Evaluation of Interpretable Machine Learning Methods in Depression Diagnosis

Ching Hong So, Guilherme Gryschek, João Rafael Calixto Carvalho, Katja Mariko Wilde, Shweta Prasad Ghaisas; Theme 3: Interpretable Machine Learning

Introduction

According to the World Health Organization (WHO), more than 280 million individuals globally experience depression, with symptoms ranging from persistent sadness and fatigue to cognitive impairments and physical ailments [1]. This condition significantly impacts quality of life and contributes to disability, economic burden, and mortality due to suicide [1]. Depression is notoriously difficult to diagnose, given its multifaceted and often overlapping symptoms [2].

The complexity of depression diagnosis has driven growing interest in artificial intelligence (AI) and machine learning (ML) technologies, which can analyze large datasets to identify subtle patterns indicative of the disorder [3, 4]. Unlike traditional diagnostic methods, which rely on qualitative assessments and are subject to human biases and constraints [5], ML offers scalable solutions capable of detecting linguistic, behavioral, and physiological markers of depression [3, 6]. Advances in Natural Language Processing (NLP) have been especially impactful, enabling the automated analysis of text and speech data to identify features such as tone, sentiment, and word choice [7].

While ML in depression detection has demonstrated impressive predictive accuracy, much of this progress has relied on complex deep-learning models that often function as "black boxes" [7-15]. These models provide little insight into how predictions are made, creating a lack of transparency that undermines trust, particularly in sensitive areas like mental health [16]. To address this challenge, Interpretable Machine Learning (IML) and Explainable AI (XAI) methods have emerged, offering tools to clarify the reasoning behind algorithmic predictions [16]. Techniques such as SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) enable clinicians to understand how specific features contribute to a diagnosis [16]. These explanations are vital for ensuring trust, ethical accountability, and informed decision-making in clinical settings [7-16]. Furthermore, regulatory frameworks like the European Union's General Data Protection Regulation (GDPR) reinforce the need for AI transparency, establishing a "right to explanation" for decisions in high-stakes domains such as healthcare [17].

Despite advancements in IML, a significant knowledge gap persists in understanding how clinicians interpret and utilize these explanations in practice [13, 16]. Many IML methods are designed with technical rigor but lack sufficient consideration of clinical workflows and user-friendliness [13, 16, 18]. Consequently, clinicians may struggle to effectively integrate these tools into their diagnostic processes, limiting the potential of IML to improve patient outcomes [18]. Addressing this challenge is vital for ensuring the practical viability of IML in real-world healthcare settings.

This study aims to bridge the gap between IML technology and clinical application in depression detection. The primary objective is to explore how clinicians would perceive, interpret, and utilize IML-generated explanations, with a focus on identifying their needs and challenges through evaluation methods. By doing so, this research seeks to inform the design of more effective IML tools that enhance clinical decision-making and foster integration into everyday practice.

The research seeks to answer the following questions:

RQ1: How do clinicians interpret explanations generated by various interpretability techniques?

RQ2: Which interpretability techniques do clinicians find most effective and understandable?

RQ3: How do different techniques impact clinicians' decision-making processes?

By addressing these questions, the study aims to contribute valuable insights to the emerging field of interpretable AI in healthcare, paving the way for future innovations that support clinicians and improve patient care.

Methods

Participants

A convenience sample was invited to participate in two phases, as explained below. The participants were psychologists, psychiatrists or general practitioners with clinical experience in mental healthcare.

Inclusion Criteria: Healthcare professionals with practical experience in diagnosis and treatment of patients with depression in their clinical practice and little/basic understanding about ML; fluent in Portuguese or English (languages team members were fluent in).
Exclusion Criteria: Healthcare professionals with large expertise in ML applied to healthcare settings.

Process

Since both qualitative and quantitative components were essential to this study, a mixed methods approach was deemed appropriate since it would allow us to get a better overall picture [19]. The study consisted of three steps:

- Preliminary interviews with the participants to gather feedback related to visualization and explainability expectations
- Development of ML models and explanation techniques
- Evaluation based on a structured questionnaire (open questions and Likert-scale) to assess prediction's and model's understanding, trustworthiness, usefulness, and cognitive workload

Preliminary interviews

Seven participants were recruited by contacting healthcare professionals known to team members who met the inclusion criteria. Initial questionnaires (developed in Portuguese and English) were shared with these participants to gather feedback on their preferences for presenting predictions and explanations (Link 1 in Appendix).

Dataset and permission

The Distress Analysis Interview Corpus- Wizard of Oz (DAIC- WOZ) dataset, containing clinical interviews of 219 patients (audio, video, and transcripts), was utilized for ML models development [20]. For the ML task, the models focused solely on plain text, which consisted of the transcripts of patients' answers during the interviews, along with time stamps and data labels. The data labels classified patients into two categories: class 1 (depression) and class 0 (no depression), based on their Patient Health Questionnaire-8 (PHQ-8) scores. This approach, which excluded audio and video data, was considered optimal for achieving the objectives of the study. Before accessing the dataset, all team members reviewed and signed the DAIC-WOZ End-User License Agreement [21], which outlines guidelines for intended users, according to the University of South Carolina Institute for Creative Technologies [20].

