A Machine Learning Approach to Mental Disorder Prediction: Handling the Missing Data Challenge

Tsholofelo Mokheleli[1](✉), Tebogo Bokaba[1], Tinofirei Museba[1], and Nompumelelo Ntshingila[2]

[1] Department of Applied Information Systems, University of Johannesburg, Johannesburg, South Africa
{tsholofelom,tbokaba,tmuseba}@uj.ac.za
[2] Department of Nursing Science, University of Johannesburg, Johannesburg, South Africa
mpumin@uj.ac.za

Abstract. In recent years, the application of Machine Learning (ML) to predict mental disorders has gained significant attention due to its potential for early prediction. This study highlights the challenges of ML in mental disorders prediction, such as missing data in mental health datasets, by comparing four data imputation methods: Mode, Multivariate Imputation by Chained Equations, Hot Deck, and K-Nearest Neighbor (K-NN) to enhance predictive accuracy; and utilizing four ML classifiers and three ensemble methods: Bagging, Boosting, and Stacking, with Mode and K-NN imputation datasets to show consistent performance. The study ultimately contributes to early mental disorder diagnosis and intervention in alignment with the United Nations Sustainable Development Goal 3 (SDG 3) for global health and well-being, by highlighting ML and data imputation's potential in mental health analysis and paving the way for further advancements in the field.

Keywords: Data Imputation · Machine Learning · Mental Disorders Prediction · Missing Values

1 Introduction

There is a growing interest in utilizing Machine Learning (ML) to predict mental disorders [1]. This surge in interest is primarily driven by the potential of ML to assist in the early identification of mental disorders [2]. ML is a subfield of artificial intelligence that involves the development of algorithms and models that can analyze and learn from data, enabling computers to make predictions or decisions without being explicitly programmed [3]. ML algorithms have showcased impressive abilities in analyzing extensive datasets encompassing medical records, physiological measurements, and patient-reported symptoms. This capability allows for predicting the likelihood of developing mental disorders and assessing the effectiveness of different treatment options [4]. Mental disorders are conditions that disrupt the thoughts, emotions, behaviors, and

M. Masinde et al. (Eds.): AFRICATEK 2023, LNICST 520, pp. 93–106, 2024.
https://doi.org/10.1007/978-3-031-63999-9_6

mental well-being of an individual, leading to distress and daily life challenges [5]. Accurately predicting mental disorders holds substantial significance, as it can lead to timely interventions, personalized treatment strategies, and an overall enhancement in the quality of life for individuals affected by these conditions [6, 7]. With the advancement of technology and data-driven approaches, the potential for harnessing predictive models to improve mental health care continues to grow progressively promising.

However, while using ML for predicting mental disorders is undoubtedly a promising avenue, it also presents challenges, particularly in the context of the unique characteristics of mental disorder datasets. Among these challenges, missing values emerge as a critical concern. Missing values introduce data inconsistencies and incompleteness, which can delay the accuracy and reliability of predictive models in this domain [8].

Mental health datasets often have missing data due to incomplete patient responses, data collection errors, data capturing, or privacy concerns. Ignoring missing data may compromise the accuracy and reliability of analyses, potentially resulting in biased outcomes. Handling missing data is thus crucial for accurate insights from ML models. Data imputation methods can be used to estimate missing data. This study used ML algorithms to address missing data, including a comparative evaluation of four data imputation methods: Mode, Multivariate Imputation by Chained Equations (MICE), Hot Deck, and K-Nearest Neighbor (K-NN).

The study aimed to shed light on the significance of mental disorder prediction and dealing with missing data in the dataset. The study's main contributions lie in comparing data imputation methods to reduce the impact of missing data on the accuracy of a mental disorder prediction model, and to enhance mental health care through data-driven approaches such as applying ML algorithms and ensemble learning methods. Therefore, the study aligns with the United Nations Sustainable Development Goal 3 (SDG 3), which strives to ensure good health and well-being for all [9].

2 Related Work

This section aims to comprehensively examine the existing literature to gain insight into the challenges posed by missing data when employing ML for predicting mental disorders. Several studies have contributed valuable insights into utilizing ML for mental disorder prediction and addressing the complexities introduced by missing data.

