mathematics

MDPI

Review

Towards Explainable Deep Learning in Computational Neuroscience: Visual and Clinical Applications

Asif Mehmood [1], Faisal Mehmood [2,*] and Jungsuk Kim [1,*]

[1] Department of Biomedical Engineering, College of IT Convergence, Gachon University, Sujeong-gu, Seongnam-si 13120, Republic of Korea; asif@gachon.ac.kr
[2] Department of AI and Software, College of IT Convergence, Gachon University, Sujeong-gu, Seongnam-si 13120, Republic of Korea
* Correspondence: faisal89@gachon.ac.kr (F.M.); jungsuk@gachon.ac.kr (J.K.)

Abstract

Deep learning has emerged as a powerful tool in computational neuroscience, enabling the modeling of complex neural processes and supporting data-driven insights into brain function. However, the non-transparent nature of many deep learning models limits their interpretability, which is a significant barrier in neuroscience and clinical contexts where trust, transparency, and biological plausibility are essential. This review surveys structured explainable deep learning methods, such as saliency maps, attention mechanisms, and model-agnostic interpretability frameworks, that bridge the gap between performance and interpretability. We then explore explainable deep learning's role in visual neuroscience and clinical neuroscience. By surveying literature and evaluating strengths and limitations, we highlight explainable models' contribution to both scientific understanding and ethical deployment. Challenges such as balancing accuracy, complexity and interpretability, absence of standardized metrics, and scalability are assessed. Finally, we propose future directions, which include integrating biological priors, implementing standardized benchmarks, and incorporating human-intervention systems. The research study highlights the position of explainable deep learning, not only as a technical advancement but represents it as a necessary paradigm for transparent, responsible, auditable, and effective computational neuroscience. In total, 177 studies were reviewed as per PRISMA, which provided evidence across both visual and clinical computational neuroscience domains.

Keywords: explainable deep learning; computational neuroscience; visual attention; clinical applications; neural networks

MSC: 92B20; 68T07

Academic Editor: Giovanny Sánchez

Received: 3 September 2025
Revised: 5 October 2025
Accepted: 11 October 2025
Published: 14 October 2025

Citation: Mehmood, A.; Mehmood, F.; Kim, J. Towards Explainable Deep Learning in Computational Neuroscience: Visual and Clinical Applications. *Mathematics* **2025**, *13*, 3286. https://doi.org/10.3390/math13203286

1. Introduction

Computational neuroscience is the study in which understanding brain function through analysis of neural processes via mathematical models is involved [1]. Neuroscience used to be rooted in theory and biophysics only, and recently it has increasingly embraced data-driven approaches with the rise of advanced technology-based neural recordings and large-scale neuroimaging [2,3]. This impacted neuroscience by enabling the adoption of convergence with DL, which is capable in uncovering patterns of complex datasets. This offers advancement and efficiency in modeling perception, cognition, and behavior.

In the neuroscience context, the DL models have shown remarkable success in replicating neural responses and have also enabled the domain in decoding brain states, especially

in visual neuroscience. It has matured in tasks like attention modeling [4,5], object recognition [6], and categorization mirroring [7] challenges faced in computer vision. Clinical neuroscience has also benefited from these DL models in neuroimaging-based diagnosis, disease progress forecasting, and treatment personalizing areas. However, these advances have major limitations because most DL models are uninterpretable systems. This shows that state-of-the-art research for neuroscience is not transparent enough in showcasing the model's internal workings [8].

This lack of interpretability poses serious challenges. In neuroscience, opaque models limit our ability to derive mechanistic or causal insight [9], making it difficult to validate findings against known biology or generate new hypotheses. In clinical settings, the risks are even greater: high-stakes decisions demand transparency, robustness, and trustworthiness, which the non-transparent models often lack. Key limitations include predictions without mechanistic understanding, overfitting to data quirks, limited support for causal inference, and vulnerability to spurious correlations or adversarial inputs.

These concerns are magnified by current trends: the rapid adoption of AI in medical diagnosis, rising demand for trustworthy AI systems [10], and an emerging regulatory emphasis on explainability [11] in healthcare technologies. Simultaneously, there's a growing awareness within neuroscience that predictive success is not enough, interpretability is now central to scientific credibility and ethical deployment. This review is motivated by the need to explore how XDL can bridge the gap between performance and understanding, particularly in the domains of visual neuroscience [12,13] and clinical applications [14–16], where both scientific insight and practical utility are critical.

The integration of DL into neuroscience has revolutionized the analysis and interpretation of complex neural data. Among various architectures, transformers have emerged as particularly influential due to their self-attention mechanisms, allowing models to capture intricate dependencies within data [17]. However, their application in neuroscience necessitates advancements in interpretability to elucidate underlying neural processes effectively. Key Applications of DL in Neuroscience:

1. Brain signal decoding: The DL models like RNNs and transformers have shown significant performance in decoding signals, including cognitive states, motor intentions, and sensory inputs [18], e.g., EEG, MEG, ECoG, and fMRI.
2. Common tasks: Tasks include perceiving, classifying, regression, brain data decoding [19], multimodal data fusion [20], and functional connectivity analysis [21] in neural signals.
3. Image segmentation: The CNNs have always played an important role in segmenting high-dimensional images that depict the neural structures. This facilitates tasks used for connectomics, i.e., mapping complicated networks of synapses and neurons [22]. This also facilitates brain atlasing, i.e., understanding the structure and function of the brain [23].
4. Disease classification & biomarker discovery: DL techniques have enhanced the diagnosis of neurological and psychiatric conditions by analyzing neuroimaging and electrophysiological data [24].
5. Neural encoding & predictive modeling: Models like CNNs and transformers are being utilized to predict the brain responses to the expected environmental stimuli. Interpreting the behavior of response and stimuli through the use of neural encoding and predictions helps in understanding and advancing the sensory processing that mimics the human brain.
6. Brain-computer interfaces: The DL approaches have proven to show improved real-time performance in communication-based feedback systems [25], rehabilitation [26],

and prosthetic control [27]. These approaches play a vital role in enhancing the BCI's ability to produce the desired result.

7. Generative modeling of brain activity: Generative models, such as GANs and VAEs, are also known to be vital. The key applications are synthesizing the neural data, reconstructing both images and speech in neuroscience [28].

Despite these advancements, challenges persist that contribute to the interpretability limitations of DL in neuroscience, which are shown in Table 1. Likely solutions are also proposed to overcome the limitations. Issues like robustness exist due to the diverse set of datasets that can easily lead the DL models to show inconsistent performance. This issue arises due to the noise and biases specific to training data. Challenges like generalizability also exist, as the models cannot always extend their findings to a broader context of the population, which needs to be addressed. Moreover, complex models that are based on transformer architecture need to be interpretable [29]. It is important to ensure that the insights derived from neuroscientific and clinical environments are meaningful and trustworthy.

Table 1. Factors contributing to the interpretability limitations of DL in neuroscience.

Factor	Description
Uninterpretable architecture	Problem: Deep models (e.g., CNNs, transformers) consist of many non-linear layers, making it difficult to trace how inputs are transformed into outputs [10,11]. Possible solution: Leveraging XDL can make the learning process more transparent. A possible solution could be to include embedding explainable attention layers into the model.
Lack of biological grounding	Problem: Most DL architectures are not inherently constrained by known neurobiological mechanisms, limiting their plausibility in neuroscience [30]. Possible solution: Tailoring the modeling phase utilizing the brain structures or pathways could guide the model toward more plausible interpretations.
High-dimensional feature spaces	Problem: DL models often operate in complex and abstract feature spaces that are difficult to interpret intuitively. This requires significant efforts to align such models with the known cognitive functions [31]. Possible solution: Concept attribution techniques allows the linkage of abstract model features to more meaningful brain functions or behaviors.
Overparameterization	Problem: Deep networks tend to have millions of parameters, increasing model complexity and obscuring the contribution of individual features or layers [32]. Possible solution: Tools like LRP enable model analysis to identify which parts actually matter the most, providing the opportunity to simplify the network.
Non-transparent decision boundaries	Problem: Unlike rule-based or linear models, DL models do not offer clear decision logic, making it hard to validate predictions or diagnoses [8]. Possible solution: Applying concepts like rule-based models can act like interpretable stand-ins as they depict the logic behind outcomes.
Limited explainability tools integration	Problem: Explainability techniques (e.g., saliency maps) are often added post hoc, and may not be tightly integrated into the model training process [33]. Possible solution: Including explainability in the training process can produce reliable and aligned results.
Inadequate evaluation metrics	Problem: There is a lack of standardized benchmarks to quantitatively assess interpretability, making it difficult to compare or trust explanations [34]. Possible solution: Standards to evaluate the training performance must be established, as they will reshape the comparison criteria to be fair. It will also impose on the models to be responsible for how well various models are explainable in their outcomes.

The limitations relevant to interpretability are shown in Table 1. These demand measures so that the neuroscience-based visual and clinical system DL models could be enhanced in terms of application scope and reliability. The primary requirement to address these limitations is to develop standard evaluation metrics.

With the DL integration increasing in the domain of computational neuroscience and clinical workflows, its explainability demand has grown into a critical concern scientifically, ethically, and practically. There is no doubt in the remarkable performance of DL for predictions. However, their opaque nature acts as a hurdle, which further leads the models to not be able to provide meaningful insights. This concern is considered critical in neural computation and clinical reasoning [8], driving the technological advancement shift towards the necessity of XDL. This shift not only comes with an ability to trade off between accuracy, complexity, and interpretability, but also comes as a foundational requirement for scientific explainability.

It is a known fact through science that explainability enables the formation of hypotheses, validation of biology, and understanding of core values. Enabling these would not have been possible with the DL alike non-transparent models. While considering a clinical aspect, explainability provides diagnostic outcomes that are reliable and standards that are compliant and auditable. Overall, explainability brings transparency and accountability to the clinical systems, which is a standardized technique. The DL-based models without explainability may produce critical errors, which could further lead to severe consequences in the real world. These consequences could be incorrect diagnosis, failure to interpret uninterpretable models' generalizability, misinterpretation of analysis, and lack of trust from scientific and medical communities.

There are several explainability methodologies, each emerging with a specific impact in the computational neuroscientific domain. As long as the brain data is presumed at a higher level of understanding, there are two most important regions in it. These regions are spatial and temporal [35], which can be analyzed through the saliency maps and activation localization techniques. All the above-mentioned techniques enable transparency, i.e., allow tracing the model decisions to specific neural signatures. Other explainable methods exist, which are concept-based [36], causal-based [9], and symbolic-based [37]. Each method has a particular impact in the space. The explainability involving mapping of internal representative explanations to human understandable abstracts, are termed as concept-based explanations. These bridge the model outputs with cognitive or clinical concepts. The second type of explainability method is causal, which emphasizes deriving the operational transparency of the model from intervention-based strategies. These techniques primarily focus on digging out the model's underlying operational mechanisms. While the third type of explainability method, symbolic-based explainability, aims at improving the audibility. This is done by translating complicated models into simpler ones by building those models' logical constructs. The fourth method is feature attribution, which has proven to be useful when considering the electrophysiological and time-series data [38].

Overall, the main purpose of borrowing the interpretability through XDL in visual and clinical applications is not just about explainability. But for building such a foundation that can lead and support the enhancement of scientific AI-based systems in neuroscience in terms of interpretability and trustworthiness. This shift advises a proactive transformation necessary in DL as it is of vital importance. Because of its contribution in understanding and providing neuroscientific computational systems that enable advancement in their clinical and visual applications, e.g., predictive analytics [39].

This study explores the XDL's impactful role in computational neuroscience, which focuses on both visual and clinical domains. Beyond improving predictions, explainability is framed as essential for trust, insight, and real-world relevance.

Section 2 provides a summary of the methodology followed to conduct this study. Section 3 (Background and Foundations) sets the stage by outlining the foundational concepts in DL and computational neuroscience. It introduces the role of explainable models and reviews key XDL techniques used to interpret brain-related data. As the knowledge of computational neuroscience for visual and clinical systems differs significantly in terms of data, objective, and applications, two separate sections are tailored to explain the visual modeling and clinical decision-making. Where Section 4 shifts focus to the visual system, exploring how XDL helps unravel neural mechanisms behind attention, categorization, and perception. It discusses both the architectures employed and the interpretability tools that bring insights into visual processing. Section 5 examines clinical applications of XDL, particularly in areas like diagnosis, outcome prediction, and clinical decision support. The emphasis is on how explainability fosters trust, supports ethical deployment, and enhances clinical usability. Section 6 (Discussion and Future Directions) brings together ongoing challenges, such as balancing accuracy and interpretability, and outlines future research priorities. These include integrating biological knowledge, developing robust evaluation benchmarks, and advancing interdisciplinary collaboration.

Finally, Section 7 concludes the review, reinforcing the importance of XDL as a pathway to more interpretable, reliable, and scientifically valuable AI applications in neuroscience.

A high-level diagram illustrating the flow from neuroscience data acquisition, e.g., EEG, fMRI, EM images, to model training using DL (e.g., CNNs, transformers), followed by explainability layers (e.g., saliency maps, concept attribution, and causal inference) that support both visual system modeling and clinical interpretation is shown in Figure 1. It illustrates the end-to-end pipeline from data acquisition (e.g., neuroimaging, electrophysiology), through DL model training, to layers of explainability that yield insight, support validation, and guide interpretation in visual or clinical contexts. This framework visually reinforces the structure of the review and clarifies how XDL sits at the intersection of data, modeling, and scientific discovery.

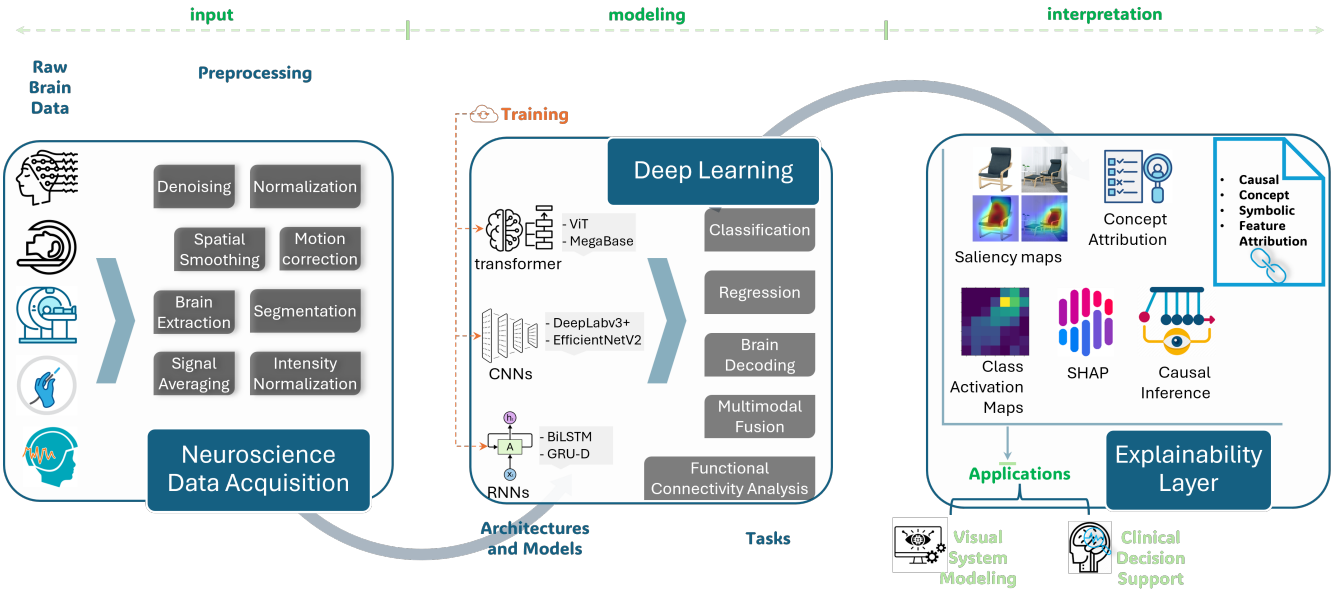


Figure 1. Conceptual overview of XDL in computational neuroscience workflows.

This review is guided by a set of research questions. This set of questions is categorized on the basis of relevance. The research questions are as follows.

- RQ1: How do data modalities differ in computational neuroscience in various resolutions, what challenges need to be addressed caused by the limitations in DL models,

how can explainability be utilized in neuroscience, and where is our review positioned in advancing the field compared to existing literature surveys?

- RQ2: What modeling techniques are mostly used in visual neuroscience, how does XDL play its role in making models interpretable and transparent, what are the attributes of state-of-the-art XDL approaches in terms of vision tasks solved, methodologies applied, and their respective outcomes achieved?
- RQ3: What XDL techniques are utilizable in clinical neuroscience, how are the techniques leveraged to diagnose, prognose, and treat the neurological and psychiatric conditions, what are the pipeline components in clinics, how do components impact diagnostic and prognostic outcomes, and how do the XDL techniques build trust and transparency through supportive applications?
- RQ4: What are the key challenges in the state-of-the-art research relevant to visual and clinical computational neuroscience, and how are the biological priors integrable?
- RQ5: What future research directions need to be followed for advancing the explainability required for building interpretable and transparent systems in visual and clinical neuroscience environments, and how can the biological priors, standards, and human-involving designs be integrated to address the key challenges?

2. Methodology

This review follows a systematic approach to ensure a comprehensive and unbiased selection of literature concerning XDL in computational neuroscience, emphasizing both visual and clinical applications. To enhance transparency, this review adheres to the PRISMA guidelines [40]. The search covered studies are recent, and most of the studies are from the last 5 years, and a very few studies are not recent, which only includes one study from the year 1988. Only two datasets are from the years 2000 to 2008, whereas most datasets are recent, lying within the range of years 2021–2025. All studies and datasets have a particular focus on recent advances in DL, neuroimaging, brain decoding, and explainability methods. The covered study distributions over the decades are as follows. The most prominent datasets were Sleep-EDF [41], OASIS-3 [42], and OpenNeuro [43]. The process involved four primary steps: database selection, search strategy, inclusion/exclusion criteria, and screening.

