Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Big Data Analytics for Uncovering Voxel Connectivity Patterns in Attention Deficit Hyperactivity Disorder

Rezzy Eko Caraka[1–4], Khairunnisa Supardi[5], Prana Ugiana Gio[6], Vijaya Isnaniawardhani[1], Rung Ching Chen[4], Bekti Djatmiko[1,7], Bens Pardamean[8,9]

[1]Engineers Profession Program, Graduate School, Universitas Padjadjaran, Bandung, West Java, 45363, Indonesia; [2]Research Center for Data and Information Sciences, Research Organization for Electronics and Informatics, National Research and Innovation Agency (BRIN), Bandung, 40135, Indonesia; [3]School of Economics and Business Telkom University, Bandung, 40257, Indonesia; [4]Department of Information Management, Chaoyang University of Technology, Taichung, 44919, Taiwan; [5]Department of Radiation Oncology, Faculty of Medicine Universitas Indonesia - Dr Cipto Mangunkusumo National General Hospital, Greater Jakarta, 10430, Indonesia; [6]Department of Mathematics, Universitas Sumatera Utara, Medan, 20155, Indonesia; [7]PT. Wiratman Cipta Manggala (WCM), Graha Simatupang, Simatupang, Jakarta, 12540, Indonesia; [8]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, 11480, Indonesia; [9]Computer Science Department, BINUS Graduate Program – Master of Computer Science Program, Bina Nusantara University, Jakarta, 11480, Indonesia

Correspondence: Rezzy Eko Caraka; Rung Ching Chen, Email r.eko.caraka@unpad.ac.id; crching@cyut.edu.tw

Introduction: Attention Deficit Hyperactivity Disorder (ADHD) is a complex neurodevelopmental condition characterized by heterogeneous brain activity patterns. Identifying key brain regions associated with ADHD remains a challenge due to the high dimensionality and complexity of neuroimaging data. This study aims to apply advanced machine learning techniques to uncover critical features and improve classification performance in ADHD diagnosis.

Methods: We analyzed 5937 brain voxels aggregated from neuroimaging records of patients diagnosed with ADHD. Feature selection was performed using Boruta, Random Forest in combination with DALEX explainability tools, and Neural Networks. Dimensionality reduction and clustering techniques including Principal Component Analysis (PCA), KMeans, and MCLUST were used to explore underlying voxel patterns. The performance of different activation functions—ReLU, Sigmoid, and Tanh—was evaluated within deep neural networks.

Results: Several key brain regions, including the Fusiform Gyrus, Thalamus, and Superior Temporal Gyrus, were identified as significant predictors for ADHD. The integration of machine learning models demonstrated improved classification accuracy, with ReLU-based neural networks outperforming others in most evaluation metrics.

Discussion: The study demonstrates the potential of a robust, integrated machine learning framework to analyze high-dimensional neuroimaging data and identify biologically relevant markers of ADHD. These findings contribute to the growing body of evidence supporting data-driven approaches in neuropsychiatric diagnosis and may inform future clinical decision-making and personalized interventions.

Keywords: ADHD, brain voxels, neuroimaging, machine learning, feature selection, deep learning, activation function

Highlight

1. Big data helps find detailed brain connectivity patterns in people with ADHD that might be missed with smaller datasets.
2. Analyzing large amounts of brain data helps identify specific patterns unique to ADHD, improving diagnosis and treatment.
3. Using big data and machine learning, scientists can create models that better predict ADHD by looking at brain connectivity, making it easier to understand and manage the disorder.

Received: 20 March 2025
Accepted: 24 July 2025
Published: 31 July 2025

Introduction

Attention-Deficit/Hyperactivity Disorder (ADHD) is a common neurodevelopmental disorder that typically begins in childhood and often persists into adolescence and adulthood. It is characterized by symptoms of inattention, hyperactivity, and impulsivity, which can significantly impair daily functioning and quality of life. ADHD is classified into three main subtypes: predominantly hyperactive-impulsive, predominantly inattentive, and the combined type, which includes all three core symptoms.[1–3]

Advances in brain imaging have greatly enhanced our understanding of the structural and functional differences in the brains of individuals with ADHD. Structural imaging techniques such as CT, MRI, MRS, and DTI are used to examine brain anatomy, while functional techniques like fMRI, PET, EEG, and MEG allow researchers to study brain activity and connectivity patterns. Previous studies have reported alterations in brain regions such as the basal ganglia and abnormalities in dopaminergic signaling in individuals with ADHD.

Resting-state functional MRI (rs-fMRI) has become a widely used tool for mapping the functional connectome of the brain.[4–6] However, the high dimensionality of neuroimaging data—often involving thousands of voxels—poses a challenge for analysis.[7,8] Feature selection techniques are essential to reduce dimensionality, improve model performance, and enhance interpretability without altering the original meaning of the variables. These methods are particularly useful in unsupervised settings, where class labels are not available.

Recent research has shown that combining advanced methodologies such as hyperbolic disc embedding and deep learning with effective feature selection strategies can yield valuable insights into brain disorders. Techniques such as Boruta, XGBoost, and other unsupervised methods have demonstrated their utility in selecting relevant features from high-dimensional neuroimaging data.

This study aims to analyze brain voxel data from patients with ADHD using a combination of feature selection and machine learning techniques. Materials and Methods presents the materials and methods, including voxel-based data preprocessing and the applied models. Dataset reports simulation results across four scenarios: dimensionality reduction, XGBoost classification, Boruta feature selection, and deep neural networks. Results And Analysis concludes with a discussion of key findings and implications for future research in the field of neuroimaging and ADHD.

Materials and Methods

K Core Percolation

The concept of k-core percolation is crucial for understanding the resilience and structural properties of networks. The k-core of a graph represents the largest subgraph in which every node has at least k connections.[6] This process has significant applications in areas such as social networks, communication systems, and biological networks, where the k-core can be used to identify influential nodes, stable communities, or robust subsystems. To analyze k -core percolation, one must consider the theoretical formulations for estimating the percolation threshold. This threshold indicates the critical point at which the k-core vanishes as nodes or edges are removed.

The starting point is a graph $G = (V, E)$ with $N = |V|$ nodes and $|E|$ edges. A subgraph $G_k \subseteq G$ is called a k-core if all nodes in $G_k$satisfy the condition $d(v) \geq k$, where $d(v)$ is the degree of node (v). The k-core percolation process iteratively removes nodes with degree less than (k) until all remaining nodes satisfy the condition.

The fraction of nodes remaining in the k-core after random removal of nodes or edges is denoted as ϕ_k. For a random graph, the percolation threshold depends on the degree distribution $P(d)$, the probability of node removal (p), and the connectivity structure of the network. The degree distribution $P(d)$, of the graph is often characterized by a generating function, defined as:

$$G_0(x) = \sum_{d=0}^{\infty} P(d)x^d$$

This function encodes the probability distribution of degrees and provides the mean degree $\langle d \rangle = G_0(1)$. The critical threshold for k-core percolation is determined by the stability condition, which ensures that a sufficient fraction of nodes retain at least k neighbors. This condition is expressed as:

$$\phi_k = \sum_{d=k}^{\infty} P(d) H_k(d)$$

Here, $H_k$represents the probability that a node with degree d retains at least k neighbors after random removal. This probability is calculated using the binomial distribution:

$$H_k(d) = \binom{d}{k} p^k (1-p)^{d-k}$$

where p is the probability that a neighbor of the node remains in the graph. For random graphs with a Poisson degree distribution, commonly generated by the Erdős–Rényi model, the degree distribution is given by

$$P(d) = \frac{e^{-\langle d \rangle} \langle d \rangle^d}{d!}$$

In such cases, the critical percolation threshold can be approximated as:

$$p_c \approx \frac{k}{\langle d \rangle}$$

This approximation assumes that the mean degree $\langle d \rangle$is large enough to sustain a k -core for a given k. The benefit of studying k-core percolation lies in its ability to predict network collapse, identify critical nodes for intervention, and optimize network designs for robustness.

Feature Selection

Feature selection is a critical preprocessing step in data analysis and machine learning, aimed at identifying the most relevant features for building effective models. In voxel-based analyses, features often include spatial properties, connectivity metrics, and intensity values. Selecting the optimal subset of these features reduces computational complexity, improves model performance, and enhances interpretability.[9]

Feature selection can be formalized as an optimization problem. Let $X = \{x_1, x_2, \ldots, x_n\}$ represent the complete set of features in a dataset, and y denote the target variable. The goal is to identify a subset $S \subseteq X$ that maximizes a predefined objective function $J(S)$, often related to predictive accuracy or statistical relevance. Mathematically, this can be expressed as

$$S^* = \arg\max_{S \subseteq X} J(S)$$

where S^* represents the optimal subset of features. The choice of $J(S)$depends on the specific problem. Common metrics include mutual information, correlation, or model performance measures such as accuracy, F1-score, or mean squared error (MSE).[10,11]

Feature selection techniques are generally categorized into three main approaches, each with distinct characteristics and use cases. Filter methods rely on statistical measures to evaluate the relevance of individual features independently of the model. These methods prioritize features based on metrics such as mutual information and correlation coefficients. Mutual information measures the shared information between a feature and the target variable. For a feature x_i and target y it is calculated as

$$I(x_i; y) = \sum_{x_i \in X} \sum_{y \in Y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i) p(y)}$$

Features with higher mutual information are considered more relevant. Similarly, correlation coefficients assess the linear relationship between a feature and the target. Features with higher absolute values of correlation are prioritized.

$$r_{x_i y} = \frac{Cov(x_i, y)}{\sqrt{Var(x_i) Var(y)}}$$

Wrapper methods evaluate subsets of features by training a machine learning model and assessing its performance. These methods iteratively add or remove features based on their impact. Forward selection begins with an empty set and progressively adds features that improve model performance. Backward elimination starts with all features and removes those contributing the least. Recursive feature elimination (RFE) ranks features by importance and eliminates the least significant ones in a stepwise manner. While wrapper methods can yield optimized feature subsets, they are computationally expensive, especially for high-dimensional data.

