Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Artificial Intelligence (XAI) in Psychology and Mental Health: A Systematic Literature Review

P.~Yu. Sereda-Kalinin

seredapj@ipran.ru

Institute of Psychology of the Russian Academy of Sciences

A.~S. Vlasova

Institute of Psychology of the Russian Academy of Sciences

A.~S. Panfilova

Institute of Psychology of the Russian Academy of Sciences

Research Article

Keywords: explainable artificial intelligence, mental health, systematic literature review, SHAP, LIME, interpretable machine learning, psychology

Posted Date: May 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9705625/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explainable Artificial Intelligence (XAI) in Psychology and Mental Health: A Systematic Literature Review

P. Yu. Sereda-Kalinin[a,1,2], A. S. Vlasova[b,1], A. S. Panfilova[c,1]

[1]Institute of Psychology of the Russian Academy of Sciences, 13/1 Yaroslavskaya st., Moscow, 129366, Russia.
[2]Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25 Alexander Solzhenitsyn st., Moscow, 109004, Russia.
E-mail: [a]seredapj@ipran.ru, [b]vlasovaas@ipran.ru, [c]panfilova87@gmail.com.

Abstract

Artificial intelligence is rapidly advancing across scientific disciplines, and its growing applicability to complex real-world problems has naturally drawn interest from psychology and mental health research. Given the error-sensitive nature of clinical practice, explainable artificial intelligence (XAI) — methods that make model predictions more transparent and interpretable — has attracted considerable attention among researchers, with a notable rise in publications. This systematic literature review (SLR) analyses **42** peer-reviewed studies (2020–2026), and addresses four research questions concerning the psychological domains targeted, the XAI methods employed, the underlying ML models and datasets, as well as the challenges and future directions in the field. The analysis shows that depression is the most studied domain, followed by stress and well-being, with SHAP being the dominant XAI method (57% of studies). Based on the identified studies, we characterise the most common ML architectures paired with XAI approaches, analyse the alignment between model types and explanation methods, and derive a set of methodological recommendations for applying XAI in psychology — including strategies for addressing the limitations inherent to each explanation technique.

Keywords: explainable artificial intelligence, mental health, systematic literature review, SHAP, LIME, interpretable machine learning, psychology

1 Introduction

Artificial intelligence (AI) is becoming an increasingly powerful tool across a wide range of scientific disciplines, and psychology is no exception. Machine learning models are already being applied to predict depression (Seok et al. 2025; Mumenin et al. 2025; Trelles et al. 2025), diagnose autism (Supekar et al. 2022; Biswas et al. 2021; Lee and Yoo 2025), assess suicide risk (de Oliveira et al. 2025; Park and Sakong 2026), and measure stress levels (Bozorgmehr and Weltermann 2023; Tariq et al. 2025; Shikha et al. 2024). Despite their strong predictive performance, most contemporary algorithms operate as *black boxes*: they produce an output without explaining the basis on which a given decision was made.

This lack of transparency is critical in all domains where automated decisions directly affect human lives — medicine, law, education, and social services. In psychology and psychiatry, automated decisions likewise directly bear on human well-being. Explainable AI (XAI) is a set of methods that make model predictions more transparent and interpretable to the end user: a clinician, a researcher, or a patient.

Interest in XAI has grown rapidly in recent years: the number of publications on the topic has shown exponential growth since 2020 (Arrieta et al. 2020). Related systematic reviews cover XAI in finance (Černevičienė and Kabašinskas 2024; Ayari et al. 2025; Yeo et al. 2025; Khan et al. 2025; Zafar and Wu 2026), medicine (Band et al. 2023), and the narrower overlapping area of diagnosing mental disorders from EEG (Zhai et al. 2025). However, no comprehensive systematic review covering XAI across the full spectrum of psychological constructs had previously existed.

The present review fills this gap: we systematically map which psychological domains are covered by XAI applications, which XAI methods are used in psychological research, which ML models and dataset characteristics are employed, and which key challenges and promising directions the existing literature identifies.

This systematic review is structured to answer four key research questions (RQs) that address the current state, shortcomings, and future directions of XAI applications in the field of psychology. These questions form the narrative backbone of this manuscript, with dedicated sections addressing them through a synthesis of evidence from the 42 selected peer-reviewed studies. Based on this synthesis, we also formulate methodological recommendations for researchers applying XAI methods in psychological research. The review follows PRISMA 2020 guidelines (Page et al. 2021) and systematic review methodology (Kitchenham and Charters 2007).

Research questions:

- **RQ1:** Which psychological constructs are investigated using XAI methods?
- **RQ2:** Which XAI methods are applied in each specific domain?
- **RQ3:** Which ML models are used and what are the characteristics of the datasets employed?
- **RQ4:** What are the key challenges, limitations, and promising directions for XAI in psychology?

Contributions of this review. Compared to prior reviews, this work makes the following contributions.

1. **The first comprehensive SLR** on XAI in psychology, covering the full spectrum of psychological constructs — from depression and anxiety to creativity and personality traits — rather than a single disorder.
2. **A unified taxonomy**: a systematisation of XAI methods along the ante-hoc/post-hoc, local/global, and model-agnostic/model-specific axes as applied to psychological tasks.
3. **Quantitative synthesis**: descriptive statistics comparing psychological domains, XAI methods, and ML algorithms across 42 selected peer-reviewed studies (2020–2026).
4. **Practical recommendations**: based on the synthesis of the reviewed studies and drawing on the authors' own experience, recommendations are formulated for researchers, clinicians, and developers of XAI systems in psychology.

This review is addressed to two audiences: psychologists and clinicians seeking to understand the capabilities and limitations of XAI tools in their field, and AI specialists wishing to bring their expertise to the domain of psychological research.

The paper is structured as follows. Section 2 provides the theoretical background: a taxonomy of XAI methods and the rationale for their application in psychological research. Section 3 describes the SLR methodology. Section 4 presents the review findings organised by research question. Section 5 discusses implications and limitations and, on that basis, formulates practical recommendations. Section 6 presents the conclusions.

2 Background

2.1 Explainable Artificial Intelligence: Definitions and Taxonomy

Imagine that a machine learning model has predicted a high risk of depression for a particular patient. Which factors led to this result? How much can this prediction be trusted? XAI methods allow researchers and clinicians to answer these questions (Arrieta et al. 2020; Adadi and Berrada 2018).

Unlike traditional statistical models, whose outputs are inherently interpretable, contemporary algorithms — gradient boosting, deep neural networks, transformers — operate as black boxes: they produce a prediction but do not explain why. XAI methods bridge this gap by providing the specialist with an intelligible justification for an automated decision (Arrieta et al. 2020).

Three key distinctions help navigate the diversity of XAI methods:

Timing of the explanation: ante-hoc vs. post-hoc. When interpretability is built into the model itself, this is referred to as an *ante-hoc* (or intrinsic) approach: decision trees, logistic regression, and rule-based systems are transparent by nature. When a complex model is trained first and the explanation is constructed separately, this is a *post-hoc* approach (SHAP, LIME, Grad-CAM, etc.). In psychological research, post-hoc methods are the most prevalent, as they allow high predictive performance to be combined with the possibility of subsequent interpretation.

Scope of the explanation: global vs. local. *Global* explanations describe model behaviour across the entire sample: for instance, which features are generally most important for predicting depression risk. This is useful for a researcher wishing to assess whether the model is consistent with theoretical expectations. *Local* explanations refer to a specific observation — an individual patient or case. They are indispensable in clinical practice, where the specialist needs to understand why the model produces a given output for a particular individual.

Dependence on model architecture: model-agnostic vs. model-specific. *Model-agnostic* methods (SHAP, LIME, counterfactual explanations) treat any model as a black box — they require only access to model inputs and outputs. *Model-specific* methods (Grad-CAM, attention mechanisms) require access to the model's weights and internal activations and are applicable only to certain architectures — typically neural networks. The choice between these approaches in practice is often determined by whether the internal computations of the model are accessible: if the model's internal structure is not available to the researcher, model-agnostic methods may be the only option.

Figure 1 illustrates the position of XAI methods within a typical machine learning pipeline across all three taxonomic dimensions.

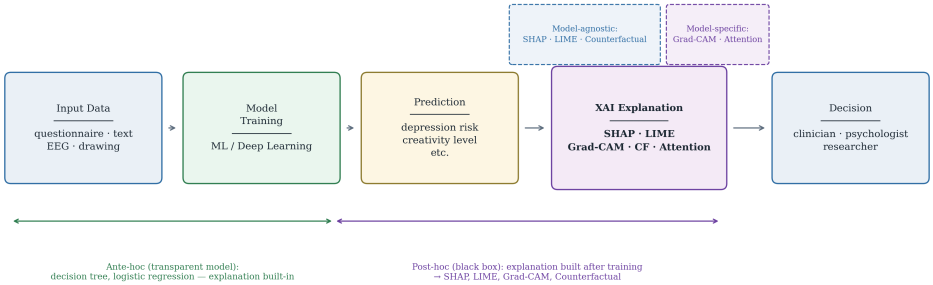


Figure 1 XAI methods in the machine learning pipeline. The pipeline comprises five stages: data input, model training, prediction, XAI explanation, and specialist decision-making. Labels indicate the stage at which methods are applied: *ante-hoc* (the model is transparent from the outset) and *post-hoc* (the explanation is constructed after training). The upper blocks above the XAI stage distinguish *model-agnostic* methods (SHAP, LIME, Counterfactual — require no access to model weights) from *model-specific* methods (Grad-CAM, Attention — require access to internal activations).

Figure 2 presents examples of explanation visualisations for a number of XAI methods in the context of psychological tasks.

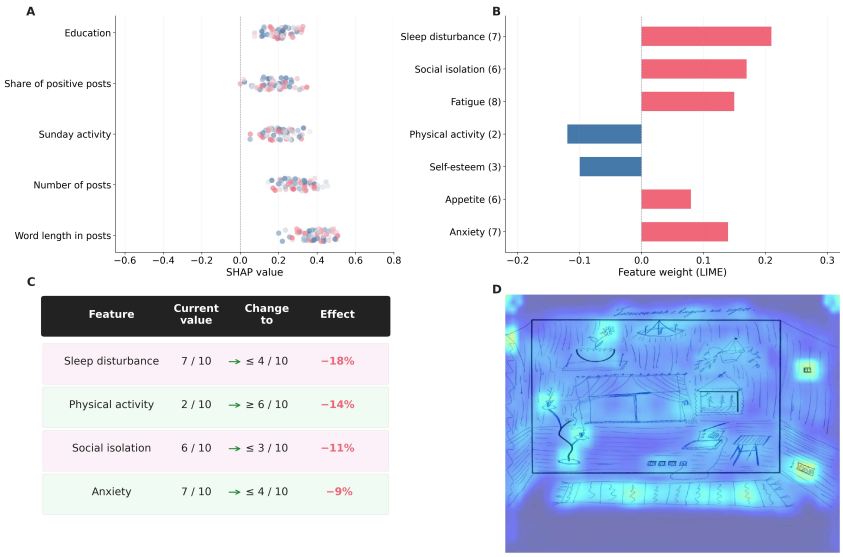


Figure 2 Examples of XAI explanation visualisations in psychological research. **A** — SHAP beeswarm (simulated on social media activity data for personality trait prediction: post length, number of posts, Sunday activity, proportion of positive posts, education level): each point corresponds to one participant; colour encodes feature value (red — high, blue — low); horizontal position reflects the feature's contribution to the prediction (rightward — increases, leftward — decreases). **B** — LIME (simulated for one patient for depression risk prediction: sleep disturbance, social isolation, fatigue, physical activity, self-esteem, appetite): local explanation for one patient; bar length reflects the feature's contribution to the predicted depression risk (74%); red — increases risk, blue — decreases. **C** — Counterfactual explanation (simulated): minimal changes to input data that reduce depression risk from 74% to 22%. **D** — Grad-CAM (Figure 3 from Panfilova et al. (Panfilova et al. 2024)); warm colours highlight the regions on which the model relied when assessing creativity.

A structured comparison of the XAI methods used in the reviewed studies, by criteria of purpose, methodology, output form, and practical considerations, is provided in Table 4 (Appendix B).

2.2 Artificial Intelligence in Psychology and Mental Health

Over the past decade, machine learning methods have found wide application in psychology and psychiatry. Classification and prediction tasks are addressed across a broad spectrum of constructs, including the detection of depression from text and speech (Kerz et al. 2023; Seok et al. 2025; Mumenin et al. 2025), autism screening from behavioural data (Supekar et al. 2022; Biswas et al. 2021; Albahri et al. 2023), suicide risk assessment in social media (de Oliveira et al. 2025; Park and Sakong 2026), and stress monitoring via wearable sensors (Shikha et al. 2024; Paul et al. 2025).

Several key data modalities are prominent in this area. *Text and speech data* (social media posts, clinical notes, interview transcripts) are processed using NLP methods, including pre-trained transformer models (BERT, RoBERTa). *Neurophysiological signals* (EEG, fMRI) are used for objective assessment of affective states and neurobiological markers of mental disorders. *Wearable biosensors* provide continuous,

ecologically valid monitoring of physiological correlates of stress and emotion. *Psychometric questionnaires* remain the most common data source owing to their accessibility and standardisation.

Among ML algorithms, ensemble methods (XGBoost, Random Forest), logistic regression, and deep neural networks (CNN, LSTM, transformers) dominate the field. The strong predictive performance of these models has created demand for XAI methods that ensure the transparency of automated decisions for clinicians and researchers.

