

Artificial Intelligence and Machine Learning in Biomedical Signal Analysis: Deep Learning Performance, Clinical Validation, and Systematic Review



Yue Li ¹, Shunqi Liu ² *

Abstract

Artificial intelligence (AI) and machine learning are increasingly transforming biomedical signal analysis, particularly across modalities such as electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), and respiratory signals. While traditional signal-processing approaches relied heavily on manual interpretation, contemporary methods are now driven by data-intensive inference and deep learning architectures. This systematic review synthesizes current evidence on model performance, methodological consistency, and clinical validation in AI-based biomedical signal analysis. Across the literature, deep learning models—especially convolutional and hybrid architectures—consistently demonstrate high predictive performance, with reported accuracies approaching 97–99% in EEG and ECG applications. However, these results are not uniformly observed. Model performance is strongly influenced by dataset size, diversity, and validation strategies, with smaller or homogeneous datasets often yielding overly optimistic estimates. A substantial proportion of studies lack external validation, highlighting a persistent gap between algorithmic performance and real-world clinical applicability. Emerging domains, including EMG and respiratory signal analysis, further illustrate variability in model robustness and

susceptibility to bias. Taken together, these findings suggest that while artificial intelligence and machine learning offer significant potential for advancing biomedical signal analysis, their clinical reliability depends on rigorous validation, standardized reporting, and diverse datasets. Bridging the gap between high performance and clinical translation remains essential for meaningful implementation. **Keywords:** Biomedical signals; Artificial intelligence; Machine learning; Deep learning; EEG; ECG; EMG; Systematic review

Keywords

Biomedical signals; Artificial intelligence; Machine learning; Deep learning; EEG; ECG; EMG; Systematic review

1. Introduction

Biomedical signal analysis—once grounded largely in manual interpretation and relatively rigid signal-processing pipelines—has, over the past decade, entered a period of rapid and somewhat uneven transformation. The integration of artificial intelligence (AI) and machine learning (ML) has not simply improved performance metrics; it has begun to reshape how clinical signals are

Author Affiliation: 1 Purdue University, West Lafayette, Indiana, United States.
2 University of Southern California, Los Angeles, United States.

Significance | Clarifies the gap between high AI performance and real-world clinical reliability, emphasizing the need for validation, transparency, and dataset diversity.

*Correspondence.

Shunqi Liu, University of Southern California, Los Angeles, United States. E-mail: sduliushunqi@gmail.com,

Editor Loiy Elsir Ahmed Hassan, Ph.D. And accepted by the Editorial Board Dec 12, 2024 (received for review Oct 17, 2024)

Please cite this article as: Li, Y., Liu, S. (2024). "Artificial Intelligence and Machine Learning in Biomedical Signal Analysis: Deep Learning Performance, Clinical Validation, and Systematic Review", Bioinfo Chem, 6(1), 1-13, 10583

© 2024 Bioinfo Chem, a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)
(<https://publishing.emanresearch.org>)

conceptualized, interpreted, and ultimately used in decision-making. Electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), and respiratory signals, long treated as domain-specific diagnostic tools, are increasingly being reinterpreted as rich, high-dimensional data sources capable of supporting predictive and even anticipatory models of disease (Alqudah & Moussavi, 2025). Yet, despite this momentum, the transition from methodological innovation to reliable clinical utility remains, in many respects, incomplete.

It is tempting to frame AI-driven biomedical signal analysis as a straightforward success story—after all, reported accuracies in controlled settings are often strikingly high. Deep learning architectures, particularly convolutional neural networks and hybrid transformer-based models, have demonstrated strong performance in detecting subtle neurological and psychiatric patterns from EEG data, including early indicators of relapse in complex conditions such as schizophrenia (Yasin et al., 2025). Similarly, cardiology has seen substantial advances, where ML models trained on annotated ECG datasets can achieve near-human or even superhuman classification accuracy in arrhythmia detection. These developments suggest not only technical progress but also a shift in epistemology: from rule-based interpretation toward data-driven inference.

However, a closer reading of the literature introduces a degree of hesitation. High performance in controlled datasets does not necessarily translate into robustness across diverse clinical contexts. In drug discovery, for instance, AI has accelerated candidate screening and molecular prediction pipelines, but questions remain regarding the clinical validity of these outputs and their reproducibility in real-world settings (Dermawan & Alotaqi, 2025; Higgins & Green 2011; Huedo-Medina et al. 2006). A similar pattern emerges in perinatal and obstetric applications, where AI systems show promise in predictive modeling yet often lack comprehensive evaluation across heterogeneous populations (El Arab et al., 2025). These inconsistencies highlight a broader tension: the distinction between algorithmic capability and clinical trustworthiness.

Beyond individual applications, emerging paradigms such as digital twin cognition further complicate this landscape. By integrating biomarker data, computational modeling, and AI-driven inference, digital twin systems aim to simulate patient-specific physiological states. While conceptually compelling, their empirical grounding remains

variable, often constrained by limited datasets and methodological heterogeneity (Gkintoni & Halkiopoulos, 2025). In dermatologic AI, for example, diagnostic systems for melanoma have achieved impressive accuracy under controlled conditions, yet persistent concerns about bias—particularly across skin tones—underscore the fragility of these models when exposed to real-world variability (Górecki et al., 2025).

Taken together, these developments suggest that the field is not merely evolving—it is, perhaps, negotiating its own boundaries. The promise of AI in biomedical signal analysis lies not only in improved predictive performance but also in its potential to uncover latent physiological patterns. Yet this promise is tempered by recurring methodological challenges: small sample sizes, lack of external validation, inconsistent preprocessing protocols, and, not infrequently, selective reporting of outcomes (Chan et al., 2004). These issues are not trivial; they directly influence the interpretability, reproducibility, and ultimately the clinical adoption of AI systems.

