International Journal of
Mental Health Promotion

Tech Science Press

http://doi.org/10.32604/ijmhp.2026.079327

REVIEW

The Contribution of Artificial Neural Networks to Psychological Assessment: A Scoping Review

Monica Casella[1], Raffaella Esposito[2,*] and Michela Ponticorvo[1]

[1]Department of Humanistic Studies, University of Naples Federico II, Naples, Italy
[2]ICAROS Center, University of Naples Federico II, Naples, Italy
*Corresponding Author: Raffaella Esposito. Email: raffaella.esposito3@unina.it

Received: 19 January 2026; Accepted: 14 July 2026; Published: 31 August 2026

ABSTRACT: Background: Artificial Neural Networks (ANNs) are increasingly used in psychological assessment to score measures, classify clinical presentations, monitor symptoms, estimate risk, and tailor interventions. **Methods:** This scoping review maps ANN applications in psychological assessment and diagnosis and examines their relevance to psychiatric practice. A structured search of Scopus and Web of Science, followed by independent screening and data charting by three reviewers, identified 31 studies published between 2020 and 2023. **Results:** The literature demonstrates broad methodological versatility across psychometric, behavioral, textual, image, and sensor data, but clinical readiness remains limited by heterogeneous samples, architectures, metrics, and validation procedures. Exceptionally high accuracy estimates require cautious interpretation, particularly in small or single-site datasets. Persistent barriers include limited external validation, insufficient interpretability, demographic and cultural bias, privacy and governance concerns, workflow integration, and unclear regulatory and medico-legal responsibility. **Conclusions:** ANNs should currently be considered clinical decision-support tools rather than substitutes for professional judgment.

KEYWORDS: Machine learning in psychology; artificial neural networks; psychological assessment

1 Introduction

Artificial Intelligence (AI) is increasingly used in psychological assessment, where observations, test responses, interview data, and behavioral indicators are translated into judgments that may influence diagnosis, treatment planning, education, employment, and access to services [1,2]. Because these decisions can materially affect people's lives, the scientific debate concerns not only efficiency and predictive performance, but also validity, interpretability, fairness, privacy, and accountability [3,4].

Applications in healthcare, education, psychological assessment, diagnosis, and personnel selection are especially consequential because model outputs can affect individual opportunities and care pathways [1,2]. In psychiatric and psychological practice, however, an accurate prediction is not equivalent to a diagnosis or a treatment decision: model output must be interpreted alongside clinical history, contextual information, patient preferences, and professional judgment.

The potential of AI to improve efficiency and support medical diagnosis, personalized recommendations, and task automation is often met with enthusiasm [5]. However, concerns about algorithmic bias, privacy, social inequality, loss of human control, and ethical dilemmas remain central [6,7]. These tensions are especially consequential in psychological assessment, where errors may shape clinical labels, referrals, interventions, or access to services.

Psychology belongs to these high-stakes fields. AI is now used across psychological research and clinical practice [8], and its rapid diffusion warrants critical reflection on what models predict, how their outputs are validated, and how they should be incorporated into professional decision-making.

In the present study, we provide a scoping review of Artificial Neural Networks (ANNs) in psychological measurement, assessment, and diagnosis. After briefly outlining the architectures and learning mechanisms needed to interpret the literature, we map their psychometric and clinical applications, distinguish prediction from clinical decision-making, and examine the methodological, ethical, and implementation requirements that currently limit routine psychiatric use.

1.1 Artificial Neural Networks in Psychological Assessment

AI refers to the simulation of human intelligence processes by machines, especially computer systems [9]. These processes include learning (the acquisition of information and rules for using the information), reasoning (using rules to reach approximate or definite conclusions), and self-correction.

Machine Learning (ML) is a subset of AI that focuses on the development of algorithms and statistical models that enable computers to perform tasks without explicit instructions, relying on patterns and inference instead [10]. It involves the construction of systems that can learn from and make predictions or decisions based on data. Among the various techniques under the broad umbrella of ML, ANNs are computational models inspired by the structure and function of biological neural networks in the brain [11,12]. In general, an ANN consists of interconnected nodes, called neurons, organized into layers. Each neuron receives input signals, processes them through an activation function, and passes the output to the next layer. Through a process of training, ANNs adjust the connection strengths (weights) between neurons to learn the patterns and relationships within the input data. ANNs are fundamental components of Deep Learning algorithms [12] and are used for pattern recognition, classification, regression, and representation learning. Deep models contain multiple layers that can extract increasingly abstract features from images, language, speech, and other complex inputs. For the purposes of this review, ANNs are treated as a coherent family of models rather than as inherently superior to alternative machine-learning or statistical approaches. ANNs are a subset of ML and are widely used for pattern recognition, classification, regression, and clustering, all of which are relevant to assessment and diagnosis. Their historical development reflects biologically inspired, distributed approaches to learning [13], although contemporary ANNs should not be interpreted as literal models of the human brain. Their practical value lies in learning complex and potentially nonlinear relations and generalizing them to previously unseen inputs, provided that training data and validation procedures are adequate. In the next section we report some essential information to introduce ANNs; for a more detailed exposition and discussion, it is possible to refer to Urban and Gates [14].

1.2 Anatomy of Artificial Neural Networks

At a topological level, the configuration of connections defines the network architecture. Information flows through the network from the input layer, where data are received externally, through one or more hidden layers—which do not have direct contact with the external environment—to the output layer, where the final outcome is produced [15].

As introduced previously, artificial neurons process incoming information using an activation function [16]. Specifically, this activation is calculated by summing all the input values multiplied by their respective weights [17]. Various types of functions can then be applied to this sum to determine the neuron's output. The most commonly used are the step function (where the output is 0 or 1 if the input exceeds a set excitability threshold), the continuous linear function (a linear transformation of the input), and the sigmoid or logistic function (which ranges from 0 to 1 and can be likened to probability distributions) [18].

Ultimately, what an ANN "learns" resides in this pattern of connections, in relation to its architecture and the information flow. For example, the flow can go in one direction only, from input to output, in a feed-forward NN, or there can be recursive feedback connections, as it happens in a recurrent NN [19,20].

1.3 Learning Mechanisms in Artificial Neural Networks

The behavior of a neural network, including its response to a given input, is determined by its architecture and the configuration of its weights. Accordingly, a network's learned representation is encoded in its weight configuration [20]. Finding an appropriate configuration for a task constitutes the learning process. Supervised learning uses labeled target outputs, whereas unsupervised learning seeks structure in data without labeled targets [21,22].

In supervised learning, an external teaching signal specifies the desired output for each training example [21]. Through repeated updates, the network adjusts its weights to reduce the discrepancy between predicted and target outputs. The delta (Widrow–Hoff/LMS) rule applies to single-layer adaptive networks, whereas backpropagation extends gradient-based learning to multilayer networks by propagating error derivatives through hidden layers [20,23].

In unsupervised learning, no target output is specified, and the network learns structure from the input data [22]. Kohonen's self-organizing maps are a canonical example of unsupervised learning [24]. Hopfield networks, by contrast, are recurrent associative-memory models designed to store and retrieve patterns, including from partial inputs [25]. In addition to these two primary paradigms, semi-supervised learning represents a hybrid approach [10]. It leverages a small amount of labeled data alongside a much larger pool of unlabeled data during the training phase. This method is particularly valuable in psychological and clinical research, where acquiring expertly labeled diagnostic data is often expensive and time-consuming, while raw, unannotated data is more readily available [26].

1.4 Overview of Artificial Neural Networks Notable Architectures

The architecture is a key element of ANNs and determines their behaviors and what learning algorithm can be applied to them. Several neural network architectures have proven particularly relevant to psychological research and assessment [14]. Feedforward Neural Networks (FNN) are the simplest type of neural network architecture, where information flows in one direction—from input nodes through intermediate nodes (hidden layers) to output nodes—without cycles or loops. They are commonly used for tasks like pattern recognition, classification, and regression; in psychometrics, they are frequently employed for predicting clinical outcomes from questionnaire data [27].

Convolutional Neural Networks (CNN) are designed specifically for processing grid-like data, such as images or videos. They consist of convolutional layers, pooling layers, and fully connected layers, leveraging convolution operations to detect patterns and spatial hierarchies in the input data. This makes them highly effective for image classification and object detection; in clinical psychology, CNNs are increasingly applied to analyze complex spatial representations, such as brain imaging scans or digitized clinical drawing tests [28]. To handle sequential data, Recurrent Neural Networks (RNN) introduce loops within the network architecture, allowing information to persist over time. A more advanced variant, Long Short-Term Memory (LSTM) networks, addresses the vanishing gradient problem by introducing gated cells. These cells allow the network to retain long-term dependencies and remember information over extended time intervals [20].

While widely used for natural language processing and sequence generation, in psychological research they are highly effective for dealing with time-series data, such as longitudinal behavioral tracking or analyzing patient speech patterns [29].

Finally, Autoencoders are a type of neural network architecture used for unsupervised learning and dimensionality reduction. They consist of an encoder network that compresses the input data into a lower-dimensional latent space representation and a decoder network that reconstructs the original input from the encoded representation. While employed in general tasks like anomaly detection and feature learning, in psychometrics they are exceptionally useful for compressing complex behavioral data or aiding in the item-reduction process of psychological scales by extracting latent features from raw inputs [30].

