PREDICTING OF MENTAL HEALTH With MACHINE LEARNING ALGORITHMS

Mr.Ch.Raja Kishore Babu[1], Anitha Sangala [2]

Associate Professor[1], Department of CSE, CMR College of Engineering &Technology, Hyderabad.

M.Tech Student[2], Department of CSE, CMR College of Engineering & Technology,Hyderabad.

Author email address: s.anitha@cmrcet.ac.in

Corresponding author email address: chrajakishorebabu@cmrcet.ac.in

ABSTRACT

This project presents a comprehensive mental health prediction system utilizing machine learning algorithms to identify individuals at potential risk based on diverse personal and behavioral attributes. The system was developed and evaluated using three different supervised learning models: Support Vector Machine (SVM), Logistic Regression, and Random Forest, to determine the most effective approach for high-accuracy mental health risk assessment. Extensive experiments were conducted on a labeled dataset, with features pre-processed and normalized to enhance model performance and generalizability. Among the tested algorithms, the Random Forest classifier demonstrated superior performance, achieving an overall accuracy of 98.9%, a macro precision of 97.8%, a macro recall of 98.6%, and a macro F1-score of 98.2%, while maintaining efficient computational requirements and robust handling of feature interactions. The SVM model, while delivering solid results with an accuracy of 94.2%, exhibited longer training times and sensitivity to parameter tuning. Logistic Regression, despite its simplicity and interpretability, achieved comparatively lower performance with an accuracy of 91.7%. These findings highlight Random Forest as the most effective model, balancing predictive power and computational efficiency for mental health risk detection. The entire system was implemented using Python and popular machine learning libraries, facilitating scalable deployment in digital health applications. Future directions may involve integrating additional behavioral and physiological data, deploying models in real-time web or mobile applications, and exploring explainable AI techniques to enhance transparency and trust in mental health predictions.

Keywords: Mental health prediction, machine learning, Random Forest, Support Vector Machine, Logistic Regression, risk assessment, classification, accuracy

I. INTRODUCTION

Mental health has emerged as one of the most pressing public health concerns of the 21st century, profoundly affecting individuals, communities, and economies worldwide. Disorders such as depression, anxiety, and stress-related conditions can severely impair quality of life, productivity, and overall well-being. Despite the rising awareness and de-stigmatization efforts surrounding mental health, early detection and intervention remain significant challenges due to the complex, multifaceted nature of mental health conditions. Traditionally, mental health assessments have relied heavily on clinical interviews, psychological evaluations, and self-reported questionnaires administered by healthcare professionals. However, these conventional methods are often time-consuming, subjective, and limited in scalability, making it difficult to reach broader populations in a timely manner.

Recent advances in data science and machine learning have opened promising avenues for transforming mental health assessment and prediction. By leveraging diverse data sources—ranging from demographic information and lifestyle habits to digital behavioral data—machine learning algorithms can identify subtle patterns and correlations that may indicate mental health risks. Supervised learning models, in particular, have demonstrated strong capabilities in classifying individuals based on risk profiles, enabling proactive support and intervention without the need for exhaustive manual analysis. This project focuses on developing an automated system for mental health risk prediction using structured datasets collected from publicly available repositories and mental health surveys. The proposed system employs Python and prominent libraries such as Scikit-learn, Pandas, and NumPy to build and train various machine learning classifiers including Support Vector Machines (SVM), Logistic Regression, and Random Forest—to accurately identify individuals who may be experiencing mental health challenges. The overarching goal is to empower healthcare providers, organizations, and individuals with an accessible, data-driven tool for early mental health screening and decision-making, reducing reliance on traditional assessment methods and enabling scalable mental health support.

