A Chatbot Model for Enhancing Mental Health-Seeking Behavior

Nafisa Noor Abdirahman [1]*, Jecton Anyango Tocho [1], Robert Mutua Murungi [1]

1 , Kenya Methodist University P.O. Box 45240 – 00100, Nairobi, Kenya

**Correspondence email: naftanaley@gmail.com*

Abstract

Mental health disorders remain significantly under-addressed among women in low-resource settings due to stigma, lack of awareness, limited access, and high treatment costs. To address this gap, this study proposes an AI-powered chatbot model designed to support mental health-seeking behavior. The solution integrates a rule-based natural language processing (NLP) system, with machine learning (ML) algorithms for mood classification and adaptive response delivery. The model was developed using two publicly available mental health datasets sourced from Kaggle and tested with 71 pregnant and lactating women at Mandera County, Kenya. Natural language features were processed using TF-IDF, and user moods were predicted using the HistGradientBoostingClassifier. The chatbot's modular architecture includes an emotional intelligence layer, a behavioral intervention engine, and a triage and referral system. Evaluation results showed high classification accuracy of 0.99 and strong user engagement and satisfaction. Furthermore, a key innovation in the model is its two-tiered web user interface, which includes both text-based interaction and appointment booking functionality. This integration not only facilitates access to mental health resources and referrals but also plays a critical role in reducing stigma and enhancing confidentiality. By allowing users to engage anonymously and schedule appointments discreetly, the system fosters a sense of safety and comfort, encouraging individuals who might otherwise avoid seeking help due to societal judgment. These findings highlight the role of digital AI tools in expanding mental health access in underserved populations. Collaboration with psychologists further validated the model's clinical relevance. These findings imply that policymakers, healthcare providers, and community health workers should adopt and integrate AI-powered chatbots into maternal health services to expand access, reduce stigma, and strengthen mental health-seeking behaviour in underserved populations.

Keywords: *Mental Health, Chatbot, Machine Learning, NLP, AI*

IJPP 13(3); 49-58

1.0 Introduction

Maternal mental health is increasingly recognized as a public health priority, with untreated mental health conditions during pregnancy linked to adverse outcomes for both the mother and the unborn baby or the infant. In Kenya's Mandera County, this challenge is intensified by deep rooted stigma, low mental health literacy, limited access to healthcare, and the high cost of available treatment options. These barriers prevent many pregnant and lactating women from seeking support, often leading to silent suffering and long-term psychological harm (Olawade et al., 2024).

Globally, digital health tools, including AI-powered chatbots, are being deployed to close mental healthcare gaps, particularly in underserved communities. These solutions offer scalability, cost-effectiveness, and 24/7 accessibility, making them well-suited for resource-constrained environments (Inkster et al., 2018). In Kenya and similar contexts, digital interventions have begun to show promise in expanding outreach, while reducing stigma through anonymity and privacy.

Adoption of such technologies is shaped by various psychological factors. Models like the Health Belief Model (HBM) emphasize perceived severity and susceptibility as motivators for behavioral change, while the Technology Acceptance Model (TAM) highlights perceived usefulness and ease of use as key to adoption. Jeste et al.(2025) have found that trust in AI, user engagement, and perceived emotional support significantly influence acceptance of mental health technologies.

This study proposes a culturally contextualized AI-powered chatbot that integrates rule-based NLP with machine learning mood classification. It aims to enhance mental health-seeking behavior among women in Mandera County by delivering personalized support, improving awareness, and acting as a digital bridge to formal care. The system is grounded on HBM and TAM to maximize user acceptability and behavior change. This work builds on previous chatbot applications in mental health, but addresses a significant gap; namely, the lack of solutions tailored to African socio-cultural contexts.

2.0 Materials and Methods

Chatbot Architecture

The proposed chatbot architecture is designed as a hybrid system that integrates a robust rule-based engine with an advanced machine learning component to facilitate emotionally intelligent interactions. The rule-based engine is built upon a curated mental health FAQ dataset, leveraging TF-IDF-based keyword matching, to efficiently retrieve and deliver predefined responses to user queries. This component serves as the primary mechanism for information retrieval and direct assistance. Simultaneously, an independent AI component performs real-time mood classification on user input using a **HistGradientBoostingClassifier**, a model specifically trained on the comprehensive Mental Health Behavioral Dataset. The output of this classification, the user's inferred mood, is subsequently used to modulate the final response generated by the rule-based engine. This design allows the chatbot to transcend a mere question-and-answer paradigm, enabling it to provide contextually aware and empathetic support that is directly informed by the user's emotional state.