Henry et al. [10] focused on predicting mental health treatment for information technology (tech) employees. Their study aimed to predict the necessity of mental health treatment among these employees using the Open Sourcing Mental Illness (OSMI) dataset. To handle the missing data in the dataset, the authors removed instances that had missing value rates exceeding 85%. The study used ensemble methods such as Bagging, Light Gradient-Boosting Machine (GBM), and Stacking to develop the predictive models. It also used Binary Particle Swarm Optimization (BPSO) for feature selection. The study found that ensemble methods do not consistently yield superior predictions.

Li [11] focused on applying ML to predict mental disorders and interpret feature importance. The study used Random Forest (RF), K-NN, and Decision Tree (DT) algorithms. The authors addressed the issue of missing data by deleting observations that exceeded half of the dataset. The RF model and Grid Search Optimization (GSO) yielded the best predictive performance.

Mitravinda et al. [12] explored mental health patterns and risk factors within the tech industry, employing the OSMI dataset. The strategy utilized XGBoost and Gradient Boosting classifiers and yielded notable predictive accuracies. The strategy entailed the removal of columns featuring over 50% missing data.

Using ML algorithms, Bajaj et al. [13] investigated non-invasive mental health prediction. The study aimed to identify the most effective predictive model among various algorithms, including Logistic Regression (LR), DT, K-NN, Adaboost, and RF. The authors used two datasets, one comprising mental health patient questionnaires and the other containing information from MRI scans of Alzheimer's patients. The authors deleted the missing data instead of data imputation. The study underscored the importance of using data that includes demographic, behavioral, and psychological factors that have the potential to further enhance prediction accuracy.

Olatunde et al. [14] formulated a classification model for mental disorders employing a variety of algorithms, such as LR, Support Vector Machine (SVM), Naïve Bayes (NB), and DT. Despite facing substantial missing data in the dataset, they managed data cleaning by discarding columns with 1 000 or more missing data and rows containing at least 11 missing data. Imputation methods were applied to categorical features, filling missing data with the most frequent value, or replacing them with a new "missing."

Duncan et al. [15] investigated the connection between mental health and academic performance in Canadian secondary school students. The missing data for all variables were addressed by employing multiple imputations using R packages and the MICE technique. This approach preserved the hierarchical structure of the data by imputing values at the individual level while considering schools as a random intercept higher-order clustering factor.

Luo [16] investigated challenges linked to missing data within clinical datasets. This study revolved around the Data Analytics Challenge of Missing Data Imputation (DACMI), which offers a shared clinical dataset and ground truth for developing and evaluating missing data imputation techniques for clinical data. The study focused on imputing missing data in 13 commonly measured blood laboratory tests. It randomly removed one recorded result per laboratory test per patient admission and used these as the ground truth for evaluation. This rigorous evaluation methodology ensured the accuracy and reliability of the imputation techniques. The ML algorithms used in the study were LightGBM, XGBoost, and MICE imputation techniques.

The significant gaps identified in the literature reviewed are related to the insufficient exploration of data imputation techniques for handling missing data within mental health datasets. Many studies have leaned towards removing rows or columns with missing data [10–14], a practice that can lead to data loss and introduce potential bias into the models. Despite the numerous investigations into predictive modeling for mental health results, a recurring pattern emerges where the resolution for missing data tends to involve removal or deletion rather than embracing more comprehensive data imputation strategies. Considering the identified gaps, this study used and compared various data imputation methods to assess their impact on enhancing predictive accuracy and overall performance in models for predicting mental disorders. While some studies primarily utilized single-year data from the OSMI dataset, this study took a more comprehensive approach by combining six different OSMI datasets to create a more extensive dataset.

3 Methodology

This section presents the dataset overview, data preprocessing and experimental setup. It also discusses the data imputation techniques applied, ML algorithms and ensemble methods, and evaluation metrics.