Embedded methods integrate feature selection into the model training process. Regularization techniques, such as Lasso (L1 regularization), penalize the magnitude of feature coefficients, effectively shrinking irrelevant features to zero.

$$\min \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ is the penalty term controlling the sparsity of the model. Tree-based models, including Decision Trees, Random Forests, and Gradient Boosting, inherently rank features by their contribution to reducing impurity or error during training. For example, feature importance in decision trees is determined by the reduction in impurity across all nodes that split on the feature, weighted by the number of samples at each node.

Deep Learning for Voxel Classification

Voxel classification has broad applications, especially in medical imaging. For example, in brain imaging, deep learning models can classify voxels to identify functionally distinct regions or detect patterns associated with neurodevelopmental disorders such as ADHD, which may aid in early diagnosis or understanding of altered brain connectivity.

In medical scans such as CT or MRI, voxel classification can help segment different tissues or organs, allowing for more accurate diagnoses and treatment planning. In 3D object recognition, voxel classification can be used to analyze 3D point clouds or voxel grids of objects, identifying different parts or categories of objects in a 3D space. This technique is widely used in robotics, autonomous vehicles, and computer vision applications.

In voxel classification, the data is represented as a 3D grid of voxels, where each voxel has attributes such as intensity, color, or other relevant features. For instance, in medical imaging, the intensity values of voxels represent different tissue types or abnormalities. This 3D volume is then processed by deep learning models to classify each voxel, providing insights into the underlying structures or patterns. The input to the neural network is a volumetric image, represented as a tensor $X \in \mathbb{R}^{D \times H \times W}$, where D, H,and, $W$are the depth, height, and width of the 3D volume, respectively. Each voxel in the volume contains a feature vector, which may include intensity values, texture information, or other derived features.

One of the most effective architectures for this task is the 3D Convolutional Neural Network (3D-CNN). In this architecture, the network applies 3D convolutions to the input volume to extract local spatial features.

$$Y = (x, y, z) = (X \times K)(x, y, z) = \sum_{i=-r}^{r} \sum_{j=-r}^{r} \sum_{k=-r}^{r} X(x+i, y+j, z+k) \cdot K(i, j, k)$$

where x is the input volume, K is the convolution kernel (filter), and $r$is the radius of the filter. The output y represents the extracted features from the input volume after applying the convolution operation. These features capture spatial relationships between neighboring voxels, which are crucial for understanding the structure of the 3D data. The architecture of a typical 3D-CNN includes multiple layers of convolutions followed by pooling layers. Pooling layers, such as 3D max-pooling, are used to downsample the feature maps and reduce dimensionality while preserving important information. After a series of convolution and pooling layers, the output feature map is passed through fully connected layers, which help in making final classifications based on the extracted features.

U-Net and V-Net for Voxel Classification

U-Net and V-Net are two popular neural network architectures specifically designed for segmentation tasks, but they can also be used for voxel classification. U-Net consists of an encoder-decoder structure, where the encoder gradually downscales the input to extract features, and the decoder upsamples the feature maps to reconstruct the spatial

dimensions of the input. The architecture is designed to preserve high-resolution information at the voxel level, which is crucial for accurate classification. Mathematically, the network can be described as follows.

$$Y = \mathrm{f}_{\mathrm{dec}}(\mathrm{f}_{\mathrm{enc}}(X))$$

where X is the input volume, $\mathrm{f}_{\mathrm{enc}}$ represents the encoding operation (convolutions and downsampling), and f_{dec} represents the decoding operation (upsampling and convolution). Skip connections between the encoder and decoder layers allow the network to retain fine-grained details during the decoding process, leading to better voxel-level predictions. V-Net is similar to U-Net but designed for 3D volumetric data. V-Net uses 3D convolutions throughout the network to process volumetric data, and its loss function is typically based on the Dice coefficient, which measures the overlap between predicted and true voxel regions. The Dice coefficient is defined as

$$D = \frac{2|A \cap B|}{|A| + |B|}$$

where A and B represent the predicted and true voxel sets, respectively. The Dice coefficient ranges from 0 to 1, with 1 indicating perfect overlap. The goal of the V-Net is to minimize the loss function based on this coefficient, thereby improving voxel-level classification accuracy.

Training deep learning models for voxel classification requires a suitable loss function and optimization algorithm. For multi-class classification tasks, the loss function commonly used is cross-entropy loss, defined as.

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(p_i)$$

where C is the number of classes, $y_i$is the ground truth label (one-hot encoded), and p_i is the predicted probability for class i. Cross-entropy loss encourages the model to predict the correct class for each voxel, minimizing the difference between the predicted and actual class distributions.

Dataset

The voxel data is investigated only for 10 subjects which contain time series of 5937 voxels (1200 time points) but not the correlation matrix itself, and the NII file containing the voxel information which refers to Brainnetome Center, Institute of Automation, Chinese Academy of Sciences (https://atlas.brainnetome.org). Brain mask image of reduced size (31*37*31) for removing background voxels and 4D data matrix (X*Y* Z*time) in every 180 subjects. For removing background voxels, we use the brain mask image size (31*37*31) to multiply this brain mask and original data. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The research question and framework to be built is to see how the best method is to classify and predict the ADHD and TDHD datasets. Figure 1 describes the construction steps of this research. The detailed statistical information presented in Appendix 1 provides a nuanced perspective on the differences between individuals with Attention Deficit Hyperactivity Disorder (ADHD) and typically developing individuals (TDHD) across multiple key variables. The ADHDIndex, a measure of ADHD symptomatology, reveals a notably lower mean for the ADHD group (−72.484) in comparison to the TDHD group (−34.852). The confidence intervals further accentuate the disparity, with the potential for negative values in the ADHD group and a wider range in the TDHD group.

Examining age, the mean age of individuals with ADHD (12.1407) is slightly lower than that of the TDHD group (12.3164). This subtle difference in age may hold significance in understanding developmental aspects associated with ADHD. Turning to cognitive abilities, individuals with ADHD consistently demonstrate higher mean scores in both Verbal IQ (VIQ) and Full IQ (FIQ) compared to their TDHD counterparts. The VIQ mean for the ADHD group is 120.62, while it is 112.49 for the TDHD group. Similarly, the FIQ mean is 118.95 for the ADHD group and 106.59 for the TDHD group. The confidence intervals provide additional context, delineating the potential ranges for the true mean values in each group.

Figure 1 Step construction. Alt text: Flowchart illustrating the process of model construction using a brain dataset. The process begins with K-fold cross-validation, where the dataset is split into K folds. A model is trained on the training data and evaluated on the validation data, with this process repeated for all folds. The average evaluation result is then calculated. Based on this result, feature selection may be applied using methods such as Random Forest Importance, Backward Selection, Forward Selection, or Stepwise Selection. If no feature selection is applied, all variables are used. Different models, including Random Forest, XGBoost, Boruta, Lasso, SVM, and Neural Networks, are trained using the selected features. If all variables are used, a Deep Neural Network is trained. The final step involves using the trained models for prediction. Diamonds represent decision points, and rectangles represent process steps, connected by arrows indicating the flow of data.

The Inattentive scores exhibit a distinctive pattern, with the ADHD group presenting a lower mean (−84.377) in contrast to the TDHD group's mean of −55.721. The wider range in the ADHD group emphasizes potential variability in attention-related traits among individuals with ADHD. These numerical insights contribute valuable information to our understanding of the cognitive, developmental, and symptomatology differences between individuals with ADHD and those who are typically developing. The comprehensive statistical analysis sheds light on the multifaceted nature of ADHD and provides a basis for further exploration and interpretation in the field of neurodevelopmental disorders.

Deep Learning is a subset of machine learning which has a function that mimics the work of the human brain. Deep learning consists of multiple layers of neurons, similar to how the human brain relays information. Thus, it is part of Artificial Neural Network (ANN), which has various layers identified as nodes with different weights. The various layers within the neural network can process a high level of abstract data, making it suitable to process unstructured data such as image, video, text, or audio. There are many types of Deep Learning with specific applications across different studies such as engineering, business, finance, health, etc. Many kinds of deep learning have been implemented in other subject areas such as Deep Multilayer perceptron (DMLP), Recurrent Neural Network (RNN),[12] Convolutional Neural Network (CNN),[13] Deep Belief Networks (DBNs).[14]

In the realm of analytical processes, the amalgamation of diverse feature selection techniques unfolds a spectrum of advantages. Principal Component Analysis (PCA) takes the lead, proficiently transforming high-dimensional data into

a more manageable, lower-dimensional space. Beyond enhancing computational efficiency, PCA captures the dataset's most significant variations and mitigates multicollinearity, elevating the interpretability of selected features.

Complementing PCA, Kmeans clustering assumes a pivotal role, unveiling intrinsic structures and patterns within the data. By exposing natural groupings of observations, Kmeans clustering facilitates discernment of subtle relationships and dependencies that might elude detection in more extensive datasets. The Gaussian Finite Mixture Model, executed through the Expectation-Maximization (EM) algorithm, proves adept at handling complex data distributions, making it a robust choice for discerning nuanced variations. Boruta acts as a central filtering mechanism in this comprehensive feature selection ensemble, ensuring the retention of all relevant variables, including those with non-linear correlations. This fortifies the reliability of the feature selection process. Concurrently, XGBoost, operating as an ensemble learning algorithm, iteratively hones predictive models based on the most informative features. This iterative approach significantly enhances predictive accuracy, particularly in scenarios marked by intricate relationships between variables.

