

KU LEUVEN

FACULTEIT PSYCHOLOGIE EN
PEDAGOGISCHE WETENSCHAPPEN

**Exploring AI Approaches to Identify Social Anxiety Disorder:
A Scoping Review**

Masterproef aangeboden tot het
verkrijgen van de graad van Master of
Science in de psychologie door

Charlotte Van Bogaert

Promotor: Prof. Dr. Sara Scheveneels

Copromotor: Dr. Tom Van Daele



Copyright informatie:

*Studiewerk van een student in het kader van de
academische opleiding en examenbeoordeling.*

*Na deze beoordeling vond geen correctie plaats
van het studiewerk.*

KU LEUVEN

FACULTEIT PSYCHOLOGIE EN
PEDAGOGISCHE WETENSCHAPPEN

**Exploring AI Approaches to Identify Social Anxiety Disorder:
A Scoping Review**

Masterproef aangeboden tot het
verkrijgen van de graad van Master of
Science in de psychologie door

Charlotte Van Bogaert

Promotor: Prof. Dr. Sara Scheveneels

Copromotor: Dr. Tom Van Daele

Summary

Social anxiety disorder (SAD) is a common mental health condition associated with substantial impairments in functioning, yet it often goes undetected in practice. Its identification has traditionally relied on clinical interviews and self-report measures, which are time-consuming and subject to limitations such as recall bias, limited insight and social desirability. Against this background, this master's thesis examines the role that artificial intelligence (AI) may play in identifying individuals with SAD or clinically relevant symptoms of social anxiety.

This master's thesis takes the form of a scoping review and maps the peer-reviewed empirical literature on the use of AI approaches to identify social anxiety. The central questions are which types of AI are used to detect or classify social anxiety on the basis of psychological and behavioral data, which data types are used as input for these models and how well these AI models perform. In addition, the review examines how these studies were designed, in which populations and contexts these approaches were investigated, and which ethical and practical considerations emerge from the literature. The review follows Joanna Briggs Institute guidelines and was reported in accordance with PRISMA-ScR. The following databases were searched: APA PsycINFO, Embase, Scopus, PubMed/MEDLINE, Web of Science and Google Scholar. Study selection was conducted in Rayyan. After deduplication, 203 studies underwent title and abstract screening, of which 46 proceeded to full-text screening. In total, 17 studies were included. Data were collected using a standardized extraction form and subsequently summarized in tables and narrative syntheses.

This review shows that AI has potential as a supportive tool for identifying social anxiety, but that the current evidence base remains early-stage and heterogeneous. The included studies used both classical machine learning methods and deep learning methods. Questionnaire data were the most commonly used input for the models, although some models also relied on more complex data input such as digital traces, interview language and eye-tracking data, thereby enabling them to capture behavioral and linguistic patterns that are less directly visible in traditional questionnaires or interviews. At the same time, many studies were still proof-of-concept in nature, used varying definitions of SAD and relied primarily on internal validation. Further progress requires external validation, more transparent reporting, more diverse datasets, careful attention to privacy, bias and margins of error, as well as sufficient AI literacy among clinicians to interpret and use AI output responsibly.

Acknowledgements

I would like to sincerely thank my supervisor, Prof. Dr. Sara Scheveneels, and my co-supervisor, Dr. Tom Van Daele, for their guidance and trust throughout the development of this master's thesis. Through their support, I was given the opportunity to work on this self-chosen topic, situated at the intersection of AI and psychology, a theme that has aligned with my academic interests for quite some time. Their valuable feedback, critical reflections and suggestions throughout the writing process helped me shape this master's thesis with greater depth and academic rigor.

I would also like to thank Dr. Dries Van Daele for taking the time to review the final version of this thesis from the perspective of his expertise in AI, and for providing targeted feedback on how the AI-related aspects of this work could be further refined.

Finally, I would also like to thank my family and friends for their support and encouragement during the writing of this master's thesis.

Approach and Personal Contribution

The student was responsible for developing and refining the research questions, designing the search strategy, preparing and updating the Open Science Framework preregistration, and conducting and documenting the literature search. The student also organized the screening process in Rayyan, performed the data extraction, synthesized the findings and drafted and revised all sections of the thesis. The supervisor and co-supervisor provided substantive and methodological guidance in refining the scope of the review, including with regard to the inclusion and exclusion criteria. The student screened all sources as the first reviewer. To ensure a methodologically rigorous process, the supervisor and co-supervisor each additionally acted as second reviewer for half of the sources. Differences in assessment and cases of uncertainty were discussed in order to reach an appropriate selection. The data extraction, interpretation of the findings, narrative synthesis and final writing were carried out independently by the student. The student received feedback from both the supervisor and co-supervisor on the full version of the master's thesis on two occasions, each time first in written form, followed by an online meeting with room for additional questions and clarification.

Table of Contents

Summary.....	I
Acknowledgements.....	II
Approach and personal contribution.....	III
Abstract.....	VIII
Introduction	1
Methods.....	7
Results	8
Search Results.....	8
Study Characteristics.....	10
Social Anxiety Definition and Measurement	11
Input Data Modalities	13
AI Methods	14
Validation and Performance Reporting	14
Ethical and Practical Considerations	16
Discussion	21
Main Findings.....	21
Primary Research Questions: Results and Interpretation.....	21
<i>RQ1. Which types of AI based on psychological and behavioral data are used to detect or classify social anxiety?.....</i>	<i>21</i>
<i>RQ2. Which data and psychological/behavioral indicators are used to identify social anxiety?</i>	<i>22</i>
<i>RQ3. How well do the models perform (e.g., accuracy, sensitivity)?</i>	<i>23</i>
Methodological and Reporting Limitations and Evidence Gaps in the Literature	24
Strengths and Limitations of the Scoping Review.....	25
Clinical Implications and Recommendations for Future Research	26
Conclusion.....	28
Funding	29

Conflicts of Interest	29
Ethics Approval	29
References	30
Appendices	38
Appendix A.....	39
Appendix B.....	41
Appendix C.....	44

List of Tables

Table 1 Overview of Key AI Terms used in This Review.....	4
Table 2 General Characteristics of the Included Studies	12
Table 3 Key Findings and Methodological Characteristics of the Included Studies.....	17
Table B1 PRISMA-ScR Checklist	41
Table C1 Full Data Extraction: Design Features and Key Methodological Characteristics	44
Table C2 Full Data Extraction: AI/ML Approaches, Validation Strategies and Performance	51

List of Figures

Figure 1 PRISMA Flow Diagram of the Study Selection Proces.....	9
---	---

Abstract

Background: Social anxiety disorder (SAD) is often under-recognized despite its substantial impact on a person's functioning. As traditional identification relies mainly on interviews and self-report questionnaires, artificial intelligence (AI) may offer added value by detecting patterns in diverse data sources that could support earlier identification and help target limited clinical resources more effectively.

Objective: The objective of this scoping review is to map and synthesize literature on AI approaches used to identify SAD or clinically significant social anxiety symptoms using binary screening outcomes.

Methods: This scoping review was preregistered on the Open Science Framework (OSF) and conducted following Joanna Briggs Institute guidance, with reporting aligned to PRISMA-ScR. On 23 September 2025, searches were run in APA PsycINFO (EBSCO), Embase, Scopus, PubMed/MEDLINE, Web of Science Core Collection, and Google Scholar. Records were screened in Rayyan. Data were charted using a piloted extraction form and summarized with tables and narrative synthesis.

Results: Seventeen studies were included. Overall, AI approaches showed promising but still preliminary performance for identifying SAD, with best-reported performance of the AI model in correctly identifying SAD ranging from modest results in small eye-tracking datasets to high internal metrics in some questionnaire-based and language-based approaches. Questionnaire-derived data were the most common input modality, although studies also used scenario-elicited self-report ratings based on simulated social situations, digital traces, interview language, eye-tracking scan paths, virtual-reality anxiety trajectories, and multimodal microblogs. Many studies were proof-of-concept, sample sizes varied widely, SAD was operationalized in different ways, and validation was predominantly internal, usually through hold-out splits or k-fold cross-validation. Reported ethical and practical concerns included privacy, bias, classification errors, and limited attention to how AI models could be implemented in practice.

Conclusions: AI-based approaches showed clear potential for identifying SAD, but the evidence base remains early-stage and too heterogeneous to support strong conclusions about reliability or added value over existing approaches.

Keywords: Social Anxiety; Artificial Intelligence; Machine Learning; Detection; Screening

Introduction

Social anxiety disorder (SAD), also referred to as social phobia, involves fear of scrutiny by other people, leading to avoidance of social interactions or situations where a person could be observed or evaluated by others (e.g., meeting new people, eating or drinking in public, or performing in front of others) (American Psychiatric Association, 2013). The main concern for people with SAD is being judged negatively, such as being embarrassed, feeling rejected, or offending others (American Psychiatric Association, 2013).

SAD is a common condition worldwide and typically begins early in life. Data from the World Mental Health Survey Initiative (28 countries; 142,405 respondents) shows a SAD prevalence of 1.3% over 30 days, 2.4% over 12 months, and a 4.0% lifetime prevalence (Stein et al., 2017). For comparison, United States diagnostic interview data, summarized by the National Institute of Mental Health, report a 7.1% past-year prevalence and a 12.1% lifetime prevalence among adults. SAD is the most common anxiety disorder and tends to emerge early in life: by age 11, about 50% of affected individuals have developed the disorder, and by age 20, this rises to about 80% (Stein & Stein, 2008). This early age of onset underscores the importance of timely identification and intervention. Consistent with this, a meta-analysis estimated global prevalences at 4.7% in children, 8.3% in adolescents, and 17% in youth (Salari et al., 2024).

The disorder is associated with clinically meaningful impairment, including role impairment and adverse effects on education, work functioning, and relationships (Stein et al., 2017). It is also frequently comorbid with other mental health disorders (e.g., depression), which can amplify the overall burden (Stein et al., 2017). Despite its impact, SAD is often under-recognized and undertreated, with substantial delays between disorder onset and the first treatment contact (Wang et al., 2005). In a clinical sample of patients with generalized SAD, the mean time from onset to first treatment contact was 15 years, and the median time was 14 years (Ertekin et al., 2015). One important factor in this treatment delay is that SAD is frequently missed in routine care. In a large primary-care cohort of adults meeting DSM-IV criteria (Diagnostic and Statistical Manual of Mental Disorders) for SAD ($n=289$), only 23.6% reported that a healthcare professional had found out and informed them about their SAD (Chapdelaine et al., 2018). More broadly, a meta-analysis of 24 studies found that general practitioners' sensitivity for identifying anxiety disorders was 44.5% overall and only 30.5% when diagnoses were unassisted by diagnostic instruments (Olariu et al., 2015). This highlights the under-detection of SAD in primary care settings, where most patients first seek help.

Traditional assessment of SAD is typically based on a clinical interview conducted by a trained clinician to determine whether a person meets the diagnostic criteria. The clinician evaluates core features such as fear, avoidance, distress, and functional impairment (American Psychiatric Association, 2013). To ensure standardized diagnostic decision-making, the Structured Clinical Interview for DSM-5 (SCID; Osório et al., 2019) and the Anxiety Disorders Interview Schedule (ADIS) (Brown et al., 2001) are often used to establish SAD diagnoses and assess comorbidity. Besides interviews, clinicians and researchers use self-report or clinician-rated questionnaires to quantify symptom severity and monitor change over time, including the Liebowitz Social Anxiety Scale (LSAS; Heimberg et al., 1999), the Social Interaction Anxiety Scale (SIAS; Mattick & Clarke, 1998), and the Social Phobia Inventory (SPIN; Connor et al., 2000).

