Bioengineering

AIMS Bioengineering, 12(1): 69–89.
DOI: 10.3934/bioeng.2025004
Received: 10 October 2024
Revised: 07 December 2025
Accepted: 07 February 2025
Published: 12 February 2025

https://www.aimspress.com/journal/Bioengineering

Research article

How artificial intelligence reduces human bias in diagnostics?

Artur Luczak*

Canadian Centre for Behavioural Neuroscience, University of Lethbridge, AB, Canada

* **Correspondence:** Email: Luczak@uleth.ca.

Abstract: Accurate diagnostics of neurological disorders often rely on behavioral assessments, yet traditional methods rooted in manual observations and scoring are labor-intensive, subjective, and prone to human bias. Artificial Intelligence (AI), particularly Deep Neural Networks (DNNs), offers transformative potential to overcome these limitations by automating behavioral analyses and reducing biases in diagnostic practices. DNNs excel in processing complex, high-dimensional data, allowing for the detection of subtle behavioral patterns critical for diagnosing neurological disorders such as Parkinson's disease, strokes, or spinal cord injuries. This review explores how AI-driven approaches can mitigate observer biases, thereby emphasizing the use of explainable DNNs to enhance objectivity in diagnostics. Explainable AI techniques enable the identification of which features in data are used by DNNs to make decisions. In a data-driven manner, this allows one to uncover novel insights that may elude human experts. For instance, explainable DNN techniques have revealed previously unnoticed diagnostic markers, such as posture changes, which can enhance the sensitivity of behavioral diagnostic assessments. Furthermore, by providing interpretable outputs, explainable DNNs build trust in AI-driven systems and support the development of unbiased, evidence-based diagnostic tools. In addition, this review discusses challenges such as data quality, model interpretability, and ethical considerations. By illustrating the role of AI in reshaping diagnostic methods, this paper highlights its potential to revolutionize clinical practices, thus paving the way for more objective and reliable assessments of neurological disorders.

Keywords: movement disorders diagnosis; bias reduction in healthcare; automated behavioral scoring; data-driven diagnostics; interpretable machine learning

1. Introduction

The diagnosis of neurological disorders often hinges on the precise assessment of behavioral patterns, particularly in conditions where motor functions are affected. Traditional diagnostic methods heavily rely on manual observation and scoring, requiring trained professionals to interpret behaviors based on predefined criteria [1,2]. While these methods have been foundational in clinical practice, they are inherently labor-intensive and susceptible to human biases and variability, which can compromise the accuracy [3,4]. The subjective nature of manual assessments may lead to inconsistencies, thus potentially affecting the treatment decisions and patient outcomes [5].

Advancements in artificial intelligence (AI), and more specifically deep neural networks (DNNs), offer promising solutions to these challenges. AI algorithms are adept at the detection of subtle data patterns that may be overlooked by human observers [6]. In the context of neurological diagnostics, AI-driven tools can automate the analysis of motor behaviors, thus providing objective and consistent assessments that reduce the impact of observer bias [7,8]. The ability of DNNs to learn from large datasets allows for an improved predictive accuracy in diagnosing conditions such as Parkinson's disease, strokes, and spinal cord injuries, and in monitoring rehabilitation progress [9].

This review makes the following contributions: (1) it provides an in-depth overview of the transformative role of AI in reducing human bias within neurological diagnostics, emphasizing its potential to automate and standardize behavioral assessments; (2) by examining how AI technologies, particularly explainable DNNs, enhance objectivity, the review highlights their ability to uncover novel, unbiased diagnostic markers, such as subtle postural changes, that are often overlooked by human experts; (3) we present key studies and advancements demonstrating AI's superior diagnostic precision and reliability in conditions like Parkinson's disease, stroke, and spinal cord injuries; (4) this paper also discusses critical challenges in implementing AI models, including concerns about the data quality, interpretability, and generalizability across diverse populations, while providing insights into overcoming these barriers; and (5) furthermore, ethical considerations surrounding AI adoption, such as transparency, accountability for errors, and patient data privacy, are addressed to ensure a responsible integration into clinical settings. Through this comprehensive analysis, we aim to demonstrate how AI is not only reshaping diagnostic practices, but is also contributing to better, more equitable patient outcomes.

2. Advantages of DNNs contributing to reducing diagnostic biases

With their multi-layer architecture, DNNs are particularly effective in analyzing complex, high-dimensional data, which is crucial for accurate diagnostics in neurological research [10]. These networks are especially suited for behavioral tests used in diagnosing movement disorders, where human bias and error can significantly influence results. By automatically extracting features from raw behavioral data, DNNs eliminate the need for manual intervention, thus reducing the subjective interpretation and enhancing the diagnostic objectivity [7,11].

The application of DNNs in behavioral analyses is driven by their ability to automatically extract and learn features from raw data without the need for manual intervention. This is particularly relevant in the context of neurological research, where behavioral data can be highly complex. For instance, in the study of Parkinson's disease, DNNs have been used to analyze brain signals and motor functions, thus providing insights that were previously unattainable using traditional methods [12–15]. The ability of DNNs to handle such intricate data makes them a valuable tool to understand the behavioral manifestations of neurological disorders [7].

There is a consensus about the efficacy of DNNs in behavioral analyses [8]. For example, the unsupervised behavior analysis (uBAM) framework, which employs DNNs, has been shown to detect subtle behavioral differences in animal models of neurological diseases that are often missed by traditional methods [11]. This study highlights the potential of DNNs to uncover nuanced behavioral patterns, thereby providing a deeper understanding of the underlying neurological conditions.

Traditional machine learning models typically require manual feature extraction, a process that can be both time-consuming and prone to human error. In contrast, DNNs are capable of automatically learning and extracting relevant features from raw data, which significantly enhances the accuracy and efficiency of the behavioral analysis [16,17]. This capability is particularly beneficial in the context of neurological disorders, where behavioral data can be intricate and multifaceted [18,19]. For instance, in the diagnosis and prognosis of Autism Spectrum Disorder (ASD), DNNs have been shown to outperform traditional methods by automating the feature extraction and improving the diagnostic accuracy [20].

Another significant advantage of DNNs is their scalability [21]. As the volume of behavioral data in neurological research continues to grow, the ability of DNNs to scale and process large datasets becomes increasingly important [22]. Traditional methods often struggle with the sheer volume of data, thus leading to potential bottlenecks in analysis. However, DNNs are designed to efficiently scale with the data, making them particularly well-suited for large-data studies [23–25]. This scalability is evident in studies such as the one that used LFP-Net, where DNNs were used to analyze extensive brain signals from Parkinson's disease patients, thus providing valuable insights into motor function and disease progression [26].

Due to the above listed advantage, DNNs have been shown to reduce biases and improve the accuracy of behavioral scoring, which is a critical aspect for neurological research [26,27]. Traditional scoring methods are often subjective, which can lead to inconsistencies in data interpretation [6]. On the other hand, DNNs can provide more objective and consistent approaches to behavioral scoring, thus reducing the likelihood of human error [28]. Studies have demonstrated that DNNs can achieve a higher accuracy in behavioral analyses compared to traditional methods, as seen in the classification of major depressive disorder [29]. Moreover, the meta-analysis on the prevalence and diagnosis of neurological disorders using deep learning techniques has highlighted the effectiveness of DNNs compared to traditional methods, further supporting their use in this field [30,31]. Thus, the ability of DNNs to enhance both the accuracy and consistency of behavioral analyses makes them a valuable tool in neurological research [8,27].