3.1 Data Collection

The dataset utilized in this study was sourced from Open Sourcing Mental Illness (OSMI), a nonprofit organization focused on promoting mental wellness within the workplace and open-source communities [17]. The OSMI website administers a survey that captures the life experiences of individuals employed in technology-oriented companies, encompassing their mental health history and consultation practices. The OSMI dataset enjoys widespread recognition within the mental health domain. This study took a unique approach by combining datasets from 2016 to 2021 to enrich the scope of data analysis. Unlike previous studies that worked with constrained data subsets [11, 18], this investigation harnessed the broader dataset to train and test ML algorithms models with improved effectiveness.

Table 1 presents a summary of the dataset, along with the features and instances in each dataset used to make up the final dataset used in the study. Table 2 shows a statistical analysis of the missing data within the dataset and the percentages of the missing data.

Table 1. Summary of the Dataset.

No.	Year	Features	Instances
1	2016	63	1433
2	2017	123	756
3	2018	123	417
4	2019	82	352
5	2020	120	180
6	2021	124	131

3.2 Data Preprocessing

The data preprocessing stage is a preparatory phase in data analysis where raw data is cleaned, transformed, and organized to ensure its quality and suitability for further analysis. During this stage, data cleaning and preparation tasks are carried out.

The process began with removing irrelevant columns from the dataset spanning 2016 to 2021. This entailed the selection and retention of important columns which were renamed for simplicity and the assurance of data quality. Textual data was transformed into numerical form using label encoding. The resulting dataset consisted of 3 269 rows and 25 columns, making it ready for ML analysis. The selection criteria for retaining columns were based on their consistent presence across all yearly datasets from 2016 to 2021.

Table 2. Overview of Data Completeness and Missing Data, Including Percentages.

Variables	Missing	Non-null	Missing in %	Non-null in %
Self Employed	0	3 269	0.0	100
Company Role	1 429	1 840	43.7	56.3
Mental Importance	546	2 723	16.7	83.3
Discuss MH	1 410	1 859	43.1	56.9
Coworkers	428	2 841	13.1	86.9
Work Interfere	2 723	546	83.3	16.7
Past Mental Health	23	3 246	0.7	99.3
Gender	28	3 241	0.9	99.1
Mental Health Diagnosed	1 080	2189	33	67

3.2.1 Data Imputation

The study employed four distinct data imputation techniques: Mode, MICE, Hot Deck, and K-NN. These techniques were chosen for their ability to effectively handle missing data by offering various advantages and trade-offs, thereby ensuring a comprehensive exploration of imputation methods.

Mode is a straightforward method to address missing data within a dataset. It involves substituting absent values with the mode, the most frequently occurring value in the corresponding variable [19, 20]. This technique suits categorical or nominal data with distinct categories or labels.

MICE stands out as a more advanced approach for handling missing data, as it employs an iterative process to impute missing data across a dataset by generating multiple complete datasets, each incorporating imputed values [21, 22]. This iterative procedure is particularly beneficial because it focuses on individual variables during each cycle while incorporating regression models with other variables as predictors, allowing for a more comprehensive and precise imputation process.

Hot Deck is a valuable method for handling missing data and operates by replacing missing data in datasets with values from similar or “neighboring” observations. This approach leverages the concept that when two observations share similarities in specific aspects, their missing data can be effectively attributed using the available data from the other observation, contributing to enhanced data completeness and accuracy [19, 23].

K-NN, an effective data imputation technique, capitalizes on the principle of data point similarity to address missing data. It is especially advantageous when dealing with numerical and continuous variables. K-NN imputation operates by identifying the ‘k’ nearest complete data points to the one with missing data and subsequently estimating the missing value by drawing insights from the values of these neighboring data points, promoting precise and context-aware imputation [24].

3.3 Experimental Setup

The experiments were performed on Google Research Colab, also referred to as Colab, which is a cloud-based platform that provides a user-friendly interface for executing Jupyter notebooks [25]. Colab offers an extensive range of pre-installed libraries and frameworks frequently utilized in tasks involving data analysis, ML, and Deep Learning. This platform was chosen for the study due to its suitability and convenience.

