ReviewPaper

Artificial Intelligence in Clinical Scalp Electroencephalogram: A Systematic Evidence Map of PubMed-indexed Studies, 2019–2026

Seyed Mohammad Reza Hadian[1], **Mahdi Saadati**[2*]

1. Department of English Language and Literature, Faculty of Humanities and Social Sciences, Golestan University, Gorgan, Iran.
2. Department of Health Information Technology, School of Allied Medical Sciences, Mazandaran University of Medical Sciences, Sari, Mazandaran, Iran.

Use your device to scan and read the article online

Citation Hadian SMR, Saadati M. Artificial Intelligence in Clinical Scalp Electroencephalogram: A Systematic Evidence Map of PubMed-indexed Studies, 2019–2026. Anatomical Sciences. 2024; 21(2):65-76.

Article info:
Received: 06 Mar 2024
Accepted: 13 Jun 2024
Available Online: 01 Jul 2024

Keywords:
Electroencephalography, Artificial intelligence, Deep learning, Machine learning, Neurologic diseases

ABSTRACT

Introduction: Artificial intelligence is increasingly being used for the analysis of clinical scalp electroencephalography data. However, current studies differ widely concerning disorders, data representations, architecture of models, validation, and reporting. In this study, we performed evidence mapping of the available literature on the shift from domain-specific deep learning to generalizable representations and clinically portable systems.

Methods: An evidence map was created based on English-language papers indexed in PubMed. The evidence map covered studies published between January 1, 2019, and August 29, 2026. Only studies that could be freely accessed through PubMed Central or other open-access PDF sources were included. Every selected study used artificial intelligence or representation-learning techniques applied to clinically relevant human scalp electroencephalography. The records were quickly checked by a single reviewer. Each identifier was verified.

Results: A total of 3,527 records were found in PubMed. Of these, 3,506 records were screened. After reviewing titles and abstracts, we retrieved 2,059 reports. Of these, 996 could not be found publicly, and 1,063 were checked for eligibility. Sixty-six were excluded. Finally, 997 publicly accessible studies were included. The most common topics were epilepsy or seizure (390 studies, 39.1%) and sleep (185 studies, 18.6%). Convolutional networks were used in 486 studies (48.7%). Recurrent models were observed in 235 studies (23.6%). Transformer or attention models were used in 176 studies (17.7%). We used subjectindependent, external, or crossdataset validation in 274 studies (27.5%). Code availability was present in 135 studies (13.5%). Explainability was present in 436 studies (43.7%).

Conclusion: In the publicly retrievable PubMed literature, clinical scalp electroencephalography research has been slower to adopt expressive architectures than to produce transferable evidence. Independent validation, leakage-free evaluation, clear reporting, resource sharing, and clinically interpretable spatiotemporal representations are crucial. This evidence map does not provide an estimate of evidence from the 996 reports that could not be publicly retrieved, and it is not meant to cover all clinical neurophysiology applications.

*** Corresponding Author:**
Mahdi Saadati, Assistant Professor.
Address: *Department of Health Information Technology, School of Allied Medical Sciences, Mazandaran University of Medical Sciences, Sari, Mazandaran, Iran.*
Tel: *+98 (911) 2461242*
E-mail: *M.Saadati@mazums.ac.ir*

Copyright © 2024 The Author(s);
This is an open access article distributed under the terms of the Creative Commons Attribution License (CC-By-NC: https://creativecommons.org/licenses/by-nc/4.0/legalcode.en), which permits use, distribution, and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Introduction

Electroencephalogram (EEG) of the scalp is a noninvasive technique to study brain dynamics in time. The usefulness of this modality in clinical practice has been shown in the fields of epilepsy investigation, sleep disorders, encephalopathies, critical care, and the characterization of various neurological and psychiatric disorders. However, the interpretation of EEG signals is difficult because of their high dimensionality, non-stationary nature, vulnerability to artifacts, and dependence on the subject, electrode montage, instrumentation, clinical context, and annotation [1, 2].

Both the diversity and shift in methods are evident in the current literature. The end-to-end neural approaches have been applied in the classification of pathological EEG recordings [3], outcome prediction after cardiac arrest [4], resilient sleep staging [5], neonatal interpretation [6], Alzheimer's disease using spatial-temporal connectivity modeling [7] and interpretable saliency analysis [8], and Parkinson's disease using spatio-oscillatory patterns [9]. Open clinical datasets have also expanded the substrate for benchmarking [10, 11]. These studies are clinically comparable but methodologically dissimilar in the following ways: The channel topology can be considered as an image, graph, sequence, or neglected; validation can be performed epoch-wise, subject-wise, site-wise, or on external data; and performance can be reported through different metrics that have clinical significance.

A second shift was also observed. Self-supervised learning, pre-training, attention models, and federated learning all seek to lessen reliance on labeled data and increase transfer between different tasks or organizations [12–17]. This shift becomes particularly important with respect to neuroanatomy. While scalp potentials do not have one-to-one correspondence with their cortical generators, electrode location, hemisphere specificity, connectivity, and spatialtemporal propagation contain information about both anatomy and function. Graph and attention models can encode these relationships, but anatomical consistency alone is not proof of clinical validity.

Existing narrative and systematic reviews established the growth of deep electroencephalography analysis [1, 2], but the rapidly expanding clinical literature requires a cross-condition map that distinguishes architectural novelty from generalization evidence. Therefore, the present systematic evidence map describes publicly retrievable PubMed-indexed literature and asks which clinical areas, architectures, representations, and learning paradigms dominate, how models are validated, how often explainability and reproducibility resources are reported, and what methodological priorities are necessary for robust clinical and neuroanatomical translation. The map does not claim to represent inaccessible reports, non-PubMed sources, or the entire engineering literature.

