engineering proceedings

MDPI

Proceeding Paper

Mental Health Classification Using Machine Learning with PCA and Logistics Regression Approaches for Decision Making †

Hendra Hendra [1,2,*], Mustafa Mat Deris [2] and Ika Safitri Windiarti [2]

[1] Informatics Engineering, Faculty of Engineering, Universitas Muhammadiyah Jakarta (UMJ), Jl. Cempaka Putih Tengah 27, Jakarta Pusat 10510, DKI Jakarta, Indonesia
[2] Faculty of Business Management and Information Technology, Universiti Muhammadiyah Malaysia (UMAM), Bangunan Wisma MAIPs, Blok S1, Kompleks Desa Siswa, Uniciti Alam, Padang Besar 02100, Perlis, Malaysia; mustafa@umam.edu.my (M.M.D.); ika.windiarti@umam.edu.my (I.S.W.)
* Correspondence: hendra@umj.ac.id or p5240017student.umam.edu.my
† Presented at the 8th Mechanical Engineering, Science and Technology International Conference, Padang Besar, Perlis, Malaysia, 11–12 December 2024.

Abstract: Mental health statistics come with numerous challenges, beginning with data integrity. Ensuring data accuracy and reliability is essential, especially if these datasets are to be used for advanced analysis or research. Additionally, privacy concerns heavily impact the management of mental health data. Protecting the privacy and confidentiality of individuals—especially those with personal or sensitive information—is paramount in system development. Robust protocols should be implemented to prevent unauthorized access and potential breaches. Another critical issue is bias in the training data, which can arise from the underrepresentation of certain demographic groups or the overrepresentation of others. Reducing bias within these datasets is essential to enhance the fairness and accuracy of the models and algorithms they support. Research on mental health classification using machine learning techniques, particularly PCA and logistic regression, is significant because it has the potential to improve decision-making in mental health care.

Keywords: mental health; machine learning; PCA; logistic regression

Academic Editors: Noor Hanita Abdul Majid, Agus Dwi Anggono, Waluyo Adi Siswanto, Tri Widodo Besar Riyadi, Mohammad Sukri Mustapa, Nur Rahmawati Syamsiyah and Afif Faishal

Published: 10 February 2025

Citation: Hendra, H.; Deris, M.M.; Windiarti, I.S. Mental Health Classification Using Machine Learning with PCA and Logistics Regression Approaches for Decision Making. *Eng. Proc.* **2025**, *84*, 47. https://doi.org/10.3390/engproc2025084047

1. Introduction

Mental health disorders represent a significant global issue, emphasizing the need for reliable tools to assist healthcare providers in diagnosis and decision-making. Traditional mental health classification methods often depend on subjective evaluations, which can result in inconsistent diagnoses [1]. Machine learning (ML) offers a promising alternative by employing data-driven techniques to analyze mental health data and enhance classification accuracy. Principal Component Analysis (PCA) and Logistic Regression are two key ML approaches that have proven effective in dimensionality reduction and improving predictive performance [2].

The importance of mental health is profound, as it significantly impacts overall health and quality of life. Poor mental health can lead to various challenges, including decreased productivity, strained relationships, and increased healthcare costs [3]. Factor analysis of neighborhood characteristics reveals the complex interplay of factors influencing mental health, demonstrating that overly simplistic or irrelevant features may obscure meaningful associations [4].

The role of PCA in mental health classification is particularly valuable due to its capacity to streamline complex datasets. PCA achieves this by converting high-dimensional

data into fewer, uncorrelated components, thereby minimizing noise and emphasizing critical features for classification [5]. Studies suggest that combining PCA with supervised learning methods like Logistic Regression can significantly enhance both the interpretability and accuracy of mental health classification models [6]. This underscores the importance of incorporating dimensionality reduction techniques in mental health analytics.

Logistic Regression, known for its robustness and interpretability, is especially effective for binary classification tasks, such as differentiating between individuals with and without mental health conditions. Its simplicity and ability to manage datasets with multicollinearity make it a preferred choice [7]. When used alongside PCA, Logistic Regression not only mitigates over fitting but also accelerates the decision-making process, enabling mental health practitioners to make informed decisions and optimize resource allocation [8].