Introducing a layer of versatility, Deep Neural Networks (DNNs) contribute to the feature selection process by capturing intricate patterns and representations within high-dimensional datasets. Their proficiency in learning hierarchical structures empowers DNNs to uncover hidden relationships and nuanced patterns in the data.The Rectified Linear Unit (ReLU) activation function, stands out in neural network architectures for introducing non-linearity and effectively addressing the vanishing gradient problem during backpropagation. Unlike the sigmoid and tanh functions, ReLU allows positive inputs to pass through unchanged, while converting negative inputs to zero.

Zooming in on the Rectified Linear Unit (ReLU) activation function in neural networks, represented as$f(x) = \max(0, x)$its distinctiveness lies in introducing non-linearity and adeptly addressing the vanishing gradient problem during backpropagation. In contrast to sigmoid and tanh functions, ReLU permits positive inputs to pass through unchanged while converting negative inputs to zero. This non-linear transformation significantly amplifies the network's capacity to capture intricate patterns, albeit with caveats such as the potential for "dead neurons" and susceptibility to the exploding gradient problem for large inputs.

Sigmoid and tanh, each with merits and demerits, present alternative considerations. Sigmoid proves advantageous in binary classification tasks due to its output range between 0 and 1 but tends to produce small gradients, contributing to the vanishing gradient problem $f(x) = \frac{1}{1+e^{-x}}$. Otherwise, Tanh, $f(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$with an output range between −1 and 1, mitigates the vanishing gradient to some extent but may grapple with challenges, especially for extreme values.

Ultimately, the selection of an activation function hinges on the specific requirements of the neural network and the nature of the data at hand. Despite their individual nuances, the integration of these diverse techniques and functions fosters a more comprehensive understanding of the dataset. Visual representations, exemplified in Figure 2, underscore the efficacy of this holistic approach by revealing discernible differences in voxel connectivity across age groups, offering profound implications for comprehending neurobiological processes associated with aging.

Results and Analysis

Scenario 1: Dimension Reduction

The most widely used feature selection technique is PCA because this technique has the advantage of a simple model that makes it easy to use.[15–17] Simply put, PCA is a linear transformation to determine the new coordinate system of a dataset. This PCA technique reduces or reduces large data information from connectivity voxels without eliminating existing information. The PCA algorithm decomposes voxel connectivity into a set of characteristic features known as "Eigenvalues".[18–20] This is then called the Principal Component in a data training set, the main feature of a PCA algorithm is the reconstruction of several connectivity voxels from the training set by combining eigenvalues as shown in Figure 3.

The information provided highlights two key features associated with age, specifically (R_7_2) and Plasticity-related gene-1 (PRG-1). These features are crucial in understanding the intricate factors contributing to attention deficit hyperactivity disorder (ADHD). PRG-1, being a brain-specific membrane protein linked to lipid phosphate phosphatases, plays a pivotal role in the hippocampus, particularly at the excitatory synapse terminating on glutamatergic neurons. This information suggests a potential connection between synaptic regulation in the hippocampus and ADHD, as disruptions in synaptic transmission have been implicated in the disorder.Moreover, the significance of specific brain regions, such as the right hemisphere area V5/MT+

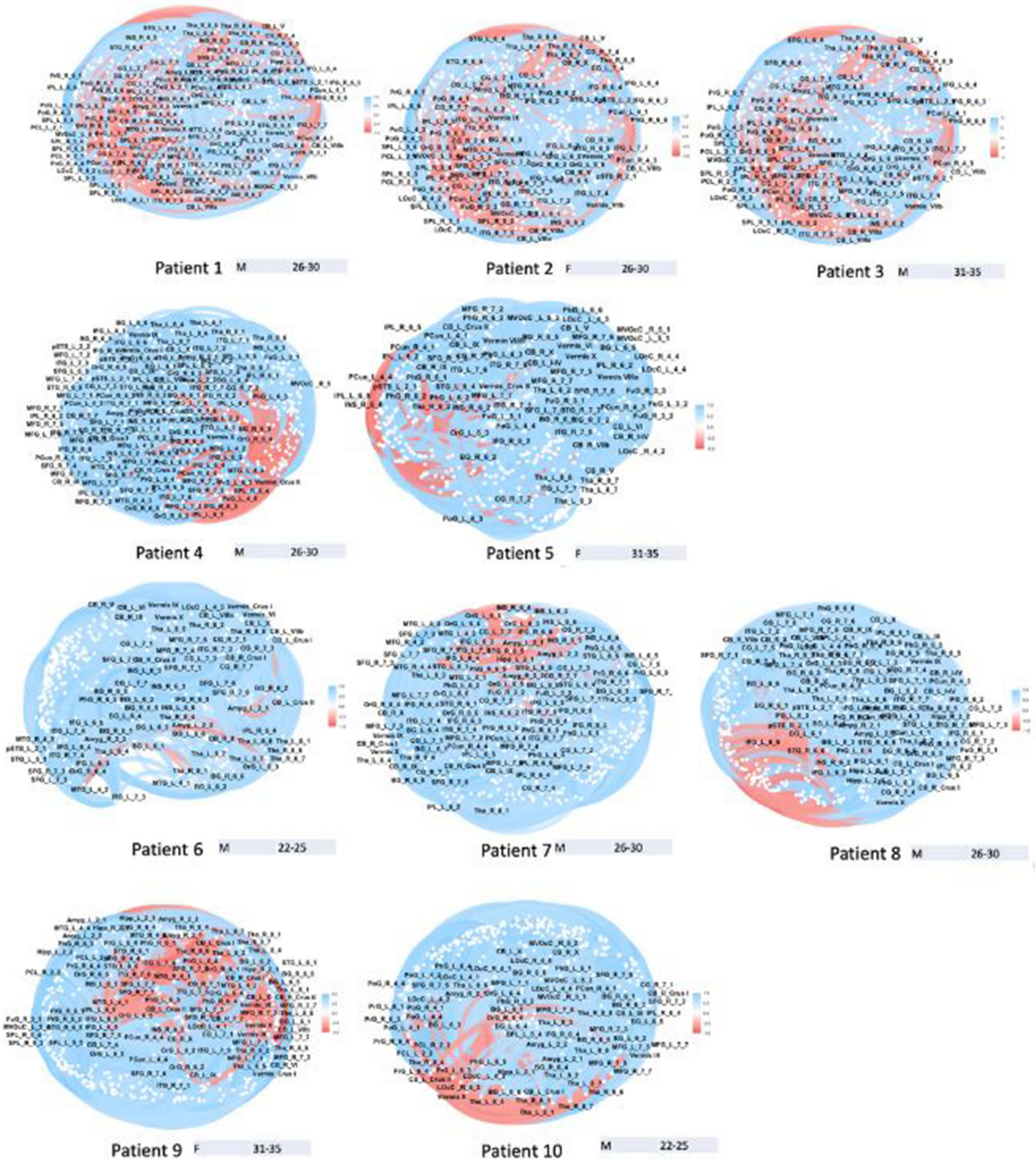

Figure 2 Brain network connectivity visualizations from selected individuals. Each panel shows functional connectivity patterns derived from neuroimaging data, illustrating inter-individual variability across the sample. The functional brain network connectivity of ten individuals is visualized using circular plots, each representing the connectivity patterns across labeled brain regions. These plots illustrate inter-regional correlations derived from neuroimaging data, with connection strength and directionality encoded through color: blue for positive correlations and red for negative ones. The plots are arranged in two rows, each containing five individual connectivity maps. Brain regions are denoted using standardized abbreviations such as "PCL_R_1_3" and "THA_L_3", and connections are depicted as lines bridging these regions. Below each plot, demographic information—sex (M or (F) and age group (ranging from 22–25 to 31–35 years)—is provided. Visual inspection reveals considerable variability in the density and distribution of network connectivity across individuals, despite similarities in age or sex. Some individuals display predominantly positive connectivity (eg, Individuals 1, 3, 7, and 9), while others exhibit more balanced or even negative connectivity patterns (eg, Individuals 5, 6, and 10). This variability underscores the heterogeneity of functional brain networks across individuals and highlights the complex interplay between demographic and possibly intrinsic neurological factors. A color scale adjacent to each plot provides reference for interpreting the range of correlation values.