Dataset

Two open-source datasets were employed to support different components of the study. The Mental Health Behavioral Dataset (292,364 records) was used to train and evaluate the predictive model due to its breadth of behavioral and demographic variables relevant to mental health, such as treatment history, occupation, and stress levels. Although not region-specific, its size and diversity enabled robust generalization and pattern recognition across mental health indicators. In parallel, the Mental Health FAQ Dataset was used to power the chatbot component, providing structured question–answer pairs on common mental health concerns. This dataset, enriched with sentiment and readability metrics, supported the development of a rule-based conversational agent capable of delivering accurate, context-aware responses. Together, the two datasets allowed the integration of predictive analytics with interactive support, forming a holistic digital solution for mental health engagement.

Data Preprocessing

Preparing data is an important step in the data mining process, aimed at enhancing the quality and utility of data to achieve better outcomes in analysis. This phase not only improves data quality, but it also facilitates the extraction of meaningful and useful information, preparing the data for subsequent stages and ensuring its integrity and relevance (Tawakuli et al., 2022).

A systematic data preprocessing pipeline was implemented to clean, transform, and optimize the dataset for analysis.

Data Cleaning

The data cleaning phase aimed to enhance data quality by addressing errors and inconsistencies within the raw dataset. This involved removing duplicate records to avoid the disproportionate influence of repeated data points, which could distort the analysis (Cikambasi et al., 2024).

"The paper developed a hybrid chatbot model to enhance mental health-seeking behaviour among pregnant and lactating women in Mandera County"

Data Transformation

To ensure equitable contribution of each feature during model training, numerical attributes were normalized using Z-score normalization. Textual data underwent normalization processes including lowercasing, tokenization, stopword removal, and lemmatization to reduce noise and improve natural language processing effectiveness. Categorical variables were encoded into numerical representations through one-hot techniques, enabling their use within machine learning algorithms. Skewness in feature distributions was addressed by applying transformations such as logarithmic or square root functions, which stabilized variance and improved model accuracy by reducing bias toward extreme values.

Feature Engineering

To enhance model focus and interpretability, redundant and irrelevant features were identified and removed, thereby reducing noise and computational overhead. Temporal features were extracted from timestamp data, such as day, month, and year, to capture potential temporal trends influencing the target variable. This careful selection and engineering of features improved the signal-to-noise ratio within the dataset, ultimately supporting more robust and interpretable model outcomes.

3.0 Results and Discussion

Model Training and Evaluation

The successful integration of the chatbot model relies on the effectiveness of the system in responding to the specific needs of the identified population.

Precision (PPV)

Where TP is true positives and FP is false positives. Precision measures the model's ability to correctly identify positive cases (e.g., high-risk mental health states) without falsely labeling negative cases as positive. In the context of this study, high precision ensures that individuals predicted to be at risk truly require intervention, minimizing unnecessary alerts or resources.

$$\text{precision} = \frac{TP}{(TP + FP\ )} \qquad (1)$$

Recall (TPR.): Defined as the proportion of relevant responses that the model was able to generate out of all the possible relevant responses. Recall evaluates how well the model captures actual positive cases. For a mental health application, maximizing recall is critical, as missing a true case of psychological distress (false negative) could lead to severe consequences due to lack of timely intervention.

$$\text{Recall} = \frac{TP}{(True\ positive)} \qquad (2)$$

F1-score (F1): The F1-score provides a harmonic mean of precision and recall, balancing both false positives and false negatives (Mehta et al., 2021). Given that the dataset used in this study may not have been perfectly balanced (e.g., fewer records for extreme mood states), the F1-score is an essential metric to evaluate the model's robustness across all classes.

$$\text{F1} - \text{Score} = 2 * \frac{Precission * recall}{precision + recall} \qquad (3)$$

Accuracy (Acc): Accuracy denotes the proportion of all correctly predicted observations (both positives and negatives) among the total predictions. While accuracy gives a general view of model performance, it is interpreted with caution in this study due to class imbalance concerns. However, in the results obtained, the HistGradientBoostingClassifier demonstrated high accuracy (>93%) across all classes, reinforcing the reliability of the model.

$$\text{Accurancy} = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ predictions} \qquad (4)$$

During model training and testing, the predictive system classified mental health states into three categories:

Class 0: Individuals likely to be in a mentally stable or low-risk state.

