End-to-end Stroke imaging analysis, using reservoir computing-based effective connectivity, and interpretable Artificial intelligence.

Wojciech Ciezobka[a,b], **Joan Falcó-Roget**[a], **Cemal Koba**[a] **and Alessandro Crimi**[a,b,*]

[a]Sano, center for computational medicine, Kraków, Poland
[b]AGH University of Krakow, Kraków, Poland
ORCID (Wojciech Ciezobka): https://orcid.org/0000-0003-2972-710X, ORCID (Joan Falcó-Roget): https://orcid.org/0000-0002-9410-6361, ORCID (Cemal Koba): https://orcid.org/0000-0001-7097-1441, ORCID (Alessandro Crimi): https://orcid.org/0000-0001-5397-6363

Abstract. In this paper, we propose a reservoir computing-based and directed graph analysis pipeline. The goal of this pipeline is to define an efficient brain representation for connectivity in stroke data derived from magnetic resonance imaging. Ultimately, this representation is used within a directed graph convolutional architecture and investigated with explainable artificial intelligence (AI) tools.

Stroke is one of the leading causes of mortality and morbidity worldwide, and it demands precise diagnostic tools for timely intervention and improved patient outcomes. Neuroimaging data, with their rich structural and functional information, provide a fertile ground for biomarker discovery. However, the complexity and variability of information flow in the brain requires advanced analysis, especially if we consider the case of disrupted networks as those given by the brain connectome of stroke patients. To address the needs given by this complex scenario we proposed an end-to-end pipeline. This pipeline begins with reservoir computing causality, to define effective connectivity of the brain. This allows directed graph network representations which have not been fully investigated so far by graph convolutional network classifiers. Indeed, the pipeline subsequently incorporates a classification module to categorize the effective connectivity (directed graphs) of brain networks of patients versus matched healthy control. The classification led to an area under the curve of 0.69 with the given heterogeneous dataset. Thanks to explainable tools, an interpretation of disrupted networks across the brain networks was possible. This elucidates the effective connectivity biomarker's contribution to stroke classification, fostering insights into disease mechanisms and treatment responses. This transparent analytical framework not only enhances clinical interpretability but also instills confidence in decision-making processes, crucial for translating research findings into clinical practice.

Our proposed machine learning pipeline showcases the potential of reservoir computing to define causality and therefore directed graph networks, which can in turn be used in a directed graph classifier and explainable analysis of neuroimaging data. This complex analysis aims at improving stroke patient stratification, and can potentially be used with other brain diseases.

* Corresponding Author. Email: a.crimi@sanoscience.org

1 Introduction

Stroke is one of the leading causes of morbidity and mortality worldwide. Accurate classification can aid in effective treatment and management. Magnetic resonance imaging (MRI) has emerged as a powerful tool for stroke diagnosis, providing detailed images of brain structures and abnormalities. However, the analysis of MRI data poses significant challenges due to its complexity and the need for efficient and reliable classification algorithms, especially when we want to understand the dynamics of the brain.

The classification of stroke using medical images has been the primary focus of previous studies [42, 2]. However, most of the approaches carried out so far are focused on the extent of lesions and limited correlation to functional damages such as aphasia and motor deficits [11]. Recent studies have started investigating the brain's inner functioning from the point of view of the influence of one brain region on another one, and how lesions compromise those interactions [3, 2]. Indeed, brain connectivity encompasses the complex interactions between neurons and their intricate network of connections. It is a broad term that encompasses connections between neurons at various levels of granularity and with different connection characteristics. Within this domain, three distinct types of connectivity have emerged: structural (SC), functional (FC), and effective connectivity (EC). Each of these holds clinical and predictive value, offering valuable insights into the brain's intricate workings [44]. Effective connectivity investigates the causal link between the time series of two regions of the brain and can be represented as directed graphs. Classification and explanation of directed graphs have not been fully investigated and the study of stroke with those tools provides the opportunity to create a pipeline exploring all those elements.

More specifically, local ischemia damages neurons and structural neural connections at the site of injury. This affects primarily subcortical regions, subsequently altering long-range functional connectivity between cortical areas. Decreases in functional connectivity alterations suggest deficits but cannot reveal the directionality or time scale of the information flow, leaving several open questions related to the directionality and functioning of the brain after a non-traumatic injury such as a stroke. Allegra and colleagues carried out previous

studies where this transfer of information view of the brain of stroke patients was investigated through Granger Causality (GC) analyses [3], where they observed a significant decrease in inter-hemispheric information transfer in stroke patients compared to matched healthy controls. GC has been used largely in computational neuroscience studies due to its low computational costs compared to other methods [20, 46]. Practically, the method estimates autoregressor variables relating to different time series which are then further validated by F-statistics to establish causality. Yet, due to the potential confounding characteristics that each autoregressor may generate [33]), there are still ongoing disagreements on whether this can help define causal interaction between brain regions [36] using this framework, and some authors consider GC as just a relation measures [16]. To overcome these limitations, researchers have explored the use of reservoir computing in a completely detached paradigm to extract causality [25, 17]. Reservoir computing is a computational framework that leverages the dynamics of recurrent neural networks to process and classify complex temporal data effectively, by exploiting the inherent memory and non-linear dynamics of reservoirs [26, 31]. It has also been used to classify electroencephalography data from stroke patients [6], though as a classifier itself, not to estimate the structure of the human brain.

Finally, capturing both spatial and temporal patterns can help understand stroke beyond traditional voxel-based lesion-symptom mapping [5] to consider specific information transfer and interactions in the brain [18, 17]. Technically, this will produce a directed graph representation that can be classified and explored with explainable AI tools.

In summary, using reservoir computing we i) defined causality in stroke patients, and, given the generated representation of causality as directed graphs, investigated ii) the value of the resulting directed maps together with their classification, and iii) the explainability of the classification to provide insights into the overall brain network disruption in stroke patients (Fig. 1). To our knowledge, no study has classified directed graphs and explained their significance in computational neuroscience and neurology. Thus, incorporating these features into classification algorithms could improve stroke diagnosis accuracy and efficiency.