These architectures are tailored to different data modalities and tasks. We selected ANNs as the organizing focus because they constitute a coherent and influential family of models across psychometric and clinical applications; this choice does not imply that they are uniformly preferable to regression, support vector machines, random forests, or boosting methods. Model selection should depend on the research question, data modality, sample size, class balance, interpretability requirements, and intended clinical use. Large Language Models were excluded because their scale, training objectives, and predominantly generative use cases require a separate review framework and would reduce comparability across the task-specific models examined here.

1.5 The Present Study

In the present study, we propose a scoping review of the use of ANNs in the field of psychological measures, assessment, and diagnosis. Rather than approaching ANNs only as abstract computational models, we examine their role in tasks that are directly relevant to applied mental health practice, including psychological evaluation, diagnostic classification, symptom monitoring, risk detection, and prediction of clinically meaningful outcomes. In this sense, ANNs are considered not simply as psychometric tools, but as methods that may help translate complex patient data into information that can support clinical interpretation and decision-making.

Moving beyond the theoretical level, we examine how these models may function as decision-support tools for clinicians and psychiatric nursing staff. ANNs may organize complex information, identify cases requiring closer assessment, and support continuous monitoring or early-warning workflows. They do not independently establish diagnoses or determine treatment, and translation into routine practice requires data standardization, algorithmic transparency, robust external validation, workflow integration, and accountable human oversight. We did not include studies whose primary objective was to connect neuroimaging or physiological patterns to clinical conditions without an explicit psychological outcome, in order to maintain a focused review of behavioral and psychometric assessment. The potential integration of validated biological, behavioral, and psychometric information is considered as a future research direction. Specifically, this review aims to summarize ANN applications and validation practices, distinguish diagnostic classification from prognosis, symptom monitoring, risk detection, and personalized intervention, and identify the conditions under which ANN outputs could responsibly support clinical assessment.

2 Methods

To address the objective of the study, we conducted a scoping review using a systematic literature search, study-selection, and data-charting process. A scoping design was selected because the review aimed to map the range, characteristics, and functions of ANN applications across heterogeneous psychological domains, participant populations, data modalities, model architectures, and validation approaches. The review was not registered, and no protocol was prepared before the study was conducted. The review is reported in accordance with the PRISMA Extension for Scoping Reviews (PRISMA-ScR) [31], and the completed checklist is provided as Supplementary Materials.

2.1 Eligibility Criteria

Studies were eligible when they: (1) were empirical articles published in English and in final publication form between 2018 and 26 January 2024; (2) were retrieved within the psychology subject area and through the psychology keyword filter applied in the databases; (3) used ANNs as a method of data analysis; and (4) investigated cognitive, emotional, or behavioral outcomes relevant to psychological measurement, psychometric development or scoring, diagnostic classification, symptom monitoring, prognosis, risk detection, or treatment-related applications. Studies were excluded if they did not use ANNs; addressed non-psychological fields, such as economics or law; analyzed physiological variables without an explicit psychological outcome; fell outside the clinical, health, or psychometric scope of the review; or were theoretical rather than empirical. The included studies were grouped according to their principal application and function in psychological assessment.

2.2 Information Sources

Scopus and Web of Science were searched. The final search in both databases was conducted on 26 January 2024. The search was restricted to the peer-reviewed literature indexed in these databases; grey literature and other complementary sources were not searched.

2.3 Search Strategy

Titles, abstracts, and keywords were searched in both Scopus and Web of Science using the same Boolean search string: “artificial” PRE/0 “neural” PRE/0 “networks” OR “artificial” PRE/0 “neural” PRE/0 “network” OR “deep” PRE/0 “learning” OR “machine” PRE/0 “learning” OR “artificial” PRE/0 “intelligence”. Results were limited to English-language articles in final publication stage, published from 2018 to the date of the final search, within the psychology subject area and using the keyword “psychology.” The identification, manual duplicate-removal, screening, and eligibility process is summarized in the PRISMA 2020 flow diagram (Fig. 1).

2.4 Selection Process

Duplicate records were identified and removed manually before screening. Three reviewers independently screened each remaining record on the basis of its abstract. At this stage, records were excluded when the abstract did not mention a machine-learning method of data analysis or a psychological variable. Full texts of potentially eligible reports were then assessed independently against the eligibility criteria. Disagreements were resolved through discussion and consensus among the three reviewers. No automation tools were used in duplicate removal or study selection.

2.5 Data Charting Process and Data Items

Three reviewers independently extracted data from each included report. A shared charting framework was used to collect the topic and area of application, authors, year, country or region, sample size and data characteristics, psychological outcome, ANN architecture and methodological approach, comparator models, training and evaluation strategy, performance metrics, and principal findings. Funding sources of the included studies were not charted.

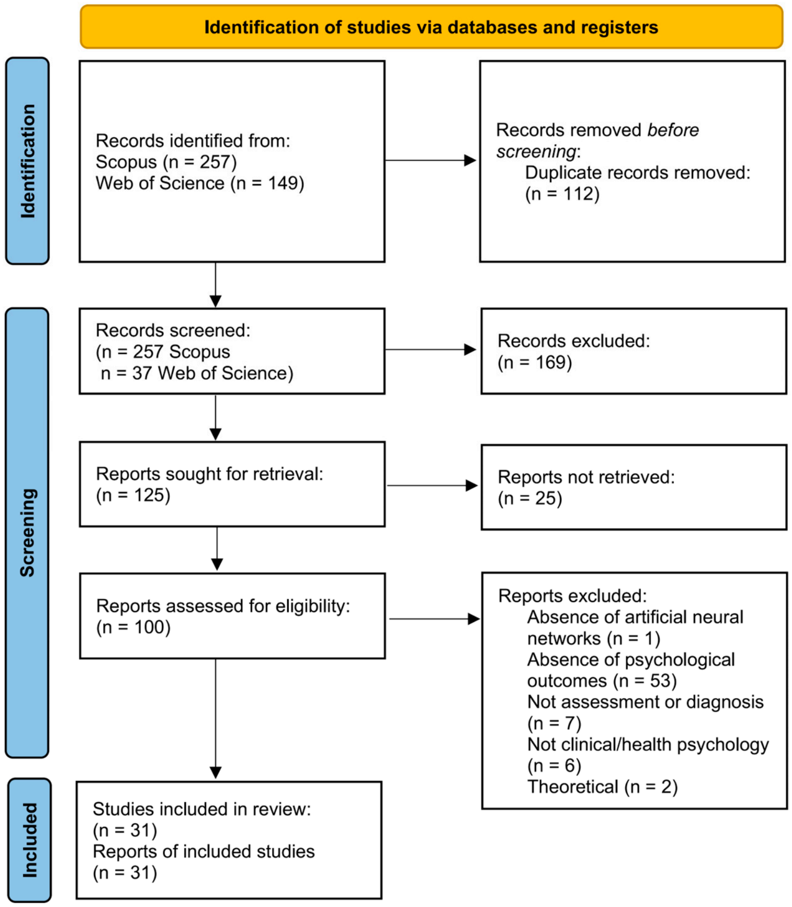

Figure 1: PRISMA 2020 flow diagram of the study-selection process for the scoping review (databases and registers only).

The charting framework was refined through discussion as the reviewers became familiar with the heterogeneous evidence base, while transparent communication and consensus were used to maintain consistency in data extraction and interpretation. Performance metrics and study findings were charted as reported by the original authors. No contact with study authors was required, and no automation tools were used in data charting.

2.6 Synthesis and Critical Appraisal

The charted data were summarized descriptively and synthesized thematically according to the principal functions served by ANN models in psychological assessment, including psychometric applications, diagnostic classification, symptom monitoring, prognosis, risk detection, and treatment-related applications. The studies were tabulated according to their general and methodological characteristics. To support thematic mapping, a co-occurrence network analysis of two-word groups extracted from the included abstracts was performed using the bibliometrix R package [32]. The resulting 25-node network represented term co-occurrences, node size reflected centrality, and communities were identified using the Walktrap algorithm. No formal critical appraisal, reporting-bias assessment, or certainty assessment was conducted because the purpose of the review was to map the range and characteristics of the evidence rather than to estimate intervention effects.

3 Thematic Synthesis of ANNs Integration in Psychological Practice

The database searches identified 406 records (Scopus, n = 257; Web of Science, n = 149). After 112 duplicate records were removed, 294 records were screened. Of 125 reports sought for retrieval, 25 could not be retrieved, and 100 full-text reports were assessed for eligibility. Sixty-nine reports were excluded for the reasons shown in Fig. 1, leaving 31 studies in the review.

Tables 1 and 2 present a summary of information on the reviewed studies. The studies include samples ranging from 22 to 44,864 participants. They were conducted between 2020 and 2023, mainly in the USA (n = 11). The most relevant sources are Frontiers in Psychology (n = 12) and the Journal of Affective Disorders (n = 4). The most reported ANN architecture is the feed-forward neural network model. Cross-entropy is the most frequently used loss function, while the minimum number of training epochs is 20 and the maximum is 5000, although some studies do not specify one or either of these characteristics.

Cross-validation appears as the main evaluation strategy, while accuracy is the most commonly reported evaluation metric. Other frequently mentioned metrics are F1-score, recall, precision, and AUC (Area Under the Receiver Operating Characteristic curve).