Model development

We extracted multiple features from the text based on literature [22] and studies that had used the same dataset [23]. Various combinations of features were used to train multiple models and the combination with the best results was used to present explainability techniques (Tables A2, A3 in Appendix). Three ML models were selected to be included in this study: Logistic Regression (LR) [24], EXtreme Gradient Boosting (XGBoost) [24], and Bidirectional Encoder Representations from Transformers (BERT) [25].

LR and XGBoost used extracted features, while BERT processed tokenized text into a dot-separated sentence format suitable for input to the transformer [25]. BERT was chosen for its ability to process raw text, requiring minimal preprocessing. Performance was assessed using accuracy (intuitive for clinicians) and F1 score (balancing precision and recall to account for false positives and false negatives) [26]. Model hyperparameters, data preprocessing and development details are presented in Appendix (Table A2 and Link 3)

Development of explanation methods

For the explainability methods, different techniques were used depending on the ML model and the expectations gathered in the preliminary interviews. We developed global (holistic) and local interpretability (case based) if the technique allowed it.

Evaluation questionnaire

An application-grounded evaluation approach was employed to assess explainability methods in IML which involves testing explanations directly in real-world scenarios or environments that closely simulate practical applications [25, 26, 27, 28]. The evaluation focused on understanding the utility of these explanations as end-users who are health professionals responsible for using clinical decision support systems that use ML for depression diagnosis.

Each participant received a unique identification number, and model explanations were evaluated separately in randomized order to minimize bias. Descriptions of the model and explainability methods were provided in written format, and clinician's understanding was assessed through a structured online questionnaire for that particular model. This was repeated for the next 2 models as well in this order (Link 2 in Appendix). Team members were available to answer questions if needed but, in order to avoid bias, only interacted with participants if asked to. Two instances from the dataset were selected as showcases - one labeled as depression (ID 641) and the other without depression (ID 471) - and their interview transcript was shown to provide a human interpretability baseline. Participants then accessed model predictions, performance, and explanations in their preferred language (Portuguese or English). After reading, they answered an open question on their understanding and Likert-scale questions on understanding, trust, and usefulness. Load of the task was measured using the NASA Task Load Index [27]. Finally, they provided feedback and suggestions for improvement.

Analysis of responses to the questionnaire

Quantitative analysis of Likert scores was used to assess the participants' understanding, trustworthiness, usefulness, and cognitive load. Qualitative analysis was conducted on open questions data to identify feedback patterns regarding model explainability techniques.

Ethical considerations

The dataset used in this study is public in which patient data has been anonymized. No extra actions were deemed necessary to ensure confidentiality and privacy [20]. For the participants, informed consent was obtained, ensuring they understood the purpose of the research, the use of their data, and the measures taken to ensure anonymity. Participants were also assured of their right to withdraw at any time without consequences, emphasizing the voluntary nature of their participation [29]. The questionnaire was made available in Portuguese and English, with participants being offered documentation in their language of fluency, seeking to address the issue of access to and production of scientific data in a multicultural context [30].

Results

Preliminary interview result

A preliminary interview was conducted to gather insights into end-user requirements regarding the explainability of ML. Refer to Table A1 in Appendix for the participants' demographic details.

Participants strongly preferred visualizations, particularly graphs, to enhance the interpretability of ML predictions. They emphasized that, besides raw text data, the ML model should incorporate language patterns as an input for prediction. Furthermore, probability-based outputs were favored indicating the likelihood of a patient developing depression, accompanied by a logical explanation of the contribution of each feature to the prediction.

Performance of ML models

Table 1 summarizes the ML model performances. BERT achieved the highest accuracy and F1 score.

Table 1 - Final Models

Models	Accuracy	F1 Score	Feature Engineering
BERT	0.92	0.91	No
LR	0.59	0.47	Yes
XGBoost	0.67	0.37	Yes

Visualisations of explanation methods

To facilitate participants' understanding of the ML models, explainability techniques like LIME, odd ratios of features and SHAP were used for BERT, LR and XGBoost respectively. Figures 1, 2, 3 are examples of various visualisations used in the evaluation questionnaire. Figures A1-A7 in the Appendix demonstrate all interpretability techniques presented to the participants.