2 Methods

2.1 Data and preprocessing

The dataset was previously collected by the School of Medicine of the Washington University in St. Louis and complete procedures can be found in [10]. They collected MRI data and behavioral examinations of stroke patients and healthy controls. The imaging data comprise structural and functional MRI from controls and patients suffering from hemorrhagic and ischemic stroke. Acquisitions were done within the first two weeks of the stroke onset (i.e., acute). Structural scans include T1-weighted, T2-weighted, and diffusion tensor images. Functional images include a resting state paradigm. Scanning was performed with a Siemens 3T Tim-Trio scanner. Briefly, we closely followed the pre-processing steps outlined in [27]. Following a quality control of *fMRIPrep* outputs, 104 stroke subjects and 26 control subjects were qualified for further analysis. For our purposes, it suffices to say that structural scans were used in combination with functional acquisitions to co-register all participants into a common template. Gray matter signal was finally obtained after artifact removal and parcellated into 100 regions of interest (ROIS) [15, 38]. For every subject and patient, these 100 time series (i.e., one for each ROI) were fed into the pipeline outlined below to obtain the subject-specific effective connectivity maps.

The dataset is not public but it is available upon request to the original authors [10]. The used code is instead available at the URL https://github.com/Wotaker/Effective-Connectivity-Reservoir-Computing.

2.2 Reservoir computing

Reservoir computing networks (RCN), despite being known for more than two decades, have been largely eclipsed by other frameworks. A reservoir network is a set of artificial neurons that are randomly connected between themselves thus forming a recurrent architecture [26, 31]. Sometimes this is also called *echo-state network* since the internal dynamics of the reservoir (or "echo state") maintain information about the system's input history. In this framework, an input series $\mathbf{u}_t$ is fed into this high dimensional dynamical system of N units through a non-linear activation function,

$$\mathbf{r}_t^{in} = f^{in}(\mathbf{W}^{in}\mathbf{u}_t), \tag{1}$$

where $\mathbf{W}^{in}$ is an N x $(N_{in}+1)$ matrix of random weights including biases, N_{in} is the dimensionality of the multivariate input, and f_{in} is the non-linearity. At each time step t the former projection is used to drive the reservoir units $\mathbf{r}_t$. The current state of each unit is a combination of the past states as well as the current input,

$$\mathbf{r}_t = (1-\lambda)\mathbf{r}_{t-1} + \lambda f(\mathbf{r}_t^{in} + \mathbf{W}\mathbf{r}_{t-1}), \tag{2}$$

where $\mathbf{W}$ is an N x N matrix of random weights, and λ is the leakage that controls the importance of the reservoir's history to the current time stamp t. The final component of the reservoir is a set of *readout* weights $\mathbf{W}^{out}$ that extract information from the hidden state and map onto specific predictions. That is,

$$\mathbf{y}_t = \mathbf{W}^{out}\mathbf{r}_t. \tag{3}$$

The predictions $\mathbf{y}_t$ might be of arbitrary dimension N^{out} and, importantly, are linear w.r.t. to the reservoir states. Within this paradigm, only that readout weights $\mathbf{W}^{out}$ are trained via incremental linear regression optimization [45, 28],

$$\mathbf{W}^{out} = (\mathbf{R}\mathbf{R}^T + \alpha\mathbf{I})^{-1}(\mathbf{Y}\mathbf{R}^T), \tag{4}$$

with α being a regularization parameter, $\mathbf{R}$ is the matrix obtained after concatenating all the reservoir states, and $\mathbf{Y}$ contains the outputs. Once again, the readout weights contain a set of N_{out} biases.

Noteworthy, as opposed to other architectures suited for time series forecasting, only a reduced set of output weights needs to be trained, thus increasing its computational efficiency. The random weights $\mathbf{W}^{in}$ are drawn from a uniform distribution bounded between -1 and 1. The recurrent connections are drawn from a standard normal distribution and are later scaled by the spectral radius. The latter largely ensures that the network possesses the echo-state property, although there is recent evidence disagreeing with this aspect [52, 13].

Briefly, the main idea behind reservoir-like computing is that a given input pushes the reservoir to specific locations in a high-dimensional manifold [26, 31]; the output weights are then optimized to retrieve information from the nearby regions. Were the input to move the reservoir away to other points, the output weights would not be able to recover meaningful information hence completely missing the prediction. Further evidence suggests that RCNs supersede

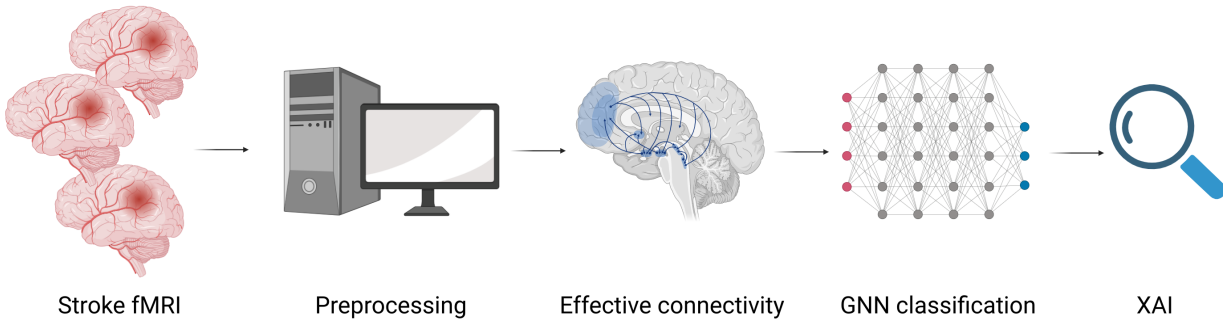

Figure 1. Overall pipeline of the study, where MRI data are preprocessed, used to define an effective connectivity representation, classified and the results are investigated by explainable AI tools.

deep learning-based models for temporal series prediction even on the verge of chaos [40]. Richer approaches aim to train the reservoir connections themselves and have been proven to be useful in understanding the dynamical properties of cortical networks [35], offering an interesting framework for similar use cases. The parameter values used in our experiments can be found in Table 1.

Object	Input-to-Node	Node-to-Node
Units (#)	50	50
Sparsity	1	1
Activation	logistic	tanh
Scaling	1	*NA*
Shift	0	*NA*
Bias scaling	1	*NA*
Bias shift	0	*NA*
Random seed	*null*	*null*
Spectral radius	*NA*	1
Leakage (λ)	*NA*	1
Bidirectional	*NA*	*false*

Table 1. Summary of the parameters chosen to train the Reservoir Computing Networks (RCNs) in this work. For the two different blocks, *NA* stands for Not Applicable, and *null* indicates that the value was left empty to be chosen by the implemented random sampler. For further details on the meaning of each one of these parameters we refer the reader to the original publication of the package [45] and documentation.