Table 1: General characteristics of the selected studies.

Reference	Authors	Year	Country/Region	Data Characteristics	Area of Interest	Outcome of Interest
[26]	Xue & Bradshaw	2021	USA	Simulated data; Data type: 20 or 30 items; Sample size: 1000, 100 replications	Diagnosis	Performance of the ANN model in diagnostic classification
[33]	Milano et al.	2023	Italy	Data type: Continuous data; Sample size: 60 children	Diagnosis	ASD classification
[34]	Kumar & Das	2022	India	Data type: Continuous data; Sample size: 701	Diagnosis	Classification of autistic patients
[35]	Lin et al.	2023	Taiwan	Data type: Discrete data from psychometric tests; Sample size: 328 records	Diagnosis	Attention Deficit Hyperactivity Disorder (ADHD) type classification
[36]	Na et al.	2021	Korea	Data type: physiological data related to heart activity; Sample size: 60 patients with panic disorder and 61 with other anxiety disorders	Diagnosis	Discriminate panic disorder from other anxiety disorders
[37]	Lee et al.	2021	Taiwan	Data type: Categorical responses; Sample size: 351 patients with schizophrenia and 101 healthy adults	Diagnosis	Prediction of diagnosis and classification performance
[38]	Jeon et al.	2022	Korea	Data type: Likert-type data; Continuous data; Sample size: 52	Diagnosis	Relation between gaze avoidance, pupil diameter, and schizophrenia psychopathology

Table 1: *Cont.*

Reference	Authors	Year	Country/Region	Data Characteristics	Area of Interest	Outcome of Interest
[39]	Savci & Tekin	2022	Turkey	Data type: social media usage data and psychometric test scores; Sample size: 309 university students	Diagnosis	Prediction of problematic social network use (PSU)
[40]	Shahzad et al.	2020	Pakistan	Data type: psychometric scale scores; Sample size: 200 OCD patients, 400 controls	Diagnosis	Classification of Obsessive Compulsive Disorder (OCD) patients vs. healthy controls
[28]	Amini et al.	2021	USA	Data type: images data; Sample size: 3263 cognitively intact individuals and 160 impaired subjects	Diagnosis	Discriminate people with cognitive impairment from cognitively intact individuals
[41]	Wu et al.	2022	USA	Data type: Continuous and categorical data; Sample size: 163,199 records (44,864 subjects)	Prognostic modeling	Prediction of probability of conversion time to next disease stage
[42]	Wang et al.	2023	USA	Data type: Likert-type data; Sample size: 30	Symptom monitoring	Daily experience of anxiety
[43]	Jacobson et al.	2021	USA	Data type: Continuous data; Sample size: 265	Symptom monitoring	Prediction of long-term deterioration in anxiety and panic disorder symptoms
[44]	Jacobson & Bhattacharya	2022	USA	Data type: sensors data and position data; Sample size: 32 undergraduate students	Symptom monitoring	Moment-to-moment prediction of anxiety and avoidance symptoms using smartphone sensor data
[29]	Little et al.	2021	UK	Data type: physical activity, neuropsychological, and recording data; Sample size: 29 participants with late-life depression and 29 age-matched controls	Symptom monitoring	Classification of speech and non-speech using acoustic features
[45]	Everaert et al.	2022	Netherlands	Data type: item responses; Sample size: 460 participants	Risk detection	Individual Beck Depression Inventory-II (BDI-II) items used as outcome variables

Table 1: *Cont.*

Reference	Authors	Year	Country/Region	Data Characteristics	Area of Interest	Outcome of Interest
[46]	Zhang et al.	2021	UK	Data type and sample size: online sources (659 suicide notes, 431 last statements, 2000 neutral posts)	Risk detection	Classification of suicide notes
[47]	Shin & Kim	2023	Korea	Data type: Continuous and dichotomic data; Sample size: 8623–9270 students	Risk detection	Presence of suicidal ideation
[48]	Singh et al.	2022	USA	Data type: Continuous data from game interaction and Likert-type data; Sample size: 118 participants;	Treatment adherence tracking	Adherence to cognitive training
[49]	Sandeep et al.	2020	USA	Data type: data from cognitive tasks; Sample size: 262 participants	Treatment personalization	Estimation of participant skill and prediction of performance at task challenges
[50]	Rennie et al.	2020	UK	Data type: cognitive and behavioral task data; baseline subgrouping sample: 616 participants; training sample: 179 participants	Treatment outcome profiling	Identification of different response profiles to training
[51]	Rao et al.	2022	China	Data type: physiological signals; Sample size: 22	Health Psychology	Emotion recognition based on physiological signals
[52]	Nursalim et al.	2023	Indonesia	Continuous and Likert-type data	Health Psychology	Resilience
[53]	Morales-Rodríguez et al.	2021	Spain	Data type: psychometric test scores (continuous data); Sample size: 337	Health Psychology	Predicting stress levels
[54]	Chadaga et al.	2023	India	Data type: Dichotomous questionnaire data; crowdsourced public dataset; sample size: 3002 participants	Health Psychology	Awareness level regarding child sexual abuse
[27]	Dolce et al.	2020	Italy	Data type: Likert-type data; Sample size: 604 adults	Psychometrics	Item selection
[30]	Casella et al.	2023	Italy	Simulated data with 4 different sample sizes (200, 500, 700, 1000)	Psychometrics	Item ranking and reconstruction

Table 1: *Cont.*

Reference	Authors	Year	Country/Region	Data Characteristics	Area of Interest	Outcome of Interest
[55]	Collier et al.	2023	USA	Data type: Likert-type data; Sample size: 771 Americans	Psychometrics	Imputation of missing data
[56]	Devine et al.	2023	England	Sample size: 1135 (Sample 1) and 1020 (Sample 2)	Psychometrics	Automated scoring system for Advanced Theory of Mind (ToM)
[57]	Thompson & Koenig	2023	USA	Sample size: 5000 ratings	Psychometrics	Hire/No hire decision
[58]	Mahajan et al.	2022	India	Data type: textual data; Sample size: 250	Psychometrics	Prediction of personality traits of real-time Twitter users using the Big Five model

Note: ADHD, Attention Deficit Hyperactivity Disorder; ANN, Artificial Neural Network; ASD, Autism Spectrum Disorder; BDI-II, Beck Depression Inventory-II; CSA, Child Sexual Abuse; GAD-7, Generalized Anxiety Disorder-7; OCD, Obsessive Compulsive Disorder; PSU, Problematic Social Media Use; ToM, Theory of Mind. For Chadaga et al. [54], 2023 denotes the online-first publication year; the article was assigned to the 2024 journal volume.

Table 2: ANN models' characteristics of the selected studies.

Reference	Methodological Approach	Top-Line Findings	Evaluation Strategy
[26]	ANNs using a Co-Training framework (combining Deterministic-Input, Noisy, "And" Gate and Deterministic-Inputs, Noisy, "Or" Gate models)	The proposed method provided classification results comparable to or higher than traditional Diagnostic Classification Models, demonstrating superior robustness to noise; when Q-matrix accuracy dropped to 90%, the latent class accuracy reduction was only 0.75% for the Artificial Neural Network, compared to 3.02% for the Generalized Deterministic-Input, Noisy, "And" Gate model and 2.98% for the Log-Linear Cognitive Diagnosis Model.	Training/testing
[33]	ANNs	Motor features (speed, max acceleration, in-acceleration, and sd-acceleration) are effective for classification (91.2% ± 5.6% accuracy).	Holdout; ten-fold cross-validation
[34]	ANNs; SVM; RF	ML techniques significantly improved ASD diagnostic accuracy achieving up to 100% accuracy (SVM) on complete data and maintaining 100% accuracy (SVM and RF) by using Recursive Feature Elimination (RFE) to handle missing values.	Training/testing
[35]	ANNs	ANN achieved 74–78% accuracy in distinguishing ADHD-Inattentive, ADHD-Combined, and control groups, and 85–90% accuracy in distinguishing ADHD-I and ADHD-C.	K-fold cross-validation
[36]	ANNs compared with SVM, GBM, RF	RF showed the best accuracy (0.784), followed by ANN (0.730), SVM (0.730), GBM (0.676); RF and ANN differentiated panic disorder from other anxiety disorders.	Ten-fold cross-validation with five repeats