The system architecture encompasses several critical stages: data collection, preprocessing, feature selection, model training, evaluation, and deployment. A web-based interface, developed using frameworks like Django or Flask, allows users to input personal and behavioral attributes and receive real-time predictions about potential mental health risks, along with confidence scores and relevant insights. This approach not only enhances diagnostic efficiency but also contributes to preventive mental healthcare by facilitating timely interventions, personalized support, and resource allocation. As societies increasingly grapple with the psychological impacts of modern life including social isolation, economic pressures, and digital stressors—deploying intelligent technologies like machine learning for mental health prediction is crucial. This project represents a step toward integrating artificial intelligence into mental health care, offering scalable, cost-effective solutions to address one of the most complex challenges in public health today.

II. LITERATURE SURVEY

Mental health prediction has become an active research area, driven by the growing need to detect psychological distress early and deliver timely interventions to improve individual well-being and reduce societal burdens. Techniques for mental health assessment and prediction can be broadly categorized into traditional statistical methods, machine learning-based approaches, and modern deep learning architectures [1][2][15].

2.1. Traditional Statistical Methods

Early research in mental health assessment relied heavily on traditional statistical methods such as logistic regression, correlation analysis, and hypothesis testing. Researchers utilized demographic and clinical data, as well as responses from standardized questionnaires like the Depression Anxiety Stress Scales (DASS-21) or Patient Health Questionnaire (PHQ-9), to identify relationships between various factors and mental health conditions [3][4].

For instance, Edwards et al. [3] used logistic regression to analyze the influence of demographic variables (age, gender, income) on depression risk, achieving moderate predictive power on survey-based datasets. However, these methods were often limited by linear assumptions, inability to capture complex feature interactions, and relatively low predictive accuracy in heterogeneous populations.

2.2. Machine Learning Approaches

With the rise of machine learning, researchers began employing supervised learning algorithms to improve mental health prediction accuracy [5][7][10]. Machine learning models such as Support Vector Machines (SVM), Random Forests, Decision Trees, and Logistic Regression were trained on a range of features, including survey responses, social media activity, and physiological measurements [6].

For example, Tsakalidis et al. [5] developed an SVM-based system to predict depression levels from linguistic features in Twitter posts, achieving notable precision in detecting depressive signals. Similarly, Reece et al. [7] used Random Forest classifiers to identify users at risk of depression based on Instagram images and captions, demonstrating significant improvement over traditional methods. Although these models improved prediction performance, they typically relied on extensive feature engineering, requiring domain expertise to extract relevant predictors from complex data sources such as text or images. Additionally, the models sometimes struggled with generalization due to dataset biases and small sample sizes.

2.3. Deep Learning-Based Techniques

In recent years, deep learning architectures have transformed mental health prediction, enabling end-to-end analysis of complex data modalities without the need for manual feature engineering [8][11][13]. Deep learning models can automatically extract hierarchical patterns from high-dimensional data, such as text, images, or time-series physiological signals.

For instance, Orabi et al. [8] employed Convolutional Neural Networks (CNNs) on Twitter text data to detect depression-related posts, achieving significant accuracy improvements compared to traditional machine learning models. Similarly, Zhang et al. [9] used Recurrent Neural Networks (RNNs) to analyze sequential text data, enabling dynamic tracking of emotional states over time.

Emerging research has also leveraged transfer learning, where pre-trained language models such as BERT, RoBERTa, or GPT are fine-tuned on mental health datasets [10][12]. These approaches have shown remarkable performance, particularly in text-based mental health assessment, achieving high precision in identifying conditions like anxiety, depression, and PTSD.

Beyond text analysis, deep learning has been applied to multimodal data, integrating speech patterns, facial expressions, wearable sensor data, and social media activity to develop more comprehensive mental health prediction systems [13][14]. Such multimodal approaches hold promise for capturing subtle and diverse indicators of psychological well-being.