2.3 *Reservoir computing networks to map causal interactions in lesioned brains*

Traditionally, effective connectivity in neuroimaging can be estimated in different ways, as dynamic causal modeling [20], GC [23], continuous-time implementations [21], or information theory [50]. Granger-like interpretations are often preferred due to their relative computational costs and implementation, though they are not exempt from controversy [24] thus justifying alternative approaches.

An unrelated proposal relies on the properties of the state-space of the dynamical system to reconstruct asymmetric mappings between delayed embeddings of each component of the system [46]. That is, it leverages Taken's theorem to find the optimal neighborhood as well as the exact delay at which the reconstruction is optimal. Recent extensions [48, 51, 7] have incorporated non-linear methods as well as reducing the number of ad-hoc parameters. Most prominently, reservoir computing has proven to be an efficient and accurate alternative to automatize the process almost in its entirety [25].

Let's consider the relationship between two one-dimensional variables, x and y, where it hypothesizes that the delay at which interactions take place is not smaller than the sampling rate (e.g., Time of Repetition in functional MRI). The prediction skill, denoted by $\rho_{x\to y}(\tau)$, is defined as the Pearson correlation between the true time series, $\mathbf{y}(t+\tau)$, and the predicted series $\hat{\mathbf{y}}(t+\tau)$ from the input $\mathbf{x}(t)$.

$$\rho_{x\to y}(\tau) := \text{corr}\ [\mathbf{y}(t+\tau), \hat{\mathbf{y}}(t+\tau)]\,. \tag{5}$$

Noteworthy, the Pearson correlation between the true and reconstructed series (ρ) is used to estimate directedness, though other metrics like mean squared error could also be used. Directionality can still be assessed using the same hypothesis testing mechanisms [48].

Moreover, the time series are fed into the reservoir *all-at-once*, letting the network project all of them. The neighboring points in the variable's embedding are then remapped to the target embedding via the training of the output weights. It should noted that this represents a deviation from more canonical usages [25, 13]. To investigate the causal relationship between variables, we first calculate both $\rho_{x\to y}(\tau)$ and $\rho_{y\to x}(\tau)$ in a given temporal domain. We then examine the values of τ at which either $\rho_{x\to y}(\tau)$ or $\rho_{y\to x}(\tau)$ reaches its peak value [51, 25]. To streamline the subsequent description, we introduce the following notation:

$$\begin{aligned} \tau_{x\to y} &:= \arg\min_{\tau} \rho_{x\to y}(\tau) \\ \tau_{y\to x} &:= \arg\min_{\tau} \rho_{y\to x}(\tau). \end{aligned} \tag{6}$$

Empirically, directionality is then defined as follows [46]:

- if $\tau_{x\to y}$ is positive, and $\tau_{y\to x}$ is negative, we say that x causes y;
- if $\tau_{x\to y}$ is negative, and $\tau_{y\to x}$ is positive, we say that y causes x;
- if both $\tau_{x\to y}$ and $\tau_{y\to x}$ are negative, we say that x and y causes each other.

Despite seeming counterintuitive, information of $\mathbf{y}$ is present in earlier observations of $\mathbf{x}$ and, consequently, that current information of the cause $\mathbf{x}$ is useful to predict future observations of the consequence $\mathbf{y}$ (see [46] for a comprehensive explanation). In certain systems, predictability scores peak at negative lags $\tau < 0$ for both directions, being the height of the peaks informative of the coupling strength [25]. However, the existence of this bidirectionality does not necessarily invalidate the former statements [17].

It was quickly noted that in large and noisy networks, such as the brain, it is unlikely that the predictability scores in Eq. 5 reach clear and distinct peaks. Functional signals are notoriously noisy [43], and indeed prediction with this approach is challenging [4]. A solution to this issue relies on assessing the minimal requirements that are needed to suggest causal interactions [17]. For that, the difference

	$\tau > 0$	$\tau < 0$
$\Delta_{x \to y}(\tau) > 0$	$x \to y$	$y \to x$
$\Delta_{x \to y}(\tau) < 0$	$y \to x$	$x \to y$

Table 2. Potential causal directions based on the sign of Δ-score and the positive or negative τ regime.

between prediction scores should be evaluated and contrasted with proper surrogate predictions [34, 39, 30]. That is,

$$\Delta_{x \to y}(\tau) := \rho_{x \to y}(\tau) - \rho_{y \to x}(\tau), \tag{7}$$

which can be interpreted as an indication of the potential causality direction (Table 2). The scores in Eqs. 5 and 7 can be contrasted against the 95% confidence interval obtained from a surrogate testing procedure [17]. It has been shown that the requirements to define causality can be compressed into a reduced set of δ-scores [17],

$$\delta_{x \to y}(\tau) := \begin{cases} (1 - p_{\rho_{x \to y}(\tau) > 0})(1 - p_{\Delta_{x \to y}(\tau) > 0}) & \text{if } \tau > 0 \\ (1 - p_{\rho_{y \to x}(\tau) > 0})(1 - p_{\Delta_{y \to x}(\tau) > 0}) & \text{if } \tau < 0 \end{cases} \tag{8}$$

for directed interactions, and

$$\delta_{y \leftrightarrow x}(\tau) := (1 - p_{\rho_{x \to y}(\tau) > 0})(1 - p_{\rho_{y \to x}(\tau) > 0}) p_{\Delta_{x \to y} \neq 0} \tag{9}$$

for bidirectional interactions. p_{H_1} is the p-value after testing the alternative hypothesis H_1 against the surrogate data (Fig. 2). For instance, $p_{\rho_{x \to y}}$ is a p-value for the hypothesis that x influences y. The values of the δ-scores range from 0 to 1, with higher values indicating greater confidence in the existence of a causal relationship with a coupling delay of τ between the examined variables.

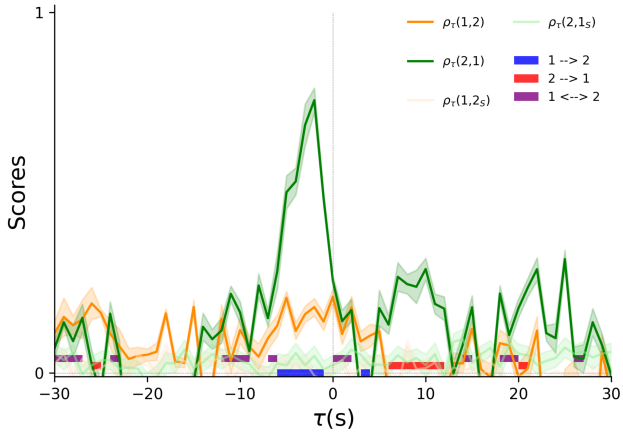

Figure 2. Predictability scores from an the same chaotic system defined in [46, 17]. Solid lines show the predictability in Eq. 5 between embeddings. Shaded regions show 1 standard error of the mean. Transparent lines show the predictability of the surrogate system, which is used to define the expected level of chance against which the hypotheses are tested. In this academic example, it can be said that strong asymmetric interactions between two time series exist at different temporal lags.

