Fully Interpretable and Adjustable Model for Depression Diagnosis: A Qualitative Approach

Author 1, Author 2
Affiliation 1
Address Line 1.1
Address Line 1.2

Author 3
Affiliation 3
Address Line 3.1
Address Line 3.2

Abstract

Recent advances in machine learning (ML) have enabled AI applications in mental disorder diagnosis, but many methods remain black-box or rely on post-hoc explanations which are not straightforward or actionable for mental health practitioners. Meanwhile, interpretable methods, such as k-nearest neighbors (k-NN) classification, struggle with complex or high-dimensional data. Moreover, there is a lack of study on users' real experience with interpretable AI. This study demonstrates a network-based k-NN model (NN-kNN) that combines the interpretability with the predictive power of neural networks. The model prediction can be fully explained in terms of activated features and neighboring cases. We experimented with the model to predict the risks of depression and interviewed practitioners in a qualitative study. The feedback of the practitioners emphasized the model's adaptability, integration of clinical expertise, and transparency in the diagnostic process, highlighting its potential to ethically improve the diagnostic precision and confidence of the practitioner.

Introduction

The booming era of Artificial Intelligence sees a rise of its application to mental health-related issues (Graham et al., 2019) and specifically to mental disorder diagnosis ("diagnosis" for short) (Bzdok and Meyer-Lindenberg, 2018; Chattopadhyay, 2017; Graham et al., 2019; Iyortsuun et al., 2023). Although various AI technologies have achieved high accuracies in diagnosis, they generally lack explanation for their decision making, and the process remains a black box for both mental health practitioners ("practitioners" for short) and clients seeking counseling services ("clients" for short) (Jarvie and Lindén, 2024; Lau, Zhu, and Chan, 2023). Therefore, there is still much caution and concern about the use of AI for diagnosis (Kerz et al., 2023; Li et al., 2024).

This gives rise to recent trends in using Explainable AI (XAI) technologies for mental health: Practitioners and clients may rely on AI tools "to the extent they can economise on human oversight, monitoring and verification of the system's outputs" (Joyce et al., 2023). Most XAI methods are based on post-hoc explanation methods such as SHAP (Shapley, 1953) and LIME (Ribeiro, Singh, and Guestrin, 2016). However, Rudin (2019) argues that post-hoc explanations are often inadequate for black-box models. The explanations may justify the model's decision after the decision is made, but not how the model reached the decision internally. In the worst cases, post-hoc explanations are excuses for a model's mistakes and offer no opportunity to debug or fine-tune a trained black-box model.

To address the interpretability and adjustability of AI models that facilitate practitioners in diagnosis, we propose to use a recently invented model, a neural network based k-nearest-neighbor algorithm (NN-kNN). As a k-nearest neighbor algorithm, NN-kNN can explain each model decision with activated cases, and each activated case can be attributed to its feature distances with the query. As a neural network, NN-kNN allows end-to-end training for both feature weights and case weights, and is compatible with other neural network methods (Ye et al., 2024).

Our study introduces a novel approach to human-machine interaction drawing on insights and methodology from both AI and psychology. Specifically, we study how NN-kNN facilitates practitioners in aspects beyond traditional diagnosis, including feature exploration, explanation of decisions, past case retrieval, and manual weight adjustment. We conducted a qualitative study, using interpretative phenomenological methods to capture the insights of the practitioner's experience with the interpretable and adjustable model. Although qualitative research is common in the field of psychology, incorporating this method in AI research offers a novel perspective and enriches the understanding of human interaction with ML models.

Related Work

Explainable and Interpretable AI

According to the survey by Joyce et al. (2023), the majority of XAI methods in mental health are based on feature importance methods such as SHAP (Shapley, 1953). Feature importance methods estimate the weights of features by measuring the influence on the model's output after perturbing a feature. Only two methods in the survey are regression-based and interpretable by design. Most studies in the wider field of XAI are post-hoc methods because they are easily

applicable to different models (Saleem et al., 2022). Post-hoc methods build an interpretable model that mimics the behavior of a black-box model and use this new model as an explanation. However, post-hoc explanations are problematic for high-stakes decision making in mental health applications: (1) post-hoc explanations may not faithfully represent the original model's computations, (2) they may lack the detail necessary to fully understand the black-box model, and (3) they do not permit manual calibration by domain experts (Rudin, 2019).

Our XAI method is interpretable by design and explains decisions using features and cases. This is similar to the explanation done in case-based reasoning (CBR) (Schoenborn et al., 2021; Gates and Leake, 2021), where decisions on queries are explained by past cases similar to queries. Most CBR systems weight features (Wettschereck, Aha, and Mohri, 1997), while fewer weight cases (Bicego and Loog, 2016). NN-kNN takes the extra step to train both feature weights and case weights at the same time in an end-to-end manner. Additionally, the model allows manual tweaking of weights by experts.