2.3 Motivation for XAI in Psychology

The first motivation is **model logic verification and clinical transparency**. The application of AI in the area of psychology raises specific ethical and practical challenges. Unlike finance or logistics, decisions in clinical psychology directly affect human life and health. This is why the transparency and justifiability of automated decisions — for instance, in suicide risk assessment or depression screening — are of particular importance (de Oliveira et al. 2025). XAI addresses this problem by allowing clinicians to verify the model's logic, identify spurious correlations, and communicate results to patients in an intelligible, interpretable form.[1]

The second motivation is **bias detection and fairness**. ML models reproduce and amplify biases present in the training data. In a psychological context this is especially significant: datasets are often formed from unrepresentative samples (predominantly young, urban, English-speaking participants), which leads to systematic underperformance of the model for underrepresented demographic groups (Lee et al. 2025). XAI methods make it possible to identify which features (e.g., age, gender, ethnicity) contribute undesirable variance to predictions, and thus serve as a tool for auditing model fairness prior to deployment.

The third motivation is **XAI as a tool for scientific discovery**. Unlike medicine, where the mechanisms of many pathologies are well understood, psychological science has only partial theoretical explanations for most constructs. Here XAI opens up additional epistemic value: a feature importance profile can generate new hypotheses about psychological mechanisms. For instance, SHAP analysis in a study of Big Five personality traits (Panfilova and Turdakov 2024) revealed non-linear interactions among social media activity features, some of which were consistent with classical personality theories and some of which diverged from them. The fact that part of the model's activations coincided with theoretically expected features serves as an argument in favour of model validity — and it is precisely this correspondence that makes the divergences a meaningful signal for theoretical reflection rather than an artefact (Panfilova et al. 2024). An analogous mechanism was demonstrated in a study of autism screening from children's drawings (Lee and Yoo 2025): analysis of saliency maps revealed that the eye region is a diagnostically significant feature, whereupon the authors explicitly incorporated this feature into the model — which led to an improvement in its

[1] The legal dimension is defined by the EU General Data Protection Regulation (GDPR), Article 22 of which restricts the use of fully automated decisions that significantly affect individuals and implies a right to explanation (European Parliament and Council of the European Union 2016). Analogous requirements regarding transparency and informed consent are contained in the Ethical Principles of Psychologists and Code of Conduct of the American Psychological Association (American Psychological Association 2017).

metric. Thus, XAI becomes not only a tool for model interpretation but also a means of generating theoretical hypotheses subject to further experimental verification.

2.3.1 XAI Method Categories Relevant to Psychology

SHAP (Lundberg and Lee 2017) is the most prevalent XAI method in psychological research: it is used in 57% of the studies in this review. The method is based on the concept of Shapley values from cooperative game theory: each feature is assigned a numerical contribution representing its effect on the deviation of a specific prediction from the sample mean. The key advantage of SHAP for psychologists is its *dual level of analysis*: global explanations show which factors generally determine risk (e.g., sleep disturbance as the leading predictor of depression across the entire sample), while local explanations allow a specific case to be examined — what contribution each feature made to the prediction for a particular patient. The method's limitation is the high computational cost of KernelSHAP (for neural networks and arbitrary models) and the potential instability of values when features are correlated, which is especially relevant for psychometric scales measuring related constructs.

LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al. 2016) constructs a local explanation by approximating the behaviour of a complex model in the neighbourhood of a single prediction with a simple linear model. For the practitioner, this method answers the question: "Which specific factors determined the prediction for this particular individual?" Compared to SHAP, the method is less computationally demanding and equally accessible to a non-specialist audience; however, its explanations can vary substantially across runs owing to the stochastic nature of the algorithm's initialisation stage — an important argument against using LIME in clinical protocols without additional verification.

Counterfactual explanations answer the question: "What is the minimum change to the input data that would cause the model to change its prediction?" This is the most *actionable* type of explanation from a clinical standpoint, as it directly points to potential intervention targets: for example, reducing a sleep disturbance score from 7 to 4 moves a patient from the high-risk to the moderate-risk group for depression (Figure 2, panel C). The key methodological challenge is ensuring the *plausibility* of the proposed changes: a counterfactual scenario must be clinically achievable, not merely mathematically optimal.

Attention mechanisms are built into transformer architectures, which are applied to a wide range of data types (e.g., text, audio, and visual data). Analysing attention weights and patterns allows visualising which parts of the input data contribute most to a given model prediction. In tasks involving analysis of social media text or clinical notes, the attention mechanism is interpreted as pointing to the lexical markers most informative for diagnosis: for example, words from the semantic fields of "hopelessness" or "fatigue" in depression detection. An important caveat: high attention weight does not imply a causal relationship — it is merely evidence of a statistical association in the training sample (Jain and Wallace 2019).

Gradient-based methods (Grad-CAM, Integrated Gradients) are applied primarily in computer vision models — convolutional and other differentiable architectures — to highlight regions of the input signal most relevant to the prediction. In

psychological research these methods are especially useful when working with images (drawings from projective tests, facial expressions) and time series (EEG): an activation heatmap clearly shows which regions of a drawing or time intervals of a signal the model identified as diagnostically significant — for example, indicators of creativity (Figure 2, panel D). Their limitation is applicability exclusively to models with accessible gradients, which precludes use with models available only for inference via API without access to weights and internal activations.

Inherent interpretability (ante-hoc) involves the use of intrinsically transparent models: decision trees, logistic regression, rule-based systems. Their principal advantage is that the explanation is not an approximation: a decision tree literally shows the path from input data to result. In psychology, such models are used where interpretability matters more than predictive performance — for example, in developing diagnostic algorithms that must be comprehensible to a clinician without additional tools. Their practical drawback is generally lower predictive performance compared to ensemble and deep learning methods, particularly on complex multimodal data.

3 Methodology

This SLR was conducted in accordance with the PRISMA 2020 guidelines (Page et al. 2021) and systematic review methodology (Kitchenham and Charters 2007). The review protocol was defined prior to data extraction. The research questions are formulated in Section 1.

3.1 Search Strategy

The search was conducted in the following databases: **PubMed/MEDLINE**, **IEEE Xplore**, **ACM Digital Library** (manual export), and **Web of Science** (WoS; search conducted on 10 April 2026). Due to the absence of institutional access, Scopus was not included; this limitation is noted in Section 5.4. The main search in PubMed, IEEE Xplore, and ACM DL was conducted on 26 March 2026. An additional manual search was carried out via Google Scholar, Semantic Scholar, and ResearchGate (primary: January 2026; update: April 2026). The manual search covered the following domains: psychophysiology, psychological profiling, and creativity (studies included in the review); psycholinguistics, educational psychology, and organisational psychology (studies identified but excluded by selection criteria); cognitive and social psychology (predominantly theoretical works examining the contribution of these disciplines to XAI rather than the reverse were found; no empirical studies meeting the inclusion criteria were identified). Notably, the few studies at the intersection of XAI with organisational and developmental psychology nevertheless focus on mental health care tasks (Mokheleli et al. 2025; Mokheleli 2025), which indirectly confirms that it is the practical significance and high risks of clinical errors that make mental health a particularly relevant domain for XAI application.

The search query followed a two-block Boolean structure:

Block A (XAI and interpretability terms) `AND` **Block B** (psychological domain terms)

Block A included: *"explainable artificial intelligence", "explainable AI", "XAI", "interpretable machine learning", "SHAP", "Shapley", "LIME", "feature importance", "counterfactual explanation", "model interpretability", "model explainability", "explainability"*.
Block B included: *"mental health", "depression", "anxiety", "stress", "PTSD", "well-being", "psychological", "psychiatric", "autism", "schizophrenia", "suicide", "personality disorder", "cognitive load", "bipolar", "OCD", "sleep disorder", "resilience", "mood disorder", "ADHD", "burnout", "mental disorder", "emotional"*.

The search was restricted to publications in **English** for the period **2015–2026**. Full search strings for each database are provided in Appendix 6.

3.2 Inclusion and Exclusion Criteria

Inclusion criteria:

- Application of an XAI or ML method to a psychological construct

Exclusion criteria:

- No open-access full text available
- Books, book chapters, technical reports, editorial articles
- Non-peer-reviewed preprints (arXiv, bioRxiv)
- Duplicate records
- Purely theoretical or methodological works on XAI
- Publications in which the subject of investigation may be psychological but the research focus is not: for example, studies defining stress exclusively through physiological indicators

3.3 Study Selection and PRISMA Flow Diagram

Database and manual searches identified a total of **4 304** records (PubMed: 1 860; IEEE Xplore Title/Abstract: 54; IEEE Xplore All Metadata: ≈1 538; ACM DL: 11; WoS: 797; manual search: 44). A substantial proportion of records from PubMed, IEEE Xplore, and WoS were not in open access and were not admitted to further screening. After deduplication, **83** unique records were retained for further screening.

Title/abstract screening led to the exclusion of **11** records for the following reasons: XAI applied outside a psychological domain ($n = 7$), no explicit XAI focus ($n = 2$), purely methodological contribution without empirical data ($n = 2$).

The remaining **72** records underwent full-text screening, conducted independently by both authors. Each study was assessed against the inclusion and exclusion criteria defined in Section 3.2. Disagreements were resolved through discussion until consensus was reached. Following full-text screening, a further **30** studies were excluded. The final sample comprised **42 studies** included in the qualitative synthesis.

The complete PRISMA 2020 flow diagram is presented in Figure 3.

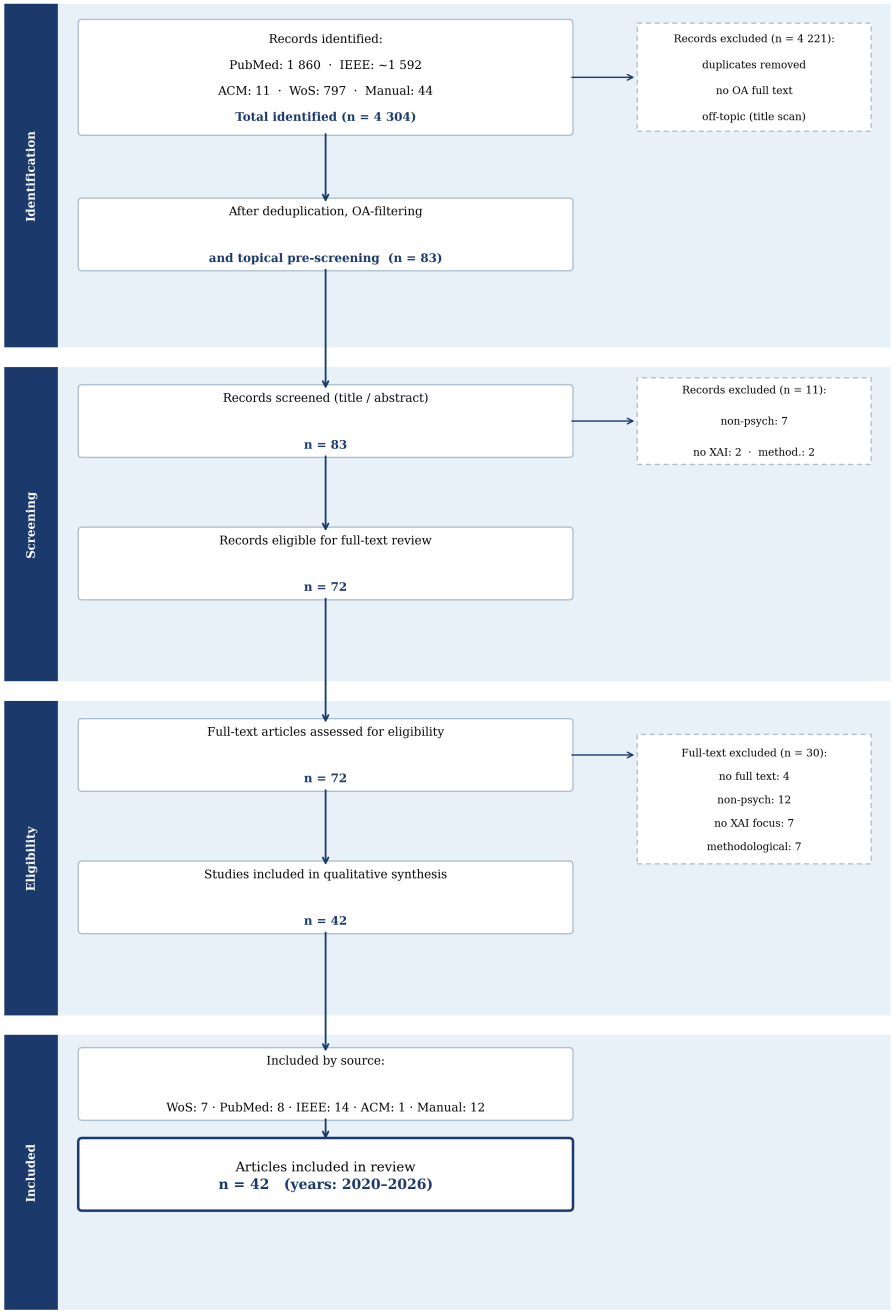


Figure 3 PRISMA 2020 flow diagram ($n = 4\,304$ records identified; **42 studies** included in the qualitative synthesis). Dashed boxes indicate excluded records.

3.4 Descriptive Analysis of Included Studies

The 42 included studies span the period 2020–2026, with a marked increase in publications from 2023 onwards. Figure 4 shows the distribution of studies by psychological domain over time: the horizontal line reflects the range of publication activity within a domain, with its thickness and colour saturation proportional to the number of included studies (n). Most domains received sustained research attention in 2023–2025; depression remains the most studied domain ($n = 13$).