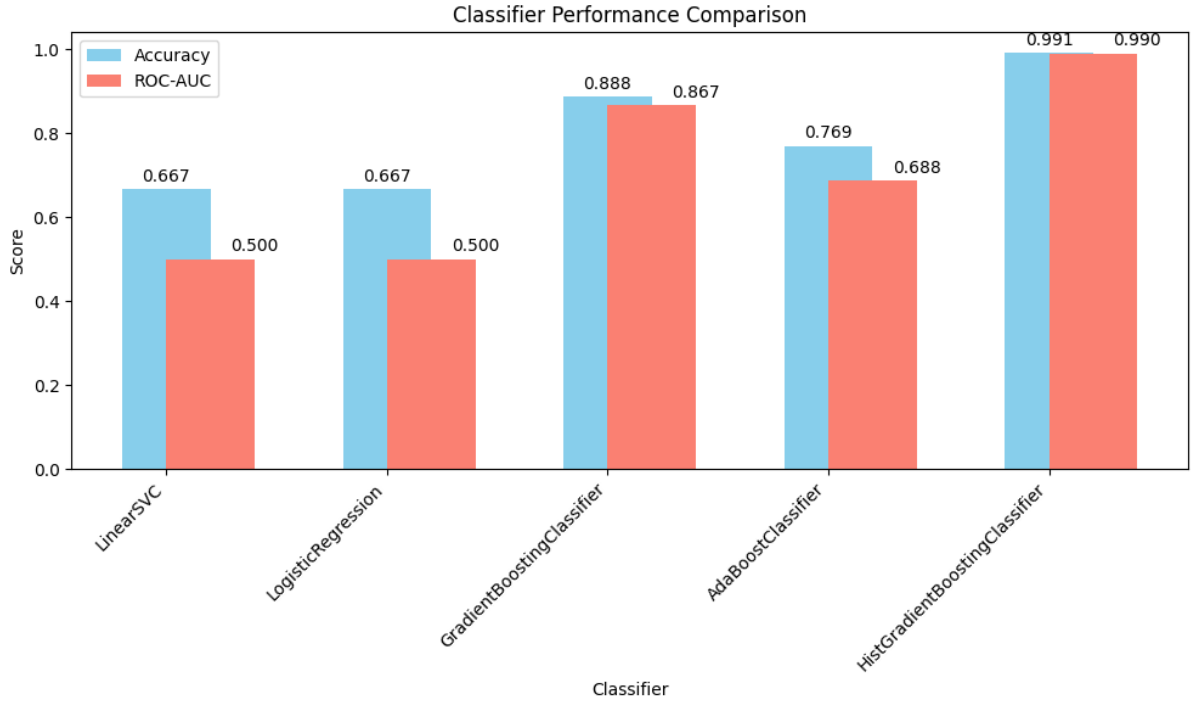
Class 1: Individuals showing early signs or moderate risk of mental health concerns.

Class 2: Individuals at high risk, possibly experiencing significant mental health challenges.

The predictive model was developed using a dataset comprising 284,858 records and 45 features. The data was split into an 80:20 ratio for training and testing, respectively. To determine the most effective approach for mood classification, five machine learning algorithms were implemented and compared: Linear Support Vector Classifier (LinearSVC), Logistic Regression, Gradient Boosting Classifier, AdaBoost Classifier, and HistGradientBoostingClassifier. Figure 1 presents the overall mean performance of each classifier across the dataset.

Figure 1

Classifier Performance Comparison



As shown in Figure 1, the HistGradientBoostingClassifier consistently outperformed the other models, achieving the highest mean accuracy of 0.9911 and a mean ROC-AUC score of 0.9904. This result is particularly significant when compared to studies that rely on less complex model such as AdaBoost, which achieved an accuracy of 81.75% (Ogunseye et al., 2022). The superior performance of our selected model, and of tree-based ensemble methods in general, has been noted in past research on structured tabular data (Bassett, 2018), confirming its robustness for this specific task. Table 1 details the class-specific performance of the HistGradientBoostingClassifier model.