Memory Augmented Networks

Weston, Chopra, and Bordes (2014) proposed the class of memory networks, where a neural network integrates an external knowledge base in its processing. Matching networks (Vinyals et al., 2016) extend the memory networks for one-shot learning. The authors do so by incorporating characteristics from non-parametric models, allowing a trained network to be directly used on a new support set. Similarly, prototypical networks (Snell, Swersky, and Zemel, 2017) learn a metric space and perform classification by computing the distances between the query and prototypes of each class. Li et al. (2018) propose a neural network model that stores auto-encoded embeddings of learned prototypes and makes prediction by comparing the query embedding with the prototype embeddings.

Mental Health Diagnosis by Practitioners

The development of efficient diagnosis has stagnated for decades, facing several challenges:

1. **Time-Consuming Diagnostic Processes:** Clinicians continue to rely on diagnostic manuals such as the DSM-V and ICD-10. The time-consuming diagnostic process takes away valuable time from direct therapeutic interventions (Perkins et al., 2018)
2. **Insufficient Training Opportunities:** Only 23% of American Psychological Association (APA)-accredited doctoral programs provide trainee clinicians with the training sites necessary for structured diagnostic interviews (Mihura, Roy, and Graceffo, 2017).
3. **Inadequacy of Diagnostic Manuals:** A systematic review involving 2,228 participants found that clinicians often view diagnostic manuals as unhelpful due to incomplete or inaccurate symptom descriptors (Perkins et al., 2018). They tend to focus on categorizing symptoms rather than identifying the underlying formations of psychological symptoms.
4. **High Rates of Misdiagnosis:** In a study of 309 psychiatric patients, 39.16% patients with severe psychiatric disorders were misdiagnosed, with schizoaffective disorder having the highest misdiagnosis rate (75%), followed by major depressive disorder (54.72%) (Ayano et al., 2021).
5. **Comorbidity and Diagnostic Complexity:** Mental health diagnoses are further complicated by comorbid conditions. For instance, individuals with autism spectrum conditions are more likely to be diagnosed with mental health disorders than non-autistic individuals, increasing the potential for misdiagnosis (Au-Yeung et al., 2019).

Given these ongoing challenges, the field of health service psychology urgently requires a more effective and accurate diagnostic system to support clinicians in making better informed decisions.

Mental Health Diagnosis by AI

Many studies have demonstrated the efficacy of AI-enhanced tools in supporting clinical diagnosis Graham et al. (2019); Lau, Zhu, and Chan (2023). For example, Zhang et al. (2022) conducted a comprehensive narrative review of 399 studies on NLP applications in mental illness detection, suggesting that deep learning approaches show better performance than traditional ML methods.

Despite the great promise of AI-assisted clinical diagnosis, both patients and clinicians remain hesitant to accept its application. In one study, patients have expressed a preference for AI tools that are tailored, person-centered, and adaptable to their individual treatment plans and expressed concern that they must adapt to technology (Li et al., 2024). In addition, clinicians often hesitate to fully integrate AI technology into their practice due to concern about their understanding of how AI systems work (Kerz et al., 2023). As such, there is a need for transparency and explainability in AI systems to foster trust with mental health practitioners. However, among the mental health XAI projects surveyed by Joyce et al. (2023), none interviewed specialists about their experience with the XAI models. Surveys in the broader field of XAI (Saeed and Omlin, 2023; Das and Rad, 2020; Molnar, Casalicchio, and Bischl, 2020) also identified evaluating the explainability of XAI models as a major challenge. In general, there is no agreed upon measure of explainability/interpretability.

In this study, we embrace the idea that explainability is not a rigorous formal concept and that an explanation is only as good as the intended audience perceives it. We evaluate explainability in a human-centered qualitative approach that is more common in social science.

Neural Network Based K-Nearest Neighbors

K-Nearest Neighbors Classifiers

K-nearest neighbors classifiers (k-NN) has been used for mental health diagnosis (Elmunsyah et al., 2019; Chahar, Dubey, and Narang, 2024; Banerjee et al., 2024). The task domain involves cases C in the form of (x, L_x). For each case $x \in C$, $f(x) =< x_1, x_2, ..., x_m >$ is a feature vector describing a client's information (survey answers, medical record information, etc.) and L_x is the diagnosis label

associated with the client (risk of depression). The function f might describe a feature extraction method or simply use the surface features of x. A naive k-NN calculates the distance between two cases as the sum of Minkowski distances between corresponding features (Dasarathy, 1991). Better k-NN methods use feature weights in the calculation of distance. Traditional methods use a global feature weighting, while others allow certain sets of cases (or even each individual case) to have their own feature weighting (Manzali et al., 2024; Aha and Goldstone, 1992; Friedman, 1994; Ricci and Avesani, 1995; Marchiori, 2013; Bonzano, Cunningham, and Smyth, 1997). Some methods assign case weights, so that certain cases contribute more in the voting of the final prediction (Bicego and Loog, 2016; Aguilera et al., 2019). The case weights may be based on the distance between the case and the query, or it may be a learned parameter of the case. Other methods such as neighborhood components analysis (NCA) and large margin nearest neighbor (LMNN) transform the feature space to extract high-level features before distance calculation (Goldberger et al., 2004; Weinberger and Saul, 2009).