Table 2: *Cont.*

Reference	Methodological Approach	Top-Line Findings	Evaluation Strategy
[37]	ANNs combined with other statistical techniques	ML-FERD screening achieved 90% accuracy, sensitivity 85–89%, specificity 93–96%.	Five-fold cross-validation
[38]	ANNs	A Residual Masking Network (accuracy >74%) successfully verified the neutrality of real-life facial stimuli, enabling the identification of significant correlations between gaze avoidance and symptom severity (e.g., PANSS total score $r = -0.401$, $p = 0.008$).	Not specified
[39]	ANNs and SVM	Most significant predictors of problematic social media use were daily usage frequency, checking behavior, desire for pleasure, exhibitionism, and Fear of Missing Out (FOMO); prediction rates: 0.62 for ANN and 0.63 for SVM).	Five-fold cross-validation
[40]	ANNs	The ANN model achieved a testing accuracy of 98.2% (AUC = 0.991) in classifying OCD patients versus healthy controls, identifying cleaning/contamination (100% normalized importance) and "worth in family" (71.1%) as the strongest predictors.	Training/testing
[28]	ANNs	Deep learning model showed promise in predicting dementia with high accuracy (average AUC and average F1 score of 91.9% ±1.1% and 94.6% ±0.4%, respectively).	Five-fold cross-validation
[41]	ANNs integrated with Cox proportional hazards survival analysis (DeepSurv)	DeepSurv predicted the probability of time to stage-specific conversion with high accuracy, achieving a concordance index (CI) of at least 86% and an integrated Brier score (IBS) of less than 0.1 across different disease stages.	Five-fold cross-validation
[42]	LSTM (Long Short-Term Memory) encoder-decoder for latent feature extraction followed by Hierarchical Clustering	The encoder–decoder reconstructed daily anxiety values with a mean correlation of $r = 0.624$ in the training set and $r = 0.602$ in the testing set (excluding missing values); latent-space clustering showed that adolescents with similar GAD-7 scores could have different day-to-day anxiety dynamics.	Training/testing split; validation loss used for early stopping and model selection
[43]	ANNs	Wearable actigraphy features predicted 17–18-year deterioration in anxiety-disorder symptoms (AUC = 0.696, 84.6% sensitivity, 52.7% specificity, balanced accuracy = 68.7%).	Four-fold cross-validation
[44]	ANNs	Personalized deep-learning models explained 74.8% of total variation across persons (robust $R^2 = 0.748$, 95% CI [0.728, 0.766]) and, on average, 38.5% of within-person hour-to-hour variation (mean robust $R^2 = 0.385$).	Cross-validation with sliding-window approach
[29]	ANNs combined with other statistical techniques	Deep learning algorithms achieved 93.8% accuracy in speech detection and 89.95% in speaker-specific classification, revealing that Late-Life Depression is associated with significantly lower speech activity and complexity.	Leave-One-Session-Out validation

Table 2: *Cont.*

Reference	Methodological Approach	Top-Line Findings	Evaluation Strategy
[45]	ANNs	Specific features of repetitive negative thinking and positive reappraisal predicted individual depressive symptoms beyond perceived stress (AUC values ranged from 0.63 to 0.81).	Training/testing
[46]	ANNs compared with other ML techniques	The transformer-based model achieved average precision of 95.0%, recall of 94.9%, and F1-score of 94.9%, outperforming the evaluated baseline approaches in suicide-note classification.	Training/testing
[47]	ANNs compared with statistical and ML methods	The convolutional neural network showed the highest prediction performance, with approximately 90% accuracy; logistic regression was used to examine the influence of independent variables.	Ten-fold cross-validation
[48]	ANNs	Deep learning models successfully predicted daily adherence to cognitive training based on past behavior, achieving highest mean F-scores of 75.5% (CNN), 75.5% (LSTM), and 74.6% (CNN-LSTM).	Training/testing
[49]	Input–output hidden Markov models, an unscented Kalman filter, and an LSTM model evaluated on cognitive-training sequences	The hidden Markov models provided the best overall prediction of participant performance; test RMSEs for the two HMM variants were 12.06% and 5.60% in Experiment 1 and 12.54% and 18.52% in Experiment 2. The study evaluated model-based skill estimation rather than deploying the models to control task difficulty.	Cross-condition testing in Experiment 1; 80:20 train/test split in Experiment 2; cross-validation for LSTM configuration
[50]	Self-Organizing Maps (SOMs) and K-means clustering to identify cognitive-performance subgroups	The study identified four cognitive-performance subgroups (High, Medium-Verbal, Medium-Visuospatial, and Low-performing) and mapped pre/post-training transitions among them; baseline fluid intelligence differed significantly across transition groups ($F(3, 122) = 28.83$, $p < 0.001$).	Cross-validation
[51]	BPNN (Backpropagation Neural Network) trained on 12-feature physiological signal time-series	The BPNN model successfully discriminated between calm, happy, and fear states, achieving an average recognition accuracy of 73% (peaking at 81% for 6-year-olds), identifying happiness as the most accurately recognized emotion.	Holdout; Training/testing
[52]	ANNs	The ANN model achieved a predictive capacity of ~65% in the holdout phase, identifying stress intensity (100% normalized importance) and mindfulness (40.9% normalized importance) as primary predictors of resilience.	Training/testing
[53]	ANNs	The model achieved a predictive accuracy >80%, identifying coping strategies (e.g., negative self-focus, positive re-evaluation) as the primary predictors of stress levels in the university community.	Training/testing
[54]	Multiple supervised ML models, including a deep neural network, and multi-level stacked ensembles combining Logistic Regression, Decision Tree, KNN, Naïve Bayes, SVM, RF, AdaBoost, XGBoost, and CatBoost	The final stacked ensemble achieved 94% accuracy, whereas the deep neural network achieved 89%. Explainability analyses identified knowledge of child grooming and recognition of signs of abuse as influential predictors.	Training/testing

Table 2: *Cont.*

Reference	Methodological Approach	Top-Line Findings	Evaluation Strategy
[27]	ANNs compared with Exploratory Data Analysis	ML procedures selected items that improved predictive accuracy (up to 88.5%) while reducing redundancy in psychometric scales.	Cross-validation
[30]	ANNs compared to PCA, Genetic Algorithms, and Ant Colony Optimization	Both autoencoders preserve the dimensionality of the original measure and preserve items with greater loadings than classical techniques.	Cross-validation
[55]	ANNs	ANN-based imputation successfully handled missing data, outperforming classical imputation methods (overall f1-score of 0.67 versus 0.22).	Training/testing
[56]	ANNs, SVM, Factor Analysis	Deep learning algorithms enabled automated scoring of open-ended responses to the Strange Stories task, demonstrating proof-of-principle for ToM assessment (Krippendorff's α = 0.86).	Training/testing
[57]	Comparative analysis of Bag-of-Words (BoW), Long Short-Term Memory (LSTM) networks, and Robustly Optimized BERT Pretraining Approach (RoBERTa) deep learning models	The RoBERTa model achieved superior performance (avg r = 0.84), showing high agreement with expert SME ratings and maintaining robust predictive validity even with limited training samples	K-fold cross-validation
[58]	CNN (Convolutional Neural Network) and BiLSTM (Bidirectional Long Short Memory)	A bidirectional LSTM Convolutional NN ensemble enabled the successful prediction of personality traits on Twitter (accuracy 75.134%).	Ten-fold cross-validation

Note: ADHD, Attention Deficit Hyperactivity Disorder; ANN, Artificial Neural Network; ASD, Autism Spectrum Disorder; CNN, Convolutional Neural Network; FOMO, Fear of Missing Out; GAD-7, Generalized Anxiety Disorder-7; GBM, Gradient Boosting Machine; KNN, K-Nearest Neighbors; LSTM, Long Short-Term Memory; ML, Machine Learning; OCD, Obsessive Compulsive Disorder; PCA, Principal Component Analysis; RF, Random Forest; SVM, Support Vector Machine; ToM, Theory of Mind.

The co-occurrence network analysis of the included abstracts is shown in Fig. 2.

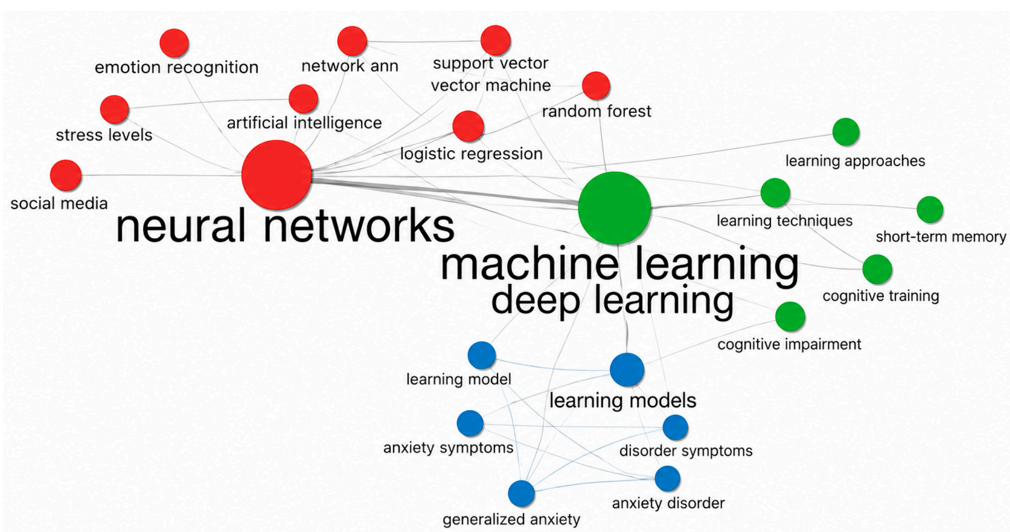

Figure 2: Co-occurrence Network Analysis of the abstracts retrieved from the papers selected.

The network comprises 25 nodes representing unique two-word groups and their respective co-occurrences.

As expected, results show that the central nodes are terms that are integral to our review's focus. Notably, "neural networks" is frequently associated with other ML algorithms. Indeed, ANNs are used as the only method 12 times, while in the other cases, they are combined with other ML methods or statistical techniques or compared with other methods. Furthermore, the network reveals three distinct clusters: the first encompasses research on emotion recognition, social media, and stress level analysis; the second is associated with clinical topics, including anxiety and panic disorders; and the third is more related to cognitive areas.

