Jurnal Matematika UNAND
Vol. **14** No. **4** Hal. 341 – 354
Edisi Oktober 2025
ISSN : 2303–291X
e-ISSN : 2721–9410
©Departemen Matematika dan Sains Data FMIPA UNAND
https://doi.org/10.25077/jmua.14.4.341-354.2025

AN EXPLAINABLE HYBRID AI FRAMEWORK USING FUZZY ROUGH SET RULES FOR MENTAL HEALTH PREDICTION

RUSTAM
Department of Telecommunication Engineering,
School of Electrical Engineering, Telkom University, Jl. Telekomunikasi No.1 Dayeuh Kolot,
40257 Kabupaten Bandung, Jawa Barat, Indonesia
email : rustamtelu@telkomuniversity.ac.id

Received July 17, 2025 Received in revised form August 13, 2025
Accepted September 9, 2025 Available online October 31, 2025

Abstract. *The increasing use of artificial intelligence (AI) in mental health prediction highlights the need for models that are both accurate and explainable. This paper proposes an explainable hybrid AI framework that integrates the K-Nearest Neighbors (KNN) classifier with fuzzy rough set (FRS) rule induction to provide transparent, human-readable explanations of predictions. The framework combines the predictive strength of KNN with the interpretability of FRS-generated fuzzy linguistic rules, enabling clear post hoc reasoning without sacrificing accuracy. A large-scale mental health dataset is utilized, comprising behavioral, psychological, and lifestyle attributes, with "coping struggles" as the target variable. The FRS-based rule induction process is formally described using fuzzy similarity relations, lower and upper approximations, and a tunable soft matching mechanism. Experimental results show that the hybrid model achieves 94.5% accuracy, 87.7% precision, 100% recall, and 93.4% F1-score, while producing high-coverage rules that closely align with the KNN predictions. Compared to a traditional fuzzy inference system (FIS), the proposed framework demonstrates superior scalability and fidelity in high-dimensional settings. These findings illustrate the potential of combining statistical learning and symbolic reasoning through FRS to develop interpretable, scalable, and trustworthy AI tools for mental health screening and decision support, addressing an underexplored area in the existing literature.*

Keywords: explainable artificial intelligence; fuzzy rough sets; interpretable machine learning; mental health prediction; rule induction

1. Introduction

The global prevalence of mental health disorders, including depression, anxiety, and stress-related conditions, has escalated in recent years, posing significant challenges to healthcare systems. These conditions affect millions worldwide and necessitate timely, accurate assessment tools to facilitate early intervention and personalized care. Artificial intelligence (AI) has emerged as a promising avenue for enhancing mental health diagnostics and treatment planning, offering capabilities to analyze complex datasets and identify patterns beyond traditional analytical methods [1,2].

A major challenge, however, lies in the opacity of many AI models—often referred to as "black-box" systems—which provide predictions without transparent reasoning pathways. This limits clinician trust and hinders adoption, especially in mental health contexts where understanding the rationale behind a decision is crucial [3,4]. Explainable Artificial Intelligence (XAI) seeks to address this issue by making AI decisions interpretable to humans, thereby enhancing clinical applicability, patient acceptance, and trust [5,6].

Among XAI methodologies, fuzzy logic stands out for its ability to handle uncertainty and mimic human reasoning [7,8]. Its capability to represent subjective mental health indicators such as "feeling down" or "experiencing stress" through fuzzy sets has been widely demonstrated. Recent advances in fuzzy clustering have also improved robustness in handling incomplete and noisy data [9], further supporting its applicability to mental health prediction.

Previous work has applied fuzzy inference systems (FIS) for depression assessment [10,11] and decision support in educational and clinical settings [12,13]. More recently, hybrid models combining fuzzy logic with modern AI methods have been proposed to integrate symbolic reasoning with statistical learning, enhancing both accuracy and interpretability [14,15]. However, most fuzzy systems rely on manually constructed rules, which are time-consuming to develop and difficult to scale in high-dimensional spaces. Data-driven rule induction techniques such as fuzzy rough sets (FRS) offer a promising alternative [16,17].

Despite these advances, there remains a lack of benchmarking of fuzzy logic-based approaches against high-performing machine learning models on large-scale mental health datasets, particularly with explicit evaluation of rule-level fidelity and coverage. Addressing these gaps, this paper introduces a hybrid explainable AI framework that integrates the predictive power of K-Nearest Neighbors (KNN) with FRS-based rule induction, alongside a transparent FIS baseline for comparison.