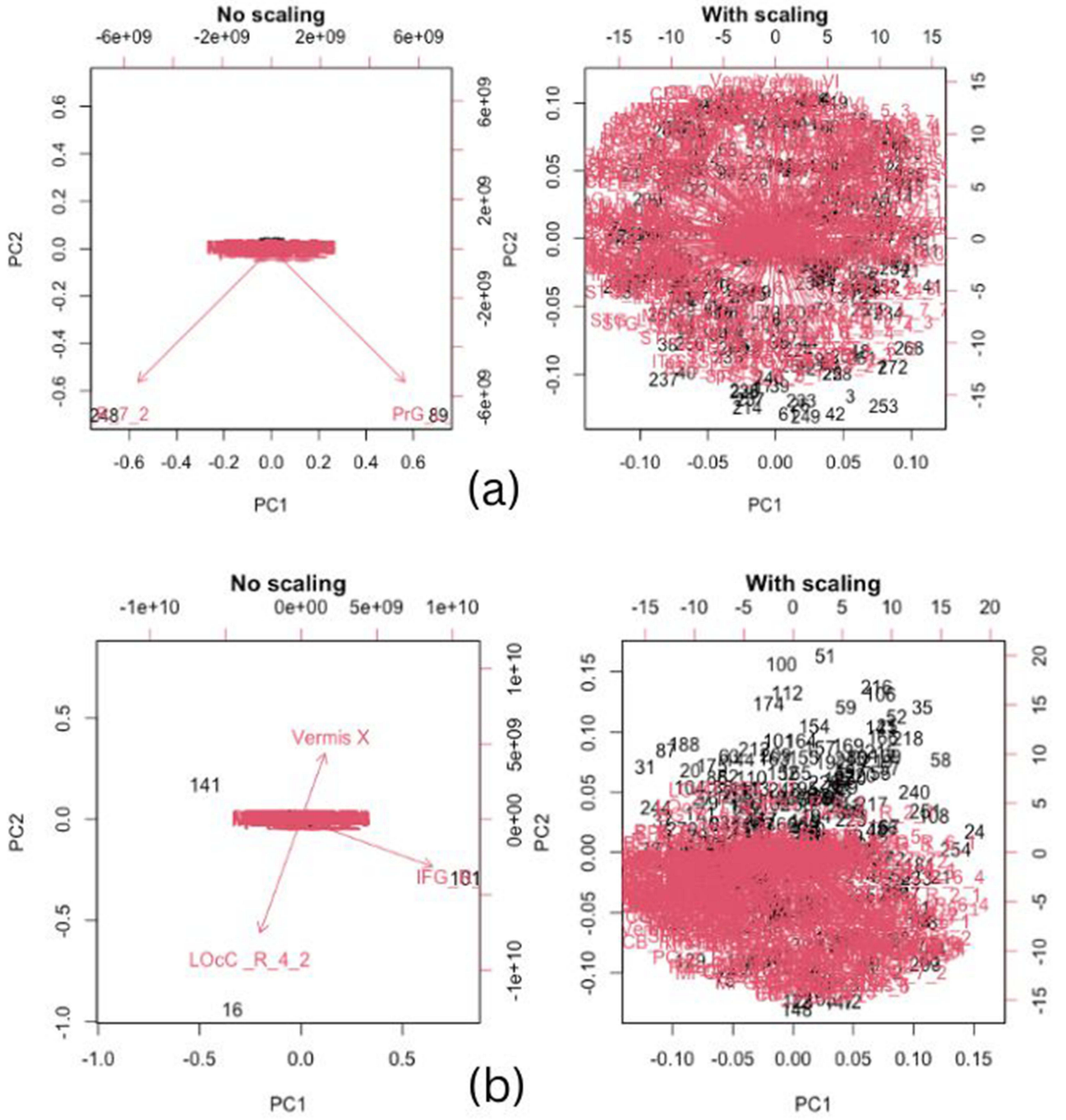

Figure 3 Best feature age (**a**) and sex (**b**) using PCA. Alt Text: Two sets of PCA scatterplots comparing data with and without scaling. (**a**) shows PCA results highlighting features related to age, while (**b**) shows PCA results highlighting features related to sex. Each panel has two plots: the left without scaling, showing data points tightly clustered with a few outliers, and the right with scaling, displaying a more dispersed distribution of points. Red arrows indicate key features.

(LOcc_R_4_2), Vermis X, and the inferior frontal gyrus (IFG_01), is underscored in the context of ADHD. These regions exert a substantial influence on cognitive processes, emphasizing their importance in understanding ADHD-related cognitive challenges. The right hemisphere area V5/MT+ is associated with visual motion processing, Vermis X is involved in motor control and coordination, while the inferior frontal gyrus plays a crucial role in executive functions and impulse control.

The mention of these brain regions implies that abnormalities or variations in these areas may contribute to the manifestation of ADHD symptoms. The interplay of factors related to age, genetic features like PRG-1, and the influence

https://doi.org/10.2147/JMDH.S523137

of specific brain regions highlights the complexity of ADHD's neurobiological underpinnings. Further research and exploration of these features can potentially lead to a more nuanced understanding of ADHD and may inform targeted interventions to address the specific neural mechanisms associated with the disorder.

The K-means method is the simplest and most common clustering method. K-means has the ability to group large amounts of data with relatively fast and efficient time computation. However, K-means has a weakness caused by determining the initial center of the cluster. The results of the cluster formed from the K-means method are very dependent on the initiation of the initial center value of the given cluster. This causes the results of the cluster to be a locally optimal solution. For this reason, the K-means collaborates with the hierarchical method for determining the initial center of the cluster. In a basic way, we run a simple *kmeans()* function, with 5 clusters then effectively duct tape the cluster numbers to each row of data and call it a day. At the same time, we justify the best cluster with the scree plot. In addition, we scale the data use the *scale()* function or another normalization technique to get more accurate and set *nstart* to 100 for simplicity. Figure 4 explains that both age and sex have 5 optimum clusters with a value of variance age 30.3% and sex 26.8%, respectively.

Gaussian mixture model is often used as unsupervised learning. In general, GMM can be written as follows:

$$f(x_i|\theta_k) = \frac{\pi_i}{\sqrt{2\pi\sigma_k^2}} e^{\frac{(x_i-\mu_k)^2}{2\sigma_k^2}}, \theta_k = (\mu_k, \sigma_k) \tag{1}$$

Assuming that the observations are independent x_i, then the likelihood function for the parameter ψ is:

$$L(\psi) = \prod_{i=1}^{N} f(x_i|\psi) = \prod_{i=1}^{N}\prod_{k=1}^{K} \frac{\pi_i}{\sqrt{2\pi\sigma_k^2}} e^{\frac{(x_i-\mu_k)^2}{2\sigma_k^2}} \tag{2}$$

Where $\psi = (\pi_1, \pi_2, \ldots, \pi_k,)$ is a vector of parameters with $\Theta = \left(\pi_1, \pi_2, \pi_3, \ldots, \sigma_1^2, \sigma_2^2, \sigma_3^3, \ldots\right)$ which is a parameter of the mixture model. The EM algorithm is often used in iterative computational approaches of maximum likelihood, for example using Newton-Raphson. In general, the EM algorithm for the gaussian mixture model is described as follows

Step 1: Mean parameter initiation μ_j, variance σ_j^2, and prior distribution π_j

Step 2: At this stage it is said to be *E*-step, which is to calculate the value of y_{ij} using parameter values using the following equation:

$$y_{ij}^{(t)} = \frac{\pi_j^{(t)} f\left(x_i|\mu_j^{(t)}, \Sigma_j^{(t)}\right)}{\sum_{j=1}^{K} \pi_j^{(t)} f\left(x_i|\mu_j^{(t)}, \Sigma_j^{(t)}\right)} \tag{3}$$

Step 3: At this stage it is said to be *M*-Step where the re-estimation of the mean parameter μ_j, variance σ_j^2, and prior distribution π_j by updating each parameter using equation (4). Therefore, t is the iteration step. Furthermore, if the likelihood function is derived, equation 4. After that, the likelihood function is reduced $L(\psi|x)$against π_j where to fulfill $0 \leq \pi_j \leq 1$ dan $\sum_{i=1}^{N} \pi_j = 1$. This can be done using a Lagrange multiplier η and maximize the derivative in equations 6 and 7

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{N} y_{ij}^{(t)} x_i}{\sum_{i=1}^{N} y_{ij}^{(t)}} \tag{4}$$

$$\left[\sigma_j^2\right]^{(t+1)} = \frac{\sum_{i=1}^{N} y_{ij}^{(t)} \left(x_i - \mu_j^{(t+1)}\right)^2}{\sum_{i=1}^{N} y_{ij}^{(t)}} \tag{5}$$

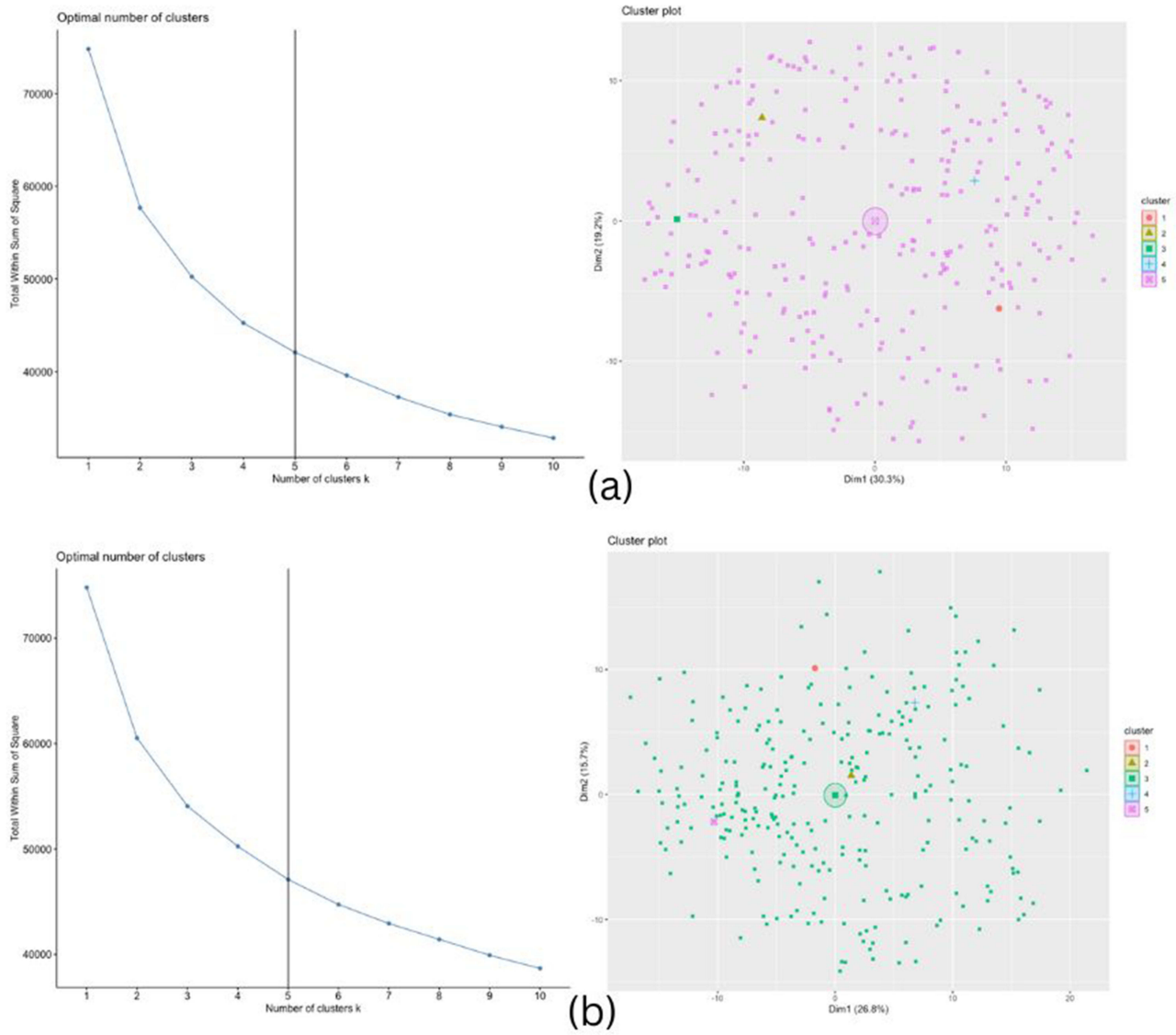

Figure 4 Best KMeans cluster age (**a**) and sex (**b**). Alt text: PCA biplot with two outliers ("2487_2" and "PrG_89"). Right: Scaled PCA with dense clusters and pink labels and PCA biplot with three outliers ("Vermis X", "LOcC_R_4_2", "IFG_161").