Building on these insights, and in order to improve the clarity and readability of the results, we decided to group the included studies based on the following areas of interest: clinical and health psychology (25 papers) and psychometrics and assessment (6 papers). This classification allows for a more structured presentation of the findings and facilitates comparison across different application contexts of neural networks in psychological research. Notably, this organization also highlights the uneven distribution of studies across domains, with a strong predominance of clinical applications, suggesting a possible research bias or a greater maturity of neural network approaches within clinical psychology compared to other areas.

3.1 Enhancing Assessment and Diagnosis through ANNs

ANNs are increasingly used to enhance psychological assessment and diagnosis across two complementary levels: (a) the methodological (psychometric) level, where ANNs support the construction, optimization, and scoring of instruments; and (b) the clinical-applicative and health psychology level, where ANNs improve diagnosis, differential diagnosis, outcome prediction, and the personalization of intervention strategies. In this section, we integrate both levels: first the methodological dimension (psychometrics) and then the applicative domains of diagnosis, differential diagnosis, monitoring and intervention.

3.1.1 ANNs in the Methodological Realm

Psychometrics constitutes the empirical foundation of psychological assessment, offering clinicians standardized and scientifically validated instruments to evaluate individual differences with accuracy and consistency. By operationalizing complex psychological constructs into measurable dimensions, psychometric methodologies ensure that diagnostic formulations are not merely subjective interpretations but empirically grounded judgments. This quantitative rigor enhances the reliability and validity of both assessment and diagnosis, facilitating more precise case conceptualizations and targeted interventions. Within clinical practice, the integration of psychometric principles thus enables a systematic and evidence-based approach to understanding psychological functioning, ultimately improving diagnostic accuracy and informing effective treatment planning.

Our scoping review encompasses six studies on the application of ANNs in Psychometrics.

Item selection is a critical aspect of psychometric research. In this field, ANNs have been utilized to improve both the efficiency and coherence of the process. Notably, the studies by Dolce et al. [27] and Casella et al. [30] demonstrate how ANNs could optimize and automate the item selection process, leading to significant reductions in time and resources; in particular, Dolce et al. [27] reported a predictive accuracy improvement of up to 88.5% while reducing redundancy in psychometric scales. These studies underscore the ability of ANNs to retain the dimensional integrity of the original tests while enhancing their predictive capabilities. This approach offers a flexible and distribution-free alternative to traditional methods, showcasing the transformative potential of these models in psychometric research.

Analyzing psychometric data can be quite challenging due to the diverse range of data modalities involved, spanning from Likert scales to open-ended responses and the analysis of social media data.

ANNs have been extensively employed in this field, demonstrating remarkable adaptability to a wide range of problems.

One of the most common types of data is Likert Scale data, which, however, presents some challenges in analysis. An interesting application of ANNs in this context is missing data imputation. Traditional methods often introduce bias, particularly with asymmetric Likert-type items. However, research such as that by Collier et al. [55], has proposed an alternative approach based on ANNs demonstrating their efficacy in improving imputation accuracy (overall f1-score of 0.67 versus 0.22 for classical logistic regression). These methods outperform traditional statistical techniques by accommodating diverse data types and capturing complex, non-linear relationships between variables without relying on distributional assumptions.

Analyzing open-ended responses demands considerable time and effort. In this domain, ANNs have proven to be effective, as demonstrated by the studies of Devine et al. [56] and Thompson et al. [57]. In particular, Devine et al. [56] focused on automating Theory of Mind (ToM) scoring, achieving a high level of agreement with human raters (yielding a mean Krippendorff's α of 0.86), while substantially reducing scoring time. This not only improves assessment reliability but also enables large-scale research by minimizing manual labor.

Similarly, Thompson et al. [57] investigated ANNs' application in automating the scoring of open-ended candidate responses in pre-hire employment selection assessments. The ANN model exhibited a strong average correlation with subject matter experts' (SME) ratings (avg r = 0.84), nearly matching SMEs' inter-rater reliability. These findings highlight ANNs' potential to closely align with human judgment in scoring responses.

Another challenge for psychometric research is to analyze social media interactions. In this domain, ANNs have been applied to explore personality traits through social media interactions. The study by Mahajan et al. [58] emphasized the predictive capability of ANNs in mapping online behaviors to real-world personality characteristics, achieving an accuracy of 75.13% in predicting personality traits within the Big Five model. This research not only confirmed the reflection of real-world personalities in online expressions but also opened possibilities for future investigations in personality psychology.

3.1.2 ANNs in Applied Domains

To synthesize practical contributions, we regrouped all applied studies by the clinical function that ANNs serve: screening/diagnosis, monitoring/prognosis, risk detection, treatment personalization/adherence, and prevention. This taxonomy, which identifies subgroups in clinical and health psychology, clarifies where ANN pipelines most directly augment assessment and decision making.

Clinical and health psychology share a common conceptual foundation in understanding the bidirectional relationship between psychological processes and physical health outcomes.

While clinical psychology traditionally emphasizes the assessment and treatment of psychopathology, health psychology extends these principles to the promotion of wellbeing and the prevention of illness within broader biopsychosocial frameworks. The convergence of clinical and health psychology is evident in their mutual focus on behavioral change, emotion regulation, and coping mechanisms as key determinants of health trajectories. Both domains advocate for evidence-based, person-centered approaches that recognize the interdependence of mind and body in the maintenance and restoration of health and both domains can benefit from ANNs contribution. The application of ANNs in clinical psychology is broadening, as evidenced by numerous studies across a range of disorders.

Twenty-two studies collectively underscore the utility of ANNs in enhancing diagnostics [26], predicting outcomes, and personalizing therapeutic interventions. These applications are diverse, analyzing various data types such as images, sensor data, and item responses, and spanning different age groups, from developmental disorders in children to cognitive disorders in older adults.

Considering the application to health psychology, ANNs have found application particularly in the context of prevention and intervention.

Diagnostic Screening and Differential Diagnosis

ANN-based models enhance differential diagnosis in developmental conditions by exploiting fine-grained behavioral signatures. In diagnosing neurodivergences such as Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactivity Disorder (ADHD), ANNs have shown significant potential. They help automate diagnostic processes and refine differential diagnoses through the analysis of unique data types, like motion patterns. For instance, Milano et al. [33] improved diagnostic precision (91.2% of accuracy) for ASD by using ANNs to identify key motion features. Kumar et al. [34] automated the diagnosis of ASD using ML models, including ANNs, achieving up to 100% accuracy with Support Vector Machines on complete data and maintaining this perfect classification even with missing data using Recursive Feature Elimination, while Lin et al. [35] applied ANNs to differentiate types of ADHD in children, achieving high accuracy (0.73 for the ANN) and highlighting the tool's potential in clinical diagnostics. In the domain of anxiety disorders, Na et al. [36] demonstrated that ML models, including ANN-based approaches, can effectively discriminate panic disorder from other anxiety disorders (accuracy of 0.73), improving diagnostic specificity in clinically overlapping conditions. In diagnosing schizophrenia, obsessive-compulsive disorder (OCD), and problematic social media usage, ANNs have enhanced diagnostic accuracy by integrating data from multiple sources. Lee et al. [37] and Jeon et al. [38] used ANNs to improve diagnostic tools and identify non-invasive biomarkers; specifically, Jeon et al. [38] applied a deep learning residual masking network (with >74% accuracy) to strictly validate neutral facial stimuli in real-life interactions, revealing significant clinical correlations between gaze avoidance and schizophrenia symptom severity (e.g., $r = -0.401$ for PANSS total score). Meanwhile, Savci et al. [39] applied ML approaches to predict problematic social media use, framing maladaptive online behaviors as clinically relevant outcomes amenable to ANN-based classification (prediction rate of 0.62). Shahzad et al. [40] successfully differentiated OCD patients from healthy controls using an ANN model that achieved an impressive 98.2% testing accuracy (AUC = 0.991). The network's normalized importance analysis identified cleaning and contamination (100%) alongside a novel "worth in family" factor (71.1%) as the strongest predictors, emphasizing the critical importance of considering family dynamics in diagnostics. Lastly, in the area of cognitive impairments, such as Alzheimer's and dementia, ANNs have been instrumental in predicting disease progression, thereby informing appropriate interventions. Wu et al. [41] used a deep learning survival approach to estimate the probability and time of cognitive impairment stage conversion, demonstrating the model's capability to handle complex longitudinal data with high predictive accuracy (achieving a concordance index of at least 86% and an integrated Brier score of less than 0.1 for each disease stage). Amini et al. [28] applied ANNs to Clock Drawing Test images to detect early signs of cognitive impairments with remarkable results (average AUC = 91.9%; F1 score = 94.6%).