Neural Network Based K-Nearest Neighbors algorithm

The neural network based k-nearest neighbors algorithm (NN-kNN) implements both feature weights and case weights by having following network layers that simulate the behavior of a k-NN:

1. The case layer stores the query q and all cases (each case is denoted as x).

$$f(q) =< q_1, q_2, ..., q_m >, f(x) =< x_1, x_2, ..., x_m > \quad (1)$$

2. The feature distance layer calculates the distance between the corresponding features as

$$\delta_i = \delta_i(q_i, x_i) \geq 0 \quad (2)$$

We choose the squared distance $\delta_i(q_i, x_i) = (q_i - x_i)^2$.

3. The case activation layer activates a case x given the query q by

$$case\_act(x|q) = \sigma(w_{x\delta_1} * \delta_1 + w_{x\delta_2} * \delta_2 + ... + b_x) \quad (3)$$

where $w_{x\delta_i}$ is the weighting of feature i and $b_x \geq 0$ is the default activation for case x. $w_{x\delta_i}$ can only be negative. We choose the sigmoid function $\sigma()$ as the activation function to limit $case\_act$ within $[0, 1]$.

4. The top k case selection layer is optional. When enabled, it will keep the top k case activations and resets other case activations to 0.

5. The target activation layer takes each case's activation to activate the corresponding class.

$$class\_act(L|q) = \Sigma_x(w_{(x,L)} * case\_act(x)) + b_L \quad (4)$$

where $w_{(x,L)}$ is the weight of the case x and b_L is the bias of the class L. $w_{(x,L)}$ is forced to be positive (by using a ReLu) if the case x is of class L and $w_{(x,L)} = 0$ otherwise, because a case should only activate its corresponding class but not other classes.

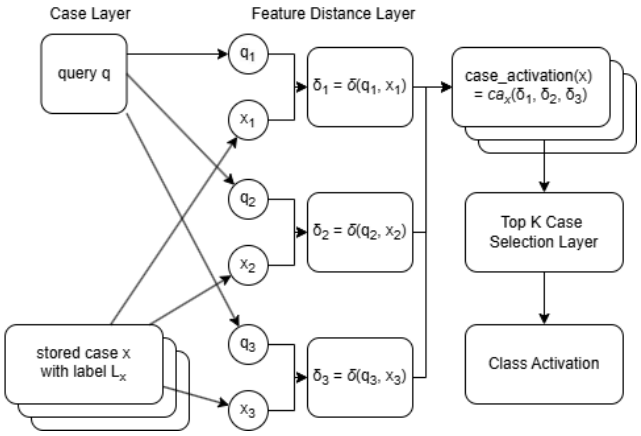

Figure 1: The Model of NN-kNN

6. The adaptation layer chooses the label L with the maximum $class\_act$.

The layers are depicted in Figure 1. NN-kNN is similar to the matching networks and prototypical networks because the layers until the target activation layer serve as a similarity metric for the cases. In fact, NN-kNN can be considered as a generalization of memory networks. It works with any case (not just prototypes) or any feature (not just embeddings extracted by an autoencoder). Its k-NN nature allows easy insertion and deletion of cases and easy explanation through activated cases and features.

Qualitative Interview and Evaluation

In previous experiments comparing NN-kNN with neural networks and state-of-the-art k-NN methods (such as LMNN), NN-kNN achieves equal or less prediction error in classification and regression on multiple datasets (Ye et al., 2024). This study focuses on the interpretability and adaptability of NN-kNN to aid practitioners in mental disorder diagnosis. We trained NN-kNN on a dataset on the prediction of depression risk and conducted a qualitative study interviewing 10 licensed practitioners about their experience with the model.

The Dataset

The model is trained on the dataset from Orozco-del Castillo et al. (2021). The original dataset contains responses to 117 true/false survey items. The survey was designed for depression screening according to the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2013) and answered by 157 undergraduate students. After data preprocessing, our final dataset contains 117 cases; Each case has 102 binary features and a class label of 0, 1 or 2, representing low, medium and high risk.