Then, for a given lag τ, a matrix $\mathbf{A}_\tau$ collects the δ-scores, where each element $[x, y]$ represents the causal relationship from node signal x to ROIs signal y,

$$A_\tau[x, y] = \begin{cases} \delta_{x \to y}(\tau) & \text{unidirectional} \\ \delta_{x \to y}(\tau) + \delta_{x \leftrightarrow y}(\tau) & \text{bidirectional.} \end{cases} \tag{10}$$

The effective connectivity (RC) matrix $\mathbf{A}_\tau$ is a final representation of the effective connectivity network of every subject; it is directed, non-symmetric, and can incorporate bidirectional causality connections. For our experiments, for every possible interaction $x \to y$, we trained 20 different reservoirs and tested against 100 shuffled targets, strictly following what was outlined in [17]. Furthermore, only unidirectional connections were kept from the adjacency matrix in Eq. 10.

In our experiments, we investigated the classification of pathological groups with the effective connectivity matrices used as features (Fig. 3 TOP), and we also compared those to the effective connectivity matrices obtained by Granger causality, representing one of the state-of-art approaches. As a last step, for each entry $A_\tau[x, y]$, we standardized all samples by subtracting the mean connectivity of the control group and dividing by the standard deviation. Finally, these standardized causal relationships (i.e., directed graphs) were fed into two simple graph classifiers to explore and explain the most informative nodes and links to detect stroke occurrence.

2.4 *Graph convolutional neural networks*

Graph convolutional neural networks (GNNs) are a variation of traditional convolutional neural networks which capitalize on graph data representations and can learn non-trivial representations by leveraging the complex topological organization of the data [49]. Intuitively, a graph constitutes a non-Euclidean geometric space where complex relationships between data points can be embedded and forwarded as inputs into a GNN [8]. More formally, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined as a set of nodes $\mathcal{V} = \{1, \ldots, n\}$ and a set of edges $\mathcal{E} = \{(i, j) \mid i, j \in \mathcal{V}\}$ where (i, j) represents a link or interaction between the i-th and j-th nodes. Initially, each node $i \in \mathcal{V}$ is associated with a column feature vector $\mathbf{h}_i^{(0)} \in \mathbb{R}^{d^{(0)}}$.

Every layer l of a GNN updates the hidden representation of each node by aggregating information from the neighborhoods:

$$\mathbf{h}_i^{(l+1)} = f_\theta \left(\mathbf{h}_i^{(l)}, \mathrm{F} \left(\{ \mathbf{h}_j^{(l)} \mid j \in \mathcal{N}_i \} \right) \right), \tag{11}$$

where $\mathbf{h}_i^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ are the new node representations, $\mathcal{N}_i$ is the neighborhood of the i-th node, f_θ denotes a nonlinearity, and F is a permutation-invariant aggregator. Several proposals exist for the aggregation operator, determining the expressive power, interpretability, learning stability, and scalability of the network [49].

The non-symmetric effective connectivity maps derived are also non-attributed, that is, there are no node features to be aggregated in Eq. 11. Although non-attributed graphs are classifiable, they dramatically increase the problem's difficulty. Fortunately, the Local Degree Profile (LDP) method effectively decreases the challenge by setting the attributes of each node to local neighboring properties [9]. Thus, we computed the *in* and *out* degree of each node as well as the minimum, maximum, mean, and standard deviation of the *in*/*out* degree of its neighbors. This created a feature vector $\mathbf{h}_i^{(0)}$ of dimension 10 that was propagated through the directed adjacency matrix for every subject. The neural network consisted of $l = 2$ hidden layers and was trained for 150 epochs with a learning rate of 0.005 to minimize the binary cross entropy between the predicted and true classes (Fig. 3 BOTTOM). The metrics were computed with a balanced class weight to account for the different number of samples in each class. The model was tested in a 10-fold cross-validation scheme and used a validation set to test for overfitting.