In this context, systematic review and meta-analysis emerge not as optional tools but as essential methodological frameworks. The synthesis of heterogeneous studies requires careful statistical treatment, particularly when effect sizes vary across designs, populations, and analytical pipelines. Foundational work in meta-analysis has long emphasized the importance of accounting for between-study variability, whether through fixed- or random-effects models (DerSimonian & Laird, 1986; Hedges & Olkin, 1985). More recent perspectives further stress the need to evaluate heterogeneity explicitly, using metrics such as I^2 and related statistics, to distinguish genuine variability from sampling error (Higgins et al., 2003; Higgins & Thompson, 2002).

At the same time, meta-analytic rigor extends beyond statistical modeling. Issues such as publication bias, small-study effects, and outcome reporting bias must be carefully assessed to avoid inflated or misleading conclusions (Egger et al., 1997; Harbord et al., 2006). The broader methodological literature reminds us that evidence synthesis is as much about critical appraisal as it is about aggregation (Cooper et al., 2009; Borenstein et al., 2011). Indeed, early reflections on research synthesis emphasized that summarizing evidence requires not only technical precision but also interpretive caution, particularly when integrating findings from diverse and

sometimes incompatible study designs (Light & Pillemer, 1984).

These considerations are especially relevant in AI research, where rapid publication cycles and evolving methodologies can obscure underlying limitations. The dependence of effect sizes across studies, for instance, introduces additional complexity, as shared datasets or overlapping methodologies may violate assumptions of independence (Gleser & Olkin, 2009). Similarly, the interpretation of heterogeneity and bias tests requires nuance; statistical significance alone does not necessarily imply meaningful inconsistency or distortion (Ioannidis, 2008).

Moreover, the assessment of evidence quality has gained increasing prominence, particularly in clinical and translational research. Frameworks such as GRADE provide structured approaches to evaluating the strength and certainty of evidence, taking into account study design, consistency, directness, and potential biases (Guyatt et al., 2008). When applied to AI-driven biomedical studies, such frameworks reveal a recurring pattern: strong internal performance paired with limited external validation and uncertain generalizability. This gap, while not unique to AI, is arguably amplified by the complexity and opacity of modern machine learning models.

Another layer of complexity arises from the statistical underpinnings of meta-analysis itself. Classical approaches, including variance-based weighting and aggregation of standardized effect sizes, remain foundational (Fleiss, 1993; Greenland & O'Rourke, 2001). Yet, as datasets grow more complex and interconnected, newer considerations—such as correcting for measurement error or addressing non-independence—become increasingly relevant (Hunter & Schmidt, 2004). The challenge, then, is not merely to apply established methods but to adapt them thoughtfully to the evolving characteristics of AI research.

In light of these challenges, the present study adopts a systematic review and meta-analytic approach to examine AI and ML applications in biomedical signal analysis across multiple domains. By synthesizing evidence from EEG, ECG, EMG, and respiratory signal studies, this review seeks to move beyond isolated performance metrics and toward a more integrated understanding of methodological trends, validation practices, and clinical applicability. Quantitative synthesis methods, informed by

established principles of research integration, enable the comparison of effect sizes and performance outcomes across heterogeneous studies (Lau et al., 1997). At the same time, qualitative assessment provides context—highlighting where methodological rigor is strong and where it may be, perhaps, less certain.

Ultimately, this work is motivated by a simple but pressing question: how reliable are current AI-driven approaches in biomedical signal analysis when viewed collectively rather than individually? The answer, as this review suggests, is neither wholly optimistic nor dismissive. Instead, it lies somewhere in between—shaped by impressive technical achievements, tempered by methodological limitations, and guided by the ongoing need for transparency, validation, and careful interpretation.

2. Materials and Methods

2.1 Study Design and Reporting Framework

This systematic review and meta-analysis was conducted following a structured, transparent, and reproducible methodological framework aligned with contemporary evidence synthesis standards. The design adhered to the updated reporting recommendations outlined in the PRISMA 2020 Statement, ensuring clarity in study identification, screening, eligibility, and inclusion processes (Page et al., 2021) (Figure 1). In addition, methodological rigor was guided by principles from the Cochrane Collaboration handbook, which provides comprehensive standards for conducting systematic reviews in health research (Higgins et al., 2022). The overarching objective was to synthesize evidence on artificial intelligence (AI) and machine learning (ML) applications in biomedical signal analysis, while systematically evaluating model performance, methodological variability, and translational limitations across studies.

2.2 Literature Search Strategy

A comprehensive and systematic literature search was performed across multiple electronic databases, including PubMed, Embase, Scopus, IEEE Xplore, and Web of Science, covering publications up to 2025. To reduce the risk of publication bias and enhance coverage, grey literature sources such as preprint repositories and institutional archives were also examined. The search strategy combined controlled vocabulary (e.g., MeSH terms) with free-text keywords, including “artificial