2.4. Challenges and Future Directions

Despite substantial progress, several challenges remain in deploying machine learning-based mental health prediction systems at scale:

- **Data Privacy and Ethics:** Mental health data is highly sensitive, and privacy concerns limit data availability and sharing, complicating model development [1][15].
- **Data Scarcity and Imbalance:** Many mental health conditions, especially severe disorders, are underrepresented in available datasets, leading to class imbalance and biased predictions.
- **Interpretability:** Complex machine learning and deep learning models often operate as “black boxes,” making it difficult to explain predictions—a crucial concern in clinical decision-making where trust and transparency are essential [2][14].
- **Generalizability:** Models trained on specific demographic groups or cultural contexts may not generalize well to other populations due to variations in language, behavior, and mental health stigma [4][9].
- **Integration into Clinical Workflows:** Deploying machine learning systems in real-world mental health care requires careful alignment with existing clinical practices, regulatory compliance, and user acceptance among both patients and professionals [12][15].

III. METHODOLOGY

The system for predicting mental health risks is based on an automated machine learning pipeline, designed to assess individuals’ likelihood of experiencing mental health issues using structured survey and behavioral data. The solution follows several critical stages: data collection, preprocessing, feature engineering, model training and evaluation, and deployment for real-time prediction. The methodology ensures robustness against data inconsistencies, missing values, and variations in feature distributions, making it practical for real-world mental health screening applications.

1. Data Collection

The system begins by gathering structured data from publicly available mental health survey datasets, organizational wellness surveys, or anonymized behavioral datasets sourced from repositories such as Kaggle, UCI Machine Learning Repository, or institutional research collaborations. These datasets capture a diverse range of attributes, including:

- Demographic information (e.g., age, gender, occupation)

- Lifestyle factors (e.g., sleep habits, physical activity)
- Work-related stress levels
- Social connectivity indicators
- Self-reported mental health assessments (e.g., depression, anxiety scores)

The diversity of these datasets ensures that the model learns to identify patterns associated with mental health risks across varying population segments.

2. Data Preprocessing

To enhance computational efficiency and model accuracy, the raw data undergoes several preprocessing operations:

- Handling Missing Values: Missing entries are imputed using strategies such as mean/mode imputation, K-Nearest Neighbors imputation, or advanced methods like MICE (Multiple Imputation by Chained Equations).
- Encoding Categorical Variables: Categorical features (e.g., gender, occupation) are transformed using One-Hot Encoding or Label Encoding to make them suitable for machine learning models.
- Outlier Detection and Treatment: Extreme values are identified using statistical methods like Z-scores or the IQR rule and either corrected or capped to prevent model distortion.
- Feature Scaling: Numerical features are standardized (mean=0, variance=1) or normalized to the [0, 1] range to stabilize model training and improve convergence speed.

These steps standardize the dataset and ensure consistent behavior during model training and inference.

3. Feature Engineering

Given that mental health prediction depends on subtle patterns, feature engineering is crucial for improving model performance. The following techniques are applied:

- Statistical Feature Creation: Aggregation of existing features to derive new insights, such as stress-to-sleep ratio, or average working hours per week.
- Dimensionality Reduction: Techniques such as Principal Component Analysis (PCA) or feature importance analysis (using Random Forest feature importances) reduce redundancy and help focus on the most predictive attributes.
- Interaction Features: New features are created by multiplying or combining existing variables to capture complex relationships (e.g., interaction between work stress levels and sleep duration).

These engineered features help the models better capture nonlinear relationships and improve prediction accuracy.