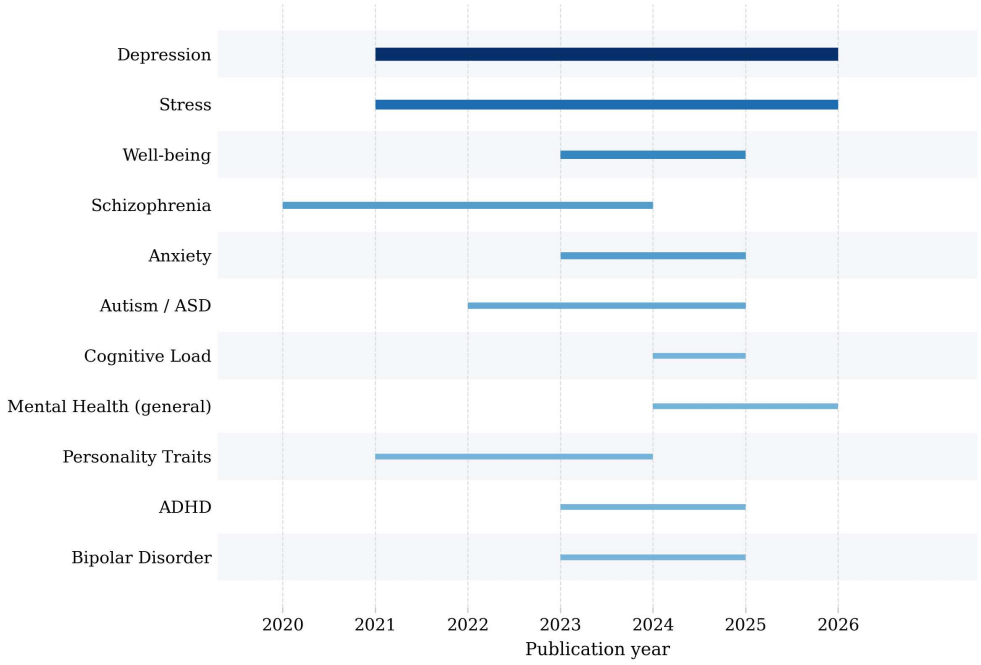


Figure 4 Temporal distribution of included studies by psychological domain ($n = 42$; only the 12 most represented domains are shown). The horizontal line indicates the range of publications in a given domain; line thickness and colour saturation are proportional to the number of studies in the domain.

The distribution of included studies by source: IEEE Xplore ($n = 14$, 33%), manual search ($n = 12$, 29%), PubMed/MEDLINE ($n = 8$, 19%), Web of Science ($n = 7$, 17%), ACM Digital Library ($n = 1$, 2%).

The most studied construct is depression ($n = 13$), followed by stress ($n = 8$) and well-being ($n = 5$). Anxiety and autism/ASD each appear in four studies; bipolar disorder, schizophrenia, suicide risk, ADHD, personality traits, and mental health in general are each covered by two studies. A number of studies address several domains simultaneously.

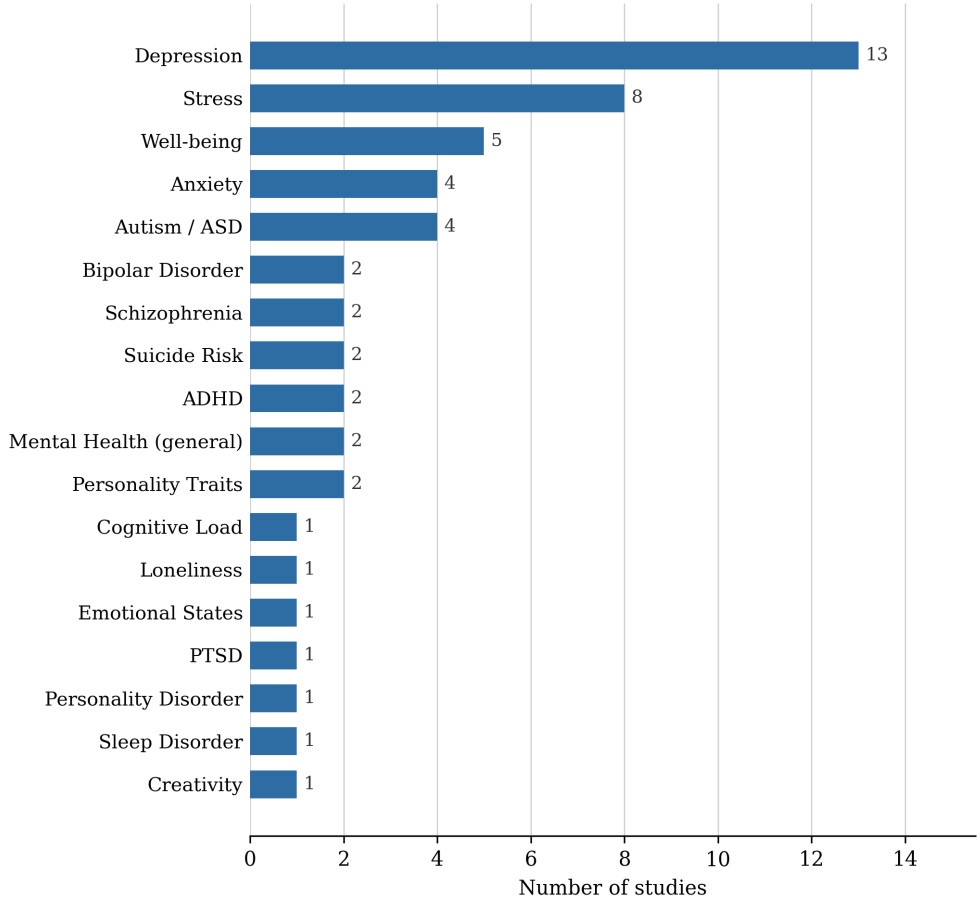


Figure 5 Distribution of included studies by psychological domain ($n = 42$). A single study may cover multiple domains; counts reflect the number of occurrences at the domain level. The most frequently studied constructs are depression, stress, and well-being.

Figure 6 presents the XAI methods used across the 42 included studies. SHAP is the most prevalent method ($n = 24$, 57%), followed by LIME ($n = 11$, 26%) and feature importance methods ($n = 7$, 17%). Several studies employed broader or data-type-specific explainability frameworks, including Grad-CAM and Integrated Gradients.

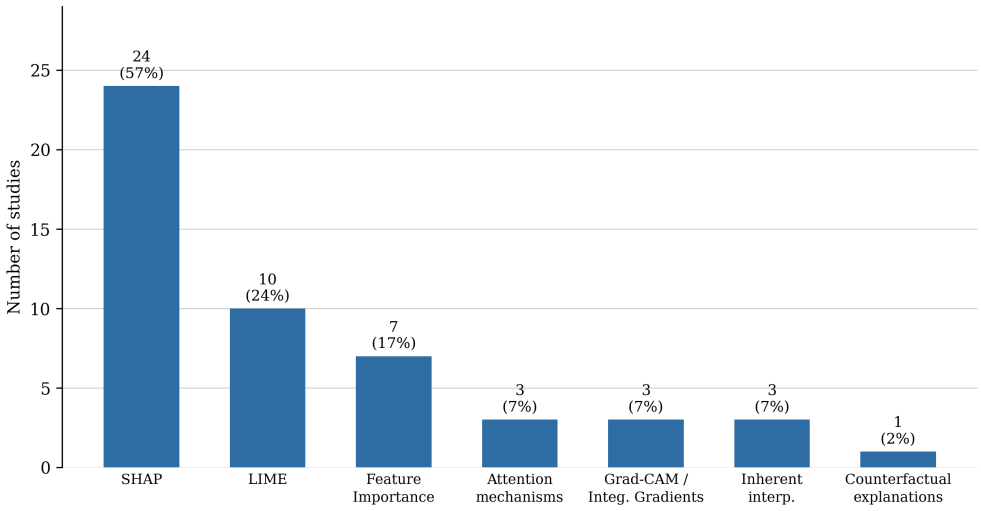


Figure 6 XAI methods used in the included studies ($n = 42$). A single study may employ multiple methods. SHAP dominates, followed by LIME and feature importance methods.

4 Results of the Review and Key Findings

4.1 RQ1: Which Psychological Constructs Are Studied Using XAI Methods?

The 42 included studies cover a broad range of psychological domains. Depression is the most extensively studied condition ($n = 13$, 31%), followed by stress ($n = 8$, 19%) and psychological well-being ($n = 5$, 13%). Anxiety disorders and autism spectrum disorder (ASD) are each represented by four studies (11%). Bipolar disorder, schizophrenia, suicide risk, attention-deficit/hyperactivity disorder (ADHD), and personality trait assessment are each addressed in two studies (5%), whereas sleep disorders, cognitive load, post-traumatic stress disorder (PTSD), personality disorders, and several other constructs are represented in only a small number of studies. The distribution across domains is presented in Table 1.

4.2 RQ2: Which XAI Methods Are Applied?

The reviewed studies employ a wide range of XAI approaches, which can be grouped into four categories: (1) feature importance-based methods, (2) local explanation methods, (3) inherently interpretable models, and (4) visualisation- and attention-based methods. The distribution of methods across application domains is presented in Table 1.

A complete breakdown of all 42 included studies by domain, XAI method, and ML algorithm is provided in Table 2.

Table 1 Distribution of included studies by psychological domain and XAI method ($n = 42$). Studies may address multiple domains and methods; counts reflect occurrences. *F.I.* = Feature Importance; *I.I.* = Inherent Interpretability; *Attn.* = Attention; *Other* = Grad-CAM, Counterfactual, Integrated Gradients, NLP Explainability, etc.

Domain	n	**%**	**SHAP**	**LIME**	**F.I.**	**I.I.**	**Attn.**	**Other**
Depression	13	31	8	4	4	1	–	–
Stress	8	19	6	2	2	–	–	2
Well-being	5	13	3	1	–	1	1	2
Anxiety	4	10	1	3	–	1	–	–
Autism / ASD	4	10	–	1	3	–	–	–
Bipolar Disorder	2	5	1	2	–	–	–	–
Schizophrenia	2	5	1	–	–	–	–	1
Suicide Risk	2	5	–	1	–	–	1	–
ADHD	2	5	1	1	–	–	–	–
Mental Health (general)	2	5	–	–	–	1	1	1
Personality Traits	2	5	2	–	–	–	–	1
Sleep Disorder	1	3	1	–	–	–	1	1
Cognitive Load	1	3	–	–	–	–	–	2
Emotional States	1	3	1	1	–	–	–	–
Loneliness	1	3	–	–	–	1	–	–
PTSD	1	3	1	–	–	–	–	–
Personality Disorder	1	3	–	–	–	–	–	1
Creativity	1	3	–	–	–	–	–	1
Total	**53**	—	**24**	**10**	**7**	**5**	**3**	**12**

Table 2: Summary of included studies by domain, XAI method, and ML algorithm ($n = 42$). Studies covering multiple domains appear on multiple rows. *Syst. review* = systematic review / meta-analysis. Underlined references mark the first occurrence of each study.

Domain	XAI Method	ML Algorithm	Ref.
Depression	Inherent interp.	Deep Learning + Decision Tree	(Liu et al. 2024)
	SHAP, Feature Importance	XGBoost, ALO, MARL, ALO-MARL	(Zhu et al. 2025)
	LIME	Deep Learning	(Kerz et al. 2023)
	LIME, Feature Importance	ML (general)	(Nnadi et al. 2026)
	SHAP	SVM	(Bilalpur et al. 2023)
	SHAP	Deep Learning	(Mumenin et al. 2025)
	SHAP	Deep Learning	(Routh and Singh 2025)
	SHAP	Ensemble (RF/XGBoost) Logistic Regression	(Seok et al. 2025)
	SHAP	Deep Learning	(Zhao et al. 2025)
	SHAP, ALE, Anchors	Ensemble (RF/AdaBoost) SVM, ANN	(Chatterjee et al. 2023)
	Syst. review	*Multiple*	(Byeon 2023)
	Syst. review	*Multiple*	(Trelles et al. 2025)
	Syst. review	*Multiple*	(Zhai et al. 2025)
Stress	Counterfactual	Ensemble (RF/XGBoost)	(Jung and Lee 2024)
	LIME	Deep Learning	(Kerz et al. 2023)
	SHAP	Ensemble (RF/XGBoost)	(Watanabe and Ohno 2021)

Continued on next page

Table 2 – *continued*

Domain	XAI Method	ML Algorithm	Ref.
	SHAP	Ensemble (RF/XGBoost)	(Bozorgmehr and Weltermann 2023)
	SHAP	ML (general)	(Shikha et al. 2024)
	SHAP	Ensemble (RF/XGBoost) SVM Logistic Regression	(Tariq et al. 2025)
	LIME, Feature Importance	ML (general)	(Nnadi et al. 2026)
	SHAP	Deep Learning	(Mumenin et al. 2025)
Well-being	Attention, Inherent interp.	Deep Learning	(Paul et al. 2025)
	SHAP	Ensemble (RF/XGBoost) Logistic Regression	(Yeh et al. 2025)
	LIME	Deep Learning	(Thushari et al. 2023)
	SHAP	Ensemble (RF/XGBoost)	(Prinzi et al. 2025)
	SHAP	Ensemble (RF/XGBoost)	(Song et al. 2025)
Anxiety	Inherent interp.	Deep Learning + Decision Tree	(Liu et al. 2024)
	LIME	Deep Learning	(Kerz et al. 2023)
	SHAP, LIME	Deep Learning SVM	(Karim et al. 2024)
	Syst. review	*Multiple*	(Zhai et al. 2025)
Autism / ASD	Feature Importance	SVM	(Biswas et al. 2021)
	Feature Importance	Deep Learning	(Supekar et al. 2022)
	Feature Importance	Deep Learning	(Lee and Yoo 2025)
	LIME	Fuzzy/Rule-based	(Albahri et al. 2023)
Bipolar Disorder	LIME	Deep Learning	(Kerz et al. 2023)

Continued on next page

Table 2 – *continued*

Domain	XAI Method	ML Algorithm	Ref.
	Syst. review	*Multiple*	(Zhai et al. 2025)
Schizophrenia	LIME	Deep Learning	(Saadatinia and Salimi-Badr 2024)
	Inherent interp.	Deep Learning	(Chang et al. 2020)
Suicide Risk	Attention	Deep Learning	(Park and Sakong 2026)
	LIME	ML (general)	(de Oliveira et al. 2025)
ADHD	LIME	Deep Learning	(Kerz et al. 2023)
	SHAP	Ensemble (RF/XGBoost)	(Lee et al. 2025)
Mental Health (general)	Attention, Inherent interp.	Deep Learning	(Paul et al. 2025)
	Counterfactual	Ensemble (RF/XGBoost)	(Jung and Lee 2024)
Cognitive Load	Grad-CAM	Deep Learning	(Hudon et al. 2021)
Loneliness	Inherent interp.	Deep Learning + Decision Tree	(Liu et al. 2024)
Emotional States	SHAP, LIME	Ensemble (RF/XGBoost)	(Houssein et al. 2024)
PTSD	SHAP	Ensemble (RF/XGBoost)	(Wang and Xu 2025)
Sleep Disorder	Attention, Counterfactual, SHAP	Deep Learning	(Ahadian et al. 2024)
Personality Disorder	NLP Explainability	Deep Learning	(Patel and Johnson 2025)

Continued on next page

Table 2 – *continued*

Domain	XAI Method	ML Algorithm	Ref.
Personality Traits	SHAP, Integrated Gradients, PDP	Ensemble (RF/XGBoost/LGBM) SVM Deep Learning	(Panfilova and Turdakov 2024)
	SHAP	Logistic Regression Random Forest Neural Network	(Ramon et al. 2021)
Creativity	Grad-CAM	Deep Learning	(Panfilova et al. 2024)
—	*Syst. review*	*Multiple*	(Joyce et al. 2023)

4.2.1 Feature Importance-Based Methods: SHAP

SHAP (Lundberg and Lee 2017) is the dominant XAI method in the reviewed literature, being employed in 24 out of 42 studies (57%). The method is grounded in Shapley values from cooperative game theory and estimates the contribution of each feature to a model's prediction at both the global and local levels of explanation.