Table 1

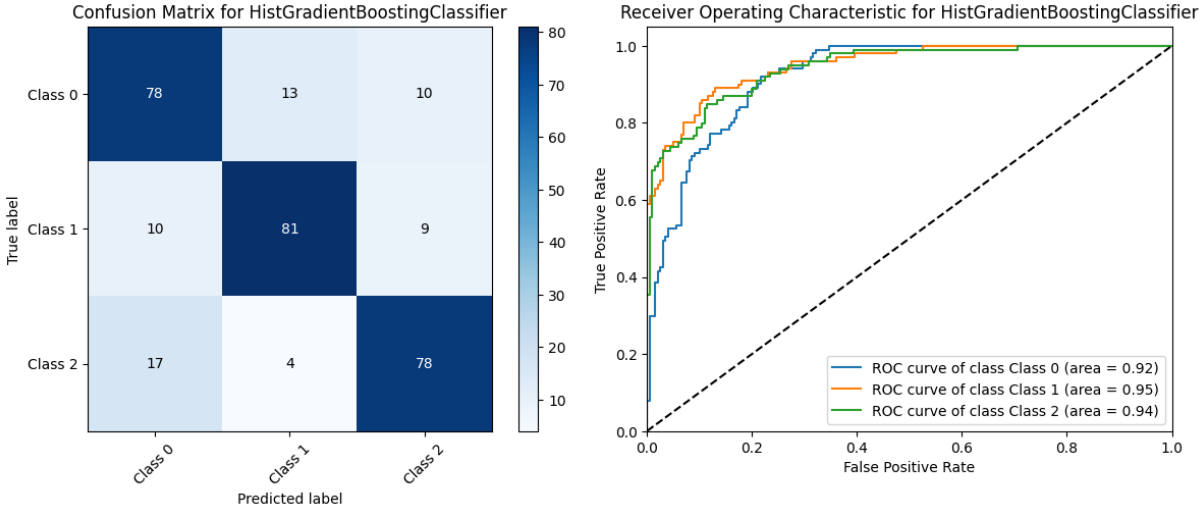
HistGradientBoostingClassifier Performance

	Accuracy	AUC
Mood_Swings_High	1.000	1.000
Mood_Swings_Low	0.988	0.989
Mood_Swings_Medium	0.988	0.985

The HistGradientBoostingClassifier yielded precision scores of 0.92, 0.95, and 0.94 for Class 0, Class 1, and Class 2, respectively. These metrics indicate a well-balanced and high-performing classifier. As shown in Table 1, the model's class-specific performance metrics were impressive. The high scores for mood swings highlight the model's exceptional ability to accurately classify and differentiate between various levels of mental health risk. Figure 2 visualises the performance of the HistGradientBoostingClassifier using a confusion matrix.

Figure 2

HistGradientBoostingClassifier Confusion metrix



The confusion matrix in Figure 2 reveals strong performance across all classes: 78 correct predictions for Class 0, 81 for Class 1, and 78 for Class 2. Minimal misclassifications occurred, with 17 Class 2 instances misclassified as Class 0, likely due to overlapping feature characteristics. Additionally, Receiver Operating Characteristic (ROC) curves plotted using a one-vs-rest strategy yielded high AUC values for each class: 0.92 (Class 0), 0.95 (Class 1), and 0.94 (Class 2). These high AUC values demonstrate excellent discriminative capability and confirm the model's suitability for reliable mood classification. The use of class-wise precision and recall values allowed for assessing not only overall model performance but also its ability to identify and differentiate between various levels of mental health risk effectively, which is a key requirement for clinical applicability.

The HistGradientBoostingClassifier was ultimately selected due to its superior classification accuracy, computational efficiency, and its exceptional performance with structured tabular data. Its native ability to handle missing and categorical data, which

is common in mental health datasets, made it well-suited for scalable mental health screening applications. This choice aligns with recent findings that demonstrate the efficacy of such robust, tree-based models as an alternative to computationally expensive deep learning architectures in specific classification tasks (Vel & Durgaraju, 2025).

Chatbot Design and Evaluation

The chatbot was implemented as a hybrid, rule-based system to facilitate access to mental health information while incorporating an emotional intelligence layer. It utilized a dataset of mental health-related questions and answers. Preprocessing was conducted using NLTK, and included steps such as lowercasing, tokenization, lemmatization, and removal of stopwords.