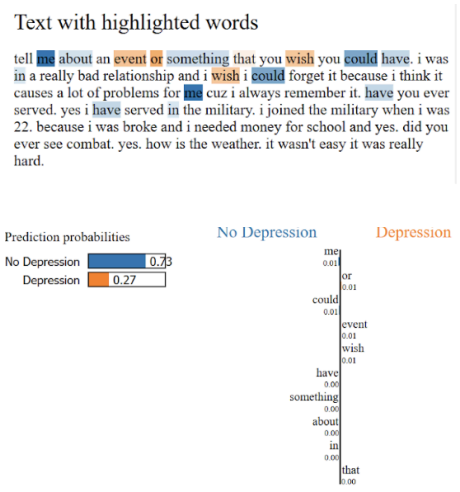


Figure 1- LIME (local interpretability) for case ID 641 (depression)

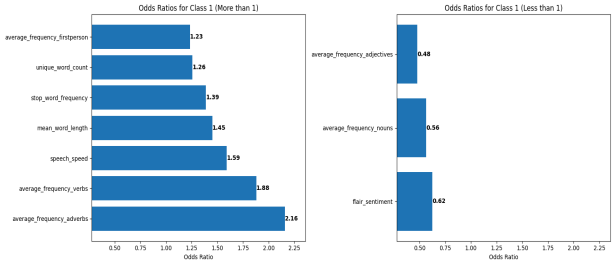


Figure 2- Odds ratios of features (global interpretability) for class 1 (depression)

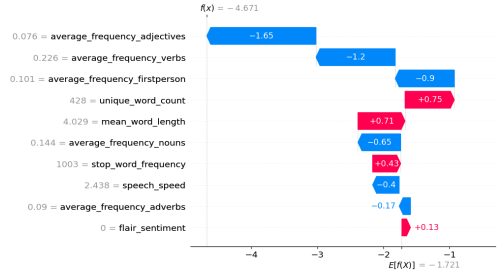


Figure 3- SHAP waterfall plot (local interpretability) for case ID 641 (depression)

Evaluation of responses to the questionnaire

Shown below are the scores of the Likert scale[1] and summary of answers[2] to the open ended questions. Refer to Table A1 in Appendix for the participants' demographic details.

Understanding

- BERT: LIME facilitated participants' understanding of BERT's prediction (4.43), and provided a clear explanation of the prediction (4.29) in the form of highlighted words considered for the prediction.
- LR: The coefficients associated with features were clear (4) and helped (4) the participants to understand the prediction. The bar chart of odds ratios demonstrated clearly how variables impact the prediction. The participants expressed concerns about requiring a certain level of mathematical knowledge to understand the model.
- XGBoost: SHAP was ranked as the least clear (3.29) and least helpful (3.71) in explaining the prediction, which participants suggested may be due to the complex and difficult SHAP graphs.

Trust

- BERT: It was ranked as the least trusted model (3.57). There was criticism about the accuracy of the reasoning provided in BERT (3.29). There was a lack of description of the variables used for prediction according to the participants.
- LR: It was the most trusted model. Its reliability confidence (4.17) and accuracy of reasoning (4.29) were the highest among the 3 ML models.
- XGBoost: It was ranked as a trustworthy model (4) due to its perceived high accuracy of reasoning (4). However there was criticism that the global explanation was complex, and the reasoning of prediction was hard to understand.

Usefulness:

- BERT: It received the lowest score in providing insights for assisting clinical decision making (3.14).
- LR: It scored moderately well (3.71) in assisting clinical decision-making.
- XGBoost: It tied with BERT for the lowest score in providing clinical insights (3.14).

NASA Task Load Index:

On the NASA Task Load Index, significant differences in scores were found on the parameter of cognitive demand.

- BERT: It was considered as the least mentally demanding model.
- LR: It was considered as a moderately mentally demanding model. There was criticism about the difficulties in understanding its abstract concept.
- XGBoost: It was ranked as the most mentally demanding ML model, requiring more effort to understand.

Discussion

In this study, we aimed to gather clinician's perspectives about the various ML models and explainability techniques used for aiding diagnosis of depression. The expectations from the predictions and explanations, collected initially from the clinicians through an online questionnaire, served as development requirements for us. Further, various ML models and explainability techniques were explored and three were chosen to be presented to the clinicians for further evaluation based on user requirements, dataset characteristics, and performance metrics.

We have employed core concepts of "user-centered design" [31] and "application-grounded evaluation approach"[26-29] by engaging with potential users from the early stages and continued with this approach while evaluating the utility of explainability techniques in clinical settings. The INTRPRT guideline for transparent ML systems in medical image analysis recommends "formative user research" and "empirical user testing" to promote trust and ensure transparency in ML algorithms [32]. Though we did not present any interactive system to the clinicians, we found some core concepts of these design approaches relevant for our enquiry like:

- Specifying how ML made a decision for a data sample from the training set [31]
- Presenting feature ranking for + and - class separately [31]

[1] average scores on the Likert scale are shown in parenthesis.
[2] answers of participants to open ended questions have been paraphrased by the researchers.