Integrating Principal Component Analysis (PCA) with Logistic Regression in mental health datasets can enhance predictive modeling by addressing dimensionality reduction and improving model accuracy. PCA effectively condenses complex datasets, such as those assessing mental health impacts during COVID-19, by identifying key variables that contribute to mental health decline, such as healthcare roles and sleep patterns [9]. Furthermore, the combination of PCA and Logistic Regression has shown promise in other health domains, such as cardiovascular disease prediction, indicating its versatility and effectiveness in handling diverse health-related datasets [10]. This methodological synergy not only enhances the reliability of predictions but also aids in identifying at-risk individuals, thereby facilitating timely interventions [11].

Machine learning has emerged as a transformative tool in the field of mental health, enabling more accurate diagnosis and personalized treatment plans. By leveraging large datasets, machine learning algorithms can identify patterns and correlations that may not be immediately apparent to clinicians. For instance, studies have shown that machine learning techniques, such as support vector machines and neural networks, can effectively classify mental health disorders based on various input features, including demographic data and clinical assessments [12].

PCA serves as a powerful dimensionality reduction tool that simplifies the data without losing critical information, facilitating more efficient processing and interpretation. This approach can lead to improved diagnostic accuracy, enabling healthcare professionals to make informed decisions based on robust data analysis.

2. Related Work

2.1. Fundamental Concepts of Mental Health

Mental health is defined as a state of emotional, psychological, and social well-being that affects how individuals think, feel, and act. It encompasses various aspects, including the ability to manage stress, relate to others, and make decisions. The World Health Organization (WHO) emphasizes that mental health is not merely the absence of mental disorders but a state in which individuals can realize their potential, cope with the normal stresses of life, work productively, and contribute to their communities [13]. As mental health issues continue to rise globally, understanding and addressing these concerns have become increasingly critical, necessitating effective frameworks for assessment and intervention.

The importance of mental health is profound, as it significantly impacts overall health and quality of life. Poor mental health can lead to various challenges, including decreased productivity, strained relationships, and increased healthcare costs [14]. Furthermore, mental health conditions often coexist with other health issues, complicating treatment and recovery processes. In this context, the application of machine learning techniques, such as PCA and logistic regression, has emerged as a promising approach for classifying and predicting mental health outcomes. These methodologies enable researchers and practitioners

to analyze complex datasets, identify patterns, and make informed decisions, ultimately contributing to more effective mental health interventions and policy development.

2.2. Machine Learning in Mental Health

Machine learning has emerged as a transformative tool in the field of mental health, enabling more accurate diagnosis and personalized treatment plans. By leveraging large datasets, machine learning algorithms can identify patterns and correlations that may not be immediately apparent to clinicians. For instance, studies have shown that machine learning techniques, such as support vector machines and neural networks, can effectively classify mental health disorders based on various input features, including demographic data and clinical assessments [15]. This capability not only enhances diagnostic accuracy but also facilitates early intervention, which is crucial for improving patient outcomes.

2.3. Principal Component Analysis

This section explores the literature and concepts Table 1 related to PCA (Principal Component Analysis), providing foundational knowledge for its application in the study. PCA is a statistical technique widely used for dimensionality reduction and feature extraction in data analysis. By transforming the original variables into a new set of uncorrelated variables known as principal components, PCA helps summarize information from the data while retaining most of its variance. This method is particularly useful in various fields, including psychology and health sciences, where researchers often deal with complex, high-dimensional datasets. PCA not only simplifies the analysis process but also enhances the interpretability of results, making it easier for researchers to identify underlying patterns [16].

Furthermore, PCA serves to address issues related to multicollinearity that can impact the performance of analytical models. By reducing the dimensionality of the data, PCA allows for more efficient computation and can improve the robustness of predictive models. This technique has been successfully applied across various domains, including mental health research, where PCA aids in identifying key factors associated with mental disorders. As the demand for data-driven insights continues to rise, PCA remains an essential tool for researchers seeking to extract meaningful information from large and complex datasets [17].