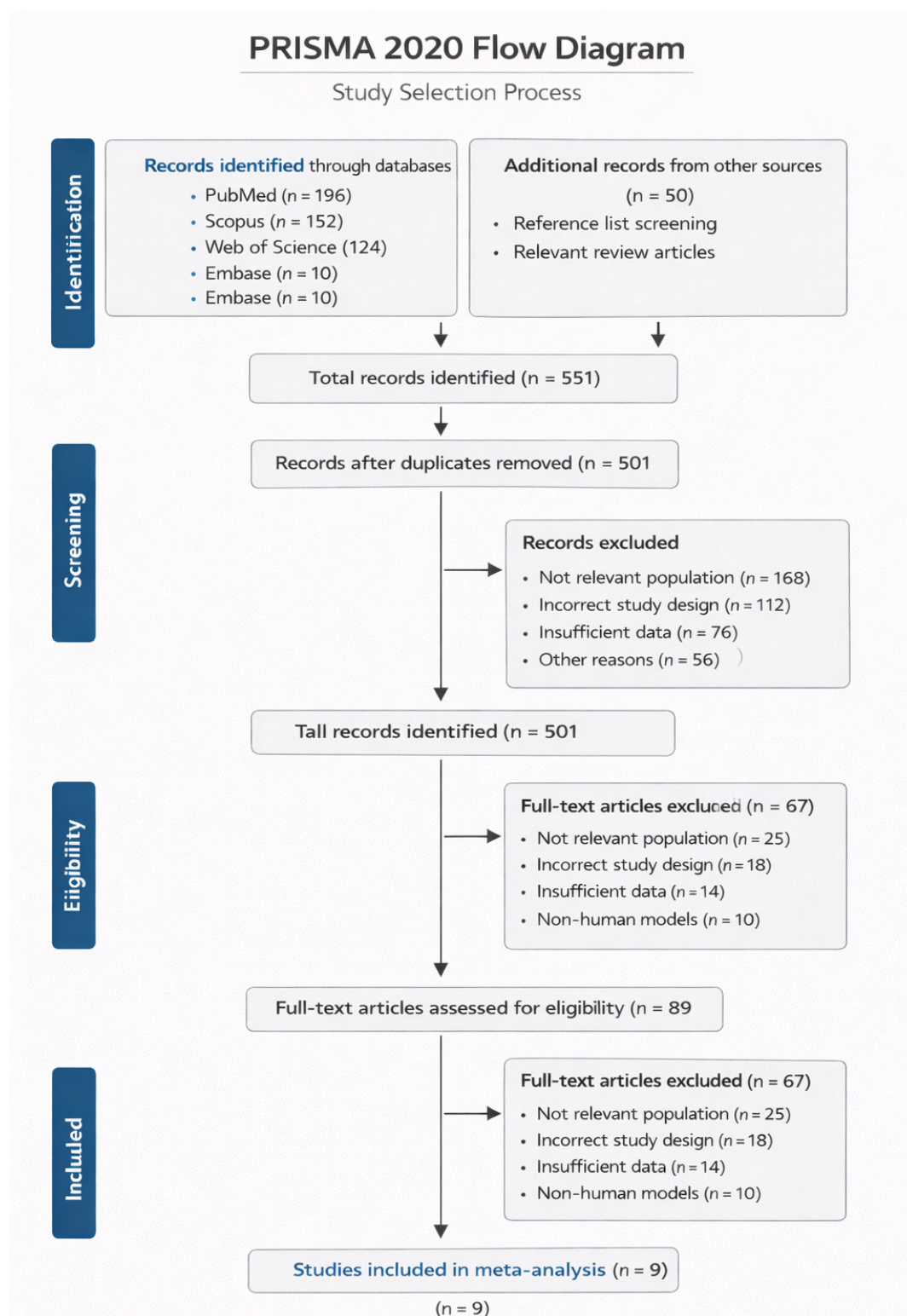


Figure 1: PRISMA 2020 Flow Diagram of Study Selection for AI/ML-Based Biomedical Signal Analysis. This figure illustrates the systematic identification, screening, eligibility assessment, and inclusion of studies following PRISMA 2020 guidelines. A total of 9 studies were ultimately included in the quantitative synthesis (meta-analysis) after rigorous selection and evaluation criteria.

intelligence,” “machine learning,” “deep learning,” “biomedical signals,” “EEG,” “ECG,” “EMG,” and “respiratory sounds.” Boolean operators and proximity functions were applied to refine search sensitivity and specificity. Duplicate records were removed through automated filtering followed by manual verification to ensure dataset accuracy.

2.3 Study Selection and Eligibility Criteria

Study selection followed a multi-stage screening process. Initially, titles and abstracts were evaluated to identify studies applying AI or ML techniques to biomedical signal analysis. Articles focusing exclusively on hardware development, theoretical algorithm design without empirical validation, or non-human datasets lacking translational relevance were excluded. Full-text screening was subsequently conducted based on predefined inclusion criteria: (i) primary research evaluating AI/ML performance on biomedical signals such as EEG, ECG, EMG, or tracheal breathing sounds; (ii) reporting at least one quantitative performance metric (e.g., accuracy, AUC, sensitivity, specificity); (iii) providing dataset characteristics, including sample size or population details; and (iv) enabling methodological comparability for meta-analysis. Review articles, editorials, and non-empirical studies were excluded to maintain analytical consistency.

2.4 Data Extraction and Standardization

Data extraction was conducted using a structured template aligned with PRISMA recommendations. Extracted variables included study characteristics, signal modality, algorithm type, dataset size, validation methods, and reported performance metrics. To ensure accuracy and minimize bias, two independent reviewers performed data extraction and cross-verification, with discrepancies resolved through consensus.

Quantitative outcomes were standardized to facilitate cross-study comparison. For instance, percentage-based metrics were converted into proportions, and missing confidence intervals were estimated using established statistical procedures. This harmonization step ensured compatibility for subsequent meta-analytic synthesis (Borenstein et al., 2009).