Although clinical interviews and questionnaires remain foundational to assessment, they are associated with important practical and methodological limitations, such as access to trained clinicians and the time demands of administering and interpreting assessments (World Health Organization, 2025). This may constrain their scalability in routine care settings. In addition, both interviews and questionnaires are limited by measurement challenges associated with retrospective self-report, including recall bias and social desirability (Van den Bergh & Walentynowicz, 2016; Tourangeau & Yan, 2007).

Artificial Intelligence (AI) may offer a complementary approach to the limitations of interviews and questionnaires. AI refers to machine-based systems that learn from inputs they receive on how to generate output such as predictions, recommendations, content, or decisions that can influence physical or virtual environments (Organisation for Economic Co-Operation and Development, 2024). AI is the umbrella term for systems that perform tasks that normally require human intelligence, and machine learning is a subset of AI systems that specifically learn patterns from data and subsequently make accurate inferences about new data (Sarker, 2021). In mental health and specifically for SAD, AI has the potential to support earlier identification by detecting patterns in heterogeneous data such as questionnaires, speech or text, and behavioral digital traces such as smartphone interaction patterns, messaging metadata, or social-media activity. Unlike traditional assessment, which often relies primarily on clinician-administered interviews and retrospective symptom reports, AI approaches can additionally draw on repeated or passively collected data collected outside the clinic, such as smartphone interaction patterns, messaging data or social media activity. AI approaches may therefore complement traditional assessment by flagging individuals who warrant clinical follow-up. If properly validated, AI tools may improve the reach and timeliness of interventions in settings strained by limited specialist capacity. However, they should be treated as decision support rather than a replacement for diagnosis and should

be governed carefully to manage risks such as privacy harms, bias, and lack of reliability in real-world deployment (National Institute of Standards and Technology, 2023).

AI terminology and its complexity may be difficult to navigate for psychologists and other readers without a technical AI background. Therefore, Table 1 (at the end of this paragraph) provides a simplified and structured overview of AI concepts and approaches relevant to this review. Rather than offering a fully exhaustive technical classification, the table is intended as a general orientation tool to facilitate understanding of approaches discussed in the following sections and to support interpretation of the results.

AI methods relevant to this review can be clustered in two broad categories: classical machine-learning methods and deep-learning methods. *Classical machine-learning methods* typically learn from predefined or engineered features, rather than learning complex internal representations directly from raw input (Faouzi & Colliot, 2023). Within this broad family, linear models learn weighted relationships between input variables and the outcome, nearest-neighbor methods classify based on similarity to nearby cases, support vector machines aim to separate groups by finding the boundary that best distinguishes cases from non-cases, tree-based methods classify cases by splitting the data using if-then decision rules, and probabilistic methods estimate the likelihood that a case belongs to a particular class (Faouzi & Colliot, 2023; Khan et al., 2024; Mahajan et al., 2024). Hybrid neuro-fuzzy methods combine neural-network learning from data with fuzzy if-then rules, allowing the model to learn from examples while handling categories more gradually than hard cutoff-based approaches (Babuška & Verbruggen, 2003; Shihabudheen & Pillai, 2018).

Deep learning is a subset of machine learning based on neural networks with multiple processing layers that can learn more complex patterns directly from data (LeCun et al., 2015). Neural network learning, in plain language, means that the model looks at examples, makes a guess, compares the guess to the correct answer, and adjusts itself using backpropagation (LeCun et al., 2015). Backpropagation is the procedure by which the model works backward through the network to calculate how its weights should be changed in order to reduce prediction error (LeCun et al., 2015). Different deep-learning methods are suited to different input types: convolutional neural networks are used for image-like data and time-series signals, transformer-based models are widely used for text and language tasks, and graph neural networks are designed for graph-structured data in which relationships between connected nodes are central (LeCun et al., 2015; Raffel et al., 2020; Zhou et al., 2020). Bayesian neural networks additionally estimate predictive uncertainty rather than producing only a single point prediction (Goan & Fookes, 2020). This dichotomy between classical machine learning and deep learning represents only one of

several possible ways to categorize AI methods. In general, classical machine learning approaches are often well suited for structured datasets with relatively limited feature spaces, whereas deep learning methods tend to excel in analysing larger, more complex, and unstructured data sources, albeit often with increased computational demands and reduced transparency (Rajkomar et al., 2019). A further consideration is that the distinction does not exhaust the ways AI models can be organized. Ensemble methods can for example combine multiple models which can rely on classical machine learning as well as deep learning to improve predictive performance beyond what any single model achieves alone. Irrespective of which models are being used, interpretability remains a significant challenge. This has led to growing interest in explainable artificial intelligence (XAI), which encompasses methods designed to make model behaviour and outputs more understandable to human users, thereby enhancing comprehension, transparency, and trust (Ali et al., 2023).

Table 1

Overview of Key AI Terms Used in This Review

Artificial Intelligence (AI)	machine-based systems that learn from inputs they receive on how to generate outputs such as predictions, recommendations, content, or decisions
Classical machine learning	refers to traditional machine-learning methods that typically learn from predefined or engineered features, often in structured data, rather than learning multilayer representations through deep neural networks
Nearest neighbor methods	This model classifies a new data point by looking at the most similar data points in the dataset (its 'neighbor') and copying their label. Example: if a new participant's symptom profile looks most similar to earlier participants with SAD, the model predicts SAD.
Linear model	This model learns how much each input variable matters and combines them into a straight-line (linear) relationship to make predictions. Example: fear of embarrassment might receive a strong positive weight, while another variable receives a smaller weight, and together they determine the predicted risk of SAD.
Support vector machines	This model uses input features to establish a decision boundary that best distinguishes between groups by maximizing the margin between them. Example: if questionnaire responses show that people with SAD usually score higher on fear and avoidance, the model can use these features to establish a decision boundary that helps discriminate between individuals with and without SAD.
Probabilistic/distribution-based methods	This model calculates the probability that a data point belongs to each class, based on assumptions about how data is distributed.

	Example: instead of only saying “SAD” or “no SAD,” the model may estimate that a participant has, for example, a 75% probability of belonging to the SAD group.
Tree-based methods	This model splits the data into branches using if-then rules Example: if avoidance is high, go left; if fear of embarrassment is also high, go left again; then predict SAD.
Hybrid neuro-fuzzy methods	This model combines neural-network learning with fuzzy inference / rule-based reasoning to model complex nonlinear relationships. Example: Suppose a person reports very high fear of embarrassment, moderate avoidance, and some physical anxiety symptoms. Instead of relying only on a hard cutoff, the model treats these as gradual categories, such as “high fear” and “moderate avoidance,” applies fuzzy if-then rules to combine them, and learns from data how strongly that overall pattern predicts SAD.
Deep learning	a type of machine learning that is especially useful for complex data such as images, text, audio, or time-series data, because it learns patterns in multiple layers, from simpler to more complex.
Graph neural networks model	This model learns from relationships between connected nodes in graph-structured data. Example: in a social network, each person is represented as a node, and the model can use not only that person’s own behavior, but also the behavior of connected nodes and the pattern of connections between nodes in the network.
Convolutional Neural Networks model	This model learns local patterns from image-like structured data, such as edges, shapes, or signal features. Example: a CNN could learn patterns from eye-tracking scanpath images or from a time-series signal recorded during a VR task.
Bayesian Neural Networks model	This model learns patterns from data while also estimating how uncertain it is about its predictions. Example: instead of only predicting “SAD,” the model may indicate that the prediction is uncertain because of limited relevant data or excessive noise in the data.
Transformer-based models	This model learns patterns by analyzing relationships between all elements in the data and capturing overall context, often used for natural language processing tasks. Example: a transformer can analyze an interview transcript and detect language patterns associated with social anxiety.

Because AI performance depends strongly on how models are evaluated, a brief clarification of common validation and performance terms is useful. Given that many studies in this review work with relatively small or heterogeneous datasets, these approaches appear frequently in the results section and are therefore worth defining here. *Cross-validation* is an internal resampling approach in which model performance is estimated by repeatedly training and testing the model across different subsets of the same dataset (James et al., 2021). By contrast, a *hold-out split*, in other words a single train-test split, evaluates model performance after one division of the data into training and test sets, making the estimate more dependent on that particular split (James et al., 2021). In *k-fold cross-validation*, the data are divided into k roughly equal parts (‘folds’) and the model is trained on k-1 folds and tested on the

remaining fold, and this process is repeated until each fold has served as the test fold once, after which performance is typically averaged across runs (James et al., 2021). *Leave-one-out cross-validation* (LOOCV) is a special case of k-fold cross-validation in which each fold contains only one observation (James et al., 2021). Related subject-wise variants, such as leave-one-subject-out evaluation, instead hold out all data from one participant at a time (Koul et al., 2018). In some studies, cross-validation is further stratified so that folds retain approximately the same class proportions as the full dataset, and repeated so that the procedure is run multiple times with different random splits, yielding a more stable internal estimate of performance (Thölke et al., 2023). Different metrics report different aspects of performance. *Accuracy* reflects the overall proportion of correct predictions, whereas *sensitivity and specificity* indicate the proportion of respectively actual cases and actual non-cases that are correctly identified, typically reported either as values between 0 and 1 or as percentages (Hicks et al., 2022). *Precision* reflects how many predicted positive cases are truly positive, whereas *recall* reflects the proportion of actual positive cases that are correctly identified. *F1-score* combines precision and recall into a single score via their harmonic mean, a form of averaging that is more sensitive to large differences between the two values, whereas area under the curve (AUC) indicates how well a model distinguishes between cases and non-cases across all possible decision thresholds (Hicks et al., 2022; Thölke et al., 2023; Çorbacioğlu & Aksel, 2023).

To understand the potential of AI for identifying SAD, it is necessary to examine which types of input data and AI approaches are being used, how accurately and reliably AI models can identify SAD, and in which populations and contexts these approaches have been studied. Therefore, this scoping review aimed to systematically map and synthesize the peer-reviewed empirical literature on AI approaches used to identify SAD or clinically significant social anxiety symptoms, operationalized through binary diagnostic or screening outcomes (SAD vs no SAD or for subclinical social-anxiety symptoms: screen-positive vs screen-negative based on validated cut-offs). Specifically, the review asked which types of AI are used to detect or classify social anxiety on the basis of psychological and behavioral data, which data types are used as input for these models and how well these AI models perform. In addition, the review examines how these studies were designed, in which populations and contexts these approaches were investigated, and which ethical and practical considerations emerge from the literature, and which gaps remain for future research.

Methods

This scoping review was preregistered on the Open Science Framework (OSF) using the Generalized Systematic Review Registration Form (<https://osf.io/by95s/overview>). The OSF registration was updated on 15 December 2025 to refine two eligibility criteria. Both criteria were implemented at the full-text screening stage, which was conducted independently by two reviewers. The review follows Joanna Briggs Institute methodology for scoping reviews and is reported in accordance with PRISMA-ScR (Hadie, 2024; Tricco et al., 2018).