In a related study, analyzed over 50 survey variables on COVID-19 psychological distress and observed a 10% increase in prediction accuracy after applying PCA. This study employed K-Nearest Neighbors as the primary classification method [18]. Focused on EEG signal analysis with more than 100 features, demonstrating that PCA enhanced classification accuracy by 12%. By leveraging methods like SVM and logistic regression, the research confirmed that PCA is a powerful tool for simplifying complex datasets while retaining crucial information. These studies collectively underscore PCA's effectiveness in reducing data complexity and improving model performance [19].

Similarly, analyzed text-based social media data and found that PCA improved feature optimization, resulting in higher accuracy when using logistic regression as the classification model [20].

Table 1. PCA for reduction dimension.

Authors	Dataset	Dimensions	Reduction Method	Classification	Result
Priya et al. (2020) [8]	DASS-42 Dataset	42 dimensions	PCA	Decision Trees, Random Forest	PCA reduced computational cost by 35% while maintaining clasification accuracy of 85%.

Table 1. *Cont.*

Authors	Dataset	Dimensions	Reduction Method	Classification	Result
Sundaraman-Stukel & Davidson (2024) [18]	Survey Data on COVID-19 Psychological Distress	50+ variables	PCA	K-Nearest Neighbors	PCA enhanced prediction accuracy by 10% and provided insights into key psychological factors.
Nguyen et al. (2023) [19]	EEG Signals	100+ features	PCA	SVM, Logistic Regression	PCA improved classification accuracy by 12% compared to raw data.
Lin et al. (2024) [20]	Social Media Posts	Text features	PCA	Logistics Regression	PCA optimized text-based feature sets, achieving 81.2% accuray in detection.

2.4. Hyperparameter

Hyperparameter optimization (HPO) is a vital process in machine learning that directly impacts a model's performance, convergence speed, and generalization ability. Several algorithms, including Random Search, Bayesian Optimization, and Cross-Entropy Optimization, have been developed for HPO, each offering distinct benefits in specific applications such as short-term load forecasting and training deep learning models [21,22]. Despite its advantages, HPO poses the risk of over fitting, where excessive fine-tuning leads to models that excel on training data but fail to generalize effectively to unseen data [23]. While HPO plays a crucial role in improving model accuracy, its application must be carefully managed to mitigate potential drawbacks like over fitting.

3. Analysis and Discussion

The methodological framework created here for the classification of mental health conditions using the main principles of a machine learning approach (more specifically PCA and logistic regression) was carefully designed to unveil how efficient, generalizable results can be achieved by systematically identifying private information on the diagnostic function of this kind of disorder. This process starts first with the assembling of a range of data from second datasets to provide a comprehensive understanding across multiple dimensions that contributes towards mental health outcomes. A rich dataset constitutes the foundational building blocks of subsequent methodologies, with an initial stage of data preprocessing involving cleaning and normalization to competently handle missing values, outliers, or inconsistencies. All of this is to prepare the dataset in such a way that it can go through feature extraction and the selection process (to reduce the number of data columns), which are equally essential as they consume a lot more time prior to building models.

In Figure 1, PCA is utilized to condense the high-dimensional dataset into a set of principal components—uncorrelated variables that capture the greatest variance—effectively simplifying the dataset's dimensionality with a minimal loss of essential information. Following this, logistic regression serves as the main classification method, chosen for its effectiveness in binary outcomes (predicting the presence or absence of a mental health condition). This process includes thorough model analysis, fitting, validation, and cross-validation to fine-tune parameters using extensive training data, assessing the model's ability to generalize. This structured approach aims not only for strong predictive accuracy but also for results that can be easily interpreted, assisting healthcare providers in making informed diagnostic and treatment decisions.