In these experiments, NN-kNN achieves an average accuracy of 0.646, outperforming both standard k-NN (0.417) and LMNN (0.492). We urge caution due to the small dataset size. We did not test extensively on multiple datasets against various models, as previous research (Ye et al., 2024) has

already conducted such comparisons. The primary focus of this study is on the interpretability and adjustability of the model for practical use by clinicians.

The Interpretability and Adjustability for Practitioners

NN-kNN is both interpretable and transparent according to the Transparency and Interpretability for Understandability framework (Joyce et al., 2023). For predicting depression risk, the model's design emphasizes the following aspects:

- All cases share the same feature weights ($w_{x\delta_i} = w_{y\delta_i} = w_{\delta_i}$ for any two cases x and y), reducing the overall number of parameters. This global feature weighting approach assigns a single weight to each feature across all cases.
- Initial settings for each case x include feature weights set to $w_{x\delta_i} = 1$, case bias at $b_x = 50$, case weight at $w_{(x,L)} = 1$, and class biases at $b_L = 1$. In this configuration, all features and cases start with equal importance. Through training, the model adjusts these weights to improve prediction accuracy.
- The calculations involved in feature distances, case activations, and class activations rely on basic operations like subtractions and summations. Practitioners can easily understand the model's operations.
- Each parameter in the model has an interpretable role. The feature weight $w_{x\delta_i}$ and the case weight $w_{(x,L)}$ respectively reflect the relevance of a feature/case to depression. The case bias b_x and the class bias b_L respectively indicate the inherent importance of a case and a class. Practitioners can identify outliers or particularly influential features/cases.
- Because each parameter has a specific semantic meaning and a preset initial value, practitioners can manually adjust the feature or case weights based on their expert knowledge. After making adjustments, the model can retrain to further refine the predictions. Practitioners can then assess whether the updated model aligns better with their clinical expertise. We demonstrated this functionality to practitioners and collected their feedback.

Qualitative Interview Design

The effectiveness of an explanation ultimately depends on how well it is understood and perceived by its intended audience. We conducted interviews with 10 U.S.-based licensed clinicians via the interpretative phenomenological analysis (IPA) approach to explore their experiences and perceptions. (Eatough and Smith, 2017). While quantitative methods are prevalent in XAI research—often involving comparisons of models across datasets and using metrics such as prediction accuracy—we chose a qualitative approach for several reasons:

- **Focusing on a specialized population**: Our target audience consists of licensed clinicians, a small and specialized group. A qualitative approach allows us to zoom in on this specific population, gaining in-depth insights into their experiences with the model, which could be missed in a broader, quantitative study.
- **Tailored model demonstrations**: Demonstrating the model to potential users individually allows us to observe first-hand how each clinician interacts with the system. This personalized approach helps capture the intricacies of their responses, including any challenges they face or specific features they find most valuable.
- **Understanding nuanced reactions**: AI models can evoke a range of responses, particularly when introduced in sensitive fields such as mental health. A qualitative approach is well suited to capturing these nuanced reactions, such as concerns about integrating AI into clinical practice, the potential for AI to aid in diagnostic processes, or hesitations about interpretability. This method allows for a deeper understanding of users' trust and confidence in the model (Maxwell, 2021).
- **Building trust and bridging theory to practice**: By engaging directly with clinicians through interviews, we can address their concerns, answer questions, and foster trust in the model's practical use. This is a crucial step in translating theoretical AI advancements into real-world applications for mental health practitioners.

The participants included four doctoral-level licensed psychologists and six master-level licensed clinicians (four men and six women). Each of them was shown the model via a Jupyter notebook over Zoom, adjusted the parameters based on their clinical judgement for depression, and spent 30 minutes answering qualitative questions. Sample questions include: *"did you feel like the model has become more useful clinically after you tuned the feature weight?" and "since the model can detect bias, if the model's explanations differ from your clinical judgements, what would you do?".*

Data Analysis through Interpretative Phenomenological Method

Following the IPA's four-step analysis, three team members (a licensed counseling psychologist, a doctoral candidate in counseling psychology, and an undergraduate psychology student) began by checking each other's biases related to AI and clinical diagnosis. Each member independently reviewed the interview transcripts, annotating their initial reactions. These annotations were translated into experiential statements, summarizing key aspects of each participant's experience. Next, we compared our individual statements to resolve any discrepancies and clustered them to generate overarching themes. Finally, we compiled eight themes that broadly reflected most participants' experiences. Before finalizing the results, the third author acted as an auditor, reviewing that the themes were grounded in the original transcripts and participants' experiences.