Symptom Monitoring and Prognostic Modelling

In the realm of mood and anxiety disorders, ANNs continue to open new avenues for predicting symptom progression. Wang et al. [42] used an LSTM encoder–decoder to represent daily anxiety dynamics; the reconstructed feature space correlated with the observed anxiety values ($r = 0.624$ in the training data), and clustering showed that adolescents with similar GAD-7 scores could display different day-to-day symptom patterns. Jacobson et al. [43] used wearable actigraphy and an ensemble deep-learning pipeline to predict 17–18-year deterioration in anxiety-disorder symptoms, reporting an AUC of 0.696, 84.6% sensitivity, 52.7% specificity, and balanced accuracy of 68.7%. Building on the potential of digital phenotyping, Jacobson and Bhattacharya [44] demonstrated that personalized (idiographic) deep-learning models using smartphone sensor data explained 74.8% of total variation across persons (robust R^2 = 0.748, 95% CI [0.728, 0.766]) and, on average, 38.5% of within-person hour-to-hour variation (mean robust R^2 = 0.385), illustrating their potential for individualized monitoring.

Risk Detection and Early Warning

Regarding depression, ANNs have proven effective in analyzing speech data and exploring various impact factors. Little et al. [29] found a correlation between reduced speech activity, related to psychomotor retardation, and late-life depression. They utilized two deep learning models to objectively detect speech as a proxy for social interaction: the first model classified speech versus non-speech (achieving 93.8% accuracy), while the second differentiated the wearer's speech from others (89.95% accuracy).

Everaert et al. [45] used ANN-based feature selection to identify specific features of repetitive negative thinking and positive reappraisal that predicted individual depressive symptoms beyond perceived stress; model AUCs ranged from 0.63 to 0.81. Zhang et al. [46] developed a transformer-based model for the narrower task of classifying suicide notes in a dataset compiled from online sources. The model achieved average precision, recall, and F1 scores of approximately 95%; these findings should not be generalized to clinical detection of suicidal tendencies. Shin and Kim [47] used several statistical and machine-learning models to predict suicidal ideation among youth, with a convolutional neural network achieving approximately 90% accuracy. Beyond affective disorders, ANNs have also been applied to emotional-state recognition in childhood. Rao et al. [51] developed a platform using a Backpropagation Neural Network (BPNN) to classify calm, happiness, and fear from 12 physiological-signal features in preschool children. The model achieved an average recognition accuracy of 73%, reaching 81% among 6-year-olds.

Treatment Personalization and Adherence

The study by Singh et al. [48] applied deep-learning models to predict daily adherence to a cognitive-training program. CNN, LSTM, and CNN-LSTM architectures achieved mean F-scores ranging from 74.6% to 75.5%. Sandeep et al. [49] compared input–output hidden Markov models, an unscented Kalman filter, and an LSTM for estimating participant skill and predicting performance at candidate task challenges. The models were evaluated as estimators and were not deployed to control task difficulty. Rennie et al. [50], by contrast, used self-organizing maps and K-means clustering to identify four response profiles in an independent baseline sample (N = 616) and then mapped the pre- and post-training trajectories of 179 children; pre-training fluid intelligence predicted profile membership. Thus, Sandeep et al. evaluated models for estimating skill and predicting appropriate future challenges, whereas Rennie et al. characterized heterogeneity in training response rather than adjusting difficulty in real time.

Population-Level Prevention and Resilience

Two studies highlight the application of ANNs in health psychology during the COVID-19 pandemic. Nursalim et al. [52] examined how demographic variables, stress intensity, and mindfulness influenced resilience. The ANN model achieved approximately 65% predictive performance in the holdout phase, with stress intensity and mindfulness emerging as the most influential predictors. Morales-Rodríguez et al. [53] used a multilayer perceptron to predict stress levels in a university community, reporting predictive performance above 80% and identifying coping strategies such as negative self-focus and positive re-evaluation as influential predictors. These studies illustrate the use of ANNs to model psychological outcomes from behavioral, psychological, and demographic variables, but their findings do not by themselves establish the effectiveness of AI-delivered interventions. In prevention-oriented health psychology, Chadaga et al. [54] compared multiple supervised machine-learning models, including a deep neural network, with multi-level stacked ensembles of non-neural classifiers. The final stacked ensemble achieved 94% accuracy, whereas the deep neural network achieved 89%; explainability analyses highlighted knowledge of child grooming and recognition of signs of abuse as influential predictors. The study therefore represents a broader machine-learning comparison that included an ANN model, not an ANN-based stacked ensemble.

4 Discussion

This scoping review shows that ANNs have been applied across the assessment pathway, from psychometric scoring to diagnostic classification, longitudinal monitoring, risk detection, and intervention tailoring. The evidence demonstrates methodological versatility, but it also exposes a gap between research performance and clinical readiness.

4.1 Evidence, Prediction, and Clinical Meaning

ANNs belong to a predictive modelling tradition that complements, rather than replaces, explanatory psychological research [59,60]. A model may accurately estimate an outcome without identifying its causes, and variables that optimize prediction are not automatically appropriate targets for intervention. Prediction should therefore be separated from clinical decision-making: diagnosis and treatment require integration of model output with history, interview findings, comorbidity, contextual factors, patient preferences, and professional standards.

The clinical value of these models cannot be inferred from discrimination metrics alone. Patient-centered outcomes include timeliness and appropriateness of referral, symptom improvement, prevention of adverse events, patient understanding and acceptability, clinician workload, access to care, and the distribution of benefits and harms across populations. ANNs may support these goals, but evidence of accurate prediction is not itself evidence of improved care.

The reviewed studies indicate that architecture should follow the data and task. Feed-forward networks were common for structured psychometric variables, CNNs for spatial or image-derived representations, and recurrent or LSTM models for longitudinal and sequential data. This task-specific flexibility is an advantage, but it does not make ANNs universally preferable. Simpler models may provide comparable performance, better calibration, and clearer explanations, particularly in small structured datasets.

The narrative synthesis distinguishes five functions that should not be conflated: psychometric construction and scoring; diagnostic classification and differential diagnosis; symptom monitoring and prognosis; risk detection; and treatment personalization or adherence. Each function entails different validation requirements, error costs, clinical pathways, and patient-centered outcomes.

4.2 Methodological Robustness and Generalizability

The promising applications must be weighed against pronounced dataset heterogeneity. Samples ranged from 22 participants to tens of thousands of individuals and included psychometric responses, text, images, speech, behavioral traces, and sensor streams. Small samples combined with high-dimensional inputs create substantial risks of overfitting, data leakage, class imbalance, unstable feature importance, and optimistic performance estimates. Exceptionally high accuracy should therefore be treated as a signal for independent replication rather than evidence of clinical certainty.

Validation procedures were also inconsistent. Many studies relied on a single training/testing split, while others used different forms of cross-validation; external validation in an independent clinical population was uncommon. Internal resampling can estimate performance within the source dataset, but it cannot establish transportability across hospitals, countries, languages, devices, demographic groups, or changes in clinical practice. Multicenter external validation, prospective temporal validation, calibration assessment, and subgroup-specific error analysis are necessary before deployment.

These requirements distinguish research-oriented applications from clinically deployable systems. A promising retrospective classifier is a proof of concept; a deployable system must additionally demonstrate reproducibility, workflow compatibility, prospective benefit, safety under distribution shift, maintenance procedures, and accountable ownership throughout its lifecycle. The predominance of clinical classification studies over other psychological domains may reflect differences in data availability as well as uneven methodological maturity.

Intervention-delivery systems such as virtual therapists, chatbots, and AI-supported psychotherapy were outside the assessment-focused scope of this review [61]. Their clinical effectiveness and safety require separate evaluation rather than inference from the performance of assessment models.

The substantial variation in outcomes, data modalities, model architectures, sample sizes, and validation strategies limits direct comparison of model performance across studies.

Architectures, activation functions, optimizers, preprocessing pipelines, class-balancing procedures, and hyperparameter choices varied considerably, limiting direct comparison across studies.

Training, validation, and testing procedures were also inconsistent. Studies used different performance metrics and data partitions, and many did not report calibration, uncertainty, threshold selection, or subgroup-specific errors.

Many studies did not provide accessible code, sufficiently detailed analytic workflows, or appropriately governed data resources, which further limits replication and independent verification.

These sources of heterogeneity should be treated as central determinants of the reported results and of the comparability of model performance across studies.

4.3 Multimodal Assessment and Objective Biomarkers

Although studies based solely on neuroimaging or physiological signals were excluded, multimodal psychiatric assessment is a logical future direction. Imaging, EEG, wearable physiology, speech, behavior, ecological self-report, and psychometric instruments may provide complementary information at different temporal and functional levels [62–69]. Objective biomarkers should complement, not displace, psychometric and clinical assessment, and their incremental validity, fairness, cost, burden, missingness, and interpretability must be evaluated against simpler pathways.

4.4 Interpretability, Fairness, and Ethical Governance

The black-box challenge is particularly consequential in psychological assessment because model outputs may influence labels, treatment, education, employment, or access to services. Post-hoc explanations can indicate which features contributed to a prediction, but they do not necessarily reveal a stable causal mechanism or prove that the model relied on clinically legitimate information [70].

Transparent reporting is essential for reproducibility. Architectures, preprocessing, feature selection, class balancing, hyperparameter tuning, threshold selection, missing-data handling, and every partition used during model development should be documented. Open code and appropriately governed data can facilitate replication, while reporting guidance such as TRIPOD+AI provides a common basis for evaluating clinical prediction-model studies [71].