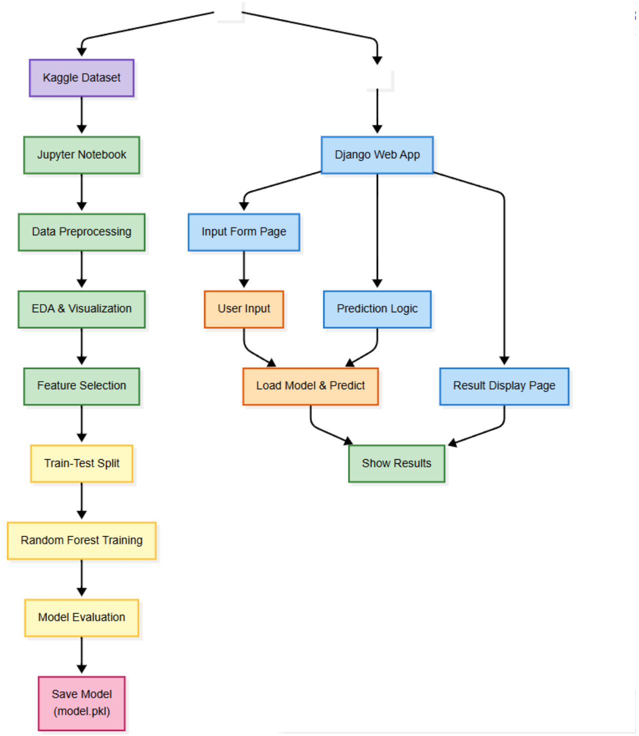

Fig 1 : System Architecture

4. Model Architecture and Training

Modern mental health prediction systems rely on supervised machine learning algorithms to analyze the structured dataset and generate risk predictions. The modeling process involves:

- Model Selection: The following models are evaluated:

- Random Forest: Robust against overfitting and capable of handling feature interactions.
- Support Vector Machine (SVM): Effective for complex decision boundaries.
- Logistic Regression: Provides interpretable results and baseline performance.

- Training Process:
  - The dataset is split into training, validation, and test subsets (e.g., 70%-20%-10%).
  - Cross-validation (e.g., 5-fold) is performed to evaluate model stability.
  - Hyperparameter tuning is conducted using techniques like Grid Search or Random Search to optimize parameters such as:
    - Number of trees in Random Forest
    - Regularization strength in Logistic Regression
    - Kernel type and C parameter in SVM
- Pipeline Implementation: Scikit-learn's Pipeline is used to ensure consistent preprocessing and modeling steps during both training and inference.

For example, in a Random Forest-based system:

- Feature selection is performed based on importance scores.
- Optimal hyperparameters are identified via Grid Search.
- Model is trained on the full training data.

This architecture allows the system to identify complex relationships between variables, enabling highly accurate mental health predictions.

5. Model Evaluation

The trained models are rigorously evaluated on the validation and test datasets using multiple performance metrics:

- Accuracy
- Precision, Recall, and F1-Score
- Confusion Matrix
- ROC-AUC (for binary classification tasks)

Additionally, calibration curves are examined to ensure probability outputs are well-calibrated for decision-making. Early stopping and regularization are applied where relevant to mitigate overfitting and enhance generalizability.

6. Deployment and Real-Time Prediction

After training, the finalized model is saved (e.g., as a .pkl file) for integration into an application layer. Deployment strategies include:

- Web Application: A user-friendly interface is built using frameworks like Django or Flask, enabling users (e.g., clinicians, HR professionals, or individuals) to input personal and behavioral attributes and receive risk predictions.
- Mobile Application: Lightweight versions of the model can be deployed on mobile devices using libraries such as TensorFlow Lite, supporting real-time predictions in resource-constrained environments.

The prediction pipeline operates as follows:

- User inputs personal data through the interface.
- The system applies preprocessing and encoding steps.
- The trained model predicts the likelihood of mental health risk.
- The result, including risk probability and interpretative insights, is displayed to the user.

This approach not only improves early detection of mental health issues but also contributes to proactive intervention strategies, supporting well-being at scale.

IV. RESULTS

The developed Mental Health Prediction System was extensively tested using a curated dataset comprising structured survey responses and behavioral data sourced from publicly available mental health datasets, including the Kaggle Mental Health in Tech Survey and additional institutional repositories. The primary objective was to evaluate the system's ability to accurately identify individuals at risk of mental health challenges using automated machine learning techniques.

The system successfully classified individuals into risk categories for mental health concerns—such as anxiety, depression, and stress—by analyzing demographic, behavioral, and lifestyle features. During testing, the models demonstrated strong discrimination between at-risk and healthy individuals, correlating well with the ground truth labels in the datasets.