Interview Findings

Our data analysis generated eight themes, endorsed by most (at least 6) participants. Following the standard way of reporting results for IPA studies (Liu et al., 2020), we combined the themes with our interpretations of participants' experiences to demonstrate a contextual exploration of the perceptions of clinicians. To ensure confidentiality, all names used in the following sections are pseudonyms.

Themes Identified and Discussions

Theme I: Building Trust and Precision in Diagnosis Participants initially perceived AI-generated diagnoses as inaccurate and unusable but changed their views after engaging with the adjustable model. They appreciated its ability to tune feature weights, adjust cases, and focus on clinically relevant aspects for more precise diagnoses. (Dr. Nate):

> "The tuning process did help somewhat in clarifying the approach to diagnosing by adjusting the feature weights. It allowed for a more targeted focus on clinically relevant aspects."

Participants highlighted the importance of transparency in the model's diagnostic process to enhance clinicians' confidence in using AI-generated results (Dr. Yun):

> "Given that the model is already explainable, and I can see the whole process regarding coming to the diagnosis. It feels more comfortable to understand how it comes to the conclusion. If I want to make some changes to certain features, I can also do that. This gives me more confidence in terms of using it in clinical situations."

Besides, participants observed the adjustment process increased accuracy of depression diagnosis:

> "A general screening can sometimes overlook important clinical features of a client. By being able to adjust the weights on specific features that impact the client's symptoms, I believe the screening becomes more clinically accurate and personalized to the individual, which enhances the overall assessment and treatment planning process."

Theme II: Potential Risk of Bias Introduced by Model Adjustment Participants identified the challenge of maintaining a balance between flexibility and accuracy, expressing concern that their adjustments might lead to less accurate diagnoses (Xing):

> "I am worried about the weight of some questions as I was tuning, which are not the signature of depression... I didn't not quite understand the decision making process of the AI because I would disagree with its clinical decision."

Participants expressed concerns about the risk of bias that clinicians might unintentionally adjust the model to confirm their pre-existing beliefs (Dr. Yong).:

> "The ability to tune the model increases the risk of introducing bias or misusing the tool, especially if there isn't enough transparency about how tuning decisions are made and under what circumstances. Without clear guidelines and a full understanding of the impact of tuning, the potential for bias could compromise trust rather than enhance it."

Theme III: Customization for Clinical Expertise and Multicultural Factors Several participants appreciated the ability to adjust feature weights. They emphasized the model's flexibility to accommodate their clinical experiences and theoretical preferences regarding diagnoses (Dr. Nate)

> "The tunable feature allows for more tailored diagnostic criteria, which can be useful in specific cases, settings, and with particular clients. It offers flexibility, helping clinicians adjust for individual circumstances."

Many clinicians highlighted the model's capacity to incorporate multicultural factors in order to make adjustments for diverse clients and cultural nuances in diagnosis (Dr. Yun)

> "After I tuned the feature weights, the AI model will run again based on my input and generate new features/cases that meets the standards, which is very intelligent. Additionally, regarding the multicultural consideration, clinicians can also tune the features related to clients from diversity background. This is pretty cool."

Furthermore, Dr. Yalin expressed that the model feels like a reliable assistant or skilled trainee, who undergoes personalized training from the clinician to ultimately save time and double-check clinical decisions:

> "I felt like I was training a reliable trainee that, once training completed, can be a great time saver and double check my clinical judgement.

Theme IV: Transparency and Ethical Trustworthiness Participants reported transparency and ethical trustworthiness as key to their willingness to use the model in practice. Additionally, they noted this model's enhanced transparency as a distinguishing feature compared to other AI technologies (Lina and Yaya).:

> "It can be ethically reliable because it ensures informed consent from clients and provides transparency in how it influences counselors' decision-making."
>
> "Additionally, transparency in how AI generates recommendations can further build trust. Ultimately, while trust varies among clinicians, the adaptability of tunable AI can significantly improve its integration into patient care."

Nonetheless, clinicians raised ethical concerns about potential misuse, particularly in the absence of clear clinical or technical guidelines. Dr. Yong expressed both concerns about the model's ethical trustworthiness despite its advancements and confusions on AI's role in replacing human judgment.

> "I think it would be helpful to state that the model is not aiming to take away the human factors when introducing the algorithm. There were moments I got confused from last and this time that the algorithm is trying to take over the human factors in the clinical decision making."

Theme V: Differential Diagnosis and Time Efficiency Participants noted unique strengths of this AI model. For example, they appreciated its ability to aid differential diagnoses and save time clinically. Also, participants were initially skeptical due to frustrating experiences with generative AI like ChatGPT, they were impressed by the model's transparency, which prompted deeper reflections on its potential for clinical diagnosis (Dr. Yun).

> "Another feature could be helpful is the differential diagnosis, such as MDD/ PPD. If AI can show why it's MDD instead of PDD, that it will certainly help the clinician to spend less time in terms of making diagnosis."

