Towards Understanding Human Functional Brain Development with Explainable Artificial Intelligence: Challenges and Perspectives

Mehrin Kiani[1]*, Javier Andreu-Perez[1] Hani Hagras[1], Silvia Rigato[1], and Maria Laura Filippetti[1]

Abstract—**The last decades have seen significant advancements in non-invasive neuroimaging technologies that have been increasingly adopted to examine human brain development. However, these improvements have not necessarily been followed by more sophisticated data analysis measures that are able to explain the mechanisms underlying functional brain development. For example, the shift from univariate (single area in the brain) to multivariate (multiple areas in brain) analysis paradigms is of significance as it allows investigations into the interactions between different brain regions. However, despite the potential of multivariate analysis to shed light on the interactions between developing brain regions, artificial intelligence (AI) techniques applied render the analysis non-explainable. The purpose of this paper is to understand the extent to which current state-of-the-art AI techniques can inform functional brain development. In addition, a review of which AI techniques are more likely to explain their learning based on the processes of brain development as defined by developmental cognitive neuroscience (DCN) frameworks is also undertaken. This work also proposes that eXplainable AI (XAI) may provide viable methods to investigate functional brain development as hypothesised by DCN frameworks.**

Index Terms—**Explainable Artificial Intelligence, Developmental Cognitive Neuroscience, xMVPA, fNIRS, EEG**

I. Introduction

Human brain development is a complex and dynamic process that begins prenatally and extends through to late adolescence [1]. The human brain has an estimated 100 billion neurons at birth [2] whose interconnections form neural networks, which become specialised over time and mediate the functional capabilities of the human brain [3]. This specialisation results not only from the structural development of the brain but also as a consequence of optimisation of inter-regional interactions in the developing brain [3]. Over the past 50 years, the field of developmental cognitive neuroscience (DCN) has examined the relations between the structural and functional development of the human brain [4], elucidating the developmental mechanisms underlying cognitive processes such as perception, attention, memory, and language.

DCN research can inform us about the influence of genetic variations and environmental factors in the specialisation of neural networks [3]. In addition, DCN studies can extend insights into how these specialised networks mediate newly acquired social and cognitive functions, shedding light on typical and atypical trajectories of human brain development [5]. A greater understanding of brain development trajectories can have profound implications for early detection and the subsequent intervention of developmental disorders [4]. Furthermore, a better understanding of the interplay between structural and functional brain development can be leveraged to inform clinical, educational and social policies [6].