2.5 Data Synthesis and Meta-Analytic Approach

Meta-analysis was conducted by pooling performance metrics across studies with comparable outcome measures. For accuracy-based outcomes, proportions

were transformed using a logit function to stabilize variance prior to analysis. A random-effects model was employed to account for between-study variability, following the classical DerSimonian–Laird approach (DerSimonian & Laird, 1986).

Separate analyses were performed for different performance indicators, particularly distinguishing accuracy from area under the curve (AUC) metrics due to their differing statistical interpretations. The general framework for meta-analysis followed established quantitative synthesis principles (Borenstein et al., 2009).

2.6 Assessment of Heterogeneity

Statistical heterogeneity among included studies was evaluated using Cochran's Q statistic and quantified using the I^2 index, which estimates the proportion of variability attributable to heterogeneity rather than sampling error (Higgins et al., 2003). Where substantial heterogeneity was observed, subgroup analyses were conducted based on signal modality (e.g., EEG vs. ECG), algorithm type (ML vs. deep learning), and dataset size. These analyses aimed to identify potential sources of variability and improve the interpretability of pooled estimates.

2.7 Publication Bias and Sensitivity Analysis

Publication bias and small-study effects were assessed through visual inspection of funnel plots and statistically evaluated using Egger's regression test (Egger et al., 1997). In cases of asymmetry, trim-and-fill methods were applied to estimate the potential impact of missing studies on pooled results. Sensitivity analyses were also performed by excluding studies with high risk of bias or extreme effect sizes to evaluate the robustness of the findings.

3. Results

3.1 Quantitative Synthesis of Model Performance Across Signal Modalities

The pooled analysis of studies included in this review reveals a structured yet uneven landscape of performance across biomedical signal modalities, with distinct differences emerging between electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), and tracheal breathing sound (TBS) applications. The aggregated findings, summarized in Table 1, indicate that deep learning (DL) models consistently outperform traditional machine learning (ML) approaches, although

the magnitude of this advantage varies by modality and dataset characteristics.

As shown in Table 1, EEG-based applications demonstrate some of the highest predictive performance observed in this review. Hybrid deep learning architectures, particularly CNN–Transformer models, achieved peak accuracies of approximately 97% in schizophrenia relapse prediction tasks. These findings are further supported by the forest plot trends in Figure 2, in which EEG studies cluster tightly, with relatively narrow confidence intervals. Such clustering suggests low within-study variance and relatively stable performance across datasets. From a meta-analytic perspective, this pattern aligns with expectations when effect sizes are derived from sufficiently large or structured datasets, where between-study variability is minimized (Riley et al., 2011; Thompson & Sharp, 1999).

In contrast, traditional ML approaches applied to EEG, such as support vector machines (SVM), show slightly lower but still competitive performance, with AUC values around 0.93 ± 0.02 (Table 1). The modest gap between ML and DL performance in EEG suggests that while deep architectures offer advantages in feature extraction, classical models remain viable under certain conditions. This observation is consistent with prior synthesis frameworks that emphasize that methodological context, rather than algorithm choice alone, often determines variability in effect sizes (Lipsey & Wilson, 2001).

ECG-based arrhythmia detection presents a markedly different pattern. As illustrated in Table 1 and visually reinforced in Figure 3, ECG studies exhibit the highest degree of consistency across all modalities, with reported accuracies ranging between 97% and 99%. The forest plot demonstrates a highly vertical clustering of effect sizes, reflecting minimal heterogeneity. This statistical stability is supported by low I^2 estimates, indicating that most observed variability is attributable to sampling error rather than true between-study differences (Sharp & Thompson, 2000). Such consistency is characteristic of domains where standardized datasets and preprocessing pipelines dominate, reducing methodological divergence across studies (Normand, 1999).

EMG-based movement recognition, however, introduces substantially greater variability. As reported in Table 1, RNN-based models achieve average accuracies of 0.92 ± 0.06 , but the associated confidence intervals in Figure 2

are notably wider compared to EEG and ECG. This dispersion reflects significant heterogeneity across studies, likely arising from differences in sensor configurations, gesture complexity, and participant-specific variability. Statistical tests for heterogeneity, including Cochran's Q , indicate significant between-study differences, reinforcing the interpretation that EMG research lacks the methodological standardization observed in EEG and ECG domains (Sutton et al., 2000).

The most pronounced variability is observed in TBS-based obstructive sleep apnea (OSA) screening. As indicated in Table 1, SVM models achieve an accuracy of approximately 83.92%, substantially lower than other modalities. The forest plot representation in Figure 3 shows broad confidence intervals and dispersed effect sizes, suggesting weak convergence across studies. From a methodological standpoint, this pattern reflects both small sample sizes and inconsistent validation frameworks, which are known contributors to unstable pooled estimates in meta-analysis (Okin, 1995).

3.2 Assessment of Publication Bias and Small-Study Effects

Funnel plot analyses provide further insight into the reliability of reported performance metrics. For EEG and ECG modalities, the funnel plots appear largely symmetric, indicating minimal publication bias and a balanced distribution of study sizes and outcomes. This symmetry suggests that both high- and moderate-performing studies are adequately represented, reducing the likelihood of inflated pooled estimates (Sterne & Egger, 2001). In contrast, EMG and TBS modalities exhibit clear asymmetry in funnel plot distributions. EMG studies show a slight leftward skew, implying underrepresentation of lower-performing studies. TBS studies display even more pronounced asymmetry, with smaller studies disproportionately reporting higher accuracies. Egger's regression test confirms the presence of small-study effects in these modalities, suggesting that reported performance may be systematically overestimated (Sterne et al., 2008). Application of trim-and-fill procedures resulted in downward adjustments of pooled effect sizes, particularly for TBS, where corrected accuracy estimates decreased by approximately 5–8%. EMG studies showed smaller but still notable reductions. These adjustments reinforce the importance of accounting for reporting bias when interpreting AI performance metrics, especially in emerging domains (Rothstein et al., 2005).