Materials and Methods

Design and reporting framework

We designed this study to create an evidence map that only uses reports that can be found in PubMed and are publicly available, and we followed a scoping review methodology. Only reports that could be reached by the public retrieval pathways were allowed to join the evidence collection. The study was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping (PRISMA-ScR) guidelines, JBI methodological guidelines, and the PRISMA 2020 flow diagram [18–20]. We checked the search logic against PRESS domains, looking at question translation, Boolean logic, subject headings, text words, syntax, and limits [21]. Screening report review and data extraction were performed with the help of a single reviewer using AI tools. This process has not been checked by a reviewer, which is considered a methodological weakness.

Eligibility criteria

The eligible articles comprised English-language primary research studies published between January 1, 2019, and August 29, 2026, that analyzed human scalp EEG as a primary and independently analyzable signal using modern neural, deep learning, or representation learning methods for clinically important applications in neuroscience, neurology, neurodevelopment, psychiatry, sleep, anesthesia, neonatology, or critical care.

Clinical uses that were allowed included detection, diagnosis, classification, staging, prediction, prognosis, monitoring, or characterization of the disease. Research using the general pre-trained model was allowed only when findings for at least one clinical downstream task involving scalp-electroencephalography could be distinguished. Further inclusion in the mapped systematic review required a comprehensive report that could be accessed through the predefined pathways. Any record without a publicly available comprehensive report was classified under the category of PRISMA reports not retrieved.

Data from animals, intracranial EEG without separable scalp data, non-EEG modes, non-clinical affective or cognitive decoding, brain-computer interfaces using healthy participant brains, editorial materials, review materials, non-peer-reviewed primary materials, and research involving only classical machine learning classifiers using hand-designed features were not considered. Classical machine learning outcomes were considered if used as a comparator within an otherwise eligible neural study.

Information source and search strategy

PubMed was the only bibliographic database used in this evidence map as well as the entire source universe of this synthesis (Appendix 1). This approach makes it easier to construct a reproducible evidence map with regard to biomedical and clinical evidence, instead of trying to exhaustively cover studies that are accessible through private channels, or even those dealing with the entire engineering discipline. For the last search, three concepts had to appear in the study: Scalp electroencephalography, modern neural, deep, or representation learning, and clinical diseases or applications. PubMed Medical Subject Headings and title/abstract words were used, limiting publications to between 2019 and 2026. The search cutoff date was August 29, 2026, and a snapshot of the entire result was obtained on September 2, 2026.

Record management and title/abstract screening

Metadata from PubMed were extracted and kept as a permanently unchangeable raw record. Duplicate records in the export process were identified through PubMed identification number and record-level comparison, resulting in only one master record for each PubMed identification number. The title/abstract review was performed based on a conservative and fast single-pass ruleset, according to pre-specified inclusion/exclusion codes. Documents identified as either 'include' or 'maybe' were then sent for the full text; the abstract-based inclusion was not considered final eligibility criteria.

Computational and generative AI tools were used to assist with specific steps in the evidence-mapping process. These steps included record triage, title and abstract screening, extracting study details in a way that put clinical areas and AI structures into categories, putting all the data together, and making tables and charts. AI was used to support the work but not to make decisions on its own. Records that looked like they might fit the rules were moved forward for a report check based on the rules that were set before. The author examined these factors. The author ensured that the final choices about including or excluding studies were correct. The author also took numbers, placed categories to which they belonged, added up how often things happened, and explained what the mapped information showed. Checking for matching identifiers and fixing duplicates was performed at the level of PubMed records. The final dataset and all results were checked by the author before writing the paper.

Full-text assessment and data charting

Full reports publicly available via PubMed identifier and digital object identifier validation were searched. Only PubMed central XML reports that had a substantial body of text and matched the focal PubMed report were selected. Additionally, publicly accessible open-access PDFs that could be located through OpenAlex served as complementary sources. Subscription and access restrictions were not circumvented. Reports that were not available through these means were designated as reports that were not retrieved according to PRISMA, were excluded from the mapped systematic review process, and were not subjected to an eligibility judgment. The retrieved reports underwent an eligibility evaluation and charting process. Fields abstracted included bibliographic details, clinical domain and task, study design, sample size, dataset, electroencephalogram acquisition, input representation, preprocessing, architecture, learning framework, validation, comparators, metrics, explainability, code accessibility, and use of public data.

Methodological reporting indicator

Each paper used was rated based on 10 automatically extractable reporting indicators: Sample size, population/condition, EEG dataset/acquisition, preprocessing/input representation, model architecture, validation approach, subject-independent/external validation, clinically relevant metric, comparator/baseline, and reproducibility resources. Ratings of 8-10, 5-7, and 0-4 were respectively categorized as strong, moderate, and limited reporting. This indicator is not a validated risk-of-bias assessment tool and is not meant to be interpreted as an indication of methodological quality. Clinical AI reporting guidelines, such as MI-CLAIM, informed the interpretation of transparency and external validity [22].

Synthesis and ethics

Descriptive statistics in terms of counts and percentages were provided for 997 publicly accessible documents that satisfied the scientific inclusion criteria. Frequency counts for architecture and representation are multi-labels and could go beyond the total number of studies,

while the highest level of validation and reporting levels are mutually exclusive. Meta-analysis was not performed due to the heterogeneity of populations, tasks, datasets, outcomes, and validation units. The current evidence map examined only scientific literature and did not involve human or animal testing.