$$\frac{\partial}{\partial \pi_j}\left[L(\psi|x) - \eta\left(\sum_{j=1}^{K}\pi_j - 1\right)\right] = 0 \tag{6}$$

Step 4: Evaluate the log-likelihood function and check the convergence of the log-likelihood function and parameter values. If the convergence criteria error value <0.0001 has not been met, then repeat Step 2.

In the context of the clustering analysis conducted using the MCLUST VVI model with a Gaussian Mixture Model (GMM) and 5 components, the results in Table 1 offer valuable insights into the optimal configuration of clusters, particularly highlighting Cluster 1 with 37 voxels as a noteworthy outcome. This determination is grounded in the maximization of the log-likelihood, indicating a higher likelihood of the observed data under the specified clustering arrangement for Cluster 1, which has a log-likelihood value of 14634.

The assessment of clustering quality goes beyond log-likelihood, incorporating additional metrics such as the Bayesian Information Criterion (BIC) and Integrated Complete Likelihood (ICL). For Cluster 1, the BIC value is 13965.58, and the ICL value is 13865.47. These metrics are pivotal in evaluating the trade-off between model fit and complexity. Lower BIC and ICL values signify a better balance between capturing the nuances of the data and avoiding

Table 1 MCLUST VVI (Diagonal, Varying Volume and Shape) Model with 5 Components Using GMM

Cluster	Voxels	Log-Likelihood	n	df	BIC	ICL
1	37	14,634	274	2744	13,965.58	13,865.47
2	29					
3	47					
4	61					
5	100					

unnecessary complexity. BIC penalizes models for increased complexity, guiding towards a more parsimonious solution, while ICL further considers the structural meaningfulness of the clustering arrangement. Additionally, the associated degrees of freedom (df) and the number of voxels in each cluster contribute to the overall characterization of the clustering solution. In this specific instance, Cluster 1 comprises 37 voxels, suggesting a focused and distinct group within the dataset.

The collective findings underscore that Cluster 1, with its 37 voxels, represents the most optimal and coherent grouping based on the MCLUST VVI model and GMM approach. Researchers and practitioners can leverage this information to delve into the underlying structure of the data associated with this specific cluster, potentially uncovering meaningful patterns or associations that contribute to a deeper understanding of the dataset. The combination of log-likelihood, BIC, ICL, and cluster characteristics provides a comprehensive basis for making informed decisions about the most suitable clustering solution for the given data.

Scenario 2: Extreme Gradient Boosting (XGBoost)

XGBoost describes the gradient boosting introduced by.[21] Gradient boosting is an approach in which a new model is created that predicts the residuals or errors of the previous model and is then added together to make a final prediction.[22–24] It's called gradient enhancement because it uses a gradient descent algorithm to minimize losses when adding a new model with the following pseudo code

1. Initialize model with a constant value $f_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$
2. For $m = 1$ to M
   a. Compute so-called pseudo-residuals: for $i = 1, \ldots, n$

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(X_i)}\right], \text{where } F(x) = F_{m-1}(x)$$

   b. Fit a base learner (or weak learner) $g_m(x)$ to pseudo-residuals including the training dataset $\{(x_i, r_{im})\}_{i=1}^{n}$
   c. Compute the multiplier value of γ_m by solving the following one-dimensional problem

$$\gamma_m = argmin_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma g_m(x_i))$$

   d. Update the final model: $F_m(x) = F_{m-1}(x) + \gamma_m g_m(x)$
   e. Finding the output of $F_m(x)$

The provided equations delineate the intricate optimization process inherent in tree boosting models, specifically delving into the nuances of stage-wise first-order functional gradient descent updates, second-order functional gradient descent updates, and the application of Nesterov's accelerated descent across diverse loss functions. Given dataset with n examples and p features

$$f_m = \arg\min_f R(F_{m-1}) + dR(F_{m-1},f) + \frac{1}{2}f^2 = \arg\min_f dR(F_{m-1},f) + \frac{1}{2}f^2$$

$$dR(F,f) = \frac{d}{d\varepsilon}R(f + \varepsilon f)$$

$$dR(F_m,f) = g_m(Y,x)f(x)$$

$$g_m(Y,x) = \frac{\partial}{\partial F}L(Y,F)|_{F=F_{m-1}(x)}$$

However, we can reduce the equation to

$$f_m = \arg\min_{f\varepsilon S} E_{Y,X}(g_m(Y,X)f(X) + \frac{1}{2}f(X)^2)$$

$$= \arg\min_{f\varepsilon S} E_{Y,X}\Big(((-g_m(Y,X) - f(X))^2\Big)$$

The equations articulate a comprehensive approach, involving the minimization of a regularized risk functional, a directional derivative term (dR), and a quadratic regularization term. However, the complexity of these equations is pragmatically distilled, particularly within the framework of tree boosting models, where the objective transforms into finding the function f_m that minimizes the expected loss based on the data and the gradient of the loss function concerning the model's predictions. This streamlined representation aligns seamlessly with the practical implementation of tree boosting models, renowned for their efficacy in capturing non-linear dependencies within data.

The optimization process further unfolds through the construction of a tree structure rooted in the gradient, identification of terminal nodes, and the nuanced determination of weights assigned to these terminal nodes utilizing a step size within the gradient descent update. This tree-based methodology eloquently encapsulates the intricate relationships inherent in the data, making it a potent tool, especially apt for addressing problems characterized by non-linear dependencies.

In the insightful Figure 4, the analytical scope extends to the interpretation of the model's output employing SHAP values. This interpretative lens offers a nuanced understanding of feature importance, with specific emphasis on voxel connectivity. The mention of 26 best features, guided by SHAP values, serves as a beacon, spotlighting the pivotal role these features play in elucidating the intricacies of voxel connectivity patterns. This strategic incorporation of SHAP values not only enhances the model's interpretability but also illuminates the significance of individual features in comprehending the complex interplay influencing voxel connectivity. In essence, the amalgamation of advanced optimization principles, tree-based structures, and SHAP values contributes to a sophisticated analytical framework, elucidating the underlying intricacies of voxel connectivity within the dataset. Given the tree Structure based on the gradient, terminal nodes are found, and the weight of terminal nodes is found based on the step gradient descent uses step size. Figure 5 explains that there are 26 best features by looking at the SHAP value between voxel connectivity.

In the intricate landscape of neuroimaging analysis, a compelling narrative emerges as key brain regions take center stage, each unveiling its unique significance within the studied framework. Topping the list is the Fusiform gyrus in the Temporal lobe (FuG_R_3_1), boasting a substantial importance value of 0.201. Nestled in the right hemisphere, specifically within Region 3, Subregion 1, this neural hub showcases its pivotal role in the observed patterns or predictive models under scrutiny. Following suit is the Thalamus in the Subcortical nuclei (Tha_R_8_2), a dynamic region in the right hemisphere's Region 8, Subregion 2, carrying a weight of 0.087. Its nuanced contribution underscores the intricate interplay between subcortical structures and the broader analytical context.

Furthermore, the Superior temporal gyrus in the Temporal lobe (STG_R_6_1) emerges as a noteworthy contributor, bringing its influence to the forefront with an importance value of 0.073. Residing in the right hemisphere's Region 6, Subregion 1, this neural locale adds a layer of complexity to the understanding of the underlying neural dynamics. Together, these top-tier brain regions, as indicated by their associated numerical values, paint a vivid picture of their