In studies on *depression*, SHAP is primarily applied in conjunction with ensemble and gradient boosting methods. Seok et al. (2025) used XGBoost combined with SHAP to analyse acoustic correlates of depression in the MODMA audio corpus, identifying predominantly spectral and timbral features, along with selected prosodic cues, as key predictors. Mumenin et al. (2025) developed a hybrid CNN+LSTM model (DDNet) for detecting depression among students and applied SHAP to interpret the contribution of individual questionnaire features. Routh and Singh (2025) used SHAP to predict depression risk among HIV-positive patients, where the method helped identify both immunological and psychosocial predictors.

In the domain of *stress*, Bozorgmehr and Weltermann (2023) applied XGBoost+SHAP to data from the large-scale national epidemiological study DEGS1 ($n = 5\,801$) to identify protective factors against chronic stress, demonstrating that good general state of health, social support, and life satisfaction have the strongest negative contributions to stress prediction. Watanabe and Ohno (2021) used LightGBM combined with SHAP on corporate survey data (Japan, $n = 6\,166$), showing that sense of coherence, quality of sleep, and job demand are leading predictors of perceived stress.

In the context of *psychological well-being*, Yeh et al. (2025) applied SHAP to an ensemble of five algorithms (XGBoost, Random Forest, LightGBM, etc.) on a Taiwanese national survey ($n = 10\,712$), demonstrating that family relationship satisfaction, health status, and interpersonal relationship satisfaction are the primary determinants of subjective well-being. Song et al. (2025) used LightGBM combined with SHAP to examine subjective well-being in urban China ($n = 1\,369$), identifying psychological security as the central determinant and revealing nonlinear threshold effects of income, with the strongest positive impact observed at middle-income levels.

For *PTSD*, Wang and Xu (2025) applied Random Forest combined with SHAP and Partial Dependence Plots (PDPs) to survey data ($n = 903$), identifying psychological resilience as the primary protective factor and revealing complex nonlinear relationships as well as interaction effects in which resilience buffers the impact of recent traumatic events, particularly among younger respondents.

In the domain of *personality traits and cognitive abilities*, Panfilova and Turdakov (2024) applied SHAP, Integrated Gradients, and Partial Dependence Plots to models for diagnosing Big Five personality traits and cognitive abilities using data from the VKontakte social network ($n = 1\,358$ users; 753 252 posts), indicating that emotional tone (joy, surprise, anger, fear) and digital footprints are key predictors of Big Five traits, logical reasoning, and verbal intelligence.

4.2.2 Local Explanation Methods: LIME

LIME (Ribeiro et al. 2016) is used in 10 out of 42 studies (24%), typically either as a complement to SHAP or as a standalone method in tasks requiring explanations of individual predictions. The method approximates the behaviour of a black-box model in the vicinity of a specific input instance using a locally interpretable model (linear regression).

Kerz et al. (2023) applied LIME in combination with a BiLSTM model to detect mental disorders (ADHD, anxiety, bipolar disorder, and depression) from Reddit texts; the method enabled the identification of lexical markers most relevant to each condition. Saadatinia and Salimi-Badr (2024) employed LIME within an EEG-based schizophrenia classification framework with generative data augmentation (WGAN-GP and VAE), using it to identify the most important regions in input spectrograms contributing to positive and negative diagnostic outcomes. de Oliveira et al. (2025) applied LIME (alongside another method) in a suicide risk screening system and conducted a user study with 78 mental health professionals, finding that the presence of LIME explanations significantly increases practitioner trust in the automated system. Nnadi et al. (2026) applied both hierarchical LIME (H-LIME) and SHAP to predict depression risk among students ($n = 27\,901$), where SHAP provided consistent global feature importance, while H-LIME offered complementary fine-grained insights by capturing threshold-dependent and subgroup-specific variations in feature effects.

4.2.3 Feature Importance: Embedded and Permutation-Based Methods

Seven studies (17%) employ feature importance methods that are either embedded within the learning algorithm (e.g., Gini impurity in decision trees and random forests) or based on permutation techniques. Permutation-based methods estimate feature importance by randomly shuffling feature values and measuring the resulting decrease in model performance. In contrast to SHAP and LIME, these approaches provide global explanations without local granularity and are computationally less demanding.

Supekar et al. (2022) used neural network-based feature importance analysis to identify functional neuroimaging markers distinguishing males and females with autism in a multi-site fMRI dataset (ABIDE + Stanford + CMI-HBN, $n = 773$), revealing sex-specific patterns of functional connectivity. Biswas et al. (2021) employed multiple complementary methods — feature importance analysis, correlation matrix visualisation, and hierarchical clustering — to interpret an SVM-based autism classifier for toddlers ($n = 1\,054$) using the Toddler Autism Dataset, constructing a multi-level explanation spanning from population-level patterns to individual-level insights. Zhu et al. (2025) used permutation feature importance (PFI), complemented by SHAP analysis, to interpret an XGBoost-based EEG depression classification model built on audio-stimulus-evoked three-channel EEG data, identifying the minimum value of power spectral density (PSD), Sample Entropy (SampEn), and Rényi Entropy as the key contributors to depression recognition.

4.2.4 Attention Mechanisms and Inherently Interpretable Models

Attention mechanisms are employed in three studies utilising transformer-based, multitask, and time-series deep learning architectures. This approach is *ante-hoc* — interpretability is embedded directly into the model architecture rather than applied post hoc.

Park and Sakong (2026) developed a unified multitask architecture for detecting suicide risk in social media texts, in which the attention mechanism supports decision sentence identification and cause sentence extraction, providing clinically interpretable explanations of risk. Paul et al. (2025) proposed a personalised mental health index (MHI) estimation system based on multitask GRU models using wearable sensor data (WESAD), where interpretability is embedded by design through temporal attention mechanisms and physiologically grounded score decomposition, enabling dynamic attribution of modality contributions over time. Ahadian et al. (2024) applied a multilevel approach to sleep disorder prediction, where LSTM, TCN, and Temporal Fusion Transformer models were equipped with temporal attention mechanisms to identify clinically relevant time windows, while SHAP and counterfactual explanations were additionally used for post-hoc interpretability ($n = 422$ patients).

4.2.5 Visualisation Methods: Grad-CAM and Saliency Maps

Gradient-based methods — Grad-CAM and saliency maps — are applied in studies involving EEG signals and neuroimaging, where it is necessary to identify temporally or spatially relevant regions of the signal that contribute to model predictions.

Hudon et al. (2021) conducted a controlled experiment ($n = 19$) to evaluate the impact of XAI visualisations (Integrated Gradients and Grad-CAM) on cognitive load and user confidence in an image classification task, showing that visualisation design factors, including adjacency and morphological clarity, significantly influence cognitive effort. Panfilova et al. (2024) applied Grad-CAM to analyse model outputs in creativity assessment using convolutional neural networks (CNNs) on drawings from the TCT-DP test ($n = 1\,823$); activation visualisations (ResNet-18, MobileNet V2, etc.) showed that the model focuses on the same regions of the drawings as human experts following TCT-DP methodological guidelines. Lee and Yoo (2025) applied CNN-based activation mapping for autism screening from children's drawings (273 sketches by 94 children), where salient visual regions were interpreted as diagnostically meaningful drawing features.

4.2.6 Specialised and Hybrid Methods

A number of studies employ non-standard or composite XAI approaches tailored to specific tasks.

Patel and Johnson (2025) developed a hybrid NLP approach with phase-wise classification for detecting Narcissistic Personality Disorder (NPD) in text, where interpretability was achieved through rule-based patterns and linguistic markers corresponding to the phases of narcissistic dynamics (idealisation, devaluation, discard, and hoovering). Chang et al. (2020) applied an Explainable Deep Neural Network

(EDNN) for schizophrenia diagnosis using structural MRI data, leveraging spatial localisation of informative brain voxels as a form of interpretability. Jung and Lee (2024) proposed a personalised mental health recommendation system using multi-objective counterfactual explanations based on Random Forest to generate realistic lifestyle interventions (EMS and mobile context data; DeepStress dataset, $n = 726$ observations). Bilalpur et al. (2023) applied SHAP to explain an SVM classifier for predicting maternal history of depression from multimodal audio–video recordings of mother–child interactions ($n = 148$ dyads), showing that maternal speech acoustic features were the most informative predictors. Karim et al. (2024) developed a hybrid CNN–SVM model (AnxPred) for predicting anxiety among university students, applying SHAP and LIME to explain predictions on the public MHP dataset ($n = 2\,028$), and demonstrating that GAD-7 elements are among the key predictors of anxiety. Lee et al. (2025) used SHAP to analyse bias in a Random Forest model predicting academic performance of students with ADHD ($n = 277$), revealing that behavioural traces from learning platforms may misrepresent actual competence and lead to systematic underestimation of performance compared to neurotypical students.

Overall, the choice of XAI method in the reviewed studies is determined by three main factors: the underlying model architecture (ensembles → SHAP/feature importance; neural networks → attention/Grad-CAM), data modality (text → LIME/attention; EEG/imaging → Grad-CAM/feature importance; tabular data → SHAP), and the required level of explanation (local vs. global). Figure 7 presents the aggregated distribution of XAI methods across psychological domains in the form of a stacked bar chart. SHAP is the dominant method across all domains, particularly in studies of depression ($n = 9$) and stress ($n = 6$). LIME ranks second and is also widely used in depression ($n = 4$) and anxiety ($n = 3$) research. Grad-CAM and Integrated Gradients are concentrated in domains involving multimodal data (EEG, imaging), including depression and stress, autism spectrum disorder, creativity assessment, and personality trait analysis. Attention-based methods are primarily associated with NLP tasks related to stress, anxiety, and suicide risk.

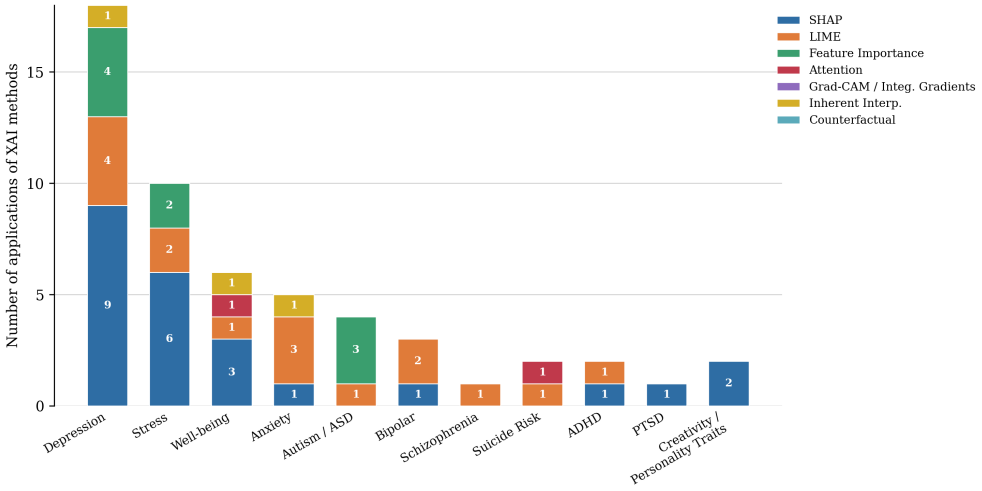


Figure 7 Stacked distribution of XAI methods across psychological domains ($n = 42$). The height of each bar represents the total number of XAI method applications within a domain; colours indicate individual methods. A single study may cover multiple domains and methods.

4.3 RQ3: Which Machine Learning Models Are Used and What Are the Characteristics of the Datasets?

Machine Learning Algorithms

Figure 8 presents the distribution of machine learning (ML) algorithms by class across the 42 included studies (a single study may employ multiple models). Deep learning architectures (CNN, LSTM, Transformer, etc.) are the most prevalent ($n = 19$, 50%), followed by ensemble methods (Random Forest, XGBoost, LightGBM; $n = 14$, 37%) and other or hybrid approaches ($n = 7$, 18%). Support Vector Machines (SVM) and hybrid models based on them appear in six studies (16%), while fuzzy and rule-based systems are used in only one study (3%).