Results

Study selection and corpus status

The search identified 3,527 PubMed records. Removal of 21 export-overlap duplicates left 3,506 unique records for title/abstract screening. A final bibliographic audit reassigned six initially advanced records to record-level exclusion, yielding 1,447 excluded records and 2,059 reports. Of these, 996 could not be retrieved through the documented public routes and therefore remained outside the publicly retrievable evidence map. Among 1,063 reports assessed for eligibility, 66 were excluded, and 997 were included in the mapped synthesis. All 56 records flagged by the initial pass were resolved through explicit PubMed-identifier-level adjudication, and no records remained in the adjudication queue (Figure 1).

Publication growth and clinical areas

The included publicly retrievable corpus increased from 23 studies in 2019 to 223 in 2025; 155 studies were indexed in the partial 2026 interval through August 29 (Figures 2 and 3). Epilepsy/seizure studies formed the largest clinical category (390/997, 39.1%), followed by sleep (185/997, 18.6%), psychiatric applications (139/997, 13.9%), dementia/Alzheimer disease/mild cognitive impairment (113/997, 11.3%), and other neurological conditions (72/997, 7.2%). Because clinical-area coding was multi-label, category counts were not mutually exclusive (Table 1).

Architectures, signal representations, and learning paradigms

Convolutional neural networks were identified in 486 studies (48.7%), recurrent neural network/long short-term memory/gated recurrent unit architectures in 235(23.6%), and transformer or attention models in 176 (17.7%). Graph models appeared in 49 studies (4.9%), while 24(2.4%) used a foundation or pretrained encoder label. Raw or minimally processed EEG was represented in 414 studies (41.5%); handcrafted descriptors (528/997, 53.0%) and entropy or nonlinear measures (417/997, 41.8%) remained common, often alongside neural components. Spectral (297/997, 29.8%), time-frequency (283/997, 28.4%), wavelet (255/997, 25.6%), and connectivity (156/997, 15.6%) representations demonstrate that end-to-end learning did not displace signal-engineered inputs (Figure 2 and Table 2).

Supervised learning or a paradigm not explicitly identified accounted for 820 studies (82.2%). Multimodal approaches appeared in 77(7.7%) and transfer learning in 67(6.7%). Unsupervised learning was identified in 16(1.6%), self-supervised learning in 13(1.3%), contrastive learning in 12(1.2%), domain adaptation/generalization in 12(1.2%), federated learning in eight (0.8%), and semi-supervised learning in seven (0.7%). These frequencies show an emerging representation-learning vocabulary but a corpus still dominated by task-specific supervised development.

Validation, outcomes, and transparency

Internal validation was the highest identified level in 516 studies (51.8%), and validation was unclear in 207 (20.8%). Subject-independent evaluation was found in 157 studies (15.7%) and external or cross-dataset validation in 117(11.7%); combined, 274 studies (27.5%) reported either form of independent evidence (Table 3). Accuracy was reported in 875 studies (87.8%), sensitivity/recall in 784(78.6%), precision/positive predictive value in 534(53.6%), F1 score in 485(48.6%), specificity in 429(43.0%), and area under the receiver operating characteristic curve in 372(37.3%). Calibration, decision-curve value, workflow impact, and prospective utility were not consistently extractable.

Explainability was identified in 436 studies (43.7%), public data in 470 (47.1%), and code availability in 135 (13.5%). The mean reporting-indicator score was 7.47 of 10; 541 studies (54.3%) were placed in the stronger reporting tier, 396 (39.7%) in the moderate tier, and 60 (6.0%) in the limited tier. The apparent reporting strength should be interpreted cautiously: The automatic indicator rewards the presence of reportable elements and cannot verify correctness, leakage prevention, or clinical usefulness (Table 4).

Subject-independent and external/cross-dataset validations were treated as distinct validation concepts in the evidence map. Subject-independent validation refers to evaluation in which the test subjects are separate from those used for model training or development, typically through subject-wise splitting or leave-one-subject-out procedures, while the data may still originate from the same dataset or acquisition setting. In contrast, external/cross-dataset validation refers to evaluation on an independent dataset or cohort that was not used for model

Table 1. Clinical application areas in the included publicly retrievable corpus (n=997) categories are multi-label

Clinical Area	No. (%)
Epilepsy/seizure	390(39.1)
Sleep	185(18.6)
Psychiatric	139(13.9)
Dementia/Alzheimer disease/MCI	113(11.3)
Other neurological	72(7.2)
Abnormal/clinical EEG	59(5.9)
Other clinical	58(5.8)
Neurodevelopmental	53(5.3)
Neonatal	37(3.7)
Consciousness/critical care	34(3.4)
Parkinsonian	30(3.0)
Stroke/traumatic brain injury	28(2.8)
Anesthesia/sedation	21(2.1)
Unclear	10(1.0)

EEG: Electroencephalogram; MCI: Mild cognitive impairment.
Percentages exceed 100% because categories are multi-label.
ANATOMICAL SCIENCES

development. Therefore, a study could be classified as subject-independent without constituting external validation, whereas external/cross-dataset validation represents a stronger assessment of generalizability across datasets or cohorts.