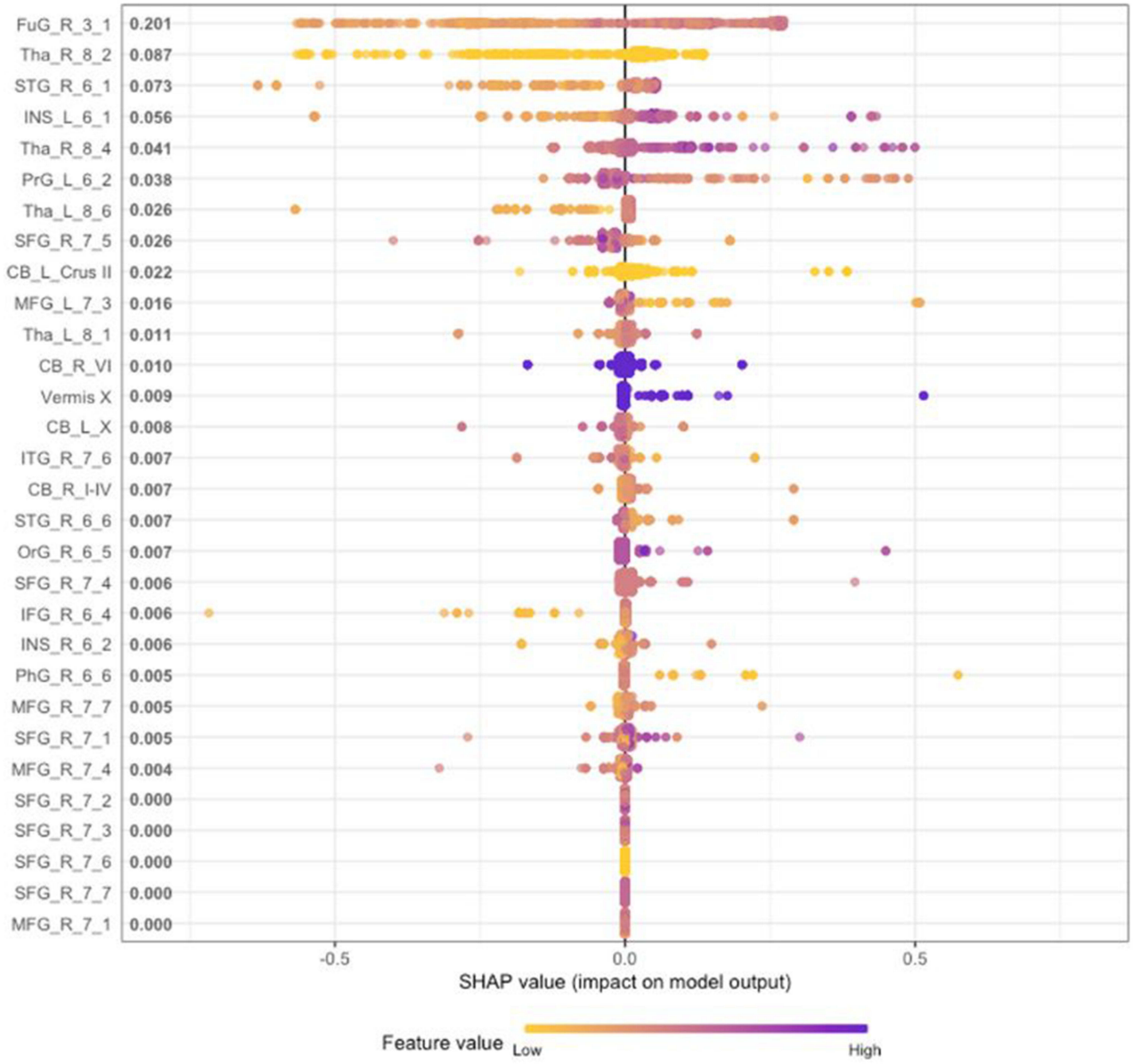

Figure 5 Best XGBoost after scaling.

respective roles in shaping the observed outcomes, providing researchers with valuable cues for further exploration. This nuanced understanding not only enhances our comprehension of neurological processes but also serves as a compass guiding future investigations in the intricate domain of neuroimaging and predictive modeling.

Alt text: Top features ranked by SHAP values. Colors represent feature values: orange (low) to purple (high). Right-side values increase model output; left-side values decrease it.

Scenario 3: Boruta

Feature selection is the process of identifying and removing features that are irrelevant and excessive. Features are considered relevant if their values vary systematically with category membership. This process is essential in the heart of machine learning because many machine learning algorithms experience decreases inaccuracy when the number of variables is significant but not optimal.[25] Boruta is one of new feature selection[26] and judges the information from all of the variable in our data set. Judge and elimination are based on distribution importance to the best shadow in Boruta.

Boruta's algorithm stands out as a robust and efficient tool for feature selection in ADHD research, significantly improving the precision and effectiveness of variable identification. Employing a novel methodology, Boruta conducts a thorough exploration of variable importance by introducing shadow attributes and utilizing a random forest classifier. This unique approach allows the algorithm not only to identify the most significant features associated with ADHD but also to streamline the dataset by eliminating less relevant or redundant attributes. The algorithm's strength lies in its robustness and consistency, achieved through an iterative refinement process that continues until reaching a predetermined limit or retaining only the "Confirmed" attribute. This iterative nature ensures stable outcomes, minimizing the impact of random fluctuations.

Crucially, Boruta's incorporation of shadow attributes serves as a preventative measure against overfitting, a vital consideration in ADHD research where model generalization is paramount. The algorithm's adaptability to diverse datasets, coupled with its user-friendly implementation and efficient resource utilization, positions Boruta as an invaluable and accessible tool for researchers. Its ability to unravel complex relationships between variables and ADHD is further exemplified by its systematic process, outlined in steps.[27–29]From augmenting the dataset with shadow attributes to employing a random forest classifier for *Z* scores and calculating Maximum *Z* scores among shadow attributes, to marking and eliminating attributes based on importance, Boruta's stepwise approach ensures a thorough exploration of variable importance. This systematic methodology guides the selection and elimination process, ultimately identifying the most relevant attributes for in-depth analysis. Overall, Boruta's versatility and methodical approach make it an indispensable asset in unraveling the intricate connections between variables and ADHD in research settings.

In the expansive realm of feature selection and predictive modeling, the Boruta algorithm has proven its mettle through 99 iterations within a remarkably efficient time span of 5.93401 minutes. The extensive outcome, featuring 270 attributes confirmed as important, reflects the algorithm's ability to discern crucial features, including notable cerebellar attributes such as `CB_L_Crus I`, `CB_L_Crus II`, `CB_L_I-IV`, `CB_R_Crus I`, and `CB_R_Crus II`, alongside a myriad of 265 additional significant attributes as shown in Figure 6. The absence of attributes labeled as unimportant underscores the robustness of the selected features in capturing the underlying complexities of the experimental data.

Exploring the intricate features of the left anterior cingulate gyrus, precuneus, and dorsomedial parietooccipital sulcus, such as `MVOcC _L_5_5`, IPL_R_6_2, `PCun_L_4_1`, and `PEr PCun_R_4_3`, holds paramount significance in understanding and addressing attention deficit hyperactivity disorder (ADHD). The left anterior cingulate gyrus,

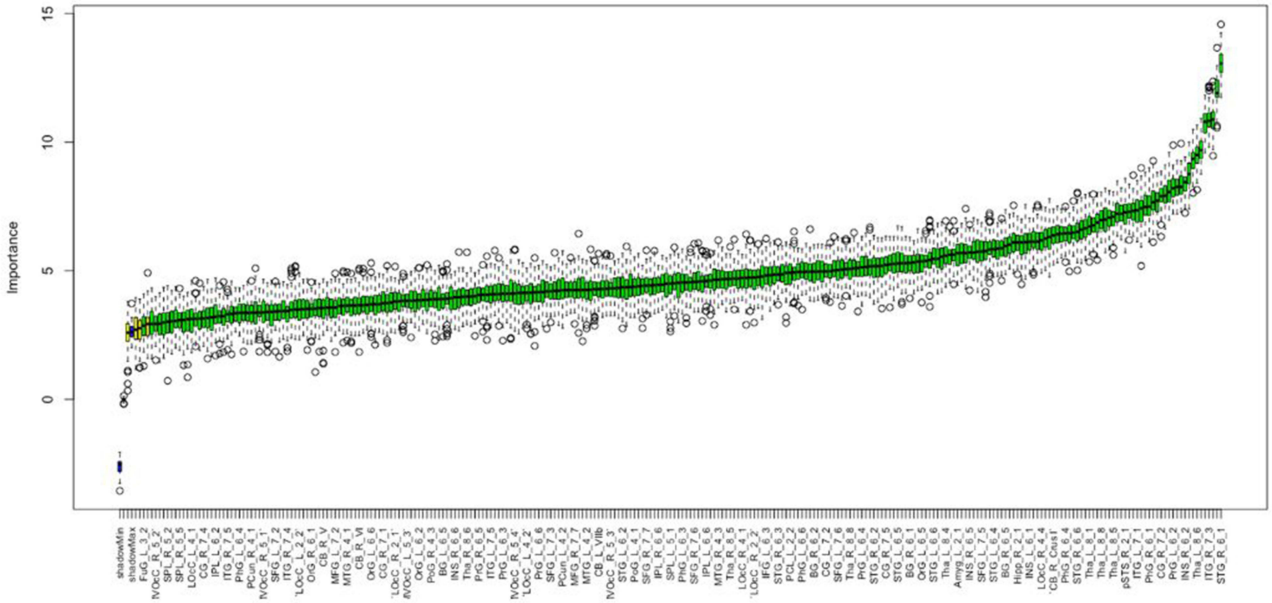

Figure 6 Best feature selection using Boruta. Alt text: Feature importance plot from the Boruta algorithm. Green points represent important features, black circles are shadow features, and yellow points are tentative features.

specifically associated with IPL_R_6_2, plays a pivotal role in error detection, conflict monitoring, and cognitive control—key aspects often impaired in individuals with ADHD. Similarly, the precuneus, represented by `PCun_L_4_1` and `PCun_R_4_3`, contributes to critical cognitive functions like self-awareness, episodic memory retrieval, and visuospatial processing, shedding light on potential neural bases for working memory challenges and attention deficits observed in ADHD patients. Additionally, the involvement of the dorsomedial parietooccipital sulcus (`PEr`) in visuospatial processing and attention orientation provides a valuable avenue for understanding the neurobiological underpinnings of attention difficulties in ADHD. Investigating and clarifying the definitive importance of these attributes in the context of ADHD not only deepens our comprehension of the disorder but also holds promise for refining targeted interventions and treatments, ultimately improving the lives of individuals affected by ADHD.