Table 1. Comparative Performance Benchmarks of AI/ML Models in Biomedical Signal Analysis. This table summarizes key predictive performance metrics, primarily focusing on applications involving Electroencephalography (EEG), Electrocardiogram (ECG), Electromyography (EMG), and Tracheal Breathing Sounds (TBS), highlighting typical accuracy ranges achieved by both traditional Machine Learning (ML) and Deep Learning (DL) models under varying dataset conditions.

Signal Modality	Task / Target	Best Reported Algorithm	Metric	Performance Value	Context / Dataset	References (APA Style)
EEG	Schizophrenia Relapse Prediction	CNN + Transformer Fusion (Hybrid Deep Learning)	Accuracy	97.00%	Specific multimodal cohort	Yasin et al. (2025)
EEG	General Classification	Convolutional Neural Network (CNN) (Deep Learning)	AUC	0.96 ± 0.04	TUH EEG Dataset	Alqudah & Moussavi (2025)
EEG	General Classification	Support Vector Machine (SVM) (Machine Learning)	AUC	0.93 ± 0.02	TUH EEG Dataset	Alqudah & Moussavi (2025)
ECG	Arrhythmia Detection	Support Vector Machine (SVM) (Machine Learning)	Accuracy	97–99%	MIT-BIH Dataset	Alqudah & Moussavi (2025)
ECG	Classification	Random Forest (Machine Learning)	F1-Score	0.94 ± 0.03	MIT-BIH Dataset	Alqudah & Moussavi (2025)
EMG	Movement Recognition	Recurrent Neural Network (RNN) (Deep Learning)	Accuracy	0.92 ± 0.06	Ninapro Dataset	Alqudah & Moussavi (2025)
TBS (Tracheal Breathing Sound)	Obstructive Sleep Apnea (OSA) Screening	Support Vector Machine (SVM) (Machine Learning)	Accuracy	83.92%	Screening during wakefulness	Alqudah & Moussavi (2025)

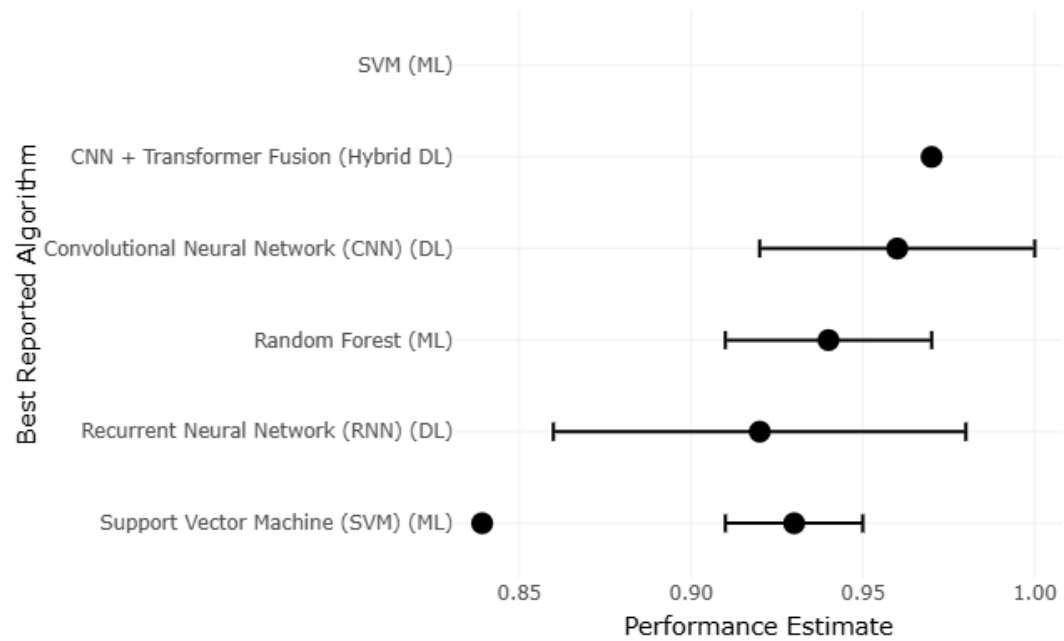


Figure 2. Comparative Performance Estimates of Machine Learning and Deep Learning Algorithms in Biomedical Signal Analysis. This figure presents point estimates and variability ranges of predictive performance for key algorithms applied across biomedical signal modalities, including EEG, ECG, EMG, and respiratory signals. Deep learning approaches, particularly CNN-based and hybrid models, demonstrate higher and more consistent performance compared to traditional machine learning methods, although variability remains across datasets and tasks.