Algorithmic bias may arise when training data underrepresent demographic, cultural, linguistic, or socioeconomic groups; when measures function differently across populations; or when historical service inequalities are encoded in labels and outcomes [3]. Overall performance can conceal clinically important subgroup failures. Studies should report sample composition, assess measurement invariance where relevant, compare calibration and error rates across groups, and test whether deployment could widen existing disparities.

Ethical implementation also requires proportionate and transparent data use, privacy and cybersecurity safeguards, meaningful human oversight, mechanisms for contesting or correcting outputs, and clear communication about uncertainty. The WHO guidance on AI for health emphasizes that ethics and human rights should be embedded in design, deployment, and governance rather than added after technical development [6].

Regulatory and medico-legal responsibility cannot be delegated to a model. In the European Union, high-risk AI systems may be subject to requirements concerning risk management, data governance, transparency, human oversight, accuracy, cybersecurity, and post-market monitoring [7]. Services must define who verifies model suitability, who acts on an alert, how disagreement between clinician and system is documented, and how patients are protected when an error occurs.

4.5 From Research Prototype to Psychiatric Service

Routine implementation requires more than installing software. Psychiatric services need secure data pipelines, interoperability with clinical records, acceptable response times, staff training, technical support, and procedures for missing or low-quality data. Decision thresholds should reflect the clinical pathway and the relative consequences of false-positive and false-negative results rather than being selected only to maximize a summary metric. ANNs should be positioned as decision-support tools that can prioritize information, prompt further assessment, or identify changes requiring review. Clinicians remain responsible for interpreting outputs in context and discussing them with patients. Prospective evaluations should examine how a system changes decisions and outcomes, including automation bias, alert fatigue, unnecessary referrals, and reduced attention to information not represented in the model. Deployment also requires continuous monitoring for calibration drift, changes in population characteristics, software updates, and emerging inequities. Early-stage evaluation frameworks such as DECIDE-AI emphasize human factors, workflow integration, and real-world safety before large-scale implementation [72]. Table 3 summarizes the principal barriers and corresponding priorities.

Table 3: Principal barriers to clinical implementation of ANN-based psychological assessment.

Priority Actions	Why It Matters	Barrier
Document provenance; assess subgroup coverage and measurement validity; use clinically justified variables.	Missingness, label error, and underrepresented groups can create unstable or biased predictions.	Data quality and representativeness
Perform multicenter external and temporal validation; report calibration and subgroup errors.	Internal accuracy may not transfer across sites, cultures, languages, devices, or time.	Validation and generalizability

Table 3: *Cont.*

Priority Actions	Why It Matters	Barrier
Evaluate explanation fidelity and usefulness; communicate uncertainty; retain clinician review.	Post-hoc explanations may be unstable or clinically misleading.	Interpretability
Conduct prospective impact studies measuring safety, symptoms, access, acceptability, and workload.	High discrimination does not demonstrate improved care.	Clinical utility and patient outcomes
Co-design with users; integrate secure data flows; define escalation and downtime procedures.	Poor integration can increase burden, alert fatigue, and workarounds.	Workflow and interoperability
Audit bias; enable contestability; involve patients and affected communities.	Models may reproduce structural inequalities or use sensitive data without adequate transparency.	Fairness and ethics
Apply data minimization, access controls, secure processing, and incident-response procedures.	Psychological and behavioral data are highly sensitive.	Privacy and cybersecurity
Assign clinical and technical ownership; document decisions; monitor drift; update and revalidate.	Accountability may be unclear as models and clinical environments change.	Regulation, liability, and maintenance

The historical relationship between psychology and AI remains relevant [73], but current clinical adoption must be governed by evidence rather than technological novelty. Professionals who use ANN outputs need sufficient understanding of model limitations, uncertainty, and failure modes to interpret them responsibly.

4.6 Limitations of the Review

This scoping review has several limitations. The search was restricted to Scopus and Web of Science and to literature available by 26 January 2024, which may have excluded relevant grey literature and later developments. Study selection and thematic extraction involved qualitative judgment, and no formal risk-of-bias instrument was applied. Excluding Large Language Models, purely neuroimaging studies, and physiological models without explicit psychological outcomes improved conceptual coherence but limits coverage of the wider AI ecosystem. Publication bias may also favor studies reporting strong performance, while heterogeneity prevented quantitative synthesis.

4.7 Future Research Priorities

Future research should prioritize multicenter and cross-cultural external validation; prospective studies of clinical impact; calibration and error analysis across demographic and socioeconomic subgroups; standardized reporting of data, training, and validation procedures; and open, reproducible workflows where privacy permits. Studies should compare ANNs with simpler baselines, evaluate multimodal models only when each modality has a justified contribution, incorporate patient and clinician perspectives, and define patient-centered endpoints before model development.

5 Conclusions

This scoping review shows that ANNs can support multiple components of psychological assessment, including psychometric scoring, diagnostic classification, symptom monitoring, risk detection, and intervention tailoring. Their principal contribution lies in modeling complex and heterogeneous data, but the current evidence base remains dominated by research-oriented, internally validated studies.

Widespread clinical implementation is therefore premature. Before ANN-based assessment can be integrated into routine psychiatric services, models require independent and prospective validation, transparent reporting, clinically meaningful interpretation, fairness auditing, secure governance, regulatory compliance, and evidence of benefit for patients and professionals. Used under these conditions, ANNs may strengthen clinical decision support; they should not be treated as autonomous diagnosticians or replacements for clinician judgment.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Michela Ponticorvo and Monica Casella; methodology, Michela Ponticorvo, Monica Casella and Raffaella Esposito; software, Monica Casella; validation, Michela Ponticorvo, Monica Casella and Raffaella Esposito; formal analysis, Michela Ponticorvo, Monica Casella and Raffaella Esposito; investigation, Michela Ponticorvo, Monica Casella and Raffaella Esposito; resources, Michela Ponticorvo, Monica Casella and Raffaella Esposito; data curation, Michela Ponticorvo, Monica Casella and Raffaella Esposito; writing—original draft preparation, Michela Ponticorvo, Monica Casella and Raffaella Esposito; writing—review and editing, Michela Ponticorvo; visualization, Michela Ponticorvo, Monica Casella and Raffaella Esposito; supervision, Michela Ponticorvo. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The extracted data through the scoping review process is presented in the manuscript and supplementary materials.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/ijmhp.2026.079327/s1.

References

1. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak. 2021;21(1):125. [CrossRef].
2. Zhou S, Zhao J, Zhang L. Application of artificial intelligence on psychological interventions and diagnosis: an overview. Front Psychiatry. 2022;13:811665. [CrossRef].
3. Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLoS Digit Health. 2023;2(6):e0000278. [CrossRef].
4. Curzon J, Kosa TA, Akalu R, El-Khatib K. Privacy and artificial intelligence. IEEE Trans Artif Intell. 2021;2(2):96–108. [CrossRef].
5. Tian M, Shen Z, Wu X, Wei K, Liu Y. The application of artificial intelligence in medical diagnostics: a new frontier. Acad J Sci Technol. 2023;8(2):57–61. [CrossRef].
6. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva, Switzerland: World Health Organization; 2021.
7. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence. Brussels, Belgium: European Parliament and Council of the European Union; 2024.
8. Stade E, Stirman SW, Ungar LH, Yaden DB, Schwartz HA, Sedoc J, et al. Artificial intelligence will change the future of psychotherapy: A proposal for responsible, psychologist-led development. PsyArXiv. 2023:1–29.

9. Zhou ZH. Machine Learning. Berlin/Heidelberg, Germany: Springer; 2021.
10. Hassoun MH. Fundamentals of Artificial Neural Networks. Cambridge, MA, USA: MIT Press; 1995. [CrossRef].
11. Zou J, Han Y, So SS. Overview of artificial neural networks. In: Artificial Neural Networks: Methods and Applications. Totowa, NJ, USA: Humana Press, 2009. p. 14–22. [CrossRef].
12. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117. [CrossRef].
13. Schank RC. The current state of AI: one man's opinion. AI Mag. 1983;4(1):3–8.
14. Urban CJ, Gates KM. Deep learning: a primer for psychologists. Psychol Methods. 2021;26(6):743–73. [CrossRef].
15. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386–408. [CrossRef].
16. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5:115–33. [CrossRef].
17. Bishop CM. Pattern recognition and machine learning. Berlin/Heidelberg, Germany: Springer; 2006 [CrossRef].
18. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA, USA: The MIT Press; 2016.
19. Elman JL. Finding structure in time. Cogn Sci. 1990;14(2):179–211. [CrossRef].
20. Haykin S. Neural networks: a comprehensive foundation. 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall; 1999.
21. Cunningham P, Cord M, Delany SJ. Supervised learning. In: Machine learning techniques for multimedia: case studies on organization and retrieval. Berlin/Heidelberg, Germany: Springer; 2008. p. 21–49. [CrossRef].
22. Barlow HB. Unsupervised learning. Neural Comput. 1989;1(3):295–311. [CrossRef].
23. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6. [CrossRef].
24. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern. 1982;43(1):59–69. [CrossRef].
25. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A. 1982;79(8):2554–8. [CrossRef].
26. Xue K, Bradshaw LP. A semi-supervised learning-based diagnostic classification method using artificial neural networks. Front Psychol. 2020;11:618336. [CrossRef].
27. Dolce P, Marocco D, Maldonato MN, Sperandeo R. Toward a machine learning predictive-oriented approach to complement explanatory modeling. An application for evaluating psychopathological traits based on affective neurosciences and phenomenology. Front Psychol. 2020;11:446. [CrossRef].
28. Amini S, Zhang L, Hao B, Gupta A, Song M, Karjadi C, et al. An artificial intelligence-assisted method for dementia detection using images from the clock drawing test. J Alzheimers Dis. 2021;83(2):581–9. [CrossRef].
29. Little B, Alshabrawy O, Stow D, Ferrier IN, McNaney R, Jackson DG, et al. Deep learning-based automated speech detection as a marker of social functioning in late-life depression. Psychol Med. 2021;51(9):1441–50. [CrossRef].
30. Casella M, Dolce P, Ponticorvo M, Milano N, Marocco D. Artificial neural networks for short-form development of psychometric tests: a study on synthetic populations using autoencoders. Educ Psychol Meas. 2024;84(1):62–90. [CrossRef].
31. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467–73. [CrossRef].
32. Aria M, Cuccurullo C. Bibliometrix: an R-tool for comprehensive science mapping analysis. J Informetr. 2017;11(4):959–75. [CrossRef].
33. Milano N, Simeoli R, Rega A, Marocco D. A deep learning latent variable model to identify children with autism through motor abnormalities. Front Psychol. 2023;14:1194760. [CrossRef].
34. Kumar CJ, Das PR. The diagnosis of ASD using multiple machine learning techniques. Int J Dev Disabil. 2022;68(6):973–83. [CrossRef].
35. Lin IC, Chang SC, Huang YJ, Kuo TBJ, Chiu HW. Distinguishing different types of attention deficit hyperactivity disorder in children using artificial neural network with clinical intelligent test. Front Psychol. 2022;13:1067771. [CrossRef].