Scenario 4: DNN

Figure 7 provides a visual representation of the optimal feature selection outcomes, showcasing a seamless integration with a neural network. The achieved accuracies are noteworthy: Boruta combined with a Neural Network yields a balanced accuracy of 85.61%. The combination of Neural Network and Random Forest elevates the balanced accuracy to an impressive 89.14%. However, the pinnacle of performance is reached with Neural Network and Random Forest

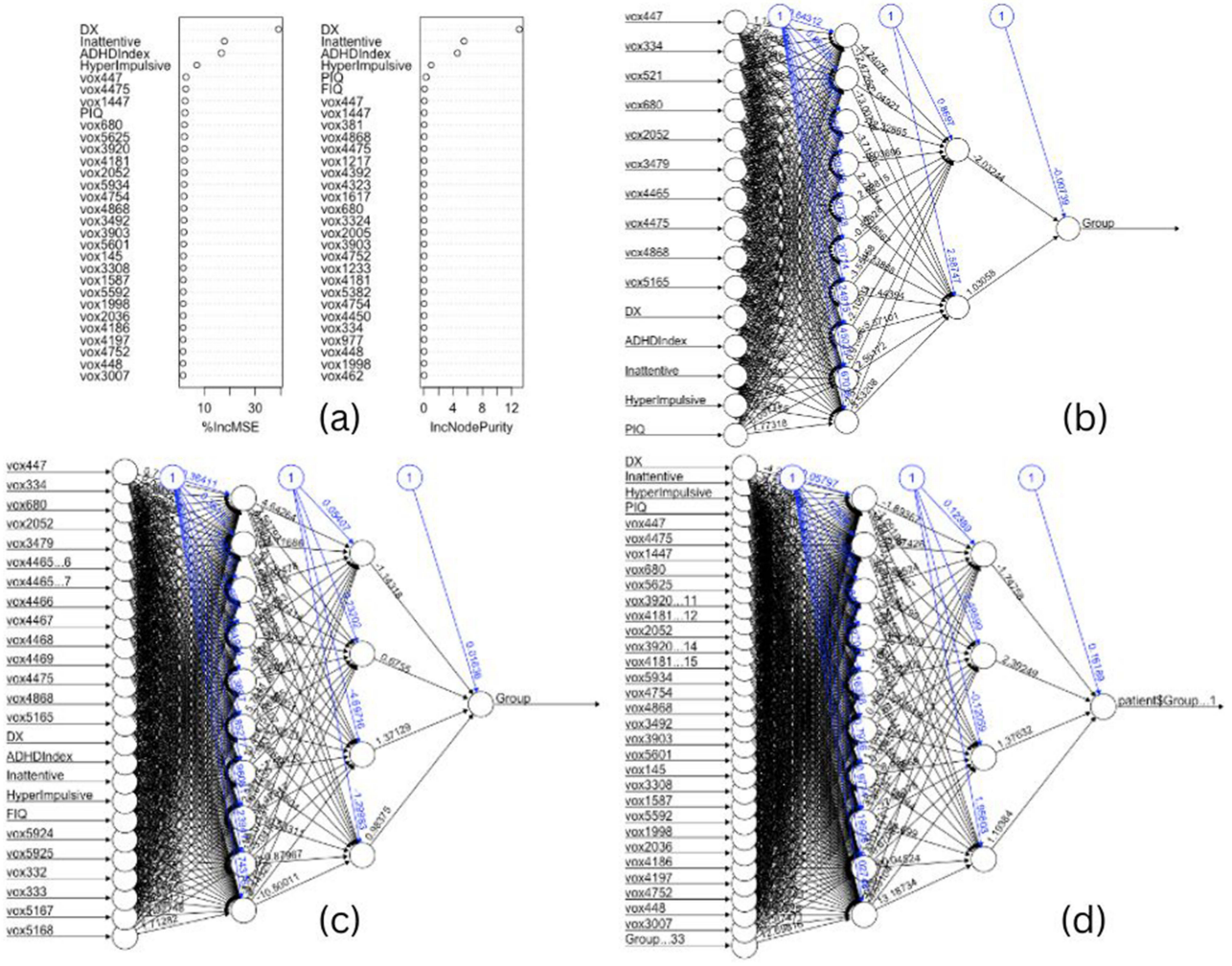

Figure 7 Best simulation using various machine learning. Alt text: This figure presents the performance and interpretability of neural networks using four different methods. (**a**) shows the SHAP plot, which illustrates the contribution of each feature to the model's predictions, highlighting both the magnitude and direction of their impact. (**b**) displays the LIME plot, offering local interpretability by showing how individual features influence specific predictions. (**c**) presents the DALEX plot, which visualizes feature importance and model performance across various feature values. Lastly, (**d**) shows the Partial Dependence Plot (PDP), illustrating the marginal effect of selected features on the predicted outcome, thereby revealing the relationships between input variables and model predictions.

DALEX, where the balanced accuracy soars to an outstanding 91.43%. This remarkable synergy underscores the potency of integrating Boruta feature selection with both neural network and Random Forest DALEX, contributing to a sophisticated and accurate predictive model.

Random Forest DALEX is an advanced extension of traditional Random Forest models, enhanced by the use of the DALEX (Descriptive mAchine Learning EXplanations) package. DALEX provides a framework for model-agnostic explanations and understanding of complex machine learning models. In the context of Random Forest, DALEX facilitates the interpretation of predictions by generating insightful visualizations and quantitative metrics. It allows for a more comprehensive understanding of the Random Forest model's behavior, feature importance, and the impact of individual features on predictions. The exceptional balanced accuracy achieved through the Neural Network and Random Forest DALEX integration underscores the efficacy of this approach in producing a highly accurate and interpretable predictive model in the experimental domain.

In our investigation, we have embarked on a thorough analysis of Deep Neural Networks (DNN) utilizing distinct activation functions, namely ReLU, Sigmoid, and Tanh. Activation functions play a pivotal role in shaping the behavior of neural networks, influencing their ability to capture complex patterns and make accurate predictions. Each activation function introduces non-linearity to the model, affecting the learning process and, consequently, the model's performance. This deliberate exploration extends beyond the mere selection of activation functions to encompass variations in the number of hidden layers, with a primary goal of discerning the profound influence of these choices on the outcomes of classification tasks. The culmination of our efforts is encapsulated in Figure 8 where we not only present crucial validation metrics but also meticulously detail the associated computing times. Remarkably, the average computing time across these scenarios stands at 7.73 hours, underscoring the comprehensive nature of our study.

Turning our attention to Figure 8 the detailed metrics for each activation function provide valuable insights. The Balanced Accuracy metrics, ranging from 89.54% to 90.14%, exemplify the nuanced impact of activation functions on the performance of the Deep Neural Network in the classification of ADHD voxel data. Each activation function contributes unique characteristics, and understanding their implications is pivotal for optimizing model performance.

Importantly, the significance of this analysis extends to the realm of ADHD voxel classification, where precision in distinguishing patterns related to ADHD is of paramount importance. The choice of activation functions and the

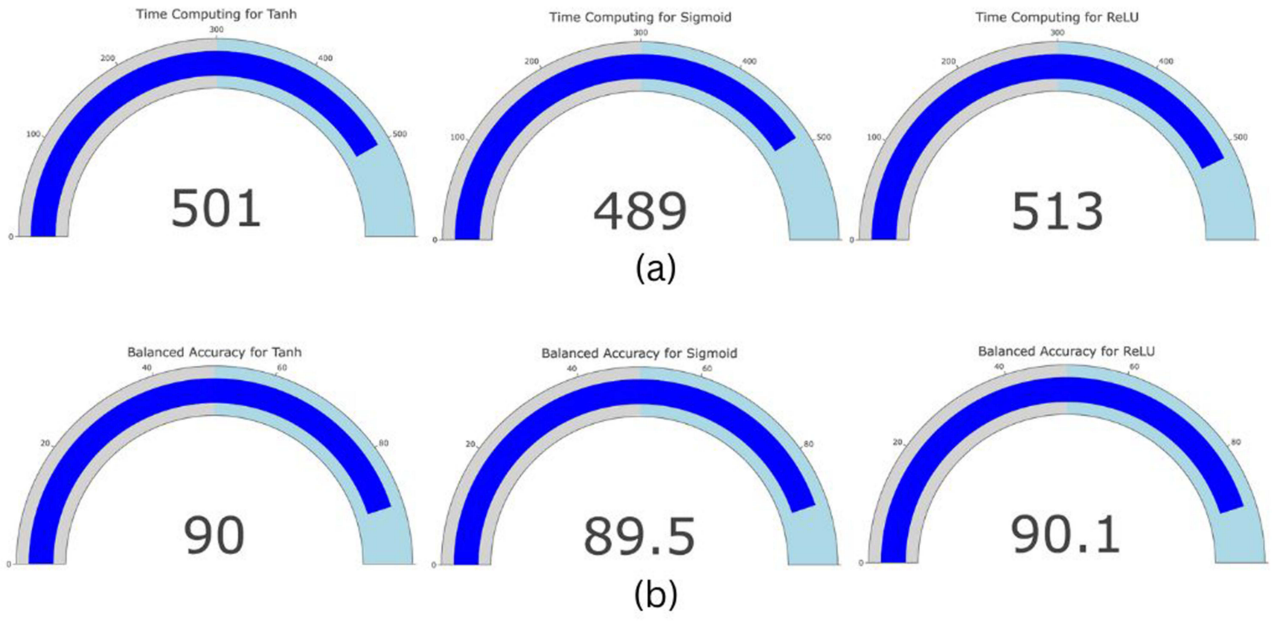

Figure 8 Deep neural network metric evaluation time computing (**a**) and balanced accuracy (**b**). Alt text: This figure compares two key performance metrics of the deep neural network. (**a**) illustrates the time required for model computation, providing insights into the efficiency and scalability of the neural network. (**b**) presents the balanced accuracy, which accounts for class imbalances, offering a more reliable measure of model performance across different categories. Together, these metrics highlight the trade-off between computational speed and predictive accuracy.

configuration of hidden layers can significantly influence the network's ability to capture intricate patterns inherent in voxel data associated with ADHD. Achieving high Balanced Accuracy, as demonstrated in our findings, not only validates the effectiveness of the chosen activation functions but also underscores their potential in enhancing the precision and reliability of ADHD classification models. In essence, our study contributes not only to the broader discourse on machine learning techniques but also holds direct implications for advancing our understanding and diagnostic capabilities in ADHD research through the analysis of voxel data.