- Early user involvement to avoid disproportionate prioritization of performance of ML algorithms over the needs of the clinical task [32]
- Empirical user evaluation [32]

In the first approach for classification, we used LIME for explaining the BERT model as it is model agnostic and locally interpretable [33]. LIME makes black box models locally interpretable by using sparse linear combinations based on data that is already explainable [34]. Besides, it highlights phrases and words that are contributing to the outcome thus providing language patterns as wished for by the users.

For classification by feature extraction method we presented two models to the participants: LR and XGBoost. LR is mathematical, thus it calculates coefficients and odds ratio to establish relationships between the features and outcome. The value of coefficients establishes feature importance globally. For local interpretability, LR calculates a weighted sum of the features to obtain the probability of belonging to the positive class [33, 35]. On the other hand, the prediction made by XGBoost can be difficult to interpret since it is an ensemble learning method. We used SHAP to explain XGBoost predictions as it captures complex feature interactions [33], calculates contribution of each feature to the decision [36], is model agnostic [37] and provides global (SHAP summary plot) and local interpretability (SHAP waterfall plot) [38]. The choice of explainability technique and visualization was based on requirements gathered from participants like feature ranking and probability of predictions.

The participants seemed to be unsatisfied with the accuracy, reliability and usefulness of LIME. A previous study evaluating explainability methods in radiology found LIME inefficient in terms of coherency and trust [39]. Two other studies in image classification have found LIME to be reliable and interpretable [36, 39]. LIME facilitated the participant's understanding of BERT's prediction by highlighting words, however it failed to explain the reasoning, especially since some of the highlighted words did not seem relevant to the context, like "wish", "have" and "could". Since the participants did not seem to trust the model, they also found it the least helpful in clinical decision making, although it had the highest accuracy and F1 score among all the models. Overall, the participants seemed to favour the interpretability provided by LR as the visualisation involves coefficients and correlations which are identical to mathematical concepts. Participants found it clear, reliable, helpful and accurate which also improved its usefulness for them in clinical decision making.

The top four features positively influencing depression diagnosis are the same for LR and SHAP, namely average frequency of adverbs and verbs, speech speed and mean word length. A previous study using ML for classification of depression had found variable feature contribution rankings amongst the various models [23]. Another study identified the same factors influencing prediction in SHAP with XGBoost and regression analysis [40]. Understanding SHAP has placed the highest mental demand and is least helpful in understanding predictions according to the participants, which could be because of the complexity of plots. A previous study exploring explainability in medical imaging concluded that SHAP was not the best for local explanations while another found it efficient for giving detailed feature importance in non imaging datasets which may be making it the most trustworthy in our case [36].

The main limitations of our study are small sample size for user evaluations, non-interactive visualisation of explainability techniques, limited dataset and use of single data type (only text) for the classification task. Further, the results of our study may not be generalizable as we tested three models paired with three different explainability techniques. We chose to evaluate participants through a questionnaire to avoid interviewer bias, but this method did not allow us to ask follow up questions on their open ended answers. We chose the models with the best performance from among the ones we had trained, but we did not use any specific methods to improve model performance other than some hyperparameter adjustments.

As much as we are aware, this is one of the first studies employing user-centered design and application-grounded evaluation approaches to evaluate explainability methods in the diagnosis of depression. We recommend that future research should be directed towards designing interfaces and testing the clinician's interactions with it empirically. We also recommend training ML models with a focus on improving their performance, aligned with their potential application in clinical settings from end users' perspective.

Conclusion

This study explored clinicians' perspectives on the use of explainability techniques in ML models for aiding depression diagnosis, emphasizing the critical role of interpretable and reliable explanations in clinical decision-making. By employing user-centered design and application-grounded evaluation approaches and engaging clinicians early through structured questionnaires, we aligned the development of models and explanations with end-user needs.

Our findings reveal that clinicians favored the interpretability of LR due to its mathematical transparency and alignment with clinical reasoning, making it the most trusted and practical model for decision-making. In contrast, the SHAP-based explanations for XGBoost, while comprehensive, were perceived as overly complex and cognitively demanding. Similarly, LIME's explanations for the BERT model, though locally interpretable, lacked coherence and contextual relevance, resulting in lower trust and utility.

These results underscore the necessity of designing explainability techniques that prioritize clarity, contextual relevance, and ease of use to meet the practical requirements of healthcare professionals. The trade-offs between interpretability, cognitive demand, and trustworthiness highlight the need for tailored solutions that integrate seamlessly into clinical workflows.

Future research should focus on developing interactive interfaces, refining explainability methods to enhance alignment with clinical logic, and improving model performance to build greater trust. By addressing these gaps, IML tools can become invaluable assets in clinical practice, advancing the diagnostic process and ultimately improving patient outcomes.

References

[1] World Health Organization. Depressive disorder (depression) [Internet]. World Health Organization; 2023. [cited 2024 Dec 09] Available from: https://www.who.int/news-room/fact-sheets/detail/depression.