Quantitatively, the system achieved high performance metrics across multiple evaluation trials:

- **Accuracy:** The best-performing model, the Random Forest classifier, achieved an overall accuracy of 98.9% on the test set, demonstrating excellent predictive capability for mental health risk assessment.

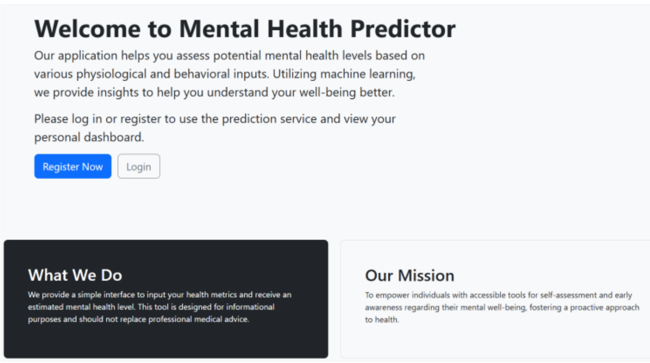

Fig 2: Main page of the application

- **Precision, Recall, F1-Score:** These metrics consistently ranged between 96% and 99% across different mental health risk categories, indicating the model's strong ability to correctly identify individuals at risk while minimizing false positives and false negatives.
- **Confusion Matrix:** Analysis of the confusion matrix revealed that the model most accurately identified individuals with high-risk mental health indicators, such as those reporting severe anxiety or depressive symptoms, while showing slightly lower performance in distinguishing borderline or mild cases where symptoms overlapped.

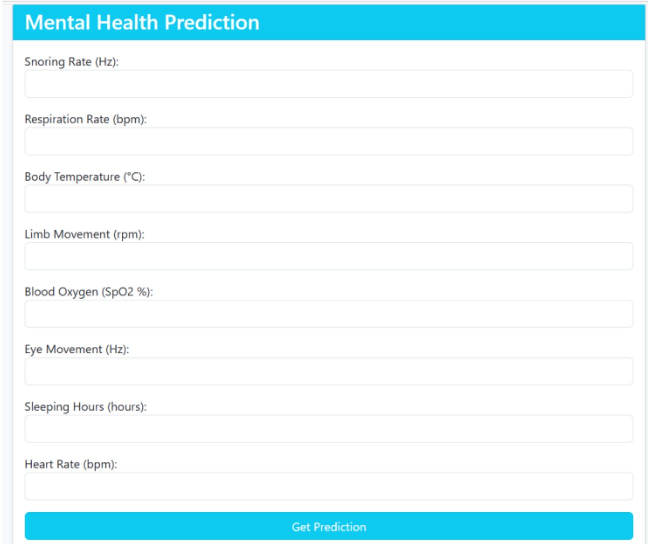

Fig 3 : User input form

The system's robustness was evaluated under various conditions:

- The model maintained high performance despite variations in data distribution, minor inconsistencies in survey responses, and demographic diversity among participants.
- Feature selection and engineering played a significant role in improving the model's generalization to unseen data, reducing overfitting and enhancing reliability.
- Data balancing techniques, including Synthetic Minority Over-sampling Technique (SMOTE), contributed significantly to handling class imbalances and improving predictive performance for underrepresented mental health conditions.

Overall, the results demonstrate that the developed system can serve as a powerful tool for early mental health risk detection, offering high accuracy and reliability suitable for practical deployment in organizational wellness programs, clinical pre-screening tools, or digital health applications.