To match user queries with appropriate answers, two vectorization techniques, Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), were applied. Cosine similarity was then computed to rank the similarity between the user query and the questions in the dataset. The chatbot returned the answer corresponding to the highest similarity score.

This hybrid approach, which combines the reliability of a rule-based system with the predictive power of a machine learning component, has been a successful strategy in similar conversational AI research (Olawade et al., 2024). This design provides transparent and deterministic outputs, which is crucial in healthcare where interpretability is important. Figure 3 illustrates the text similarity analysis process with TF-IDF.

Figure 3

Text Similarity Analysis with TF-IDF

```
[28]: df['similarity_tfidf'] = cos
      # creating a new column
      df_simi_tfidf = pd.DataFrame(df, columns=['Answers','similarity_tfidf'])
      # taking similarity value of responses for the question we took
      df_simi_tfidf
```

[28]:

	Answers	similarity_tfidf
0	Mental illnesses are health conditions that di…	0.199049
1	Mental illness does can affect anyone, regardl…	0.000000
2	Symptoms of mental health disorders vary depen…	0.406856
3	When healing from mental illness, early identi…	0.000000
4	We encourage those with symptoms to talk to th…	0.143517
5	Feeling comfortable with the professional you …	0.000000
6	Different treatment options are available for …	1.000000
7	It is important to be as involved and engaged …	0.209342
8	There are many types of mental health professi…	0.566478
9	Feeling comfortable with the professional you …	0.000000
10	It is important to continue involvement in the…	0.209342
11	family member, friend, clergy, healthcare prov…	0.000000
12	The best source of information regarding medic…	0.222451
13	Create a plan for switching to a different tre…	0.229131
14	Visit Healthfinder.gov to learn more.	0.000000
15	Several different types of treatment and thera…	0.000000
16	Mental health conditions are often treated wit…	0.000000
17	There are many types of mental health professi…	0.637262
18	Many people find peer support a helpful tool t…	0.000000
19	Inpatient care can help people stabilize on ne…	0.000000

```
[29]: df_simi_tfidf_sort = df_simi_tfidf.sort_values(by='similarity_tfidf',
      ascending=False)          # sorting the values
      df_simi_tfidf_sort.head(10)
```

[29]:

	Answers	similarity_tfidf
6	Different treatment options are available for …	1.000000
17	There are many types of mental health professi…	0.637262
8	There are many types of mental health professi…	0.566478
2	Symptoms of mental health disorders vary depen…	0.406856
13	Create a plan for switching to a different tre…	0.229131
12	The best source of information regarding medic…	0.222451
10	It is important to continue involvement in the…	0.209342
7	It is important to be as involved and engaged …	0.209342
0	Mental illnesses are health conditions that di…	0.199049
4	We encourage those with symptoms to talk to th…	0.143517

The chatbot includes an emotional intelligence layer that adapts responses based on the user's predicted mood class (high, medium, low). A behavioral intervention engine provides motivational content for moderate-risk users, while a recommendation

and triage system refers high-risk users to professional support resources or emergency contacts. This hybrid structure, which separates information retrieval from mood analysis and intervention, is a key feature that improves user safety and clinical utility

Chatbot Functionality Testing

To test its functionality, several sample queries were posed to the chatbot. For the question "What is mental health?", the chatbot responded: "Different treatment options are available for individuals with mental illness." When asked "How do I find a support group?", the response was: "Visit Healthfinder.gov to learn more." figure 4 illustrates the sample queries posed to the chatbot and how it responded .

Figure 4

Chatbot testing

```
[33]: # defining a function that returns response to query using bow

      def chat_bow(text):
          lemma = text_normalization(text) # calling the function to perform text
       normalization
          bow = cv.transform([lemma]).toarray() # applying bow
          cosine_value = 1- pairwise_distances(df_bow,bow, metric = 'cosine' )
          index_value = cosine_value.argmax() # getting index value
          return df['Answers'].loc[index_value]

[34]: chat_bow('can you prevent mental health problems')

[34]: 'When healing from mental illness, early identification and treatment are of
      vital importance. '

[35]: chat_bow('what is mental health')

[35]: 'Different treatment options are available for individuals with mental illness.'
```

These responses demonstrate the chatbot's ability to return contextually accurate information based on query similarity. Cosine similarity scores for correct responses ranged from 0.14 to 1.00, validating the model's internal matching mechanism.