The main contribution of this study lies in the design and evaluation of a transparent fuzzy inference system (FIS) using expert-defined rules based on core psychological indicators, and in the development of a hybrid explainable AI framework that combines the predictive strength of the K-Nearest Neighbors (KNN) classifier with the interpretability of fuzzy rough set (FRS) rule induction. The proposed framework not only achieves strong predictive performance, as measured by standard classification metrics such as accuracy, precision, recall, and F1-score, but also provides linguistically interpretable explanations through compact rule sets whose fidelity and coverage with respect to the underlying classifier are explicitly evaluated. By leveraging the theoretical rigor of FRS for automated, data-driven rule induction, the framework overcomes the scalability limitations of manually constructed fuzzy systems, enabling its application to multi-feature, high-dimensional mental health datasets while preserving interpretability for real-world decision support.

2. Related Work

AI adoption in mental health aims to improve diagnostic accuracy, treatment personalization, and early intervention [18,19]. Fuzzy logic and XAI have emerged as key approaches for addressing the inherent uncertainty of mental health assessment [20,21].

Fuzzy logic offers a mathematical framework for handling imprecision, making it effective for modeling subjective indicators. Studies have demonstrated its use in depression assessment with interpretable reasoning [22,23] and in decision support for various populations [24,25]. In parallel, XAI methods have been applied to mental health prediction from wearables, questionnaires, and social media data, enabling actionable insights [26,27,28,29,30,31,32].

Hybrid systems integrating fuzzy logic with AI combine the adaptability of machine learning with the transparency of rule-based models [33,34]. These have proven valuable in domains where interpretability is essential [35,36], with automated rule extraction reducing reliance on manual rule design [37,38]. Rough set theory and association rule mining have also supported interpretable healthcare models [39].

FRS is a particularly promising yet underutilized tool for explainability, offering a principled way to handle vagueness in rule extraction while maintaining consistency with observed data [40]. FRS-based approaches can produce concise, human-readable rules that approximate complex classifiers [16,41,42]—a key advantage in mental health, where symptom patterns are often overlapping and context-dependent.

Challenges remain in ensuring generalizability across contexts [43,44,45] and integrating XAI into clinical workflows [46,47]. Nonetheless, combining fuzzy logic and XAI aligns with responsible AI goals [48,49,50,51]. This study addresses the underexplored integration of high-performing ML with FRS-based explainability, aiming to achieve both scalability and interpretability for mental health screening and intervention.

3. Dataset and Proposed Methodology

3.1. *Overview of the Proposed Hybrid Explainable AI Framework*

The main objective of this study is to develop a hybrid explainable AI framework that integrates the predictive strength of the K-Nearest Neighbors (KNN) classifier with the interpretability of fuzzy rough set (FRS) rule induction. This architecture maintains high classification performance while generating transparent, human-readable explanations in the form of fuzzy linguistic rules.

The framework consists of three main components:

(1) **Predictive Core (KNN)**: Trained on a curated subset of features from the dataset to maximize predictive accuracy for the target variable *Coping Struggles*.
(2) **FRS Rule Induction Layer**: Extracts a compact set of interpretable fuzzy if–then rules that approximate the decision boundaries of the KNN model.
(3) **Explainability Metrics**: Uses coverage and fidelity to evaluate how well the

extracted rules explain the underlying KNN predictions.

The proposed framework is a novel hybrid model that explicitly integrates a high-performing predictive component with a transparent explanation layer. The predictive core employs the K-Nearest Neighbors (KNN) classifier, selected based on its superior accuracy in baseline evaluations. The explainability layer is constructed using a mathematically grounded Fuzzy Rough Set (FRS) rule induction process, which follows a systematic pipeline: computing fuzzy similarity relations between instances, constructing lower and upper approximations, and generating fuzzy *if–then* rules that are both interpretable and clinically meaningful. Each rule is evaluated using quantitative metrics such as support and confidence, ensuring that only reliable patterns are retained. The quality of the generated explanations is further assessed using coverage and fidelity metrics, which measure the proportion of instances explained by the rules and the degree of agreement with the predictive core, respectively. A tunable soft matching threshold is incorporated to balance the trade-off between interpretability and predictive precision. This explicit integration of KNN with FRS rule induction results in a well-defined architecture that goes beyond a simple combination of existing methods, providing a coherent and reproducible hybrid model capable of delivering high predictive performance while maintaining transparent and mathematically traceable explanations for mental health prediction.

3.2. *Dataset and Preprocessing*

This study uses a curated version of the "Mental Health in Tech Survey" dataset, consisting of 10,000 records with 17 categorical attributes describing demographic, behavioral, and mental health-related factors.

3.2.1. *Dataset Structure*

The attributes are grouped as follows:

- Demographic: gender, country, occupation, self-employed.
- Mental health status/history: family history, treatment, mental health history, coping struggles, mood swings.
- Behavioral and lifestyle: days indoors, growing stress, changes habits, work interest, social weakness.
- Healthcare access and preferences: mental health interview, care options.