Review questions and eligibility criteria were defined using the Population-Concept-Context framework recommended for scoping reviews (Hadie, 2024). The population was individuals of all ages, including participants with and without SAD or subclinical social-anxiety symptoms. The concept was AI-based detection or classification using binary diagnostic or screening outcomes. The context was any setting (e.g., clinical, community, laboratory, online) in any country. The primary review questions are: ‘Which types of AI based on psychological and behavioral data are used to detect or classify social anxiety?’ ‘Which data and psychological/behavioral indicators are used to identify social anxiety?’ and ‘How well do the models perform (e.g., accuracy, sensitivity)?’ The secondary questions are: ‘How are studies designed, who are the participants, and what data sources are used?’ ‘What ethical risks and real-world constraints are reported, and how are they handled?’ and ‘What gaps remain in the existing literature, and what areas should future research address?’

A structured search was conducted on 23 September 2025, across APA PsycINFO (EBSCO), Embase, Scopus, PubMed/MEDLINE, and Web of Science Core Collection. Google Scholar was used for supplementary searching, but no grey literature was included. The master Boolean strategy combined three concept blocks: (1) social anxiety disorder terms, (2) artificial intelligence terms, and (3) assessment/detection/classification terms. A generic master Boolean string was: (“social anxiety disorder” OR “social phob*” OR “social anx*” OR “social fear”) AND (“artificial intelligence” OR “machine learning” OR “deep learning” OR “natural language processing” OR “speech analysis” OR “computer vision”) AND (assess* OR detect* OR diagnos* OR predict* OR identif* OR classif*). The string was translated for each database to match its syntax. Full search strings for each source are provided in Appendix A.

All records identified through the database searches were exported from the databases and imported into Rayyan for deduplication and screening (Ouzzani et al., 2016). Screening was conducted in

two stages: title and abstract screening, followed by full-text screening. In each stage, articles were screened independently by two researchers using Rayyan with decision blinding. Bibliographic fields were not hidden. Disagreements were resolved through discussion, and if consensus could not be reached, a third reviewer adjudicated. Reasons for full-text exclusions are reported in the PRISMA flow diagram (Figure 1).

Included studies were peer-reviewed, published in English, and applied AI methods to detect or classify social anxiety. Following the OSF protocol update (15 December 2025), the outcome criterion was refined to require binary diagnostic or screening outcomes (e.g., SAD vs no SAD, or for subclinical social anxiety symptoms: screen-positive vs screen-negative based on validated cut-offs), ensuring that all included studies address the same core detection/classification task. Multi-class studies were included only when results for the social anxiety class were reported separately. Studies were excluded if they (1) did not include an AI component, (2) focused on general anxiety or other disorders without reporting separate results for social anxiety, (3) used neuroimaging/brain-scan data (e.g., fMRI, EEG) as model input or (4) were non-empirical publications (review, opinion articles, editorials or case reports without original empirical data). Following the OSF protocol update (15 December 2025), one exclusion criterion was expanded to also exclude studies using only physiological/biomedical measures or only socio-demographic/socio-economic variables as model inputs, consistent with the focus of this review on psychological and behavioral AI approaches.

Data were charted using a standardized extraction form, based on JBI guidelines. The charting form was piloted on five studies and refined. Extracted items include: authors, year of publication, country of dataset, study characteristics, sample characteristics, SAD definition/measure, data source/modality, funding, AI approach, validation, and ethical/real-world considerations. Data extraction was performed by the first author, followed by a brief verification check by a second reviewer to ensure completeness and accuracy. Findings were summarized using descriptive tables and narrative synthesis.

Results

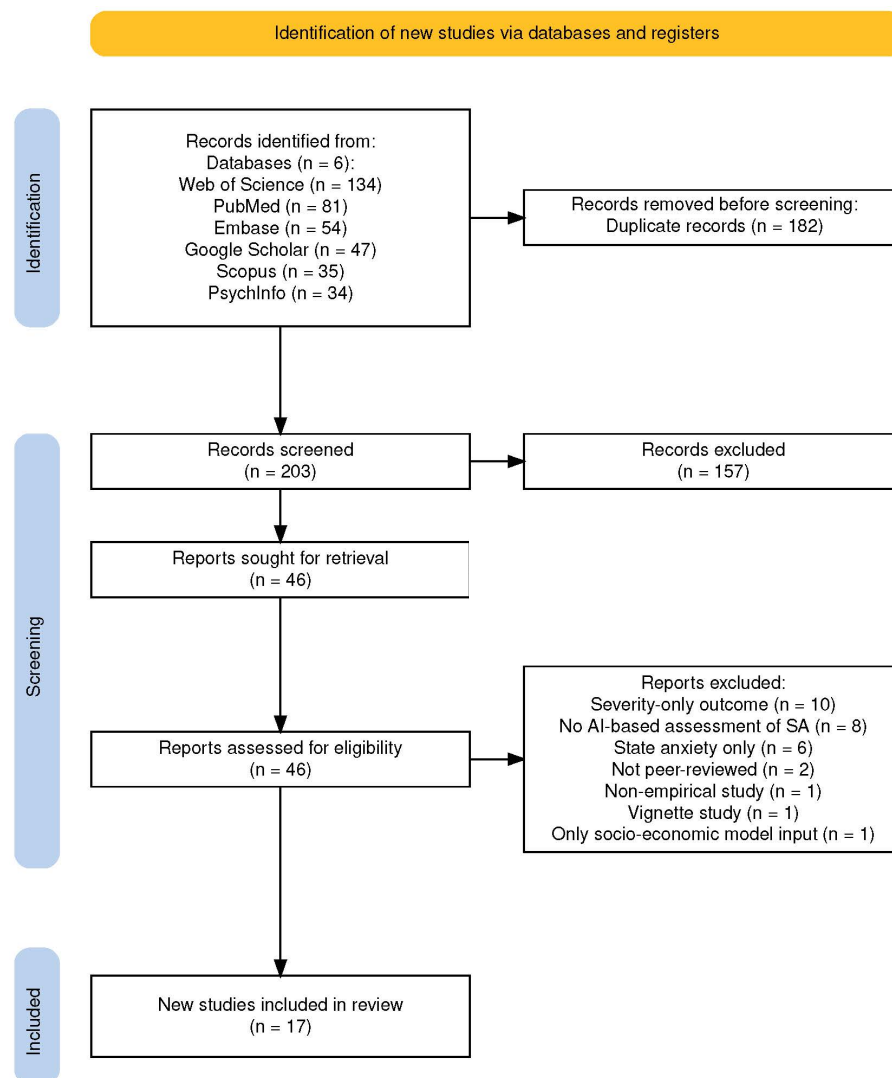
Search Results

In total, 385 records were imported in Rayyan. Of the 385 records, 182 (47%) were removed as duplicates, leaving 203 unique records for screening. Two reviewers independently screened the 203 titles and abstracts. Of the 203 titles and abstracts, there was agreement on 171 records (84%) and there

were 32 conflicts (16%) resolved through assessment by a third reviewer and discussion. A consensus was reached on all records and 46 records were included. The 46 included records (23% of screened records) underwent full-text screening, conducted by two reviewers, to determine final eligibility. Conflicts were again resolved through assessment by a third reviewer and discussion. 17 articles were included (37% of full texts screened), and 29 articles (63% of full texts screened) were excluded. Figure 1 presents the PRISMA 2020 flow diagram, summarizing the study selection process and reporting the reasons for exclusion at the full-text screening stage (Page et al., 2021; Haddaway et al., 2022).

Figure 1

PRISMA Flow Diagram of the Study Selection Process



Note. Generated using the PRISMA Shiny app (Haddaway et al., 2022)

Study Characteristics

All retrieved studies were published between 2020 and 2025, despite no restrictions on the timeframe of the search, with most appearing in 2024 ($n = 6/17$, 35%) and 2025 ($n = 4/17$, 24%). Datasets originated from multiple regions (Table 2).

Sample sizes varied widely. The smallest datasets were from experimental and lab-based studies ($48 \leq N \leq 63$; Chlasta et al., 2024; Tlachac & Ogden, 2024; Ohse et al., 2024; Pardini et al., 2025). Medium-sized studies included those relying on the aforementioned questionnaire dataset ($N = 214$; $N = 151$ after cleaning; Fathi et al., 2020; Cesa et al., 2022; Bathula et al., 2024; Chadaga et al., 2024). Other medium-sized samples relied on clinical or community-based datasets ($104 < N < 236$; Ahmadi Majd et al., 2025; Xiong et al., 2021; Demetriou et al., 2020), while Chang and Tseng (2020) reported a Facebook-based sample of >200 users. Larger datasets included a Massive Multiplayer Online Game (MMOG) user sample ($N = 819$; Yokotani et al., 2025) and a multimodal dataset of 1000 microblog posts from a social media platform (Fan et al., 2025). The largest evidence base came from a large-scale epidemiological cohort using survey-based baseline risk factors, including sociodemographic, prior mental and physical health, personality traits, stressful life events, childhood adversity/trauma and family history ($N = 32902$; Rosellini et al., 2020). One study did not report its participant sample size (Mezzi et al., 2022).

Studies recruited participants from different populations, including web-recruited self-report questionnaire datasets (and secondary analyses of the same dataset) (Fathi et al., 2020; Cesa et al., 2022; Bathula et al., 2024; Chadaga et al., 2024), university and young adult samples (Li et al., 2024; Tlachac & Ogden, 2024; Chlasta et al., 2024), and community volunteers recruited outside clinical services (e.g., via online ads with scale-based pre-screening) (Pardini et al., 2025). Other studies drew on children and adolescents assessed via parent-report with a clinician-interview reference standard (Xiong et al., 2021), clinic-based or clinic-referred samples (Demetriou et al., 2020; Mezzi et al., 2022), interview/transcript-based adult research samples (Ohse et al., 2024), and population-based survey cohort data (Rosellini et al., 2020). Several studies ($n = 4/17$, 24%) also sampled from a digital context, including a web-application user sample (Ahmadi Majd et al., 2025) and digital-platform users such as social network users (Chang & Tseng, 2020), microblog users (Fan et al., 2025), and multiplayer online game players (Yokotani et al., 2025). A notable pattern is the repeated use of a questionnaire dataset originally reported by Fathi and Ahmadi (2020) across four included studies ($n = 4/17$, 24%) (Fathi et al., 2020; Bathula et al., 2024; Chadaga et al., 2024; Cesa et al., 2022). The reuse was not confined to a single author team, although Bathula et al. (2024) and Chadaga et al. (2024) showed partial overlap in

authorship. As a result, a substantial part of the current evidence base is concentrated around one single data source, increasing the risk of overestimating generalizability, and means that several ‘independent’ model results are not independent in terms of underlying participants or data input. This dataset (Fathi & Ahmadi, 2020) also does not explicitly report the participants’ country of recruitment. The corresponding author was contacted for clarification, but no response was received. The same uncertainty about the country of the dataset applied to Chang and Tseng (2020), who collected data via the Facebook Application Programming Interfaces (APIs). The corresponding author was contacted to verify the recruitment country, but no response was received.