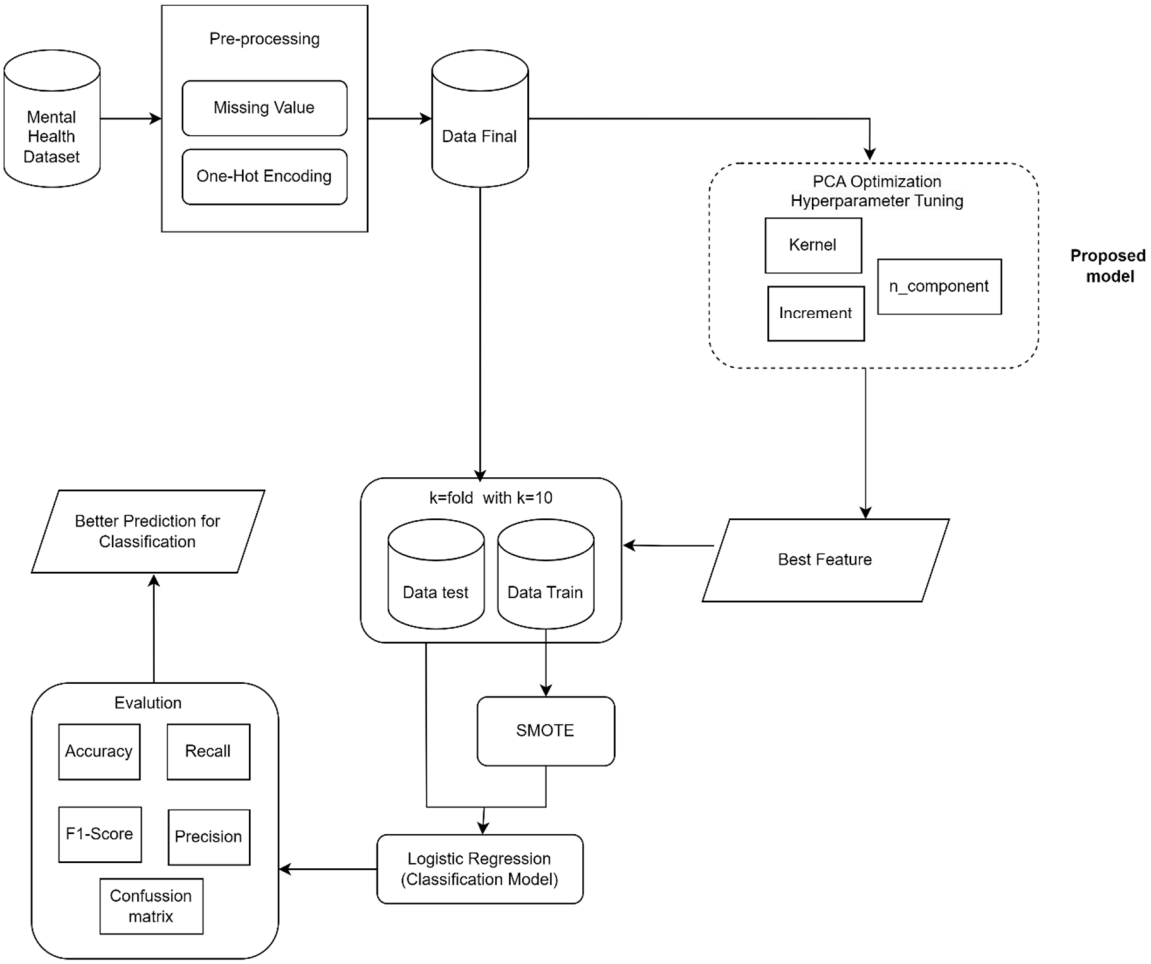

Figure 1. Research Frame Work PCA and Logistic Regression.

Regarding mental health classification with machine learning, evaluation methods are essential for measuring the model's performance accurately. Cross-validation is a key evaluation technique, where the dataset is divided into multiple subsets to ensure the model is tested on diverse data points. This approach helps reduce overfitting by offering a more reliable estimate of model performance. This method improves evaluation reliability and offers insight into the model's performance on new data.