3. Applications of DNNs in neurological diagnostics to increase accuracy and objectivity

The application of DNNs in the diagnosis of neurological disorders has revolutionized how we assess motor impairments and other functional deficits. By analyzing data from diagnostic tests and patient assessments, including those that simulate conditions such as Parkinson's disease, strokes, and spinal cord injuries, DNNs offer critical insights into the progression of neurological disorders and the effectiveness of various interventions. DNNs have proven invaluable in automating and enhancing the assessment of motor functions, predicting neurological impairments, and monitoring recovery during rehabilitation. This section explores the specific applications of DNNs in these diagnostic areas, thereby highlighting key studies that have demonstrated their potential to replace traditional methods and improve the accuracy and objectivity of diagnostic assessments.

One of the most significant applications of DNNs in neurological diagnostics is the assessment of motor function. Traditionally, motor function has been evaluated through manual scoring, a process that is both time-consuming and prone to human error, as mentioned earlier. However, DNNs have emerged as a powerful tool to automate and enhance this diagnostic process, thus providing more accurate and reliable assessments. For example, in the diagnosis of Parkinson's disease, DNNs have been used to analyze gait patterns and tremors, thus offering a detailed, objective evaluation of motor impairments [32]. These models have been shown to outperform traditional scoring methods, particularly in detecting subtle motor deficits that may be missed by manual observation [33]. Additionally, in stroke and spinal cord injury cases, DNNs have been employed to assess motor recovery, thus providing a continuous and precise evaluation of motor function over time, which is essential for accurate diagnostics and tracking a patient's progress [34].

The accuracy and reliability of DNNs in motor function assessment have been highlighted in several studies. For instance, a study on spinal cord injuries demonstrated that DNNs could accurately classify different stages of motor recovery, thus providing a more nuanced understanding of the rehabilitation process compared to traditional methods [35]. These findings underscore the potential of DNNs to not only replace manual scoring, but also enhance the precision of motor function assessments, ultimately leading to better insights into the progression of neurological disorders.

Moreover, DNNs have been increasingly applied to predict the onset and progression of neurological impairments based on behavioral data. For example, DNNs have been used to analyze early behavioral changes in animal models of Parkinson's disease, such as altered movement patterns, that precede the clinical manifestation of motor symptoms [36]. These models have demonstrated a superior accuracy in predicting disease onset compared to traditional machine learning methods, making them a promising tool for early diagnosis [37].

In addition to predicting the disease onset, DNNs have been employed to forecast the progression of neurological impairments over time. For instance, in models of Alzheimer's disease, DNNs have been used to predict cognitive decline based on behavioral data, thus offering the potential for early intervention before significant impairments occur [27]. In this work, the authors used a DNN with two fully connected hidden layers (100 and 40 nodes) and two output nodes for genotype classification (two genotype classes). The predictive capabilities of such models are crucial, as they allow for timely therapeutic interventions that may slow or even halt the progression of neurological disorders. The ability of DNNs to integrate and analyze large datasets, including behavioral, genetic, and imaging data, makes them particularly well-suited for this task [38].

Another critical application of DNNs in neurological diagnostics is their use in tracking recovery during rehabilitation. In clinical settings following strokes or spinal cord injury, DNNs have been used to monitor the patient's recovery progress by continuously analyzing the motor behavior over time [35]. These models offer a more detailed and continuous assessment of recovery compared to traditional methods, which often rely on periodic manual evaluations. By providing real-time feedback on motor functions, DNNs can help optimize the rehabilitation protocols, thus ensuring that interventions are tailored to the individual needs of each patient [33].

One of the key advantages of using DNNs in rehabilitation tracking is their ability to personalize treatment plans [39]. For example, DNNs are frequently used analyze motor function data and adjust rehabilitation exercises based on the specific progress of each animal [40]. This personalized approach not only improves the efficacy of rehabilitation, but also reduces the required time for recovery. Furthermore, DNNs have been shown to detect subtle improvements in motor function that may not be apparent through manual observation, thus providing a more comprehensive understanding of the recovery process [41].

Additionally, DNNs have significantly advanced other fields, such as oncology, where DNN models have been developed to predict the status of epidermal growth factor receptor (EGFR) mutations in non-small cell lung cancer (NSCLC) patients. A systematic review and meta-analysis evaluated the performance of AI algorithms and found that DNNs achieved a higher predictive accuracy compared to conventional machine learning approaches, highlighting the potential of DNNs in non-invasive cancer diagnostics [42]. Similarly, in pharmacology, DNNs have been instrumental in predicting drug-drug interactions (DDIs), which are crucial for patient safety and effective treatment planning. Traditional methods often require extensive known DDI information, which is scarce for emerging drugs. To address this, Zhang et al. [43] developed graph neural network (GNN) approaches. This study introduced EmerGNN, a flow-based GNN that leverages biomedical networks to predict interactions for emerging drugs. By extracting paths between drug pairs and incorporating relevant biomedical concepts, EmerGNN demonstrated a higher accuracy in predicting DDIs compared to the existing methods [44]. These advancements underscore the transformative role of AI in enhancing diagnostic precision and drug safety across various medical disciplines.

It is also important to notice that other machine learning (ML) algorithms have shown promising results in medical diagnostics, often matching or surpassing human performance. For instance, for diabetes prediction, a hybrid k-means-PCA model combined with Random Forest achieved a 95.2% accuracy [45]. In liver disease diagnosis, Random Forest outperformed other methods with a 98.14% accuracy [46]. Moreover, Random Forest and Naive Bayes models achieved high accuracy levels when applied to classify many disease datasets such as diabetes, heart disease, and cancer [47]. Various ML algorithms, including Support Vector Machines, Gradient Boosting Machines, Random Forest, Naïve Bayes, and K-Nearest Neighborhood, have been applied in mental health diagnostics [48]. The k-Nearest Neighbors (kNN) algorithm has also proven successful in diagnostics such as cervical cancer prediction, achieving remarkable results with an accuracy of 0.9941, a precision of 0.98, a recall of 0.96, and an F1 score of 0.97 [49]. Adding to these developments, a recent study proposed an ML-based framework for the early prediction of acute coronary syndrome (ACS), leveraging Gradient

Boosting Machines, Deep Forest, and Logistic Regression [50]. This approach achieved an accuracy of 94.5% and demonstrated a superior performance across metrics such as precision, recall, and F1-score when compared to traditional methods. Furthermore, it employed a Shapley Additive Explanations (SHAP) analysis to ensure transparency, thus providing interpretative insights into the risk factor significance for personalized patient care. These studies highlight the potential of a wide range of ML algorithms in enhancing medical diagnoses and reducing human bias in healthcare settings [48].

The above examples show that AI have proven to be a transformative tool in the diagnosis of neurological disorders, particularly in the assessment of motor function, the prediction of neurological impairments, and the monitoring of recovery during rehabilitation. By automating and enhancing these diagnostic processes, AI offers the potential for more accurate, reliable, and personalized approaches to diagnosing and managing neurological disorders. AI may also help in the future by improving the analysis accuracy of other neuroscience datasets, such as electrophysiological recordings [51–53], histology [54], or cognitive processes [55]. As research in DNNs field continues to advance, it is likely that AI will play an increasingly central role in both the diagnosis and treatment of neurological conditions, ultimately improving patient outcomes and enhancing the precision of medical care.

4. Explainable DNNs for improving diagnostics

While DNNs offer an unprecedented accuracy in predicting and classifying behaviors, their “black-box” nature often limits the interpretability of the results [56]. However, by employing explainability techniques such as the Layer-wise Relevance Propagation (LRP) [57,58], it is possible to uncover the specific features or behaviors that contribute most to the diagnostic decisions made by DNNs [41,59]. This approach not only enhances our understanding of the neural mechanisms underlying specific behaviors, but also provides actionable insights that can guide further experimental research and potentially lead to earlier and more accurate diagnoses of neurological impairments. Explainable models can be pivotal in developing new diagnostic tests that are more sensitive and specific, as they provide a deeper understanding of how certain behaviors are linked to neurological impairments [38,60]. For example, by employing techniques such as the LRP mentioned above, the decision-making process of DNNs can be deconstructed, thus identifying the specific features that contribute to their predictions [61].