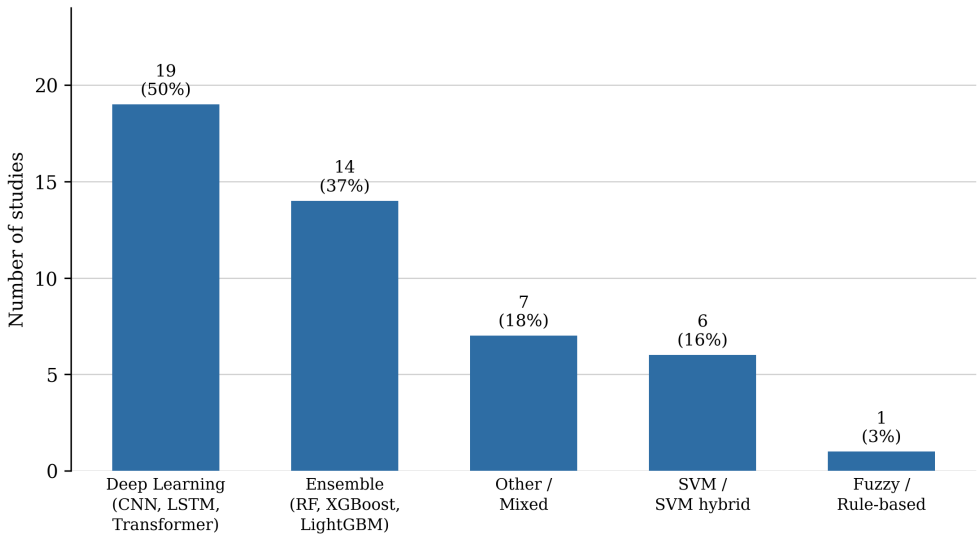


Figure 8 Distribution of ML algorithms by class across the included studies. A single study may employ multiple models. Deep learning architectures and ensemble methods dominate.

Data Modalities

Self-report questionnaires and surveys are the most commonly used modality (n = 24, 57%), reflecting the dominant role of psychometric assessment in psychological research. Text data from social media and clinical documentation (NLP-based approaches) are used in 7 studies (17%), images (drawings, neuroimaging) in 7 (17%), EEG in 5 (12%), speech audio in 5 (12%), and wearable sensors in 4 (10%). One study (Supekar et al. 2022) used fMRI data. Several studies combine multiple modalities, such as audio and EEG (Zhao et al. 2025), or wearable sensor data with self-reported emotional labels (Houssein et al. 2024).

Datasets and Availability

In terms of data availability, 15 studies (36%) used publicly available datasets, 13 (31%) relied on proprietary corporate or clinical data, 6 (14%) used mixed sources, and in 8 cases availability was not specified.

Among publicly available datasets, the most frequently used include Reddit-based datasets (e.g., SWMH, Dreaddit, and Kaggle collections) for detecting depression, anxiety, and suicide risk from social media texts (Kerz et al. 2023; Park and Sakong 2026); MODMA, an audio corpus for depression screening from speech (n = 52 participants) (Seok et al. 2025); DEGS1, a German national health survey (n = 5 801 participants) (Bozorgmehr and Weltermann 2023); and ABIDE, an fMRI dataset for autism research (n = 773 participants) (Supekar et al. 2022).

Sample Sizes

Sample sizes vary substantially across studies. The smallest samples are typically found in laboratory-based physiological studies (e.g., $n = 30$ for stress classification using wearable biosensors (Shikha et al. 2024), and $n = 52$ in the MODMA audio study (Seok et al. 2025)). The largest datasets are derived from social media, with over 150 000 posts used in suicide risk detection (Park and Sakong 2026). The EnvBodySens dataset (Houssein et al. 2024) includes 82 729 samples from only 30 participants, illustrating a common pattern in sensor-based research: small participant cohorts with high-frequency measurements. Survey-based studies typically occupy an intermediate range, from approximately 200 to over 10 000 participants. This variability complicates direct comparison of model performance across studies.

Languages and Geographic Coverage

The vast majority of studies are conducted on English-language data. Exceptions include Chinese (2 studies (Liu et al. 2024; Song et al. 2025)), German (1 (Bozorgmehr and Weltermann 2023)), Japanese (1 (Watanabe and Ohno 2021)), Brazilian Portuguese (1 (de Oliveira et al. 2025)), and Russian (1 (Panfilova and Turdakov 2024)). Geographic coverage is explicitly reported only in a subset of studies (e.g., the United States, China, Germany, Taiwan, Brazil, and Russia), while many studies do not specify the country of data collection, indicating a lack of geographic transparency in the field.

4.4 RQ4: Challenges, Limitations, and Future Directions of XAI in Psychology

Small and Imbalanced Datasets

The most frequently reported limitation is the insufficient size of datasets. Physiological studies (EEG, audio) are particularly vulnerable, with several works relying on samples ranging from 30 to 200 participants (Seok et al. 2025; Zhu et al. 2025; Zhao et al. 2025). Class imbalance — such as substantially fewer cases of suicide risk or schizophrenia compared to control groups — is documented in multiple studies (Saadatinia and Salimi-Badr 2024; Park and Sakong 2026) and is recognised as a source of model bias. To address imbalance, some studies apply synthetic resampling techniques (ADASYN, SMOTE); however, these approaches may introduce additional distributional bias (Routh and Singh 2025).

Limited Clinical Validation

Most studies evaluate models on a single dataset, often without external validation on independent samples from other institutions. Only a few studies involve mental health professionals (MHPs) in the evaluation process. For example, de Oliveira et al. (2025) conducted a user study involving psychologists and psychiatrists to assess the impact of XAI explanations on trust in the system. Several review papers (Trelles et al. 2025; Zhai et al. 2025) note that none of the studies analysed performed true external validation in independent clinical settings.

Lack of Standardised Evaluation of Explanations

The field of XAI in psychology lacks standardised metrics for evaluating explanation quality. Most studies rely on qualitative visualisations (e.g., SHAP plots or LIME explanations) without quantitative assessment of their fidelity, stability, or perceived usefulness for end users.

Biases and Ethical Risks

Several studies report gender and demographic biases in both datasets and models. The PTSD study by Wang and Xu (2025) included 98% male participants. Reddit-based datasets predominantly represent English-speaking users, limiting the generalisability of findings to other cultural contexts (Kerz et al. 2023; Park and Sakong 2026). Moreover, the cross-cultural validity of psychological constructs (e.g., stress, well-being) remains largely unexplored when transferring XAI models across different populations.

Future Directions

Analysis of future research sections reveals several key directions. *Multimodal XAI*: many studies emphasise the need to integrate multiple data sources (e.g., EEG + audio, wearable sensors + self-reports) with unified explanations (Zhao et al. 2025; Paul et al. 2025). *Longitudinal and multi-centre studies*: the predominance of cross-sectional designs limits the ability to capture temporal dynamics; longitudinal validation across institutions is needed (Nnadi et al. 2026; Trelles et al. 2025; Zhai et al. 2025). *Real-world deployment*: integration of models into clinical information systems and evaluation in real-world settings remain largely unexplored (de Oliveira et al. 2025; Park and Sakong 2026). *Human-centred evaluation*: there is a growing need for XAI evaluation metrics aligned with the needs of clinicians and patients, rather than relying solely on predictive performance (de Oliveira et al. 2025).

5 Discussion

5.1 Summary of Findings

This systematic review analyzed 42 peer-reviewed studies applying explainable artificial intelligence (XAI) methods to psychological constructs.

In response to **RQ1**, depression emerges as the most extensively studied domain ($n = 13$, 31% of all studies), followed by stress ($n = 8$, 19%) and well-being ($n = 5$, 12%). Anxiety disorders and autism spectrum disorder (ASD) are each represented by four studies (11%). Domains such as sleep disorders, PTSD, personality disorders, and cognitive load are represented by single studies, indicating substantial gaps in the coverage of psychological constructs.

Regarding **RQ2**, SHAP is identified as the dominant XAI method ($n = 24$, 57%), followed by LIME ($n = 10$, 24%). The choice of method is largely determined by the underlying model architecture and data modality: attention mechanisms are predominantly employed in NLP tasks, while gradient-based methods are used in EEG and neuroimaging studies.

Findings related to **RQ3** indicate that self-report questionnaires and surveys constitute the most prevalent data modality (approximately 50% of studies), while EEG and audio data are used less frequently. Sample sizes vary substantially, ranging from 30 to over 150,000 records. Where data availability is reported, it is distributed approximately equally between publicly available and proprietary datasets.

The synthesis of findings for **RQ4** highlights several recurring limitations: small and imbalanced datasets, lack of external clinical validation, absence of standardized metrics for evaluating explanation quality, and demographic biases. Authors of the included studies consistently identify multimodal approaches, longitudinal designs, and human-centered evaluation as priority directions for future research.

5.2 Comparison with Related Reviews

The present review exhibits thematic overlap with existing XAI surveys in adjacent domains, yet differs in both focus and scope.

Černevičienė and Kabašinskas (2024) conducted a systematic review of XAI in the financial domain and reported similar findings regarding the dominance of SHAP and the lack of standardized metrics for evaluating explanations. However, financial applications are more oriented towards regulatory compliance (GDPR, MiFID II), whereas psychological applications foreground clinical safety and ethical considerations when working with vulnerable populations.

Compared to the systematic review of XAI applications in healthcare (Band et al. 2023), psychological studies demonstrate substantially greater heterogeneity in data modalities (ranging from EEG to social media texts and drawings), smaller sample sizes, and greater subjectivity in target variables (psychological constructs are measured indirectly through self-report or behaviour).

The existence of a dedicated systematic review on XAI for depression assessment (Macaulay et al. 2026) aligns with our finding that this domain is the most extensively studied. The key distinction of the present review lies in its coverage of a broader spectrum of psychological disorders. An analogous conclusion is reached in comparing our work with a review focused on the use of XAI in neurocognitive disorder diagnosis (Lazarjani et al. 2025), which addresses an adjacent but narrower domain.

5.3 Implications

For psychological theory. The findings of this review point to a dual role of XAI in relation to psychological theory. On the one hand, several included studies demonstrate *convergent validity*: XAI methods reproduce theoretically expected predictors — sleep disturbances and anhedonia in depression, behavioural rigidity in autism — thereby indirectly supporting the plausibility of the model. On the other hand, several studies report *unexpected* feature importance profiles that diverge from established theoretical assumptions. This raises a productive question for psychological science: is the divergence attributable to a model artefact — or does it signal a limitation of existing theory? Systematic comparison of XAI-derived explanations with theoretical predictions could emerge as a distinct research direction, bridging machine learning methodologies with psychological theory. Furthermore, the dominance of correlational

rather than causal methods across the corpus indicates the need for caution in theoretical interpretation: feature importance in a model is not equivalent to causal role in a psychological process.

For researchers. The review identifies several methodological gaps. First, a number of psychologically and clinically relevant constructs — PTSD, personality disorders, and cognitive load — are represented by single studies despite their clinical significance. Second, cross-sectional designs predominate; longitudinal studies tracking the dynamics of psychological states and the stability of XAI explanations over time are extremely rare. Third, there is no consensus on the choice of metrics: researchers are encouraged to apply established explanation quality measures (e.g., faithfulness, stability, comprehensibility) in addition to standard ML performance metrics.

For practitioners and clinicians. XAI explanations can serve as instruments for hypothesis testing about psychological mechanisms: for example, confirmation of the role of sleep disturbances as a key predictor of depression is consistent with clinical theory. Caution is nonetheless required in interpretation: feature importance in a model reflects statistical association, not causal relationship. The single study that involved actual practitioners in the evaluation of an XAI system (de Oliveira et al. 2025) demonstrated a positive effect of explanations on trust in the system; however, full-scale randomised clinical trials of XAI-assisted decision support tools in psychology are yet to be conducted.

For AI developers. The findings suggest that XAI systems should be designed with sensitivity to the specifics of the psychological domain: overly technical explanations should be avoided, visualisations should be adapted to the expertise level of the end user (clinician, researcher, or patient), and mechanisms should be provided that allow the specialist to critically evaluate and, if necessary, contest the automatic recommendations of the system.

5.4 Limitations

The present review has several limitations. First, the absence of institutional access to Scopus means that some relevant publications indexed exclusively in that database may not have been captured. Web of Science was included as an additional source (April 2026), which partially compensates for this limitation. Second, the substantial initial number of records from PubMed (n = 1,860) and IEEE Xplore ($n \approx$ 1,538 All Metadata) includes a considerable proportion of closed-access publications. Only open-access records were admitted to screening, which may have led to the omission of relevant studies available exclusively through institutional subscriptions. Third, records from ACM Digital Library were collected manually, which may have introduced selection bias. Fourth, the search was restricted to publications in English, potentially excluding relevant work in other languages. Fifth, the supplementary manual search via Google Scholar, Semantic Scholar, and ResearchGate is not systematic and cannot guarantee comprehensive coverage of the existing literature; moreover, source selection during manual search is inevitably subject to researcher preferences, introducing additional selection bias.

Finally, the field of XAI in psychology is developing exceptionally rapidly; studies published or deposited in preprint archives after March 2026 are not covered by this review.

5.5 Future Directions

Based on the synthesis of responses to RQ4, the following priority directions for future research can be identified.

1. **Standardisation of explanation evaluation.** There is a need to develop widely accepted, psychologically valid metrics for assessing the quality of XAI explanations (e.g., perceived utility, clinical trust, diagnostic value) that go beyond technical performance indicators.
2. **Data scaling and diversification.** Large, multicentre, representative datasets spanning diverse cultures, languages, and demographic groups are critically needed. Open repositories (analogous to ABIDE for autism) are required for tasks related to stress, PTSD, and personality disorders.
3. **Multimodal XAI.** Integrating physiological signals (EEG, wearable sensors), speech, and text data with a unified explanation layer has the potential to substantially improve the diagnostic accuracy and clinical applicability of XAI systems.
4. **Longitudinal and intervention-based designs.** Cross-sectional studies predominate in the corpus; a shift to longitudinal designs will allow tracking of the dynamics of psychological states and evaluation of the stability of XAI explanations over time.
5. **Human-centered evaluation with clinician involvement.** Collaborative research involving psychologists, psychiatrists, and patients is essential for assessing the real-world utility of XAI tools and their integration into clinical decision-making protocols.
6. **Generative LLMs and hybrid explanations.** Text-based explanations generated by large language models offer the prospect of producing explanations adapted to the knowledge level of the end user. Systematic evaluation of this approach in the psychological domain has yet to be conducted.