3.2.2. *Target Variable*

The target variable *Coping Struggles* is binary (Yes/No) and indicates whether a respondent experiences challenges in coping with psychological, social, or emotional pressures. It is clinically meaningful, well-distributed, and suitable for fuzzy rule-based inference.

3.2.3. *Data Quality and Transformation*

The dataset is complete for core predictive features, with only 249 missing values in the *self-employed* attribute, handled via imputation or exclusion. For fuzzy modeling, categorical responses are mapped to a 0–10 numeric scale to define interpretable membership functions. A subset of seven features was selected based on clinical relevance and predictive contribution: *Growing Stress*, *Mood Swings*, *Social Weakness*, *Changes Habits*, *Mental Health History*, *Work Interest*, and *Days Indoors*.

3.3. ***Predictive Core: K-Nearest Neighbors (KNN)***

KNN is chosen as the predictive core based on baseline comparisons with Logistic Regression, Naive Bayes, and Support Vector Machine. It classifies instances based on the majority label among nearest neighbors in the feature space, making it well-suited for the dataset's distribution.

3.4. ***FRS Rule Induction Layer***

The FRS layer extracts rules that approximate the KNN classifier's decision function.

Fuzzy Indiscernibility Relation

For attribute $a \in A$ with fuzzy partition $\{T_a^1, T_a^2, \cdots, T_a^k\}$, the fuzzy similarity between x_i and x_j is:

$$R_a(x_i, x_j) = \max_{j=1,2,\cdots,k} \min\left(\mu_{T_a^j}(x_i), \mu_{T_a^j}(x_j)\right). \tag{3.1}$$

The overall relation over A is:

$$R_A(x_i, x_j) = \min_{a \in A} R_a(x_i, x_j). \tag{3.2}$$

Fuzzy Lower and Upper Approximations

For decision class d with membership:

$$\mu_X(x) = \begin{cases} 1 & \text{if } D(x) = d, \\ 0 & \text{otherwise,} \end{cases} \tag{3.3}$$

the lower and upper approximations are:

$$\mu_{\underline{R}(X)}(x) = \inf_{y \in U} \max\left(1 - R_A(x, y), \mu_X(y)\right), \tag{3.4}$$

$$\mu_{\overline{R}(X)}(x) = \sup_{y \in U} \min\left(R_A(x, y), \mu_X(y)\right). \tag{3.5}$$

Rule Construction and Matching

A fuzzy rule is:

$$\text{IF } \bigwedge_{a \in R'} (a \text{ is } T_a) \text{ THEN } D = d, \tag{3.6}$$

with degree of matching:

$$\mu_{Cond}(x) = \min_{a \in R'} \mu_{T_a}(x), \tag{3.7}$$

and consequent match:

$$\mu_{D=d}(x) = \begin{cases} 1 & \text{if } D(x) = d, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

Rule Support and Confidence

Support:

$$Supp(r) = \sum_{x \in U} \mu_{Cond}(x) \cdot \mu_{D=d}(x), \tag{3.9}$$

Confidence:

$$Conf(r) = \frac{Supp(r)}{\sum_{x \in U} \mu_{Cond}(x)}. \tag{3.10}$$

3.5. *Explainability Metrics*

Coverage:

$$Coverage = \frac{|\{x \in Test \mid \mu_{Cond}(x) \geq \theta\}|}{|Test|}. \tag{3.11}$$

Fidelity:

$$Fidelity = \frac{|\{x \in Test \mid \mu_{Cond}(x) \geq \theta \ \wedge \ D(x) = d_{rule}\}|}{|\{x \in Test \mid \mu_{Cond}(x) \geq \theta\}|}. \tag{3.12}$$

Coverage measures the proportion of test instances to which at least one fuzzy rule applies. Fidelity measures agreement between rule-based predictions and the KNN classifier among covered instances.

3.6. *Baseline Explainable Model: Fuzzy Inference System (FIS)*

For comparison, a baseline FIS was built using three variables: *Growing Stress*, *Mood Swings*, and *Social Weakness*. The output *Coping Struggles* is defined over [0,10] with triangular membership functions:

$$\mu(x; a, b, c) = \begin{cases} 0 & \text{if } x \leq a \text{ or } x \geq c, \\ \dfrac{x-a}{b-a} & \text{if } a < x \leq b, \\ \dfrac{c-x}{c-b} & \text{if } b < x < c. \end{cases} \tag{3.13}$$

Out of $3^3 = 27$ possible rules, seven clinically relevant ones were selected to prioritize detection of high-risk cases. The FIS serves as a transparent benchmark but exhibits scalability issues in high-dimensional settings.