The ability to extract and interpret the features driving DNN predictions is particularly valuable in neuroscience, where understanding the biological basis of behavior is paramount [62]. In recent years, explainable DNNs have uncovered new behavioral components, such as posture and movement initiation patterns, that are strongly associated with specific neurological impairments [41,63]. These findings offer more than just a classification of behavioral data; they provide a deeper understanding of how certain behaviors are linked to neural processes, potentially guiding the development of more targeted interventions [64]. Moreover, the insights gained from explainable DNN models can inform future experimental designs, thus leading to more refined hypotheses and a better understanding of complex brain-behavior relationships [65]. As such, explainable AI stands at the forefront of a new era in neuroscience research, where machine learning models do not just predict outcomes, but also enhance our comprehension of the brain’s intricate workings [66].

For example, Ryait et al., [41] used the LRP method to find to which parts of video frames the DNN was paying the most attention to classify healthy vs stroke animals. Videos were recorded from parallel-beam-walking task, where fine inaccuracies in the paw placement and paw slips were counted to provide a measure of movement impairment after stroke (Figure 1). Similarly to human experts, the DNN paid the most attention to food slips to identify stroke animals. However, for the control animals, the network mostly focused on the rat posture (Figure 1A), rather than details of the foot placements. This was interesting because the human experts did not notice this before the importance of posture in evaluating the recovery from strokes. Thus, the use of an explainable network allowed them to discover novel behavioral features to help evaluate stroke deficits in a more unbiased and data-driven way.

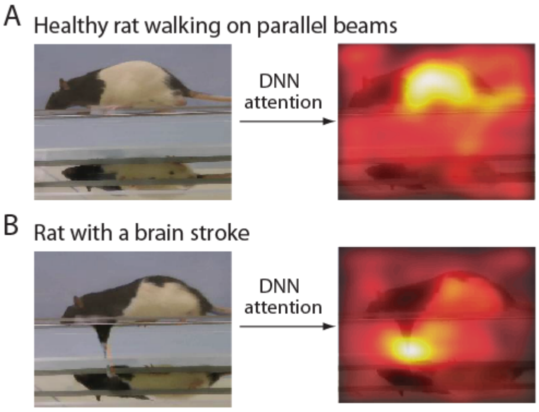

Figure 1. Explainable DNN can help to identify behavioral deficits in an unbiased and data-driven way. (A) Sample video frame showing a rat on the parallel-beam-walking task. Note the mirror below the rat is showing an additional view of paw placement. (B) Sample frame of animal with a brain stroke. Arrows point to "attention" maps superimposed on frames: parts of frames most informative for a network decision (marked in lighter colors). It shows that similarly to experts, the network uses foot slips to score stroke deficits, but it also discovered that body posture is important to identify healthy animals (modified from [41]).

In another example, Ryait et al. [41] used the same explainable DNN to analyze motor behavior in rats while performing a skilled reaching task (Figure 2). One of the key findings was that the DNN scores were more strongly correlated with the stroke lesion volume than the human expert scores ($R_{DNN} = 0.78$, $p = 0.0003$ vs $R_{Human\ Expert} = 0.6$, $p = 0.015$), thus showing the model's ability to discover the most informative movement patterns autonomously. For instance, while experts traditionally scored individual movement components such as limb lift, pronation, and grasp, the DNN identified additional behavioral cues - such as reaching for food with the mouth instead of the hand. These network-discovered patterns were robustly associated with stroke severity and provided a finer-grained understanding of impairment. The study also illustrated that the explainable DNN allowed for the visualization of movements that were most informative for decision-making. This highlights the potential for explainable DNNs to complement human scoring in clinical and research settings, thus providing a scalable, objective, and interpretable approach to diagnosing and tracking neurological impairments.

The DNN architecture used in this study combined a convolutional neural network (ConvNet) and a recurrent neural network (RNN) to effectively capture both spatial and temporal aspects of the animals' behavior. First, a pre-trained ConvNet (Inception-V3) extracted 2048 high-level features from each video frame, capturing spatial information related to posture and movement. Then, these features were processed sequentially by a long short-term memory (LSTM) layer in the RNN, which analyzed temporal patterns across frames. This enabled them to identify the most informative movements for distinguishing between healthy and animals with neurological deficits.

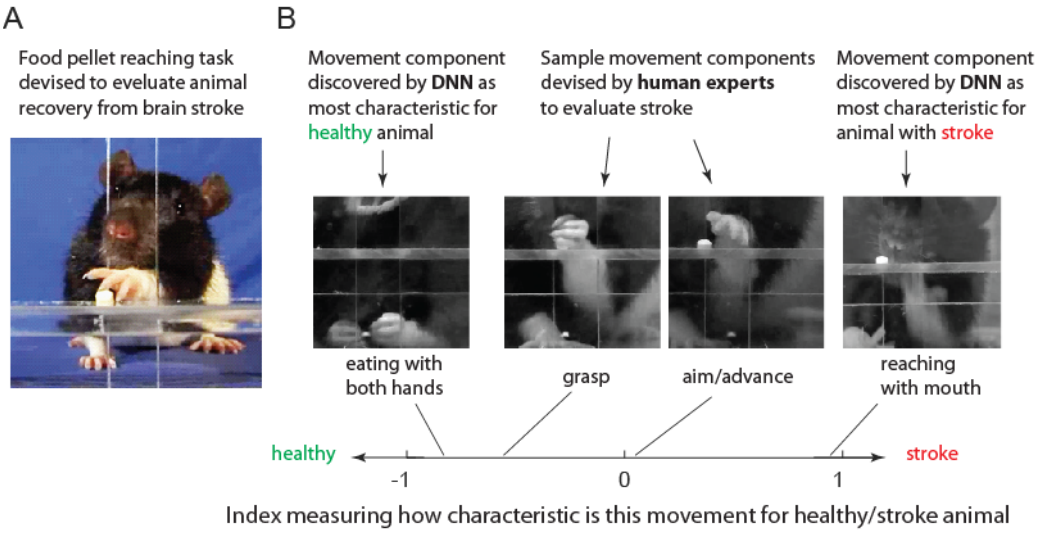
A
Food pellet reaching task devised to eveluate animal recovery from brain stroke
B
Movement component discovered by DNN as most characteristic for healthy animal
Sample movement components devised by human experts to evaluate stroke
Movement component discovered by DNN as most characteristic for animal with stroke
eating with both hands
grasp
aim/advance
reaching with mouth
healthy
stroke
-1
0
1
Index measuring how characteristic is this movement for healthy/stroke animal

Figure 2. Reaching for a food pellet: an example of how explainable DNNs can reduce human bias in diagnostic of stroke-induced movement impairments. In humans, 80% of strokes affect hand use [67]; therefore, quantifying hand use in animal models of stroke provides an important information for evaluating the effectiveness of different therapies [68]. (A) A rat reaching through an opening to retrieve a sucrose pellet placed on a shelf attached to a Plexiglas cage (image courtesy of IQ Wishaw). (B) Video frames illustrating different movement components during reaching for food task (modified from [41]).