The experiments began with data preprocessing, including dataset concatenation and label encoding to convert text to numbers (see Fig. 1). Inconsistencies and ambiguous categories were resolved by observation; for instance, ages below 16 were removed and treated as missing data. The four data imputation techniques outlined above were applied to create clean datasets for baseline model construction. Baseline models employed four classifiers and five evaluation metrics, detailed in the subsequent section. For the construction and testing of the baseline models, and the ensemble models, the data was split into an 80% training set and a 20% testing set. This split was crucial to provide ample data for the training phase while ensuring a robust evaluation of the models on unseen data during the testing phase. K-fold cross-validation was employed to assess how well the ML model will generalize and perform on different folds of data.

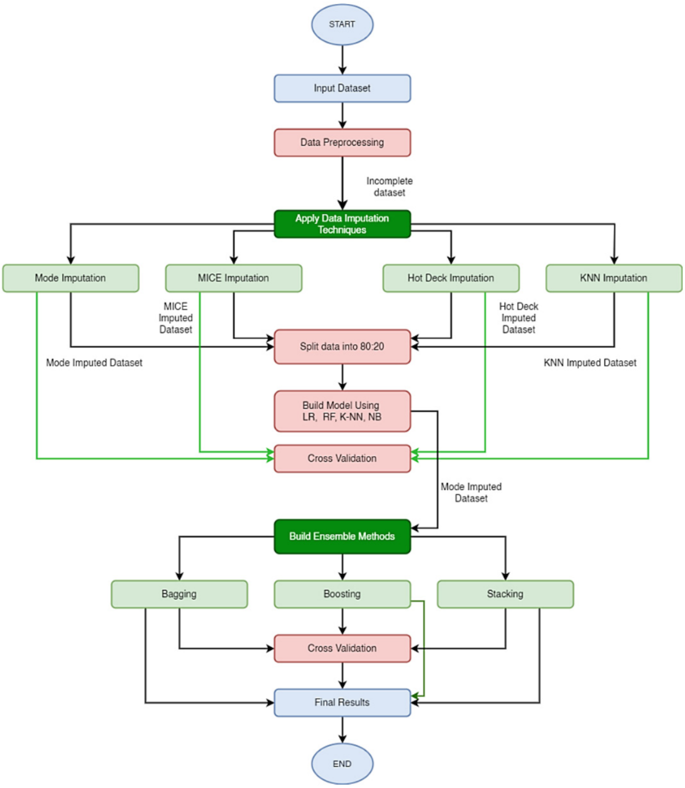

Fig. 1. Experimental Process Flow

3.4 ML Algorithms and Ensemble Methods

This section discusses the ML algorithms employed in the study, encompassing various approaches. The initial baseline model featured four classifiers: LR, RF, K-NN, and NB.

LR is a statistical method used in binary classification and regression tasks. It models the probability of an instance belonging to a particular class based on its independent variables [26]. LR is widely recognized for its user-friendly nature.

RF is a supervised learning method that constructs a "forest" using multiple decision trees trained through the "Bagging" technique. Bagging is based on the concept that aggregating multiple learning models enhances the overall predictive capability [27].

K-NN is a flexible ML algorithm for classification and regression. It predicts by considering the k closest data points and uses majority class (classification) or average value (regression) among them [26].

NB is a powerful, simple classification algorithm. It uses Bayes' theorem to calculate class probabilities based on observed features, using feature mean and standard deviation estimates for each class [28].

The study employed three prominent ensemble methods to enhance the predictive capabilities of the models: Bagging, Boosting, and Stacking.

Bagging and Boosting were selected due to their proven effectiveness in the literature for improving model accuracy. Stacking was introduced as an innovative approach not previously explored in the reviewed literature to enhance model robustness by combining multiple learners for superior predictions. Bagging, short for Bootstrap Aggregating, involves training multiple instances of a base model on different subsets of the training data and combining their predictions to reduce variance and improve overall model stability [29].

Boosting focuses on iterative training models, assigning more weight to instances misclassified in previous iterations. This adaptive learning process helps create a strong, accurate model by giving higher importance to challenging data points [30].

Stacking combines the strengths of various base models by training a meta-model that learns from their collective predictions [31]. These ensemble techniques provide opportunities for improving model performance and offer valuable insights into how different algorithms complement each other to tackle complex predictive tasks effectively.