4. Results and Discussion

This section presents and discusses the experimental results obtained from multiple evaluation scenarios designed to assess both predictive performance and explainability. The analysis proceeds in stages. First, we establish baseline performance using conventional machine learning models, including Logistic Regression, Naive Bayes, Support Vector Machine, and K-Nearest Neighbor. Second, we evaluate a baseline Fuzzy Inference System (FIS) as an inherently interpretable model. Third, we introduce and analyze the proposed Hybrid Explainability Framework, which combines a high-performing KNN model with Fuzzy Rough Set (FRS)-based rule induction to provide interpretable explanations of model predictions.

For the hybrid framework, we compare two matching strategies—Crisp Matching and Soft Matching—examining the trade-offs between rule coverage and fidelity, calculated according to Eq. (3.11) and Eq. (3.12), respectively. The results demonstrate the strengths and limitations of each approach, providing insights into their applicability across different practical scenarios.

4.1. *Baseline Machine Learning Models*

As a foundation for comparison, we first evaluate the classification performance of several commonly used machine learning (ML) algorithms on the target mental health dataset. The models examined include Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes (NB). Each model was trained and tested using consistent preprocessing and evaluation protocols, and their performance was assessed based on four standard metrics: Accuracy, Precision, Recall, and F1 Score.

Table 1 summarizes the results. Among the evaluated models, the K-Nearest Neighbors classifier achieved the highest overall performance, with an impressive accuracy of **98.85%**, precision of **98.52%**, recall of **99.35%**, and F1 score of **98.93%**. These results suggest that KNN is highly effective in capturing the underlying structure of the data, consistent with the fuzzy similarity aggregation in Eq. (3.2), where classification depends on proximity relationships among instances.

Support Vector Machine followed with moderate performance, achieving an accuracy of 76.80% and an F1 score of 78.57%. Although SVM exhibited better recall than precision, it struggled to generalize as effectively as KNN. Logistic Regression and Naive Bayes showed notably lower accuracy (67.15% and 66.95%, respectively), with corresponding F1 scores below 70%. These results reflect their limitations in handling non-linear and potentially overlapping class distributions.

The dominance of KNN in this setting may be attributed to its instance-based nature, which is conceptually related to the fuzzy indiscernibility relation in Eq. (3.1), where similarity between objects drives decision-making. However, KNN lacks inherent interpretability, making it challenging to extract human-readable justifications for its predictions. This motivates the integration of KNN with interpretable rule-based reasoning from FRS.

Table 1: Performance of Baseline Machine Learning Models

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.6715	0.6859	0.7105	0.6982
K-Nearest Neighbors	**0.9885**	**0.9852**	**0.9935**	**0.9893**
Support Vector Machine	0.7680	0.7738	0.7998	0.7856
Naive Bayes	0.6695	0.6925	0.6866	0.6895

4.2. *Fuzzy Inference System (Baseline Explainable Model)*

To establish a transparent and interpretable baseline for mental health screening, a fuzzy inference system (FIS) was developed using expert-derived linguistic rules as in Eq. (3.6) and triangular membership functions defined in Eq. (3.13). Unlike black-box classifiers, the FIS provides a rule-based reasoning framework that translates psychological input variables into interpretable if–then statements.

The system was designed to operate based on core psychological indicators, and its performance is summarized in Table 2. The FIS achieved an accuracy of **52.03%**, precision of **52.88%**, recall of **90.91%**, and an F1 score of **66.86%**.

Table 2: Performance of Fuzzy Inference System (FIS)

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
FIS	0.5203	0.5288	0.9091	0.6686

While accuracy and precision are relatively low, the system demonstrates high recall—reflecting an emphasis on inclusiveness, even at the expense of false positives. In terms of rule metrics, the expert-crafted rules inherently balance support (Eq. (3.9)) and confidence (Eq. (3.10)) differently from data-driven models, prioritizing coverage of high-risk cases.

However, scaling the FIS to all seven selected features led to rule explosion per Eq. (3.6), resulting in sparse rule activation and degraded performance. This limitation motivates adopting a data-driven fuzzy rule induction process to handle high-dimensional inputs.

4.3. *Proposed Hybrid Explainability Framework (KNN + FRS Explain)*

The hybrid framework integrates KNN's predictive performance with FRS-based interpretable rules. Using fuzzy lower and upper approximations (Eq. (3.4) and Eq. (3.5)), the FRS layer approximates KNN's decision regions. Rule activation follows Eq. (3.7), with rule quality assessed via Eq. (3.9) and Eq. (3.10).

The framework achieved accuracy of **94.5%**, precision of **87.7%**, recall of

100%, and an F1 score of **93.4%** (Table 3), outperforming all baselines and preserving interpretability.