Social Anxiety Definition and Measurement

Most studies ($n = 9/17$, 53%) operationalized SAD as screen-positive vs screen-negative using self-report scales with thresholds, so model performance primarily reflects agreement with an instrument cut-off rather than an independently established diagnosis (Chlasta et al., 2024; Ohse et al., 2024; Pardini et al., 2025; Tlachac & Ogden, 2024; Yokotani et al., 2025; Fathi et al., 2020; Bathula et al., 2024; Cesa et al., 2022; Chadaga et al., 2024). One study ($n = 1/17$, 6%) used questionnaires to inform clinician labeling, but did not report instrument names or cut-offs (Chang & Tseng, 2020). A smaller subset ($n = 4/17$, 24%) relied on diagnostic interviews, where the outcome most closely reflected diagnostic status, including the Anxiety Disorders Interview Schedule for DSM-IV, the Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV, the Mini International Neuropsychiatric Interview and a clinician interview with the exact instrument not clearly reported (Demetriou et al., 2020; Rosellini et al., 2020; Xiong et al., 2021; Mezzi et al., 2022). One study ($n = 1/17$, 6%) used microblog posts labeled against DSM-5 criteria (Fan et al., 2025). Another study ($n = 1/17$, 6%) used a single self-rated social anxiety item scored from 0 (“not having it”) to 3 (“definitely having it”), which was subsequently dichotomized into non-case (0-1) versus social anxiety (2-3) (Li et al., 2024). Finally, one study ($n = 1/17$, 6%) used an extreme-groups definition in which participants were classified as SAD only if they self-reported a prior clinician diagnosis and endorsed high or very high anxiety, and classified as controls if they reported no diagnosis and endorsed low or very low anxiety, with moderate levels excluded and SPIN scores used for post hoc validation (Ahmadi Majd et al., 2025). Overall, this heterogeneity limits the direct comparability of reported model performance across papers because algorithms are learning different label constructions and reference standards (Table 2).

Table 2*General Characteristics of the Included Studies (n = 17).*

Study	Country of dataset	Sample (N)	SAD definition / measure
Ahmadi Majd et al. (2025)	Iran	104	self-reported clinician diagnosis + SPIN for validation
Bathula et al. (2024)	Iran*	214	Fathi dataset label; original classification based on SPIN ≥ 21 , with optional LSAS ≥ 35 for validation
Cesa et al. (2022)	Iran*	214	Fathi dataset label; original classification based on SPIN ≥ 21 , with optional LSAS ≥ 35 for validation
Chadaga et al. (2024)	Iran*	214	Fathi dataset label; original classification based on SPIN ≥ 21 , with optional LSAS ≥ 35 for validation
Chang & Tseng (2020)	Taiwan*	>200	Clinician labeling after questionnaire (questionnaire name + cut-off not reported)
Chlasta et al. (2024)	Poland	48	LSAS cut-offs (exact cut-off not stated)
Demetriou et al. (2020)	Australia	236	ADIS/SCID
Fan et al. (2025)	China	1000 posts	DSM-5-aligned researcher labels
Fathi et al. (2020)	Iran*	151	SPIN ≥ 21 ; optional LSAS ≥ 35 for validation/triangulation
Li et al. (2024)	China	2134	Self-rated social anxiety with binary threshold
Mezzi et al. (2022)	Tunisia	NR	MINI module / clinical diagnosis
Ohse et al. (2024)	Germany	51	SPIN cut-off (>25)
Pardini et al. (2025)	Italy	63	LSAS cut-off (≥ 55)
Rosellini et al. (2020)	USA	32,902 [†]	AUDADIS-IV (DSM-IV) incident onset
Tlachac & Ogden (2024)	USA	50 [‡]	Cut-off SIAS (≥ 36) + DASS-D cutoff (≥ 5)
Xiong et al. (2021)	Australia	132	Clinician interview (not further specified)
Yokotani et al. (2025)	Japan	819 users: 144,147 communication logs and 613,618 social edges	LSAS-J cut-off (≥ 60)

Note. ADIS = Anxiety Disorders Interview Schedule; AUDADIS-IV = Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV; DASS-D = Depression Anxiety Stress Scales–Depression subscale; LSAS = Liebowitz Social Anxiety Scale; LSAS-J = Japanese version of the Liebowitz Social Anxiety Scale; MINI = Mini International Neuropsychiatric Interview; SAD = social anxiety disorder; SCID = Structured Clinical Interview for DSM Disorders.

* Country not explicitly reported; listed as most plausible based on affiliation/platform context.

[†] Eligible sample for incident-onset analysis.

[‡] Participants with complete SIAS/DASS-D were used in the final analysis.

Input Data Modalities

Input data modalities are clustered into six broad types (Table 3). First, questionnaire-derived features were the most common modality ($n = 7/17$, 41%), including repeated secondary analyses of the same Fathi dataset (Bathula et al., 2024; Cesa et al., 2022; Chadaga et al., 2024; Fathi et al., 2020), other self-report item batteries (Li et al., 2024), parent-completed diagnostic questionnaires (Xiong et al., 2021) and large-scale epidemiological survey predictors used for longitudinal risk modeling, including sociodemographic characteristics, personality traits, pre-existing mental disorders/symptoms, physical health problems, stressful life events, childhood adversity/trauma and family history (Rosellini et al., 2020). Second, two studies ($n = 2/17$, 12%) used scenario-elicited self-report inputs. Ahmadi Majd et al. (2025) used a web app with 10 multimedia social scenarios, after which participants rated the extent to which suppression, avoidance and reappraisal matched their likely response on a Likert scale. Pardini et al. (2025) used continuous anxiety ratings recorded in real time with a joystick during a virtual reality waiting-room scenario that started empty and progressively filled with virtual avatars. Third, one study ($n = 1/17$, 6%) used a mixed clinical assessment battery combining objective neuropsychological/social-cognition task performance with self-report scales (Demetriou et al., 2020). Fourth, a smaller subset relied on behavioral and digital trace data ($n = 4/17$, 24%), including Facebook-derived indicators such as profile characteristics, self-disclosure behaviors (e.g., describing a person's feelings), language use (i.e., features extracted from written post content) and temporal activity patterns (e.g., log-in/log-out timing and bursts of activity) (Chang & Tseng, 2020), microblog posts containing both written text and accompanying images (Fan et al., 2025), SMS metadata logs (timestamps and hashed sender/recipient identifiers) (Tlachac & Ogden, 2024) and online game communication logs plus social-network structure (Yokotani et al., 2025). Fifth, one study ($n = 1/17$, 6%) used eye-tracking scanpaths, recorded while participants viewed slides with emotional facial expressions, as input (Chlasta et al., 2024). Finally, two studies ($n = 2/17$, 12%) used clinical language modalities, deriving inputs from interview speech or transcripts (Mezzi et al., 2022; Ohse et al., 2024). Overall, self-report-based inputs broadly defined, including questionnaire-based and scenario-elicited self-report data, accounted for the largest share of studies, while passive digital traces, a mixed clinical assessment battery, eye-tracking data and clinical language data remain less common.

AI Methods

Consistent with the simplified overview of AI method families presented in the Introduction (Table 1), the included studies can be grouped into classical machine learning and deep learning. Most studies ($n = 10/17$, 59%) used classical machine learning approaches, typically applied to structured input features such as questionnaire data or other self-report data (Fathi et al., 2020; Cesa et al., 2022; Bathula et al., 2024; Chadaga et al., 2024; Li et al., 2024; Ahmadi Majd et al., 2025; Tlachac & Ogden, 2024; Rosellini et al., 2020; Demetriou et al., 2020; Chang & Tseng, 2020). These studies usually compared several standard classifiers, such as logistic regression, support vector machines, k-nearest neighbors, random forests or ensemble approaches and then reported the best-performing model under internal validation. Within this group, two studies ($n = 2/17$, 12%) additionally applied explainability methods, such as SHAP (Shapley additive explanations) and LIME (local interpretable model-agnostic explanations), to make model outputs easier to interpret (Bathula et al., 2024; Chadaga et al., 2024).

The remaining studies ($n = 7/17$; 41%) used deep learning approaches, mainly in settings involving more complex data types. These included convolutional neural networks for eye-tracking scan paths images and virtual-reality anxiety trajectories (Chlasta et al., 2024; Pardini et al., 2025), transformer-based models and large language model approaches for interview text or transcripts (Mezzi et al., 2022; Ohse et al., 2024), multimodal models combining text and images from microblogs (Fan et al., 2025), a graph neural network for social-network structure derived from game communication data (Yokotani et al., 2025) and a Bayesian neural network applied to questionnaire-based screening data (Xiong et al., 2021). A more detailed overview of the individual models used in each study is provided in Appendix C.

Validation and Performance Reporting

As shown in Table 3 and using the terms introduced in the Introduction, reported model evaluation approaches were heterogeneous and mostly relied on internal validation. For studies that trained predictive models, four main validation strategies could be distinguished: single hold-out splits, k-fold cross-validation, combined hold-out and repeated stratified cross-validation, and leave-one-out style evaluations. Single hold-out splits were used in six studies ($n = 6/17$, 35%) (Bathula et al., 2024; Chadaga et al., 2024; Chlasta et al., 2024; Fan et al., 2025; Mezzi et al., 2022; Yokotani et al., 2025). K-fold cross-validation was used in seven studies ($n = 7/17$, 41%) (Cesa et al., 2022; Chang & Tseng, 2020;

Demetriou et al., 2020; Fathi et al., 2020; Li et al., 2024; Rosellini et al., 2020; Xiong et al., 2021). Notably, Rosellini et al. (2020) uniquely adopted a prospective cohort prediction framing, using Wave-1 predictors to predict incident (new-onset) social phobia by Wave 2. However, because model performance was still estimated via internal validation within the same cohort, it is grouped under k-fold cross-validation, rather than treated as a separate validation strategy. One study ($n = 1/17$, 6%) combined a hold-out split with repeated stratified cross-validation (Ahmadi Majd et al., 2025). Two studies ($n = 2/17$, 12%) used leave-one-out style evaluations: Pardini et al. (2025) applied leave-one-subject-out cross-validation to participant-level VR time-series data, whereas Tlachac and Ogden (2024) used leave-one-out-cross-validation on participant-level feature vectors derived from text-message metadata. One study ($n = 1/17$, 6%) differed from the other studies because it evaluated a pre-existing language-model scoring approach, rather than training and validating a study-specific predictive model. In this study, GPT-4, a Generative Pre-trained Transformer language model was prompted to infer SPIN scores directly from interviews without fine-tuning, and the outputs were compared with self-reported SPIN scores (Ohse et al., 2024). This approach is therefore better understood as a direct evaluation of a pre-existing model's scoring ability, rather than as a conventional validation strategy based on training and testing a newly developed model.

Performance reporting used a heterogeneous set of metrics, most commonly accuracy, sensitivity, specificity, F1-score, and area under the curve (AUC), sometimes alongside a confusion matrix (Table 3). As outlined in the introduction, these metrics capture different aspects of performance and should not be interpreted as interchangeable (Hicks et al., 2022). Importantly, 'model performance' in this context primarily reflects how well an algorithm reproduces the study's chosen outcome definition (e.g., a questionnaire cut-off or a clinician interview label) under the given validation design. Across the studies, best-reported performances span a wide range, from modest multi-class accuracies in small exploratory perceptual datasets (e.g., ~.64 accuracy for 3-class classification including an anxiety class in eye tracking) (Chlasta et al., 2024) to some studies using structured questionnaire-based data or language-based scoring approaches reporting accuracies and AUC values in the mid-.80s to upper .90s (e.g., Bathula et al., 2024; Cesa et al., 2022; Chadaga et al., 2024; Fathi et al., 2020; Xiong et al., 2021). However, these best-reported metrics are not directly comparable across studies and should be treated primarily as evidence of feasibility, rather than as a reliable ranking of methods (Tables 2-3). This is because studies differed in how social anxiety was defined, which input modalities were used, and how performance was validated and reported. Therefore, it is not possible to directly compare the models or input modalities across the included studies, nor determine which approach is the most suitable for

detecting SAD based on the current evidence. The subgroup of studies based on the Fathi and Ahmadi dataset (2020) may seem to provide a more related basis for comparison. However, the studies differed in feature selection, treatment of SPIN/LSAS variables, anomaly or outlier handling, validation procedures and reporting detail, meaning that differences in performance cannot be attributed to AI models alone (Fathi et al., 2020; Cesa et al., 2022; Bathula et al., 2024; Chadaga et al., 2024).