[2] Liu X, Jiang K. Why is Diagnosing MDD Challenging? Shanghai Arch Psychiatry. 2016 Dec 25;28(6):343-345. doi: 10.11919/j.issn.1002-0829.216073.

[3] Bhadra S, Kumar CJ. An insight into diagnosis of depression using machine learning techniques: a systematic review. Curr Med Res Opin. 2022;38(5):749-71.

[4] Byeon H. Advances in Machine Learning and Explainable Artificial Intelligence for Depression Prediction. IJACSA. 2023;14(6):520-6.

[5] Maurer DM, Raymond TJ, Davis BN. Depression: Screening and Diagnosis. Am Fam Physician. 2018;98(8):508-15.

[6] Richter T, Fishbain B, Richter-Levin G, Okon-Singer H. Machine Learning-Based Behavioral Diagnostic Tools for Depression: Advances, Challenges, and Future Directions. J Pers Med. 2021 Sep 26;11(10):957. doi: 10.3390/jpm11100957.

[7] Kerz E, Zanwar S, Qiao Y, Wiechmann D. Toward explainable AI (XAI) for mental health detection based on language behavior. Front Psychiatry. 2023 Dec 7;14:1219479. doi: 10.3389/fpsyt.2023.1219479.

[8] Asma SA, Akhter N, Sharmin S, Rahman MS, Hosen ASMS, Lee OS, et al. Hierarchical Explainable Network for Investigating Depression From Multilingual Textual Data. IEEE Access. 2024;12:131915-27.

[9] Zogan H, Razzak I, Wang XZ, Jameel S, Xu GD. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web-Internet and Web Information Systems. 2022;25(1):281-304.

[10] Kaur I, Kamini, Kaur J, Gagandeep, Singh SP, Gupta U. Enhancing explainability in predicting mental health disorders using human-machine interaction. Multimedia Tools and Applications. 2024.

[11] Hosseinzadeh Kasani P, Lee JE, Park C, Yun CH, Jang JW, Lee SA. Evaluation of nutritional status and clinical depression classification using an explainable machine learning method. Front Nutr. 2023;10:1165854.

[12] Uddin MZ, Dysthe KK, Folstad A, Brandtzaeg PB. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing & Applications. 2022;34(1):721-44.

[13] Nguyen HV, Byeon H. Explainable Deep-Learning-Based Depression Modeling of Elderly Community after COVID-19 Pandemic. Mathematics. 2022;10(23).

[14] Ahmed MS, Ahmed N. A Fast and Minimal System to Identify Depression Using Smartphones: Explainable Machine Learning-Based Approach. JMIR Form Res. 2023;7:e28848.

[15] Bertl M, Bignoumba N, Ross P, Yahia SB, Draheim D. Evaluation of deep learning-based depression detection using medical claims data. Artificial Intelligence in Medicine. 2024;147:102745.

[16] Abdullah TAA, Zahid MSM, Ali W. A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. Symmetry. 2021;13(12):2439.

[17] Regulation (EU) 2024/1689 of the European Parliament and of the Council [Internet]. Brussels: Official Journal of the European Union; 2024 [cited 2024 Dec 09]. Available from: http://data.europa.eu/eli/reg/2024/1689/oj.

[18] Al-Ansari N, Al-Thani D, Al-Mansoori RS. User-Centered Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review. Human Behavior and Emerging Technolo-gies. 2024;2024:1-28. Available from: https://doi.org/10.1155/2024/4628855.

[19] Almeida F. Strategies to perform a mixed methods study. Eur J Educ Stud. 2018;5(1):137. doi: 10.5281/zenodo.1406214.

[20] Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, et al. The Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) [Internet]. Los Angeles: University of Southern California Institute for Creative Technologies; 2014 [cited 2025 Jan 8]. Available from: https://dcapswoz.ict.usc.edu/

[21] USC Institute for Creative Technologies. DAIC-WOZ Database [Internet]. Los Angeles: University of Southern California; [cited 2024 Dec 13]. Available from: https://dcapswoz.ict.usc.edu/daic-woz-database-download/

[22] Rude S, Gortner EM, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cogn Emotion. 2004;18(8):1121-33.

[23] Lorenzoni G, Tavares C, Nascimento N, Alencar P, Cowan D. Assessing ML classification algorithms and NLP techniques for depression detection: An experimental case study. arXiv [preprint]. 2024 [cited 2024 Dec 20]. Available from: http://arxiv.org/abs/2404.04284.

[24] Databricks. Machine Learning Models [Internet]. Databricks, Inc.; [cited 2024 Dec 13]. Available from: https://www.databricks.com/glossary/machine-learning-models

[25] Novikova J, Shkaruta K. DECK: Behavioral tests to improve interpretability and generalizability of BERT models detecting depression from text. arXiv [preprint]. 2022 Sep 10 [cited 2024 Dec 13];arXiv:2209.05286.