Table 3: Performance of Proposed Hybrid Explainability Framework (KNN + FRS Explain)

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
KNN + FRS Explain	0.9450	0.8770	1.0000	0.9340

In addition to predictive performance, the framework extracted ten high-coverage, linguistically interpretable rules using the procedure in Eq. (3.6)–(3.10). These rules are shown in Table 4.

Table 4: Top 10 Extracted Fuzzy Rough Set Rules (FRS)

No.	**FRS Rule (Antecedents)**	**Conclusion (Coping_Struggles)**
1	Growing_Stress is Low AND Mood_Swings is Low AND Social_Weakness is High AND Changes_Habits is Medium AND Mental_Health_History is High AND Work_Interest is High AND Days_Indoors is Low	No
2	Growing_Stress is Medium AND Mood_Swings is Low AND Social_Weakness is Low AND Changes_Habits is High AND Mental_Health_History is Low AND Work_Interest is Medium AND Days_Indoors is Low	Yes
3	Growing_Stress is High AND Mood_Swings is High AND Social_Weakness is Medium AND Changes_Habits is Medium AND Mental_Health_History is Medium AND Work_Interest is High AND Days_Indoors is Low	No
4	Growing_Stress is High AND Mood_Swings is High AND Social_Weakness is High AND Changes_Habits is Medium AND Mental_Health_History is Medium AND Work_Interest is Low AND Days_Indoors is Low	No
5	Growing_Stress is High AND Mood_Swings is Medium AND Social_Weakness is High AND Changes_Habits is Low AND Mental_Health_History is High AND Work_Interest is Low AND Days_Indoors is Low	No
6	Growing_Stress is High AND Mood_Swings is Medium AND Social_Weakness is High AND Changes_Habits is Medium AND Mental_Health_History is Medium AND Work_Interest is Medium AND Days_Indoors is Low	Yes
7	Growing_Stress is High AND Mood_Swings is High AND Social_Weakness is High AND Changes_Habits is Medium AND Mental_Health_History is Medium AND Work_Interest is Low AND Days_Indoors is Low	Yes
8	Growing_Stress is High AND Mood_Swings is Low AND Social_Weakness is Low AND Changes_Habits is Low AND Mental_Health_History is High AND Work_Interest is Low AND Days_Indoors is Low	Yes
9	Growing_Stress is High AND Mood_Swings is Medium AND Social_Weakness is Medium AND Changes_Habits is Low AND Mental_Health_History is Low AND Work_Interest is Low AND Days_Indoors is Low	No
10	Growing_Stress is High AND Mood_Swings is Low AND Social_Weakness is High AND Changes_Habits is Medium AND Mental_Health_History is High AND Work_Interest is Medium AND Days_Indoors is Low	Yes

4.4. *Rule Coverage and Fidelity: Crisp vs Soft Matching*

Coverage and fidelity are computed using Eq. (3.11) and Eq. (3.12). In crisp matching ($\theta = 0$), antecedents must fully satisfy the membership conditions in Eq. (3.7). In soft matching, partial satisfaction is allowed when $\mu_{Cond}(x) \geq \theta$.

Table 5: Crisp vs Soft Matching Summary: Rule Coverage and Fidelity

Matching Type	**Coverage (%)**	**Fidelity (%)**
Crisp	85.12	71.20
Soft $\theta = 0.1$	85.12	71.20
Soft $\theta = 0.2$	58.68	54.93
Soft $\theta = 0.3$	58.68	54.93
Soft $\theta = 0.4$	11.57	100.00
Soft $\theta = 0.5$	11.57	100.00
Soft $\theta = 0.6$	6.06	100.00
Soft $\theta = 0.7$	0.00	0.00
Soft $\theta = 0.8$	0.00	0.00
Soft $\theta = 0.9$	0.00	0.00

Lower θ increases the denominator in Eq. (3.12), raising coverage but potentially reducing agreement. Higher θ yields high agreement but reduced explanatory reach, reflecting the precision–coverage trade-off in interpretable models.

Novelty and Contribution

The novelty lies in combining symbolic reasoning from FRS (Eq. (3.6)–(3.10)) with statistical classification (KNN), decoupling prediction from explanation. This modular approach maintains accuracy while generating rules aligned with the base model's decision boundaries.

Comparative Insights

Compared to the standalone FIS using Eq. (3.13), the hybrid approach scales to seven features without rule explosion. Against black-box ML models, it offers transparency via rules computed per Eq. (3.7)–(3.10), validated by coverage and fidelity in Eq. (3.11)–(3.12).