2.1. Databases Searched

The following databases were selected for their comprehensive indexing of literature in neuroscience, artificial intelligence, and biomedical applications [44]:

- PubMed indexes peer-reviewed biomedical and clinical research, including studies involving neuroimaging, cognitive neuroscience, and brain disorders. It was essential for retrieving medically relevant studies using DL models in clinical neuroscience. Coverage ensured comprehensive literature retrieval.
- Scopus provided multidisciplinary coverage, including computational modeling, neuroscience, and machine learning. Its citation tools aided in identifying influential papers and landmark studies.
- Web of Science offered access to high-impact journals in both science and technology, helping identify interdisciplinary contributions in XDL research.
- IEEE Xplore offers access to high-impact technical literature in AI, machine learning, and signal processing. It was crucial for capturing studies on DL model architectures, BCI, and signal decoding from EEG or MEG.
- SpringerLink provides a list of prestigious journals such as Cognitive Neurodynamics, Computational visual media, BioMed Central Medical Ethics, BioData Mining, BioMed

Central Psychiatry, and Journal of neuroengineering and rehabilitation. It supports diverse neuroscience research dissemination worldwide.
- ACM Digital Library offers a variety of disciplinary journals related to evolutionary learning and optimization.
- MDPI also offers a list of high-ranked journals, including Mathematics, Diagnostics, Sensors which gave us the knowledge about clinical and visual applications relevant to multidisciplinary domains, including IoT, medical, and mathematics.
- arXiv provides early-access and research. These preprints are not peer-reviewed yet but serve as early dissemination of recent in-progress research work. It facilitates rapid sharing among neuroscience researchers.
- Books/Other Sources provide knowledge on theoretical concepts, standardized methods, and motivational aspects of the research.

2.2. Search Strategy and Keywords

A structured search strategy was designed using combinations of controlled vocabulary and free-text terms. The literature search was last conducted on 3 September 2025. The following primary terms and combinations were used:

Core concepts
- Explainable deep learning
- Interpretable deep neural networks
- Transparent AI in neuroscience
- Black-box models in neuroscience
- Uninterpretability in neuroscience
- Model interpretability in clinical AI

Neuroscience focus
- Brain decoding with AI
- Visual neuroscience and deep learning
- Clinical neuroimaging models
- Neurobiologically plausible models
- Functional brain connectivity analysis
- Electrophysiological signal modeling
- Spatio-temporal neural dynamics

Applications
- Cognitive state prediction
- Motor intention decoding
- Neurological disease diagnosis
- Psychiatric biomarker extraction
- Brain-computer interface explainability
- Neural signal reconstruction

Modalities
- EEG deep learning explainability
- MEG and fMRI decoding
- ECoG-based neural interfaces
- Multimodal neuroimaging fusion
- Time-series neural data interpretation

Techniques
- Feature attribution (SHAP, LRP, saliency maps)
- Concept-based explainability (TCAV, ACE)
- Causal explanation models
- Transformer interpretability in neuroscience
- Post-hoc vs integrated explainable methods
- Benchmarking interpretability metrics

Boolean operators used in the search are as follows
- AND (e.g., “explainable deep learning AND fMRI”)
- OR (e.g., “brain decoding OR neural signal interpretation”)

We used a combination of the below list items in our search. The Boolean operator ‘OR’ is used for word synonyms. Whereas, the Boolean operator ‘AND’ is used with different concepts, such that the most relevant studies appear in the search result. The list items are as follows:

- (“Explainable deep learning” OR “Interpretable deep neural networks”

OR "Transparent AI" OR "Black-box models" OR "Uninterpretability"
OR "Model interpretability")

- ("Brain decoding" OR "Visual neuroscience" OR "Clinical neuroimaging"
OR "Neurobiologically plausible models" OR "Functional brain connectivity"
OR "Electrophysiological signal modeling" OR "Spatio-temporal neural dynamics")
- ("Cognitive state prediction" OR "Motor intention decoding"
OR "Neurological disease diagnosis" OR "Psychiatric biomarker extraction"
OR "Brain-computer interface explainability" OR "Neural signal reconstruction")
- ("EEG" OR "MEG" OR "fMRI" OR "ECoG"
OR "Multimodal neuroimaging" OR "Time-series neural data")
- ("Feature attribution" OR "SHAP" OR "LRP" OR "Saliency maps"
OR "Concept-based explainability" OR "TCAV" OR "ACE" OR "Causal explanation"
OR "Transformer interpretability" OR "Post-hoc explainable methods"
OR "Benchmarking interpretability metrics")
- ("Kappa observer agreement for categorical data")

Searches were conducted individually on each database to maximize breadth and eliminate indexing limitations.

2.3. Inclusion and Exclusion Criteria

Time windows:

- One study is from 1988, while most are from 2020 to 2025
- Two datasets are from 2000 to 2008, while most are from 2019 to 2025

Inclusion criteria:

- Peer-reviewed journal articles or conference proceedings
- Articles focused on XDL, neuroscience modeling, neuroimaging, or clinical diagnosis
- Research using brain data modalities (EEG, fMRI, MEG, MRI, ECoG)
- Papers on interpretable/explainable brain and cognitive modeling

Exclusion criteria:

- Non-English publications and preprints without peer review
- Editorials, white papers, and non-peer-reviewed articles
- Theoretical AI papers not involving neuroscience/brain data
- Duplicates or articles without full-text access
- Reviews without methodological description or reproducible implementation

2.4. Screening Procedure

All references were managed in order of topics. After importing, duplicate entries were removed. The Cohen's Kappa was used to evaluate the screening procedure [45]. Two reviewers (A.Mehmood. and F.Mehmood.) independently screened titles and abstracts of 268 records identified from 9 databases for potential inclusion based on predefined eligibility criteria. Reviewer 1 and reviewer 2 agreed to include 165 and exclude 75 studies. Whereas the studies over which reviewer 1 included and reviewer 2 excluded are 12, and which reviewer 1 excluded and reviewer 2 included are 16. Discrepancies between reviewers were resolved through discussion, and Cohen's kappa ($\kappa = 0.764$) was calculated to quantify inter-rater agreement, with observed agreement $P_o = 0.895$ and expected agreement by chance $P_e = 0.556$.

Figure 2 shows the PRISMA flow diagram detailing the screening and inclusion process in the computational neuroscience study. In this approach, 347 records were identified from 9 different indexing databases. After the removing 79 duplicate records and excluding irrelevant 87 records, 177 studies were included in this literature.

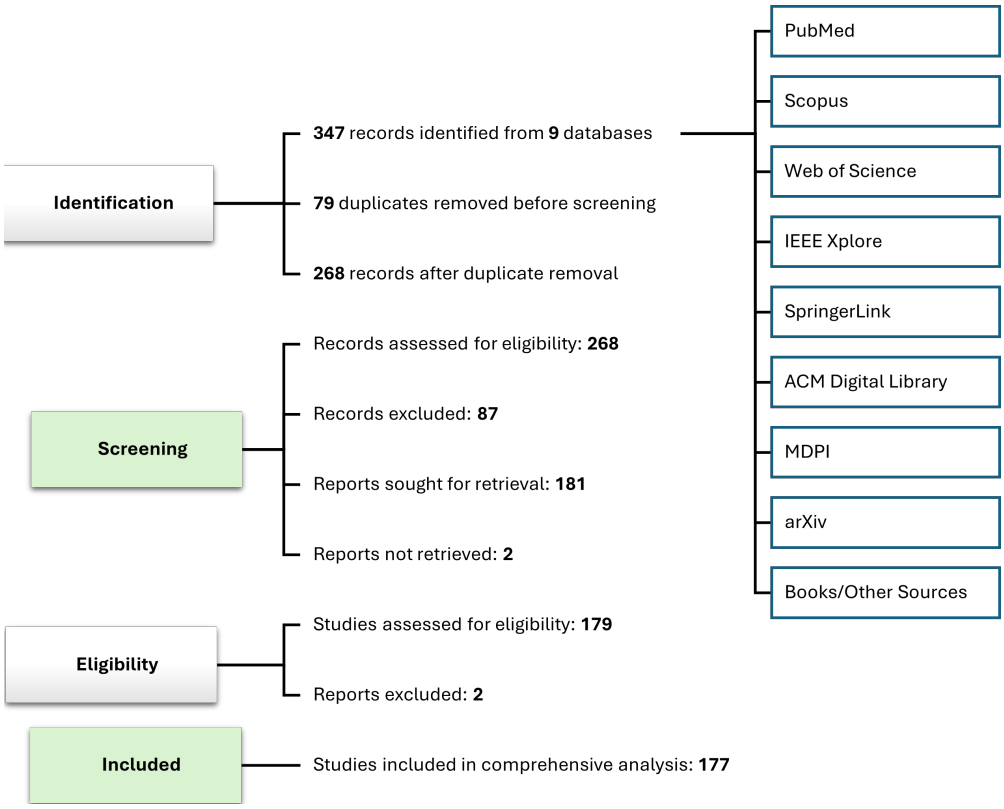


Figure 2. PRISMA-based systematic study selection for XDL in computational neuroscience.

3. Background and Foundations

In this Section, an overview of the foundations of computational neuroscience and DL is provided, followed up by the examination of XDL methods involved in aiming to bridge the gap between performance and interpretability [46].

3.1. Computational Neuroscience: A Brief Overview

Computational neuroscience is an interdisciplinary field that employs mathematical models, theoretical analysis, and simulations to understand the functions of the nervous system. It bridges the gap between experimental neuroscience and theoretical modeling, providing insights into how neural circuits process information.

The nominal research in computational neuroscience includes Dayan and Abbott's Theoretical Neuroscience [47]. This work offers a comprehensive introduction to modeling neural systems mathematically. In addition, methods for neuronal modeling by Koch and Segev et al. [48] put their primary focus on delving into the bio-physical modeling of neurons and networks. The model presented in Izhikevich et al.'s work focuses on the analysis of balancing between biological plausibility and computational efficiency, which makes it suitable for extensive simulations [49]. The above-mentioned models act as foundations that have initiated the way for more complicated simulative experiments, thus, improving and advancing the understanding of neural dynamics.

Table 2 summarizes a comparison of various brain data modalies' temporal and spatial resolutions. The provided statistics depict the invasiveness and applicability of various DL applications. These characteristics highlight why fMRI could be a considerable modality in computational neuroscience research because of the higher spatial resolution. In contrast, the EEG can not be neglected, as it is considered best due to its applicability in real-time systems. The ECoG is also a modality with high temporal and spatial resolution but is less preferable in practice as compared to the former because it requires surgery to the surface of the brain [50].

Table 2. Comparison of brain data modalities and their suitability. Modalities such as EEG, MEG, fMRI, fNIRS [51], and ECoG [52] depict their characteristics for DL applications.

Modality	Temporal Resolution	Spatial Resolution	Invasiveness	Suitability for DL
EEG	High (1–5 ms)	Low (~11 mm)	Non-invasive	High
MEG	High (<1 ms)	Medium (2–4 mm)	Non-invasive	Medium–High
fMRI	Low (500–2000 ms)	High (1.5–3.5 mm)	Non-invasive	High
ECoG	High (~1 ms)	High (1–4 mm)	Invasive (minimal)	High
fNIRS	Medium (>100 ms)	Low (~10 mm)	Non-invasive	Medium

3.2. Deep Learning Models in Neuroscience

The field of neuroscience has been made technologically advanced by leveraging the DL models, by enabling the analysis of sophisticated, multi-dimensional data. DL models, particularly CNNs and RNNs, have been applied to various neuroimaging modalities, including MRI, fMRI, and EEG. More recently, transformers [17] have gained traction due to their powerful attention mechanisms and scalability, especially in applications requiring long-range dependencies across time or space, such as predicting brain responses to naturalistic stimuli. Transformers, originally developed for NLP tasks, are being widely adopted for complicated data in the biological domain, i.e., genomics [53].

For instance, Plis et al. demonstrated the efficacy of DL in classifying brain states from fMRI data [54], while Storrs and Kriegeskorte explored the use of DL models to simulate cognitive processes [55]. Livezey and Glaser reviewed DL architectures for decoding neural signals [56], highlighting their potential in translating brain activity into behavioral predictions. Despite these advancements, challenges remain, such as the need for large datasets [57], model interpretability [8], and generalizability across subjects and tasks.

Figure 3 illustrates a typical DL pipeline for neuroimaging data, showcasing the flow from raw data input through preprocessing, feature extraction using CNNs or RNNs, and finally classification or regression outputs. This figure highlights several modalities used in neuroscience. It includes fMRI, EEG, MEG, DTI, MRI, etc. The corresponding tasks shown are motion correction, artifact removal, and noise reduction, which are associated with cognitive or clinical neuroscience. In a typical experimental design, the raw data is processed to normalize it, which is further sent to convolutional layers to extract features such that the model is able to learn patterns from it and make associations leveraging the layers and mechanisms defined with respect to the associated model architecture. In the final layer of DL architecture, the outcomes are normalized and converted into a human-readable form dependent on the use case.

To enable robust modeling of brain structure, function, and pathology, DL in computational neuroscience heavily relies on large-scale, well-annotated datasets. Table 3 presents a curated selection of representative brain datasets frequently used in both visual and clinical neuroscience applications. These datasets span a wide range of imaging and electrophysiological modalities, including MRI, PET, EEG, MEG, and EOG, and support diverse tasks such as Alzheimer's diagnosis, e.g., ADNI [58,59], OASIS-3 [42], brain connectivity mapping, e.g., HCP [60], motor imagery decoding (e.g., BCI Competition) [61,62], and tumor segmentation (e.g., BraTS) [63–65]. In contrast to physiological modalities, PPG-based non-neuro but biologically meaningful signals like pulse wave velocity are also utilized to enrich systematic modeling and neural interaction [66,67]. Some datasets, like UK Biobank [68] and OpenNeuro [43], also offer extensive metadata or multi-domain recordings, making them particularly valuable for developing generalizable models. The variability in sample size and data richness across datasets reflects both their specialized use cases and the ongoing need for harmonization in benchmarking and evaluation. These datasets form the empirical foundation for advancing XDL approaches in neuroscience by

supporting reproducibility, enabling comparative studies, and grounding interpretability in biologically and clinically meaningful patterns.

Table 3. Representative brain datasets used in DL for visual and clinical neuroscience applications.

Dataset (Modality)	**Application Domain**	**Sample Size**	**General Remarks & Bias/Harmonization Notes**
ADNI (MRI, PET) [57]	Alzheimer's Disease diagnosis	1000 subjects	Widely used, multi-modal neuroimaging. Multi-site variability and intensity normalization applied [69]. Recent works leverage MRI and PET for Alzheimer diagnosis [70]
OASIS-3 (MRI) [42]	Aging and dementia studies	1098 subjects	Longitudinal, includes cognitive scores. Demographic imbalance (age/sex) and preprocessing harmonized across scans [71].
HCP (fMRI, DWI, MEG) [60,72]	Brain connectivity and functional mapping	1200 subjects	High-quality, useful for deep encoding. Scanner differences mitigated and standardized preprocessing pipelines [73]. Multimodal combinations (e.g., EEG and fMRI) explored in epilepsy studies [74].
CamCAN (MEG, MRI) [75]	Lifespan cognitive aging	650 subjects	Broad age range, resting-state and task. Age distribution bias and artifact removal and normalization applied [76].
OpenNeuro (fMRI, EEG) [43]	Cognitive tasks and perception	23 subjects (median) 928 subjects (largest)	Open-access, highly diverse tasks. Small sample bias for some tasks and task standardization recommended [43].
BCI Competition IV (EEG) [61,77]	Motor imagery	9 subjects	Standard benchmark for BCI modeling. Very small sample, class imbalance, and cross-subject normalization suggested [78].
TUH EEG Corpus [79]	Epilepsy detection	24,000 records	Largest public clinical EEG dataset. Multi-center recordings, site-specific variations, and preprocessing standardization applied [80].
Sleep-EDF Expanded (EEG, EOG) [41]	Sleep staging	197 records	Used for sleep research and BCI studies. Limited sample, demographic bias, artifact correction applied [81].
UK Biobank (MRI Genetics) [68,82]	Population-level analysis	500,000 records	Imaging data with extensive phenotypic, genetic, and clinical metadata. Multimodal neuroimaging-genomics used for dementia prediction [83]. Site differences, large demographic variability, harmonized preprocessing and QC pipeline [84].
BraTS (MRI) [64,85]	Brain tumor segmentation	800 cases	Challenge dataset, used in DL segmentation. Recent multimodal MRI+genomics frameworks extend brain tumor applications [86]. Multi-institutional variability, intensity normalization, and augmentation applied [87].