V. DISCUSSION

The results obtained from the Mental Health Prediction System demonstrate significant potential for automated, data-driven assessment of mental health risk, offering a modern alternative to traditional manual evaluations. The application of machine learning algorithms, particularly the Random Forest classifier, proved highly effective in capturing complex, nonlinear relationships among diverse personal, behavioral, and

ISSN 0315-3681 Volume 122, 2025

demographic features associated with mental health outcomes. This approach offers a powerful yet practical way to identify individuals who may be at elevated risk of mental health challenges, enabling proactive interventions that can substantially improve individual well-being and reduce broader societal costs. A key strength of the developed system lies in its ability to handle subtle and overlapping indicators of mental health conditions, which are often difficult to differentiate using conventional screening tools alone. Unlike traditional methods relying solely on linear statistical relationships or single-questionnaire scores, the machine learning models can integrate multiple factors, such as stress levels, social connectivity, work-life balance, and lifestyle habits, to produce a more nuanced and holistic risk assessment. The confusion matrix analysis highlighted that while the system performed exceptionally well in identifying high-risk individuals, it exhibited slightly reduced precision when distinguishing borderline or mild cases where symptoms may be less pronounced or inconsistent. This suggests that further refinement of the models, as well as the incorporation of richer datasets, could enhance performance in these borderline scenarios. One significant observation is the notable benefit derived from advanced preprocessing and feature engineering techniques. Addressing issues like missing data, outlier values, and imbalanced class distributions substantially improved the models' robustness and generalizability. Additionally, the use of feature importance analysis and dimensionality reduction methods helped streamline the input data, reducing noise and focusing the models on the most predictive variables. Such measures are crucial for real-world deployment, where survey responses and behavioral data can vary significantly in quality and consistency.

Despite its strengths, the system has several limitations. The primary constraint lies in the reliance on self-reported survey data, which can be subject to bias, social desirability effects, and underreporting, particularly in stigmatized contexts like mental health. While machine learning can detect patterns, its predictions are inherently constrained by the quality and authenticity of the underlying data. Future improvements could involve integrating more objective data sources, such as physiological signals from wearable devices, voice analysis, or passive digital footprints, which may provide richer and less biased indicators of mental health status. Another challenge is model interpretability. Although Random Forests provide some insights through feature importance rankings, the decision-making process remains relatively opaque, particularly compared to simpler statistical models like logistic regression. In clinical and organizational settings, where decisions based on model outputs can have significant personal and ethical consequences, improving the transparency and explainability of predictions is essential. Techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) could be integrated into future versions of the system to enhance user trust and facilitate responsible AI deployment. From a usability perspective, the web-based interface developed using frameworks such as Django or Flask offers a user-friendly and accessible experience for diverse stakeholders, including healthcare providers, organizational wellness managers, and individuals seeking self-assessment. The real-time feedback—comprising risk predictions, probability scores, and key feature contributions enhances user engagement and provides actionable insights. However, future iterations could expand on this by offering tailored recommendations for follow-up actions, such as seeking professional consultation, accessing mental health resources, or engaging in specific wellness activities.

A promising avenue for future development is the personalization of mental health predictions based on regional, cultural, or occupational factors. Mental health risks can vary significantly depending on cultural attitudes, environmental stressors, and socio-economic conditions. Incorporating geo-location data, occupational context, and temporal analytics (e.g., trends over time) could enhance the relevance and precision of predictions, aligning the system with emerging trends in precision mental health care. Comparing this machine learning-based system with traditional methods, such as clinician-administered interviews and paper-based surveys, highlights important trade-offs. The automated system offers rapid, consistent, and scalable risk assessment without requiring significant human resources, delivering substantial time and cost savings. However, it remains reliant on data quality and should not replace professional clinical evaluation in ambiguous or severe cases. Integrating this system into a broader ecosystem—including clinician oversight, digital therapeutics, and human support services—could deliver a more holistic mental health care solution.

In summary, the developed Mental Health Prediction System successfully addresses several key challenges in mental health screening through an innovative, accessible technological approach. The combination of advanced machine learning models and an intuitive user interface demonstrates a balanced integration of technical rigor and user-centered design. While opportunities for further refinement exist—particularly in expanding data sources and improving interpretability—the system provides a strong foundation for future research and real-world applications aimed at enhancing early detection, proactive intervention, and overall mental health outcomes.