5. Conclusion

This paper introduces a well-defined hybrid explainable AI framework that integrates the predictive capability of the K-Nearest Neighbors (KNN) classifier with the interpretability of fuzzy rough set (FRS) rule induction for mental health prediction. Unlike a conventional fuzzy inference system (FIS), the proposed framework is architecturally composed of two distinct but integrated layers: a predictive core, implemented via KNN to achieve high baseline accuracy on the *Coping Struggles* classification task, and an explainability layer, formulated through fuzzy similarity relations, lower and upper approximations, fuzzy *if–then* rule construction, and rule quality measures. Experimental evaluation on a large-scale mental health dataset demonstrates that the hybrid model achieves **94.5%** accuracy, **87.7%** precision, **100%** recall, and **93.4%** F1-score, while maintaining strong rule coverage and fidelity under both crisp and soft matching schemes. The extracted fuzzy rules are semantically coherent, clinically interpretable, and exhibit high agreement with the KNN predictions. By decoupling the learning and explanation layers mathematically yet integrating them functionally, the proposed architecture mitigates the rule explosion problem commonly faced by traditional FIS in high-dimensional spaces, while preserving interpretability. These results confirm that fuzzy rough set theory provides a principled mathematical foundation for generating faithful, human-readable explanations of statistical learning models. The framework offers a scalable and trustworthy solution for AI-based clinical decision support, bridging the gap between predictive performance and interpretability in real-world mental health applications.

6. Acknowledgment

The author gratefully acknowledges the financial support provided by Telkom University.

Bibliography

[1] Ahmed, M.Z., Ahmed, O., Aibao, Z., Hanbin, S., Siyu, L., Ahmad, A., 2020, Epidemic of COVID-19 in China and associated psychological problems, *Asian Journal of Psychiatry*, Vol. **51**: 102092.

[2] Iyortsuun, N.K., Kim, S.H., Jhon, M., Yang, H.J., Pant, S., 2023, A review of machine learning and deep learning approaches on mental health diagnosis, *Healthcare*, Vol. **11**(3): 285.

[3] Tjoa, E., Guan, C., 2020, A survey on explainable artificial intelligence (XAI): Toward medical XAI, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. **32**(11): 4793–4813.

[4] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019, Causability and explainability of artificial intelligence in medicine, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. **9**(4): e1312.

[5] Arrieta, A.B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Herrera, F., 2020, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, Vol. **58**: 82–115.

[6] Adadi, A., Berrada, M., 2018, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), *IEEE Access*, Vol. **6**: 52138–52160.

[7] Mohammadi Motlagh, H.A., Minaei Bidgoli, B., Parvizi Fard, A.A., 2018, Design and implementation of a web-based fuzzy expert system for diagnosing depressive disorder, *Applied Intelligence*, Vol. **48**: 1302–1313.

[8] Rajawat, A.S., Bedi, P., Goyal, S.B., Bhaladhare, P., Aggarwal, A., Singhal, R.S., 2023, Fusion fuzzy logic and deep learning for depression detection using facial expressions, *Procedia Computer Science*, Vol. **218**: 2795–2805.

[9] Rustam, Usman, K., Kamaruddin, M., Chamidah, D., Nopendri, Saleh, K., Eliskar, Y., Marzuki, I., 2021, Modified possibilistic fuzzy c-means algorithm for clustering incomplete data sets, *Acta Polytechnica*, Vol. **61**(2): 364–377.

[10] Tadalagi, M., Joshi, A.M., 2021, AutoDep: automatic depression detection using facial expressions based on linear binary pattern descriptor, *Medical & Biological Engineering & Computing*, Vol. **59**(6): 1339–1354.

[11] Amin, M., Ullah, K., Asif, M., Waheed, A., Haq, S.U., Zareei, M., Biswal, R.R., 2022, ECG-based driver's stress detection using deep transfer learning and fuzzy logic approaches, *IEEE Access*, Vol. **10**: 29788–29809.

[12] Sharkadi, M., Dorovtsi, A., 2024, Building a fuzzy model for determining the level of social well-being of the population, *Eastern-European Journal of Enterprise Technologies*, Vol. **130**(4).

[13] Papadimitriou, S., Virvou, M., 2025, Fuzzy Logic and Applications in Education and Games: Theory, Practical Implementations and a Literature Review, *Artificial Intelligence—Based Games as Novel Holistic Educational Environments to Teach 21st Century Skills*, pp. 95–127.

[14] Mohamed, E.S., Naqishbandi, T.A., Bukhari, S.A.C., Rauf, I., Sawrikar, V., Hussain, A., 2023, A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms, *Healthcare Analytics*, Vol. **3**: 100185.

[15] Kumar, A., Sangwan, S.R., Arora, A., Menon, V.G., 2022, Depress-DCNF: A deep convolutional neuro-fuzzy model for detection of depression episodes using IoMT, *Applied Soft Computing*, Vol. **122**: 108863.