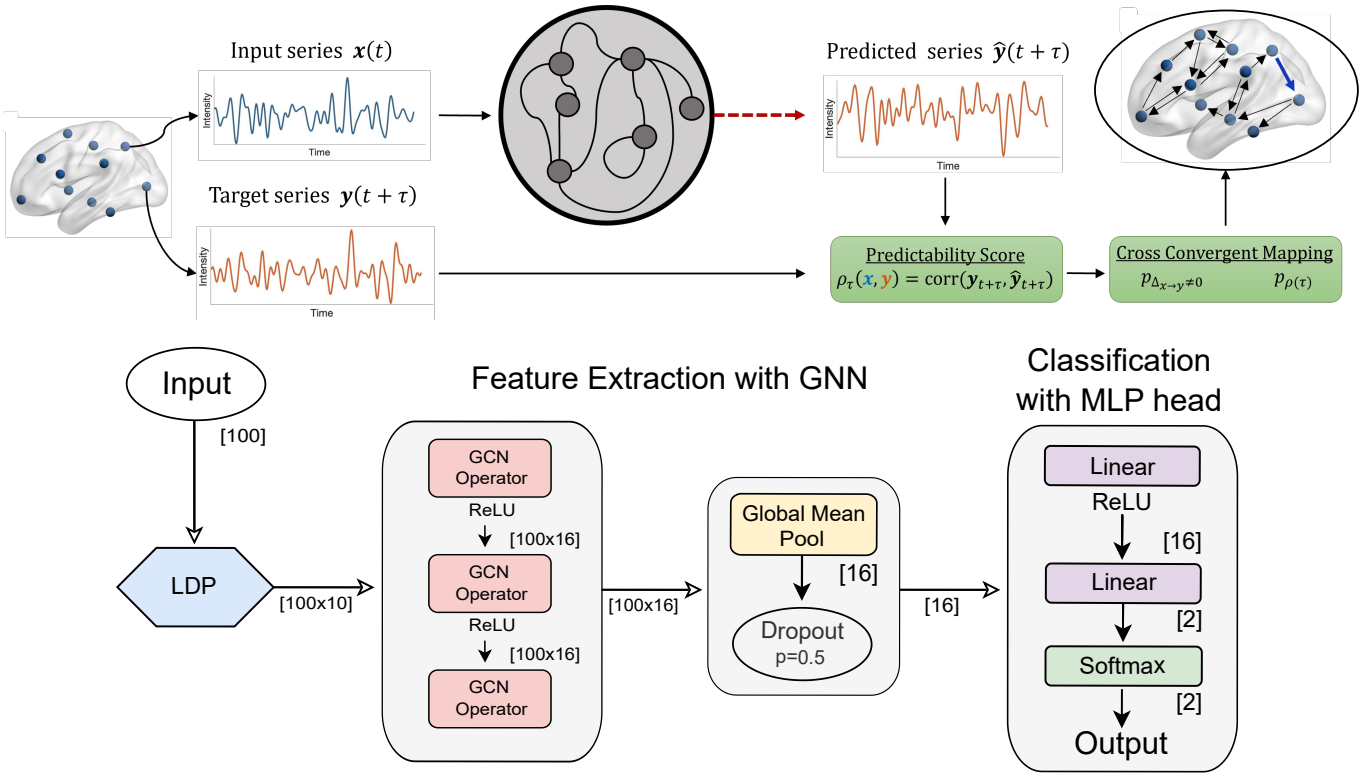

Figure 3. Working diagram of causality given by the reservoir computing (TOP) and graph convolutional architecture (BOTTOM).

2.5 *Local Topology Profile*

A recent extension of the LDP attribution outlined before incorporates other local properties to the already-mentioned descriptors. This Local Topology Profile [1] has been shown to improve the accuracy over its parent version, namely LDP. Following the original proposal, we extended the feature vector $\mathbf{h}_i^{(0)}$ with the edge betweenness centrality [22], the overlap between node neighborhoods (i.e., Jaccard index), and the local degree score [29].

However, as an attempt to further reduce the complexity of the workflow, we used the 13 LTP features with a random forest classifier of max depth 2 and a maximum number of features equal to 5. As in the GCN classifier, we used class weights to balance the dataset and used a 10-fold cross-validation scheme. The architecture used in practice is summarized in Figure 3.

2.6 *Local Interpretable Machine-Agnostic Explanations*

To explain the features allowing the classification we used the LIME (Local Interpretable Model-agnostic Explanations) approach. This technique explains the prediction of any classifier by learning the model locally around the prediction [37]. In our case, this was used to highlight the edges that contributed to the classification performance the most. LIME assigns a coefficient to each edge on the EC matrix based on the contribution to the final classification score.

Positive values were useful in identifying the stroke group, whereas negative values were consistent in identifying the control group. The total explainability values of each ROI were calculated for both groups separately. These values were thresholded with the arbitrary threshold of 0.02 for the stroke group and -0.02 for the control group (because these directions helped the correct decisions). Edges associated with wrong decisions were not studied due to their lack of meaning in neurological terms.

3 Results

3.1 *Effective connectivity maps derived from Reservoir Computing*

EC maps were not readily interpretable given the complex interactions expected to occur at different spatial and temporal scales. Consensus stipulates that information transfer is obscured by the hemodynamic response function, which effectively masks the corresponding temporal delay between cause-consequence associations. We computed effective connectivity maps between 100 ROIs at two different delays (Time of Repetition = 1 and 2; see Fig. 4). The average maps showed clear patterns of hemispheric segregation while at the same time exhibiting strong connectivity between homotopic regions. In canonical functional connectivity studies, this *a priori* segregated structure can be considered as an initial quality assessment of the resulting maps, forming the basis for an accurate description of the functional relationships expected to occur in brain disease.

Even though stroke occurrence is not entirely random [10, 47], their exact morphologies and functional disconnection patterns are highly variable. We further examined the properties of the directed networks by computing the average directed connectivity for controls, subjects suffering from right-hemispheric stroke, and subjects suffering from a stroke located on the left hemisphere (Fig. 5).

Global hemispheric connectivity was computed by averaging the EC maps within and between hemispheres. That is, averaging the val-

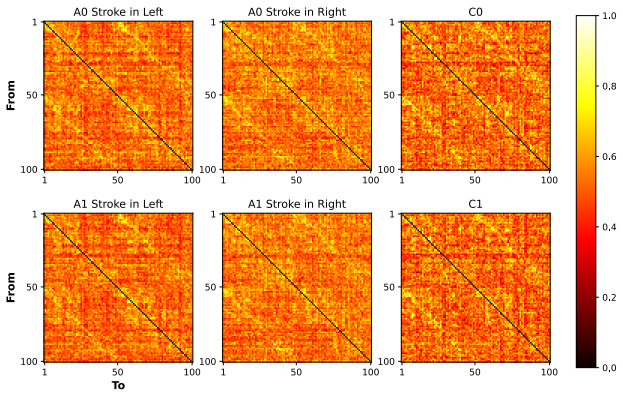

Figure 4. Group averaged effective connectivity matrices for two different Times of Repetition. Top: -1 TR. Bottom: -2 TRs. The left column is the average of subjects suffering from a stroke located in the left hemisphere. The middle column is the average of subjects suffering from a stroke located in the right hemisphere. The right column is the average of the control group.

ues in each on of the 4 visible squares in the average EC maps (Fig. 4). Briefly, intra- and inter-hemispheric connectivity was severely altered in all patients, showing a clear break of symmetric communication w.r.t. the control group, especially for right-impaired subjects [27].

3.2 Classification results

The results of the classification are reported in Tables 4 and 3 respectively for the GCN and LTP classifiers. Results are reported for both the proposed method and Granger Causality: Average AUC, accuracy, precision, recall, and F1 are reported. As expected, the LTP (augmented with a random forest classifier) generally increased the classification metrics, although both models are comparable. It should be noted that classifying effective connectivity graphs is a complicated task due to sample heterogeneity [12, 2], and that very similar scores compared to the chance levels (e.g., an increase of 0.2-0.3) are found in the literature [1].

Table 3. Classification performance of the GCN model. Results are shown by comparing the classification of EC networks derived with the whole-brain RCC method and the GC method.

Metric	whole-brain RCC	Granger Causality
AUC score	0.6866 ± 0.0830	0.6074 ± 0.0588
Accuracy	0.6816 ± 0.0551	0.5386 ± 0.1610
Precision	0.9253 ± 0.0654	0.9178 ± 0.0585
Recall	0.6870 ± 0.0991	0.4968 ± 0.2184
F1 score	0.7808 ± 0.0511	0.6143 ± 0.1922

Table 4. Classification performance of the LTP model. Comparing the classification of EC networks derived with the whole-brain RCC method and the GC method.