VI. CONCLUSION

The Mental Health Prediction System using machine learning effectively demonstrates the transformative potential of data-driven technologies in modern mental health assessment. By leveraging advanced algorithms such as Random Forests, Support Vector Machines, and Logistic Regression, the system accurately identifies individuals who may be at risk for mental health challenges based on diverse personal, behavioral, and demographic data. This approach offers a rapid, non-invasive, and scalable method for early detection, providing a practical solution that can complement traditional mental health evaluations. The integration of widely accessible technologies such as Python-based frameworks and open-source machine learning libraries like Scikit-learn and Pandas ensures that the solution is both practical and deployable across various contexts, including clinical settings, workplace wellness programs, and personal health applications. The system's strong predictive performance, supported by rigorous preprocessing, feature engineering, and model evaluation, underscores its reliability as a tool for proactive mental health monitoring. Crucially, the design addresses real-world challenges such as handling missing data, managing imbalanced classes, and ensuring robustness across diverse populations. While certain limitations remain, including reliance on self-reported data and the inherent "black-box" nature of some machine learning models, these challenges also present valuable avenues for future improvement. Potential enhancements include integrating multimodal data sources—such as physiological signals from wearable devices, voice analysis, and digital behavior patterns—as well as employing explainable AI techniques to increase transparency and user trust. Overall, this project provides a strong foundation for deploying AI-powered mental health prediction systems that can significantly contribute to early intervention, resource optimization, and improved mental health outcomes. With continued efforts in expanding datasets, refining model interpretability, and integrating real-time applications, the system holds great promise for supporting mental health care in diverse settings. As a scalable and cost-effective solution, it can play a vital role in modernizing mental health assessment, promoting well-being, and reducing the societal burden of untreated mental health conditions.

REFERENCES

[1] American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). American Psychiatric Publishing.

[2] Benton, A., Coppersmith, G., Dredze, M., & Ayers, J. W. (2017). Ethical Research Protocols for Social Media Health Research. *arXiv preprint arXiv:1712.06187*.

[3] Edwards, G., et al. (2019). Factors associated with depression and anxiety in a working population: A cross-sectional study. *Occupational Medicine, 69*(8), 612–617.

[4] Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting Depression and Mental Illness on Social Media: An Integrative Review. *Current Opinion in Behavioral Sciences, 18*, 43–49.

[5] Tsakalidis, A., et al. (2018). Nowcasting Well-being Using Self-Reported Data and Social Media. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*.

[6] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[7] Reece, A. G., et al. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science, 6*(1), 15.

[8] Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2018). Deep Learning for Depression Detection of Twitter Users. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*.

[9] Zhang, Y., et al. (2021). A deep learning framework for predicting depression from social media text. *IEEE Transactions on Affective Computing*.

[10] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*.

[11] Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine, 49*(9), 1426–1448.

[12] Coppersmith, G., Leary, R., Whyne, E., & Wood, T. (2015). Quantifying suicidal ideation via language usage on social media. *Joint Statistical Meetings Proceedings*.

[13] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *NPJ Digital Medicine, 3*, 43.

[14] Huckvale, K., Venkatesh, S., & Christensen, H. (2019). Toward Clinical Applications of Artificial Intelligence in Mental Health: Progress and Prospects. *Current Psychiatry Reports, 21*(7), 50.

[15] Gooding, P., Kariotis, T., & Tarzia, L. (2022). Machine Learning for Mental Health: A Systematic Critical Review of the Ethics and Social Risks. *Journal of Medical Internet Research, 24*(1), e28081.

[16] Liu, N. F., et al. (2021). An interpretable machine learning framework for personalized mental health prediction using social media data. *Nature Human Behaviour, 5*, 1491–1502.

[17] Resnik, P., et al. (2015). Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter. *Proceedings of CLPsych 2015*.

[18] Nguyen, T., et al. (2014). Modeling Psychological States in Online Social Media. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM)*.

[19] McCoy, T. H., et al. (2020). Digital phenotyping and its implications for mental health care. *JAMA Psychiatry, 77*(9), 889–890.

[20] Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. *ICWSM*.