3.5 Evaluation Metrics

A diverse set of evaluation metrics is essential to assess the performance of ML models in predicting mental disorder outcomes. This study employed a comprehensive suite of metrics, including accuracy, precision, recall, f1 score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Multiple metrics are imperative, particularly when addressing mental disorders, as they provide a more holistic assessment beyond accuracy. The formulas for all the metrics are shown in Eq. 1–5. The formulas involve these keywords: True Positive (TP) for correctly identifying a positive outcome, False Positive (FP) for incorrectly identifying a positive outcome that is not there, True Negative (TN) for correctly identifying a negative outcome, and False Negative (FN) for incorrectly identifying a negative outcome that exists [32, 33].

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ instances}{Total\ number\ of\ instances} \tag{1}$$

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)} \tag{2}$$

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)} \tag{3}$$

$$F1score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{4}$$

$$AUC - ROC = sum(\frac{(TPR[i] + TPR[i+1]) * (FPR[i+1] - FPR[i])}{2}) \tag{5}$$

where true positive rate (TPR) and false positive rate (FPR) are calculated:

$$TPR = \frac{TP}{(TP + FN)} \tag{6}$$

$$FPR = \frac{FP}{(FP + TN)} \tag{7}$$

4 Experimental Results and Discussion

This section showcases the experimental results obtained through the utilization of data imputation techniques and ML algorithms previously detailed in the methodology section. The 80:20 data split discussed in Sect. 3.3 was used during the generation of the results presented in Sects. 4.1 and 4.2.

4.1 Initial Models

Four models were constructed using four classifiers from the imputed datasets. The results of the top-performing models are presented in Tables 3 and 4. The best data imputation methods were determined based on cumulative metrics scores and cross-validation results, including mean cross-validation accuracy and standard deviation in cross-validation accuracy.

Mode imputation exhibited mean cross-validation accuracy ranging from 74% to 92%, with LR achieving 89%, RF 92%, K-NN 74%, and NB 88%, all having a standard deviation in cross-validation accuracy of no more than 0.02. Precision ranged from 74% to 93%, while Recall ranged from 90% to 96%. F1 scores consistently exceeded 82%, as shown in Table 3.

K-NN imputation showed results close to Mode, with LR at 89%, RF at 92%, K-NN at 73%, and NB at 89%, maintaining a standard deviation in cross-validation accuracy of no more than 0.02. K-NN imputation yielded favorable Precision results from 72% to 93%. Recall also demonstrated a strong performance, from 90% to 97%. The F1 Score also showed robust results, with the lowest score being 82% (Table 4).

Mode and K-NN emerged as the top data imputation methods, demonstrating consistent performance across metrics and cross-validation.

Table 3. Model Performance with Mode-Imputed Data.

	Classifiers			
Metrics	LR	RF	K-NN	NB
Accuracy	0.89	0.92	0.73	0.89
Precision	0.89	0.92	0.74	0.93
Recall	0.96	0.96	0.93	0.90
F1 Score	0.92	0.94	0.82	0.91
AUC-ROC	0.86	0.90	0.64	0.88
Cross-validation Acc (%)	89	92	74	88

Table 4. Model Performance with K-NN-Imputed Data.

	Classifiers			
Metrics	LR	RF	K-NN	NB
Accuracy	0.89	0.92	0.72	0.89
Precision	0.88	0.92	0.72	0.93
Recall	0.96	0.97	0.94	0.90
F1 Score	0.92	0.94	0.82	0.92
AUC-ROC	0.85	0.90	0.62	0.88
Cross-validation Acc (%)	89	92	73	89

Overall, among the models, the best classifiers can also be identified. RF and NB consistently demonstrated strong performance across various imputation methods and metrics. While the models performed well, there is room for improvement by incorporating ensemble methods to enhance their overall performance.

4.2 Ensemble Methods

Three ensemble methods in ML were discussed in the methodology section. The ensemble methods utilized the Mode imputation dataset as it exhibited encouraging results considering the cumulative score of the metrics. In the Bagging analysis, LR and RF were employed as base models. RF had a high mean cross-validation accuracy (0.92) and low standard deviation (0.01), making it a strong candidate for Bagging. LR also had a relatively high mean accuracy (0.89) and a low standard deviation (0.01).