Ethical and Practical Considerations

Ethical as well as practical considerations were reported in most studies ($n = 15/17$, 88%), but the level of detail and specificity varied substantially (Table 3). Several articles explicitly framed AI as a screening or decision-support aid intended to support earlier identification, rather than as a substitute for clinical diagnosis (Cesa et al., 2022; Xiong et al., 2021; Rosellini et al., 2020; Tlachac & Ogden, 2024). One highlighted ethical concern is the consequence of classification errors (Cesa et al., 2022). False positives may trigger unnecessary follow-up and anxiety, whereas false negatives risk missed detection and delayed care, implying that threshold choice and error trade-offs matter for real-world deployment. Another concern is the need for privacy-preserving data handling in continuous monitoring contexts, because passive screening may involve repeated or ongoing collection of potentially sensitive behavioral data (Tlachac & Ogden, 2024). The authors therefore emphasize approaches that avoid message content and instead rely on log metadata, noting that mobile devices prevent log collection without explicit user permission (Tlachac & Ogden, 2024). Broader translational limitations were also raised, including the need for external validation on independent samples before deployment (Cesa et al., 2022; Demetriou et al., 2020). External validation would involve testing a model on an independent dataset from new participants or settings, rather than only evaluating it within the dataset used to develop the model. Concerns about generalizability were also raised in relation to homogeneous samples, cultural differences, and platform-specific data sources (Fan et al., 2025; Ohse et al., 2024). In this context, cultural or platform-specific bias refers to the possibility that patterns learned from one cultural, linguistic or online-platform context may not transfer equally well to other populations or digital environments. Multiple studies proposed plausible deployment settings (e.g., schools, workplaces, hospitals, organizational monitoring, or college-orientation prevention), but these were often aspirational and not matched with detailed implementation plans or governance safeguards (Chadaga et al., 2024; Chlasta et al., 2024; Li et al., 2024; Pardini et al., 2025; Rosellini et al., 2020).

Table 3

Key Findings and Methodological Characteristics of the Included Studies (n = 17).

Authors (Year)	Data source / modality	AI/ML approach	Performance (best reported model)	Ethical/real-world considerations
Ahmadi Majd et al. (2025)	Scenario-elicited self-report ratings across multimedia social scenarios	Classical ML (linear, tree-based, nearest-neighbor, and support vector methods)	K-Nearest Neighbors: Acc .84; Prec .80; Rec .81; F1 .79; AUC .88. Validation: 80/20 train-test split; repeated 5-fold stratified cross-validation.	Notes online self-report verification limits because of extreme groups design; discusses future richer behavioral cues (e.g., robots; gaze/posture/prosody).
Bathula et al. (2024)	Questionnaire (data originally reported by Fathi and Ahmadi (2020)): demographic, emotional, physical symptom variables (LSAS-score included as a variable; SPIN-score excluded as a feature)	Classical ML (tree-based, linear, nearest-neighbor and ensemble methods combining the former three method types) Attention for XAI	Random Forest: Acc .84; Prec .86; Rec .86; F1 .86; AUC .95. Validation: Internal hold-out 80/20.	ML can assist in better diagnosis and treatment.
Cesa et al. (2022)	Questionnaire (data originally reported by Fathi and Ahmadi (2020)): demographic, emotional, physical symptom variables (LSAS-score and SPIN-score included as variables)	Classical ML (tree-based, linear, support vector, and nearest-neighbor method types)	Support vector machine: Acc .96; Prec .97; Rec .95; F1 .96; AUC .97. Validation: K-fold cross-validation.	Supports (not replaces) diagnosis; stresses harms of false positives and false negatives; small validation with 5 psychologists; recommends external validation and more datasets.
Chadaga et al. (2024)	Questionnaire (data originally reported by Fathi and Ahmadi (2020)): demographic, emotional, physical symptom variables (LSAS-score included as a variable and SPIN-score excluded as a feature)	Classical ML (tree-based, linear, and nearest-neighbor and ensemble combining the former three method types) Attention for XAI	Logistic Regression: Acc .88; Prec .89; Rec .89; F1 .88; AUC .94. Adaptive Boosting: Acc .88; Prec .88; Rec .88; F1 .88; AUC .92. Validation: 80/20 train-test split.	Suggests deployment in schools, workplaces and hospitals.

Authors (Year)	Data source / modality	AI/ML approach	Performance (best reported model)	Ethical/real-world considerations
Chang & Tseng (2020)	Social network profile and activity traces such as number of interests and emoticon use, word-use patterns, signs of one-sided online relationships and bursts of online activity	Classical ML (support vector methods and ensemble methods combining one-class support vector methods)	One-class Support Vector Machine: Acc NR; Prec .66; Rec 1.00; F1 .79; AUC NR. Validation: 5-fold cross-validation	Frames online social network data as a complementary alternative to the conventional approaches in psychology
Chlasta et al. (2024)	Eye-tracking scanpath images derived from gaze recordings during the viewing of slides with emotional facial expressions	Deep learning (convolutional neural networks)	Best 3-class classification model (anxious/control/depressive): Acc .64; Prec NR; F1 NR; AUC NR. Validation: 80/20 train-test split.	Envisions affordable eye-trackers in VR for proactive monitoring/support in organizations
Demetriou et al. (2020)	Neuropsychological (objective) and self-report (subjective) measures of social cognition, cognitive and executive function, as well as self-report measures of affective states (depression, anxiety, and stress)	Classical ML (tree-based and linear methods)	Bayesian Additive Regression Trees (for the SAD vs neurodevelopmental condition comparison, best reported model performance): Acc .71; Prec .79; Rec .68; F1 NR; AUC .78. Validation: repeated 10-fold cross-validation.	Explicitly calls for external validation with an independent large test set of completely unseen data to assess the generalizability.
Fan et al. (2025)	Microblog combining written text and images	Deep learning (convolutional neural networks and transformer-based models)	Weighted fusion model: Acc .83; Prec NR; Rec NR; F1 .82; AUC .89. Validation: 70/15/15 train/validation/test split.	Flags cultural bias, ethical issues, and cross-platform generalizability as key for responsible deployment; the framework's resilience to noisy and incomplete data makes it suitable for real-world applications.
Fathi et al. (2020)	Questionnaire (data originally reported by Fathi and Ahmadi (2020): 11 attributes used: DSM-5/ICD-10 symptom items and demographics (gender, family history of anxiety or depressive disorder)	Classical ML (hybrid neuro-fuzzy methods)	Adaptive Neuro-Fuzzy Inference System, cleaned dataset: Acc .99; Prec NR; Rec .97; Spec 1.00; F1 NR; AUC NR. Validation: 5-fold cross-validation.	Notes that social class and education may affect how attributes are interpreted, potentially producing anomalous responses; recommends future decision-support use for severity assessment.

Authors (Year)	Data source / modality	AI/ML approach	Performance (best reported model)	Ethical/real-world considerations
Li et al. (2024)	Questionnaire: 7 Likert-type items (0–3; e.g., fear of embarrassment/negative evaluation, social media–related anxiety, internet addiction, parenting/childhood experience) and 1 binary item (only-child status, 0/1).	Classical ML (linear, nearest-neighbor, probabilistic, tree-based, support vector and ensemble methods combining the former three method types)	Ensemble model: Acc NR; Prec NR; Rec NR; F1 .83; AUC .85. Validation: 10-fold cross-validation (10 runs).	Suggests possible school applicability due to low data needs; acknowledges data/sample limitations and missing multidimensional factors.
Mezzi et al. (2022)	Speech-to-text responses from a Mini International Neuropsychiatric Interview (MINI) questions delivered by a 3D avatar, including the social phobia module	Deep learning (transformer-based models)	Bidirectional Encoder Representations from Transformers: Acc .94; Prec .95; Rec .99; F1 .97; AUC .86. Validation: 80/10/10 train/validation/test split.	NR
Ohse et al. (2024)	Speech-to-text transcripts from LSAS-based semi-structured interviews on behavior, thoughts, emotions, fear, and avoidance	Deep learning (transformer-based models)	Zero-shot evaluation against self-reported SPIN cut-off (≥ 25). GPT-4: Acc NR; Prec .84; Rec .84; F1 .84; AUC .93.	Calls for validation against expert-rated ground truth; notes homogeneous sample and recruitment limits generalizability.
Pardini et al. (2025)	Scenario-elicited continuous joystick-rated anxiety during a virtual reality waiting-room scenario that gradually filled with virtual avatars	Deep learning (convolutional neural networks)	1D Convolutional Neural Network, anxious class: Acc .78; Prec .77; Rec .89; F1 .82; AUC NR. Validation: Leave-one-subject-out cross-validation (LOSO) across $n = 63$, with repeated stratified 80/20 train/validation splits within each LOSO iteration for tuning.	Next step: embed model in VR for real-time adaptation and personalized intervention (dynamic stimulus modulation).

Authors (Year)	Data source / modality	AI/ML approach	Performance (best reported model)	Ethical/real-world considerations
Rosellini et al. (2020)	Survey predictors, including sociodemographic, prior mental and physical health, personality traits, stressful life events, childhood adversity/trauma and family history	Classical ML (ensemble methods combining linear, probabilistic, nearest-neighbor, support-vector and tree-based method types)	Super learner ensemble: Acc NR; Prec NR; Rec NR; F1 NR; AUC .80. Validation: 10-fold cross-validation	Suggests online risk calculators and use in settings like college orientation for prevention and referral.
Tlachac & Ogden (2024)	SMS metadata such as when messages were sent, reply delays and number of sent/received messages, without message content	Classical ML (probabilistic, support vector methods and tree-based ensemble)	Random forest: Acc NR; Prec NR; Rec .69; Spec .86; F1 NR; AUC NR. Validation: Leave-one-out cross-validation.	Ethical issues: continuous data collection; recommend privacy protections (no content; hashed numbers); optional use; avoid diagnostic use; could be deployed pre-appointment or in patient portals.
Xiong et al. (2021)	Questionnaire (Youth Online Diagnostic Assessment completed by parents): 28 questions; binary yes/no and severity/frequency-scale answers.	Deep learning (Bayesian neural networks)	Feature-Ensemble Bayesian Neural Network: Acc NR; Prec NR; Rec .76; Spec .85; F1 NR; AUC .91. Validation: Internal 5-fold cross-validation	Positions model as a screening and triage tool; not a standalone diagnostic replacement.
Yokotani et al. (2025)	Online game communication logs and social-network features, such as virtual house visits, friend connections, avatar profile data and timing patterns of virtual house visits	Deep learning (graph neural networks)	GraphSAGE graph neural network: F1 .72. Derived from Figure 4B: Acc .63; Prec .66; Rec .79. AUC NR. Validation: Internal hold-out validation (80/10/10 train/validation/test split).	NR

Note. Abbreviations: Acc = accuracy; Prec = precision; Rec = recall; Spec = specificity; F1 = F1-score; AUC = area under the curve; ML = machine learning; XAI = explainable artificial intelligence; NR = not reported; SPIN = Social Phobia Inventory; LSAS = Liebowitz Social Anxiety Scale; ANFIS = Adaptive Neuro-Fuzzy Inference System; GPT = Generative Pre-trained Transformer; LOSO = leave-one-subject-out; GraphSAGE = Graph Sample and Aggregate. Performance metrics are reported as proportions and rounded to two decimal places where applicable. In Bathula et al. (2024), Random Forest performance metrics were reported inconsistently across the abstract, Table I, and the results text. The values extracted here were taken from the results text and should therefore be interpreted with caution.