Discussion

Principal findings

In the published scientific literature available on PubMed, an evidence map shows a topic that is vast, fast-growing, and inconsistent from a methodological perspective. Electroencephalography artificial intelligence in clinical practice is predominantly associated with epilepsy and sleep disorders, whereas research on other conditions, including psychiatry, neurodegenerative, neonatal, critical care, and peri-anesthesia indications, is on the rise. Convolutional and recurrent neural networks are prevalent, while attention, graph, pre-trained, federated, and self-supervised learning models represent another, albeit smaller but theoretically relevant front of research. The key result lies in translation: Architectural utilization is wide, but only 27.5% of papers have cross-dataset validation.

From engineered features to learned representations

The results do not support a simple historical narrative in which deep learning replaced signal engineering. Raw-input systems coexist with spectral, time-frequency, wavelet, entropy, nonlinear, connectivity, and topographic representations. This coexistence can be productive. Carefully chosen representations may improve data efficiency and interpretability, while learned encoders can reduce dependence on brittle feature sets. Comparative pathology studies [3], robust sleep models [5], and open clinical datasets [10, 11] demonstrate why benchmarking across representations is more informative than reporting a single optimized pipeline.

Pretraining and self-supervision offer a possible route beyond task-specific modeling. Banville et al. used self-supervised objectives to uncover structures in clinical EEG [12]. Later work applied self-supervised representation learning to sleep staging [13], and attention-based pretraining has reached seizure and consciousness applications [14, 15]. Foundation-model proposals have articulated a broader agenda for reusable EEG encoders [17]. However, the evidence map found explicit self-

Table 2. Model architectures and EEG input representations (n=997)

Domain	Category	No. (%)
Architecture	CNN	486(48.7)
Architecture	RNN/LSTM/GRU	235(23.6)
Architecture	Neural family unclear	214(21.5)
Architecture	Transformer/attention	176(17.7)
Architecture	MLP/DNN/ANN	117(11.7)
Architecture	GNN/ GCN/ GAT	49(4.9)
Architecture	Autoencoder/VAE	36(3.6)
Architecture	Foundation/pretrained encoder	24(2.4)
Architecture	Generative/GAN /diffusion	24(2.4)
Architecture	Spiking neural network	12(1.2)
Architecture	State-space/Mamba	5(0.5)
Representation	Handcrafted descriptors	528(53.0)
Representation	Entropy/nonlinear	417(41.8)
Representation	Raw/minimally processed EEG	414(41.5)
Representation	Spectral	297(29.8)
Representation	Time-frequency	283(28.4)
Representation	Wavelet	255(25.6)
Representation	Connectivity	156(15.6)
Representation	Unclear	133(13.3)
Representation	Topographic map/image	16(1.6)

ANATOMICAL SCIENCES

Abbreviations: ANN: Artificial neural network; CNN: Convolutional neural network; DNN: Deep neural network; GAN: Generative adversarial network; GAT: Graph attention network; GCN: Graph convolutional network; GNN: Graph neural network; GRU: Gated recurrent unit; LSTM: Long short-term memory; MLP: Multilayer perceptron; RNN: Recurrent neural network; VAE: Variational autoencoder; EEG: Electroencephalogram.
All categories are multi-label.

supervision in 1.3% of the included studies and foundation/pretrained encoders in 2.4%. The methodological direction is visible, but it should not be mistaken for current clinical maturity.

Neuroanatomical relevance

Clinical scalp EEG is a functional signal influenced by cortical generators, volume conduction, reference selection, electrode configuration, and network-level propagation. Spatial representations that preserve channel specificity, hemispheric relationships, topography, or functional connectivity may provide useful information regarding broad cortical activity patterns. Spatial-temporal graph modeling in Alzheimer disease [7] and spatio-oscillatory convolutional analysis in Parkinson disease [9] illustrate this potential. However, these representations do not constitute direct or definitive measures of activity from specific cortical regions, and learned saliency maps, graph weights, or attention maps should not be interpreted as precise anatomical localization without appropriate source-localization or complementary neuroimaging validation. Therefore, interpretability methods should distinguish predictive associations from underlying physiological mechanisms.

Table 3. Learning paradigms and highest validation level in the included publicly retrievable corpus (n=997)

Domain	Category	No. (%)
Learning paradigm*	Supervised/not explicitly stated	820(82.2)
Learning paradigm*	Multimodal	77(7.7)
Learning paradigm*	Transfer learning	67(6.7)
Learning paradigm*	Unsupervised	16(1.6)
Learning paradigm*	Self-supervised	13(1.3)
Learning paradigm*	Contrastive	12(1.2)
Learning paradigm*	Domain adaptation/generalization	12(1.2)
Learning paradigm*	Federated	8(0.8)
Learning paradigm*	Semi-supervised	7(0.7)
Highest validation†	Internal validation	516(51.8)
Highest validation†	Unclear	207(20.8)
Highest validation†	Subject-independent	157(15.7)
Highest validation†	External/cross-dataset	117(11.7)

ANATOMICAL SCIENCES

*Multi-label; †Mutually exclusive categories. Combined subject-independent or external/cross-dataset evidence: 274/997 (27.5%).

Table 4. Outcomes, transparency, and automated reporting indicators (n=997)

Domain	Indicator	No. (%)
Reported metric*	Accuracy	875(87.8)
Reported metric*	Sensitivity/recall	784(78.6)
Reported metric*	Precision/PPV	534(53.6)
Reported metric*	F1 score	485(48.6)
Reported metric*	Specificity	429(43.0)
Reported metric*	AUROC	372(37.3)
Transparency*	Explainability detected	436(43.7)
Transparency*	Public data used	470(47.1)
Transparency*	Code availability detected	135(13.5)
Reporting tier†	Stronger (8–10)	541(54.3)
Reporting tier†	Moderate (5–7)	396(39.7)
Reporting tier†	Limited (0–4)	60(6.0)

AUROC: Area under the receiver operating characteristic curve; PPV: Positive predictive value.