Traditionally, human experts assess the quality of reaching by evaluating multiple movement sub-components, such as advancing the hand towards the aim and grasping [69,70]. However, using an explainable convolutional DNN trained on 1500 videos of reaching trials allowed them to discover that other movement components were more informative about stroke impairment. Specifically, the DNN found that the act of eating food with both hands after the end of each trial, and not the reaching movement itself, was the most predictive of healthy rat. This finding indicated that the stroke-affected rats had a significant difficulty with coordinating hand movements. Moreover, the trials typically discarded by human experts, where the rat attempted to "cheat" by reaching with its mouth instead of its hand, turned out to be the most predictive of stroke impairment. These examples demonstrate that DNNs using a data-driven approach can identify movement components that are more informative about impairments than traditional human-designed scoring systems, which may be biased by preconceived notions about which movements are most significant. This is an important point because behavioral tests are widely used to diagnose various neurological disorders, and incorporating AI-driven approaches could enhance the accuracy and reduce biases in human-designed metrics. (Figure modified from [41]).

Using a similar approach, Torabi et al. [59] investigated the use of explainable DNNs to enhance the detection of neurodevelopmental effects in rat pups whose moms were exposed to nicotine. Traditional human-designed metrics rely on counting the number of squares crossed by a rat after being placed in the center of the cage (Figure 3). However, such measures of walked distance have a limited accuracy, achieving a classification rate of around 64% between the control and the nicotine-exposed animals. In contrast, the DNN achieved a much higher classification accuracy of 87% by analyzing video recordings without relying on predefined metrics. Remarkably, the DNN was able to make an accurate classification from just the first frame of each video. Using the LRP method to gain insight into the DNN's decision-making process revealed that the healthy and nicotine-exposed pups had different postures when placed in the center of the cage. The control pups exhibited a more collected posture with limbs under the body, while the nicotine-exposed pups showed extended limbs that indicated problems with balance, a difference that was not previously described in the literature. These findings underscore the DNN's capacity to autonomously discover subtle movement characteristics that correlate with the exposure history, thus presenting an objective, data-driven approach for assessing motor development and reducing the risk of human biases in diagnostics.

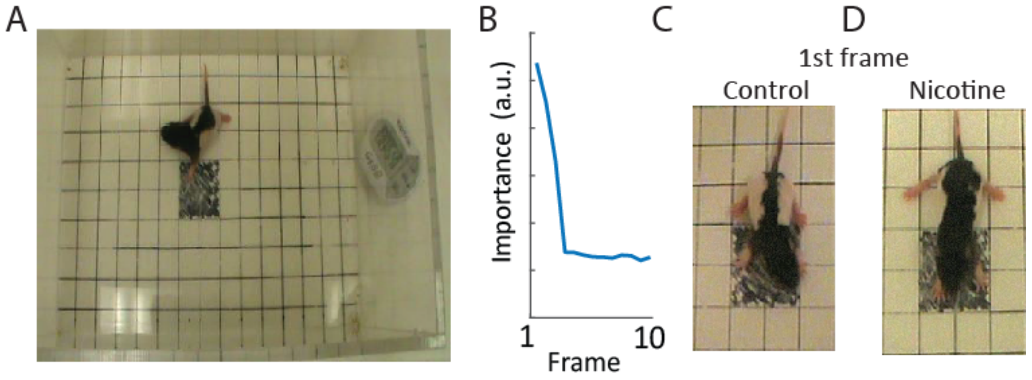

Figure 3. Explainable DNN discovering behavioral deficits in the open field walking task. (A) A rat pup is placed in the center of the cage, and the distance it walks (number of crossed squares) is measured by a human observer to assess neurodevelopmental deficits. (B) Explainable DNN trained to discriminate videos of control vs nicotine-exposed pups discovers that the most informative is the first frame in each video. This is opposed to expert expectations where later frames should be more related to the distance covered by pups. However, a closer examination of initial frames revealed that the starting postures (in the 1st frame) of the control and the nicotine animals significantly differ. The healthy (control) animals had legs close to the body (C). In contrast, the nicotine group had widely extended legs indicating problems with balance (D). This postural difference was missed by human experts and was discovered in a data-driven way only by explainable DNN. This approach allows for identifying more accurate and unbiased measures for behavioral tests.

5. Challenges and limitations in the application of deep neural networks in diagnostics

The application of DNNs in behavioral analyses represents a significant advancement in neuroscience research, thus offering new avenues for understanding and diagnosing neurological disorders [6,23]. However, this approach is not without its challenges and limitations. This section discusses the primary obstacles that researchers face when applying DNNs to behavioral data, including issues related to the data quality and availability, the model interpretability, the generalization across models and species, and the ethical considerations inherent in using AI for this purpose.

5.1. Model interpretability

The "black box" nature of DNNs is a well-known limitation, particularly in fields such as neuroscience, where understanding the underlying mechanisms is as important as the predictive accuracy of the models [71,72]. DNNs, especially deep architectures, often operate in ways that are not easily interpretable, making it difficult for researchers to understand how the model is making its decisions. This lack of transparency can be a significant barrier to the adoption of DNNs in behavioral analyses, as it undermines trust in the model's outputs and limits the insights that can be drawn about the neural processes being studied [65,73]. In response to this challenge and as described in the section above, the field of explainable AI (XAI) has emerged, offering tools and techniques designed to make DNNs more interpretable [74,75]. Layer-wise Relevance Propagation (LRP) is one such method that has been successfully applied in neuroscience [76,77]. LRP helps to deconstruct the decision-making process of a DNN by highlighting the contributions of individual features to the model's output, thereby providing insights into which aspects of the behavioral data are most influential in the model's predictions [78]. Studies that utilized the LRP and similar methods have demonstrated that it is possible to gain meaningful interpretations from DNNs, thereby revealing new patterns and relationships in the data that were previously obscured [79,80]. These advancements in interpretability are critical for making DNNs more accessible and valuable in neuroscience research [81].

5.2. Generalization across models and species

Another significant challenge in applying DNNs to behavioral analyses is the issue of generalization. DNNs trained on data from specific animal models or experimental conditions may not generalize well to other models, species, or even slightly different experimental setups [82,83]. This limitation is particularly concerning when attempting to translate the findings from animal models to human conditions, as differences in behavior and neural architecture can lead to discrepancies in the model performance. Research has shown mixed results in this area. Some studies have successfully developed DNNs that generalize well across different species, particularly when the behaviors being analyzed are highly conserved across species [84,85]. However, other studies have highlighted the limitations of this approach, showing that models often require fine-tuning or retraining on new datasets to adequately perform in different contexts [6]. These findings underscore the need for continued research into the factors that influence generalization and the development of strategies to improve the transferability of DNNs across different models and species.

5.3. Ethical considerations

The integration of AI into healthcare introduces several ethical challenges that necessitate careful consideration to ensure patient safety, trust, and equitable care:

Transparency in AI Decision-Making: AI systems, particularly those that utilize DNNs, often function as "black boxes", making it difficult to understand how they arrive at specific decisions. This opacity can hinder a clinicians' ability to trust and effectively utilize AI recommendations. Developing explainable AI models as described in the previous section can provide clear insights into their decision-making processes, thereby enhancing the transparency and trust in clinical settings [86].

Accountability for Errors: Determining responsibility when AI systems provide errors is complex. If an AI system provides incorrect advice leading to patient harm, it raises questions about whether the fault lies with the developers, the healthcare providers, or the institution deploying the technology. Thus, there should be introduced clear guidelines delineating the accountability in AI-assisted healthcare to address this issue [87].

Privacy of Patient Data: AI systems require vast amounts of data, often including sensitive patient information. Ensuring the privacy and security of this data is paramount. Concerns regarding data anonymization and the potential for re-identification, robust data protection measures, and compliance with regulations should be developed [88].