Discussion

Main Findings

Traditional assessment of social anxiety disorder (SAD) relies heavily on clinician-administered interviews and retrospective self-report, which is time-intensive, vulnerable to recall and social-desirability biases, and can contribute to under-detection and long delays before treatment contact, especially in settings with limited specialist capacity (American Psychiatric Association, 2013; Tourangeau & Yan, 2007; Van den Bergh & Walentynowicz, 2016; Wang et al., 2005; Ertekin et al., 2015; World Health Organization, 2025; Chapdelaine et al., 2018; Olariu et al., 2015). This scoping review mapped empirical AI approaches used to identify SAD or subclinical social anxiety symptoms across heterogeneous populations, case definitions, input modalities, and validation strategies. Following a structured search and two-stage screening, 17 peer-reviewed empirical studies published between 2020 and 2025 were included in this scoping review. Overall, the evidence base suggests feasibility because all approaches can, to some degree, discriminate between ‘screen-positive’ and ‘screen-negative’ groups under study-specific conditions. However, the literature remains early-stage in terms of generalizability, comparability across studies, and readiness for responsible deployment in real screening or decision-support contexts.

Primary Research Questions: Results and Interpretation

RQ1. Which types of AI based on psychological and behavioral data are used to detect or classify social anxiety?

Across included studies, two broad types of AI approaches were used. Most papers applied traditional supervised machine learning to structured or engineered features, typically comparing several standard classifiers and selecting the best-performing model under internal validation (Fathi et al., 2020; Cesa et al., 2022; Bathula et al., 2024; Chadaga et al., 2024; Li et al., 2024; Ahmadi Majd et al., 2025; Rosellini et al., 2020; Tlachac & Ogden, 2024; Demetriou et al., 2020; Chang & Tseng, 2020). A smaller subset used deep learning methods often tailored to more complex inputs such as eye tracking scan paths, time-series trajectories in virtual reality, interview language, multimodal social media content (Chlasta et al., 2024; Pardini et al., 2025; Mezzi et al., 2022; Ohse et al., 2024; Fan et al., 2025; Yokotani et al., 2025) but sometimes also used for questionnaire-based screening data (Xiong et al., 2021). The predominance of approaches that operate on structured data suggests that the literature remains largely feasibility-driven: these approaches are common because they are practical on small datasets and straightforward to benchmark. The added value of AI methods should

be interpreted cautiously when both inputs and the social anxiety label are questionnaire-based. However, deep learning approaches are more plausibly value-adding because inputs are less directly scorable with simple rules.

RQ2. Which data and psychological/behavioral indicators are used to identify social anxiety?

Overall, self-report based inputs broadly defined, including questionnaire-based and scenario elicited self-report data, accounted for the largest share of studies, with many studies effectively learning a predictive mapping from symptom profiles (and related correlates) to a binary label defined by a questionnaire threshold or closely related criteria (Fathi et al., 2020; Cesa et al., 2022; Bathula et al., 2024; Chadaga et al., 2024; Li et al., 2024; Ahmadi Majd et al., 2025). A notable pattern was the reuse of a single small questionnaire dataset, originally reported by Fathi and Ahmadi (2020), across multiple later papers (Fathi et al., 2020; Bathula et al., 2024; Chadaga et al., 2024; Cesa et al., 2022), which concentrates evidence around one data source and limits generalizability. More modality-rich approaches, including digital traces, language (interview/transcripts), experimental and perceptual signals, and VR time-series, appear less frequently but represent the clearest departure from conventional screening, because they operate on inputs that are difficult to score transparently using traditional rule-based methods (Chang & Tseng, 2020; Mezzi et al., 2022; Yokotani et al., 2025; Fan et al., 2025). For that reason, these approaches are arguably the most promising for an added value of AI approaches alongside traditional assessment, because they can leverage behavioral or communicative signals that standard interviews and questionnaires do not capture directly.

Taken together, the predominance of questionnaire-based inputs and the more limited use of richer behavioral, language-based, and digital data modalities suggest that the added value of AI depends on the modality and clinical purpose. In several studies, SAD identification is framed as supervised prediction from structured, questionnaire-derived/tabular self-report features with outcomes defined via scale cut-offs or closely related rules. In these designs, ‘AI’ often functions as an alternative scoring layer over conventional screening inputs and may have limited incremental benefit. In other words, when the inputs are already symptom ratings and the ground truth is a questionnaire-defined label, high performance may reflect optimized reproduction of that questionnaire’s decision boundary rather than improved identification of clinically meaningful SAD in real practice. This is important because conventional SAD assessment often relies heavily on retrospective self-report, which can be shaped by recall, limited ecological validity, and the fact that it captures symptoms only at discrete time points rather than as they unfold in daily life. Nevertheless, questionnaire-based AI approaches may still add some value if they clearly

demonstrate added value over established scoring procedures, for example by identifying interpretable item patterns or combinations that would not be captured through standard manual scoring.

By contrast, the rationale for AI is stronger when unstructured, passive, high-dimensional, time-series, or network-based inputs are used, because these are often modalities that cannot be straightforwardly scored using traditional rules. These include online social network traces, SMS metadata, MMOG communication graphs, eye-tracking scan paths, VR anxiety trajectories, multimodal microblog content, and interview-derived language data (Chang & Tseng, 2020; Tlachac & Ogden, 2024; Yokotani et al., 2025; Chlasta et al., 2024; Pardini et al., 2025; Fan et al., 2025; Mezzi et al., 2022; Ohse et al., 2024). In such contexts, AI can plausibly extend assessment beyond traditional assessment based on interviews and questionnaires that rely on retrospective self-report and toward behavioral and communicative patterns, potentially enabling earlier flagging or monitoring, provided that governance, consent, and clinical safeguarding are in place.

RQ3. How well do the models perform (e.g., accuracy, sensitivity)?

Overall, reported model performance should be interpreted as promising but study-bound. Although studies reported moderate to strong within-study results, these estimates are difficult to compare across papers because studies used different outcome definitions or reference standards (e.g. a scale cut-off, a clinician interview outcome or an expert labeling) (Fathi et al., 2020; Demetriou et al., 2020; Fan et al., 2025) and model testing relied on internal evaluation within a single dataset. Reported model performance therefore often reflects how well a model predicts the study-specific operationalization of social anxiety disorder, rather than demonstrating that it can generalize across settings, populations or reference standards. As a result, the current evidence does not allow conclusions about which AI model or input modality combination performs best for identifying social anxiety across contexts.

In addition, the literature reports heterogeneous metrics, which capture different aspects of performance and should not be treated as interchangeable indicators of clinical utility (Hicks et al., 2022; Thölke et al., 2023). Therefore, evidence supports feasibility under study-specific conditions, while claims about reliable screening performance in practice remain premature without stronger evidence of generalization and decision-relevant interpretation of errors (Cesa et al., 2022; Tlachac & Ogden, 2024).

Methodological and Reporting Limitations and Evidence Gaps in the Literature

Despite promising feasibility findings, the current evidence base is constrained by recurring methodological and reporting limitations. Much of the current evidence remains proof-of-concept and relies on internal evaluation within single datasets. This means that models were generally trained and tested on different subsets of the same dataset, rather than being evaluated on an entirely independent sample from a new population or setting. This limits generalizability, comparability, and real-world readiness. Additionally, studies use different reference standards to label 'social anxiety', and as a result, comparisons between studies are fragile and should be interpreted with caution.

This scoping review also shows that the literature is skewed toward structured self-report input, where the incremental value of ML approaches over simpler scoring approaches is not always demonstrated (Fathi et al., 2020; Cesa et al., 2022; Bathula et al., 2024; Chadaga et al., 2024; Li et al., 2024; Ahmadi Majd et al., 2025; Xiong et al., 2021; Rosellini et al., 2020). By contrast, genuinely ecological or passive modalities remain relatively rare and are often based on small or setting-specific samples (Chang & Tseng, 2020; Tlachac & Ogden, 2024; Chlasta et al., 2024; Ohse et al., 2024; Pardini et al., 2025; Fan et al., 2025; Yokotani et al., 2025). At the same time, it was not always clear to what extent model input was fully independent from the outcome being predicted. For example, in the dataset reported by Fathi and Ahmadi (2020), the SAD label was derived from SPIN and LSAS screening results, while several model inputs consisted of closely related SAD symptom indicators, such as fear of public speaking, fear of being the center of attention and fear of meeting strangers. In later analyses of the same dataset, SPIN was sometimes removed because of its strong association with the target, while LSAS or closely related symptom indicators were still retained or considered as model inputs. This raises a concern that high model performance may partly reflect overlap between predictors and label construction, rather than fully independent identification of clinically diagnosed SAD.

A separate but related limitation is that the same Fathi and Ahmadi (2020) dataset was reused across several studies which further concentrates the evidence base around one underlying data source (Fathi et al., 2020; Cesa et al., 2022; Bathula et al., 2024; Chadaga et al., 2024). Greater transparency regarding label construction, predictor-outcome independence and reuse is therefore essential for evaluating added value, comparability and clinical relevance. More broadly the process by which models were developed and evaluated was not always reported in sufficient detail, particularly with regard to data preparation, feature selection and class-imbalance handling. This limited reporting transparency may contribute to overly optimistic performance estimates. In one

study, performance metrics were reported inconsistently across different sections of the same article, further complicating interpretation and comparison across studies (Bathula et al., 2024). In addition, reporting gaps remained around implementation-relevant issues, such as privacy and the way these tools could be realistically (operational rather than aspirational) integrated in the clinical workflows (Tlachac & Ogden, 2024; Chadaga et al., 2024).

A further concern relates to the contextual transparency of the evidence base and representativeness. Although psychological research has long been criticized for its overreliance on WEIRD (Western, Educated, Industrialized, Rich and Democratic) populations (Henrich et al., 2010; Thalmayer et al., 2021), several studies in this review drew on datasets from non-Western settings, including Iran, China, Japan and Tunisia. In that sense, the included literature partly diverges from a broader pattern in psychology in which Western perspectives and samples remain disproportionately influential (Henrich et al., 2010; Thalmayer et al., 2021). However, this should be interpreted cautiously. Geographical spread alone does not guarantee generalizable evidence. In the present review, recruitment context was sometimes incompletely reported. As a result, the country of origin of some datasets could not be verified with certainty.

Strengths and Limitations of the Scoping Review

This scoping review has several strengths. To our knowledge, at the time this review was initiated (September 2025), no scoping review around this specific subject had yet been conducted, but literature had started emerging, supporting the need for a focused mapping of this emerging literature. The review was pre-registered on OSF and followed established scoping-review guidance (JBI) with reporting aligned to PRISMA-ScR, supporting transparency and reproducibility of the search and charting process (Tricco et al., 2018; Hadie, 2024). A further strength is the explicit focus on detection/classification (binary outcomes) after protocol refinement, which improves conceptual coherence across included studies when synthesizing validation and performance reporting.