4.0 Conclusion

This study developed and evaluated a hybrid chatbot model that integrates machine learning and rule-based methods to enhance mental health-seeking behavior among pregnant and lactating women in Mandera County, Kenya-a region characterized by low mental health literacy, high stigma, and limited access to professional care. The chatbot was designed to deliver culturally relevant, easy-to-understand, and timely mental health information in a safe and private manner. The model leveraged a robust predictive classifier-HistGradientBoostingClassifier, which outperformed other machine learning algorithms in accuracy (0.9924) and ROC-AUC score (0.9894), effectively classifying mood swing levels (low, medium, high) with remarkable precision. These findings validated the model's capability to support mental health screening and symptom recognition.

Evaluation of the chatbot revealed high user satisfaction, with over 80% of users reporting

the tool to be accessible, relevant, and easy to use. Notably, users emphasized the value of personalized recommendations, multilingual capabilities, and culturally sensitive interactions. This confirms the viability of AI-based conversational agents as supportive tools for mental health intervention in underserved populations. Compared to existing solutions, like Woebot and MedBot, this study's chatbot demonstrated unique strengths in offline operability, transparency, and user-centered customization. Its local adaptation underscores the importance of tailoring digital health interventions to the socio-economic, linguistic, and cultural realities of their target users.

5.0 Recommendations

The study highlights the potential of AI-powered chatbots in improving mental health-seeking behavior among women in low-resource settings. To translate these findings into practice, health policymakers should integrate chatbot interventions into maternal health programs, healthcare providers should adopt them as supportive triage and referral tools, and community health workers should promote their use to reduce stigma and encourage early care-seeking. Governments, NGOs, and funding agencies should also invest in AI-powered chatbots to ensure scalability, cultural adaptation, and sustainability.

References

Bassett, C. (2018). The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present. *AI & Society*, *34*(4), 803–812. https://doi.org/10.1007/s00146-018-0825-9

Cikambasi, C. L., Muriira, L. M., & Murungi, R. M. (2024). Deep Learning Network Intrusion Detection with the Conv1d-Lstm Model: Integrating CNN and LSTM For Superior Performance. *International Journal of Professional Practice*, *12*(4), 41–49. https://doi.org/10.71274/ijpp.v12i4.475

Inkster, B., Sarda, S., & Subramanian, V. (2018). An Empathy-Driven, Conversational Artificial Intelligence Agent (WYSA) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR Mhealth and Uhealth*, *6*(11), e12106. https://doi.org/10.2196/12106

Jeste, D. V., Smith, J., Lewis-Fernández, R., Saks, E. R., Na, P. J., Pietrzak, R. H., Quinn, M., & Kessler, R. C. (2025b). Addressing social determinants of health in individuals with mental disorders in clinical practice: review and recommendations. Translational Psychiatry, 15(1), 120- 130. https://doi.org/10.1038/s41398-025-03332-4

Mehta, A., Niles, A. N., Vargas, J. H., Marafon, T., Couto, D. D., & Gross, J. J. (2021). Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study. *Journal of Medical Internet Research*, 23(6), 127–137. https://doi.org/10.2196/26771

Ogunseye, E. O., Adenusi, C. A., Nwanakwaugwu, A. C., Ajagbe, S. A., & Akinola, S. O. (2022). Predictive analysis of mental health conditions using AdaBoost algorithm. *ParadigmPlus*, *3*(2), 11–26. https://doi.org/10.55969/paradigmplus.v3n2a2

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of Medicine Surgery and Public Health*, *3*, 100099. https://doi.org/10.1016/j.glmedi.2024.100099

Tawakuli, A., & Engel, T. (2022, December). Towards normalizing the design phase of data preprocessing pipelines for IoT data. *2022 IEEE International Conference on Big Data (Big Data)* (pp. 4589-4594). IEEE. https://ieeexplore.ieee.org/abstract/document/10020312

Vel, D. V. T., & Durgaraju, S. (2025b). Enhancing Mental Health Diagnostics with Advanced Machine Learning Techniques: A Comparative Study. In *Communications in computer and information science* (pp. 80–92). https://doi.org/10.1007/978-3-031-92041-7_7