ANATOMICAL SCIENCES

*Multi-label; †Mutually exclusive. Mean reporting-indicator score 7.47/10.

The indicator is descriptive and is not a validated risk-of-bias assessment.

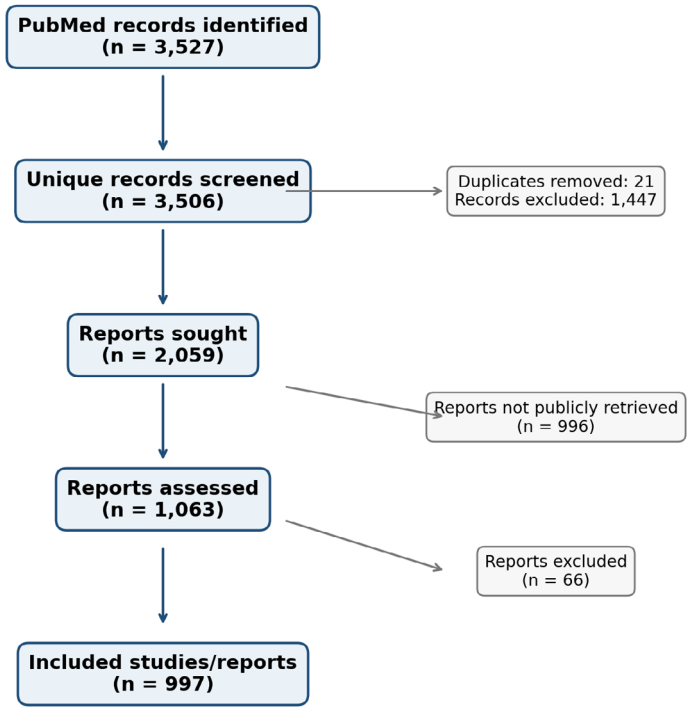


Figure 1. Study flow for the systematic evidence map of publicly retrievable PubMed literature
Reports not retrieved are retained as an explicit PRISMA access outcome, are outside the mapped synthesis, and are not counted as eligibility exclusions.

Generalization is the principal bottleneck

Random division at an epoch level may produce optimistic results if data from one person are used in both training and testing datasets. Subject-independence splitting is a minimum standard for generalization to new subjects, while site-independence or cross-dataset evaluation is needed for transportability to different data collection modalities, populations, and medical procedures. The 11.7% cross-dataset/external splitting rate suggests that almost all reported performance numbers are still locally dependent. Future publications need to clarify the unit of splitting, avoid preprocessing across the train-test boundary, explain hyperparameter tuning separately from testing, and quantify uncertainty in clinically meaningful metrics.

Clinical evaluation and reproducibility

The sole consideration of accuracy is not adequate when the prevalence is unbalanced, or the impact of errors is unequal. The use of sensitivity, specificity, predictive value, area under the receiver operating characteristic curve, precision-recall curve, calibration, and utility based on threshold criteria will be necessary according to the clinical application in which it will be put into practice. The comparison of results to suitable baseline data and, where feasible, to clinician performance is crucial.

Public data were identified in 47.1% of included studies, but only 13.5% had detectable code availability. Public benchmarks can accelerate progress; however, repeated optimization on a small set of datasets can create hidden test-set overfitting. Dataset documentation, immutable test sets, external replication, executable preprocessing, trained weights, and complete reporting of channel mappings are necessary. Federated approaches may improve multi-site collaboration without centralizing raw EEG, as illustrated in dementia classification [16]; however, privacy preservation does not remove the need for harmonization, auditing, and external validation.

Implications for research

The following five points emerge from the data presented above. First, validation must be based on the