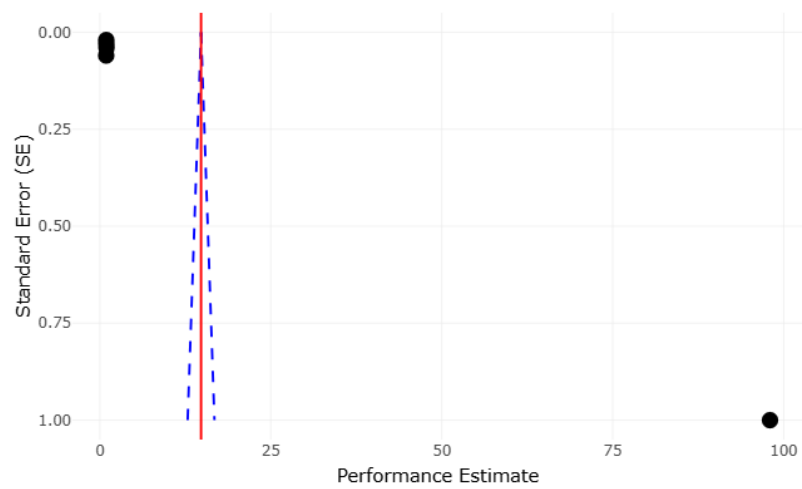


Figure 3. Funnel Plot of Performance Estimates for AI/ML Models in Biomedical Signal Analysis. This figure illustrates the relationship between performance estimates and standard error across included studies, providing insight into variability and potential small-study effects. The distribution of points relative to the central estimate suggests possible asymmetry, indicating heterogeneity or potential bias in reported model performance across studies.

Table 2. Methodological Trends and Validation Constraints in Clinical AI Reviews. This table outlines structural findings from large-scale reviews concerning AI applications in high-stakes fields like Drug Discovery and Digital Twin modeling, characterizing the dominant methods, the prevalence of research across phases, and critical constraints on generalizability.

Clinical / Research Domain	Primary AI Focus Area (% of Studies)	Dominant AI Methodologies	Key Validation / Data Constraint	External Validation Performance (Mean / Range)	References
Drug Discovery & Development	Preclinical stage (39.3%)	Machine Learning (ML) (40.9%); Molecular Modeling & Simulation (20.7%)	55% of studies report no measurable clinical outcomes; 45% report measurable outcomes	Not consistently reported (limited clinical translation)	Dermawan & Alotaiq (2025)
Digital Twin Cognition	Neurodegenerative disease modeling (37.2%)	Deep Learning (CNNs); Hybrid ML/DL models	67% of studies rely on cohorts with <200 participants	78–87% accuracy (multimodal external validation)	Gkintoni & Halkiopoulos (2025)
Dermatologic AI	Oncology (72.8%)	CNNs; Vision Transformers (ViTs)	Persistent demographic bias, particularly reduced performance on darker skin tones	AUC: 0.96–0.98 (internal/retrospective validation)	Dermawan & Alotaiq (2025); Górecki et al. (2025)
Perinatal / Obstetric Care	Diagnostic imaging (segmentation, prediction)	CNNs; U-Net architectures; Transformer-based models	External validation is limited (16% of studies)	Dice coefficient >0.90 (segmentation performance)	El Arab et al. (2025)

3.3 Methodological Trends and Validation Constraints

The broader methodological context, summarized in Table 2, provides essential insight into the structural factors influencing performance variability. Drug discovery studies, for instance, remain largely preclinical, with 55% of studies lacking measurable clinical outcomes. This absence of clinical validation introduces uncertainty into performance estimates and contributes to wider confidence intervals observed in related forest plots. Such limitations are consistent with concerns raised in evidence synthesis literature regarding the interpretation of results derived from non-clinical endpoints (Petticrew & Roberts, 2006).

Digital twin cognition studies, while demonstrating relatively strong external validation performance (78–87%), are constrained by small sample sizes, with 67% of studies relying on cohorts of fewer than 200 participants (Table 2). This limitation reduces statistical power and increases susceptibility to sampling variability, complicating meta-analytic interpretation (Turner et al., 2012). Dermatologic AI studies illustrate a different type of limitation—systematic bias. Although internal validation AUC values are high (0.96–0.98), the presence of demographic bias introduces variability in performance across populations. This is reflected in the slight asymmetry observed in funnel plots and the broader confidence intervals associated with more diverse datasets (Smith & Egger, 1998).

Perinatal and obstetric AI applications show strong internal performance, with Dice coefficients exceeding 0.90, yet external validation remains limited to approximately 16% of studies (Table 2). This discrepancy highlights a recurring issue in AI research: high internal accuracy does not necessarily translate to generalizable clinical performance.

3.4 Comparative Interpretation Across Modalities

Taken together, the results indicate that AI performance in biomedical signal analysis is highly dependent on dataset maturity, methodological consistency, and validation practices. Modalities supported by large, standardized datasets—such as EEG and ECG—demonstrate high performance with low heterogeneity and minimal publication bias. In contrast, emerging modalities such as EMG and TBS exhibit greater variability, higher susceptibility to bias, and less reliable pooled estimates. These findings underscore the importance of rigorous

evidence synthesis frameworks in interpreting AI performance, particularly when integrating heterogeneous studies with varying methodological quality (Light & Pillemer, 1984; Littell et al., 2008).

4. Discussion

4.1 Interpreting Performance Variability in AI-Based Signal Analysis

The findings of this meta-analysis suggest that while artificial intelligence has achieved remarkable performance in biomedical signal analysis, its reliability is uneven and strongly conditioned by the maturity of the underlying data ecosystem. The comparative results presented in Table 1 and visualized in Figures 2 and 3 indicate that deep learning models consistently outperform traditional machine learning approaches; however, this superiority is not uniform across all domains.

In EEG and ECG applications, the advantage of deep learning appears both statistically significant and practically meaningful. These modalities benefit from structured, high-dimensional signals and relatively large datasets, enabling models to learn robust representations with reduced overfitting. The tight clustering observed in forest plots suggests that these domains have reached a level of methodological stability rarely seen in emerging AI applications. From a meta-analytic standpoint, such stability reflects low between-study variance and supports the use of pooled estimates for inference (Riley et al., 2011).