[16] Zhu, X., Wang, D., Pedrycz, W., Li, Z., 2022, Fuzzy rule-based local surrogate models for black-box model explanation, *IEEE Transactions on Fuzzy Systems*, Vol. **31**(6): 2056–2064.

[17] Suzuki, A., Negishi, E., 2024, Fuzzy Logic Systems for Healthcare Applications, *Journal of Biomedical and Sustainable Healthcare Applications*, Vol. **4**(1).

[18] Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F., Acharya, U.R., 2022, Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022), *Computer Methods and Programs in Biomedicine*, Vol. **226**: 107161.

[19] Le Glaz, A., Haralambous, Y., Kim-Dufor, D.H., Lenca, P., Billot, R., Ryan, T.C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., et al., 2021, Machine learning and natural language processing in mental health: systematic review, *Journal of Medical Internet Research*, Vol. **23**(5): e15708.

[20] Jayalakshmi, M., Garg, L., Maharajan, K., Jayakumar, K., Srinivasan, K., Bashir, A.K., Ramesh, K., 2021, Fuzzy logic-based health monitoring system for COVID'19 patients, *Computers, Materials and Continua*, Vol. **67**(2): 2431–2447.

[21] Khoo, L.S., Lim, M.K., Chong, C.Y., McNaney, R., 2024, Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches, *Sensors*, Vol. **24**(2): 348.

[22] Rad, D., Rad, G., Maier, R., Demeter, E., Dicu, A., Popa, M., Alexuta, D., Floroian, D., Mărineanu, V.D., 2022, A fuzzy logic modelling approach on psychological data, *Journal of Intelligent & Fuzzy Systems*, Vol. **43**(2): 1727–1737.

[23] Bakhtavar, E., Valipour, M., Yousefi, S., Sadiq, R., Hewage, K., 2021, Fuzzy cognitive maps in systems risk analysis: a comprehensive review, *Complex & Intelligent Systems*, Vol. **7**: 621–637.

[24] Qazi, S., Raza, K., 2021, Fuzzy logic-based hybrid knowledge systems for the detection and diagnosis of childhood autism, in: *Handbook of Decision Support Systems for Neurological Disorders*, Elsevier, pp. 55–69.

[25] Bai, L., Han, X., 2023, Design of a mental health assessment system based on fuzzy evaluation, in: *Proceedings of the 2023 International Conference on Data Science and Network Security (ICDSNS)*, IEEE, pp. 1–6.

[26] Chaddad, A., Peng, J., Xu, J., Bouridane, A., 2023, Survey of explainable AI techniques in healthcare, *Sensors*, Vol. **23**(2): 634.

[27] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A., 2024, Interpreting black-box models: a review on explainable artificial intelligence, *Cognitive Computation*, Vol. **16**(1): 45–74.

[28] Chen, Z., Xiao, F., Guo, F., Yan, J., 2023, Interpretable machine learning for building energy management: A state-of-the-art review, *Advances in Applied Energy*, Vol. **9**: 100123.

[29] Darko, A.P., Antwi, C.O., Adjei, K., Zhang, B., Ren, J., 2024, Predicting determinants influencing user satisfaction with mental health app: An explainable machine learning approach based on unstructured data, *Expert Systems with Applications*, Vol. **249**: 123647.

[30] Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., Kurakin, A., 2021, Adamatch: A unified approach to semi-supervised learning and domain adaptation, *arXiv preprint arXiv:2106.04732.*

[31] Long, N., Lei, Y., Peng, L., Xu, P., Mao, P., et al., 2022, A scoping review on monitoring mental health using smart wearable devices, *Mathematical Biosciences and Engineering*, Vol. **19**(8): 7899–7919.

[32] Koh, J.X., Liew, T.M., 2022, How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds, *Journal of Psychiatric Research*, Vol. **145**: 317–324.

[33] Kumar, M., Nigam, A., Singh, S.V., 2024, Analysis of Hybrid Fuzzy-Neuro Model for Diagnosis of Depression, in: *Recent Advances in Computational Intelligence and Cyber Security*, CRC Press, pp. 36–42.

[34] Murad, N.Y., Hasan, M.H., Azam, M.H., Yousuf, N., Yalli, J.S., 2024, Unraveling the black box: A review of explainable deep learning healthcare techniques, *IEEE Access*, Article in press.

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

[36] Thieme, A., Hanratty, M., Lyons, M., Palacios, J., Marques, R.F., Morrison, C., Doherty, G., 2023, Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment, *ACM Transactions on Computer-Human Interaction*, Vol. **30**(2): 1–50.

[37] Papadopoulos, P., Soflano, M., Chaudy, Y., Adejo, W., Connolly, T.M., 2022, A systematic review of technologies and standards used in the development of rule-based clinical decision support systems, *Health and Technology*, Vol. **12**(4): 713–727.