Addressing these ethical considerations is crucial for the responsible adoption of AI in healthcare, thus ensuring that technological advancements translate into improved patient outcomes without compromising ethical standards.

5.4. Data quality and availability

One of the most significant challenges in applying DNNs to develop new diagnostic methods is the need for large, high-quality datasets. DNNs require vast amounts of data to train effectively, and the quality of this data directly impacts the model's performance. In neuroscience, collecting such datasets is particularly challenging due to the complexity and variability of the behavioral data, as well as the labor-intensive nature of manual data collection and the ethical constraints on using large number of experimental animals [89,90]. Low-quality data—whether due to noise, inconsistencies, or insufficient sample sizes—can lead to a poor model performance and unreliable results [91]. Several studies have highlighted these challenges and proposed strategies to address them. For example, researchers have developed data augmentation techniques that artificially expand the size of available datasets by generating modified versions of existing data, such as through transformations or simulations. This approach can help to improve the robustness of DNNs, as seen in studies where augmented data has led to a better generalization and a higher accuracy of predictions [92–94]. Additionally, efforts to standardize data collection protocols across laboratories and animal models have been crucial in ensuring that datasets are more consistent and reliable, as demonstrated by initiatives that create large, shared databases of annotated behavioral data [95].

5.5. *Insights from distributed and adaptive control systems*

Incorporating insights from distributed and adaptive control systems can also offer useful strategies for mitigating human biases in the diagnosis of neurological disorders. Just as distributed control architectures enable robots to operate flexibly and autonomously in complex environments [96], decentralized AI frameworks could help ensure that the diagnostic models are not overly reliant on one narrow data source or clinical perspective. By learning from diverse patient populations and adapting their parameters in response to new information, these AI-driven diagnostic tools could better capture the multifaceted nature of neurological conditions and reduce the reliance on potentially biased heuristics. Applying data-driven learning methods similar to those used in advanced control systems, such as adaptive parameter tuning and continuous data integration [97], can strengthen the AI's ability to generalize across different clinical scenarios and populations. This, in turn, can promote a more robust decision-making process that is less susceptible to the subjective influences that often challenge human clinicians.

Furthermore, adopting the principles behind distributed real-time control architectures could help seamlessly integrate explainable AI models into clinical workflows. Just as modular and decentralized controllers enable humanoid robots to balance real-time responsiveness with adaptive learning, compartmentalizing the diagnostic AI into interpretable sub-components could make the decision-making process more transparent, traceable, and fair. These sub-components could be individually analyzed, validated, and updated to address the potential biases while still operating within an integrated system that ensures timely and accurate diagnoses. By drawing on the concepts of scalability, adaptability, and modularity from distributed control systems, the development of explainable and bias-mitigating AI tools could foster greater trust and adoption in clinical settings, ultimately improving the patient outcomes.

5.6. *AI implementation challenges*

Implementing AI models in clinical diagnostics presents several challenges, particularly concerning the integration into existing healthcare workflows and regulatory compliance. Integrating AI into clinical settings requires a seamless incorporation into established processes without disrupting the patient care. This necessitates comprehensive training for healthcare professionals to effectively utilize AI tools and adapt to new workflows [98]. Regulatory concerns also pose substantial hurdles. The lack of standardized guidelines for AI applications in medicine complicates the approval process and may delay the deployment [99]. Additionally, AI systems are susceptible to errors, including false positives and false negatives, potentially resulting in serious safety outcomes for patients [100]. Addressing these challenges requires a multifaceted approach, including developing robust validation protocols, establishing clear regulatory guidelines, and ensuring that AI tools are designed to complement, rather than replace, human clinical judgment.

5.7. AI-generated biases

AI has the potential to revolutionize healthcare by enhancing diagnostics, personalizing treatments, and streamlining administrative tasks. However, AI systems are also susceptible to biases that can adversely affect medical applications. These biases often stem from the data used to train AI models. For instance, an AI system trained predominantly on data from a specific demographic may underperform when applied to other groups, thus leading to disparities in healthcare outcomes [101]. Therefore, variability in training datasets is crucial for mitigating biases in AI systems. When the training data lacks diversity or disproportionately represents certain categories, AI models can inadvertently learn and perpetuate such biases, thus leading to inaccurate outcomes [102]. Addressing AI bias in medical applications requires a multifaceted approach. First, it's crucial to ensure that the training datasets are representative of the diverse populations the AI will serve. Second, implementing fairness-aware algorithms that actively detect and mitigate biases during the model development process is essential. Techniques such as reweighting, resampling, and adversarial debiasing can help achieve this. Third, the continuous monitoring and evaluation of AI systems in real-world settings are necessary to identify and correct biases that may emerge over time [103].

Thus, while DNNs offer powerful tools for advancing diagnostic methods in neurological disorders, their application is accompanied by significant challenges and limitations [71,73]. Addressing these issues—through an improved data quality, an enhanced interpretability, a better generalization, and careful ethical considerations—will be essential for fully realizing the potential of DNNs in this field [104,105]. As the technology continues to evolve, ongoing efforts to overcome these challenges will play a crucial role in shaping the future of AI-driven diagnostics.

6. Conclusions

A significant advancement in the use of DNNs for neurological diagnostics has been their ability to provide automated scoring with human-expert accuracy, while also offering insights into the underlying mechanisms of neurological impairments. Recent studies have demonstrated how explainable DNNs can be employed to design diagnostic tests in a data-driven and unbiased manner, significantly improving the sensitivity and specificity of these tests in detecting neurological disorders. These networks are capable of identifying subtle motor deficits and behavioral changes that human observers may overlook, thereby reducing the observer bias and enhancing the objectivity of diagnostic assessments. For instance, DNNs have been used to automatically detect motor deficits in clinical and experimental settings, while also revealing nuanced changes in behaviors that traditional methods often miss (Figure 1). This automated approach not only minimizes the workload for clinicians and researchers, but also eliminates the variability caused by human error, resulting in more reliable and reproducible diagnostic results. Furthermore, leveraging knowledge extraction techniques from DNNs allows researchers and clinicians to develop new diagnostic assays that are more sensitive and tailored to detect early signs of neurological and motoric impairments. This capability is particularly valuable for neurological disorders such as strokes, Parkinson's disease, and autism spectrum disorders, where early detection and precise, unbiased monitoring are crucial to improve the patient prognosis and evaluating the efficacy of therapeutic interventions.

Acknowledgments

The author developed AI agents to work with them, like with a good MSc student. Agents helped to identify the most relevant literature, design a plan for the paper, implement suggested improvements, and draft paper sections and rewrite them based on comments provided by the author. The author assumes full responsibility for the accuracy of the content presented here.

Conflict of interest

The author has no conflicts of interest to declare.