Metric	whole-brain RCC	Granger Causality
AUC score	0.6900 ± 0.0652	0.7240 ± 0.1186
Accuracy	0.6972 ± 0.0552	0.7921 ± 0.1377
Precision	0.9228 ± 0.0523	0.9121 ± 0.0451
Recall	0.7041 ± 0.0757	0.8233 ± 0.1493
F1 score	0.7947 ± 0.0443	0.8606 ± 0.1040

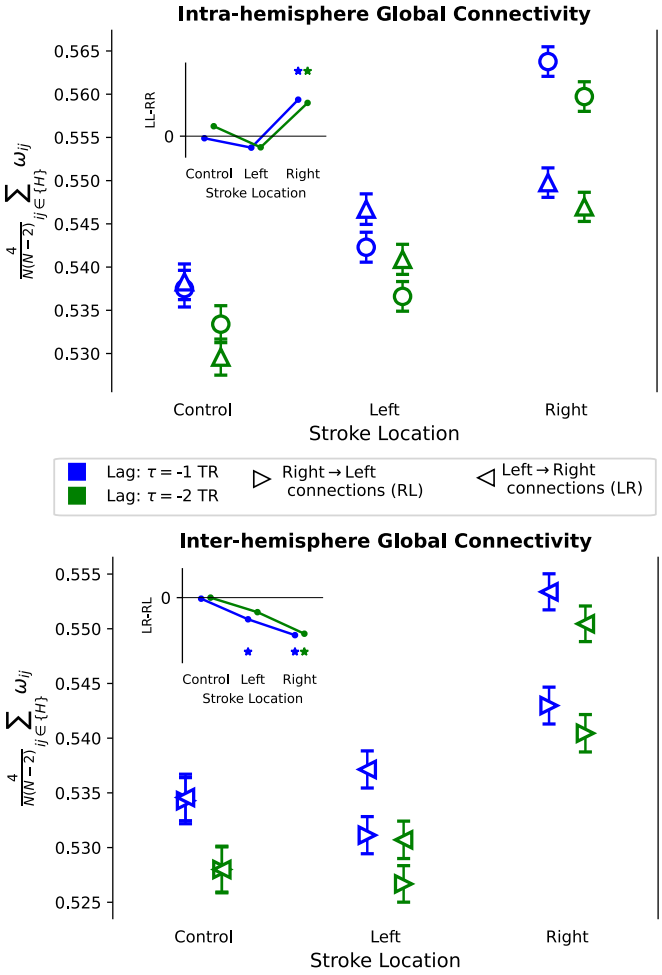

Figure 5. Global effective connectivity alterations between regions located in the same hemisphere (top) and between regions located in different hemispheres (bottom). Error bars depict 1 standard error of the mean. Insets show the average difference between left-left and right-right effective connectivity (top) and between left-right and right-left effective connectivity (bottom). Statistical significance was assessed via a two-sample t-test ('*' p<0.05). Global connectivities were obtained by averaging the weight value over the connections belonging to the corresponding hemispheres H.

3.3 Node and edge importance in stroke detection

We used the LIME explainability framework on the LTP classifier due to its slightly better performance and higher computational efficiency to highlight the most descriptive ROIs and edges related to stroke onset. Importantly, the explanations were done on top of the EC matrices obtained with the reservoir method and not the granger one. For each node in the EC networks, we summed all the explainability coefficients to assess the contribution of each connection arising in each node to the correct classification (i.e., sum over all columns). Lastly, binarized and thresholded explainability values were projected back to the surface mesh (Fig. 6; see also Methods and [27]). The resulting maps show that regions in visual, dorsal, and ventral attention have the most contribution to the classification performance for stroke subjects, while ventral attention and frontoparietal networks contributed the most to the detection of control subjects.

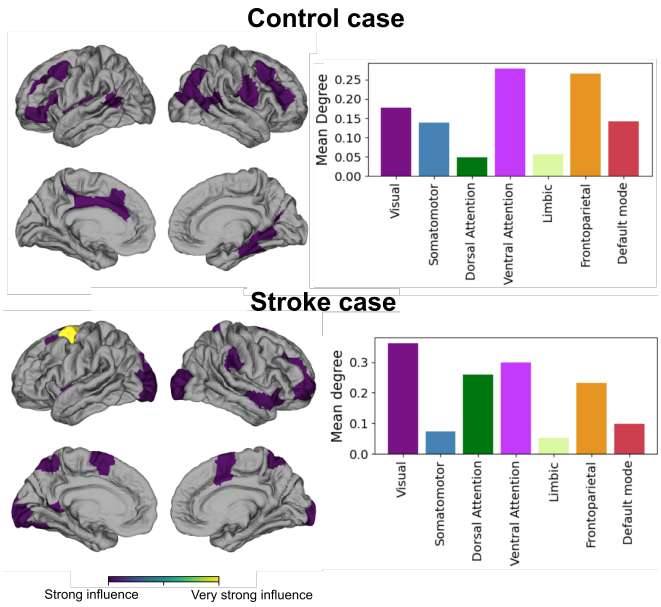

Figure 6. Interpretation of the LIME explainability outputs for each group. Cortical projection of the total contribution of each ROI (left) and its association with one of the 7 resting-state networks. The Dorsal attention network is distinctively necessary to discriminate the presence of a lesion.

4 Discussion

This study addresses the critical need for precise diagnostic tools in stroke management, highlighting the complexity and variability of MRI data and the limitations of conventional machine learning approaches in capturing dynamic network disruptions. The proposed pipeline begins by employing reservoir computing to define effective connectivity of the brain [20]. Effective connectivity using reservoir computing has been recently proposed to unravel more precise interactions in large neural systems [17]. However, studies that thoroughly assess the quality of the resulting causal mappings remain unseen. We propose to evaluate them by first studying existing asymetries in brain information transfer. These maps lead to directed graph representations, which have been loosely explored by graph convolutional network classifiers. Later, we used these directed maps in a AI classification and explainability paradigm; that is, disentangling regions and connections that are important for each control or stroke group.

Functional and effective connectivity asymmetries have been previously characterized in two different formats. Using a Granger-based methodology, Allegra and colleagues [3] described a connectivity imbalance between lesioned and healthy hemispheres. With the maps obtained with the whole brain reservoir computing causality methodology, we observed a similar pattern which was exacerbated in subjects suffering from right-sided lesions (Figs. 4 and 5). Furthermore, upon examining the connectivity between hemispheres, the same type of broken balance was significantly visible as well. Future work could assess how this asymmetry relates to subject behavior. With respect to this, Koba and colleagues [27] explored hemispheric asymmetry in functional connectivity gradients [32] finding a slightly higher correlation between behavior and functional aberrancy in subjects with right-sided lesions. Hence, our findings agree with the fact that the location of the stroke conveys different functional and effective information at a connectomic scale strengthening the need for a more accurate characterization of the expected behavioral dysfunctions and prognosis [18].