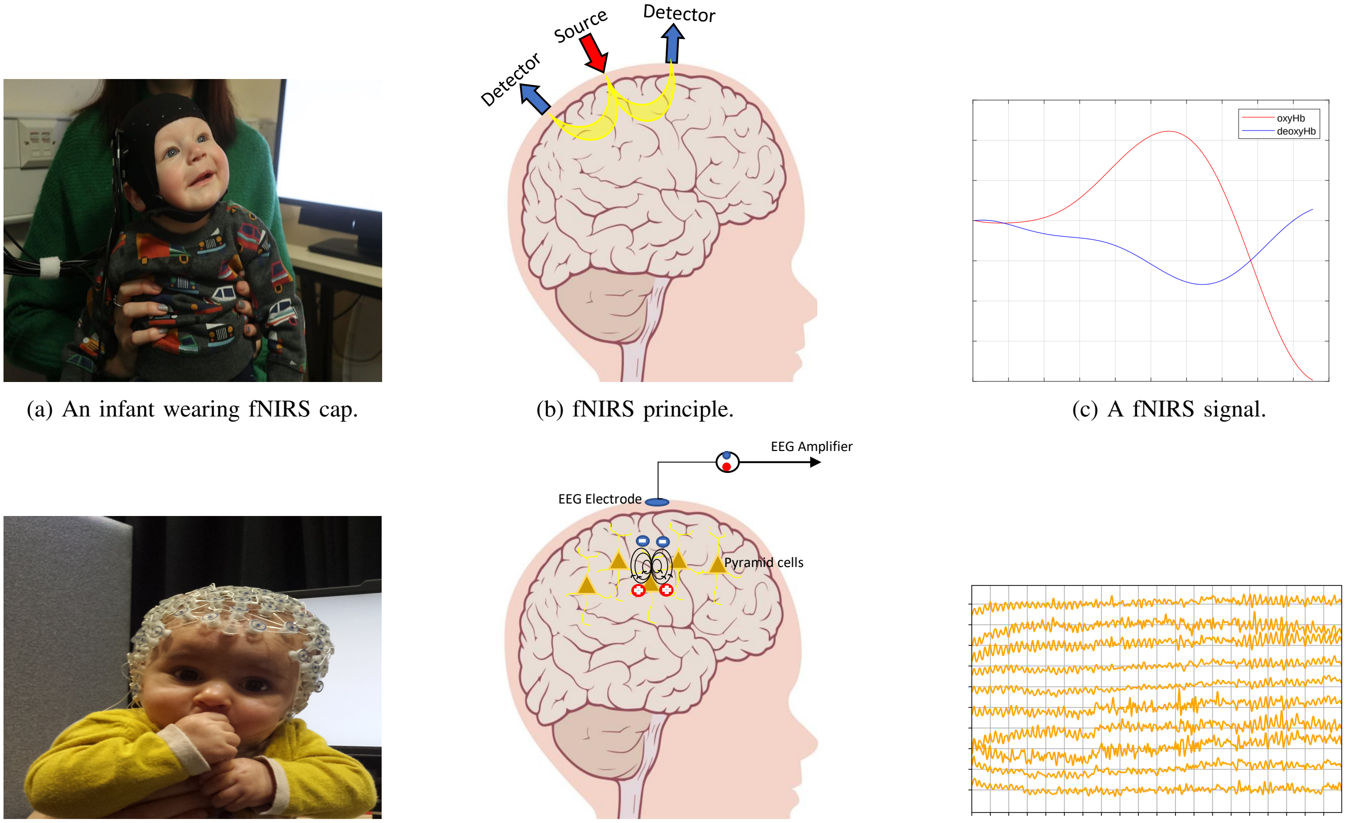
In order to examine the neural underpinnings of cognitive processes and their changes across development, functional Near-Infrared Spectroscopy (fNIRS) [7] [8], and Electroencephalogram (EEG) [9] have been widely used in DCN studies with infants and children. These neuroimaging modalities are both non-invasive, portable, wearable, and relatively inexpensive compared to functional magnetic resonance imaging (fMRI), which has instead proved pivotal in adult brain neuroimaging. In particular, fNIRS and EEG allow for the young participants to stay engaged in tasks whilst recording their brain activity in more naturalistic postures (e.g., sitting upright vs laying down) and ecologically valid settings such as their homes [8]. Nevertheless, fMRI has been successfully used in developmental studies with asleep infants [10, 11], and more recently with awake infants [12, 13]. As fNIRS and EEG are considered the most commonly used and 'infant-friendly' modalities to investigate neural substrates in DCN studies, the present review paper will focus on these two modalities and their respective data analysis paradigms.

fNIRS is an optical neuroimaging modality that uses Near-Infrared (NIR) light on the scalp to record changes in blood haemoglobin that occur as a result of cerebral activity. More specifically, fNIRS measures the relative changes in haemoglobin (Hb) concentration in the blood, based on NIR light absorption by the Hb molecules, which is inferred as a measure of the cortical brain activity [7]. The fNIRS cap, comprising of pairs of sources and detectors, can be flexibly adapted based on the brain areas of interest (see for an example Fig. 1a). The strength of fNIRS lies in its good spatial localisation (within 2cm) that allows for conclusions to be drawn about the localised cortical activity from different anatomical locations of the cortical structures, as recorded by the fNIRS channels located on the participant's head (Fig. 1a). An illustration of the fNIRS principle (Fig. 1b) along with a representative signal (oxy-Hb in red, and deoxy-Hb in blue is shown in Fig. 1c) is shown in Fig. 1.

While fNIRS relies on changes in blood oxygenation to measure brain activity, EEG measures electrophysiological brain activation. More specifically, EEG records electrical changes on the scalp, allowing the measurement of rapid

* Corresponding author: mehrin.kiani@essex.ac.uk

[1] University of Essex, United Kingdom



(a) An infant wearing fNIRS cap.

(b) fNIRS principle.

(c) A fNIRS signal.

(d) An infant wearing EEG net.

(e) EEG principle.

(f) 10 EEG signals.

Fig. 1: (a) An infant wearing an fNIRS cap. The placement of the fNIRS channels on the fNIRS cap is dependent on the areas of interest for investigating the underlying brain activity. (b) An illustration of the fNIRS principle: fNIRS is an optical neuroimaging modality that reads underlying cerebral activity using fNIRS channels formed by a pair of sources and detectors. Based on the area under investigation, a fNIRS source shines Near InfraRed (NIR) light at the point on the surface of the head, and the diffusely refracted NIR light is recorded by a fNIRS detector. The relative changes in haemoglobin (Hb) concentration in the blood measured by fNIRS channels is inferred as cortical brain activity [7]. The delay in the fNIRS haemodynamic response measurement varies significantly depending on the age of the participants and the specific cognitive/motor task undertaken [14]. (c) A representative fNIRS signal. The y-axis has Δ concentration Hb values recorded at time (x-axis) post stimulus presentation. The red signal is Δ concentration in oxy-haemoglobin values whereas the blue signal is Δ concentration in deoxy-haemoglobin values. (d) An infant wearing an EEG net. The EEG electrode placement has been standardised using an international 10–20 system that uses anatomical landmarks on the skull [15]. (e) EEG measures brain electrical activity, with the electrodes placed on the scalp, which reflect the summated postsynaptic potentials of cortical neurons in response to changing cognitive or perceptual states [16]. EEG activity is mainly generated by pyramidal neurons in the cerebral cortex that are perpendicular to the brain's surface/electrode on the scalp [17]. (f) 10 illustrative EEG signals with measured voltages on the y-axis and time on the x-axis. EEG signals have temporal resolution in milliseconds, providing a near real-time display of ongoing cerebral activity, but with limited spatial resolution due to effects of electric field spread [18].