Dr. Yalin disclosed that the traceability and documentation of the model's steps make it a valuable resource for complex decision-making, especially around ethical issues and when dealing with differential diagnoses:

Other participants believed that with refinement, the model has the potential to simplify complex tasks and

improve diagnostic efficiency, highlighting the excitement around using AI to streamline clinical processes:

> "The potential for simplifying complex tasks and improving diagnostic efficiency is exciting...I look forward to seeing how it develops further and hope that with some improvements, it can greatly assist clinicians."

Theme VI: Psychotherapy Insights for Future Directions Along with the excitement of using our AI model for differential diagnosis, some clinicians offered expectations of its future clinical utility. Participants noted that tracking mood symptoms over time provides valuable insights so that clinicians is able to adjust treatment plans based on evolving patient data (Lina).

> "Also, it is beneficial for tracking mood symptoms over time with more thorough contextual insights. I would review the model's explanations and compare them to my clinical insights. Ongoing psychotherapy will allow counselors to gain a deeper understanding of their clients, enabling them to enhance their clinical judgment and make more informed decisions moving forward."

Other participants thought that AI could enhance diagnostic reasoning by flagging symptoms that don't fit a particular diagnosis. Dr. Yun noted that AI can not only justify a diagnosis but also explain why it rules out others: Last but not least, one participant (Xing) suggest the model to capture key mental health factors, like affect and history, and summarize the client's mental health background:

> "I think there are so many important factors that the model is not detecting yet. eg. affects, history of mental health. I would also be curious of a feature that could summarize client's mental health history."

Theme VII: Peer Consultation and Training for Clinicians Some participants discussed the potential utility of the model for peer consultation and clinical training. Clinicians began to envision how they might incorporate the model into their decision-making processes, suggesting a growing willingness to see themselves as potential users.

Xing mentioned that if the model's output differed from their judgment, they would seek guidance from peers or supervisors, acknowledging the importance of human expertise in complex clinical cases:

> "I am worried about the weight of some questions which are not the signature of depression, but it also adds in personal bias to the tool. I would seek peers/supervisors for a second opinion when the model differs from my clinical judgement."

Lina discussed that effective training on model tuning is crucial to avoid bias and ensure the model's appropriate use. Ongoing guidance would help clinicians incorporate evidence-based practices into their work:

> "Will there be more guidance on using this tool to enhance clinical judgment or incorporate evidence-based criteria? Counselors' biases can be identified by comparing similar clients' depression symptoms and analyzing their similarities and differences."

Theme VIII: Varied Potentials in the Future When considering the future of the model, participants expressed a mix of concerns and excitement. Two participants discussed the importance of having an user-friendly and visual interface to enhance the user experience, allowing clinicians to see the impact of weight changes clearly. This would improve engagement and the overall utility of the model:

> "I would like to make this model more user friendly, visual, and show more clear changes/interpretation after each feature weights change."
>
> "Ensuring that the tool remains user-friendly and offers high fidelity in its outputs is crucial for its future success. I look forward to seeing how it develops further and hope that with some improvements, it can greatly assist clinicians."

Wang raised concerns about if allowing all users to adjust feature weights might compromise accuracy. Other participants echoed this worry that excessive manipulation could overlook important peripheral data leading to potential inaccurate diagnoses.

> "I am not sure what's the usage of this model. It's like allowing every clinician to change the algorithm, then how is that applicable to large amount of clients?"
>
> "Over-manipulation by the clinician that may overlook the importance of some peripheral data. It's adjustable but not sure how accurate the result is."

It is noteworthy that observed differences between doctorate- and master-level clinicians during interviews. Doctorate-level clinicians were more proactive, asking clarifying questions, creatively adjusting features, and exploring integration into their work. In contrast, master-level clinicians were more defensive, raising concerns that led to shorter and less detailed interviews. These differences may stem from doctorate clinicians' broader roles and theoretical training, while master-level clinicians, burdened by heavy caseloads and limited support, prioritized practical applications and struggled with the prototype's unpolished interface.

Conclusion

We propose NN-kNN for mental disorder diagnosis not as a solution to the inherent challenges of diagnosis, but as a tool to enhance understanding and transparency in the diagnostic process. All models, including NN-kNN, are biased because they are at most as effective as the data they are trained on. What sets NN-kNN apart is its full interpretability and adjustability, enabling practitioners to detect and correct biases within both the model and the data. By offering the ability to adjust model parameters based on expert knowledge, NN-kNN empowers clinicians to make more informed and ethical decisions. Practitioners can manually adjust the model parameters, discover unforeseen patterns or biases, and decide when it is appropriate to rely on AI in clinical settings.