Regarding the classification paradigm, graph-structured data is ubiquitous across various disciplines, yet the use of specific graph convolutional neural networks is relatively recent (see [54] for an extensive review). Extensions of methods for directed graph analysis have also been proposed [53], modifying the architecture to perform node classification or link prediction. In this study instead, we focused on overall directed graph classification which was achieved by using conventional graph convolutions with directed adjacency matrices. We are then aggregating these Local Degrees and Topological Profiles based on the message passing across these directed connections.

The pipeline achieves promising results, yielding an area under the curve of 0.69, superior to the state-of-art method (GC) using the GCN classification model. This should be considered a promising result given the highly heterogeneous dataset (stroke lesions were present in different parts of the brain), where similar scores relative to chance levels are often observed [2]. Furthermore, it was also possible to employ explainable AI tools to interpret disrupted networks despite these diversified lesions across brain networks. This elucidates the contribution of effective connectivity biomarkers that can capture aspects at a general level despite those individual differences, offering insights into disease mechanisms and treatment responses.

Previous studies on structural connectome of stroke patients highlighted network dysfunctions [41]. Stroke-related modulations in inter- and intra-hemispheric coupling were recently investigated highlighting asymmetry and inter-areal interactions after stroke, related to broad changes in inter-areal communication and resulting in several deficits [3]. Moreover, Erdogan and colleagues argued that the global fMRI signal is affected by the stroke lesion generating a delay of the blood-oxygen-level-dependent (BOLD) signal depending on the lesions [14]. Our results were in line with those previous analyses. We found inter-hemispheric connectivity was severely altered in all patients, showing a clear break of symmetric communication w.r.t. the control group. The differences were particularly pronounced in the case of stroke lesions in the right hemisphere. This can hypothesized as the integrity of the within-hemispheric networks is sustained through language-related connections, as the right hemisphere is less involved in speech generation and suffers more from the injury. [19]. Indeed, the explainability maps of the control subjects resemble the vision and language networks. It is possible that the algorithm abused the connections from/to the language network to detect control subjects. Aphasia is a common symptom in the case of ischemic stroke, therefore the connections of the language network in the stroke group may show different characteristics. A similar hypothesis can be suitable also for stroke subjects because the supplementary motor area, which plays an important role in language processing, was also useful for accurate classification. Importantly, alterations in the ventral and dorsal attention networks are often present in stroke [10, 11, 2, 27], which are in line with our explainable maps in Fig. 6. Nevertheless, these claims should be confirmed with larger datasets.

Undoubtedly, there are several ways to discriminate control subjects from stroke patients which are less computationally demanding [42], and previous studies also showed a correlation between functional and effective connectivity with the first being easier to compute than the latter [3]. Here, we emphasized the use of a classification task for two reasons 1) to further assess the effective connectivity maps and 2) to provide a strong basis for which to implement explainability pipelines. With this we also propose an approach to classify directed graphs. However, we showed the need to use fur-

ther mapping into anatomical atlases to allow acceptable explainability. Although, in conclusion, this proposes an end-to-end pipeline for studying effective connectivity brain disorders, capitalizing on a specific approach for directed graph and explainability.

This analytical framework enhances clinical interpretability but also can inspire confidence in decision-making processes, crucial for translating research findings into clinical practice as it can translate complex neuroimaging features into simple visualizations. The study lays the groundwork for improved patient stratification in other brain diseases as well, with the ultimate goal of assisting doctors, demonstrating also the potential of reservoir computing causality, graph convolutional networks, and explainable analysis.

Acknowledgements

Authors thank Prof. Maurizio Corbetta for sharing the dataset used in this study. The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533. This publication is supported by Sano project carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund. This research was supported in part by the PLGrid infrastructure. Computations have been partially performed on the ARES supercomputer at ACC Cyfronet AGH.

References

[1] J. Adamczyk and W. Czech. Strengthening structural baselines for graph classification using local topological profile. In *International Conference on Computational Science*, pages 597–611. Springer, 2023.

[2] M. H. Adhikari, J. Griffis, J. S. Siegel, M. Thiebaut de Schotten, G. Deco, A. Instabato, M. Gilson, and M. Corbetta. Effective connectivity extracts clinically relevant prognostic information from resting state activity in stroke. *Brain communications*, 3(4):fcab233, 2021.

[3] M. Allegra, C. Favaretto, N. Metcalf, M. Corbetta, and A. Brovelli. Stroke-related alterations in inter-areal communication. *NeuroImage: Clinical*, 32:102812, 2021.

[4] S. Avvaru and K. K. Parhi. Effective brain connectivity extraction by frequency-domain convergent cross-mapping (FDCCM) and its application in Parkinson's disease classification. *IEEE Transactions on Biomedical Engineering*, 2023.

[5] E. Bates, S. M. Wilson, A. P. Saygin, F. Dick, M. I. Sereno, R. T. Knight, and N. F. Dronkers. Voxel-based lesion–symptom mapping. *Nature neuroscience*, 6(5):448–450, 2003.

[6] S. Bouazizi and H. Ltifi. Novel diversified echo state network for improved accuracy and explainability of EEG-based stroke prediction. *Information Systems*, 120:102317, 2024.

[7] L. Breston, E. J. Leonardis, L. K. Quinn, M. Tolston, J. Wiles, and A. A. Chiba. Convergent cross sorting for estimating dynamic coupling. *Scientific reports*, 11(1):1–10, 2021.

[8] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. doi: 10.1109/MSP.2017.2693418.

[9] C. Cai and Y. Wang. A simple yet effective baseline for non-attributed graph classification. *arXiv preprint arXiv:1811.03508*, 2018.

[10] M. Corbetta, L. Ramsey, A. Callejas, A. Baldassarre, C. D. Hacker, J. S. Siegel, S. V. Astafiev, J. Rengachary, K. Zinn, C. E. Lang, et al. Common behavioral clusters and subcortical anatomy in stroke. *Neuron*, 85 (5):927–941, 2015.

[11] M. Corbetta, J. S. Siegel, and G. L. Shulman. On the low dimensionality of behavioral deficits and alterations of brain network connectivity after focal injury. *Cortex*, 107:229–237, 2018.