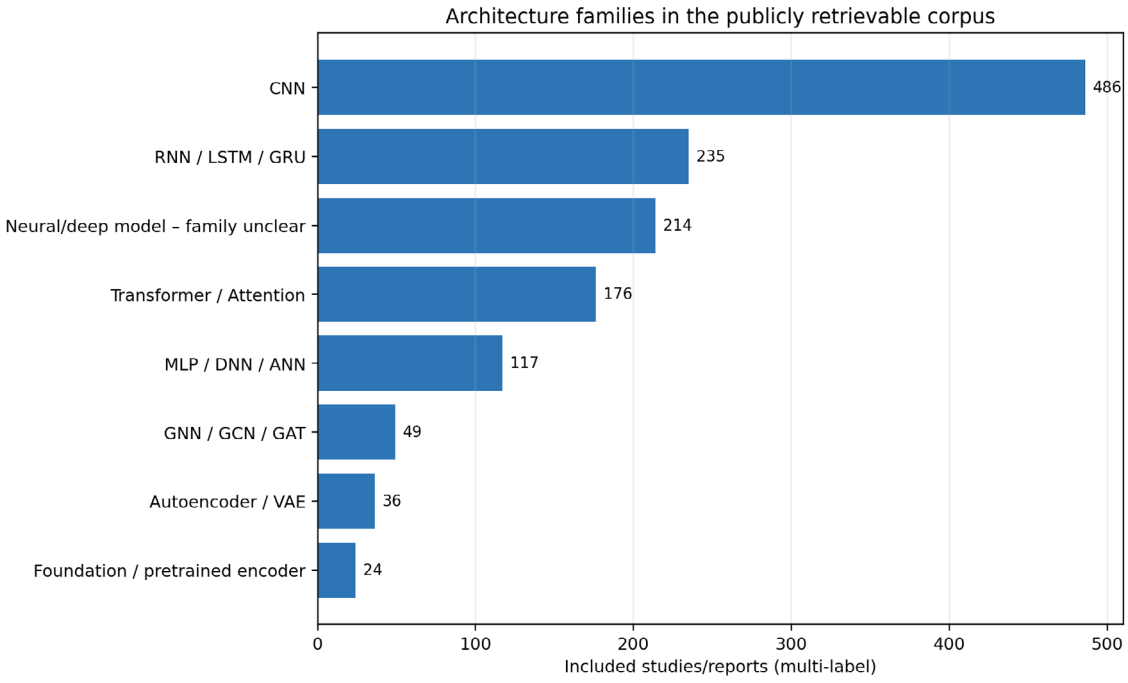


Figure 2. Distribution of major architecture families in the included publicly retrievable corpus
ANATOMICAL SCIENCES
Categories are multi-label; therefore, counts can exceed the number of included reports.

desired clinical generalization unit: Patient, session, device, hospital, or population. Second, the spatial models need to describe what they assume regarding anatomy or connectivity. Third, the foundation and self-supervised models need to be tested not only against strong baselines but also on several clinical datasets. Fourth, calibration, performance on subgroups, measures of uncertainty, and clinically relevant operating points must be reported. Fifth, code, preprocessing, and model artifacts must be shared when ethical and legal considerations are allowed.

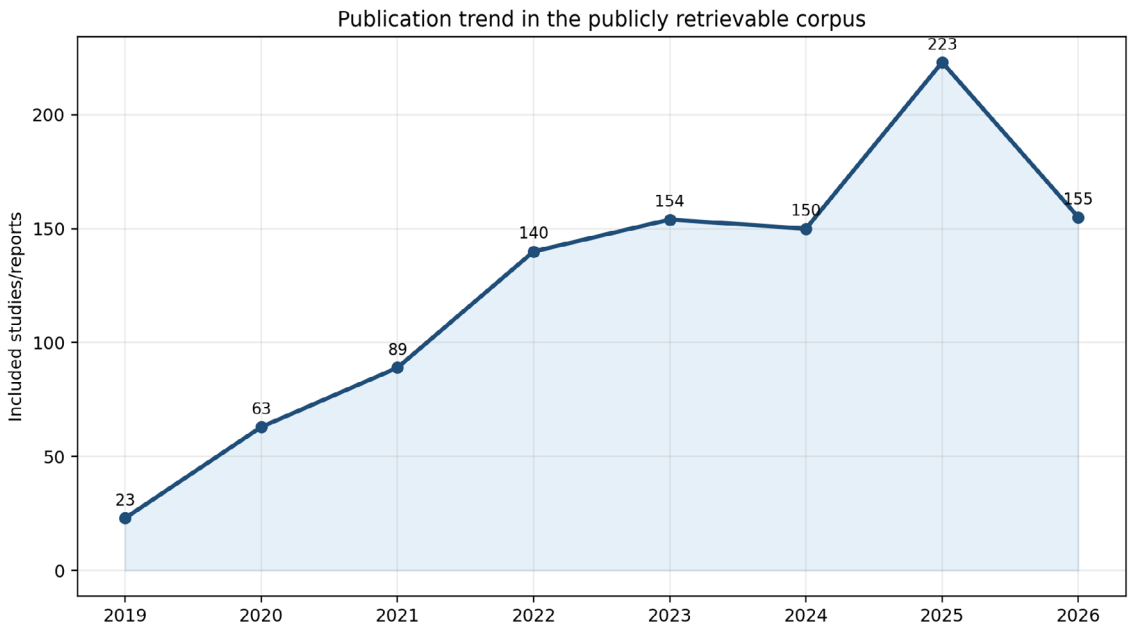


Figure 3. Publication-year distribution of the included publicly retrievable corpus
ANATOMICAL SCIENCES
The 2026 count covers publications through 29 August and should not be compared as a complete calendar year.

Conclusion

In the publicly retrievable PubMed literature, artificial intelligence for clinical scalp electroencephalography expanded rapidly from 2019 through 2026, with convolutional, recurrent, and increasingly attention- or graph-based systems applied across epilepsy, sleep, psychiatric, neurodegenerative, neonatal, and critical-care contexts. The field is beginning to explore pretrained and self-supervised representations, but task-specific supervised learning remains dominant. The principal weakness is not a shortage of architectures; it is limited evidence that performance survives new patients, sites, devices, and datasets. Progress toward clinically credible and neuroanatomically interpretable artificial intelligence will require leakage-resistant independent validation, transparent spatial assumptions, clinically appropriate outcomes, and reproducible resources. The 997-report evidence corpus is final for the prespecified public-retrieval workflow; it does not characterize the 996 reports that were not publicly retrieved and should not be generalized to all clinical electroencephalography artificial-intelligence literature.

Future clinical translation should focus heavily on external validation, reproducibility, and tests across independent patient groups, clinical sites, and EEG devices. Future clinical translation will be secure and broadly useful if this method works everywhere.

Strengths and limitations

Strengths include an explicit and reproducible public-retrieval scope, a frozen PubMed query, preservation of raw records, explicit eligibility codes, a large cross-condition evidence map, distinction between internal and independent validation, and transparent labeling of automated extraction. The review maps architecture, representation, clinical application, validation, transparency, and reporting within one framework, while separating reports assessed from reports that were not publicly retrievable.