Through our qualitative interviews, practitioners expressed appreciation for the model's transparency, flexibility, and the way it integrates their clinical expertise into the diagnostic process. This balance of AI-driven insights and human judgment has the potential to build greater trust and utility in AI-based diagnostic tools. Looking ahead, we plan to extend NN-kNN by incorporating complex data from additional modalities, further supporting clinicians in delivering more holistic and data-rich diagnoses.

References

Aguilera, J.; González, L. C.; Montes-y Gómez, M.; and Rosso, P. 2019. A new weighted k-nearest neighbor algorithm based on newton's gravitational force. In Vera-Rodriguez, R.; Fierrez, J.; and Morales, A., eds., *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 305–313. Cham: Springer International Publishing.

Aha, D. W., and Goldstone, R. L. 1992. Concept learning and flexible weighting. In *In Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 534–539. Erlbaum.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Au-Yeung, S. K.; Bradley, L.; Robertson, A. E.; Shaw, R.; Baron-Cohen, S.; and Cassidy, S. 2019. Experience of mental health diagnosis and perceived misdiagnosis in autistic, possibly autistic and non-autistic adults. *Autism* 23(6):1508–1518.

Ayano, G.; Demelash, S.; Yohannes, Z.; Haile, K.; Tulu, M.; Assefa, D.; Tesfaye, A.; Haile, K.; Solomon, M.; Chaka, A.; et al. 2021. Misdiagnosis, detection rate, and associated factors of severe psychiatric disorders in specialized psychiatry centers in ethiopia. *Annals of general psychiatry* 20:1–10.

Banerjee, A.; Mishra, S.; Malik, V.; and Albuquerque, V. H. C. d. 2024. Assessing psychological risks severity level using variants of k-nearest neighbor algorithm. *AIP Conference Proceedings* 2919(1):090007.

Bicego, M., and Loog, M. 2016. Weighted k-nearest neighbor revisited. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 1642–1647.

Bonzano, A.; Cunningham, P.; and Smyth, B. 1997. Using introspective learning to improve retrieval in CBR: A case study in air traffic control. In Leake, D., and Plaza, E., eds., *Proceedings of the 2nd International Conference on Case-Based Reasoning (ICCBR-97)*, volume 1266 of *LNAI*, 291–302. Berlin: Springer.

Bzdok, D., and Meyer-Lindenberg, A. 2018. Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3(3):223–230.

Chahar, R.; Dubey, A. K.; and Narang, S. K. 2024. An efficient knn algorithm for the mental health performance assessment using k-means clustering. In Shetty, N. R.; Prasad, N. H.; and Nagaraj, H. C., eds., *Advances in Communication and Applications*, 575–586. Singapore: Springer Nature Singapore.

Chattopadhyay, S. 2017. A neuro-fuzzy approach for the diagnosis of depression. *Applied Computing and Informatics* 13(1):10–18.

Das, A., and Rad, P. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Dasarathy, B. V. 1991. Nearest neighbor (nn) norms : Nn pattern classification techniques. *IEEE Computer Society Tutorial*.

Eatough, V., and Smith, J. A. 2017. Interpretative phenomenological analysis. *The Sage handbook of qualitative research in psychology* 193–209.

Elmunsyah, H.; Mu'awanah, R.; Widiyaningtyas, T.; Zaeni, I. A.; and Dwiyanto, F. A. 2019. Classification of employee mental health disorder treatment with k-nearest neighbor algorithm. In *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, volume 6, 211–215.

Friedman, J. H. 1994. Flexible metric nearest neighbor classification. Technical report, Stanford University.

Gates, L., and Leake, D. B. 2021. Evaluating cbr explanation capabilities: Survey and next steps. In *ICCBR Workshops*.

Goldberger, J.; Hinton, G. E.; Roweis, S.; and Salakhutdinov, R. R. 2004. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Graham, S.; Depp, C.; Lee, E. E.; Nebeker, C.; Tu, X.; Kim, H.-C.; and Jeste, D. V. 2019. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports* 21:1–18.

Iyortsuun, N. K.; Kim, S.-H.; Jhon, M.; Yang, H.-J.; and Pant, S. 2023. A review of machine learning and deep learning approaches on mental health diagnosis. In *Healthcare*, volume 11, 285. MDPI.

Jarvie, H., and Lindén, H. 2024. Exploring human therapists' perspectives on artificial intelligence therapists in mental health care.

Joyce, D. W.; Kormilitzin, A.; Smith, K. A.; and Cipriani, A. 2023. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine* 6(1):6.

Kerz, E.; Zanwar, S.; Qiao, Y.; and Wiechmann, D. 2023. Toward explainable ai (xai) for mental health detection based on language behavior. *Frontiers in psychiatry* 14:1219479.