[26] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning [Internet]. arXiv; 2017 [cited 2024 Nov 25]. Available from: http://arxiv.org/abs/1702.08608

[27] Ochella S, Shafiee M. Performance metrics for artificial intelligence (AI) algorithms adopted in prognostics and health management (PHM) of mechanical systems. J Phys Conf Ser. 2021;1828(1):012005. doi: 10.1088/1742-6596/1828/1/012005.

[28] Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems [Internet]. arXiv; 2020 [cited 2024 Nov 26]. Available from: http://arxiv.org/abs/1811.11839.

[29] Nii Laryeafio M, Ogbewe OC. Ethical consideration dilemma: systematic review of ethics in qualitative data collection through interviews. J Ethics Entrep Technol. 2023;3(2):94-110. doi: 10.1108/JEET-09-2022-0014.

[30] Schembri N, Jahić Jašić A. Ethical issues in multilingual research situations: a focus on interview-based research. Res Ethics. 2022;18(3):210-25. doi: 10.1177/17470161221085857.

[31] Petkovic D, Altman R, Wong M, Vigil A. Improving the explainability of Random Forest classifier - user centered approach. Pac Symp Biocomput. 2018;23:204-15. PMID: 29218882

[32] Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med. 2022 Oct 19;5(1):156. doi: 10.1038/s41746-022-00699-2. PMID: 36261476; PMCID: PMC9581990.

[33] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. arXiv.2016. arXiv:1602.04938. Available from: https://arxiv.org/pdf/1602.04938.

[34] Choi B, Shim G, Jeong B, Jo S. Data-driven analysis using multiple self-report questionnaires to identify college students at high risk of depressive disorder. Sci Rep. 2020;10(1):7867.

[35] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. 1st ed. New York: Springer; 2013.

[36] Lundberg, S., and S. Lee. A Unified Approach to Interpreting Model Predictions: 31st Conference on Neural Information Processing Systems Advances in neural information processing systems. 2017;30.

[37] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610; 2019.

[38] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Aug; p. 785–94.

[39] Sun J, Chakraborti T, Noble JA. A Comparative Study of Explainer Modules Applied to Automated Skin Lesion Classification. In: XI-ML@ KI; 2020.

[40] Jin Y, Xu S, Shao Z, Luo X, Wang Y, Yu Y, et al. Discovery of depression-associated factors among childhood trauma victims from a large sample size: Using machine learning and network analysis. J Affect Disord. 2024;345:300–10.

Author's contributions

- Data preprocessing: all.
- Predictive models and performance: Ching, Shweta, Katja.
- Explainability methods: Joao, Shweta, Katja.
- Questionnaire design and data collecting: Guilherme, Shweta.
- Introduction, methods, and results in the report: Katja, Joao, Guilherme, Ching.
- Discussion and conclusions in the report: Shweta, Katja.
- Revision of the entire report: all

Appendix

Tables

Table A1: Demographic Characteristics of Participants

Characteristics	Distribution
Participants, number	7
Age (years), mean	44
Gender, number	Males: 1 Females: 6
Professional Qualification, number	General Practitioners: 1 Psychiatrists: 1 Psychologists: 5
Geographical location, number	Australia: 1 India: 2 Brazil: 4

Table A2: Models with hyperparameters, features and performance metrics

Model	Hyperparameters	Features	Accuracy	F1 Score
BERT	{'k_folds': 5, 'seed': 6, 'model_name':'bert-base-uncased', 'num_labels': 2, 'batch_size': 8, 'num_epochs': 20, 'learning_rate': 1e-5, 'optimizer': 'Adam' , 'loss_function': 'Sparse CategoricalCrossentropy', 'early_stop_monitor': 'val_loss', 'early_stop_patience': 5, 'f1_average': 'weighted'}	NA	0.92	0.91
LR	{ 'class_weight': balanced}	engineered features*	0.59	0.47
XGBoost	{'learning_rate': 0.2, 'max_depth': 2, 'n_estimators': 200}	engineered features*	0.67	0.37

* features used for LR and XGBoost: mean_word_length, flair_sentiment, average_frequency_nouns, average_frequency_verbs, average_frequency_adjectives, average_frequency_adverbs, average_frequency_firstperson, speech_speed, unique_word_count, stop_word_frequency