References

1. Bakeman R, Quera V (2011) *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139017343
2. Metz GA, Whishaw IQ (2002) Cortical and subcortical lesions impair skilled walking in the ladder rung walking test: a new task to evaluate fore-and hindlimb stepping, placing, and co-ordination. *J Neurosci Meth* 115: 169–179. https://doi.org/10.1016/S0165-0270(02)00012-2
3. Spano R (2005) Potential sources of observer bias in police observational data. *Soc Sci Res* 34: 591–617. https://doi.org/10.1016/j.ssresearch.2004.05.003
4. Asan O, Montague E (2014) Using video-based observation research methods in primary care health encounters to evaluate complex interactions. *J Innov Health Inform* 21: 161–170. https://doi.org/10.14236/jhi.v21i4.72
5. Moran RW, Schneiders AG, Major KM, et al. (2016) How reliable are functional movement screening scores? A systematic review of rater reliability. *Brit J Sport Med* 50: 527–536. https://doi.org/10.1136/bjsports-2015-094913
6. Mathis MW, Mathis A (2020) Deep learning tools for the measurement of animal behavior in neuroscience. *Curr Opin Neurobiol* 60: 1–11. https://doi.org/10.1016/j.conb.2019.10.008
7. Gautam R, Sharma M (2020) Prevalence and diagnosis of neurological disorders using different deep learning techniques: a meta-analysis. *J Med Syst* 44: 49. https://doi.org/10.1007/s10916-019-1519-7
8. Singh KR, Dash S (2023) Early detection of neurological diseases using machine learning and deep learning techniques: a review. *Artif Intell Neurol Diso* 2023: 1–24. https://doi.org/10.1016/B978-0-323-90277-9.00001-8
9. Arac A, Zhao P, Dobkin BH, et al. (2019) DeepBehavior: A deep learning toolbox for automated analysis of animal and human behavior imaging data. *Front Syst Neurosci* 13: 20. https://doi.org/10.3389/fnsys.2019.00020
10. Sewak M, Sahay SK, Rathore H (2020) An overview of deep learning architecture of deep neural networks and autoencoders. *J Comput Theor Nanos* 17: 182–188. https://doi.org/10.1166/jctn.2020.8648
11. Brattoli B, Büchler U, Dorkenwald M, et al. (2021) Unsupervised behaviour analysis and magnification (uBAM) using deep learning. *Nat Mach Intell* 3: 495–506. https://doi.org/10.1038/s42256-021-00326-x

12. ul Haq A, Li JP, Agbley BLY, et al. (2022) A survey of deep learning techniques based Parkinson's disease recognition methods employing clinical data. *Expert Syst Appl* 208: 118045. https://doi.org/10.1016/j.eswa.2022.118045
13. Nilashi M, Abumalloh RA, Yusuf SYM, et al. (2023) Early diagnosis of Parkinson's disease: a combined method using deep learning and neuro-fuzzy techniques. *Comput Biol Chem* 102: 107788. https://doi.org/10.1016/j.compbiolchem.2022.107788
14. Shahid AH, Singh MP (2020) A deep learning approach for prediction of Parkinson's disease progression. *Biomed Eng Lett* 10: 227–239. https://doi.org/10.1007/s13534-020-00156-7
15. Chintalapudi N, Battineni G, Hossain MA, et al. (2022) Cascaded deep learning frameworks in contribution to the detection of parkinson's disease. *Bioengineering* 9: 116. https://doi.org/10.3390/bioengineering9030116
16. Almuqhim F, Saeed F (2021) ASD-SAENet: a sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data. *Front Comput Neurosci* 15: 654315. https://doi.org/10.3389/fncom.2021.654315
17. Zhang L, Wang M, Liu M, et al. (2020) A survey on deep learning for neuroimaging-based brain disorder analysis. *Front Neurosci* 14: 779. https://doi.org/10.3389/fnins.2020.00779
18. Uddin MZ, Shahriar MA, Mahamood MN, et al. (2024) Deep learning with image-based autism spectrum disorder analysis: a systematic review. *Eng Appl Artif Intel* 127: 107185. https://doi.org/10.1016/j.engappai.2023.107185
19. Gupta C, Chandrashekar P, Jin T, et al. (2022) Bringing machine learning to research on intellectual and developmental disabilities: taking inspiration from neurological diseases. *J Neurodev Disord* 14: 28. https://doi.org/10.1186/s11689-022-09438-w
20. Saleh AY, Chern LH (2021) Autism spectrum disorder classification using deep learning. *IJOE* 17: 103–114. https://doi.org/10.3991/ijoe.v17i08.24603
21. Koppe G, Meyer-Lindenberg A, Durstewitz D (2021) Deep learning for small and big data in psychiatry. *Neuropsychopharmacology* 46: 176–190. https://doi.org/10.1038/s41386-020-0767-z
22. Gütter J, Kruspe A, Zhu XX, et al. (2022) Impact of training set size on the ability of deep neural networks to deal with omission noise. *Front Remote Sens* 3: 932431. https://doi.org/10.3389/frsen.2022.932431
23. Sturman O, von Ziegler L, Schläppi C, et al. (2020) Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacology* 45: 1942–1952. https://doi.org/10.1038/s41386-020-0776-y
24. He T, Kong R, Holmes AJ, et al. (2020) Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage* 206: 116276. https://doi.org/10.1016/j.neuroimage.2019.116276
25. Chen M, Li H, Wang J, et al. (2019) A multichannel deep neural network model analyzing multiscale functional brain connectome data for attention deficit hyperactivity disorder detection. *Radiol Artif Intell* 2: e190012. https://doi.org/10.1148/ryai.2019190012
26. Golshan HM, Hebb AO, Mahoor MH (2020) LFP-Net: a deep learning framework to recognize human behavioral activities using brain STN-LFP signals. *J Neurosci Meth* 335: 108621. https://doi.org/10.1016/j.jneumeth.2020.108621

27. Sutoko S, Masuda A, Kandori A, et al. (2021) Early identification of Alzheimer's disease in mouse models: Application of deep neural network algorithm to cognitive behavioral parameters. *Iscience* 24: 102198. https://doi.org/10.1016/j.isci.2021.102198
28. Tarigopula P, Fairhall SL, Bavaresco A, et al. (2023) Improved prediction of behavioral and neural similarity spaces using pruned DNNs. *Neural Networks* 168: 89–104. https://doi.org/10.1016/j.neunet.2023.08.049
29. Uyulan C, Ergüzel TT, Unubol H, et al. (2021) Major depressive disorder classification based on different convolutional neural network models: deep learning approach. *Clin EEG Neurosci* 52: 38–51. https://doi.org/10.1177/1550059420916634
30. Wen J, Thibeau-Sutre E, Diaz-Melo M, et al. (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med Image Anal* 63: 101694. https://doi.org/10.1016/j.media.2020.101694
31. Karthik R, Menaka R, Johnson A, et al. (2020) Neuroimaging and deep learning for brain stroke detection-A review of recent advancements and future prospects. *Comput Meth Prog Bio* 197: 105728. https://doi.org/10.1016/j.cmpb.2020.105728
32. Iqbal MS, Heyat MBB, Parveen S, et al. (2024) Progress and trends in neurological disorders research based on deep learning. *Comput Med Imag Grap* 116: 102400. https://doi.org/10.1016/j.compmedimag.2024.102400
33. Kim S, Pathak S, Parise R, et al. (2024) The thriving influence of artificial intelligence in neuroscience, In: Bhupathyraaj M, Reeta Vijayarani K, Dhanasekaran M, et al., *Application of Artificial Intelligence in Neurological Disorders,* Singapore: Springer Nature Singapore, 157–184. https://doi.org/10.1007/978-981-97-2577-9_9
34. Lima AA, Mridha MF, Das SC, et al. (2022) A comprehensive survey on the detection, classification, and challenges of neurological disorders. *Biology* 11: 469. https://doi.org/10.3390/biology11030469
35. Mulpuri RP, Konda N, Gadde ST, et al. (2024) Artificial intelligence and machine learning in neuroregeneration: a systematic review. *Cureus* 16: e61400 https://doi.org/10.7759/cureus.61400
36. Keserwani PK, Das S, Sarkar N (2024) A comparative study: prediction of parkinson's disease using machine learning, deep learning and nature inspired algorithm. *Multimed Tools Appl* 83: 69393–69441. https://doi.org/10.1007/s11042-024-18186-z
37. Fatima A, Masood S (2024) Machine learning approaches for neurological disease prediction: a systematic review. *Expert Syst* 41: e13569. https://doi.org/10.1111/exsy.13569
38. Surianarayanan C, Lawrence JJ, Chelliah PR, et al. (2023) Convergence of artificial intelligence and neuroscience towards the diagnosis of neurological disorders—a scoping review. *Sensors* 23: 3062. https://doi.org/10.3390/s23063062
39. Lombardi A, Diacono D, Amoroso N, et al. (2021) Explainable deep learning for personalized age prediction with brain morphology. *Front Neurosci* 15: 674055. https://doi.org/10.3389/fnins.2021.674055
40. Choo YJ, Chang MC (2022) Use of machine learning in stroke rehabilitation: a narrative review. *Brain Neurorehab* 15: e26. https://doi.org/10.12786/bn.2022.15.e26