Limitations abound. This is an evidence map of PubMed-published literature, not a comprehensive review of all clinical electroencephalography artificial intelligence research. The use of PubMed only understates the literature at engineering conferences, and the required clinical search filter may overlook general artificial intelligence models whose abstracts do not mention a disease. The 996 studies that could not be found via any publicly documented route may vary systematically from the 997 that were found. The scientific eligibility and results of these are unknown, introducing access-related selection bias.

Ethical Considerations

Compliance with ethical guidelines

Ethics approval and informed consent were not required because this review used published literature and involved no new human or animal data collection.

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

All authors contributed equally to the conception and design of the study, data collection and analysis, interpretation of the results, and drafting of the manuscript. Each author approved the final version of the manuscript for submission.

Conflict of interest

The authors declared no conflict of interest.

Acknowledgments

The authors gratefully acknowledge the Artificial Intelligence Center of Mazandaran University of Medical Sciences for its support of this study.

References

[1] Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: A review. Journal of Neural Engineering. 2019; 16(3):031001. [DOI:10.1088/1741-2552/ab0ab5] [PMID]

[2] Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J. Deep learning-based electroencephalography analysis: A systematic review. Journal of Neural Engineering. 2019; 16(5):051001. [DOI:10.1088/1741-2552/ab260c] [PMID]

[3] Gemein LAW, Schirrmeister RT, Chrabąszcz P, Wilson D, Boedecker J, Schulze-Bonhage A, et al. Machine-learning-based diagnostics of EEG pathology. Neuroimage. 2020; 220:117021. [DOI:10.1016/j.neuroimage.2020.117021] [PMID]

[4] Jonas S, Rossetti AO, Oddo M, Jenni S, Favaro P, Zubler F. EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. Human Brain Mapping. 2019; 40(16):4606-17. [DOI:10.1002/hbm.24724] [PMID]

[5] Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: Resilient high-frequency sleep staging. NPJ Digital Medicine. 2021; 4(1):72. [DOI:10.1038/s41746-021-00440-5] [PMID]

[6] Gomez-Quintana S, O'Shea A, Factor A, Popovici E, Temko A. A method for AI assisted human interpretation of neonatal EEG. Scientific Reports. 2022; 12(1):10932. [DOI:10.1038/s41598-022-14894-4] [PMID]

[7] Shan X, Cao J, Huo S, Chen L, Sarrigiannis PG, Zhao Y. Spatial-temporal graph convolutional network for Alzheimer classification based on brain functional connectivity imaging of electroencephalogram. Human Brain Mapping. 2022; 43(17):5194-209. [DOI:10.1002/hbm.25994] [PMID]

[8] Lopes M, Cassani R, Falk TH. Using CNN saliency maps and EEG modulation spectra for improved and more interpretable machine learning-based Alzheimer's disease diagnosis. Computational Intelligence and Neuroscience. 2023; 2023:3198066. [DOI:10.1155/2023/3198066] [PMID]

[9] Shabanpour M, Kaboodvand N, Iravani B. Parkinson's disease is characterized by sub-second resting-state spatio-oscillatory patterns: A contribution from deep convolutional neural network. NeuroImage. Clinical. 2022; 36:103266. [DOI:10.1016/j.nicl.2022.103266] [PMID]

[10] Khan HA, Ul Ain R, Kamboh AM, Butt HT, Shafait S, Alamgir W, et al. The NMT scalp EEG dataset: an open-source annotated dataset of healthy and pathological EEG recordings for predictive modeling. Frontiers in Neuroscience. 2022; 15:755817. [DOI:10.3389/fnins.2021.755817] [PMID]

[11] Kiessner AK, Schirrmeister RT, Gemein LAW, Boedecker J, Ball T. An extended clinical EEG dataset with 15,300 automatically labelled recordings for pathology decoding. NeuroImage. Clinical. 2023; 39:103482. [DOI:10.1016/j.nicl.2023.103482] [PMID]

[12] Banville H, Chehab O, Hyvärinen A, Engemann DA, Gramfort A. Uncovering the structure of clinical EEG signals with self-supervised learning. Journal of Neural Engineering. 2021; 18(4):046020. [DOI:10.1088/1741-2552/abca18] [PMID]

[13] Yang C, Xiao C, Westover MB, Sun J. Self-supervised electroencephalogram representation learning for automatic sleep staging: Model development and evaluation study. JMIR AI. 2023; 2(1):e46769. [DOI:10.2196/46769] [PMID]

[14] Huang Y, Chen Y, Xu S, Wu D, Wu X. Self-supervised learning with adaptive frequency-time attention Transformer for seizure prediction and classification. Brain Sciences. 2025; 15(4):382. [DOI:10.3390/brainsci15040382] [PMID]

[15] Wang Z, Yu J, Gao J, Bai Y, Wan Z. MutaPT: A multi-task pre-trained Transformer for classifying state of disorders of consciousness using EEG signal. Brain Sciences. 2024; 14(7):688. [DOI:10.3390/brainsci14070688] [PMID]

[16] Umair M, Khan MS, Hanif M, Ghaban W, Nafea I, Qasem SN, et al. Privacy-preserving dementia classification from EEG via hybrid-fusion EEGNetv4 and federated learning. Frontiers in Computational Neuroscience. 2025; 19:1617883. [DOI:10.3389/fncom.2025.1617883] [PMID]