Another key evaluation method involves using performance metrics to measure the classification model's effectiveness. Common metrics include accuracy, precision, recall, and F1-score, each offering unique insights into model performance. Accuracy assesses the overall correctness, while precision and recall evaluate the model's ability to detect true positive cases among predicted positives and actual positives, respectively. The F1-score, as the harmonic mean of precision and recall, provides a balanced assessment and is especially valuable for imbalanced datasets, which are common in mental health classification. Examining these metrics gives researchers a thorough understanding of the model's strengths and areas for improvement.

Additionally, confusion matrices serve as a visual tool for assessing the classification model's performance. A confusion matrix offers a detailed comparison of the model's predictions against actual results, showing the number of correct and incorrect classifications. This breakdown helps pinpoint specific challenges the model may face, such as difficulties in distinguishing between various mental health conditions. Reviewing the confusion matrix along with performance metrics allows researchers to make informed adjustments to the model or preprocessing steps to enhance the classification accuracy.

External validation is a vital part of the evaluation process, and involves testing the model on an independent dataset not used during training or initial evaluation. This step is essential for determining the model's generalizability to real-world situations. By applying the trained model to new data, researchers can evaluate its practical effectiveness, such as in clinical decision-making within mental health contexts. This thorough evaluation approach ensures that the logistic regression model, optimized with PCA, is both accurate and dependable for real-world mental health classification tasks.

4. Conclusions

The study focuses on integrating Principal Component Analysis (PCA) for dimensionality reduction and logistic regression for predictive modeling. The results indicate that the combination of PCA and logistic regression effectively handles the challenges of high-dimensional mental health data. PCA enhances the model's efficiency and clarity by filtering out noise and isolating the most critical features. Logistic regression further ensures accurate classification, making the framework both computationally efficient and clinically relevant.

The contributions of this study include efficient dimensionality reduction through PCA, the accurate classification of mental health conditions using logistic regression, and a practical framework to support clinicians in making data-driven decisions. While the approach shows promise, it has some limitations, such as potential biases due to linear assumptions and challenges with imbalanced datasets. Future research could incorporate advanced methods, such as ensemble techniques or neural networks, to address these issues while maintaining interpretability. Overall, this study presents a robust machine learning framework that bridges the gap between data analysis and clinical practice, contributing to better decision-making and improved mental health outcomes.

Author Contributions: Conceptualization, H.H.; methodology, M.M.D.; formal analysis, I.S.W. writing-original draft preparation and editing, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was founded by BAZNAS and PP Muhammadiyah No. 1304/I.3/D/2024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Forthman, K.L.; Colaizzi, J.M.; Yeh, H.; Kuplicki, R.; Paulus, M.P. Latent variables quantifying neighborhood characteristics and their associations with poor mental health. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1202. [CrossRef] [PubMed]
2. Iyortsuun, N.K.; Kim, S.-H.; Jhon, M.; Yang, H.-J.; Pant, S. A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare* **2023**, *11*, 285. [CrossRef] [PubMed]
3. Hassantabar, S.; Zhang, J.; Yin, H.; Jha, N.K. MHDeep: Mental Health Disorder Detection System based on Body-Area and Deep Neural Networks. *arXiv* **2021**, arXiv:2102.10435. [CrossRef]
4. Mancini, E.; Tanevska, A.; Galassi, A.; Galatolo, A.; Ruggeri, F.; Torroni, P. Promoting fairness and diversity in speech datasets for mental health and neurological disorders research. *arXiv* **2024**, arXiv:2406.04116. [CrossRef]
5. Doran, C.M.; Kinchin, I. The role of mental health on workplace productivity: A critical review of the literature. *Appl. Health Econ. Health Policy* **2023**, *21*, 167–193. [CrossRef]
6. Chung, J.; Teo, J. Single classifier vs. ensemble machine learning approaches for mental health prediction. *Brain Inform.* **2023**, *10*, 2–10. [CrossRef] [PubMed]