41. Ryait H, Bermudez-Contreras E, Harvey M, et al. (2019) Data-driven analyses of motor impairments in animal models of neurological disorders. *PLoS Biol* 17: e3000516. https://doi.org/10.1371/journal.pbio.3000516
42. Nguyen HS, Ho DKN, Nguyen NN, et al. (2024) Predicting EGFR mutation status in non-small cell lung cancer using artificial intelligence: a systematic review and meta-analysis. *Acad Radiol* 31: 660–683. https://doi.org/10.1016/j.acra.2023.03.040
43. Zhang Y, Yao Q, Yue L, et al. (2023) Emerging drug interaction prediction enabled by a flow-based graph neural network with biomedical network. *Nat Comput Sci* 3: 1023–1033. https://doi.org/10.1038/s43588-023-00558-4
44. Le NQK (2023) Predicting emerging drug interactions using GNNs. *Nat Comput Sci* 3: 1007–1008. https://doi.org/10.1038/s43588-023-00555-7
45. Abed Mohammed A, Sumari P (2024) Hybrid k-means and principal component analysis (PCA) for diabetes prediction. *Int J Comput Dig Syst* 15: 1719–1728. https://doi.org/10.12785/ijcds/1501121
46. Mostafa F, Hasan E, Williamson M, et al. (2021) Statistical machine learning approaches to liver disease prediction. *Livers* 1: 294–312. https://doi.org/10.3390/livers1040023
47. Jackins V, Vimal S, Kaliappan M, et al. (2021) AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *J Supercomput* 77: 5198–5219. https://doi.org/10.1007/s11227-020-03481-x
48. Cho G, Yim J, Choi Y, et al. (2019) Review of machine learning algorithms for diagnosing mental illness. *Psychiat Invest* 16: 262. https://doi.org/10.30773/pi.2018.12.21.2
49. Aljrees T (2024) Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning. *Plos One* 19: e0295632. https://doi.org/10.1371/journal.pone.0295632
50. Hajare S, Rewatkar R, Reddy KTV (2024) Design of an iterative method for enhanced early prediction of acute coronary syndrome using XAI analysis. *AIMS Bioeng* 11: 301–322. https://doi.org/10.3934/bioeng.2024016
51. Schjetnan AGP, Luczak A (2011) Recording large-scale neuronal ensembles with silicon probes in the anesthetized rat. *J Vis Exp* 56: e3282. https://doi.org/10.3791/3282-v
52. Luczak A, Narayanan NS (2005) Spectral representation-analyzing single-unit activity in extracellularly recorded neuronal data without spike sorting. *J Neurosci Meth* 144: 53–61. https://doi.org/10.1016/j.jneumeth.2004.10.009
53. Luczak A, Hackett TA, Kajikawa Y, et al. (2004) Multivariate receptive field mapping in marmoset auditory cortex. *J Neurosci Meth* 136: 77–85. https://doi.org/10.1016/j.jneumeth.2003.12.019
54. Luczak A (2010) Measuring neuronal branching patterns using model-based approach. *Front Comput Neurosci* 4: 135. https://doi.org/10.3389/fncom.2010.00135
55. Luczak A, Kubo Y (2022) Predictive neuronal adaptation as a basis for consciousness. *Front Syst Neurosci* 15: 767461. https://doi.org/10.3389/fnsys.2021.767461
56. Lepakshi VA (2022) Machine learning and deep learning based AI tools for development of diagnostic tools, In: Parihar A, Khan R, Gohel H, et al., *Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection,* Academic Press, 2022: 399–420. https://doi.org/10.1016/B978-0-323-91172-6.00011-X

57. Montavon G, Binder A, Lapuschkin S, et al. (2019) Layer-wise relevance propagation: an overview, In: Samek W, Montavon G, Vedaldi A, et al., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning,* Cham: Springer, 193–209. https://doi.org/10.1007/978-3-030-28954-6_10
58. Nazir S, Dickson DM, Akram MU (2023) Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Comput Biol Med* 156: 106668. https://doi.org/10.1016/j.compbiomed.2023.106668
59. Torabi R, Jenkins S, Harker A, et al. (2021) A neural network reveals motoric effects of maternal preconception exposure to nicotine on rat pup behavior: a new approach for movement disorders diagnosis. *Front Neurosci* 15: 686767. https://doi.org/10.3389/fnins.2021.686767
60. Shahtalebi S, Atashzar SF, Patel RV, et al. (2021) A deep explainable artificial intelligent framework for neurological disorders discrimination. *Sci Rep* 11: 9630. https://doi.org/10.1038/s41598-021-88919-9
61. Morabito FC, Ieracitano C, Mammone N (2023) An explainable artificial intelligence approach to study MCI to AD conversion via HD-EEG processing. *Clin EEG Neurosci* 54: 51–60. https://doi.org/10.1177/15500594211063662
62. Goodwin NL, Nilsson SRO, Choong JJ, et al. (2022) Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience. *Curr Opin Neurobiol* 73: 102544. https://doi.org/10.1016/j.conb.2022.102544
63. Lindsay GW (2024) Grounding neuroscience in behavioral changes using artificial neural networks. *Curr Opin Neurobiol* 84: 102816. https://doi.org/10.1016/j.conb.2023.102816
64. Dan T, Kim M, Kim WH, et al. (2023) Developing explainable deep model for discovering novel control mechanism of neuro-dynamics. *IEEE T Med Imaging* 43: 427–438 https://doi.org/10.1109/TMI.2023.3309821
65. Fellous JM, Sapiro G, Rossi A, et al. (2019) Explainable artificial intelligence for neuroscience: behavioral neurostimulation. *Front Neurosci* 13: 1346. https://doi.org/10.3389/fnins.2019.01346
66. Bartle AS, Jiang Z, Jiang R, et al. (2022) A critical appraisal on deep neural networks: bridge the gap between deep learning and neuroscience via XAI. In: Angelov PP, *HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation*, 2022: 619–634. https://doi.org/10.1142/9789811247323_0015
67. Lemon RN (1997) Mechanisms of cortical control of hand function. *Neuroscientist* 3: 389–398. https://doi.org/10.1177/107385849700300612
68. Alaverdashvili M, Whishaw IQ (2013) A behavioral method for identifying recovery and compensation: hand use in a preclinical stroke model using the single pellet reaching task. *Neurosci Biobehav R* 37: 950–967. https://doi.org/10.1016/j.neubiorev.2013.03.026
69. Metz GAS, Whishaw IQ (2000) Skilled reaching an action pattern: stability in rat (Rattus norvegicus) grasping movements as a function of changing food pellet size. *Behav Brain Res* 116: 111–122. https://doi.org/10.1016/S0166-4328(00)00245-X
70. Faraji J, Gomez-Palacio-Schjetnan A, Luczak A, et al. (2013) Beyond the silence: bilateral somatosensory stimulation enhances skilled movement quality and neural density in intact behaving rats. *Behav Brain Res* 253: 78–89. https://doi.org/10.1016/j.bbr.2013.07.022