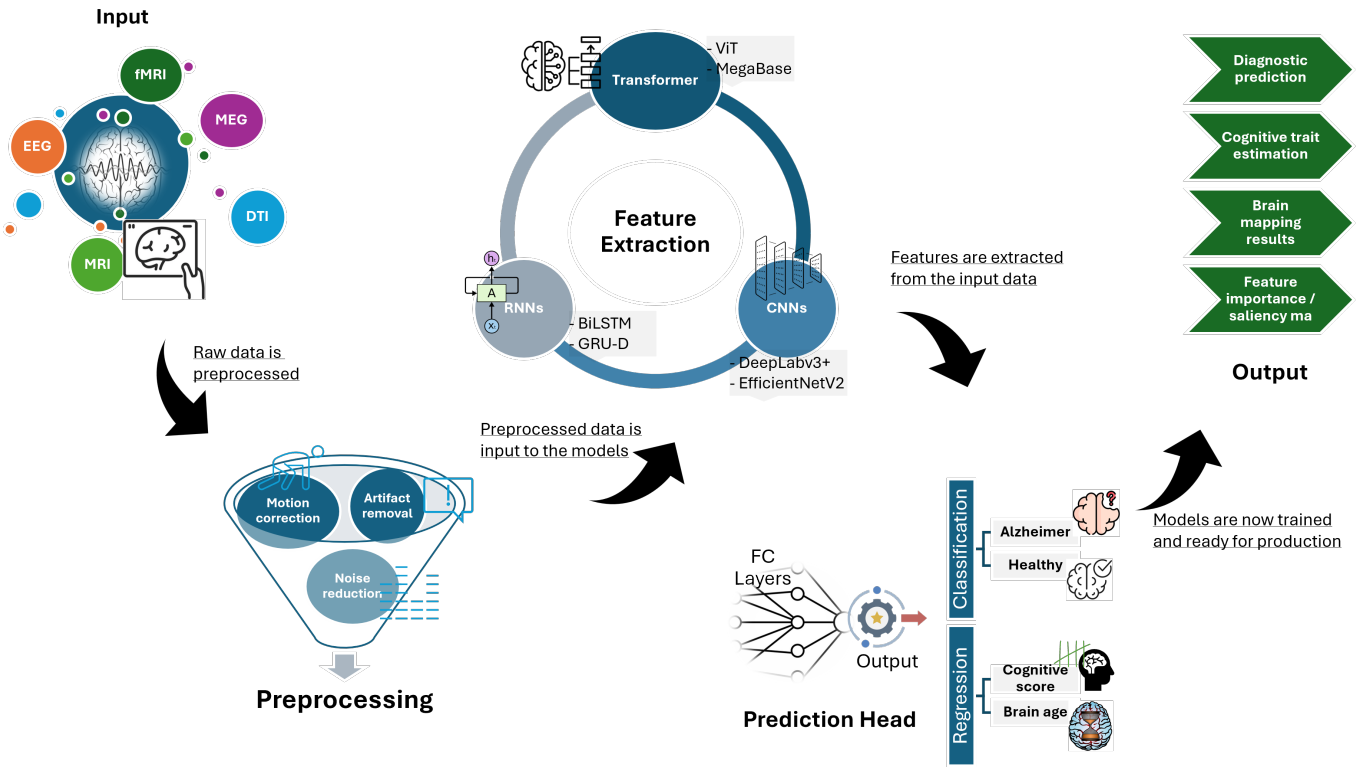


Figure 3. DL architecture for neuroimaging or brain signals.

3.3. Explainable Deep Learning: Concepts and Techniques

As DL models become increasingly prevalent in neuroscience, the need for interpretability has grown [8]. The XAI aims to make DL models transparent, ensuring that their decisions can be understood and trusted. Such models are termed XDL, which includes the development of techniques such as LRP [88], saliency maps [12], and concept activation vectors [89] to provide insights into model decisions.

Samek et al. introduced methods like LRP to visualize DL decisions, enhancing interpretability [90]. Schwalbe et al. provided a taxonomy of explainable methods, guiding practitioners in selecting appropriate techniques [37]. Ramachandram et al. explored the relationship between interpretability and trust in DL models, emphasizing the need for reliable explanations [91]. However, challenges persist, including trade-offs between accuracy and interpretability, computational complexity, and the lack of standardized evaluation metrics.

Multimodal explainability in DL, is another aspect to be considered while discussing the feature attribution and attention mechanisms, which are able to provide interpretable insights across multiple data modalities [92]. Another example of multimodal, which aligns with the current study methodologically, showcases the human activity recognition with RGB, skeletal tracking, and pose estimation [93], which provides a blueprint for combining neural modalities such as EEG, fMRI, and fNIRS.

Table 4 summarizes four major explainable methods. It highlights the attributes of each in terms of its interpretability and limitations along with their respective focal application areas. Saliency methods are usually found to visualize the model's performance but only through the correlations, whereas perturbation methods provide feature importance, which is found to be highly interpretable but comes with a cost. The concept-based method requires human efforts in preparing data needing semantics and is found to be highly interpretable. Intrinsic explainable methods use specialized techniques such as attentions that offer transparency of the model to the stakeholder, which is also known to be efficient

and highly interpretable. Saliency maps enable the provision of correlational insights but are not able to provide causal links between features and outcomes. Due to this limitation of not being able to provision causal links, concept-based approaches allow testing the model performance (sensitivity) in accordance with the human-defined concepts. While causal validation methods like perturbation assess whether the highlighted regions (predictions) are true and valid. Overall, each method is unique in its role and applications. By keeping the given aspects of each category in mind, it makes it easier to choose the most suitable candidate method.

Table 4. Categories of XDL methods, their interpretability levels, and application areas, along with the limitations in computational neuroscience (further methodological details are illustrated in Section 5.3).

Category	Techniques	Interpretable	Application Areas	Limitations
Saliency [12]	LRP, Grad-CAM [94]	Medium	Visualizing model attention	Highlights correlation only, not causal
Perbutation [95]	LIME, SHAP	High	Feature importance analysis	Computationally expensive, post-hoc
Concept [96]	TCAV, ACE	High	Understanding concepts	Requires semantic definition
Causal validation [97]	MR analysis, ablation, counterfactuals, interventions	High	Testing causal contribution of features	Complex to design, limited scalability
Intrinsic XDL [98]	Attention mechanisms, interpretable architectures	High	Model transparency	Architectural limits accuracy

This comprehensive overview of computational neuroscience, DL applications, and explainable AI techniques sets the foundation for understanding the integration of these domains. The subsequent sections will delve deeper into specific applications in visual and clinical neuroscience, addressing the challenges and future directions in the field.

3.4. Comparison with Existing Surveys

Table 5 compares our review with existing surveys in the field. Previous works include explainability application in neurostimulation [99], in cognitive functions and dysfunctions [100], and in healthcare for general interpretable systems [101]. Each of the surveys depicted is either too broad in topic or puts its focus on explainability only in a specific area. These studies direct our research somehow to support explainable models specific to computational neuroscience. Unlike prior surveys, our work is the first systematic PRISMA-guided review for the visual and clinical computational neuroscience domain and the second systematic PRISMA-guided review for neuroscience in general. It specifically focuses on the integration of XDL into computational neuroscience. Moreover, this study identifies the research gaps, proposes a structured future direction, and highlights the key challenges in provisioning explainable and interpretable systems in computational neuroscience.

Table 5. Comparison of existing surveys in neuroscience.

Survey Focus	Methodology	Applications	Explainable?	Datasets	Novelty	Limitations	Our Review
Application of explainability in neurostimulation [99]	Literature-based conceptual overview, and no formal/PRISMA analysis	DBS, computational psychiatry, broader potential in other medical fields	Yes (mechanistic, functional)	neural, behavioral, imaging	Framework for integrating Explainability. Discussion on closed-loop behavioral modulation	Lacks structured evidence synthesis and fewer technical insights. Mainly focused on what is needed rather than solutions.	Provides structured and technical evidence, and covers visual & clinical domains. Addresses explainability trade-offs & need for benchmarks
Focuses on explainability applied to cognitive functions and dysfunctions [100]	Joanna Briggs and PRISMA	Normal cognition and impaired cognition	Yes (good)	Anatomical datasets. Emphasis on correlation finding	Qualitative framework to evaluate explainability in cognitive neuroscience	Limited number of studies (12). Oversimplification	Larger evidence studies (177), discussed key issues, explainability trade-offs, and biological priors integration
Broad application of DL in medical imaging [102]	Narrative review of 300+ studies	Cancer, radiology, pathology	No	Imaging datasets -MRI -CT -PET	Comprehensive early review of DL in medicine	Lack of interpretability. Minimal focus on neuroscience	Extends DL towards neuroscience with explainability
DL in fMRI Neuroimaging [103]	Systematic DL function survey of neuroimaging	Cognitive neuroscience, brain state decoding	Limited	fMRI datasets	Highlights DL potential in functional brain mapping	Does not cover explainable AI approaches	Integrates multimodal datasets and explainable tools
DL for structural and functional brain MRI [104]	Summarized DL pipelines	Alzheimer's, Multiple sclerosis, tumor segmentation	No	MRI datasets	Overview of DL pipelines in neuroimaging	No interpretability focus. Clinical translation briefly discussed	Focuses on explainable clinical trust and pipelines
DL across biomedical signals and health data [105]	Review over EEG, MRI, EHR, multimodal	EEG, electronic health records, wearable signals	Minimal	EEG, MRI, multimodal data	Broad biomedical scope including signals	Not neuroscience-specific. Lacks transparency discussion	Centers on neuroscience with interpretability emphasis
DL applied to EEG data [106]	Focused review on DL models	BCI, seizure prediction	Limited	EEG datasets	Detailed DL coverage in EEG domain	No significant focus on XDL techniques	Part of broader multimodal explainable DL perspective
CNN applications in neural imaging [107]	Overview of CNN methods	Object recognition, vision tasks, neuropathology	No	MRI, histology	Early CNN-centric neuroimaging focus	Excludes transformer and XDL methods	Covers CNN, RNN, Transformers with explainability
Interpretability and explainability in healthcare AI [101]	Narrative and taxonomy of LIME, SHAP	Healthcare broadly, not neuroscience-specific	Yes	Clinical datasets (EHR, imaging)	Defines interpretability frameworks in AI	Not neuroscience-centric	Bridges XDL specifically into visual and clinical neuroscience

4. XDL in Visual Computational Neuroscience

4.1. Modeling Visual Attention and Categorization

This subsection explores how DL models are employed to simulate and understand the neural mechanisms of visual attention and object categorization, two fundamental cognitive processes governed by the visual system. The CNNs and ViTs are particularly instrumental in modeling how the brain processes visual stimuli, replicating hierarchical patterns seen in the human visual cortex. These models are trained on tasks such as scene recognition, saliency detection, and object categorization, thereby providing computational analogues to cognitive processes like selective attention and feature binding. For example, recent developments in wearable devices (for eye-tracking) [108] provide tools for validating and standardizing DL models for real-time environments.

Recent innovations in attention mechanisms within DL architectures (e.g., self-attention in ViTs) serve as bridges between artificial and biological vision systems. These mechanisms enable offering meaningful associations of data relevant to biological systems that require a deep focus. This feature distinguishes between top-down and bottom-up attentional processes, for example, objective-driven and stimulus-driven, respectively. Specific to visual attention, hardware innovations are also being advanced, including the study on the integration of a real-time eye tracker with in-pixel IP, where authors explore the neural mechanisms of attention, further supporting the explainability in visual computational neuroscience [109].

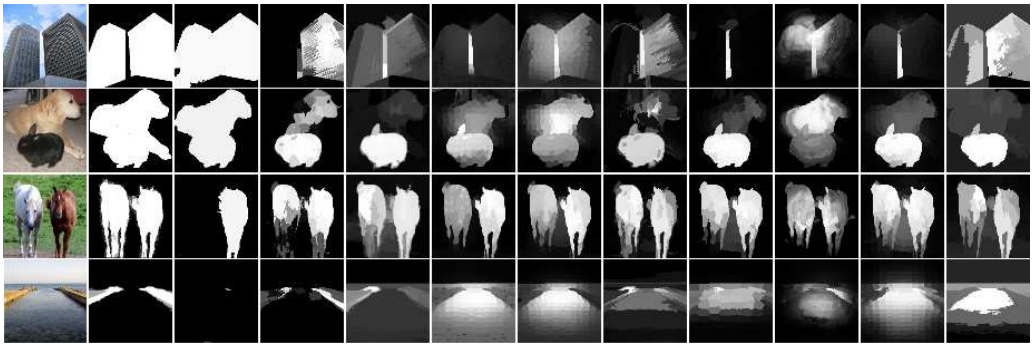
To illustrate the mechanisms applied for explainability, Figure 4 depicts how the generated attention maps based on saliency methods are prioritized by the model when processing visuals. This prioritization is because these methods proviide better insights into the data, which play a key role in making categorical decisions. These methods are used as validating tools or hypothesis-generating tools in the domain of neuroscience. Though with some limitations, the mentioned techniques show comparisons to other techniques depicted in tables provided beside Figure 4a–d.

4.2. XDL for Interpreting Visual Models

As DL models achieve remarkable accuracy on visual perception tasks, their explainability becomes essential for neuroscientific interpretation and validation. This subsection focuses on the application of XDL tools to interpret these models in visual neuroscience contexts. Saliency maps, CAM, integrated gradients, and concept-based explanations such as TCAV are among the key interpretability tools that reveal which spatial or semantic features contribute to a model's output.

These methods make it possible to align DL behavior with biological plausibility, assessing whether models rely on the same image features that human vision systems prioritize. For instance, comparing CAM-generated heatmaps with fMRI activation regions allows researchers to evaluate how closely artificial networks mirror brain processing. Moreover, concept attribution tools help isolate abstract semantic categories that models use internally, paralleling the brain's hierarchical feature representations.

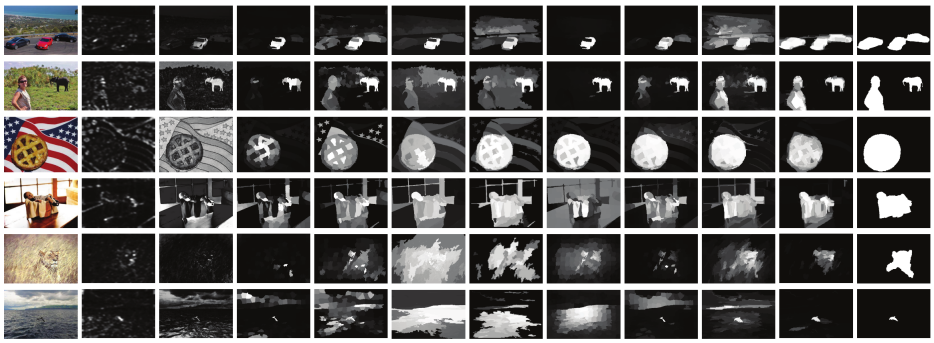
Graphical summaries of visual XDL approaches are shown in Figure 5. X-axis depicts the model families, while the y-axis depicts the visual tasks. Each study's respective brain-data alignment, XDL method, and selection criteria are comprehensively highlighted.



| Input image | CNN-DecNN+RACDNN | MCDL | MDF | RRWR | BSCA | DRFI | RBD | DSR | MC | HS |

CNN-DecNN Performance		
Dataset	**F-M**	**MAE**
HKUIS	85.57	7.03
ECSSD	87.81	8.12
SED2	85.35	9.29

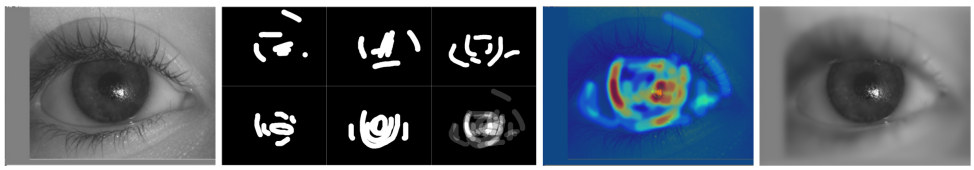
(**a**) Qualitative saliency results.
(Reprinted from 'Recurrent attentional networks for saliency detection,' arXiv:1604.03227, under the CC BY 4.0.)



| Source | SR | FT | SF | GS | HS | RC | MR | wCtr* | DRFI | MDF | GT |

Dataset	Enhanced Performance	
MSRA-B	Precision:	86.4%
	Recall:	87.0%
	↑ F-M:	5.0%
	↓ MAE:	5.7%
HKUIS	↑ Precision by:	9%
	↑ Recall:	5.7%
	↑ F-M:	13.2%
	↓ MAE:	35.1%
SOD	↓ MAE:	17.1%
iCoSeg	↓ MAE:	26.3%

(**b**) Visual comparison of saliency maps. ↓ denotes an increase, whereas ↓ denotes decrease. wCtr* refers to optimized weighted contrast.
(Reprinted from 'Visual saliency based on multiscale deep features,' arXiv:1503.08663, under the CC BY 4.0.)

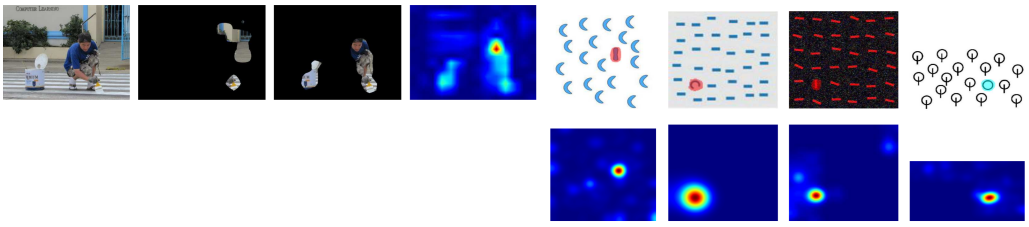


| Input image | Five individual correct annotated for saliency map | Heatmap | Blurred after training |

PAD Classification Error Rates	
Metric	**Values**
$APCER_{average}$	24.14%
BPCER	8.61%
ACER	16.37%

(**c**) Human saliency maps.
(Reprinted from 'Human-aided saliency maps improve generalization of deep learning,' arXiv:2105.03492, under the CC BY 4.0.)

| Input image | After fine-tuning | Activation map after fine-tuning | Fine-tuning with highest mean NSS |



| Top row: annotated masks | Bottom row: Activation maps from layer conv4-1 |

Saliency model performance	
Model	**NMM (unit-less)**
DeepGaze	1.66
SAM	1.25
DVA	1.19
GBVS	2.57
BMS	3.65

(**d**) Activation maps and synthetic search arrays.
(Reprinted from 'Understanding and visualizing deep visual saliency models,' arXiv:1903.02501, under the CC BY 4.0.)

Figure 4. Saliency/attention maps from DL in visual tasks with accompanying faithfulness evaluations [110–113].