5.6 Emerging Trends (2025–2026)

A substantial proportion of the included studies ($n = 19$, 45%) were published in 2025–2026, allowing the identification of several emerging trends not captured in earlier reviews.

Multimodal XAI.

Recent work demonstrates a shift from unimodal to multimodal architectures: the integration of EEG and audio (Zhao et al. 2025) and the combination of wearable sensors with self-reported emotional labels (Houssein et al. 2024). Explanations, however, are still constructed within individual modalities; a unified cross-modal XAI layer remains an unresolved challenge.

Bias and Fairness (Fairness-Aware XAI).

Lee et al. (2025) are among the first in the psychological domain to apply XAI specifically for *auditing model bias*, rather than solely for interpreting predictions. This approach — using SHAP as a diagnostic tool for model unfairness — reflects a broader trend of responsible AI in high-stakes domains.

Counterfactual Explanations as Intervention Tools.

Jung and Lee (2024) demonstrate that counterfactual explanations can go beyond answering "what should have been changed" to providing concrete, clinically actionable recommendations for patient lifestyle modifications — a first step towards XAI as an instrument for personalised interventions.

5.7 Methodological Recommendations for XAI Research in Psychology

Based on the synthesis of the included studies and the authors' experience in this domain, we propose three recommendations addressed to researchers applying XAI in psychological contexts. Table 3 summarises common issues for each class of XAI method and concrete strategies for addressing them; detailed justification for each point is provided in the paragraphs below.

Table 3: Common issues in XAI methods for psychological research and mitigation strategies.

Method / approach	**Common issue**	**Possible mitigation strategies**
SHAP (TreeSHAP, KernelSHAP)	Instability of top features on small samples; interpreting correlations as causal relationships	Bootstrap / repeated cross-validation with tracking of top-N features; confidence intervals for SHAP values; ablation: removal of key features and verification of accuracy drop
LIME	High variance of explanations due to stochastic sampling; sensitivity to neighbourhood size	Multiple runs and averaging of explanations; verification of surrogate model accuracy on held-out data; comparison with SHAP to identify discrepancies
Grad-CAM / Attribution maps	Saliency maps may highlight statistically active but psychologically irrelevant input regions	Masking / removal of high-importance regions: accuracy should decrease; assessment of map stability under small input perturbations; comparison with expert criteria of the assessment methodology

Table 3 – *continued*

Method / approach	**Common issue**	**Possible mitigation strategies**
Counterfactual explanations	Proposed changes may be unrealistic or may not cause the predicted class to change	Validity check: changes actually alter the prediction; verification of minimality and realism of changes relative to the original instance
Inherent interpretability (decision tree, logistic regression)	Model simplicity reduces predictive accuracy; risk of overestimating interpretability with deep trees	Explicit comparison of accuracy with a black-box model; limitation of tree depth and verification of rule stability across different data splits
Example-based explanations	Selected prototypes may not reflect group diversity; proximity in feature space may be non-intuitive	Assessment of distance between prototype and explained instance; verification of predictive consistency: similar instances should receive similar predictions
Chain-of-Thought in LLMs	Plausible reasoning with incorrect or non-falsifiable prediction mechanism (plausibility $\neq$ faithfulness)	Comparison with an independent XAI method (SHAP/LIME); stability check of CoT under repeated queries; ablation: removal of elements identified as key; expert clinical evaluation
Any XAI method (general)	Theoretically ungrounded top features accepted as a finding without further verification; convergence of multiple methods mistakenly taken as evidence of validity	Explicit discussion of alignment between explanations and psychological theory; replication on an independent sample as the gold standard; distinction between consistency and validity of explanations

1. Chain-of-Thought in Large Language Models Is Not a Reliable Source of Explanation.

The emergence of LLMs with Chain-of-Thought (CoT) reasoning creates an illusion of built-in interpretability: models generate coherent textual justifications for their predictions. However, prior work demonstrates that CoT reasoning does not always reflect the actual computational path of the model (Turpin et al. 2023; Lanham et al. 2023): models can produce correct predictions accompanied by incorrect or deliberately perturbed explanations, and the content of CoT changes substantially under different prompt formulations even when the final prediction remains unchanged. Thus, CoT provides *plausibility* but not *faithfulness* of the explanation in the sense of Jacovi & Goldberg (Jacovi and Goldberg 2020).

From a practical perspective, this implies that using LLMs to interpret psychological states does not eliminate the need for verification using standard XAI methods

— and introduces additional computational costs. Researchers using LLMs in psychological applications are encouraged to treat CoT as an auxiliary narrative tool rather than as independent evidence of the prediction mechanism.

The following approaches can be proposed for validating CoT reasoning.

- **Comparison with an independent XAI method.** Applying a model-agnostic method (SHAP, LIME) to the same task and comparing the identified factors with those referenced in the CoT explanation. Systematic divergence between what the model states and what an independent method shows serves as a signal of post-hoc rationalisation rather than genuine explanation.
- **Assessment of reasoning stability.** Repeated queries with an identical prompt under non-zero generation temperature allow evaluation of the stability of the reasoning chain. High variability in CoT while the final prediction remains stable indicates that the reasoning is not a necessary condition for the prediction.
- **Expert clinical evaluation.** Involving a specialist in psychology or psychiatry to assess the clinical plausibility of the CoT reasoning chain represents the most direct, if resource-intensive, form of verification. Such evaluation is especially important in applied clinical decision support systems, where an erroneous but persuasive explanation may influence the practitioner's actions.
- **Ablation of CoT reasoning components.** If CoT identifies specific elements of the input data (symptoms, questionnaire responses, speech patterns) as critical to the prediction, their sequential removal from the input should produce a noticeable drop in model accuracy. The absence of such a drop indicates that the reasoning describes not the actual dependency but merely a plausible version of it. This approach constitutes an adaptation of the classical *sufficiency* test (Jacovi and Goldberg 2020) to the LLM architecture and allows quantitative assessment of the extent to which a CoT explanation genuinely carries information about the prediction mechanism.
- **Comparison of explanations across different LLMs.** Running the same task on multiple models (e.g., GPT, Gemini, LLaMA) and comparing the resulting CoT chains allows evaluation of the stability of explanations across model choices. Convergent results constitute a weak argument in favour of explanation faithfulness, but are not evidence of it: if models are trained on similar corpora or have undergone analogous optimisation (RLHF), their explanations may systematically reproduce the same plausible narratives regardless of the true prediction mechanism. The extent to which biases in training data and reward hacking constitute a common source of conflict between plausibility and faithfulness remains an open question requiring further investigation.

2. Small Samples Increase the Risk of Artefactual Explanations: Stability Verification Is Essential.

Psychological research often relies on relatively small datasets, particularly in physiological studies (EEG, audio). This introduces a critical risk: a model trained on small data may overfit to artefacts of the specific sample, and the XAI method will faithfully "explain" precisely those artefacts. SHAP values obtained on one train/test split may change substantially on a different split; the narrative "feature X turned out to be the

key predictor of condition Y" is frequently reproduced in published articles without caveats about instability.

The natural desire to find an explanation — a characteristic research motivation, especially in a domain where the mechanisms of mental processes remain contested — combined with small sample sizes, increases the likelihood of treating an unstable artefact as a meaningful finding. The temptation to support a finding with corroborating literature is further complicated by publication bias and inconsistency across studies: finding publications that align with a discovered feature is not difficult, but this reduces the evidentiary value of such alignment as an argument.

To mitigate this risk, it is recommended to distinguish between the *consistency* of explanations and their *validity*. High consistency is a necessary but not sufficient condition: multiple methods may converge on the same artefact. Specific verification strategies depend on the class of XAI method used.

- **Feature importance–based methods (SHAP, permutation).** Assess the stability of top features using bootstrapping or repeated cross-validation; report confidence intervals for SHAP values; train multiple models on the same dataset and compare importance profiles; introduce small perturbations to the input data and evaluate result stability. The key validity check is the removal of features identified as important: model accuracy should decrease noticeably.
- **Local surrogate methods (LIME, KernelSHAP).** These methods involve sources of randomness (perturbation sampling, random subsets), so repeated runs on the same data allow estimation of explanation variance. For surrogate models, the key validity criterion is the approximation accuracy of the original model on held-out data.
- **Counterfactual explanations.** Verify whether the proposed changes actually cause the model's prediction to change, and assess their realism and minimality relative to the original instance.
- **Attribution maps / saliency maps (Grad-CAM, Integrated Gradients).** Apply masking or progressive removal of the most important input regions: a substantial degradation in prediction performance upon removal of "important" regions is evidence of map correctness. Additionally assess the stability of maps under small perturbations.
- **Example-based explanations.** Evaluate validity through the proximity of selected examples to the instance being explained in feature space, and through predictive consistency: similar instances should receive similar predictions.

Whenever possible, explanations should be reproduced on an independent dataset, which remains the most reliable criterion for their generalisability.

3. Theoretical Grounding as a Criterion for Trust in Explanations.

For XAI results to have scientific and practical relevance, they must be interpreted in alignment with established psychological theory and empirical evidence. Such alignment constitutes an essential part of the validation of the obtained results.

If a model identifies sleep disturbances as key predictors of depression or social isolation as a predictor of suicide risk, this is consistent with well-established psychological frameworks (Beck's cognitive model, Joiner's interpersonal theory of suicide) and thereby increases trust in the explanation. For instance, in a study predicting Big Five personality traits and intelligence level from users' social network texts (Panfilova and Turdakov 2024), the contribution of age to the model-predicted level of verbal intelligence is consistent with empirical findings in this area.

The inverse logic also holds: XAI can serve as a tool for generating theoretical hypotheses. A Grad-CAM analysis in a study on automated creativity assessment (Panfilova et al. 2024) illustrates this potential: the regions of participants' drawings highlighted as important largely overlapped with the criteria of the expert TCT-DP methodology — and criteria that were clearly formulated and had high inter-rater agreement were "learned" more readily by the model. It is precisely this alignment with theoretically grounded features that lends reliability to the interpretation of *non-overlapping* regions: divergences between model activations and expert criteria cease to be artefacts and become meaningful signals — a basis for formulating new research hypotheses about aspects of creativity that lie beyond the scope of existing theory.

Similarly, Lee and Yoo (2025) identified the eye region as an additional salient area in a drawing-based autism screening model, and subsequently incorporated this feature explicitly into the model architecture, which led to a measurable improvement in performance. This example illustrates how XAI-driven discovery can directly inform theoretical and empirical refinement.

One important caveat applies here: the successful automation of a methodology based on expert judgement is not equivalent to validation of the theoretical underpinnings of that methodology. We can only establish the extent to which the mechanisms learned by the model replicate the expert's decision process — not whether that process reflects an underlying psychological construct. This limitation should be stated explicitly in the discussion of results.

Top-ranked features that lack an obvious theoretical grounding warrant particular attention: they may point to a novel phenomenon requiring further investigation — but are equally likely to reflect a methodological artefact. Explicit discussion of the theoretical interpretability of the obtained explanations represents a useful norm for publications in this field.

6 Conclusion

This systematic literature review examined the application of explainable artificial intelligence (XAI) to psychological constructs and mental health care tasks, based on 42 peer-reviewed studies published between 2020 and 2026.

In response to **RQ1**, depression, stress, and psychological well-being emerge as the most extensively studied domains, while several clinically significant constructs — PTSD, personality disorders, and ADHD — remain substantially underrepresented.

Regarding **RQ2**, SHAP and LIME are identified as the dominant XAI methods. The choice of method is systematically associated with the underlying model architecture and data modality: attention mechanisms are characteristic of NLP tasks, while gradient-based methods are associated with EEG and neuroimaging.

The analysis of **RQ3** reveals a wide diversity of data sources (self-report questionnaires and surveys, social media text, EEG, audio, and images) and substantial variability in sample sizes, ranging from tens to hundreds of thousands of records; data availability is distributed approximately equally between open and proprietary sources.

Findings related to **RQ4** indicate that key unresolved challenges include small and imbalanced datasets, lack of external clinical validation, absence of standardized metrics for evaluating explanation quality, and demographic biases; future research is oriented towards multimodal XAI, longitudinal designs, and human-centered evaluation of systems.

Overall, the results suggest that XAI holds significant potential for enhancing transparency, trust, and clinical utility of AI systems in psychology and mental health care. Realizing this potential requires systematic efforts — from developing standards for explanation evaluation and open data repositories to conducting multicentre clinical trials of XAI-assisted decision support tools. Only through close collaboration between AI researchers, clinical psychologists, and experts in human–computer interaction can XAI transition from academic experimentation to tools that meaningfully improve psychological care.

For researchers entering this field, the recommendations outlined in Section 5.7 — namely, validation of Chain-of-Thought reasoning, control of explanation stability in small-sample settings, and reliance on psychological theory as a criterion for validity — may serve as a starting point for designing the methodology of future studies.

Declarations

Funding. The author(s) declare that financial support was received for the research and/or publication of this article. The research was supported by RF state assignment 0138-2024-0020.

Author contributions. All authors contributed equally to this work.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. Not applicable. This study is a systematic literature review of publicly available published research and did not involve human participants or animals.

Data availability. Full references for all 42 included studies are provided in the main text (Table 2). Complete Boolean search strings are provided in Appendix A. Screening decisions are available from the corresponding author upon reasonable request. No new datasets were generated during this study.