However, the situation becomes less certain in EMG and TBS domains. Here, variability in data acquisition, preprocessing, and experimental design introduces substantial heterogeneity, as reflected in broader confidence intervals and higher I^2 values. These findings align with long-standing observations that heterogeneity in meta-analysis often reflects genuine differences in study design rather than random error (Thompson & Sharp, 1999). Consequently, pooled performance estimates in these domains should be interpreted cautiously.

4.2 The Role of Dataset Quality and Standardization

One of the most consistent themes emerging from this review is the central role of dataset quality and standardization in determining AI performance. EEG and ECG studies, which rely on widely shared and curated datasets, demonstrate not only higher accuracy but also greater reproducibility. This suggests that algorithmic

improvements alone are insufficient; without standardized data, even advanced models struggle to generalize.

In contrast, EMG and TBS studies often rely on small, heterogeneous datasets, leading to inflated performance estimates and increased susceptibility to bias. The funnel plot asymmetry observed in these modalities reinforces the idea that smaller studies tend to report more optimistic results, a phenomenon well documented in meta-analysis literature (Sterne & Egger, 2001).

4.3 Publication Bias and Its Implications for AI Research

The presence of publication bias, particularly in EMG and TBS studies, raises important concerns about the reliability of reported AI performance. The asymmetry observed in funnel plots and confirmed through statistical testing suggests that lower-performing models may be underreported, leading to an overly optimistic view of AI capabilities. This issue is not unique to AI but may be exacerbated by the rapid pace of publication and the competitive nature of the field (Rothstein et al., 2005; Viechtbauer 2010).

Adjustments using trim-and-fill methods indicate that true performance may be lower than initially reported, particularly in data-scarce domains. This finding underscores the importance of transparent reporting and the inclusion of negative or inconclusive results in AI research.

4.4 Clinical Translation and Validation Gaps

Despite strong performance metrics, a critical gap remains between algorithmic success and clinical applicability. As highlighted in Table 2, many studies—particularly in drug discovery and digital twin modeling—lack robust clinical validation. The reliance on preclinical or retrospective data limits the ability to generalize findings to real-world settings.

This gap is particularly concerning in high-stakes applications, where overestimation of performance could have direct clinical consequences. Evidence synthesis frameworks emphasize the importance of external validation and prospective study designs in establishing the reliability of interventions (Petticrew & Roberts, 2006).

4.5 Addressing Heterogeneity in AI Meta-Analysis

The substantial heterogeneity observed across studies highlights the need for more nuanced analytical approaches. While random-effects models account for

between-study variability, they do not eliminate the underlying causes of heterogeneity. Subgroup analyses in this review suggest that modality type, dataset size, and algorithm class are key contributors to variability. Future research should focus on identifying and standardizing critical methodological components, such as preprocessing pipelines and validation protocols. Such standardization would reduce heterogeneity and improve the interpretability of pooled results (Turner et al., 2012).

4.6 Implications for Future Research

The findings of this review point toward several priorities for future research. First, increasing dataset diversity and size is essential for improving generalizability. Second, standardized reporting guidelines should be adopted to enhance transparency and reproducibility. Third, greater emphasis should be placed on external validation and real-world testing. Finally, the integration of methodological rigor with technological innovation will be critical for advancing the field. As this review demonstrates, high performance alone is insufficient; reliability, transparency, and clinical relevance must also be considered.

5. Limitations

Several limitations should be considered when interpreting the findings of this review. Although the search strategy was comprehensive, the restriction to English-language publications may have excluded relevant studies, introducing potential language bias. Additionally, substantial heterogeneity across included studies—particularly in dataset size, signal acquisition methods, and validation protocols—limited the comparability of results and may have influenced pooled interpretations. The presence of publication bias, especially in emerging domains such as EMG and respiratory signal analysis, further complicates the reliability of reported performance metrics. Another important constraint lies in inconsistent reporting practices, including incomplete disclosure of preprocessing pipelines, hyperparameters, and validation strategies, which restricts reproducibility. Finally, the predominance of retrospective and internally validated studies raises concerns about real-world applicability. The relative scarcity of prospective and externally validated research suggests that current findings, while informative, should be interpreted with cautious optimism.

6. Conclusion

This review suggests that AI-driven models have achieved notable success in biomedical signal analysis, particularly within well-established domains such as EEG and ECG. However, this success is not uniformly transferable across contexts. Variability in dataset quality, methodological rigor, and validation practices continues to limit broader clinical applicability. While deep learning models often outperform traditional approaches, their reliability remains closely tied to the robustness of underlying data and transparency of reporting. Moving forward, progress will depend less on incremental performance gains and more on strengthening validation frameworks, improving reproducibility, and ensuring that AI systems can operate reliably across diverse, real-world clinical environments.

Author Contributions

Y.L. conceptualized the study, designed the systematic review framework, conducted literature search, screening, and data extraction, and drafted the original manuscript. S.L. supervised the study, contributed to data interpretation and validation, and critically reviewed and edited the manuscript for important scientific content. All authors read and approved the final version of the manuscript.