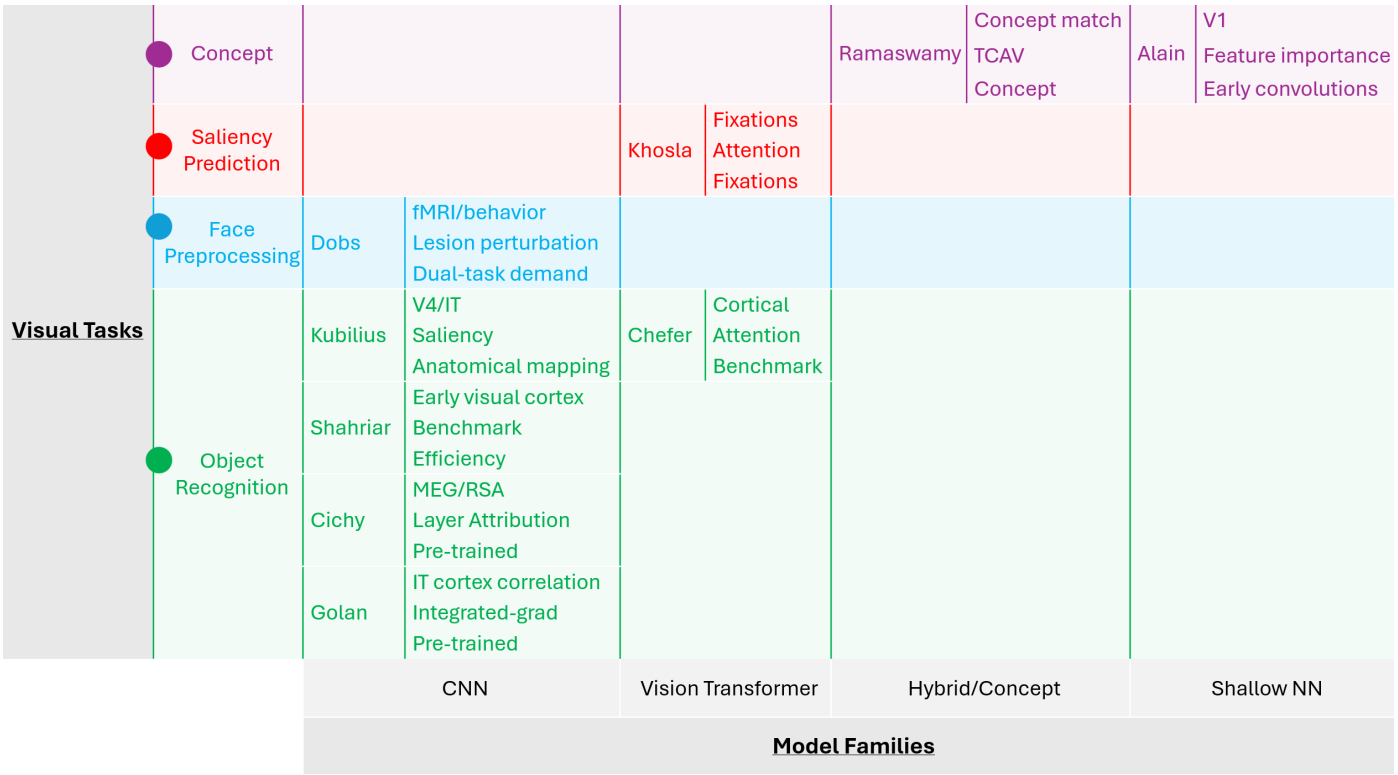


Figure 5. Comparison of visual XDL approaches: model families and visual tasks.

Table 6 is provided with a comparison of visual XDL approaches, mentioning the model type, task, interpretability tool, and their outcome. This table categorizes recent studies by the DL architecture used (e.g., CNN, ViT), task domain (e.g., object recognition, Image saliency prediction), XDL-based interpretability technique applied (e.g., Saliency maps, CAM, Layer-wise feature attribution), and the outcome or neuroscientific insight gained (e.g., feature correspondence with visual cortex, improved transparency for downstream analysis). The comparative view helps assess the effectiveness of different XDL techniques in contributing to visual neuroscience. By analysis of the attributes of each reference in this table, it can be concluded that the research work by Kubilius et al., Dobs et al. [114], Chefer et al. [115], Golan et al. [116], Khosla et al. [117], and Ramaswamy et al. [96] triumph, as each directly links model type, task, interpretability tool, and the respective associated outcome.

Table 6. Comparison of visual XDL approaches: model, task, interpretability, outcome, criteria, and brain-data alignment.

Reference	DL Model Type	Visual Task	XDL Method	Outcome/Neuroscientific Insight
Kubilius [118]	CNN (CORnet-S)	Object recognition	Saliency maps, CAM	Layer-wise features map to ventral stream hierarchy.
Selection criteria: Anatomical mapping and recurrence Brain-data alignment: Composite Brain-Score (V4, IT, behavior)				
Dobs [114]	CNN	Face detection Object detection	Lesion experiments	Networks spontaneously segregate face and object processing similar to brain functional specialization. Dual-task networks resemble human visual behavior.
Selection criteria: Chosen to test dual-task demands and emergence of specialized subsystems Brain-data alignment: Aligned CNN representations with fMRI/behavioral face–object patterns				

Table 6. *Cont.*

Reference	DL Model Type	Visual Task	XDL Method	Outcome/Neuroscientific Insight
Chefer [115]	Vision transformer	Image classification	Attention visualization	Attention heads correlate with semantic segment boundaries and objects. This aids in transformer interpretability.
Selection criteria: Selected for benchmark performance and interpretability Brain-data alignment: Aligned attention maps with cortical object-selective activations				
Shahriar [119]	CNN (lightweight)	Image classification	Performance benchmarking	Transfer learning enhances CNNs for memory-limited devices via accuracy-efficiency trade-offs.
Selection criteria: Selected for low parameters and efficiency Brain-data alignment: Feature similarity with early visual cortex				
Cichy [120]	CNN (AlexNet)	Brain–DL model alignment	Layer-wise feature attribution	Temporal dynamics of brain activity matched CNN layer activation.
Selection criteria: Pre-trained CNN on image recognition benchmarks Brain-data alignment: RSA between CNN layers and MEG data				
Golan [116]	CNN (VGG-16)	Visual categorization	Integrated gradients	Feature importance maps show overlap with IT cortex response patterns.
Selection criteria: High-performing CNN trained on object recognition Brain-data alignment: Correlation of saliency maps with primate IT activity				
Khosla [117]	Vision transformer	Image saliency prediction	Self-attention visualization	Predictive saliency regions correspond with human fixation maps.
Selection criteria: Alignment with human eye-tracking datasets Brain-data alignment: Correlation with fixation distributions				
Alain [121]	Shallow networks	Feature visualization (class)	Feature importance	Reveals shallow network focus on high-frequency, edge-related features.
Selection criteria: Early convolutional layers with clear feature localization Brain-data alignment: Feature activations compared with early visual cortex (V1).				
Ramaswamy [96]	CNN and TCAV	Concept learning in vision tasks	Concept activation vectors	Extracted abstract concepts matched labeled categories, e.g., striped patterns.
Selection criteria: Predefined visual concepts with clear semantic labels Brain-data alignment: Learned concept vectors compared with labeled categories				

Together, the advancements in modeling visual attention and categorization using DL, and the corresponding developments in explainable techniques, underscore a powerful convergence between computational modeling and neuroscience. By replicating the hierarchical and attention-driven mechanisms of the visual system, models such as CNNs and vision transformers offer valuable insights into how the brain processes complex visual environments. However, it is through XDL methods, ranging from saliency maps to concept-based explanations, that these models become more than just high-performing predictors. They transform into tools for scientific discovery, enabling researchers to link artificial visual computations with biologically plausible mechanisms. These explainability tools are not only able to improve the interpretability of neural models but are also capable of enhancing their credibility, facilitating their hypothesis, and promoting their trust in various domain applications. As we move forward, it will be necessary to integrate the performance-driven modeling and interpretability-driven analysis because enabling fair understanding of the brain through visual processing systems is vital in driving the future of neuroscience.

5. XDL in Clinical Computational Neuroscience

5.1. Clinical Use Cases of DL Models

The DL has emerged as a valuable tool in neuroscience, specifically clinical, as it offers substantial improvements in the diagnosis and recovery of neurological and psychiatric conditions. Models trained on neuroimaging-based datasets, including MRI [122], PET, EEG, and retinal scans, have always achieved a higher accuracy in pattern detection of various diseases like alzheimer [58], epilepsy [123], retinal degenerative [124], and major depressive disorder [125,126]. Multimodal approaches to provide richer clinical insights exist, which are able to capture complementary signals from different modalities. This approach enables a more comprehensive understanding of the neuron-based mechanisms [127]. Its importance can not be neglected, as it plays a key role in several applications such as alzheimer (MRI and PET) [70], epilepsy (EEG and fMRI) [74], brain tumors (MRI and genomics) [86], parkinson disease (MRI and clinical scores) [128], population densities (MRI and genetics) [83].

Among neurodegenerative disorders, neuroimaging has also been utilized for predicting Alzheimer's disease [129]. For instance, ensemble learning approaches suggest clock drawing tests that leverage the fusion of EEG signals and cognitive assessment data, which show satisfactory performance in alzheimer disease classification [130]. Another example of its usage is to train a CNN for classifying structural brain images, further allowing the detection of early alzheimer referring to an increasing sensitivity and specificity. Similarly, multi-scale CNN models for SaMD have proven to be effective in translational potential in clinical diagnostics, specifically the multi-class brain MRI classification [131], although the requirement for interpretability remains limited, which drives the necessity for adopting XDL approaches [131]. Apart from this, DL models trained on EEG data have been used to detect epileptic seizures with real-time performance [132], and generative models trained on MRI data have been utilized to synthesize missing brain data [133].

In addition to diagnostic applications, DL models are increasingly employed in predictive modeling and outcome forecasting. Recurrent and temporal models have been used to anticipate the progression of neurodegenerative diseases, predict response to treatment in depression, or forecast seizure likelihood in epilepsy patients. These predictive capabilities offer a new paradigm in clinical care, transitioning from retrospective to proactive and personalized interventions. For example, a recent article explores how hybrid anomaly detection approaches can be used to improve the predictability in rehabilitation by highlighting the key roles of explainable systems [134].

A recent study on MRI-based alternating magnetic field treatment systems discusses the role and impact of neuroimaging beyond diagnostics towards therapeutics [135]. Another example of applying DL over fNIRS signals is demonstrated to classify the disease of anesthesia showing the impact of neuroimaging [136] applications. We emphasize integrating the XDL approaches to support such systems, which can significantly enhance the treatment planning and monitoring, leveraging its explainable attributes by enabling safety validation.

Figure 6 presents a pipeline of a clinical XDL system, mapping the workflow from raw imaging input (e.g., brain MRI, retinal scan), through a DL model (e.g., CNN, transformer), to an explainable output that highlights relevant biomarkers or risk indicators. This pipeline emphasizes where and how explainability tools are integrated, such as feature attribution maps or concept-based justifications used by clinicians.

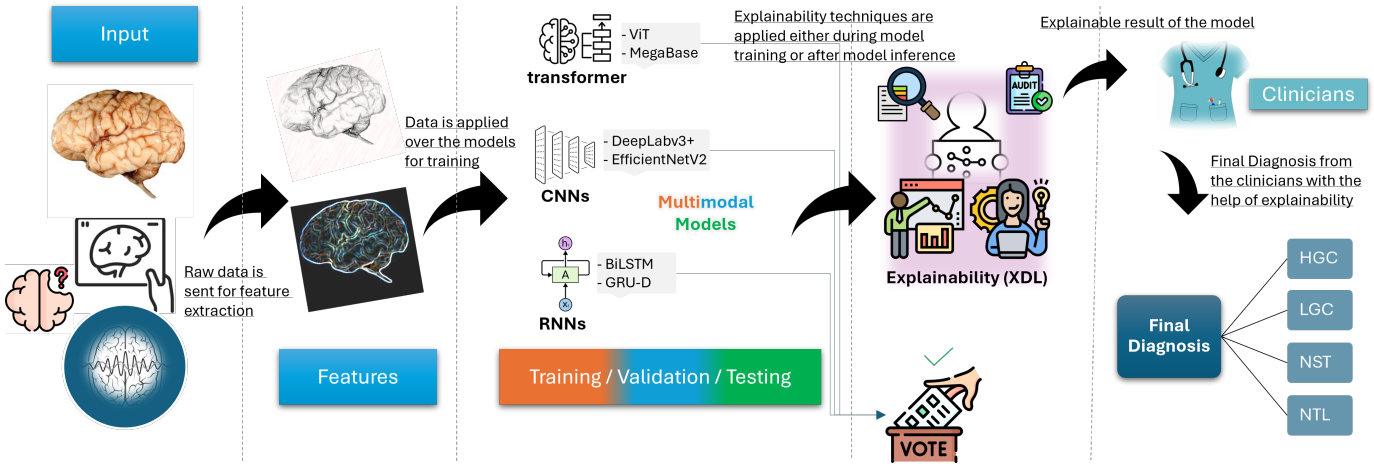


Figure 6. Clinical XDL pipeline: imaging -> model -> explanation.

5.2. Enhancing Trust in Clinical Models with XDL

While DL models have shown promise, their "uninterpretable" nature presents critical barriers in clinical settings where decisions must be interpretable, auditable, and ethically sound. The XDL addresses this concern by making model outputs transparent and comprehensible to both clinicians and patients. Methods such as LRP, SHAP, saliency maps, and rule-based surrogate models are increasingly used in clinical DL systems. These tools help identify which input features, like specific brain regions or image patterns, played the biggest role in guiding a diagnosis or prediction. These approaches also support patient-centered healthcare, and also enable the ethical deployment of AI systems [137].

Interpretability plays a key role in building trust, especially in sensitive fields like medical healthcare [138]. It goes beyond just technical insight, clear, explainable models also help meet growing legal and regulatory expectations, such as those set by the EU AI Act and FDA guidelines. From a practical standpoint, interpretability allows clinicians to better understand, challenge, or confirm model decisions, ultimately supporting more confident use in real settings.

OCTNet is another example of attention-based DL architectural enhancement, which demonstrates the optimization of interpretability and performance leveraging the retinal imaging, thus, aligns with the XDL goals for clinical neuroscience [139].

To offer a snapshot of how these ideas are being applied, Table 7 outlines recent clinical DL studies, along with the XDL methods used and the insights they provided. Among all references, the most applicable are He et al. [140], Zhang et al. [141], and Yang et al. [142], due to their ability to provide interpretable insights into functional brain connectivity with a higher accuracy and to provide tangible and explainable outcomes relevant to the retinal regions for diabetic retinopathy, respectively.

Table 7. Clinical DL applications: XDL methods and interpretability outcomes.

Reference	Clinical Domain	DL Model	XDL Method	Interpretability Outcome
Lundervold [143]	Alzheimer	CNN (3D-MRI)	Saliency map, LRP	Identified hippocampal and cortical atrophy as key-markers.
External Validation: Reported performance drop on external datasets [144] Calibration: Matched predicted with actual outcome rates [145] Uncertainty Quantification: Aleatoric and epistemic uncertainties estimated [145]				
Roy [146]	Epilepsy, EEG	RNN, 1D-CNN	Feature attribution	Revealed seizure-indicative frequency-bands.
External Validation: Not performed Calibration: Not specifically reported Uncertainty Quantification: Not quantified				

Table 7. *Cont.*

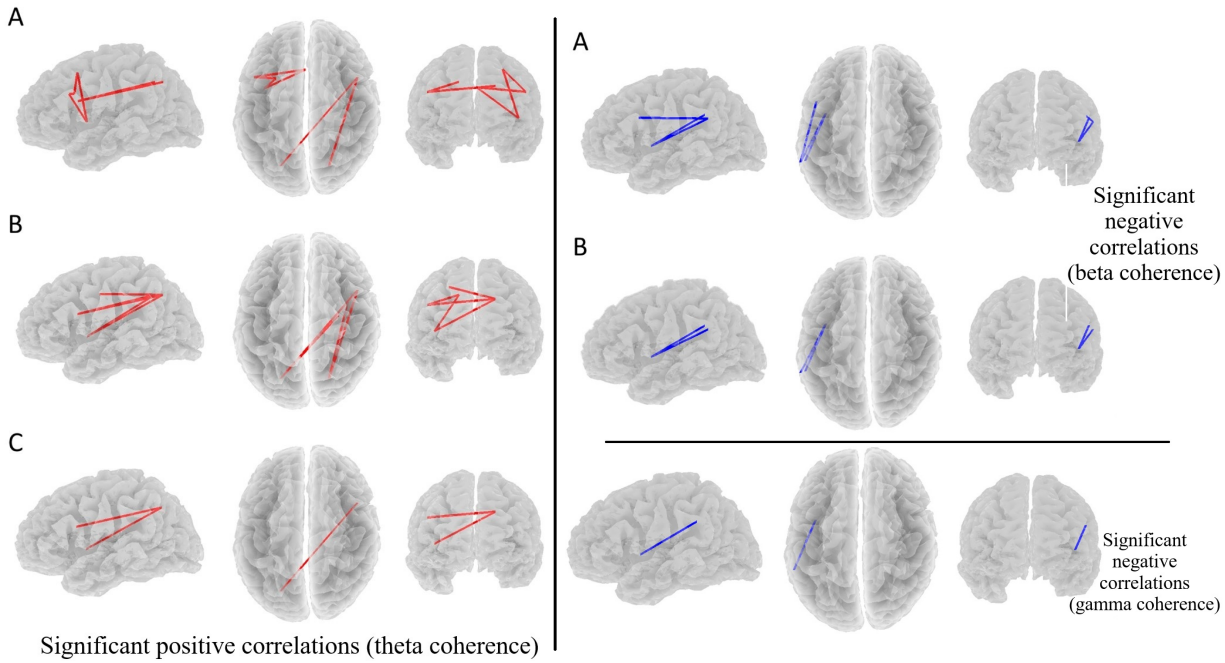
Reference	Clinical Domain	DL Model	XDL Method	Interpretability Outcome
Zhang [141]	Retinal disease	RetinaNet, ResNet	Grad-CAM	Highlighted retinal areas most associated with diabetic retinopathy through visual mapping.
External Validation: Tested on external datasets from EyePACS and Messidor-2 Calibration: Not explicitly reported Uncertainty Quantification: Not performed				
He. [140]	Depression, fMRI	Deep graph CNN	SortPooling layer	Achieved 72.1% classification accuracy on large multi-site MDD dataset. Improved interpretability by identifying salient brain regions through graph vertex sorting and functional connectivity patterns.
External Validation: 10-fold cross-validation on large multi-site REST-meta-MDD dataset Calibration: Not explicitly reported Uncertainty Quantification: Not performed				
Ahmed [147]	Brain tumor, MRI	ViT with GRU		Extracted essential features via ViT and captured feature relationships via GRU for classification.
External Validation: Validated on Brain Tumor Kaggle Dataset alongside primary MRI data from BSMMCH, Faridpur, Bangladesh Calibration: Calibration achieved with cross-entropy loss and performance curves Uncertainty Quantification: Uncertainty assessed through confidence scores and model robustness checks				
Jiang [148]	Glioma Diagnosis, Molecular Marker Discovery	Transformer	Attention maps, Saliency maps, Feature attribution	Visual explanations of tumor regions and molecular marker associations.
External Validation: Yes, tested on independent datasets (TCGA and Xiangya) Calibration: Not explicitly reported Uncertainty Quantification: Not explicitly reported				