[12] A. Crimi, L. Dodero, F. Sambataro, V. Murino, and D. Sona. Structurally constrained effective brain connectivity. *NeuroImage*, 239: 118288, 2021.

[13] M. Cucchi, S. Abreu, G. Ciccone, D. Brunner, and H. Kleemann. Hands-on reservoir computing: a tutorial for practical implementation. *Neuromorphic Computing and Engineering*, 2(3):032002, 2022.

[14] S. B. Erdoğan, Y. Tong, L. M. Hocke, K. P. Lindsey, and B. deB Frederick. Correcting for blood arrival time in global mean regression enhances functional connectivity analysis of resting state fMRI-BOLD signals. *Frontiers in human neuroscience*, 10:311, 2016.

[15] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods*, 16(1):111–116, 2019.

[16] A. Etkin. Addressing the causality gap in human psychiatric neuroscience. *JAMA psychiatry*, 75(1):3–4, 2018.

[17] J. Falco-Roget, A. I. Onicas, F. Akwasi-Sarpong, and A. Crimi. Directed networks and resting-state effective brain connectivity with state-space reconstruction using reservoir computing causality. *bioRxiv*, pages 2023–06, 2023.

[18] M. D. Fox. Mapping symptoms to brain networks with the human connectome. *New England Journal of Medicine*, 379(23):2237–2245, 2018.

[19] A. D. Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 2011.

[20] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[21] M. Gilson, R. Moreno-Bote, A. Ponce-Alvarez, P. Ritter, and G. Deco. Estimation of directed effective connectivity from fMRI functional connectivity hints at asymmetries of cortical connectome. *PLoS computational biology*, 12(3):e1004762, 2016.

[22] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99 (12):7821–7826, 2002.

[23] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[24] G. Grassmann. New considerations on the validity of the Wiener-Granger causality test. *Heliyon*, 6(10):e05208, 2020.

[25] Y. Huang, Z. Fu, and C. L. Franzke. Detecting causality from time series in a machine learning framework. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(6), 2020.

[26] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34):13, 2001.

[27] C. Koba, J. Falco-Roget, and A. Crimi. Reshaped functional connectivity gradients in acute ischemic stroke. *bioRxiv*, pages 2024–04, 2024.

[28] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan. A fast and accurate online sequential learning algorithm for feedforward networks. *IEEE Transactions on neural networks*, 17(6):1411–1423, 2006.

[29] G. Lindner, C. L. Staudt, M. Hamann, H. Meyerhenke, and D. Wagner. Structure-preserving sparsification of social networks. In *Proceedings of the 2015 IEEE/ACM International conference on advances in social networks analysis and mining 2015*, pages 448–454, 2015.

[30] J. Lucio, R. Valdés, and L. Rodríguez. Improvements to surrogate data methods for nonstationary time series. *Physical Review E*, 85(5): 056202, 2012.

[31] W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560, 2002.

[32] D. S. Margulies, S. S. Ghosh, A. Goulas, M. Falkiewicz, J. M. Huntenburg, G. Langs, G. Bezgin, S. B. Eickhoff, F. X. Castellanos, M. Petrides, et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 113(44):12574–12579, 2016.

[33] M. Maziarz. A review of the granger-causality fallacy. *The journal of philosophical economics: Reflections on economic and social issues*, 8 (2):86–105, 2015.

[34] J. M. McCracken and R. S. Weigel. Convergent cross-mapping and pairwise asymmetric inference. *Phys. Rev. E*, 90:062903, Dec 2014.

[35] N. Parga, L. Serrano-Fernández, and J. Falcó-Roget. Emergent computations in trained artificial neural networks and real brains. *Journal of Instrumentation*, 18(02), 2023.

[36] A. T. Reid, D. B. Headley, R. D. Mill, R. Sanchez-Romero, L. Q. Uddin, D. Marinazzo, D. J. Lurie, P. A. Valdés-Sosa, S. J. Hanson, B. B. Biswal, et al. Advancing functional connectivity research from association to causation. *Nature neuroscience*, 22(11):1751–1760, 2019.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[38] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral cortex*, 28(9):3095–3114, 2018.

[39] T. Schreiber and A. Schmitz. Improved surrogate data for nonlinearity tests. *Physical review letters*, 77(4):635, 1996.

[40] S. Shahi, F. H. Fenton, and E. M. Cherry. Prediction of chaotic time series using recurrent neural networks and reservoir computing techniques: A comparative study. *Machine learning with applications*, 8: 100300, 2022.

[41] J. S. Siegel, A. Z. Snyder, L. Ramsey, G. L. Shulman, and M. Corbetta. The effects of hemodynamic lag on functional connectivity and behavior after stroke. *Journal of Cerebral Blood Flow & Metabolism*, 36(12): 2162–2176, 2016.

[42] A. G. Smith and C. Rowland Hill. Imaging assessment of acute ischaemic stroke: a review of radiological methods. *The British journal of radiology*, 91(1083):20170573, 2017.

[43] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modelling methods for fMRI. *Neuroimage*, 54(2):875–891, 2011.

[44] O. Sporns. Structure and function of complex brain networks. *Dialogues in clinical neuroscience*, 15(3):247–262, 2013.

[45] P. Steiner, A. Jalalvand, S. Stone, and P. Birkholz. PyRCN: A toolbox for exploration and application of reservoir computing networks. *Engineering Applications of Artificial Intelligence*, 113:104964, 2022.

[46] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting causality in complex ecosystems. *Science*, 338 (6106):496–500, 2012.

[47] M. Thiebaut de Schotten, C. Foulon, and P. Nachev. Brain disconnections link structural connectivity with function and behaviour. *Nature communications*, 11(1):5094, 2020.

[48] A. A. Tsonis, E. R. Deyle, H. Ye, and G. Sugihara. Convergent cross mapping: theory and an example. *Advances in nonlinear geosciences*, pages 587–600, 2018.

[49] P. Veličković. Everything is connected: Graph neural networks. *Current Opinion in Structural Biology*, 79:102538, 2023.

[50] R. Vicente, M. Wibral, M. Lindner, and G. Pipa. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67, 2011.

[51] H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports*, 5(1):14750, 2015.

[52] I. B. Yildiz, H. Jaeger, and S. J. Kiebel. Re-visiting the echo state property. *Neural networks*, 35:1–9, 2012.

[53] X. Zhang, Y. He, N. Brugnone, M. Perlmutter, and M. Hirn. Magnet: A neural network for directed graphs. *Advances in neural information processing systems*, 34:27003–27015, 2021.

[54] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.