[38] Muhammad, L.J., Garba, E.J., Oye, N.D., Wajiga, G.M., Garko, A.B., 2021, Fuzzy rule-driven data mining framework for knowledge acquisition for expert system, in: *Translational Bioinformatics in Healthcare and Medicine*, Elsevier, pp. 201–214.

[39] Mwangi, I.K., Nderu, L., Mwangi, R.W., Njagi, D.G., 2023, Hybrid interpretable model using roughset theory and association rule mining to detect interaction terms in a generalized linear model, *Expert Systems with Applications*, Vol. **234**: 121092.

[40] Wang, X., Lu, F., Zhou, M., Zeng, Q., 2022, A synergy-effect-incorporated fuzzy Petri net modeling paradigm with application in risk assessment, *Expert Systems with Applications*, Vol. **199**: 117037.

[41] Maria, A.J., Castiello, C., Luis, M., Mencar, C., et al., 2021, Explainable fuzzy systems: Paving the way from interpretable fuzzy systems to explainable AI systems, *Studies in Computational Intelligence*, Vol. **970**: 1–253.

[42] Moradi, M., Samwald, M., 2021, Post-hoc explanation of black-box classifiers using confident itemsets, *Expert Systems with Applications*, Vol. **165**: 113941.

[43] Sogancioglu, G., Mosteiro, P., Salah, A.A., Scheepers, F., Kaya, H., 2024, Fairness in AI-based mental health: Clinician perspectives and bias mitigation, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. **7**: 1390–1400.

[44] Sharma, E., De Choudhury, M., 2018, Mental health support and its relationship to linguistic accommodation in online communities, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

[45] Sharma, A., Biloria, D., 2025, Sociological Perspective on the Future of AI in Health: Trends and Prospects, in: *AI Technologies and Advancements for*

Psychological Well-Being and Healthcare, IGI Global, pp. 53–76.

[46] Suresh, H., Guttag, J., 2021, A framework for understanding sources of harm throughout the machine learning life cycle, in: *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–9.

[47] Grote, T., Keeling, G., 2022, Enabling fairness in healthcare through machine learning, *Ethics and Information Technology*, Vol. **24**(3): 39.

[48] Upendran, S.V., 2024, Explainable AI in Healthcare: A Multi-Disciplinary Perspective, in: *Analyzing Explainable AI in Healthcare and the Pharmaceutical Industry*, IGI Global, pp. 58–71.

[49] Tornero-Costa, R., Martinez-Millana, A., Azzopardi-Muscat, N., Lazeri, L., Traver, V., Novillo-Ortiz, D., et al., 2023, Methodological and quality flaws in the use of artificial intelligence in mental health research: systematic review, *JMIR Mental Health*, Vol. **10**(1): e42045.

[50] Valente, F., Henriques, J., Paredes, S., Rocha, T., de Carvalho, P., Morais, J., 2021, A new approach for interpretability and reliability in clinical risk prediction: Acute coronary syndrome scenario, *Artificial Intelligence in Medicine*, Vol. **117**: 102113.

[51] Shan, Y., Ji, M., Xie, W., Lam, K.Y., Chow, C.Y., 2022, Public trust in artificial intelligence applications in mental health care: topic modeling analysis, *JMIR Human Factors*, Vol. **9**(4): e38799.

[52] Rustam, Gunawan, A.Y., Kresnowati, M.T.A.P., 2022, Data dimensionality reduction technique for clustering problem of metabolomics data, *Heliyon*, Vol. **8**(6).

[53] Rustam, 2025, Pretreatment methods for enhancing machine learning performance on metabolomics data, *IEEE Access*, Vol. **13**: 80133–80148.

[54] Rustam, Gunawan, A.Y., Kresnowati, M.T.A.P., 2020, Artificial neural network approach for the identification of clove buds origin based on metabolites composition, *Acta Polytechnica*, Vol. **60**(5): 440–447.

[55] Kerz, E., Zanwar, S., Qiao, Y., Wiechmann, D., 2023, Toward explainable AI (XAI) for mental health detection based on language behavior, *Frontiers in Psychiatry*, Vol. **14**: 1219479.

[56] Naqvi, S., Shaikh, A.Z., Altaf, T., Singh, S., 2021, Fuzzy logic enabled stress detection using physiological signals, *Emerging Technologies in Computing: Proceedings of the 4th EAI/IAER International Conference, iCETiC 2021*, Springer, pp. 161–173.

[57] Ahmed, U., Lin, J.C.W., Srivastava, G., 2021, Fuzzy explainable attention-based deep active learning on mental-health data, *Proceedings of the 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, pp. 1–6.