5.3. Salient, Concept, and Causal-Based Validation in Clinical XDL

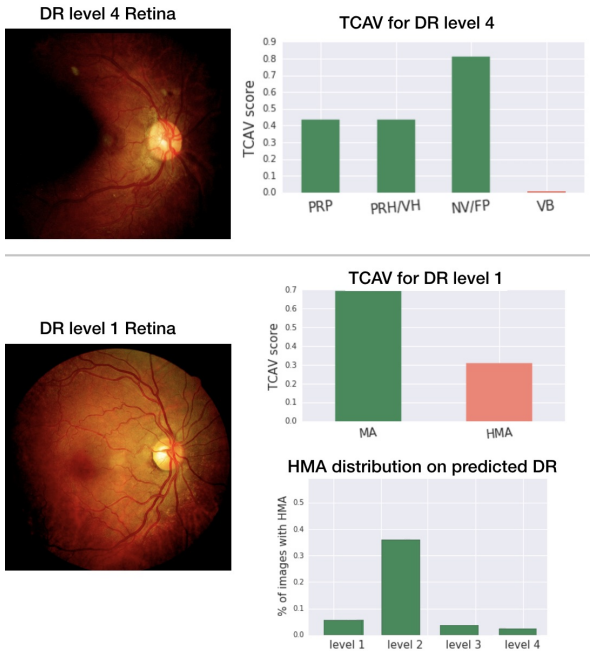
This section highlights the saliency, concept, and causal-based techniques utilized in clinical neuroscience. Figure 7 depicts the techniques, involving both XDL and heuristic types, which are considered for the evaluation of biological priors in clinical neuroscience. The techniques like saliency are evaluated on the basis of correlations (positive and negative) in terms of theta, beta, and gamma coherence. Other techniques like concept-based explanations, are useful in quantifying human-interpretable attributes that influence the performance of models. While causal validation approaches include MR analysis, ablation, and counterfactuals. Although MR analysis is not based on DL but serves as a powerful causal inference technique. The salient, conceptual, and causal-based validations are depicted in Figure 7a. In addition, the concept-based explainable technique is shown in Figure 7b, which explains the TCAV technique applied for DR level 4 and level 1, along with the depiction of the HMA distribution.

In Table 8, the results of MR analysis are shown. These were derived from the IVW method. The number of SNPs is shown, which are depicted in red dots, i.e., point estimates. Each exposure metric is listed to show its specific contribution to the causal estimate. The findings of MR analysis identify fresh fruit intake as a protective factor. In addition, it identifies lamb intake to be a risk factor. While most dietaries do not show significant causal effects.

In short, XDL has brought major advancements to clinical neuroscience, especially in diagnosing conditions and predicting outcomes. But despite these benefits, a lack of transparency can limit how these models are used in real-world care. To enable transparency of models, XDL plays a helping role in bridging this gap, which further encourages the trust, supports the ethical standards, and aligns with the regulatory needs. All the developments related to explainability in various domains are driving the future systems to make more reliable, focused, and explainable decision-making.



(**a**) Significant positive and negative correlations in view of theta, beta, gamma coherence.
(Reprinted from 'Differences in EEG functional connectivity in the dorsal and ventral attentional and salience networks across multiple subtypes of depression,' Appl. Sci. 2025, 15, 1459, under the CC BY 4.0.)



Top For DR level 4 images, the $TCAV_Q$ scores are high for features that are relevant to this severity level (highlighted in green), while features unrelated to DR are scored low (highlighted in red).

Middle For DR level 1 (mild) cases, TCAV reveals that the model sometimes confuses level 1 with level 2. This misclassification can be better understood through TCAV: features typically associated with level 1 (green, MA) show high $TCAV_{QS}$, but features related to level 2 (red, HMA) also contribute, indicating the model's overlapping attention.

Bottom The HMA feature appears more frequently in DR level 2 images than in level 1, helping to explain why the model occasionally predicts a higher severity than the ground truth.

(**b**) DR level 4 image and TCAV results.
(Reprinted from 'Interpretability Beyond Feature Attribution: Quantitative TCAV,' arXiv:1711.11279v5, under the CC BY 4.0.)

Figure 7. Salient, concept, and causal-based validations in clinical XDL [149,150].

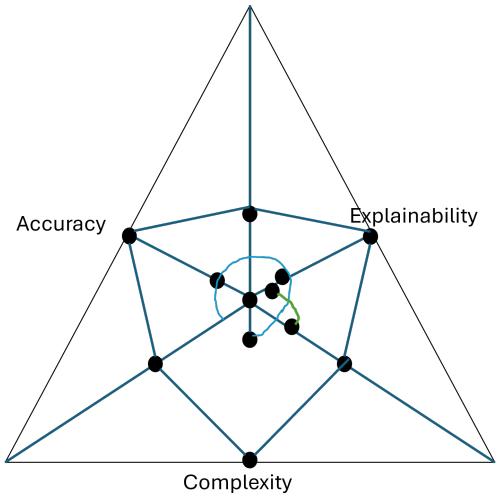


Figure 8. Triangle plot of accuracy, complexity, and explainability trade-offs.

Table 8. The MR analysis using the IVW method. The red dots depict the point estimate [151]. (Reprinted from 'The Causal Relationship Between Dietary Factors and the Risk of Intracranial Aneurysms: A Mendelian Randomization Analysis,' Biomedicines 2025, 13, 533, under the CC BY 4.0).

Exposure	SNPs	Odds Ratio (95%CI)	Odds Ratio (95%CI)	*p* Value	Significant
Alcoholic drinks per week	34		1.057 (0.788∼1.42)	0.711	No
Alcoholic intake frequency	96		1.084 (0.909∼1.29)	0.369	No
Processed meat intake	23		1.354 (0.55664∼3.29)	0.504	No
Poultry intake	7		1.06 (0.157∼7.19)	0.949	No
Beef intake	14		0.5945 (0.114∼3.1)	0.537	No
Non-oily fish intake	11		0.72101 (0.156∼3.34)	0.676	No
Oily fish intake	61		0.815 (0.4796∼1.39)	0.45	No
Pork intake	14		4.375 (0.717∼26.7)	0.11	No
Lamb/mutton intake	31		2.831 (1.0304∼7.78)	0.044	Yes
Bread intake	30		1.097 (0.518∼2.32)	0.81	No
Cheese intake	61		1.21 (0.697∼2.1)	0.499	No
Cooked vegetable intake	17		0.66 (0.18357∼2.37)	0.524	No
Tea intake	39		1.086 (0.66∼1.79)	0.745	No
Fresh fruit intake	53		0.209 (0.0893∼0.491)	0.000327	Yes
Cereal intake	38		0.591 (0.3001∼1.17)	0.129	No
Salad/raw vegetable intake	19		0.807 (0.1841∼3.54)	0.776	No
Coffee intake	38		1.149 (0.575∼2.3)	0.694	No
Dried fruit intake	40		0.734 (0.3367∼1.6)	0.437	No
		0 $5 \cdot 10^{-1}$ 1 1.5 2 2.5 3 3.5 4 4.5 5 5.5 6			

6. Discussion and Future Directions

This section highlights the basic hurdles and emerging opportunities in utilizing the XDL for solving computational neuroscience problems. As XDL tools have recently become widely adopted in both visual and clinical domain applications, carefully assessing where they fall short and where they hold the most promise has become even more crucial.

6.1. Key Challenges in XDL for Neuroscience

While XDL has made significant progress in this domain, a variety of key issues still exist and need to be addressed, as they limit the capabilities and advancement of neuroscience. The issues are as follows:

1. Balancing accuracy and clarity: Some of the most accurate models, like deep CNNs and transformers, and are often difficult to interpret. On the other hand, models that are easier to explain typically do not perform as well as the complex models. Therefore, trading off between these two goals is a major challenge in computational neuroscience, especially when considering the critical clinical decisions. Beyond individual strengths, these XDL methods have some limitations that need to be addressed. These limitations include instability, reproducibility, and clinical adaptability. Computational cost is another big challenge, as high-dimensional neuroscientific imaging data requires improved models for producing better results. The techniques mentioned in this survey are domain-dependent, making the current explainable systems unable to solve a problem in a generalized way as the real-world demands.
2. Missing evaluation standards: Currently, there is no widely accepted approach or method for measuring or validating the ability of a model to explain its outcome. Tools vary in what they aim to show significantly, some focus on pixel-level importance, others on broader concepts, making it hard to evaluate results or reproduce findings across studies.
3. Issues with generalization and real-time use: Many models work well on specific datasets but do not hold up when tested across different groups, imaging types, or use cases. Real-time performance is also a problem, especially in clinical areas like brain-computer interfaces. On top of that, we still do not fully understand how robust these models are to noise or unexpected input changes. Recent studies also show that hybrid anomaly detection strategies can improve robustness by the removal of noise in data and by the regularization of data. This work highlights the importance of resilience in clinical XDL systems [152].
4. Dataset characteristics can introduce biases. These properties can affect the XDL model's interpretability and generalization. The most common issues are class imbalance, local-specific variations, and demographic disparities. To tackle these challenges, various harmonization strategies could be applied. Preprocessing normalization, domain adaptation, class balancing, and multimodal feature standardization are a few strategies that can be applied. Being aware of such strategies and reporting them are both crucial. These steps need to be taken to ensure reliable and generalizable explanations in computational neuroscience. Recent multimodal studies still demand enhancements in several aspects for addressing clinical challenges for alzheimer (MRI and PET) [70], epilepsy (EEG and fMRI) [74], brain tumors (MRI and genomics) [86], parkinson disease (MRI and clinical scores) [128], population densities (MRI and genetics) [83] problem domains. Identified gaps of each multimodal research work mentioned here are discussed, along with their proposed solutions, in Section 6.3.

To visually capture the relationship among these challenges, Figure 8 presents a triangle plot illustrating the trade-offs between the model's three metrics [153,154], i.e., accuracy, complexity, and explainability. This highlights the tensions faced in XDL model design. The closer the value moves towards a corner, the more that the associated metric (be it accuracy, explainability, or complexity) will be maximized. So, overall, moving towards one corner affects the other inversely proportionally. Hence, central regions suggest a compromise zone. In this zone model can moderate the trade-off for accuracy, explainability, and complexity. Altogether, the decision is left to the decision-maker so that the requirement of a case-specific scenario is fulfilled.

6.2. Future Research Directions

To address these limitations, several research directions offer promise for advancing XDL in neuroscience:

1. Integration of biological priors into model architectures: Embedding anatomical, functional, or physiological knowledge (sEMG-based muscle fatigue detection [155]) into DL models can enhance alignment with neural mechanisms and improve both interpretability and generalization. Another innovative advancement is the integration of ultra-high-frequency biosignal retrieval [156], which enhances the systems to show real-time interpretable DL. Further details on biological priors are depicted in Section 6.2.1.
2. Standardized frameworks for XDL evaluation: Developing rigorous benchmarks, grounded in neuroscientific and clinical relevance, will facilitate objective assessment of explainability methods and model comparisons across studies. Recent works that utilize transformer-based clinical emotion classification emphasize the need to incorporate evaluation pipelines to enable trustworthy deployment [157].
3. Human-in-the-loop and neuro-symbolic models: Combining human domain expertise with automated learning systems can improve interpretability and trust. Neuro-symbolic approaches, which integrate symbolic reasoning with neural representations, offer a promising hybrid paradigm.
4. Collaborative and interdisciplinary research models: Progress in XDL will require active collaboration between AI researchers, neuroscientists, clinicians, and ethicists. Such partnerships can accelerate development while ensuring ethical deployment and practical utility.

6.2.1. Integration of Biological Priors in XDL

This Subsection discusses the future directions relevant to the integration of XDL techniques that can opt for biological priors for advancing and enhancing computational neuroscience. The knowledge derived from the structure, function, and organization of the neurons in the brain drives the modeling and design of computational models. This approach also includes the anatomical constraints, neurophysiological attributes, and more. Innovating the DL models to adopt this approach enhances interpretability via ensuring the model design alignment is similar to the neurological principles. This also improves generalization across modalities and increases the plausibility of model outcomes, both in clinical and visual neuroscience applications.

We discuss the necessary techniques that need to be incorporated in order to appropriately opt for the biological priors.

- Architectural constraints: The biological constraints are a key part in modeling neuroscientific models. A recent research focuses on the necessity of integration of biological priors into current DL model designs. Authors propose a brain-constrained modeling approach to incorporate the principles of brain structure internals like synaptic plasticity, local and long-range connectivity, and area-specific compositions into the design of the model [158]. This can lead to significant improvements in current DL techniques to improve the accuracy and explanatory abilities of such models for neuroscience applications.

  Another study investigates DL based networks in which hidden layers correspond to and mimic the structure of a biological entity, i.e., genes, pathways, and their connections, to encode the relationships of brain-mimicking nodes [159].

  Another study emphasizes leveraging the biological implausibility in backpropagation. The updating of errors while training these models is inspired by the actual human behavior when correcting mistakes [160].

- Regularization and loss functions: The enhancements in regularization and loss functions are a necessity for minimizing the prediction errors considering the integration of biological priors in both visual and clinical neuroscience [161]. Experiments in this work show that utilizing the structured loss can improve the MCTs like incremental learning, few-shot learning, and long-tailed classification.
  Another study proposes a model named bio-RNN that is trained with multiple loss terms [162], e.g., neuron-level loss, trial-matching loss, and regularization loss. These losses jointly optimize the updating weights of the model network.
- Data augmentation with priors: Data augmentation improves and extends the existing dataset, and is much needed, as the clinical datasets, specifically, are very limited in size. A study on one such augmentation technique called phylogenetic is applied, which is homologous sequence from related species [163]. This is particularly the best technique for small datasets.
- Architectural Hybrid neuro-symbolic models: Neuro-symbolic models are enabled with the ability to reason [164]. This property is embedded into models by combining NN components with symbolic reasoning components. Reasoning components include logic, rules, planning, high-level intents in hierarchical form.
  Another study presents a hybrid architecture to combine deep neural components. This work emphasizes the integration of pattern recognition extraction components with symbolic reasoning, leveraging the medical ontologies, rule-based inference, and statistics and probabilities [165].
  Another study on enforcing structural priors involves the embedding of symbolic modules in the form of known anatomical circuits and functional hierarchies into the model architecture [166].

The integration of the above-mentioned biological prior techniques is much needed, as they allow combining data-driven feature extraction better in performance and make the clinical systems able to interpret and explain their outcomes clearly and efficiently, due to the nature of such models mimicking neuroscience.

Despite the techniques' promising results, integrating biological priors also presents challenges and remains ever-evolving due to the never-ending upgrades caused by the technological advancements. Over-constraining is another aspect to analyze, as it may lead to reduced flexibility and may compromise the predictive performance. Trading off between biological realism and model explainability is still a challenge that will definitely require important steps to be taken in advance. Future directions in computational neuroscience are never-ending but require the aforementioned proposed techniques, which are comprehensively concluded in the following sections.

6.2.2. Strategic Directions for Advancing XDL

This Subsection discusses the strategic directions for advancing XDL in the domain of computational neuroscience. Table 9 outlines four strategic research directions aimed at addressing current limitations in the use of XDL in computational neuroscience. These directions reflect a cross-cutting emphasis on biological plausibility, evaluation consistency, human-in-the-loop design, and interdisciplinary integration. Each of the directions contributes to the robust application of XDL across visual and clinical neuroscience.

Table 9. Strategic directions for advancing XDL in computational neuroscience.