References

Adadi A, Berrada M (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahadian P, et al (2024) Adopting trustworthy AI for sleep disorder prediction: Deep time series analysis with temporal attention mechanism and counterfactual explanations. In: 2024 IEEE International Conference on Big Data (BigData). IEEE, https://doi.org/10.1109/BigData62323.2024.10825687

Albahri AS, Joudar SS, Hamid RA, et al (2023) Explainable artificial intelligence multimodal of autism triage levels using fuzzy approach-based multi-criteria decision-making and lime. International Journal of Fuzzy Systems 26(1):274–303. https://doi.org/10.1007/s40815-023-01597-9

American Psychological Association (2017) Ethical Principles of Psychologists and Code of Conduct. American Psychological Association, Washington, DC, including 2010 and 2016 amendments

Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Ayari H, Guetari R, Kraïem N (2025) Machine learning powered financial credit scoring: a systematic literature review. Artificial Intelligence Review 59(1):13. https://doi.org/10.1007/s10462-025-11416-2

Band SS, Yarahmadi A, Hsu CC, et al (2023) Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked 40:101286. https://doi.org/10.1016/j.imu.2023.101286

Bilalpur M, et al (2023) SHAP-based prediction of mother's history of depression to understand the influence on child behavior. In: Proceedings of the ACM International Conference on Multimodal Interaction (ICMI 2023). ACM, https://doi.org/10.1145/3577190.3614136

Biswas M, et al (2021) An XAI based autism detection: The context behind the detection. In: Research Challenges in Information Science. Springer, https://doi.org/10.1007/978-3-030-86993-9_40

Bozorgmehr A, Weltermann B (2023) Prediction of chronic stress and protective factors in adults: Development of an interpretable prediction model based on xgboost and shap using national cross-sectional degs1 data. JMIR AI 2:e41868. https://doi.org/10.2196/41868

Byeon H (2023) Advances in machine learning and explainable artificial intelligence for depression prediction. International Journal of Advanced Computer Science and Applications 14(6). https://doi.org/10.14569/ijacsa.2023.0140656

Černevičienė J, Kabašinskas A (2024) Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review 57:216. https://doi.org/10.1007/s10462-024-10854-8

Chang YW, Tsai SJ, Wu YF, et al (2020) Development of an al-based web diagnostic system for phenotyping psychiatric disorders. Frontiers in Psychiatry 11. https://doi.org/10.3389/fpsyt.2020.542394

Chatterjee S, Mishra J, Sundram F, et al (2023) Towards personalised mood prediction and explanation for depression from biophysical data. Sensors 24(1):164. https://doi.org/10.3390/s24010164

European Parliament, Council of the European Union (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). Tech. Rep. L 119, Official Journal of the European Union, article 22: Automated individual decision-making, including profiling

Houssein EH, Mohsen S, Emam MM, et al (2024) Leveraging explainable artificial intelligence for emotional label prediction through health sensor monitoring. Cluster Computing https://doi.org/10.1007/s10586-024-04804-w

Hudon A, et al (2021) Explainable artificial intelligence (XAI): How the visualization of AI predictions affects user cognitive load and confidence. In: Information Systems and Neuroscience (NeuroIS 2021). Springer, https://doi.org/10.1007/978-3-030-88900-5_27

Jacovi A, Goldberg Y (2020) Towards faithfully interpretable NLP systems: On the calibration of attention weights. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp 4203–4213

Jain S, Wallace BC (2019) Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Association for Computational Linguistics, Minneapolis, Minnesota, pp 3543–3556

Joyce DW, Kormilitzin A, Smith KA, et al (2023) Explainable artificial intelligence for mental health through transparency and interpretability for understandability. npj Digital Medicine 6(1). https://doi.org/10.1038/s41746-023-00751-9

Jung M, Lee U (2024) Counterfactual explanations in personal informatics for personalized mental health management. In: Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, https://doi.org/

10.1145/3675094.3678427

Karim MR, et al (2024) AnxPred: A hybrid CNN-SVM model with XAI to predict anxiety among university students. In: 2024 International Conference on Informatics, Electronics and Vision (ICIEV). IEEE, https://doi.org/10.1109/Informatics62280.2024.10900931

Kerz E, Zanwar S, Qiao Y, et al (2023) Toward explainable ai (xai) for mental health detection based on language behavior. Frontiers in Psychiatry 14. https://doi.org/10.3389/fpsyt.2023.1219479

Khan FS, Mazhar SS, Mazhar K, et al (2025) Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. Artificial Intelligence Review 58(8). https://doi.org/10.1007/s10462-025-11215-9

Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Tech. Rep. EBSE 2007-001, Keele University and Durham University

Lanham T, Chen A, Radhakrishnan A, et al (2023) Measuring faithfulness in chain-of-thought reasoning. In: arXiv preprint arXiv:2307.13702

Lazarjani HG, Zarei S, Aboutorabi F, et al (2025) Application of explainable artificial intelligence for predicting and diagnosing neurocognitive disorders using tabular data: A systematic review. Archives of Computational Methods in Engineering https://doi.org/10.1007/s11831-025-10471-w

Lee J, et al (2025) XAI reveals the causes of attention deficit hyperactivity disorder (ADHD) bias in student performance prediction. In: Proceedings of the 15th International Conference on Learning Analytics and Knowledge (LAK25). ACM, https://doi.org/10.1145/3706468.3706521

Lee TG, Yoo JH (2025) Interpretable autism screening via object-based analysis of child-drawn sketches. IEEE Access 13:211138–211150. https://doi.org/10.1109/access.2025.3642950

Liu F, Wang P, Hu J, et al (2024) A psychologically interpretable artificial intelligence framework for the screening of loneliness, depression, and anxiety. Applied Psychology: Health and Well-Being 17(1). https://doi.org/10.1111/aphw.12639

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30

Macaulay A, Din FU, Nguyen PA, et al (2026) Using explainable AI for assessment of depression: A systematic literature review. Next-Gen Healthcare pp 441–464

Mokheleli T (2025) Age-stratified mental health prediction using SHAP: An explainable artificial intelligence framework. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 14:e32910. https://doi.org/10.14201/adcaij.32910

Mokheleli T, Bokaba T, Mbunge E (2025) Explainable artificial intelligence for workplace mental health prediction. Informatics 12(4):130. https://doi.org/10.3390/informatics12040130

Mumenin N, Abu Yousuf M, Alassafi MO, et al (2025) Ddnet: A robust, and reliable hybrid machine learning model for effective detection of depression among university students. IEEE Access 13:49334–49353. https://doi.org/10.1109/access.2025.3552041

Nnadi LC, Isiwu CP, Ding D, et al (2026) Multi-level explainable ai for predicting student depression risk: Global, subgroup, and individual insights. IEEE Access 14:6271–6286. https://doi.org/10.1109/access.2026.3652631

de Oliveira AC, Azevedo JPC, Ruback L, et al (2025) Effect of explainable artificial intelligence on trust of mental health professionals in an ai-based system for suicide prevention. IEEE Access 13:60987–61005. https://doi.org/10.1109/access.2025.3556245

Page MJ, McKenzie JE, Bossuyt PM, et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71

Panfilova AS, Turdakov DY (2024) Applying explainable artificial intelligence methods to models for diagnosing personal traits and cognitive abilities by social network data. Scientific Reports https://doi.org/10.1038/s41598-024-56080-8

Panfilova AS, Valueva EA, Ilyin IY (2024) The application of explainable artificial intelligence methods to models for automatic creativity assessment. Frontiers in Artificial Intelligence https://doi.org/10.3389/frai.2024.1310518

Park K, Sakong W (2026) A unified multi-task architecture for suicide risk detection via attention-based decision and cause sentence extraction. IEEE Access 14:39154–39170. https://doi.org/10.1109/access.2026.3669351

Patel D, Johnson AP (2025) Detecting narcissistic personality disorder (npd): A hybrid regex and nlp-based ai approach with phase-aware classification. IEEE Access 13:185011–185024. https://doi.org/10.1109/access.2025.3624864

Paul A, Dhanushkodi K, Alphonse S (2025) The mental health index (mhi): A personalized multi-task deep learning framework for real-time and explainable mental well-being assessment. IEEE Access 13:203125–203144. https://doi.org/10.1109/access.2025.3637389

Prinzi F, Barbiero P, Greco C, et al (2025) Using ai explainable models and handwriting/drawing tasks for psychological well-being. Information Systems 127:102465. https://doi.org/10.1016/j.is.2024.102465

Ramon Y, Farrokhnia R, Matz SC, et al (2021) Explainable ai for psychological profiling from behavioral data: An application to big five personality predictions from financial transaction records. Information 12(12):518. https://doi.org/10.3390/info12120518

Ribeiro MT, Singh S, 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, pp 1135–1144, https://doi.org/10.1145/2939672.2939778

Routh A, Singh J (2025) Enhancing explainability of quantum-based stacking for mental health assessment: Predicting depression risk in aids patients. IEEE Access 13:206490–206503. https://doi.org/10.1109/access.2025.3639110

Saadatinia M, Salimi-Badr A (2024) An explainable deep learning-based method for schizophrenia diagnosis using generative data-augmentation. IEEE Access 12:98379–98392. https://doi.org/10.1109/access.2024.3428847

Seok KH, Shin J, Bae SM (2025) Exploring acoustic correlates of depression and preliminary screening models using xgboost and shap. Behavioral Sciences 15(12):1648. https://doi.org/10.3390/bs15121648

Shikha S, Sethia D, Indu S (2024) Optimization of wearable biosensor data for stress classification using machine learning and explainable ai. IEEE Access 12:169310–169327. https://doi.org/10.1109/access.2024.3463742

Song J, Pan F, Liu H (2025) Income, psychological security, and subjective well-being in urban china: a machine learning analysis with shap interpretation. BMC Psychology 13(1). https://doi.org/10.1186/s40359-025-03591-2

Supekar K, de Los Angeles C, Ryali S, et al (2022) Deep learning identifies robust gender differences in functional brain organization and their dissociable links to clinical symptoms in autism. The British Journal of Psychiatry 220(4):202–209. https://doi.org/10.1192/bjp.2022.13

Tariq R, Orozco-del Castillo MG, Zamir MT, et al (2025) Explainable artificial intelligence for predictive modeling of student stress in higher education. Scientific Reports 15(1). https://doi.org/10.1038/s41598-025-22171-3

Thushari PD, Aggarwal N, Vajrobol V, et al (2023) Identifying discernible indications of psychological well-being using ml: explainable ai in reddit social media interactions. Social Network Analysis and Mining 13(1). https://doi.org/10.1007/s13278-023-01145-1

Trelles A, Ruiz TF, Rojo AP (2025) Systematic review and meta-analysis of explainable machine learning models for clinical depression detection. Behavioral Sciences 15(11):1476. https://doi.org/10.3390/bs15111476

Turpin M, Michael J, Perez E, et al (2023) Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems 36. NeurIPS 2023

Wang J, Xu X (2025) Unveiling the drivers of ptsd: An interpretable machine learning approach with shap. In: Proceedings of the 2nd International Conference on Intelligent Computing and Data Analysis. ACM, pp 782–788, https://doi.org/10.1145/3772726.3772849

Watanabe S, Ohno Y (2021) Analysis of the importance for perceived stress using lightgbm and shap values. Japanese Journal of Applied IT Healthcare 16(1):21. https://doi.org/10.11204/ithc.16.21

Yeh MHS, Kuo CT, Hsieh SH, et al (2025) Determinants of subjective well-being in taiwan: A machine learning and shap analysis. medRxiv https://doi.org/10.1101/2025.10.05.25337343

Yeo WJ, Van Der Heever W, Mao R, et al (2025) A comprehensive review on financial explainable AI. Artificial Intelligence Review 58(6):189. https://doi.org/10.1007/s10462-024-11077-7

Zafar U, Wu F (2026) Methodological challenges in explainable AI for fraud detection: a systematic literature review. Artificial Intelligence Review 59(4):115. https://doi.org/10.1007/s10462-026-11516-7

Zhai L, Zhao M, Zhang J, et al (2025) A systematic review of eeg-based biomarkers for depression, anxiety, and bipolar disorder: trends in explainable artificial intelligence (xai). BMC Psychiatry 26(1). https://doi.org/10.1186/s12888-025-07589-3

Zhao Q, Jiang H, Wu Z, et al (2025) E-dann: An enhanced domain adaptation network for audio-eeg feature decoupling in explainable depression recognition. IEEE Transactions on Neural Systems and Rehabilitation Engineering 33:3647–3661. https://doi.org/10.1109/tnsre.2025.3608181

Zhu L, Wang R, Jin X, et al (2025) Explainable depression classification based on eeg feature selection from audio stimuli. IEEE Transactions on Neural Systems and Rehabilitation Engineering 33:1411–1426. https://doi.org/10.1109/tnsre.2025.3557275

Appendix A: Full Database Search Strings

This appendix provides the complete Boolean search strings used in each bibliographic database, as required by PRISMA 2020 reporting standards (item 6). All searches were conducted on 26 March 2026. The search period covered publications from 1 January 2015 to 26 March 2026; all searches were restricted to English-language records.