Boosting aims to improve model performance by focusing on misclassified instances. RF and LR were again chosen as base models. Of the possible combinations of classifiers that were tested, RF and LR were the ones that performed the best.

Stacking focuses on the diversity of base classifiers that complement each other. All four classifiers offered diversity, and LR was used as the meta-learner to combine the predictions of the diverse base models.

Table 5 shows the results of the ensemble methods. The methods achieved similar accuracy, precision, and F1 score results. However, Boosting with a recall of 0.97 slightly outperformed Bagging and Stacking with recalls of 0.96. Stacking also slightly outperformed Boosting and Bagging regarding AUC ROC, with 0.90 compared to 0.89 for the latter. Stacking may be considered the best ensemble method as it embraces diversity, and as AUC-ROC is a critical metric for assessing the overall discriminative power of a model, is especially effective in binary classification tasks. The cross-validation results demonstrate that Stacking consistently performed well across all folds of the dataset, achieving an impressive mean cross-validation accuracy of 92%. In comparison, Bagging achieved a slightly lower accuracy of 90%, while Boosting attained 85%. All three ensemble methods exhibited a low standard deviation of 0.01 in cross-validation accuracy, indicating their stability and reliability.

Table 5. Ensemble Method Results Using Mode-Imputed Data

	Ensemble Method		
Metrics	Bagging	Boosting	Stacking
Accuracy	0.92	0.92	0.92
Precision	0.92	0.92	0.92
Recall	0.96	0.97	0.96
F1 Score	0.94	0.94	0.94
AUC-ROC	0.89	0.89	0.90
Cross-validation Acc (%)	90	85	92

4.3 Comparative Analysis

In comparing the performance of initial models and ensemble methods for handling missing data in mental disorder datasets, the study observed that Mode and K-NN imputation methods consistently demonstrated strong performance, with RF and NB as the standout classifiers across various imputation techniques and metrics. The initial models showed that Mode imputation consistently outperformed K-NN imputation in terms of mean cross-validation accuracy, precision for LR, RF and NB classifiers, and had comparable Recall scores. Mode imputation can be considered the best-performing method among the two for handling missing data in mental disorder datasets.

However, while these initial models showed promise, the introduction of ensemble methods further enhanced their overall performance. Among the ensemble methods, Stacking exhibited the most balanced performance, achieving comparable accuracy, precision, and F1 score results to Bagging and Boosting. Stacking excelled in terms of AUC-ROC, a critical metric for assessing discriminative power in binary classification. Stacking's consistently high cross-validation accuracy of 92% and low standard deviation indicated its stability and reliability, suggesting its potential as an effective approach to handling missing data in mental disorder datasets.

5 Contributions of the Study

This study has significantly addressed the escalating worldwide challenges associated with mental disorders. It aimed to bridge a void in the current body of literature by employing ML algorithms and data imputation techniques to impute missing data and predict and analyze the occurrence of mental disorders. The study has not only made significant advancements in theoretical contributions but has also generated implications, both theoretical and practical, that are relevant to the field.

Theoretical Contributions: The study contributes to understanding how data imputation techniques can effectively be used in mental disorders prediction using ML. It also adds to the growing research examining how ML can be applied to mental disorder prediction.

Theoretical Implications: The study's results could impact how mental healthcare data are analyzed, highlighting the significance of data imputation techniques. This paper could inspire further research to improve ML algorithms specifically designed for mental disorder prediction purposes.

Practical Implications: The ML data imputation compared and discussed has the potential to enhance the accuracy of predictive models for mental disorders, which can be valuable for early diagnosis and treatment. The findings could also impact the development of systems that promote improved decision-making based on data.

6 Limitation and Future Work

A limitation of the study was the presence of class imbalance within the dataset. The unequal distribution of data across different classes can pose challenges for ML models, potentially leading to biased predictions. As a result of class imbalance, the predictive accuracy of the models may not have fully realized their potential. Incorporating hyperparameter tuning can be a critical consideration for future research in this domain.

Future work should prioritize addressing class imbalance as a foundational step in model improvement. Techniques like oversampling, undersampling, or advanced resampling methods can help rebalance the dataset. Implementing hyperparameter tuning and feature selection processes can further boost model performance, especially in mental disorder prediction, where accuracy is crucial. Future work can significantly enhance predictive models for mental health outcomes by systematically addressing these limitations and refining modeling approaches.

