning method using majority based voting of multiple selective decision trees," Computing, vol. 107, no. 1, p. 42, Jan. 2025, doi: 10.1007/s00607-024-01394-8.

[44] Y. Resti, C. Irsan, J. F. Latif, I. Yani, and N. R. Dewi, "A Bootstrap-Aggregating in Random Forest Model for Classification of Corn Plant Diseases and Pests," Science and Technology Indonesia, vol. 8, no. 2, pp. 288–297, Apr. 2023, doi: 10.26554/sti.2023.8.2.288-297.

[45] S. Gaïffas, I. Merad, and Y. Yu, "WildWood: A New Random Forest Algorithm," IEEE Trans. Inf. Theory, vol. 69, no. 10, pp. 6586–6604, Oct. 2023, doi: 10.1109/TIT.2023.3287432.