Table A3: Features and their descriptions

Feature	Explanation
Mean_word_length	Average number of characters in words within a speech sample. Depression can manifest in simpler, less varied language use.
Average_frequency_nouns	Average occurrence of nouns in the responses.
Average_frequency_verbs	Average occurrence of verbs in the responses.
Average_frequency_adjectives	Average occurrence of adjectives in the responses.
Average_frequency_adverbs	Average occurrence adverbs in the responses.
Average_frequency_firstperson	Average frequency of first-person pronouns (like "I," "we," "us", etc) across all responses.
Speech_speed	Calculated by taking the word count for each response, dividing it by the time taken to respond, and then averaging this value across all responses in the conversation.
Unique_word_count	Calculated by determining the number of unique (non-repeated) words in each response, dividing it by the total word count in that response, and then averaging this value across all responses.
Stop_word_count	Counts the frequency of stop words in each response (e.g. and, to, a, the).
Flair_sentiment	Average sentiment across all responses for one patient. Sentiment refers to the emotional tone or attitude expressed in a piece of text.
Flair_score*	A sentiment analysis score that classifies text as positive or negative along with associated probabilities.
Frequency_of_speech*	Counts the number of times a person says something in an interview. Can capture interaction intensity and verbosity. Lower engagement, reflected by less frequent speech, is often linked to depressive states.

* excluded for the final evaluated models

Figures depicting the visualizations provided by the various explainability techniques:

1. BERT:

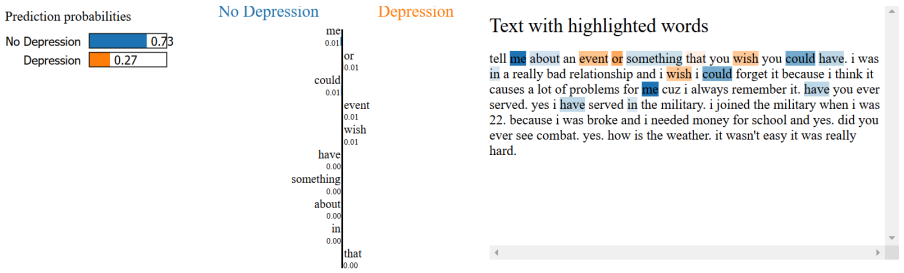


Figure A1- LIME (local explainability) for case ID 641 (depression)

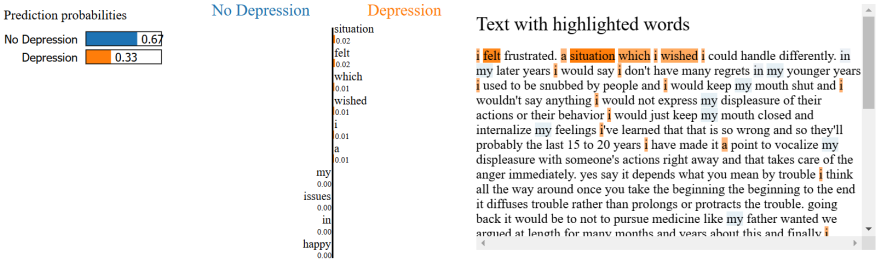


Figure A2- LIME (local explainability) for case ID 471 (no depression)

2. LR:

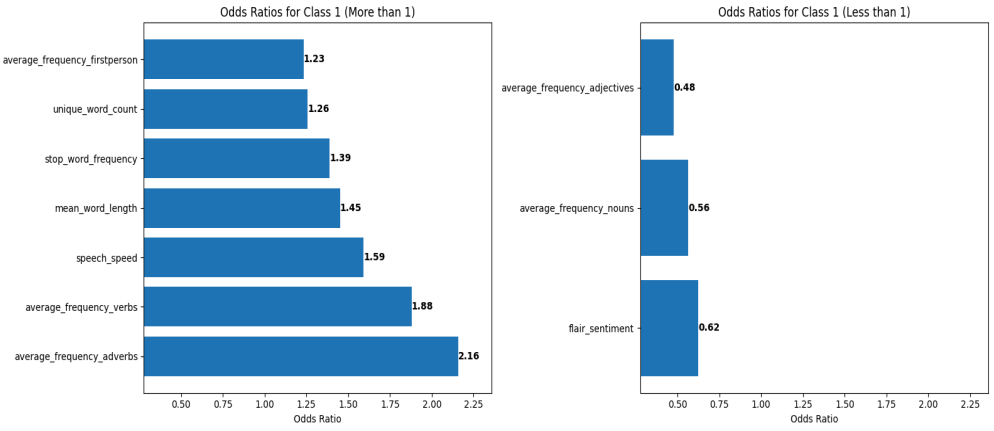


Figure A3- Odds ratios of features (global explainability) for class 1 (depression)