Conclusion

In this paper, our focus revolves around the application of diverse machine learning and data science methodologies for feature selection, aiming to identify crucial variables in the predictive analysis of brain voxels related to Attention Deficit Hyperactivity Disorder (ADHD) in patients. The utilization of multiple methods yields varying accuracy rates and provides insights into the significance of different variables. Notably, our forthcoming research endeavors will delve deeper into the application of Convolutional Neural Networks (CNNs) with a more diverse pooling approach, aiming to enhance the precision and effectiveness of ADHD detection through advanced neural network architectures.

ADHD, characterized by symptoms of restlessness, hyperactivity, low attention span, impulsivity, and destructiveness, poses a significant challenge in children, potentially impacting their academic achievements and learning processes at school. The imbalance syndrome associated with ADHD underscores the necessity for special attention and therapeutic interventions to ensure the proper development of affected children. Our ongoing research aims to identify dominant factors contributing to the disorder, exploring genetic predispositions, exposure to chemicals and viruses, potential complications during pregnancy and childbirth, and other conditions that may affect brain tissue development.

Beyond heredity, the role of the social environment emerges as a significant factor in ADHD, as highlighted in various studies. Improper utilization of audio-visual information technology, including television, computers, and gadgets, is also under scrutiny as a potential exacerbating factor for the onset of ADHD symptoms. It's important to note that these symptoms can manifest in children with normal neurological conditions, and parenting practices may contribute to their occurrence. As we navigate the multifaceted landscape of ADHD research, our holistic approach encompasses not only the refinement of predictive models through advanced machine learning techniques but also the exploration of environmental, genetic, and lifestyle factors contributing to the disorder. By integrating diverse methodologies and addressing the complexity of ADHD, our research endeavours aim to advance our understanding and inform effective therapeutic strategies for individuals affected by this prevalent neurodevelopmental condition.

Data Sharing Statement

Data that support the findings of this study are available upon reasonable request.

Ethics Approval

Not applicable; we are utilizing secondary data for our research, which has been previously collected and does not involve any direct interaction with human subjects.

Informed Consent

Not Applicable; are utilizing secondary data for our research, which has been previously collected and does not involve any direct interaction with human subjects.

Acknowledgments

Rezzy Eko Caraka and Rung Ching Chen contributed equally as corresponding authors for this study.

Author Contributions

Rezzy Eko Caraka: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Project Administration, Visualization. Khairunnisa Supardi: Methodology, Data Curation, Writing – Original Draft, Project Administration. Prana Ugiana Gio: Writing – Review & Editing. Vijaya Isnaniawardhani: Writing –

Review & Editing. Rung Ching Chen: Writing – Review & Editing. Bekti Djatmiko: Writing – Review & Editing. Bens Pardamean: Writing – Review & Editing. All authors contributed to the interpretation of data, took part in drafting, revising, or critically reviewing the article, approved the final version to be published, agreed on the journal to which the article has been submitted, and agree to be accountable for all aspects of the work.

Funding

REC is partially supported by the Padjadjaran University and Telkom University. Rung Ching Chen is partially supported by the National Science and Technology Council, Taiwan, under Grant NSTC-112-2221-E-324-003-MY3 and Grant NSTC-112-2221-E-324-011-MY2.

Disclosure

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bledsoe J, Xiao D, Chaovalitwongse A, et al. Diagnostic classification of ADHD versus control: support vector machine classification using brief neuropsychological assessment. *J Atten Disord*. 2017;21:1040–1049. doi:10.1177/1087054716649666
2. Agustini M, Yufiarti, Wuryani, Yufiarti Y, Wuryani W. Development of learning media based on android games for children with attention deficit hyperactivity disorder. *Int J Interactive Mobile Technol*. 2020;14(6):205–213. doi:10.3991/IJIM.V14I06.13401
3. Cho SC, Kim JW, Choi HJ, et al. Associations between symptoms of attention deficit hyperactivity disorder, depression, and suicide in Korean female adolescents. *Depress Anxiety*. 2008;25(11):E142–E146. doi:10.1002/da.20399
4. Castanho EN, Aidos H, Madeira SC. Biclustering fMRI time series: a comparative study. *BMC Bioinf*. 2022;23(1). doi:10.1186/s12859-022-04733-8
5. Whi W, Ha S, Kang H, Lee DS. Hyperbolic disc embedding of functional human brain connectomes using resting state fMRI. *bioRxiv*. 2022;6(3):745–64. doi:10.1101/2021.03.25.436730
6. Whi W, Huh Y, Ha S, Kang H, Lee H, Lee DS. Characteristic functional cores revealed by hyperbolic disc embedding and k-core percolation on resting-state fMRI. *Sci Rep*. 2022;12(1). doi:10.1038/s41598-022-08975-7
7. Rahaman MA, Rodrigue A, Glahn D, Turner J, Calhoun V. Shared sets of correlated polygenic risk scores and voxel-wise grey matter across multiple traits identified via bi-clustering. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*.; 2021:2201–2206. doi:10.1109/EMBC46164.2021.9630825.
8. Zanderigo F, Ogden RT, Bertoldo A, Cobelli C, Mann JJ, Parsey RV. Empirical Bayesian estimation in graphical analysis: a voxel-based approach for the determination of the volume of distribution in PET studies. *Nucl Med Biol*. 2010;37(4):443–451. doi:10.1016/j.nucmedbio.2010.02.004
9. Caraka RE, Lee Y, Chen RC, Toharudin T. Using hierarchical likelihood towards support vector machine: theory and its application. *IEEE Access*. 2020;8:194795–194807. doi:10.1109/ACCESS.2020.3033796
10. Tsoumakas G, Katakis I. Multi-label classification: an overview. *Int J Data Warehousing Mining*. 2007;3(3):1–13. doi:10.4018/jdwm.2007070101
11. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. *Data Min Knowl Discov*. 2019;33(4):917–963. doi:10.1007/s10618-019-00619-1
12. Caterini AL, Chang DE. Recurrent neural networks. In: *Deep Neural Networks in a Mathematical Framework*. Springer International Publishing; 2018:59–79. doi:10.1007/978-3-319-75304-1_5
13. Kim Y. Convolutional neural networks for sentence classification. In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.; 2014. doi:10.3115/v1/d14-1181.
14. Kamada S, Ichimura T. An adaptive learning method of deep belief network by layer generation algorithm. 2018;2–5.
15. Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: *Proceedings of 2014 Science and Information Conference, SAI 2014*.; 2014:372–378. doi:10.1109/SAI.2014.6918213.
16. Ferizal R, Wibirama S, Setiawan NA. Gender recognition using PCA and LDA with improve preprocessing and classification technique. In: *Proceedings - 2017 7th International Annual Engineering Seminar, InAES 2017*.; 2017:1–6. doi:10.1109/INAES.2017.8068547.
17. Sałat R, Osowski S, Siwek K. Principal component analysis (PCA) for feature selection at the diagnosis of electrical circuits. *Przeglad Elektrotechniczny*. 2003;79(10):667–670.
18. Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput*. 1998;10(5):1299–1319. doi:10.1162/089976698300017467
19. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;2(4):433–459. doi:10.1002/wics.101
20. Vigneau E, Qannari EM. Clustering of variables around latent components. *Commun Statistics - Simulation Computation*. 2003;32(4):1131–1150. doi:10.1081/SAC-120023882
21. Liu Y, Just A. SHAPforxgboost: SHAP plots for "XGBoost". 2020;1–21.
22. Cenggoro TW, Mahesworo B, Budiarto A, Baurley J, Suparyanto T, Pardamean B. Features importance in classification models for colorectal cancer cases phenotype in Indonesia. *Procedia Comput Sci*. 2019;157:313–320. doi:10.1016/j.procs.2019.08.172
23. Nielsen D. Tree Boosting With XGBoost – why Does XGBoost Win "Every" Machine Learning Competition. *Tree Boosting with XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition?* 2016.
24. Gumus M, Kiran MS. Crude oil price forecasting using XGBoost. *2nd International Conference on Computer Science and Engineering, UBMK 2017*. 2017:1100–1103. doi:10.1109/UBMK.2017.8093500.
25. Ma H, Smith LA. Practical Feature Subset Selection for Machine Learning. 1998.

26. Kursa MB, Jankowski A, Rudnicki WR. Boruta - A system for feature selection. *Fundam Inform*. 2010;101(4):271–285. doi:10.3233/FI-2010-288
27. Caraka RE, Nugroho NT, SKuo T, RChing C, Toni T, Bens P. Feature importance of the aortic anatomy on endovascular aneurysm repair (EVAR) using boruta and bayesian MCMC. *Commun Math Biol Neurosci*. 2020;2020(1–23). doi:10.28919/cmbn/4584
28. Ahmadpour H, Bazrafshan O, Rafiei-sardooi E, Zamani H, Panagopoulos T. Gully erosion susceptibility assessment in the kondoran watershed using machine learning algorithms and the boruta feature selection. *Sustainability*. 2021;13(10110):1–23. doi:10.3390/su131810110
29. Ebrahimi-Khusfi Z, Nafarzadegan AR, Dargahian F. Predicting the number of dusty days around the desert wetlands in southeastern Iran using feature selection and machine learning techniques. *Ecol Indic*. 2021;125(November 2020):107499. doi:10.1016/j.ecolind.2021.107499