Several limitations should be considered when interpreting findings. First, consistent with scoping-review methodology, no formal critical appraisal/risk-of-bias assessment was conducted. Therefore, the review maps the landscape of methods and reporting but cannot determine the quality of evidence or rank approaches as ‘best’ for clinical use. Second, the search was limited to peer-reviewed studies in English, and no systematic grey-literature search was performed, which may have excluded relevant preprints, technical reports, or industry evaluations, potentially important in a fast-moving AI domain. Third, evidence synthesis is constrained because studies define ‘social anxiety’ differently (e.g., questionnaire cut-offs vs clinician interviews) and evaluate

models using different validation designs and metrics, which makes results difficult to compare and prevents meaningful quantitative pooling. Finally, despite efforts to document dataset provenance, some primary studies provided incomplete reporting about recruitment context or participant characteristics.

Clinical Implications and Recommendations for Future Research

The findings of this review have several clinical implications, but these should be interpreted cautiously given the early-stage nature of the evidence. As outlined in the Introduction, SAD is often under-recognized and traditional assessment depends on trained clinicians, semi-structured interviews and self-report measures, which can be time-intensive and difficult to scale in settings with limited specialist capacity or funding. In this context the future of AI may be to support, not replace, clinician-led diagnosis by flagging individuals who may warrant clinical follow-up. However, at present, the evidence base is not sufficiently mature to support direct implementation of AI-based SAD identification tools in routine clinical practice. If adequately validated, AI could contribute to more efficient and potentially more consistent assessment pathways by helping clinicians identify which individuals require more comprehensive evaluation. In this way, AI may help limited clinical resources to be directed more effectively toward assessment, follow-up and more time for treatment. This could be particularly relevant in primary care, schools or universities, where many individuals with social anxiety may first present and where access to specialized psychological assessment can be limited.

The potential added value of AI appears strongest when models use data that are difficult to interpret with traditional scoring methods, such as language, digital traces, eye tracking or multimodal behavioral data. AI could complement clinical interviews and questionnaires by identifying behavioral or communicative patterns that may not be immediately visible in standard assessment. This possible role is not limited to SAD. Broader reviews of AI in mental health care suggest that AI is increasingly being studied for diagnosis, monitoring and interventions across mental health conditions, while also highlighting similar concerns about data quality, generalizability, transparency, interpretability and ethical implementation (Cruz-Gonzalez et al., 2025). The present review therefore fits within a wider movement toward AI-supported mental health assessment, but also shows that SAD-specific evidence is not yet strong enough for routine clinical use. Before such tools can be responsibly implemented, future research should move from proof-of-concept within one dataset toward clinically credible, reproducible and deployable evidence.

Future research should first strengthen case definitions, reference standards and validation design. As shown in the results section on how social anxiety was defined and measured, studies used heterogeneous ground truths, including questionnaire cut-offs, clinical interviews, DSM-5-aligned researchers' labels and dichotomized self-ratings, which limit comparability and interpretability across papers. Future studies should therefore compare AI predictions with clinically meaningful reference standards, ideally clinically diagnosed SAD by a clinician-administered semi-structured diagnostic interview such as the Structured Clinical Interview for DSM-5 (SCID-5-CV) or the Anxiety Disorder Interview Schedule (ADIS), rather than merely reproducing study-specific screening labels. This comparison with clinical judgement is not the same as external validation. External validation refers to testing a model on an independent sample or setting beyond the development dataset. However, both are needed for clinical credibility: AI models should be tested in independent samples, and their predictions should be evaluated against the current clinical reference standard for diagnosing SAD.

Evaluation should also clarify which type of input or combination of inputs, and modeling leads to most reliable and clinically useful outcomes. This does not necessarily require broad comparison of all possible algorithms. Instead, stronger evidence could come from carefully designed studies that select one clinically plausible modality, such as interview transcript or digital-trace data, develop a transparent modeling approach, preregister the validation strategy and test whether the model generalizes across new participants and recruitment contexts. Future work should also examine whether multimodal approaches provide added value over single-modality models. AI may be particularly relevant when it can combine multiple sources of information, such as interview language, clinician-rated information, eye-tracking, digital traces, social-media data and other naturalistic behavioral data. Such multimodal approaches could help clinicians to identify patterns across data sources that are difficult to structure or weigh simultaneously. As highlighted in the study characteristics results, repeated reuse of a single questionnaire dataset across multiple papers concentrates the evidence and risks overstating generalizability. Future work should therefore expand to new datasets with clearly reported recruitment contexts and broader demographic and cultural coverage.

Future work should also improve reporting transparency, reproducibility and ethical governance. As shown in the results on validation and performance reporting, studies used heterogeneous metrics and often insufficiently detailed modeling processes. Future studies should transparently document preprocessing and modeling steps, including feature selection, imbalance handling, tuning and the treatment of overlapping predictors and outcomes. They should also report performance using complementary metrics that are interpretable under class imbalances. For

ecological or passive modalities, the ethical and practical issues summarized in the results section on ethical and practical considerations should be addressed more systematically, including privacy-preserving data practices, explicit consent procedures, transparency about what data are collected, careful consideration of false positives and false negatives, and clear governance around clinical follow-up after a positive AI-generated flag. These safeguards are especially important when tools rely on digital traces, language or continuous monitoring, because users may not always realize how sensitive these data are or what psychological information could be inferred from them.

Finally, responsible implementation will require a human-in-the-loop approach in which psychologists and other clinicians remain centrally involved in interpreting outputs, weighing error trade-offs and deciding on appropriate follow-up. This implies that future progress depends not only on better models, but also on greater AI literacy among clinicians. Explainable AI may support this process, but it should not be treated as a sufficient solution itself. Rather, its value depends on whether explanations are understandable, clinically relevant and capable of supporting an appropriate level of trust among clinicians, rather than either overreliance or undue skepticism (Rosenbacke et al., 2024).

Conclusion

The scoping review shows that AI approaches have potential to support the identification of social anxiety disorder, but the current literature is still too limited and heterogeneous to support routine use in practice. Although several studies reported promising results, these findings were usually based on internal validation, differing definitions of social anxiety and often questionnaire-based input, which makes the evidence difficult to compare and limits conclusions about generalizability and real clinical added value. The most promising direction appears to be the use of richer data sources, such as language, digital traces, virtual reality, eye-tracking and other multimodal behavioral data, where AI may offer more than a refined version of conventional screening. At present, these tools should therefore be viewed as a possible support to clinical decision-making or screening for SAD, rather than as a substitute for diagnosis. Further progress will require external validation, clearer reporting, careful attention to privacy and error consequences and sufficient AI literacy among clinicians to ensure that such tools are interpreted and used appropriately.

Funding

This scoping review itself received no external funding. Nine of the seventeen included studies reported receiving external funding, two explicitly stated that no external funding was received, and three did not report any funding information. The remaining three studies reported publication-related support (e.g., open-access/APC funding) or mixed statements indicating no research funding but support for publication costs. Full funding statements and any available grant details are provided in the full extraction table (Appendix C).

Conflicts of Interest

The author declares no conflict of interest.

Ethics Approval

Not applicable, as this review used only published literature and collected no new participant data.

References

- Ahmadi Majd, S., Parsaeian, M. R., Madani, M., Moradi, H., & Mohammadi, A. (2025). *A machine learning web application for screening social anxiety disorder based on participants' emotion regulation (ML-SAD)*. *Frontiers in Robotics and AI*, 12, Article 1620609. <https://doi.org/10.3389/frobt.2025.1620609>
- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, Article 101805. <https://doi.org/10.1016/j.inffus.2023.101805>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>
- Babuška, R., & Verbruggen, H. (2003). Neuro-fuzzy methods for nonlinear system identification. *Annual Reviews in Control*, 27(1), 73–85. [https://doi.org/10.1016/S1367-5788\(03\)00009-9](https://doi.org/10.1016/S1367-5788(03)00009-9)
- Bathula, S., Chadaga, K., Prabhu, S., & Sampathila, N. (2024). *Classification of social anxiety disorder using explainable machine learning and Pearson's correlation technique*. In *Proceedings of the 8th IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER 2024)* (pp. 292–297). IEEE. <https://doi.org/10.1109/DISCOVER62353.2024.10750623>
- Brown, T. A., Di Nardo, P. A., Lehman, C. L., & Campbell, L. A. (2001). *Reliability of DSM-IV anxiety and mood disorders: Implications for the classification of emotional disorders*. *Journal of Abnormal Psychology*, 110(1), 49–58. <https://doi.org/10.1037//0021-843X.110.1.49>
- Cesa, H. A., Lawsins, J. D., Chiongbian, C. J., Zaragoza, M. A., Pepito, G., & Álvarez, R. (2022). *Memores: A web-based intelligent screening tool for predicting social anxiety disorder using a machine learning model*. *Journal of Engineering Science and Technology*, Special Issue on ICITE2022, 32–47. https://jestec.taylors.edu.my/Special%20Issue%20ICITE2022/ICITE_2022_03.pdf
- Chapdelaine, A., Carrier, J.-D., Fournier, L., Duhoux, A., & Roberge, P. (2018). Treatment adequacy for social anxiety disorder in primary care patients. *PLOS ONE*, 13(11), e0206357. <https://doi.org/10.1371/journal.pone.0206357>
- Chadaga, K., Prabhu, S., Sampathila, N., Chadaga, R., Bhat, D., Sharma, A. K., & Swathi, K. S. (2024). SADXAI: Predicting social anxiety disorder using multiple interpretable artificial intelligence techniques. *SLAS Technology*, 29(2), 100129. <https://doi.org/10.1016/j.slact.2024.100129>

- Chang, M.-Y., & Tseng, C.-Y. (2020). *Detecting social anxiety with online social network data*. In 2020, 21st IEEE International Conference on Mobile Data Management (MDM) (pp. 333–336). IEEE. <https://doi.org/10.1109/MDM48529.2020.00073>
- Chlasta, K., Wisiecka, K., Krejtz, K., & Krejtz, I. (2024). *AI-based screening for depression and social anxiety through eye tracking: An exploratory study*. *International Journal of Marketing, Communication and New Media*, 15, 75–91. <https://doi.org/10.54663/2182-9306.2024.SpecialIssueMBP.75-91>
- Connor, K. M., Davidson, J. R. T., Churchill, L. E., Sherwood, A., Foa, E., & Weisler, R. H. (2000). Psychometric properties of the Social Phobia Inventory (SPIN): New self-rating scale. *The British Journal of Psychiatry*, 176(4), 379–386. <https://doi.org/10.1192/bjp.176.4.379>
- Çorbacioğlu, Ş. K., & Aksel, G. (2023). *Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value*. *Turkish Journal of Emergency Medicine*, 23(4), 195–198. https://doi.org/10.4103/tjem.tjem_182_23
- Cruz-Gonzalez, P., He, A. W.-J., Lam, E. P. P., Ng, I. M. C., Li, M. W., Hou, R., Chan, J. N.-M., Sahni, Y., Vinas-Guasch, N., Miller, T., Lau, B. W.-M., & Sánchez Vidaña, D. I. (2025). Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications. *Psychological Medicine*, 55, Article e18. <https://doi.org/10.1017/S0033291724003295>
- Demetriou, E. A., Park, S. H., Ho, N., Pepper, K. L., Song, Y. J. C., Naismith, S. L., Thomas, E. E., Hickie, I. B., & Guastella, A. J. (2020). *Machine learning for differential diagnosis between clinical conditions with social difficulty: Autism spectrum disorder, early psychosis, and social anxiety disorder*. *Frontiers in Psychiatry*, 11, 545. <https://doi.org/10.3389/fpsyt.2020.00545>
- Ertekin, E., Çelebi, F., Koyuncu, A., Uysal, Ö., Demir, E. Y., & Tükel, R. (2015). *Predictors of early or late treatment seeking in patients with social anxiety disorder*. *Annals of Clinical Psychiatry*, 27(4), 236–241. <https://doi.org/10.1177/104012371502700402>
- Fan, Y., Zhang, L., Yang, X., Zhou, Z., Zhang, Y., & Fang, T. (2025). Multimodal detection of social anxiety disorder: A deep learning and large language model approach using microblog data. In 2025 5th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) (pp. 1–4). <https://ieeexplore.ieee.org/document/11100179>