Research Direction	Objective	Primary Appl. Domain	Key Challenges
Integration of biological priors	Enhance model fidelity by embedding anatomical, physiological, or functional constraints into DL architectures	Visual attention modeling, brain-signal decoding, neural encoding	Balancing realism and flexibility. Limited availability of detailed neural priors
Standardized frameworks for XDL evaluation	Establish consistent benchmarks for assessing interpretability and trustworthiness of XDL models	Clinical diagnostics, regulatory validation, cross-study comparisons	Metric inconsistency. Data heterogeneity. Lack of community consensus
Human-in-the-loop and neuro-symbolic systems	Combine expert knowledge with machine learning to enhance interpretability and domain relevance	Interactive BCIs, explainable clinical decision-support tools	Real-time integration, hybrid model design complexity
Collaborative interdisciplinary research models	Foster synergy among AI, neuroscience, clinical, and ethical domains for responsible innovation	Translational neuroscience, XDL healthcare deployment	Communication barriers. Differing research cultures and standards

6.3. Research Gaps and Proposed Approaches

Table 10 provides a summary of the major research gaps identified in the surveyed literature and highlights the proposed approaches and their potential impacts. Standard evaluation metrics, biologically integrated DL models, transparent and interpretable clinical adoption frameworks, scalable real-time XDL methods, and interpretable visual and clinical neuroscience integration are all comprehensively depicted in this section. The aim is to provide a roadmap to further explore and advance XDL in computational neuroscience.

Table 10. Research gaps, proposed approaches, and their potential impacts.

Identified Gap	Proposed Approach (References)	Potential Impact
Lack of standardized evaluation metrics for XDL (EEG-based BCI) in neuroscience.	Development of unified benchmark datasets and evaluation pipelines, acceleration of benchmarking practice adoption across the field [167].	Ensures reproducibility and comparability across XDL pipelines. Standard reporting evaluation. metrics for BCI/EEG based XDL studies.
Limited integration of biological priors into DL models.	Exploration of knowledge-informed and neurobiological-constrained cognitive systems (e.g., brain-inspired models, neural priors) [168].	Improves biological plausibility and interpretability of models.
Insufficient trust in clinical adoption of XDL models.	Human-in-the-loop systems, clinician-centered interpretability studies, regulatory frameworks to achieve safety and transparency [169].	Builds trust, facilitates clinical integration, and ensures ethical deployment.
Limited real-world applicability (multimodal) [70].	Thorough real-world clinical evaluation with large multi-site longitudinal datasets.	Enhanced clinical decision-making reliability.
Optimal integration challenge [74] (multimodal).	Advanced algorithm development.	Enhanced clinical insights.
Limited interpretability transparency [86] (multimodal).	Integrate explainable AI techniques.	Improves clinical trust and decision-making.
Limited clinical interpretability [128] (multimodal).	Integrate advanced explainable AI techniques for clearer clinical insights.	Improved diagnostic trust and clinical decision-making.
Limited clinical explainability [83] (multimodal).	Integrate tailored imputation and nuanced interpretation	Enhance personalized and effective clinical decisions

Table 10. *Cont.*

Identified Gap	Proposed Approach (References)	Potential Impact
Lack of real-time, scalable XDL methods for large datasets.	Efficient model architectures for multimodal fusion (coupled SSM model) hardware-aware XDL techniques. This approach uses an inter-modal hidden states transition scheme for efficiency, and formulates a global convolution kernel for enabling parallelism [170].	Enables deployment in time-sensitive neuroscience and clinical applications for provisioning real-time and scalable XDL models. Another study demonstrates deployable XDL for portable clinical neuroscience with limited resources, much needed for real-time and scalable systems [171].
Fragmented research between visual and clinical neuroscience.	Interdisciplinary frameworks combining cognitive neuroscience, clinical practice, personalized treatment through and XDL [172].	Encourages holistic, cross-domain advancements for the XDL the neuroscience.

The proposed XDL benchmark framework details are shown in Table 11 for the purpose of promoting reproducibility and standardization. This includes the tasks, datasets (curated), metrics (quantifiable), and mechanisms to govern. Classification, prognosis, and other tasks are depicted to illustrate various domain problems like disease vs. control and disease progression. Benchmark datasets, including the ADNI, UK Biobank, and OpenNeuro, are essential to ensure multimodal and multi-domain coverage. Evaluation metrics, including faithfulness, sensitivity, calibration, and clinical utilization support, are used in quantifying the qualitative metrics, such as interpretability, reliability, and relevance. These metrics are crucial for neuroscience applications. Data sharing standards among the business stakeholders, open code, and checklists built through community participation ensure the transparent comparison across the studies. Overall, this provides the necessary steps given in a structured direction to the visual and clinical application developers and the stakeholders.

Table 11. Proposed XDL benchmark framework.

Component	Purpose	Illustrative Examples
Task	Define evaluation objectives	Classification, prognosis. Benchmarks supervised learning tasks (classification, prognosis) using clinical/biomedical tabular datasets [173]. Evaluates AI benchmarking for predictive modeling (object detection/classification), applicable to clinical differentiation tasks [174].
Dataset	Ensure multimodal coverage	ADNI, UK Biobank, OpenNeuro. Demonstrates use of diverse biomedical datasets for benchmarking and generalization [173]. Uses electrophysiological datasets from multiple sources to ensure multimodal, multi-site evaluations [175].
Metrics	Quantify interpretability	Faithfulness, sensitivity, calibration, and clinical utility. Proposes a benchmarking framework for XAI, focusing on interpretability, faithfulness, and clinical relevance [176]. Benchmarks frameworks with a focus on accuracy, reliability, robustness, including calibration and sensitivity [177].
Governance	Promote reproducibility	Open code/data, reporting standards. Advocates open data/code, reproducible evaluation protocols, and reporting standards in benchmarking studies [173]. Discusses transparency, standardized evaluation, and community-driven checklists for reproducibility and governance [176].

7. Conclusions

The XDL has emerged as a pivotal approach in advancing computational neuroscience, particularly in the domains of visual processing and clinical applications. This review has explored how DL models, especially convolutional neural networks and vision transform-

ers, are being employed to simulate mechanisms of visual attention and categorization, offering computational analogues to brain processes. In visual neuroscience, XDL techniques such as saliency mapping, class activation, and concept-based explanations not only enhance our interpretability of models but also deepen our scientific understanding of visual perception in the brain.

In clinical neuroscience, DL models are being applied to diagnose neurological and psychiatric disorders, forecast patient outcomes, and support precision medicine. However, given the high-stakes nature of clinical decision-making, the need for explainable models is not merely academic but essential for building trust, ensuring regulatory compliance, and facilitating ethical deployment. XDL plays a critical role here by enabling clinicians and researchers to validate model predictions, identify potential biases, and improve transparency in diagnosis and treatment.

Across both domains, the review emphasizes the core limitations and challenges facing XDL, particularly the trade-offs between performance and interpretability, the lack of standardized evaluation frameworks, and difficulties in generalizing across populations and modalities. Future progress lies in embedding biological priors into model architectures, advancing neuro-symbolic and human-in-the-loop approaches, and fostering stronger interdisciplinary collaboration among AI researchers, neuroscientists, and clinicians.

Ultimately, XDL is not just a technical advancement but a paradigm shift towards responsible, transparent, and impactful neuroscience. As the field evolves, interpretable DL models will be foundational to bridging computational tools with scientific discovery and real-world healthcare transformation.

Author Contributions: Conceptualization, A.M.; investigations, A.M. and F.M.; writing—original draft preparation, A.M. and F.M.; writing—review and editing, A.M. and F.M.; resources, J.K.; supervision, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Research Foundation of Korea grant (RS-2024-00419269). This work was also supported by the Gachon University research fund of 2023 (GCU-202303650001).

Data Availability Statement: No new data were created or analyzed in this study.

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

Abbreviations

ACE	Automated Concept-based Explanation
ACER	Average Classification Error Rate
ADNI	Alzheimer's Disease Neuroimaging Initiative
APCER	Attack Presentation Classification Error Rate
BCI	Brain-Computer Interface
BMS	Bayesian Model Selection
BPCER	Bona Fide Presentation Classification Error Rate
BraTS	Brain Tumor Segmentation
BSMMCH	Bangabandhu Sheikh Mujib Medical College Hospital
CAM	Class Activation Map
CAMCAN	Cambridge Centre for Ageing and Neuroscience
CI	Confidence Interval
CNN	Convolutional Neural Network
CORnet-S	Core Object Recognition Network Shallow

DBS	Deep Brain Stimulation
DL	Deep Learning
DR	Diabetic Retinopathy
DTI	Diffusion Tensor Imaging
DVA	Deep Visual Analytics
DWI	Diffusion-Weighted Imaging
ECoG	Electrocorticography
ECSSD	Extended Complex Scene Saliency Dataset
EDF	European Data Format
EEG	Electroencephalography
EHR	Electronic Health Records
EM	Electron Microscopy
EOG	Electrooculography
F-M	F-measure
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional Near-Infrared Spectroscopy
GANs	Generative Adversarial Networks
GBVS	Graph-Based Visual Saliency
HCP	Human Connectome Project
HGC	High-Grade Condition
HKUIS	Hong Kong University Image Saliency
HMA	Hard Exudate Microaneurysms
iCoSeg	Interactive Co-segmentation
IoT	Internet of Things
IP	Image Processing
IT	Inferior Temporal cortex
IVW	Inverse-Variance Weighted
LGC	Low-Grade Condition
LRP	Layer-wise Relevance Propagation
MA	Exudate Microaneurysms
MAE	Mean Absolute Error
MCT	Machine-challenging Task
MDD	Major Depressive Disorder
MEG	Magnetoencephalography
MR	Mendelian Randomization
MRI	Magnetic Resonance Imaging
MSRA-B	Microsoft Research Asia (version B)
NLP	Natural Language Processing
NMM	Normalized Mean value under the annotated Mask
NN	Neural Network
NST	Non-Specific Tumor
NTL	Normal Tissue Label
OASIS	Open Access Series of Imaging Studies
OCTNet	Optical Coherence Tomography Network
PAD	Presentation Attack Detection
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PET	Positron Emission Tomography
REST	Resting-State
PPG	Photoplethysmogram
RNN	Recurrent Neural Network

RSA	Representational Similarity Analysis
SAM	Segment Anything Model
SaMD	Software as a Medical Device
SED	Social Event Detection
sEMG	Surface Electromyography
SNP	Single Nucleotide Polymorphism
SOD	Salient Object Detection
SSM	State Space Model
TCAV	Testing with Concept Activation Vectors
TCGA	The Cancer Genome Atlas
TUH	Temple University Hospital
V1/V4	Visual Cortex Area 1/Visual Cortex Area 4
ViT	Vision Transformer
VAE	Variational Autoencoder
VGG	Visual Geometry Group
XAI	Explainable Artificial Intelligence
XDL	Explainable Deep Learning

References

1. Guimarães, K.; Duarte, A. A Network-Level Stochastic Model for Pacemaker GABAergic Neurons in Substantia Nigra Pars Reticulata. *Mathematics* **2023**, *11*, 3778. [CrossRef]
2. Iwama, S.; Tsuchimoto, S.; Mizuguchi, N.; Ushiba, J. EEG decoding with spatiotemporal convolutional neural network for visualization and closed-loop control of sensorimotor activities: A simultaneous EEG-fMRI study. *Hum. Brain Mapp.* **2024**, *45*, e26767. [CrossRef]
3. Huang, Y.; Huan, Y.; Zou, Z.; Wang, Y.; Gao, X.; Zheng, L. Data-driven natural computational psychophysiology in class. *Cogn. Neurodynamics* **2024**, *18*, 3477–3489. [CrossRef]
4. Elmoznino, E.; Bonner, M.F. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLoS Comput. Biol.* **2024**, *20*, e1011792. [CrossRef]
5. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
6. Watanabe, N.; Miyoshi, K.; Jimura, K.; Shimane, D.; Keerativittayayut, R.; Nakahara, K.; Takeda, M. Multimodal deep neural decoding reveals highly resolved spatiotemporal profile of visual object representation in humans. *NeuroImage* **2023**, *275*, 120164. [CrossRef] [PubMed]
7. Farzmahdi, A.; Zarco, W.; Freiwald, W.A.; Kriegeskorte, N.; Golan, T. Emergence of brain-like mirror-symmetric viewpoint tuning in convolutional neural networks. *eLife* **2024**, *13*, e90256. [CrossRef] [PubMed]
8. Salahuddin, Z.; Woodruff, H.C.; Chatterjee, A.; Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput. Biol. Med.* **2022**, *140*, 105111. [CrossRef]
9. Danks, D.; Davis, I. Causal inference in cognitive neuroscience. *Wiley Interdiscip. Rev. Cogn. Sci.* **2023**, *14*, e1650. [CrossRef]
10. Žlahtič, B.; Završnik, J.; Kokol, P.; Blažun Vošner, H.; Sobotkiewicz, N.; Antolinc Schaubach, B.; Kirbiš, S. Trusting AI made decisions in healthcare by making them explainable. *Sci. Prog.* **2024**, *107*, 00368504241266573. [CrossRef] [PubMed]
11. Freyer, N.; Groß, D.; Lipprandt, M. The ethical requirement of explainability for AI-DSS in healthcare: A systematic review of reasons. *BMC Med. Ethics* **2024**, *25*, 104. [CrossRef]
12. Brima, Y.; Atemkeng, M. Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis. *BioData Min.* **2024**, *17*, 18. [CrossRef] [PubMed]
13. Yang, L.; Qiao, C.; Zhou, H.; Calhoun, V.D.; Stephen, J.M.; Wilson, T.W.; Wang, Y. Explainable multimodal deep dictionary learning to capture developmental differences from three fMRI paradigms. *IEEE Trans. Biomed. Eng.* **2023**, *70*, 2404–2415. [CrossRef]
14. Sattiraju, A.; Ellis, C.A.; Miller, R.L.; Calhoun, V.D. An explainable and robust deep learning approach for automated electroencephalography-based schizophrenia diagnosis. In Proceedings of the 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE), Dayton, OH, USA, 4–6 December 2023; pp. 255–259.
15. Rodríguez Mallma, M.J.; Zuloaga-Rotta, L.; Borja-Rosales, R.; Rodríguez Mallma, J.R.; Vilca-Aguilar, M.; Salas-Ojeda, M.; Mauricio, D. Explainable Machine Learning Models for Brain Diseases: Insights from a Systematic Review. *Neurol. Int.* **2024**, *16*, 1285–1307. [CrossRef]

16. Hussain, T.; Shouno, H. Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping. *Information* **2023**, *14*, 642. [CrossRef]
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
18. Li, Y.; Lou, A.; Xu, Z.; Zhang, S.; Wang, S.; Englot, D.; Kolouri, S.; Moyer, D.; Bayrak, R.; Chang, C. Neurobolt: Resting-state eeg-to-fmri synthesis with multi-dimensional feature mapping. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 23378–23405.
19. Song, L.; Ren, Y.; Xu, S.; Hou, Y.; He, X. A hybrid spatiotemporal deep belief network and sparse representation-based framework reveals multilevel core functional components in decoding multitask fMRI signals. *Netw. Neurosci.* **2023**, *7*, 1513–1532. [CrossRef] [PubMed]
20. Hu, W.; Meng, X.; Bai, Y.; Zhang, A.; Qu, G.; Cai, B.; Zhang, G.; Wilson, T.W.; Stephen, J.M.; Calhoun, V.D.; et al. Interpretable multimodal fusion networks reveal mechanisms of brain cognition. *IEEE Trans. Med. Imaging* **2021**, *40*, 1474–1483. [CrossRef] [PubMed]
21. Ji, Y.; Yang, C.; Liang, Y. A multiview deep learning method for brain functional connectivity classification. *Comput. Intell. Neurosci.* **2022**, *2022*, 5782569. [CrossRef]
22. Germann, J.; Zadeh, G.; Mansouri, A.; Kucharczyk, W.; Lozano, A.M.; Boutet, A. Untapped neuroimaging tools for neuro-oncology: Connectomics and spatial transcriptomics. *Cancers* **2022**, *14*, 464. [CrossRef]
23. Luo, X.; Zeng, Z.; Zheng, S.; Chen, J.; Jannin, P. Statistical Multiscore Functional Atlas Creation for Image-Guided Deep Brain Stimulation. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2025**, *33*, 818–828. [CrossRef]
24. Li, R. Integrative diagnosis of psychiatric conditions using ChatGPT and fMRI data. *BMC Psychiatry* **2025**, *25*, 145. [CrossRef]
25. Khan, S.; Kallis, L.; Mee, H.; El Hadwe, S.; Barone, D.; Hutchinson, P.; Kolias, A. Invasive Brain–Computer Interface for Communication: A Scoping Review. *Brain Sci.* **2025**, *15*, 336. [CrossRef]
26. Baniqued, P.D.E.; Stanyer, E.C.; Awais, M.; Alazmani, A.; Jackson, A.E.; Mon-Williams, M.A.; Mushtaq, F.; Holt, R.J. Brain–computer interface robotics for hand rehabilitation after stroke: A systematic review. *J. Neuroeng. Rehabil.* **2021**, *18*, 1–25. [CrossRef]
27. Zhang, X.; Zhang, T.; Jiang, Y.; Zhang, W.; Lu, Z.; Wang, Y.; Tao, Q. A novel brain-controlled prosthetic hand method integrating AR-SSVEP augmentation, asynchronous control, and machine vision assistance. *Heliyon* **2024**, *10*, e26521. [CrossRef] [PubMed]
28. Berezutskaya, J.; Freudenburg, Z.V.; Vansteensel, M.J.; Aarnoutse, E.J.; Ramsey, N.F.; van Gerven, M.A. Direct speech reconstruction from sensorimotor brain activity with optimized deep learning models. *J. Neural Eng.* **2023**, *20*, 056010. [CrossRef]
29. Lai, T. Interpretable medical imagery diagnosis with self-attentive transformers: A review of explainable AI for health care. *BioMedInformatics* **2024**, *4*, 113–126. [CrossRef]
30. Sinz, F.H.; Pitkow, X.; Reimer, J.; Bethge, M.; Tolias, A.S. Engineering a less artificial intelligence. *Neuron* **2019**, *103*, 967–979. [CrossRef] [PubMed]
31. Stringer, C.; Pachitariu, M.; Steinmetz, N.; Carandini, M.; Harris, K.D. High-dimensional geometry of population responses in visual cortex. *Nature* **2019**, *571*, 361–365. [CrossRef]
32. Räuker, T.; Ho, A.; Casper, S.; Hadfield-Menell, D. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In Proceedings of the 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SATML), Raleigh, NC, USA, 8–10 February 2023; pp. 464–483.
33. Retzlaff, C.O.; Angerschmid, A.; Saranti, A.; Schneeberger, D.; Roettger, R.; Mueller, H.; Holzinger, A. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cogn. Syst. Res.* **2024**, *86*, 101243. [CrossRef]
34. Salih, A.M.; Galazzo, I.B.; Gkontra, P.; Rauseo, E.; Lee, A.M.; Lekadir, K.; Radeva, P.; Petersen, S.E.; Menegaz, G. A review of evaluation approaches for explainable AI with applications in cardiology. *Artif. Intell. Rev.* **2024**, *57*, 240. [CrossRef]
35. Liu, Y.; Ge, E.; Kang, Z.; Qiang, N.; Liu, T.; Ge, B. Spatial-temporal convolutional attention for discovering and characterizing functional brain networks in task fMRI. *NeuroImage* **2024**, *287*, 120519. [CrossRef]
36. El Shawi, R. Conceptglassbox: Guided concept-based explanation for deep neural networks. *Cogn. Comput.* **2024**, *16*, 2660–2673. [CrossRef]
37. Schwalbe, G.; Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Min. Knowl. Discov.* **2024**, *38*, 3043–3101. [CrossRef]
38. Wang, Z.; Huang, C.; Li, Y.; Yao, X. Multi-objective feature attribution explanation for explainable machine learning. *ACM Trans. Evol. Learn. Optim.* **2024**, *4*, 1–32. [CrossRef]
39. Lee, S.; Lee, K.S. Predictive and Explainable Artificial Intelligence for Neuroimaging Applications. *Diagnostics* **2024**, *14*, 2394. [CrossRef] [PubMed]
40. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int. J. Surg.* **2010**, *8*, 336–341. [CrossRef] [PubMed]
41. Kemp, B.; Zwinderman, A.H.; Tuk, B.; Kamphuisen, H.A.; Oberye, J.J. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Trans. Biomed. Eng.* **2000**, *47*, 1185–1194. [CrossRef]