The search logic followed a three-block Boolean structure:

Block A (XAI / interpretability terms) `AND` **Block B** (ML / AI terms) `AND` **Block C** (psychological domain terms)

A.1 PubMed / MEDLINE

Search date: 26 March 2026 **Fields:** Title / Abstract **Records retrieved:** 1 860

```
(
  (
    "explainable artificial intelligence"[Title/Abstract] OR
    "explainable AI"[Title/Abstract] OR
    "XAI"[Title/Abstract] OR
    "interpretable machine learning"[Title/Abstract] OR
    "interpretable AI"[Title/Abstract] OR
    "model interpretability"[Title/Abstract] OR
    "model explainability"[Title/Abstract] OR
    "explainability"[Title/Abstract] OR
    "SHAP"[Title/Abstract] OR
    "Shapley"[Title/Abstract] OR
    "SHapley Additive exPlanations"[Title/Abstract] OR
    "LIME"[Title/Abstract] OR
    "Local Interpretable Model-agnostic Explanations"[Title/Abstract] OR
    "feature importance"[Title/Abstract] OR
    "counterfactual explanation"[Title/Abstract] OR
    "attention mechanism"[Title/Abstract] OR
    "saliency map"[Title/Abstract] OR
    "permutation importance"[Title/Abstract] OR
    "transparent AI"[Title/Abstract] OR
    "black-box explanation"[Title/Abstract]
  )
  AND
  (
    "mental health"[Title/Abstract] OR
    "depression"[Title/Abstract] OR
    "anxiety"[Title/Abstract] OR
    "stress"[Title/Abstract] OR
    "PTSD"[Title/Abstract] OR
    "post-traumatic stress"[Title/Abstract] OR
```

```
    "wellbeing"[Title/Abstract] OR
    "well-being"[Title/Abstract] OR
    "psychological"[Title/Abstract] OR
    "psychiatric"[Title/Abstract] OR
    "autism"[Title/Abstract] OR
    "ASD"[Title/Abstract] OR
    "schizophrenia"[Title/Abstract] OR
    "suicide"[Title/Abstract] OR
    "suicidal ideation"[Title/Abstract] OR
    "personality disorder"[Title/Abstract] OR
    "personality trait"[Title/Abstract] OR
    "cognitive load"[Title/Abstract] OR
    "bipolar"[Title/Abstract] OR
    "OCD"[Title/Abstract] OR
    "obsessive-compulsive"[Title/Abstract] OR
    "sleep disorder"[Title/Abstract] OR
    "resilience"[Title/Abstract] OR
    "mood disorder"[Title/Abstract] OR
    "ADHD"[Title/Abstract] OR
    "burnout"[Title/Abstract] OR
    "psycholinguistic"[Title/Abstract] OR
    "mental disorder"[Title/Abstract] OR
    "emotional"[Title/Abstract] OR
    "narcissistic"[Title/Abstract] OR
    "loneliness"[Title/Abstract] OR
    "grief"[Title/Abstract]
  )
)
AND ("2015/01/01"[Date - Publication] : "3000/12/31"[Date - Publication])
AND (English[Language])
```

Note: Of the 1 860 records retrieved, a substantial proportion are closed-access publications. Only open-access records were retained for screening (54 records). Closed-access studies were automatically excluded; this is acknowledged as a study limitation (Section 5.4).

A.2 IEEE Xplore

Two complementary searches were performed in IEEE Xplore.

A.2.1 Title / Abstract search (primary, 54 records)

Search date: 26 March 2026 **Fields:** Document Title + Abstract **Records retrieved:** 54

```
(
  (
```

```
  "Document Title":"explainable artificial intelligence" OR
  "Document Title":"explainable AI" OR
  "Document Title":"XAI" OR
  "Document Title":"interpretable machine learning" OR
  "Document Title":"SHAP" OR
  "Document Title":"LIME" OR
  "Document Title":"feature importance" OR
  "Document Title":"model interpretability" OR
  "Abstract":"explainable artificial intelligence" OR
  "Abstract":"explainable AI" OR
  "Abstract":"XAI" OR
  "Abstract":"interpretable machine learning" OR
  "Abstract":"SHAP" OR
  "Abstract":"LIME" OR
  "Abstract":"feature importance" OR
  "Abstract":"model interpretability"
)
AND
(
  "Document Title":"mental health" OR
  "Document Title":"depression" OR
  "Document Title":"anxiety" OR
  "Document Title":"stress" OR
  "Document Title":"PTSD" OR
  "Document Title":"wellbeing" OR
  "Document Title":"psychological" OR
  "Document Title":"psychiatric" OR
  "Document Title":"autism" OR
  "Document Title":"schizophrenia" OR
  "Document Title":"suicide" OR
  "Document Title":"personality disorder" OR
  "Document Title":"cognitive load" OR
  "Document Title":"bipolar" OR
  "Document Title":"sleep disorder" OR
  "Document Title":"emotion" OR
  "Document Title":"mood disorder" OR
  "Document Title":"ADHD" OR
  "Document Title":"burnout" OR
  "Document Title":"psycholinguistic" OR
  "Abstract":"mental health" OR
  "Abstract":"depression" OR
  "Abstract":"anxiety" OR
  "Abstract":"stress" OR
  "Abstract":"PTSD" OR
  "Abstract":"wellbeing" OR
```

```
    "Abstract":"psychological" OR
    "Abstract":"psychiatric" OR
    "Abstract":"autism" OR
    "Abstract":"schizophrenia" OR
    "Abstract":"suicide" OR
    "Abstract":"personality disorder" OR
    "Abstract":"cognitive load" OR
    "Abstract":"bipolar" OR
    "Abstract":"sleep disorder" OR
    "Abstract":"emotion" OR
    "Abstract":"mood disorder" OR
    "Abstract":"ADHD" OR
    "Abstract":"burnout" OR
    "Abstract":"psycholinguistic"
  )
)
Filters: Years 2015-2026; Content Types: Journals, Conference Publications
```

A.2.2 All Metadata search (supplementary, ≈1 538 records)

Search date: 26 March 2026 **Fields:** All Metadata **Records retrieved:** ≈1 538

```
(
  (
    "All Metadata":"explainable artificial intelligence" OR
    "All Metadata":"explainable AI" OR
    "All Metadata":"XAI" OR
    "All Metadata":"interpretable machine learning" OR
    "All Metadata":"SHAP" OR
    "All Metadata":"LIME" OR
    "All Metadata":"feature importance" OR
    "All Metadata":"model interpretability"
  )
  AND
  (
    "All Metadata":"machine learning" OR
    "All Metadata":"deep learning" OR
    "All Metadata":"neural network"
  )
  AND
  (
    "All Metadata":"mental health" OR
    "All Metadata":"depression" OR
    "All Metadata":"anxiety" OR
    "All Metadata":"stress" OR
    "All Metadata":"wellbeing" OR
```

```
    "All Metadata":"psychological" OR
    "All Metadata":"psychiatric"
  )
)
Filters: Years 2015-2026
```

Note: The All Metadata search (≈1 538 records) was used as a reference to ensure broad coverage. The majority of retrieved records are closed-access; only open-access records from this pool were carried forward. This query was treated as supplementary to the Title/Abstract search.

A.3 ACM Digital Library

Search date: 26 March 2026 **Fields:** Abstract **Records retrieved:** 11 (manual collection)

```
(
  (
    Abstract:("explainable artificial intelligence") OR
    Abstract:("explainable AI") OR
    Abstract:("XAI") OR
    Abstract:("interpretable machine learning") OR
    Abstract:("SHAP") OR
    Abstract:("Shapley") OR
    Abstract:("LIME") OR
    Abstract:("feature importance") OR
    Abstract:("model interpretability") OR
    Abstract:("model explainability") OR
    Abstract:("counterfactual explanation")
  )
  AND
  (
    Abstract:("mental health") OR
    Abstract:("depression") OR
    Abstract:("anxiety") OR
    Abstract:("stress") OR
    Abstract:("PTSD") OR
    Abstract:("wellbeing") OR
    Abstract:("well-being") OR
    Abstract:("psychological") OR
    Abstract:("psychiatric") OR
    Abstract:("autism") OR
    Abstract:("schizophrenia") OR
    Abstract:("suicide") OR
    Abstract:("personality disorder") OR
    Abstract:("cognitive load") OR
```

```
    Abstract:("bipolar") OR
    Abstract:("sleep disorder") OR
    Abstract:("emotion") OR
    Abstract:("mood disorder") OR
    Abstract:("ADHD") OR
    Abstract:("burnout") OR
    Abstract:("psycholinguistic") OR
    Abstract:("mental disorder") OR
    Abstract:("resilience") OR
    Abstract:("grief") OR
    Abstract:("narcissistic")
  )
)
Filters: Published 2015-2026;
         Content Types: Research Article, Short Paper, Conference Paper
```

Note: Automated bulk export from the ACM Digital Library requires institutional access not available to the authors. Records were collected manually by entering query terms and recording matching articles; this approach may introduce selection bias, which is acknowledged as a study limitation (Section 5.4).

A.4 Web of Science (WoS)

Search date: 10 April 2026 **Fields:** Topic (Title, Abstract, Keywords, Keywords Plus) **Records retrieved:** 797

```
TS=(
  ("explainable artificial intelligence" OR "explainable AI" OR "XAI"
   OR "interpretable machine learning" OR "SHAP" OR "Shapley"
   OR "LIME" OR "feature importance" OR "model interpretability"
   OR "model explainability" OR "explainability"
   OR "counterfactual explanation" OR "attention mechanism"
   OR "saliency map" OR "Grad-CAM")
  AND
  ("mental health" OR "depression" OR "anxiety" OR "stress" OR "PTSD"
   OR "wellbeing" OR "well-being" OR "psychological" OR "psychiatric"
   OR "autism" OR "schizophrenia" OR "suicide" OR "personality disorder"
   OR "cognitive load" OR "bipolar" OR "sleep disorder" OR "mood disorder"
   OR "ADHD" OR "burnout" OR "mental disorder" OR "emotional")
)
Filters: Publication Years: 2015-2026; Language: English
```

Note: Of the 797 retrieved records, 32 were identified as topically relevant (containing both XAI/interpretability and psychological terms in their title). After open-access filtering and full-text screening of the 32 candidate records, 7 studies were included in the final corpus.

A.5 Manual / Web Search

A supplementary manual search was conducted in **Google Scholar**, **Semantic Scholar**, and **ResearchGate** using the keyword combinations below. The primary manual search was performed in February 2025 and updated in March 2026. Backward citation tracking was applied to the most relevant retrieved studies.
Records retrieved: 42

```
Query variants used (entered as free-text):

1. "explainable AI" AND ("mental health" OR "depression" OR "anxiety")
2. "XAI" AND "psychology"
3. "SHAP" AND ("depression" OR "stress" OR "well-being")
4. "LIME" AND ("mental health" OR "autism" OR "schizophrenia")
5. "interpretable machine learning" AND "psychiatry"
6. "feature importance" AND ("psychological" OR "burnout" OR "PTSD")
7. "explainable artificial intelligence" AND "suicide"
8. "explainable AI" AND "personality"
9. "SHAP" AND "resilience"
10. "XAI" AND "cognitive load"
```

A.6 Databases Not Searched

Scopus was not searched due to the absence of institutional access at the time of the primary search. Its omission is acknowledged as a limitation of this study (Section 5.4). **PsycINFO** (APA) was similarly unavailable. Future updates to this review should include these databases.

Appendix B: Summary Tables

Table 4: Comparison of XAI methods applied in the reviewed studies ($n = 42$). *Stage*: ante-hoc = intrinsic transparency; post-hoc = explanation after training. *Scope*: L = local (per-prediction); G = global (model-level). MA = model-agnostic; MS = model-specific.

Group	Feature	SHAP	LIME	Feat. Imp.	Grad-CAM / IG	Attention	Counterfactual
Objective	Goal	Quantify feature contributions	Approximate locally with interpretable surrogate	Rank features by model-derived score	Localise discriminative regions in input	Weight relevant tokens or time steps	Find minimal change for different outcome
	Target audience	Researchers & clinicians	Non-technical users, clinicians	Researchers	Domain experts with visual data	NLP researchers	Clinicians, patients
	ML task	Classification & regression	Classification & regression	Classification & regression	Classification (image/signal)	Sequence classification	Classification
	Input modality	Tabular, text, image, EEG	Tabular, text, image	Tabular	Image, EEG, fMRI	Text, EEG sequences	Tabular, text
Methodology	Stage	Post-hoc	Post-hoc	Post-hoc (or ante-hoc for tree models)	Post-hoc	Ante-hoc (built-in)	Post-hoc

(Table 4 continued)

Group	Feature	SHAP	LIME	Feat. Imp.	Grad-CAM / IG	Attention	Counterfactual
	Applicability	MA (also MS for TreeSHAP)	MA	MA (or MS for RF/XGBoost)	MS (gradient-based networks)	MS (attention-based architectures)	MA
	Method basis	Shapley values (game theory)	Local linear surrogate	Tree/gradient impurity or permutation	Gradient × activation map	Learned attention weights	Optimisation (minimal perturbation)
	Scope	L + G	L	G	L	L + G	L
Output	Result form	Numeric contribution per feature	Numeric weight per feature (local)	Ranked feature list or bar chart	Heat map overlaid on input	Attention weight matrix / heat map	Alternative input instance
	Presentation	Beeswarm / bar / waterfall plot	Bar chart (local)	Bar chart / table	Visual overlay (saliency map)	Attention map / heat map	Textual or tabular diff
	Evaluation metric	Faithfulness, stability, consistency	Faithfulness, fidelity to black-box	Predictive accuracy of surrogate	Pointing game, localisation score	Gradient-based sanity checks	Proximity, sparsity, plausibility

(Table 4 continued)

Group	Feature	SHAP	LIME	Feat. Imp.	Grad-CAM / IG	Attention	Counterfactual
Practical notes	Key strength	Theoretically grounded; L+G; widely adopted	Simple; modality-flexible; human-readable	Fast; no extra library needed for tree models	Intuitive visual output for imaging data	Explanations inherent to model; no extra cost	Actionable; causal framing
	Key limitation	Computationally expensive (KernelSHAP); can be unstable	High variance; sensitive to neighbourhood sampling	Can be misleading for correlated features	Limited to gradient-capable architectures; may miss context	Attention ≠ explanation debate; architecture-dependent	Hard to guarantee plausibility; optimisation-heavy
Usage in corpus	Studies (n = 42)	24 (57%)	10 (24%)	8 (19%)	3 (7%)	3 (7%)	2 (5%)
	Main psychological domains	Depression, stress, well-being, anxiety	Depression, stress, well-being	Depression, well-being, stress	Autism/ASD, creativity, personality traits	Depression (NLP), stress (EEG)	Depression, mental health (general)