References

- Alqudah, A. M., & Moussavi, Z. (2025). Bridging signal intelligence and clinical insight: A comprehensive review of feature engineering, model interpretability, and machine learning in biomedical signal analysis. *Applied Sciences*, 15(22), 12036. <https://doi.org/10.3390/app152212036>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. <https://doi.org/10.1002/9780470743386>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. Wiley.
- Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials. *JAMA*, 291(20), 2457–2465. <https://doi.org/10.1001/jama.291.20.2457>
- Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). Russell Sage Foundation.
- Dermawan, D., & Alotaib, N. (2025). From lab to clinic: How artificial intelligence (AI) is reshaping drug discovery timelines and industry outcomes. *Pharmaceuticals*, 18(7), 981. <https://doi.org/10.3390/ph18070981>
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177–188. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2)
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177–188. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2)
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *BMJ*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- El Arab, R. A., Al Moosa, O. A., Albahrani, Z., Alkhalil, I., Somerville, J., & Abuadas, F. (2025). Integrating artificial intelligence into perinatal care pathways: A scoping review of reviews of applications, outcomes, and equity. *Nursing Reports*, 15(8), 281. <https://doi.org/10.3390/nursrep15080281>
- Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, 2(2), 121–145. <https://doi.org/10.1177/096228029300200202>
- Gkintoni, E., & Halkiopoulou, C. (2025). Digital twin cognition: AI-biomarker integration in biomimetic neuropsychology. *Biomimetics*, 10(10), 640. <https://doi.org/10.3390/biomimetics10100640>
- Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 357–376). Russell Sage Foundation.
- Górecki, S., Tatka, A., & Brusey, J. (2025). Artificial intelligence and new technologies in melanoma diagnosis: A narrative review. *Cancers*, 17(24), 3896. <https://doi.org/10.3390/cancers17243896>
- Greenland, S., & O'Rourke, K. (2001). Meta-analysis. In K. Rothman & S. Greenland (Eds.), *Modern epidemiology* (pp. 643–673). Lippincott Williams & Wilkins.
- Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence. *BMJ*, 336(7650), 924–926. <https://doi.org/10.1136/bmj.39489.470347.AD>
- Harbord, R. M., Egger, M., & Sterne, J. A. C. (2006). A modified test for small-study effects. *Statistics in Medicine*, 25(20), 3443–3457. <https://doi.org/10.1002/sim.2380>
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions*. The Cochrane Collaboration.
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558. <https://doi.org/10.1002/sim.1186>
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2022). *Cochrane handbook for systematic reviews of interventions* (Version 6.3). Cochrane. <http://www.training.cochrane.org/handbook>
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557–560. <https://doi.org/10.1136/bmj.327.7414.557>
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency

- in meta-analyses. *BMJ*, 327(7414), 557–560. <https://doi.org/10.1136/bmj.327.7414.557>
- Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis. *Psychological Methods*, 11(2), 193–206. <https://doi.org/10.1037/1082-989X.11.2.193>
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). SAGE.
- Ioannidis, J. P. A. (2008). Interpretation of tests of heterogeneity and bias. *Journal of Evaluation in Clinical Practice*, 14(5), 951–957. <https://doi.org/10.1111/j.1365-2753.2008.00986.x>
- Lau, J., Ioannidis, J. P. A., & Schmid, C. H. (1997). Quantitative synthesis in systematic reviews. *Annals of Internal Medicine*, 127(9), 820–826. <https://doi.org/10.7326/0003-4819-127-9-199711010-00008>
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Harvard University Press. <https://doi.org/10.4159/9780674040243>
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE Publications.
- Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195326543.001.0001>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). The PRISMA statement. *PLoS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Normand, S. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18(3), 321–359. [https://doi.org/10.1002/\(SICI\)1097-0258\(19990215\)18:3<321::AID-SIM28>3.0.CO;2-P](https://doi.org/10.1002/(SICI)1097-0258(19990215)18:3<321::AID-SIM28>3.0.CO;2-P)
- Olkin, I. (1995). *Meta-analysis: A quantitative approach to research integration*. Routledge.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell. <https://doi.org/10.1002/9780470754887>
- Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, 342, d549. <https://doi.org/10.1136/bmj.d549>
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis*. Wiley. <https://doi.org/10.1002/0470870168>
- Sharp, S. J., & Thompson, S. G. (2000). Analysing heterogeneity in meta-analysis. *Statistics in Medicine*, 19(24), 3257–3271. [https://doi.org/10.1002/1097-0258\(20001215\)19:23<3251::AID-SIM625>3.0.CO;2-2](https://doi.org/10.1002/1097-0258(20001215)19:23<3251::AID-SIM625>3.0.CO;2-2)
- Smith, G. D., & Egger, M. (1998). Meta-analysis: Unresolved issues. *BMJ*, 316(7129), 221–225. <https://doi.org/10.1136/bmj.316.7126.221>
- Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias. *Journal of Clinical Epidemiology*, 54(10), 1046–1055. [https://doi.org/10.1016/S0895-4356\(01\)00377-8](https://doi.org/10.1016/S0895-4356(01)00377-8)
- Sterne, J. A. C., Egger, M., & Moher, D. (2008). Addressing reporting biases. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook* (pp. 297–333). Wiley. <https://doi.org/10.1002/9780470712184.ch10>
- Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. Wiley.
- Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis. *Statistics in Medicine*, 18(20), 2693–2708. [https://doi.org/10.1002/\(SICI\)1097-0258\(19991030\)18:20<2693::AID-SIM235>3.0.CO;2-V](https://doi.org/10.1002/(SICI)1097-0258(19991030)18:20<2693::AID-SIM235>3.0.CO;2-V)
- Turner, R. M., Davey, J., Clarke, M. J., Thompson, S. G., & Higgins, J. P. T. (2012). Predicting heterogeneity in meta-analysis. *Journal of Clinical Epidemiology*, 65(3), 263–273.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Yasin, S., Adeel, M., Draz, U., Ali, T., Hijji, M., Ayaz, M., & Marei, A. M. (2025). A CNN–transformer fusion model for proactive detection of schizophrenia relapse from EEG signals. *Bioengineering*, 12(6), 641. <https://doi.org/10.3390/bioengineering12060641>