42. LaMontagne, P.J.; Benzinger, T.L.; Morris, J.C.; Keefe, S.; Hornbeck, R.; Xiong, C.; Grant, E.; Hassenstab, J.; Moulder, K.; Vlassenko, A.G.; et al. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv* **2019**. [CrossRef]
43. Markiewicz, C.J.; Gorgolewski, K.J.; Feingold, F.; Blair, R.; Halchenko, Y.O.; Miller, E.; Hardcastle, N.; Wexler, J.; Esteban, O.; Goncavles, M.; et al. The OpenNeuro resource for sharing of neuroscience data. *eLife* **2021**, *10*, e71774. [CrossRef]
44. Mehmood, A.; Ko, J.; Kim, H.; Kim, J. Optimizing Image Enhancement: Feature Engineering for Improved Classification in AI-Assisted Artificial Retinas. *Sensors* **2024**, *24*, 2678. [CrossRef] [PubMed]
45. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, 159–174. [CrossRef]
46. Lai-Dang, Q.V.; Kang, T.; Son, S. Adaptive Transformer Programs: Bridging the Gap Between Performance and Interpretability in Transformers. In Proceedings of the The Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.
47. Dayan, P.; Abbott, L.F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*; MIT Press: Cambridge, MA, USA, 2005; Volume 15, pp. 563–577.
48. Koch, C.; Segev, I. *Methods in Neuronal Modeling: From Synapses to Networks*; Computational Neuroscience; MIT Press: Cambridge, MA, USA, 1988; p. 524.
49. Moshruba, A.; Poursiami, H.; Parsa, M. Izhikevich-Inspired Temporal Dynamics for Enhancing Privacy, Efficiency, and Transferability in Spiking Neural Networks. *arXiv* **2025**, arXiv:2505.04034. [CrossRef]
50. Wojcik, G.M.; Masiak, J.; Kawiak, A.; Kwasniewicz, L.; Schneider, P.; Polak, N.; Gajos-Balinska, A. Mapping the human brain in frequency band analysis of brain cortex electroencephalographic activity for selected psychiatric disorders. *Front. Neuroinform.* **2018**, *12*, 73. [CrossRef]
51. Yang, Y.; Luo, S.; Wang, W.; Gao, X.; Yao, X.; Wu, T. From bench to bedside: Overview of magnetoencephalography in basic principle, signal processing, source localization and clinical applications. *Neuroimage Clin.* **2024**, *42*, 103608. [CrossRef]
52. Kaiju, T.; Doi, K.; Yokota, M.; Watanabe, K.; Inoue, M.; Ando, H.; Takahashi, K.; Yoshida, F.; Hirata, M.; Suzuki, T. High spatiotemporal resolution ECoG recording of somatosensory evoked potentials with flexible micro-electrode arrays. *Front. Neural Circuits* **2017**, *11*, 20. [CrossRef]
53. Sadad, T.; Aurangzeb, R.A.; Safran, M.; Imran.; Alfarhood, S.; Kim, J. Classification of highly divergent viruses from DNA/RNA sequence using transformer-based models. *Biomedicines* **2023**, *11*, 1323. [CrossRef]
54. Abrol, A.; Fu, Z.; Salman, M.; Silva, R.; Du, Y.; Plis, S.; Calhoun, V. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. Commun.* **2021**, *12*, 353. [CrossRef] [PubMed]
55. Storrs, K.R.; Kriegeskorte, N. Deep learning for cognitive neuroscience. *arXiv* **2019**, arXiv:1903.01458. [CrossRef]
56. Livezey, J.A.; Glaser, J.I. Deep learning approaches for neural decoding across architectures and recording modalities. *Briefings Bioinform.* **2021**, *22*, 1577–1591. [CrossRef]
57. Jack, C.R., Jr.; Bernstein, M.A.; Fox, N.C.; Thompson, P.; Alexander, G.; Harvey, D.; Borowski, B.; Britson, P.J.; L. Whitwell, J.; Ward, C.; et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging Off. J. Int. Soc. Magn. Reson. Med.* **2008**, *27*, 685–691. [CrossRef]
58. Mehmood, F.; Mehmood, A.; Whangbo, T.K. Alzheimer's Disease Detection in Various Brain Anatomies Based on Optimized Vision Transformer. *Mathematics* **2025**, *13*, 1927. [CrossRef]
59. Rehman, A.; Yi, M.K.; Majeed, A.; Hwang, S.O. Early diagnosis of alzheimer's disease using 18f-fdg pet with soften latent representation. *IEEE Access* **2024**, *12*, 87923–87933. [CrossRef]
60. Makropoulos, A.; Robinson, E.C.; Schuh, A.; Wright, R.; Fitzgibbon, S.; Bozek, J.; Counsell, S.J.; Steinweg, J.; Vecchiato, K.; Passerat-Palmbach, J.; et al. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *Neuroimage* **2018**, *173*, 88–112. [CrossRef]
61. Liao, W.; Wang, W. Eegencoder: Advancing bci with transformer-based motor imagery classification. *arXiv* **2024**, arXiv:2404.14869. [CrossRef]
62. Shahwar, T.; Rehman, A.U. Automated detection of brain disease using quantum machine learning. In *Brain-Computer Interfaces*; Elsevier: Amsterdam, The Netherlands, 2025; pp. 91–114.
63. Ishfaq, Q.U.A.; Bibi, R.; Ali, A.; Jamil, F.; Saeed, Y.; Alnashwan, R.O.; Chelloug, S.A.; Muthanna, M.S.A. Automatic smart brain tumor classification and prediction system using deep learning. *Sci. Rep.* **2025**, *15*, 14876. [CrossRef] [PubMed]
64. Hashmi, S.; Lugo, J.; Elsayed, A.; Saggurthi, D.; Elseiagy, M.; Nurkamal, A.; Walia, J.; Maani, F.A.; Yaqub, M. Optimizing brain tumor segmentation with mednext: Brats 2024 ssa and pediatrics. *arXiv* **2024**, arXiv:2411.15872.
65. Abbas, K.; Khan, P.W.; Ahmed, K.T.; Song, W.C. Automatic brain tumor detection in medical imaging using machine learning. In Proceedings of the 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 16–18 October 2019; pp. 531–536.
66. Bhutta, M.A.M.; Kim, S.; Kim, Y.J. Dual Photoplethysmography Pulse Wave Velocity-Based Blood Pressure Prediction for Wearable Healthcare Internet of Things. *IEEE Internet Things J.* **2025**, *12*, 20773–20786. [CrossRef]

67. Ali, S.T.A.; Kim, S.; Kim, Y.J. Towards Reliable ECG Analysis: Addressing Validation Gaps in the Electrocardiographic R-Peak Detection. *Appl. Sci.* **2024**, *14*, 10078. [CrossRef]
68. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J.; et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **2018**, *562*, 203–209. [CrossRef]
69. Landau, S.M.; Harrison, T.M.; Baker, S.L.; Boswell, M.S.; Lee, J.; Taggett, J.; Ward, T.J.; Chadwick, T.; Murphy, A.; DeCarli, C.; et al. Positron emission tomography harmonization in the Alzheimer's Disease Neuroimaging Initiative: A scalable and rigorous approach to multisite amyloid and tau quantification. *Alzheimers Dement.* **2025**, *21*, e14378. [CrossRef]
70. Abdelaziz, M.; Wang, T.; Anwaar, W.; Elazab, A. Multi-scale multimodal deep learning framework for Alzheimer's disease diagnosis. *Comput. Biol. Med.* **2025**, *184*, 109438. [CrossRef]
71. Matsushita, S.; Tatekawa, H.; Ueda, D.; Takita, H.; Horiuchi, D.; Tsukamoto, T.; Shimono, T.; Miki, Y. The Association of Metabolic Brain MRI, amyloid PET, and clinical factors: A study of Alzheimer's disease and Normal controls from the open access series of imaging studies dataset. *J. Magn. Reson. Imaging* **2024**, *59*, 1341–1348. [CrossRef]
72. Van Essen, D.C.; Smith, S.M.; Barch, D.M.; Behrens, T.E.; Yacoub, E.; Ugurbil, K.; WU-Minn HCP Consortium. The WU-Minn human connectome project: An overview. *Neuroimage* **2013**, *80*, 62–79. [CrossRef]
73. Germani, E.; Fromont, E.; Maurel, P.; Maumet, C. The HCP multi-pipeline dataset: An opportunity to investigate analytical variability in fMRI data analysis. *arXiv* **2023**, arXiv:2312.14493.
74. Warbrick, T. Simultaneous EEG-fMRI: What have we learned and what does the future hold? *Sensors* **2022**, *22*, 2262. [CrossRef]
75. Shafto, M.A.; Tyler, L.K.; Dixon, M.; Taylor, J.R.; Rowe, J.B.; Cusack, R.; Calder, A.J.; Marslen-Wilson, W.D.; Duncan, J.; Dalgleish, T.; et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: A cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol.* **2014**, *14*, 204. [CrossRef]
76. Taylor, J.R.; Williams, N.; Cusack, R.; Auer, T.; Shafto, M.A.; Dixon, M.; Tyler, L.K.; Henson, R.N. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage* **2017**, *144*, 262–269. [CrossRef] [PubMed]
77. Tangermann, M.; Müller, K.R.; Aertsen, A.; Birbaumer, N.; Braun, C.; Brunner, C.; Leeb, R.; Mehring, C.; Miller, K.J.; Müller-Putz, G.R.; et al. Review of the BCI competition IV. *Front. Neurosci.* **2012**, *6*, 55. [CrossRef] [PubMed]
78. Zhang, H.; Guan, C.; Ang, K.K.; Wang, C.; Chin, Z.Y. BCI competition IV–data set I: Learning discriminative patterns for self-paced EEG-based motor imagery detection. *Front. Neurosci.* **2012**, *6*, 7.
79. Obeid, I.; Picone, J. The temple university hospital EEG data corpus. *Front. Neurosci.* **2016**, *10*, 196. [CrossRef]
80. Mellot, A.; Collas, A.; Rodrigues, P.L.; Engemann, D.; Gramfort, A. Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling. *Imaging Neurosci.* **2023**, *1*, 1–23. [CrossRef]
81. Ujma, P.P.; Dresler, M.; Bódizs, R. Comparing Manual and Automatic Artifact Detection in Sleep EEG Recordings. *Psychophysiology* **2025**, *62*, e70016. [CrossRef] [PubMed]
82. Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **2015**, *12*, e1001779. [CrossRef] [PubMed]
83. Mirabnahrazam, G.; Ma, D.; Lee, S.; Popuri, K.; Lee, H.; Cao, J.; Wang, L.; Galvin, J.E.; Beg, M.F.; Initiative, A.D.N. Machine learning based multimodal neuroimaging genomics dementia score for predicting future conversion to alzheimer's disease. *J. Alzheimers Dis.* **2022**, *87*, 1345–1365. [CrossRef]
84. Starck, S.; Sideri-Lampretsa, V.; Ritter, J.J.; Zimmer, V.A.; Braren, R.; Mueller, T.T.; Rueckert, D. Using uk biobank data to establish population-specific atlases from whole body mri. *Commun. Med.* **2024**, *4*, 237. [CrossRef]
85. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2014**, *34*, 1993–2024. [CrossRef]
86. Lyu, Q.; Costen, F. A twin-tower model using MRI and gene for prediction on brain tumor patients' response to therapy. *Bioinform. Adv.* **2025**, *5*, vbaf041. [CrossRef] [PubMed]
87. Sidume, F.; Soula, O.; Wacira, J.M.; Zhu, Y.; Muhammad, A.R.; Zeraii, A.; Kalejaye, O.; Ibrahim, H.; Gaddour, O.; Halubanza, B.; et al. Resource-Efficient Glioma Segmentation on Sub-Saharan MRI. *arXiv* **2025**, arXiv:2509.09469.
88. Bhati, D.; Neha, F.; Amiruzzaman, M.; Guercio, A.; Shukla, D.K.; Ward, B. Neural network interpretability with layer-wise relevance propagation: Novel techniques for neuron selection and visualization. In Proceedings of the 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2025; pp. 441–447.
89. Nicolson, A.; Schut, L.; Noble, J.A.; Gal, Y. Explaining Explainability: Recommendations for Effective Use of Concept Activation Vectors. *arXiv* **2024**, arXiv:2404.03713.
90. Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.R. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2660–2673. [CrossRef]
91. Ramachandram, D.; Joshi, H.; Zhu, J.; Gandhi, D.; Hartman, L.; Raval, A. Transparent AI: The Case for Interpretability and Explainability. *arXiv* **2025**, arXiv:2507.23535. [CrossRef]

92. Mehmood, F.; Rehman, S.U.; Choi, A. Vision-AQ: Explainable Multi-Modal Deep Learning for Air Pollution Classification in Smart Cities. *Mathematics* **2025**, *13*, 3017. [CrossRef]
93. Rehman, S.U.; Yasin, A.U.; Ul Haq, E.; Ali, M.; Kim, J.; Mehmood, A. Enhancing human activity recognition through integrated multimodal analysis: A focus on RGB imaging, skeletal tracking, and pose estimation. *Sensors* **2024**, *24*, 4646. [CrossRef] [PubMed]
94. Park, E.S.; Kim, S.; Mujtaba, G.; Ryu, E.S. Detection of Power Transmission Equipment in Image using Guided Grad-CAM. In Proceedings of the 2020 Summer Conference of the Korean Society of Broadcast and Media Engineers (KIBME), Online, 13–15 July 2020; pp. 709–713.
95. Mi, X.; Zou, B.; Zou, F.; Hu, J. Permutation-based identification of important biomarkers for complex diseases via machine learning models. *Nat. Commun.* **2021**, *12*, 3008. [CrossRef]
96. Ramaswamy, V.V.; Kim, S.S.; Fong, R.; Russakovsky, O. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 10932–10941.
97. Dai, Z.; Gifford, D.K. Ablation based counterfactuals. *arXiv* **2024**, arXiv:2406.07908. [CrossRef]
98. Saranya, A.; Subhashini, R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decis. Anal. J.* **2023**, *7*, 100230. [CrossRef]
99. Fellous, J.M.; Sapiro, G.; Rossi, A.; Mayberg, H.; Ferrante, M. Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. *Front. Neurosci.* **2019**, *13*, 1346. [CrossRef]
100. Mahmood, S.; Teo, C.; Sim, J.; Zhang, W.; Muyun, J.; Bhuvana, R.; Teo, K.; Yeo, T.T.; Lu, J.; Gulyas, B.; et al. The application of eXplainable artificial intelligence in studying cognition: A scoping review. *Ibrain* **2024**, *10*, 245–265. [CrossRef]
101. Sathyan, A.; Weinberg, A.I.; Cohen, K. Interpretable AI for bio-medical applications. *Complex Eng. Syst.* **2022**, *2*, 18. [CrossRef]
102. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
103. Warren, S.L.; Moustafa, A.A. Functional magnetic resonance imaging, deep learning, and Alzheimer's disease: A systematic review. *J. Neuroimaging* **2023**, *33*, 5–18. [CrossRef]
104. Avberšek, L.K.; Repovš, G. Deep learning in neuroimaging data analysis: Applications, challenges, and solutions. *Front. Neuroimaging* **2022**, *1*, 981642. [CrossRef] [PubMed]
105. Feng, B.; Yu, T.; Wang, H.; Liu, K.; Wu, W.; Long, W. Machine learning and deep learning in biomedical signal analysis. *Front. Hum. Neurosci.* **2023**, *17*, 1183840. [CrossRef] [PubMed]
106. Cui, J.; Yuan, L.; Wang, Z.; Li, R.; Jiang, T. Towards best practice of interpreting deep learning models for EEG-based brain computer interfaces. *Front. Comput. Neurosci.* **2023**, *17*, 1232925. [CrossRef]
107. Mathi, S.; Deb, D.; Chaudhuri, A.K.; Joseph, L.; Biswas, S. A Review Of Convolutional Neural Networks For Medical Image Analysis: Trends And Future Directions. *J. Neonatal Surg.* **2025**, *14*. [CrossRef]
108. Chang, S.J.; Lee, J.Y.; Kim, S.I.; Lee, S.Y.; Lee, D.W. Simple Wearable Eye Tracker With Mini-Infrared Point Sensors. *IEEE Access* **2024**, *12*, 41019–41026. [CrossRef]
109. Khaki, A.M.Z.; Choi, A. A real-time integrated eye tracker with in-pixel image processing in 0.18-μm CMOS technology. *Integration* **2025**, *105*, 102526. [CrossRef]
110. Kuen, J.; Wang, Z.; Wang, G. Recurrent Attentional Networks for Saliency Detection. *arXiv* **2016**, arXiv:1604.03227. [CrossRef]
111. Li, G.; Yu, Y. Visual Saliency Based on Multiscale Deep Features. *arXiv* **2015**, arXiv:1503.08663. [CrossRef]
112. Boyd, A.; Bowyer, K.; Czajka, A. Human-Aided Saliency Maps Improve Generalization of Deep Learning. *arXiv* **2021**, arXiv:2105.03492. [CrossRef]
113. He, S.; Tavakoli, H.R.; Borji, A.; Mi, Y.; Pugeault, N. Understanding and Visualizing Deep Visual Saliency Models. *arXiv* **2019**, arXiv:1903.02501. [CrossRef]
114. Dobs, K.; Martinez, J.; Kell, A.J.; Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. Adv.* **2022**, *8*, eabl8913. [CrossRef]
115. Chefer, H.; Gur, S.; Wolf, L. Transformer interpretability beyond attention visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 782–791.
116. Golan, T.; Raju, P.C.; Kriegeskorte, N. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 29330–29337. [CrossRef] [PubMed]
117. Khosla, M.; Ngo, G.; Jamison, K.; Kuceyeski, A.; Sabuncu, M. Neural encoding with visual attention. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 15942–15953.
118. Kubilius, J.; Schrimpf, M.; Kar, K.; Rajalingham, R.; Hong, H.; Majaj, N.; Issa, E.; Bashivan, P.; Prescott-Roy, J.; Schmidt, K.; et al. Brain-like object recognition with high-performing shallow recurrent ANNs. *arXiv* **2019**, arXiv:1909.06161. [CrossRef]
119. Shahriar, T. Comparative Analysis of Lightweight Deep Learning Models for Memory-Constrained Devices. *arXiv* **2025**, arXiv:2505.03303. [CrossRef]