7 Conclusion

By addressing the critical issues of missing data in mental health datasets and employing various data imputations and ensemble methods, this study has made significant strides in aligning with the United Nations SDG 3, namely, to improve global health and well-being. It has laid a strong foundation for developing predictive models for mental disorders. The study contributes to early diagnosis and intervention and opens the door to more advanced algorithms and evaluation metrics that can substantially enhance the accuracy and reliability of these predictive models.

This study marks an important milestone in leveraging the power of ML within the realm of mental health. It offers promising avenues for the early detection and management of mental disorders, aligning with the broader goals of SDG 3. Furthermore, it underscores the potential of ML and data imputation techniques in mental health analysis while highlighting future research and improvement areas. As cutting-edge technologies are embraced, and approaches are refined, the potential for making even greater strides in the crucial mental health analysis and care field is evident.

References

1. Garriga, R., et al.: Machine learning model to predict mental health crises from electronic health records. Nat. Med. **28**(6), 1240–1248 (2022). https://doi.org/10.1038/s41591-022-01811-5
2. Awal, G.K., Rao, K.: Can machine learning predict an employee's mental health? Commun. Comput. Inform. Sci. **1417**, 235–247 (2021). https://doi.org/10.1007/978-3-030-88378-2_19
3. Ercan, U.K., Özdemir, G.D., Özdemir, M.A., Güren, O.: Plasma medicine: the era of artificial intelligence. Plasma Processes Polym. (2023). https://doi.org/10.1002/ppap.202300066
4. Vahdat, M., Oneto, L., Anguita, D., Funk, M., Rauterberg, M.: Can machine learning explain human learning? Neurocomputing **192**, 14–28 (2016). https://doi.org/10.1016/j.neucom.2015.11.100
5. Galderisi, S., Heinz, A., Kastrup, M., Beezhold, J., Sartorius, N.: Toward a new definition of mental health. World Psych. **14**(2), 231–233 (2015). https://doi.org/10.1002/wps.20231
6. Allahyari, E., Roustaei, N.: Applying artificial neural-network model to predict psychiatric symptoms. Biomedicine (Taipei) **12**(1) (2022). https://doi.org/10.37796/2211-8039.1149
7. Yi, Y., Park, Y.-H.: Structural equation model of the relationship between functional ability, mental health, and quality of life in older adults living alone. PLoS ONE **17**(8), e0269003 (2022). https://doi.org/10.1371/journal.pone.0269003
8. Ridzuan, F., Wan Zainon, W.M.N.: A review on data cleansing methods for big data. Procedia Comput. Sci. **161**, 731–738 (2019). https://doi.org/10.1016/j.procs.2019.11.177