- Faouzi, J., & Colliot, O. (2023). Classic machine learning methods. In O. Colliot (Ed.), *Machine learning for brain disorders* (Neuromethods, Vol. 197, pp. 25–75). Humana.
https://doi.org/10.1007/978-1-0716-3195-9_2
- Fathi, S., & Ahmadi, M. (2020). *Data for: Development and use of a clinical decision support system for the diagnosis of social anxiety disorder* (Version 2) [Data set]. Mendeley Data.
<https://doi.org/10.17632/4jycfwhb4g.2>
- Fathi, S., Ahmadi, M., Birashk, B., & Dehnad, A. (2020). Development and use of a clinical decision support system for the diagnosis of social anxiety disorder. *Computer Methods and Programs in Biomedicine*, 190, 105354. <https://doi.org/10.1016/j.cmpb.2020.105354>
- Goan, E., & Fookes, C. (2020). Bayesian neural networks: An introduction and survey. In K. L. Mengersen, P. Pudlo, & C. P. Robert (Eds.), *Case studies in applied Bayesian data science: CIRM Jean-Morlet Chair, fall 2018* (pp. 45–87). Springer International Publishing.
https://doi.org/10.1007/978-3-030-42553-1_3
- Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews*, 18, e1230.
<https://doi.org/10.1002/cl2.1230>
- Hadie, S. N. H. (2024). *ABC of a scoping review: A simplified JBI scoping review guideline*. *Education in Medicine Journal*, 16(2), 185–197. <https://doi.org/10.21315/eimj2024.16.2.14>
- Heimberg, R. G., Horner, K. J., Juster, H. R., Safren, S. A., Brown, E. J., Schneier, F. R., & Liebowitz, M. R. (1999). Psychometric properties of the Liebowitz Social Anxiety Scale. *Psychological Medicine*, 29(1), 199–212. <https://doi.org/10.1017/S0033291798007879>
- Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). *On evaluation metrics for medical applications of artificial intelligence*. *Scientific Reports*, 12(1), Article 5979. <https://doi.org/10.1038/s41598-022-09954-8>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. <https://doi.org/10.1007/978-1-0716-1418-1>
- Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*, 9, Article 1117. <https://doi.org/10.3389/fpsyg.2018.01117>

- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
<https://doi.org/10.1038/nature14539>
- Li, C., Xu, B., Chen, Z., Huang, X., He, J., & Xie, X. (2024). *A stacking model-based classification algorithm is used to predict social phobia*. *Applied Sciences*, 14(1), 433.
<https://doi.org/10.3390/app14010433>
- Mahajan, P., Uddin, S., Hajati, F., Moni, M. A., & Gide, E. (2024). A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets. *Health and Technology*, 14, 597–613. <https://doi.org/10.1007/s12553-024-00835-w>
- Mattick, R. P., & Clarke, J. C. (1998). Development and validation of measures of social phobia, scrutiny fear, and social interaction anxiety. *Behaviour Research and Therapy*, 36(4), 455–470. [https://doi.org/10.1016/S0005-7967\(97\)10031-6](https://doi.org/10.1016/S0005-7967(97)10031-6)
- Mezzi, R., Yahyaoui, A., Krir, M. W., Boulila, W., & Koubaa, A. (2022). *Mental health intent recognition for Arabic-speaking patients using the Mini International Neuropsychiatric Interview (MINI) and BERT model*. *Sensors*, 22(3), 846. <https://doi.org/10.3390/s22030846>
- National Institute of Mental Health. (n.d.). *Social anxiety disorder*. Retrieved January 11, 2026, from <https://www.nimh.nih.gov/health/statistics/social-anxiety-disorder>
- National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). <https://doi.org/10.6028/NIST.AI.100-1>
- Ohse, J., Hadžić, B., Mohammed, P., Peperkorn, N., Fox, J., Krutzki, J., Lyko, A., Fan, M., Zheng, X., Rättsch, M., & Shiban, Y. (2024). *GPT-4 shows potential for identifying social anxiety from clinical interview data*. *Scientific Reports*, 14(1), 30498. <https://doi.org/10.1038/s41598-024-82192-2>
- Olariu, E., Forero, C. G., Castro-Rodríguez, J. I., Rodrigo-Calvo, M. T., Álvarez, P., Martín-López, L. M., Sánchez-Toto, A., Adroher, N. D., Blasco-Cubedo, M. J., Vilagut, G., Fullana, M. A., & Alonso, J. (2015). Detection of anxiety disorders in primary care: A meta-analysis of assisted and unassisted diagnoses. *Depression and Anxiety*, 32(7), 471–484.
<https://doi.org/10.1002/da.22360>
- Organisation for Economic Co-operation and Development. (2024). *Explanatory memorandum on the updated OECD definition of an AI system* (OECD Artificial Intelligence Papers No. 8). OECD Publishing. <https://doi.org/10.1787/623da898-en>

- Osório, F. L., Loureiro, S. R., Hallak, J. E. C., Machado-de-Sousa, J. P., Ushirohira, J. M., Baes, C. V. W., Apolinario, T. D., Donadon, M. F., Bolsoni, L. M., Guimarães, T., Fracon, V. S., Silva-Rodrigues, A. P. C., Pizeta, F. A., Souza, R. M., Sanches, R. F., dos Santos, R. G., Martin-Santos, R., & Crippa, J. A. S. (2019). Clinical validity and intrarater and test-retest reliability of the Structured Clinical Interview for DSM-5—Clinician Version (SCID-5-CV). *Psychiatry and Clinical Neurosciences*, 73(12), 754–760. <https://doi.org/10.1111/pcn.12931>
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5, Article 210. <https://doi.org/10.1186/s13643-016-0384-4>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J. M., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Pardini, M., Frumento, S., Martini, M., Rho, G., Vatteroni, V., Tharun, K., Alaimo, M., Galatolo, F. A., De Marinis, M., Scilingo, E. P., Menicucci, D., Cimino, M. G. C. A., Chessa, M., & Greco, A. (2025). *Deep learning-based classification of social anxiety disorder using continuous self-reported anxiety in virtual reality*. In *2025 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (pp. 1–6). IEEE. <https://doi.org/10.1109/MEMEA65319.2025.11068082>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer*. *Journal of Machine Learning Research*, 21(140), 1–67. <https://jmlr.org/papers/v21/20-074.html>
- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358. <https://doi.org/10.1056/NEJMr1814259>
- Rosellini, A. J., Liu, S., Anderson, G. N., Sbi, S., Tung, E. S., & Knyazhanskaya, E. (2020). *Developing algorithms to predict adult-onset internalizing disorders: An ensemble learning approach*. *Journal of Psychiatric Research*, 121, 189–196. <https://doi.org/10.1016/j.jpsychires.2019.12.006>
- Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. *JMIR AI*, 3, e53207. <https://doi.org/10.2196/53207>

- Salari, N., Heidarian, P., Hassanabadi, M., Babajani, F., Abdoli, N., Aminian, M., & Mohammadi, M. (2024). Global prevalence of social anxiety disorder in children, adolescents, and youth: A systematic review and meta-analysis. *Journal of Prevention*, 45(5), 795–813.
<https://doi.org/10.1007/s10935-024-00789-9>
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications, and research directions. *SN Computer Science*, 2(3), 160. <https://doi.org/10.1007/s42979-021-00592-x>
- Shihabudheen, K. V., & Pillai, G. N. (2018). Recent advances in neuro-fuzzy system: A survey. *Knowledge-Based Systems*, 152, 136–162. <https://doi.org/10.1016/j.knosys.2018.04.014>
- Stein, D. J., Lim, C. C. W., Roest, A. M., de Jonge, P., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Benjet, C., Bromet, E. J., Bruffaerts, R., de Girolamo, G., Florescu, S., Gureje, O., Haro, J. M., Harris, M. G., He, Y., Hinkov, H., Horiguchi, I., Hu, C., ... WHO World Mental Health Survey Collaborators. (2017). The cross-national epidemiology of social anxiety disorder: Data from the World Mental Health Survey Initiative. *BMC Medicine*, 15, Article 143.
<https://doi.org/10.1186/s12916-017-0889-2>
- Stein, M. B., & Stein, D. J. (2008). Social anxiety disorder. *The Lancet*, 371(9618), 1115–1125.
[https://doi.org/10.1016/S0140-6736\(08\)60488-2](https://doi.org/10.1016/S0140-6736(08)60488-2)
- Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist*, 76(1), 116–129.
<https://doi.org/10.1037/amp0000622>
- Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Hadid, V., Combrisson, E., O’Byrne, J., & Jerbi, K. (2023). *Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data*. *NeuroImage*, 277, 120253.
<https://doi.org/10.1016/j.neuroimage.2023.120253>
- Tlachac, M. L., & Ogden, S. S. (2024). *Student mental health screening with text message metadata*. In *2024 International Conference on Machine Learning and Applications (ICMLA)* (pp. 1337–1342). IEEE. <https://doi.org/10.1109/ICMLA61862.2024.00208>
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883. <https://doi.org/10.1037/0033-2909.133.5.859>

- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., ... Straus, S. E. (2018). *PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Annals of Internal Medicine, 169*(7), 467–473. <https://doi.org/10.7326/M18-0850>
- Wang, P. S., Berglund, P., Olfson, M., Pincus, H. A., Wells, K. B., & Kessler, R. C. (2005). Failure and delay in initial treatment contact after the first onset of mental disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 603–613. <https://doi.org/10.1001/archpsyc.62.6.603>
- World Health Organization. (2025, September 2). *Over a billion people living with mental health conditions – services require urgent scale*. <https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up>
- Xiong, H., Berkovsky, S., Romano, M., Sharan, R. V., Liu, S., Coiera, E., & McLellan, L. F. (2021). *Prediction of anxiety disorders using a feature ensemble-based Bayesian neural network. Journal of Biomedical Informatics, 123*, 103921. <https://doi.org/10.1016/j.jbi.2021.103921>
- Yokotani, K., Takano, M., Abe, N., & Kato, T. A. (2025). *Predicting social anxiety disorder based on communication logs and social network data from a massively multiplayer online game: Using a graph neural network. Psychiatry and Clinical Neurosciences, 79*(5), 274–281. <https://doi.org/10.1111/pcn.13804>
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open, 1*, 57–81. <https://doi.org/10.1016/j.aiopen.2021.01.001>