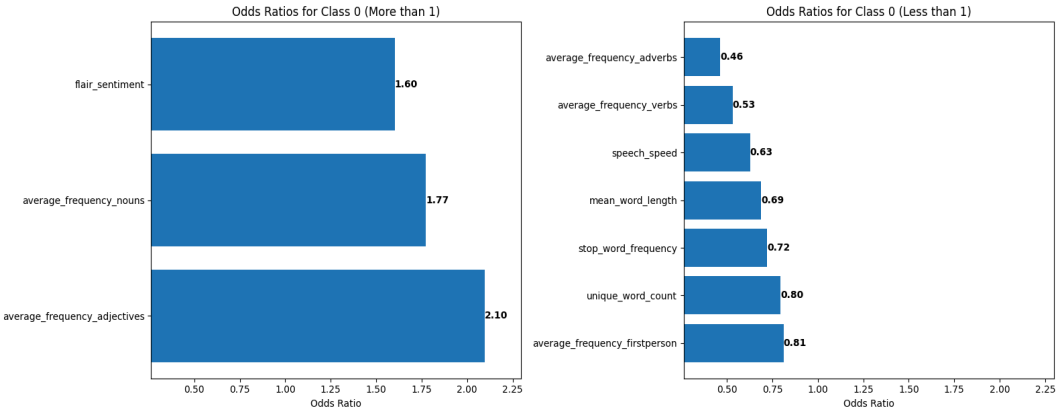


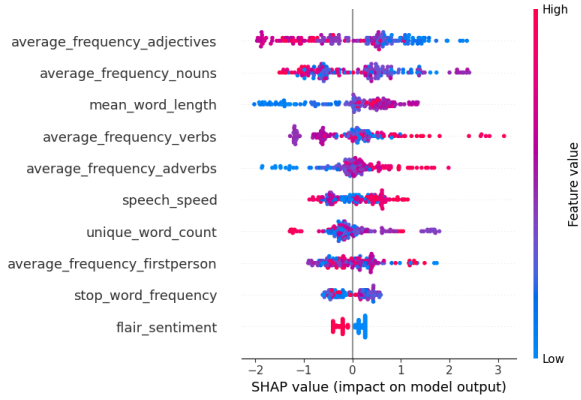
Figure A4 - Odds ratios of features (global explainability) for class 0 (no depression)

*LR presents local explainability as a probability value, which represents how likely it is that a given data point belongs to the positive class (class 1)

For Participant ID 471 (No depression or class 0 by label), predicted probability (p): [[0.34955431]]

For Participant ID 641 (depression or class 1 by label), predicted probability (p): [[0.52261688]]

3. XGBoost



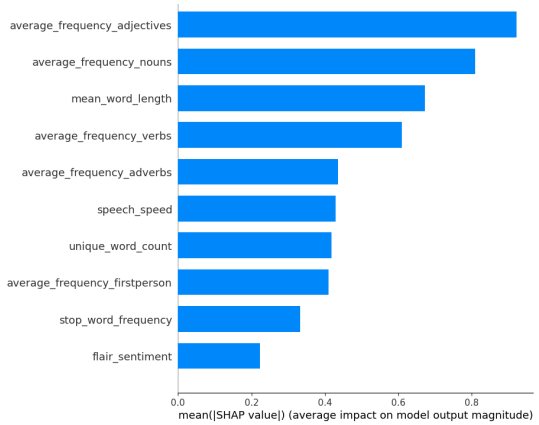


Figure A5- SHAP summary plot (global interpretability) *Figure A6- SHAP feature importance (global interpretability)*

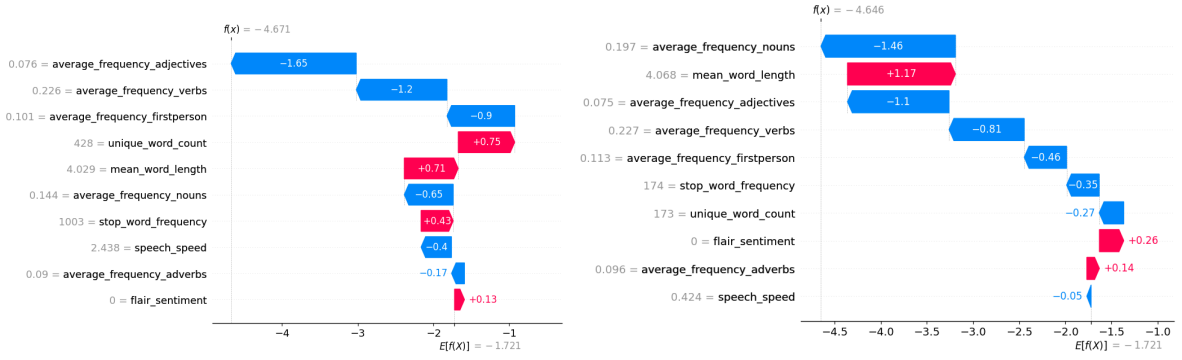


Figure A7- SHAP waterfall plot (local interpretability), left: case ID 641 (depression), right: case ID 471 (no depression)

Links:

Link 1- Link to the questionnaire used in the preliminary interviews to gather user expectations:
https://olqnpvjg.forms.app/usersystemrequirements

Link 2- Link to the evaluation questionnaire used to assess explainability methods in XAI:
https://olqnpvjg.forms.app/questionnaire-evaluation-iml

Link 3 - Link to the GitHub repository that contains all the code used for the project:

https://github.com/sochinghong/SU_project.git