[17] Cui S, Lee D, Wen D. Toward brain-inspired foundation model for EEG signal processing: our opinion. Frontiers in Neuroscience. 2024; 18:1507654. [DOI:10.3389/fnins.2024.1507654] [PMID

[18] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of Internal Medicine. 2018; 169(7):467-73. [DOI:10.7326/M18-0850] [PMID]

[19] Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evidence Synthesis. 2020; 18(10):2119-26. [DOI:10.11124/JBIES-20-00167] [PMID]

[20] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021; 372:n71. [DOI:10.1136/bmj.n71] [PMID]

[21] McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. Journal of Clinical Epidemiology. 2016; 75:40-6. [DOI:10.1016/j.jclinepi.2016.01.021] [PMID]

[22] Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nature Medicine. 2020; 26(9):1320-4. [DOI:10.1038/s41591-020-1041-y] [PMID]

Appendix A. Exact PubMed Search Strategy

Search period: 1 January 2019–29 August 2026 · Snapshot: 2 September 2026 · Database: PubMed

("Electroencephalography"[Mesh] OR EEG [tiab] OR electroencephalogra*[tiab] OR "scalp EEG"[tiab]) AND ("Deep Learning"[Mesh] OR "Neural Networks, Computer"[Mesh] OR "deep learning"[tiab] OR "deep neural network*"[tiab] OR "artificial neural network*"[tiab] OR "neural network*"[tiab] OR DNN [tiab] OR "convolutional neural network*"[tiab] OR "convolutional network*"[tiab] OR CNN [tiab] OR "recurrent neural network*"[tiab] OR RNN [tiab] OR "long short-term memory"[tiab] OR LSTM [tiab] OR "gated recurrent unit*"[tiab] OR GRU [tiab] OR "temporal convolutional network*"[tiab] OR TCN [tiab] OR transformer*[tiab] OR "attention mechanism*"[tiab] OR "attention-based"[tiab] OR "self-attention"[tiab] OR "multi-head attention"[tiab] OR "graph neural network*"[tiab] OR "graph convolutional network*"[tiab] OR GNN [tiab] OR GCN [tiab] OR autoencoder*[tiab] OR "variational autoencoder*"[tiab] OR "representation learning"[tiab] OR "representation-learning"[tiab] OR "self-supervised learning"[tiab] OR "self supervised learning"[tiab] OR "contrastive learning"[tiab] OR "semi-supervised learning"[tiab] OR "transfer learning"[tiab] OR "domain adaptation"[tiab] OR "domain generalization"[tiab] OR pretrain*[tiab] OR "pre-trained model*"[tiab] OR "pretrained model*"[tiab] OR "foundation model*"[tiab] OR "vision-language model*"[tiab] OR VLM [tiab] OR "large language model*"[tiab] OR LLM [tiab] OR mamba [tiab] OR "generative model*"[tiab] OR "generative adversarial network*"[tiab] OR "diffusion model*"[tiab] OR "federated learning"[tiab] OR "spiking neural network*"[tiab] OR "extreme learning machine*"[tiab] OR "multilayer perceptron*"[tiab] OR "multi-layer perceptron*"[tiab] OR MLP [tiab] OR "Kolmogorov-Arnold network*"[tiab] OR KAN [tiab]) AND ("Epilepsy"[Mesh] OR "Seizures"[Mesh] OR "Sleep Wake Disorders"[Mesh] OR "Sleep Stages"[Mesh] OR "Dementia"[Mesh] OR "Alzheimer Disease"[Mesh] OR "Parkinson Disease"[Mesh] OR "Stroke"[Mesh] OR "Brain Injuries, Traumatic"[Mesh] OR diagnos*[tiab] OR "clinical EEG"[tiab] OR "routine EEG"[tiab] OR "EEG interpretation"[tiab] OR "electroencephalogram interpretation"[tiab] OR "EEG report*"[tiab] OR "electroencephalogram report*"[tiab] OR "EEG pathology"[tiab] OR "pathological EEG"[tiab] OR "abnormal EEG"[tiab] OR epilep*[tiab] OR seizure*[tiab] OR ictal [tiab] OR interictal [tiab] OR epileptiform [tiab] OR "EEG spike*"[tiab] OR IED [tiab] OR sleep*[tiab] OR polysomnograph*[tiab] OR insomnia*[tiab] OR narcolep*[tiab] OR apnea [tiab] OR apnoea[tiab] OR dementia [tiab] OR Alzheimer*[tiab] OR Parkinson*[tiab] OR "mild cognitive impairment"[tiab] OR MCI [tiab] OR "cognitive impairment"[tiab] OR neurodegenerat*[tiab] OR stroke [tiab] OR concussion*[tiab] OR "traumatic brain injur*"[tiab] OR TBI [tiab] OR encephalopath*[tiab] OR coma [tiab] OR "disorder* of consciousness"[tiab] OR autism [tiab] OR autistic [tiab] OR ADHD [tiab] OR "attention deficit hyperactivity"[tiab] OR dyslex*[tiab] OR schizophren*[tiab] OR depress*[tiab] OR anxiety [tiab] OR "generalized anxiety disorder"[tiab] OR PTSD [tiab] OR "posttraumatic stress"[tiab] OR "posttraumatic stress"[tiab] OR bipolar [tiab] OR migraine*[tiab] OR "multiple sclerosis"[tiab] OR ICU [tiab] OR "intensive care"[tiab] OR anesthe*[tiab] OR anaesthe*[tiab] OR sedation[tiab] OR delirium [tiab]) AND ("2019/01/01"[Date - Publication]: "2026/08/29"[Date - Publication])