120. Cichy, R.M.; Khosla, A.; Pantazis, D.; Torralba, A.; Oliva, A. Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. *arXiv* **2016**, arXiv:1601.02970. [CrossRef]
121. Alain, G.; Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv* **2016**, arXiv:1610.01644.
122. Iqbal, S.; Qureshi, A.N.; Aurangzeb, K.; Alhussein, M.; Wang, S.; Anwar, M.S.; Khan, F. Hybrid parallel fuzzy CNN paradigm: Unmasking intricacies for accurate brain MRI insights. *IEEE Trans. Fuzzy Syst.* **2024**, *32*, 5533–5544. [CrossRef]
123. Berrich, Y.; Guennoun, Z. EEG-based epilepsy detection using CNN-SVM and DNN-SVM with feature dimensionality reduction by PCA. *Sci. Rep.* **2025**, *15*, 14313. [CrossRef]
124. Jafarbeglou, F.; Ahmadieh, H.; Soleimani, F.; Karimi, A.; Daftarian, N.; Fekri, S.; Motevasseli, T.; Naderan, M.; Kamali Doust Azad, B.; Sheikhtaheri, A.; et al. A deep learning model for diagnosis of inherited retinal diseases. *Sci. Rep.* **2025**, *15*, 22523. [CrossRef]
125. Noda, Y.; Sakaue, K.; Wada, M.; Takano, M.; Nakajima, S. Development of artificial intelligence for determining major depressive disorder based on resting-state EEG and single-pulse transcranial magnetic stimulation-evoked EEG indices. *J. Pers. Med.* **2024**, *14*, 101. [CrossRef]
126. Liaqat, H.; Parveen, A.; Kim, S.Y. Neuroprotective natural products' regulatory effects on depression via gut–brain axis targeting tryptophan. *Nutrients* **2022**, *14*, 3270. [CrossRef]
127. Gallucci, A.; Varoli, E.; Del Mauro, L.; Hassan, G.; Rovida, M.; Comanducci, A.; Casarotto, S.; Lo Re, V.; Romero Lauro, L.J. Multimodal approaches supporting the diagnosis, prognosis and investigation of neural correlates of disorders of consciousness: A systematic review. *Eur. J. Neurosci.* **2024**, *59*, 874–933. [CrossRef]
128. Dentamaro, V.; Impedovo, D.; Musti, L.; Pirlo, G.; Taurisano, P. Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: Insights from the PPMI database. *Sci. Rep.* **2024**, *14*, 20941. [CrossRef]
129. Rahman, A.U.; Ali, S.; Saqia, B.; Halim, Z.; Al-Khasawneh, M.; AlHammadi, D.A.; Khan, M.Z.; Ullah, I.; Alharbi, M. Alzheimer's disease prediction using 3D-CNNs: Intelligent processing of neuroimaging data. *SLAS Technol.* **2025**, *32*, 100265. [CrossRef] [PubMed]
130. Huh, Y.J.; Park, J.H.; Kim, Y.J.; Kim, K.G. Ensemble Learning-Based Alzheimer's Disease Classification Using Electroencephalogram Signals and Clock Drawing Test Images. *Sensors* **2025**, *25*, 2881. [CrossRef] [PubMed]
131. Yazdan, S.A.; Ahmad, R.; Iqbal, N.; Rizwan, A.; Khan, A.N.; Kim, D.H. An efficient multi-scale convolutional neural network based multi-class brain MRI classification for SaMD. *Tomography* **2022**, *8*, 1905–1927. [CrossRef] [PubMed]
132. Shen, M.; Yang, F.; Wen, P.; Song, B.; Li, Y. A real-time epilepsy seizure detection approach based on EEG using short-time Fourier transform and Google-Net convolutional neural network. *Heliyon* **2024**, *10*, e31827. [CrossRef]
133. Conte, G.M.; Weston, A.D.; Vogelsang, D.C.; Philbrick, K.A.; Cai, J.C.; Barbera, M.; Sanvito, F.; Lachance, D.H.; Jenkins, R.B.; Tobin, W.O.; et al. Generative adversarial networks to synthesize missing T1 and FLAIR MRI sequences for use in a multisequence brain tumor segmentation model. *Radiology* **2021**, *299*, 313–323. [CrossRef] [PubMed]
134. Khan, M.A.; Jang, J.H.; Iqbal, N.; Jamil, H.; Naqvi, S.S.A.; Khan, S.; Kim, J.C.; Kim, D.H. Enhancing patient rehabilitation predictions with a hybrid anomaly detection model: Density-based clustering and interquartile range methods. *CAAI Trans. Intell. Technol.* **2025**, *10*, 983–1006. [CrossRef]
135. Paek, S.H.; Cho, Z.H.; Son, Y.D.; Kwon, D.H.; Park, S.B. Magnetic Resonance Imaging-Based Alternating Magnetic Field Treatment System. U.S. Patent Application No. 19/052,776, 5 June 2025.
136. Liu, Z.; Si, L.; Shi, S.; Li, J.; Zhu, J.; Lee, W.H.; Lo, S.L.; Yan, X.; Chen, B.; Fu, F.; et al. Classification of three anesthesia stages based on near-infrared spectroscopy signals. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 5270–5279. [CrossRef]
137. Mehmood, F.; Mumtaz, N.; Mehmood, A. Next-Generation Tools for Patient Care and Rehabilitation: A Review of Modern Innovations. *Actuators* **2025**, *14*, 133. [CrossRef]
138. Ahmed, M.; Dar, A.R.; Helfert, M.; Khan, A.; Kim, J. Data provenance in healthcare: Approaches, challenges, and future directions. *Sensors* **2023**, *23*, 6495. [CrossRef] [PubMed]
139. Khalil, I.; Mehmood, A.; Kim, H.; Kim, J. OCTNet: A modified multi-scale attention feature fusion network with InceptionV3 for retinal OCT image classification. *Mathematics* **2024**, *12*, 3003. [CrossRef]
140. Zhu, M.; Quan, Y.; He, X. The classification of brain network for major depressive disorder patients based on deep graph convolutional neural network. *Front. Hum. Neurosci.* **2023**, *17*, 1094592. [CrossRef]
141. Jiang, H.; Xu, J.; Shi, R.; Yang, K.; Zhang, D.; Gao, M.; Ma, H.; Qian, W. A multi-label deep learning model with interpretable Grad-CAM for diabetic retinopathy classification. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 1560–1563.
142. Yang, J.; Wang, G.; Xiao, X.; Bao, M.; Tian, G. Explainable ensemble learning method for OCT detection with transfer learning. *PLoS ONE* **2024**, *19*, e0296175. [CrossRef] [PubMed]
143. Lundervold, A.S.; Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z. Med. Phys.* **2019**, *29*, 102–127. [CrossRef]
144. Yu, A.C.; Mohajer, B.; Eng, J. External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiol. Artif. Intell.* **2022**, *4*, e210064. [CrossRef] [PubMed]

145. Faghani, S.; Moassefi, M.; Rouzrokh, P.; Khosravi, B.; Baffour, F.I.; Ringler, M.D.; Erickson, B.J. Quantifying uncertainty in deep learning of radiologic images. *Radiology* **2023**, *308*, e222217. [CrossRef]
146. Roy, Y.; Banville, H.; Albuquerque, I.; Gramfort, A.; Falk, T.H.; Faubert, J. Deep learning-based electroencephalography analysis: A systematic review. *J. Neural Eng.* **2019**, *16*, 051001. [CrossRef]
147. Ahmed, M.M.; Hossain, M.M.; Islam, M.R.; Ali, M.S.; Nafi, A.A.N.; Ahmed, M.F.; Ahmed, K.M.; Miah, M.S.; Rahman, M.M.; Niu, M.; et al. Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh. *Sci. Rep.* **2024**, *14*, 22797. [CrossRef]
148. Jiang, R.; Yin, X.; Yang, P.; Cheng, L.; Hu, J.; Yang, J.; Wang, Y.; Fu, X.; Shang, L.; Li, L.; et al. A transformer-based weakly supervised computational pathology method for clinical-grade diagnosis and molecular marker discovery of gliomas. *Nat. Mach. Intell.* **2024**, *6*, 876–891. [CrossRef]
149. Evans, I.D.; Sharpley, C.F.; Bitsika, V.; Vessey, K.A.; Williams, R.J.; Jesulola, E.; Agnew, L.L. Differences in EEG Functional Connectivity in the Dorsal and Ventral Attentional and Salience Networks Across Multiple Subtypes of Depression. *Appl. Sci.* **2025**, *15*, 1459. [CrossRef]
150. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In Proceedings of the International Conference on Machine Learning. PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2668–2677.
151. Li, L.; Li, J.; Chang, M.; Wu, X.; Yin, Z.; Chen, Z.; Wang, Z. The Causal Relationship Between Dietary Factors and the Risk of Intracranial Aneurysms: A Mendelian Randomization Analysis. *Biomedicines* **2025**, *13*, 533. [CrossRef] [PubMed]
152. Khan, M.A.; Iqbal, N.; Jamil, H.; Qayyum, F.; Jang, J.H.; Khan, S.; Kim, J.C.; Kim, D.H. Enhanced abnormal data detection hybrid strategy based on heuristic and stochastic approaches for efficient patients rehabilitation. *Future Gener. Comput. Syst.* **2024**, *154*, 101–122. [CrossRef]
153. Schug, D.; Yerramreddy, S.; Caruana, R.; Greenberg, C.; Zwolak, J. Explainable classification techniques for quantum dot device measurements. In Proceedings of the XAI4Sci: Explainable Machine Learning for Sciences Workshop (AAAI 2024), Vancouver, CA, Canada, 26 February 2024; pp. 1–6.
154. Agarwal, C.; Queen, O.; Lakkaraju, H.; Zitnik, M. Evaluating explainability for graph neural networks. *Sci. Data* **2023**, *10*, 144. [CrossRef] [PubMed]
155. Ou, J.; Li, N.; He, H.; He, J.; Zhang, L.; Jiang, N. Detecting muscle fatigue among community-dwelling senior adults with shape features of the probability density function of sEMG. *J. Neuroeng. Rehabil.* **2024**, *21*, 196. [CrossRef]
156. Aïssa, B.; Sinopoli, A.; Ali, A.; Zakaria, Y.; Zekri, A.; Helal, M.; Nedil, M.; Rosei, F.; Mansour, S.; Mahmoud, K. Nanoelectromagnetic of a highly conductive 2D transition metal carbide (MXene)/Graphene nanoplatelets composite in the EHF M-band frequency. *Carbon* **2021**, *173*, 528–539. [CrossRef]
157. Kumar, M.; Khan, L.; Choi, A. RAMHA: A Hybrid Social Text-Based Transformer with Adapter for Mental Health Emotion Classification. *Mathematics* **2025**, *13*, 2918. [CrossRef]
158. Pulvermüller, F.; Tomasello, R.; Henningsen-Schomers, M.R.; Wennekers, T. Biological constraints on neural network models of cognitive function. *Nat. Rev. Neurosci.* **2021**, *22*, 488–502. [CrossRef]
159. Esser-Skala, W.; Fortelny, N. Reliable interpretability of biology-inspired deep neural networks. *NPJ Syst. Biol. Appl.* **2023**, *9*, 50. [CrossRef]
160. Ororbia, A.; Mali, A.; Kohan, A.; Millidge, B.; Salvatori, T. A review of neuroscience-inspired machine learning. *arXiv* **2024**, arXiv:2403.18929. [CrossRef]
161. Gütlin, D.; Auksztulewicz, R. Predictive Coding algorithms induce brain-like responses in Artificial Neural Networks. *bioRxiv* **2025**. [CrossRef]
162. Sourmpis, C.; Petersen, C.C.; Gerstner, W.; Bellec, G. Biologically informed cortical models predict optogenetic perturbations. *bioRxiv* **2024**. [CrossRef]
163. Duncan, A.G.; Mitchell, J.A.; Moses, A.M. Improving the performance of supervised deep learning for regulatory genomics using phylogenetic augmentation. *Bioinformatics* **2024**, *40*, btae190. [CrossRef]
164. Oltramari, A.; Francis, J.; Henson, C.; Ma, K.; Wickramarachchi, R. Neuro-symbolic architectures for context understanding. *arXiv* **2020**, arXiv:2003.04707. [CrossRef]
165. Gandhirajan, S. Neurosymbolic Approaches for Knowledge-Infused Clinical Reasoning in Explainable Artificial Intelligence for Differential Diagnosis. *J. Asian Sci. Res. (JASR)* **2025**, *15*, 24–30.
166. Kiruluta, A. A Novel Architecture for Symbolic Reasoning with Decision Trees and LLM Agents. *arXiv* **2025**, arXiv:2508.05311. [CrossRef]
167. Chevallier, S.; Carrara, I.; Aristimunha, B.; Guetschel, P.; Sedlar, S.; Lopes, B.; Velut, S.; Khazem, S.; Moreau, T. The largest EEG-based BCI reproducibility study for open science: The MOABB benchmark. *arXiv* **2024**, arXiv:2404.15319. [CrossRef]
168. Mumuni, F.; Mumuni, A. Improving deep learning with prior knowledge and cognitive models: A survey on enhancing explainability, adversarial robustness and zero-shot learning. *Cogn. Syst. Res.* **2024**, *84*, 101188. [CrossRef]

169. Nasarian, E.; Alizadehsani, R.; Acharya, U.R.; Tsui, K.L. Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Inf. Fusion* **2024**, *108*, 102412. [CrossRef]
170. Li, W.; Zhou, H.; Yu, J.; Song, Z.; Yang, W. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 59808–59832.
171. Mehmood, F.; Ahmad, S.; Whangbo, T.K. Object detection based on deep learning techniques in resource-constrained environment for healthcare industry. In Proceedings of the 2022 International Conference on Electronics, Information, and Communication (ICEIC), Jeju, Republic of Korea, 6–9 February 2022; pp. 1–5.
172. Lundervold, A.J. Precision neuropsychology in the area of AI. *Front. Psychol.* **2025**, *16*, 1537368. [CrossRef]
173. Shmuel, A.; Glickman, O.; Lazebnik, T. A comprehensive benchmark of machine and deep learning across diverse tabular datasets. *arXiv* **2024**, arXiv:2408.14817. [CrossRef]
174. Cantero, D.; Esnaola-Gonzalez, I.; Miguel-Alonso, J.; Jauregi, E. Benchmarking object detection deep learning models in embedded devices. *Sensors* **2022**, *22*, 4205. [CrossRef]
175. Jiao, M.; Xian, X.; Wang, B.; Zhang, Y.; Yang, S.; Chen, S.; Sun, H.; Liu, F. XDL-ESI: Electrophysiological sources imaging via explainable deep learning framework with validation on simultaneous EEG and IEEG. *Neuroimage* **2024**, *299*, 120802. [CrossRef] [PubMed]
176. Canha, D.; Kubler, S.; Främling, K.; Fagherazzi, G. A Functionally-Grounded Benchmark Framework for XAI Methods: Insights and Foundations from a Systematic Literature Review. *ACM Comput. Surv.* **2025**, *57*, 1–40. [CrossRef]
177. Pochelu, P. Deep learning inference frameworks benchmark. *arXiv* **2022**, arXiv:2210.04323. [CrossRef]