71. Sheu Y (2020) Illuminating the black box: interpreting deep neural network models for psychiatric research. *Front Psychiatry* 11: 551299. https://doi.org/10.3389/fpsyt.2020.551299
72. Fan FL, Xiong J, Li M, et al. (2021) On interpretability of artificial neural networks: a survey. *IEEE T Radiat Plasma* 5: 741–760. https://doi.org/10.1109/TRPMS.2021.3066428
73. Smucny J, Shi G, Davidson I (2022) Deep learning in neuroimaging: overcoming challenges with emerging approaches. *Front Psychiatry* 13: 912600. https://doi.org/10.3389/fpsyt.2022.912600
74. Kohlbrenner M, Bauer A, Nakajima S, et al. (2020) Towards best practice in explaining neural network decisions with LRP, *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020: 1–7. https://doi.org/10.1109/IJCNN48605.2020.9206975
75. Farahani FV, Fiok K, Lahijanian B, et al. (2022) Explainable AI: a review of applications to neuroimaging data. *Front Neurosci* 16: 906290. https://doi.org/10.3389/fnins.2022.906290
76. Böhle M, Eitel F, Weygandt M, et al. (2019) Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front Aging Neurosci* 11: 456892. https://doi.org/10.3389/fnagi.2019.00194
77. Marques dos Santos JD, Marques dos Santos JP (2023) Path-weights and layer-wise relevance propagation for explainability of ANNs with fMRI data, *International Conference on Machine Learning, Optimization, and Data Science,* Cham: Springer Nature Switzerland, 433–448. https://doi.org/10.1007/978-3-031-53966-4_32
78. Filtjens B, Ginis P, Nieuwboer A, et al. (2021) Modelling and identification of characteristic kinematic features preceding freezing of gait with convolutional neural networks and layer-wise relevance propagation. *BMC Med Inform Decis Mak* 21: 341. https://doi.org/10.1186/s12911-021-01699-0
79. Li H, Tian Y, Mueller K, et al. (2019) Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation. *Image Vision Comput* 83: 70–86. https://doi.org/10.1016/j.imavis.2019.02.005
80. Nam H, Kim JM, Choi W, et al. (2023) The effects of layer-wise relevance propagation-based feature selection for EEG classification: a comparative study on multiple datasets. *Front Hum Neurosci* 17: 1205881. https://doi.org/10.3389/fnhum.2023.1205881
81. Korda AI, Ruef A, Neufang S, et al. (2021) Identification of voxel-based texture abnormalities as new biomarkers for schizophrenia and major depressive patients using layer-wise relevance propagation on deep learning decisions. *Psychiat Res-Neuroim* 313: 111303. https://doi.org/10.1016/j.pscychresns.2021.111303
82. von Ziegler L, Sturman O, Bohacek J (2021) Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology* 46: 33–44. https://doi.org/10.1038/s41386-020-0751-7
83. Marks M, Jin Q, Sturman O, et al. (2022) Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. *Nat Mach Intell* 4: 331–340. https://doi.org/10.1038/s42256-022-00477-5
84. Bohnslav JP, Wimalasena NK, Clausing KJ, et al. (2021) DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. *Elife* 10: e63377. https://doi.org/10.7554/eLife.63377

85. Wang PY, Sapra S, George VK, et al. (2021) Generalizable machine learning in neuroscience using graph neural networks. *Front Artif Intell* 4: 618372. https://doi.org/10.3389/frai.2021.618372
86. Watson DS, Krutzinna J, Bruce IN, et al. (2019) Clinical applications of machine learning algorithms: beyond the black box. *Bmj* 364: l886. https://doi.org/10.2139/ssrn.3352454
87. Jain A, Salas M, Aimer O, et al. (2024) Safeguarding patients in the AI era: ethics at the forefront of pharmacovigilance. *Drug Safety* 48: 119–127. https://doi.org/10.1007/s40264-024-01483-9
88. Murdoch B (2021) Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics* 22: 1–5. https://doi.org/10.1186/s12910-021-00687-3
89. Ziesche S (2021) AI ethics and value alignment for nonhuman animals. *Philosophies* 6: 31. https://doi.org/10.3390/philosophies6020031
90. Bossert L, Hagendorff T (2021) Animals and AI. the role of animals in AI research and application-an overview and ethical evaluation. *Technol Soc* 67: 101678. https://doi.org/10.1016/j.techsoc.2021.101678
91. Gong Y, Liu G, Xue Y, et al. (2023) A survey on dataset quality in machine learning. *Inform Software Tech* 162: 107268. https://doi.org/10.1016/j.infsof.2023.107268
92. Bolaños LA, Xiao D, Ford NL, et al. (2021) A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. *Nat Methods* 18: 378–381. https://doi.org/10.1038/s41592-021-01103-9
93. Lashgari E, Liang D, Maoz U (2020) Data augmentation for deep-learning-based electroencephalography. *J Neurosci Methods* 346: 108885. https://doi.org/10.1016/j.jneumeth.2020.108885
94. Barile B, Marzullo A, Stamile C, et al. (2021) Data augmentation using generative adversarial neural networks on brain structural connectivity in multiple sclerosis. *Comput Meth Prog Bio* 206: 106113. https://doi.org/10.1016/j.cmpb.2021.106113
95. Memar S, Jiang E, Prado VF, et al. (2023) Open science and data sharing in cognitive neuroscience with MouseBytes and MouseBytes+. *Sci Data* 10: 210. https://doi.org/10.1038/s41597-023-02106-1
96. Jleilaty S, Ammounah A, Abdulmalek G, et al. (2024) Distributed real-time control architecture for electrohydraulic humanoid robots. *Robot Intell Automat* 44: 607–620. https://doi.org/10.1108/RIA-01-2024-0013
97. Zhao J, Wang Z, Lv Y, et al. (2024) Data-driven learning for H∞ control of adaptive cruise control systems. *IEEE Trans Veh Technol* 73: 18348–18362 https://doi.org/10.1109/TVT.2024.3447060
98. Kelly CJ, Karthikesalingam A, Suleyman M, et al. (2019) Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 17: 1–9. https://doi.org/10.1186/s12916-019-1426-2
99. Kulkarni PA, Singh H (2023) Artificial intelligence in clinical diagnosis: opportunities, challenges, and hype. *Jama* 330: 317–318. https://doi.org/10.1001/jama.2023.11440
100. Choudhury A, Asan O (2020) Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR Med inf* 8: e18599. https://doi.org/10.2196/18599
101. Ratwani RM, Sutton K, Galarraga JE (2024) Addressing AI algorithmic bias in health care. *Jama* 332: 1051–1052. https://doi.org/10.1001/jama.2024.13486

102. Chen C, Sundar SS (2024) Communicating and combating algorithmic bias: effects of data diversity, labeler diversity, performance bias, and user feedback on AI trust. *Hum-Comput Interact* 2024: 1–37. https://doi.org/10.1080/07370024.2024.2392494
103. Chen F, Wang L, Hong J, et al. (2024) Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *J Am Medl Inform Assn* 31: 1172–1183. https://doi.org/10.1093/jamia/ocae060
104. Ienca M, Ignatiadis K (2020) Artificial intelligence in clinical neuroscience: methodological and ethical challenges. *AJOB Neurosci* 11: 77–87. https://doi.org/10.1080/21507740.2020.1740352
105. Avberšek LK, Repovš G (2022) Deep learning in neuroimaging data analysis: applications, challenges, and solutions. *Front Neuroimag* 1: 981642. https://doi.org/10.3389/fnimg.2022.981642