7. Deb, A.; Samadder, B.; Chowdhury, S.; Das, S.; Banarjee, S. Measuring mental health condition using logistic regression. *Int. J. Eng. Technol. Manag. Sci.* **2023**, *7*, 327–338. [CrossRef]
8. Priya, A.; Garg, S.; Tigga, N.P. Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Comput. Sci.* **2020**, *167*, 1258–1267. [CrossRef]
9. Haq, A.K. UI, Khattak, A.; Jamil, N.; Naeem, M.A.; Mirza, F. Data analytics in mental healthcare. *Sci. Program.* **2020**, *2020*, 2024160. [CrossRef]
10. Reddy, K.V.V.; Elamvazuthi, I.; Aziz, A.A.; Paramasivam, S.; Chua, H.N.; Pranavanand, S. An efficient prediction system for coronary heart disease risk using feature reduction and hyperparameter optimization. *Appl. Sci.* **2023**, *13*, 118. [CrossRef]
11. Rezapour, M.; Hansen, L. A machine learning analysis of COVID-19 mental health data. *Sci. Rep.* **2022**, *12*, 14965. [CrossRef] [PubMed]
12. Rahman, M.A.; Hossain, M.F.; Hossain, M.; Ahmmed, R. Employing PCA and t-statistical approach for feature extraction and classification of emotion from multichannel EEG signal. *Egypt. Inform. J.* **2020**, *21*, 23–35. [CrossRef]
13. Gautam, S.; Jain, A.; Chaudhary, J.; Gautam, M.; Gaur, M.; Grover, S. Concept of mental health and mental well-being, its determinants and coping strategies. *Indian J. Psychiatry* **2024**, *66*, S231–S244. [CrossRef] [PubMed]
14. Dunbar, O.R.A.; Nelsen, N.H.; Mutic, M. Hyperparameter optimization for randomized algorithms: A case study for random features. *arXiv* **2024**, arXiv:2407.00584. [CrossRef]
15. Radwan, A.; Amarneh, M.; Alawneh, H.; Ashqar, H.I.; AlSobeh, A.; Magableh, A.A. Predictive analytics in mental health leveraging LLM embeddings and machine learning models for social media analysis. *Int. J. Web Serv. Res.* **2024**, *21*, 1–22. [CrossRef]
16. Kim, K.; Oh, H.S. Principal component analysis in the graph frequency domain. *arXiv* **2024**, arXiv:2410.08422. [CrossRef]
17. Reinbott, F.; Janßen, A. Principal component analysis for max-stable distributions. *arXiv* **2024**, arXiv:2408.10650. [CrossRef]
18. Sundaram-Stukel, R.; Davidson, R.J. Associational and plausible causal effects of COVID-19 public health policies on economic and mental distress. *arXiv* **2021**, arXiv:2112.11564. [CrossRef]
19. Nguyen, T.H.T.; Le, B.; Nguyen, P.; Tran, L.G.H.; Nguyen, T.; Nguyen, B.T. Principal Components Analysis Based Imputation for Logistic Regression. In *Advances and Trends in Artificial Intelligence. Theory and Applications, Proceedings of the IEA/AIE, Shanghai, China, 19–22 July 2023*; IEA: Paris, France, 2023; pp. 28–36. [CrossRef]
20. Lin, C.; Hu, P.; Su, H.; Li, S.; Mei, J.; Zhou, J.; Leung, H. Sensemood: Depression detection on social media. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020; pp. 407–411. [CrossRef]
21. Hakyemez, T.C.; Adar, O. Testing the efficacy of hyperparameter optimization algorithms in short-term load forecasting. *arXiv* **2024**, arXiv:2410.08803. [CrossRef]
22. Li, K.; Li, F. Cross-entropy optimization for hyperparameter optimization in stochastic gradient-based approaches to train deep neural networks. *arXiv* **2024**, arXiv:2409.09240. [CrossRef]
23. Tetko, I.V.; van Deursen, R.; Godin, G. Be aware of overfitting by hyperparameter optimization! *arXiv* **2024**, arXiv:2407.20786. [CrossRef]