Lau, C.; Zhu, X.; and Chan, W.-Y. 2023. Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry* 14:1160291.

Li, O.; Liu, H.; Chen, C.; and Rudin, C. 2018. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 3530–3537. AAAI Press.

Li, E.; Kealy, D.; Aafjes-van Doorn, K.; McCollum, J.; Curtis, J. T.; Luo, X.; and Silberschatz, G. 2024. "it felt like i was being tailored to the treatment rather than the treatment being tailored to me": Patient experiences of helpful and unhelpful psychotherapy. *Psychotherapy Research* 1–15.

Liu, H.; Wong, Y. J.; Mitts, N. G.; Li, P. J.; and Cheng, J. 2020. A phenomenological study of east asian international students' experience of counseling. *International Journal for the Advancement of Counselling* 42:269–291.

Manzali, Y.; Barry, K. A.; Flouchi, R.; Balouki, Y.; and Elfar, M. 2024. A feature weighted k-nearest neighbor algorithm based on association rules. *Journal of Ambient Intelligence and Humanized Computing* 1–14.

Marchiori, E. 2013. *Class Dependent Feature Weighting and K-Nearest Neighbor Classification*. Berlin, Heidelberg: Springer Berlin Heidelberg. 69–78.

Maxwell, J. A. 2021. Why qualitative methods are necessary for generalization. *Qualitative Psychology* 8(1):111.

Mihura, J. L.; Roy, M.; and Graceffo, R. A. 2017. Psychological assessment training in clinical psychology doctoral programs. *Journal of Personality assessment* 99(2):153–164.

Molnar, C.; Casalicchio, G.; and Bischl, B. 2020. Interpretable machine learning – a brief history, state-of-the-art and challenges. In Koprinska, I.; Kamp, M.; Appice, A.; Loglisci, C.; Antonie, L.; Zimmermann, A.; Guidotti, R.; Özgöbek, Ö.; Ribeiro, R. P.; Gavaldà, R.; Gama, J.; Adilova, L.; Krishnamurthy, Y.; Ferreira, P. M.; Malerba, D.; Medeiros, I.; Ceci, M.; Manco, G.; Masciari, E.; Ras, Z. W.; Christen, P.; Ntoutsi, E.; Schubert, E.; Zimek, A.; Monreale, A.; Biecek, P.; Rinzivillo, S.; Kille, B.; Lommatzsch, A.; and Gulla, J. A., eds., *ECML PKDD 2020 Workshops*, 417–431. Cham: Springer International Publishing.

Orozco-del Castillo, M. G.; Orozco-del Castillo, E. C.; Brito-Borges, E.; Bermejo-Sabbagh, C.; and Cuevas-Cuevas, N. 2021. An artificial neural network for depression screening and questionnaire refinement in undergraduate students. In Mata-Rivera, M. F., and Zagal-Flores, R., eds., *Telematics and Computing*, 1–13. Cham: Springer International Publishing.

Perkins, A.; Ridler, J.; Browes, D.; Peryer, G.; Notley, C.; and Hackmann, C. 2018. Experiencing mental health diagnosis: a systematic review of service user, clinician, and carer perspectives across clinical settings. *The Lancet Psychiatry* 5(9):747–764.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Ricci, F., and Avesani, P. 1995. *Learning a local similarity metric for case-based reasoning*. Berlin, Heidelberg: Springer Berlin Heidelberg. 301–312.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1:206–215.

Saeed, W., and Omlin, C. 2023. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* 263:110273.

Saleem, R.; Yuan, B.; Kurugollu, F.; Anjum, A.; and Liu, L. 2022. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing* 513:165–180.

Schoenborn, J. M.; Weber, R. O.; Aha, D. W.; Cassens, J.; and Althoff, K.-D. 2021. Explainable case-based reasoning: a survey. In *AAAI-21 Workshop Proceedings*.

Shapley, L. S. 1953. A value for n-person games. *Contribution to the Theory of Games* 2.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems* 29.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* 10:207–244.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *CoRR* abs/1410.3916.

Wettschereck, D.; Aha, D.; and Mohri, T. 1997. A review and empirical evaluation of feature-weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review* 11(1-5):273–314.

Ye, X.; Leake, D.; Wang, Y.; Zhao, Z.; and Crandall, D. 2024. Towards network implementation of cbr: Case study of a neural network k-nn algorithm. In *Case-Based Reasoning Research and Development: 32nd International Conference, ICCBR 2024, Merida, Mexico, July 1–4, 2024, Proceedings*, 354–370. Berlin, Heidelberg: Springer-Verlag.

Zhang, T.; Schoene, A. M.; Ji, S.; and Ananiadou, S. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine* 5(1):1–13.