9. Salvo, D., et al.: Physical activity promotion and the United Nations Sustainable Development Goals: building synergies to maximize impact. J. Phys. Act. Health **18**(10), 1163–1180 (2021). https://doi.org/10.1123/jpah.2021-0413
10. Henry, M., Isa, S.M.: Mental health treatment prediction for tech employee with the implementation of ensemble methods. J. Theor. Appl. Inf. Technol. **100**(8), 2675–2685 (2022)
11. Li, Y.: Application of machine learning to predict mental health disorders and interpret feature importance. In: 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS), pp. 257–261. IEEE (2023). https://doi.org/10.1109/ISCTIS58954.2023.10213032
12. Mitravinda, K.M., Nair, D.S., Srinivasa, G.: Mental health in tech: Analysis of workplace risk factors and impact of COVID-19. SN Comput. Sci. **4**(2) (2023). https://doi.org/10.1007/s42979-022-01613-z
13. Bajaj, V., Bathija, R., Megnani, C., Sawara, J., Ansari, N.: Non-invasive mental health prediction using machine learning: an exploration of algorithms and accuracy. In: 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 313–321. IEEE (2023). https://doi.org/10.1109/ICICCS56967.2023.10142504
14. Olatunde, O., Falola, B.: Classification of mental health disorders. Tech. Rep. (2021). https://doi.org/10.13140/RG.2.2.34918.60483
15. Duncan, M.J., Patte, K.A., Leatherdale, S.T.: Mental health associations with academic performance and education behaviors in Canadian secondary school students. Can. J. Sch. Psychol. **36**(4), 335–357 (2021). https://doi.org/10.1177/0829573521997311
16. Luo, Y.: Evaluating the state of the art in missing data imputation for clinical data. Brief Bioinform. **23**(1) (2022). https://doi.org/10.1093/bib/bbab489
17. OSMI. About OSMI. Open Sourcing Mental Health (2023). https://osmihelp.org/about/about-osmi.html. Accessed 8 Sep 2023
18. Tate, A.E., McCabe, R.C., Larsson, H., Lundström, S., Lichtenstein, P., Kuja-Halkola, R.: Predicting mental health problems in adolescence using machine learning techniques. PLoS One **15**(4) (2020). https://doi.org/10.1371/journal.pone.0230389
19. Xu, X., Xia, L., Zhang, Q., Wu, S., Wu, M., Liu, H.: The ability of different imputation methods for missing values in mental measurement questionnaires. BMC Med. Res. Methodol. **20**(1), 42 (2020). https://doi.org/10.1186/s12874-020-00932-0
20. Makaba, T., Dogo, E.: A comparison of strategies for missing values in data on machine learning classification algorithms. In: 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), pp. 1–7. IEEE (2019). https://doi.org/10.1109/IMITEC45504.2019.9015889
21. Alruhaymi, A.Z., Kim, C.J.: Why can multiple imputations and how (MICE) algorithm work? Open J. Stat. **11**(05), 759–777 (2021). https://doi.org/10.4236/ojs.2021.115045
22. Ganapathy, S., Bhaskarapillai, B., Dandge, S.: The effect of multiple imputations by chained equations on the factors associated with immunization coverage in India. Int. J. Health Sci. Res. **11**(6), 249–262 (2021). https://doi.org/10.52403/ijhsr.20210638
23. Myers, T.A.: Goodbye, listwise deletion: presenting Hot Deck imputation as an easy and effective tool for handling missing data. Commun. Methods Meas. **5**(4), 297–310 (2011). https://doi.org/10.1080/19312458.2011.624490
24. Liao, S.G., et al.: Missing value imputation in high-dimensional phenomic data: imputable or not, and how?. BMC Bioinform. **15**(1) (2014). https://doi.org/10.1186/s12859-014-0346-6
25. Google Colab Team. Google Colaboratory. Google (2023). https://colab.google/. Accessed 16 June 2023
26. Reddy, U.S., Thota, A.V., Dharun, A.: Machine learning techniques for stress prediction in working employees. In: 2018 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2018, Institute of Electrical and Electronics Engineers Inc. (2018). https://doi.org/10.1109/ICCIC.2018.8782395

27. Mohamed, E.S., Naqishbandi, T.A., Bukhari, S.A.C., Rauf, I., Sawrikar, V., Hussain, A.: A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms. Healthcare Anal. **3**, 100185 (2023). https://doi.org/10.1016/j.health.2023.100185
28. Ali, L., et al.: A feature-driven decision support system for heart failure prediction based on $\chi 2$ Statistical Model and Gaussian Naive Bayes. Comput. Math. Methods Med. **2019**, 1–8 (2019). https://doi.org/10.1155/2019/6314328
29. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996). https://doi.org/10.1007/BF00058655
30. Wang, G., Sun, J., Ma, J., Xu, K., Gu, J.: Sentiment classification: the contribution of ensemble learning. Decis. Support. Syst. **57**, 77–93 (2014). https://doi.org/10.1016/j.dss.2013.08.002
31. Aboneh, T., Rorissa, A., Srinivasagan, R.: Stacking-based ensemble learning method for multi-spectral image classification. Technol. (Basel) **10**(1), 17 (2022). https://doi.org/10.3390/technologies10010017
32. Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., Mononen, J.: Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. Behav. Proc. **148**, 56–62 (2018). https://doi.org/10.1016/j.beproc.2018.01.004
33. Chicco, D., Tötsch, N., Jurman, G.: The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min **14**(1), 13 (2021). https://doi.org/10.1186/s13040-021-00244-z