36. Na KS, Cho SE, Cho SJ. Machine learning-based discrimination of panic disorder from other anxiety disorders. J Affect Disord. 2021;278:1–4. [CrossRef].
37. Lee SC, Chen KW, Liu CC, Kuo CJ, Hsueh IP, Hsieh CL. Using machine learning to improve the discriminative power of the FERD screener in classifying patients with schizophrenia and healthy adults. J Affect Disord. 2021;292:102–7. [CrossRef].
38. Jeon G, Choi HS, Jung DU, Moon S, Kim G, Kim SJ, et al. Evaluation of the correlation between gaze avoidance and schizophrenia psychopathology with deep learning-based emotional recognition. Asian J Psychiatr. 2022;68:102974. [CrossRef].
39. Savci M, Tekin A, Elhai JD. Prediction of problematic social media use (PSU) using machine learning approaches. Curr Psychol. 2022;41(5):2755–64. [CrossRef].
40. Shahzad MN, Suleman M, Ahmed MA, Riaz A, Fatima K. Identifying the symptom severity in obsessive-compulsive disorder for classification and prediction: an artificial neural network approach. Behav Neurol. 2020;2020:2678718. [CrossRef].
41. Wu X, Peng C, Nelson PT, Cheng Q. Machine learning approach predicts probability of time to stage-specific conversion of Alzheimer's disease. J Alzheimers Dis. 2022;90(2):891–903. [CrossRef].
42. Wang B, Nemesure MD, Park C, Price GD, Heinz MV, Jacobson NC. Leveraging deep learning models to understand the daily experience of anxiety in teenagers over the course of a year. J Affect Disord. 2023;329:293–9. [CrossRef].
43. Jacobson NC, Lekkas D, Huang R, Thomas N. Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17-18 years. J Affect Disord. 2021;282:104–11. [CrossRef].
44. Jacobson NC, Bhattacharya S. Digital biomarkers of anxiety disorder symptom changes: personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. Behav Res Ther. 2022;149:104013. [CrossRef].
45. Everaert J, Benisty H, Gadassi Polack R, Joormann J, Mishne G. Which features of repetitive negative thinking and positive reappraisal predict depression? An in-depth investigation using artificial neural networks with feature selection. J Psychopathol Clin Sci. 2022;131(7):754–68. [CrossRef].
46. Zhang T, Schoene AM, Ananiadou S. Automatic identification of suicide notes with a transformer-based deep learning model. Internet Interv. 2021;25:100422. [CrossRef].
47. Shin S, Kim K. Prediction of suicidal ideation in children and adolescents using machine learning and deep learning algorithm: a case study in South Korea where suicide is the leading cause of death. Asian J Psychiatr. 2023;88:103725. [CrossRef].
48. Singh A, Chakraborty S, He Z, Tian S, Zhang S, Lustria MLA, et al. Deep learning-based predictions of older adults' adherence to cognitive training to support training efficacy. Front Psychol. 2022;13:980778. [CrossRef].
49. Sandeep S, Shelton CR, Pahor A, Jaeggi SM, Seitz AR. Application of machine learning models for tracking participant skills in cognitive training. Front Psychol. 2020;11:1532. [CrossRef].
50. Rennie JP, Zhang M, Hawkins E, Bathelt J, Astle DE. Mapping differential responses to cognitive training using machine learning. Dev Sci. 2020;23(4):e12868. [CrossRef].
51. Rao Z, Wu J, Zhang F, Tian Z. Psychological and emotional recognition of preschool children using artificial neural network. Front Psychol. 2021;12:762396. [CrossRef].
52. Nursalim M, Saroinsong WP, Boonroungrut C, Wagino, da Costa A. Prediction of students' COVID-19 resilience: an artificial neural network based on gender, age, stress intensity, and mindfulness variables. Psychol Sch. 2023;60(9):3253–65. [CrossRef].
53. Morales-Rodríguez FM, Martínez-Ramón JP, Méndez I, Ruiz-Esteban C. Stress, coping, and resilience before and after COVID-19: a predictive model based on artificial intelligence in the university environment. Front Psychol. 2021;12:647964. [CrossRef].
54. Chadaga K, Prabhu S, Sampathila N, Chadaga R, Bairy M, Swathi KS. An explainable framework to predict child sexual abuse awareness in people using supervised machine learning models. J Technol Behav Sci. 2024;9(2):346–62. [CrossRef].
55. Collier ZK, Kong M, Soyoye O, Chawla K, Aviles AM, Payne Y. Deep learning imputation for asymmetric and incomplete likert-type items. J Educ Behav Stat. 2024;49(2):241–67. [CrossRef].

56. Devine RT, Kovatchev V, Grumley Traynor I, Smith P, Lee M. Machine learning and deep learning systems for automated measurement of "advanced" theory of mind: reliability and validity in children and adolescents. Psychol Assess. 2023;35(2):165–77. [CrossRef].
57. Thompson I, Koenig N, Mracek DL, Tonidandel S. Deep learning in employee selection: evaluation of algorithms to *Automate* the scoring of open-ended assessments. J Bus Psychol. 2023;38(3):509–27. [CrossRef].
58. Mahajan R, Mahajan R, Sharma E, Mansotra V. "Are we tweeting our real selves?" personality prediction of Indian Twitter users using deep learning ensemble model. Comput Hum Behav. 2022;128:107101. [CrossRef].
59. Shmueli G. To explain or to predict? Statist Sci. 2010;25(3):289–310. [CrossRef].
60. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci. 2017;12(6):1100–22. [CrossRef].
61. Sadeh-Sharvit S, Del Camp T, Horton SE, Hefner JD, Berry JM, Grossman E, et al. Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: randomized clinical trial. J Med Internet Res. 2023;25:e46781. [CrossRef].
62. Caria A, Grecucci A. Neuroanatomical predictors of real-time fMRI-based anterior insula regulation. A supervised machine learning study. Psychophysiology. 2023;60(5):e14237. [CrossRef].
63. Williams KS. Evaluations of artificial intelligence and machine learning algorithms in neurodiagnostics. J Neurophysiol. 2024;131(5):825–31. [CrossRef].
64. Hussain I, Jany R, Boyer R, Azad A, Alyami SA, Park SJ, et al. An explainable EEG-based human activity recognition model using machine-learning approach and LIME. Sensors. 2023;23(17):7452. [CrossRef].
65. Orrù G, Monaro M, Conversano C, Gemignani A, Sartori G. Machine learning in psychometrics and psychological research. Front Psychol. 2020;10:2970. [CrossRef].
66. Shajari S, Kuruvinashetti K, Komeili A, Sundararaj U. The emergence of AI-based wearable sensors for digital health technology: a review. Sensors. 2023;23(23):9498. [CrossRef].
67. Jafari M, Shoeibi A, Khodatars M, Bagherzadeh S, Shalbaf A, García DL, et al. Emotion recognition in EEG signals using deep learning methods: a review. Comput Biol Med. 2023;165:107450. [CrossRef].
68. Vempati R, Sharma LD. A systematic review on automated human emotion recognition using electroencephalogram signals and artificial intelligence. Results Eng. 2023;18:101027. [CrossRef].
69. Khare SK, Blanes-Vidal V, Nadimi ES, Acharya UR. Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations. Inf Fusion. 2024;102:102019. [CrossRef].
70. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–50. [CrossRef].
71. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. [CrossRef].
72. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28:924–33. [CrossRef].
73. Cordeschi R. AI turns fifty: revisiting its origins. Appl Artif Intell. 2007;21(4–5):259–79. [CrossRef].