cognitive processes [19] with high temporal accuracy in the order of milliseconds [20]. The EEG principle is represented in Fig. 1e. The EEG net has electrodes fitted in it using a standardised 10/20 electrode placement system that covers the whole head (see Fig. 1d). Since EEG can record activity on the time scale of underlying neuronal activity, EEG signals (representative EEG signal shown in Fig. 1f) are best suited for connectivity analysis. However, EEG is also more sensitive to motion artifacts [21] and due to its limited spatial resolution [18], EEG makes it difficult to map brain electric activity read by the electrodes to their corresponding anatomical regions in the brain.

One of the most recent advancements in neuroscience research is the combined use of neuroimaging techniques (e.g., [22]). In particular, in DCN, multimodal imaging can provide a wider picture of functional brain activity by benefiting from the advantages of different neural measures (e.g., EEG-fNIRS [23]). Given the complementary characteristics of EEG and fNIRS, i.e., EEG records highly accurate temporal information whilst fNIRS is more spatially localised, multimodal fNIRS-

EEG studies enable greater information to be recorded regarding the underlying brain activity. While this represents a step further to better study the developing brain, it does not prove sufficient to translate DCN research to inform typical and atypical trajectories of functional brain development. More specifically, if functional brain trajectories using infant's neuroimaging data can be established with the help of explainable Artificial Intelligence (XAI) methods, it may assist in early identification of, and thus intervention in, neurodevelopmental disorders [24], as well as shape policies in the context of typical neurocognitive development.

Indeed, the fundamental question in DCN of how cognitive development is mediated by structural maturation (i.e., the emergence of faculties through growth processes) and optimised interactions remains open. In this regards an understanding of the theoretical frameworks that can explain the bidirectional relation between the structural and functional development of the human brain is critical. Therefore, in Section II, we firstly summarise the key concepts of the DCN frameworks including the 'Interactive Specialisation' (IS) theory [3, 25, 26] and the neuroconstructivist approach [27, 28]. A review of the current artificial intelligence (AI) algorithms, as applied to fNIRS and EEG data both in infancy and adulthood is undertaken in Section III. The aim of the review is to investigate the extent to which these AI methods can explain human functional brain development in light of the theoretical frameworks of DCN. Implications of explainable AI methods, that also mimics the mechanisms proposed by DCN frameworks, and the conclusion are presented in Section IV and V respectively.

II. Developmental Cognitive Neuroscience (DCN) Frameworks

The developed adult human brain, both in terms of structure and function, is a 'small world' network [29]. A small world network is typically characterised with concentrated local activity, decreased short-range interconnections (segregation), and increased long-range connections (integration) rendering it cost efficient. Repeated processing of certain types of input leads to certain brain networks becoming increasingly proficient and fine-tuned to process that specific information [28]. In particular, developmental change in the varying levels of activity across different cortical regions leads to gradual specialisation and localisation observed in the developed human brain [28], as illustrated in Fig. 2.

A developed brain is also modular with respect to functional organisation, i.e., it has a hierarchical network that has the ability to feed processed information from one layer (module) to another. The hypothesis of a more modular developed brain is based on the evidence of top-down and bottom-up information flow. For example, during visual processing, the information in the adult brain flows from the primary area of visual processing (such as occipital cortex) to higher hierarchical levels (such as pre-frontal cortex (PFC)) where the information processed by lower hierarchical levels is integrated [30, 31].

The three main DCN frameworks, namely 1) Maturational perspective, 2) Interactive Specialisation (IS), and 3) Skill learning, aim to answer the question of how these optimised, hierarchical networks emerge during postnatal development. For the purpose of this work, we will focus on the IS perspective, which is largely supported by DCN studies [32]. The IS framework proposes that both feed-forward and feedback connections between different cortical regions affect the functional specialisation of cortical regions [33]. More specifically, the IS theory provides a description of the following three major processes that occur in the developing brain:

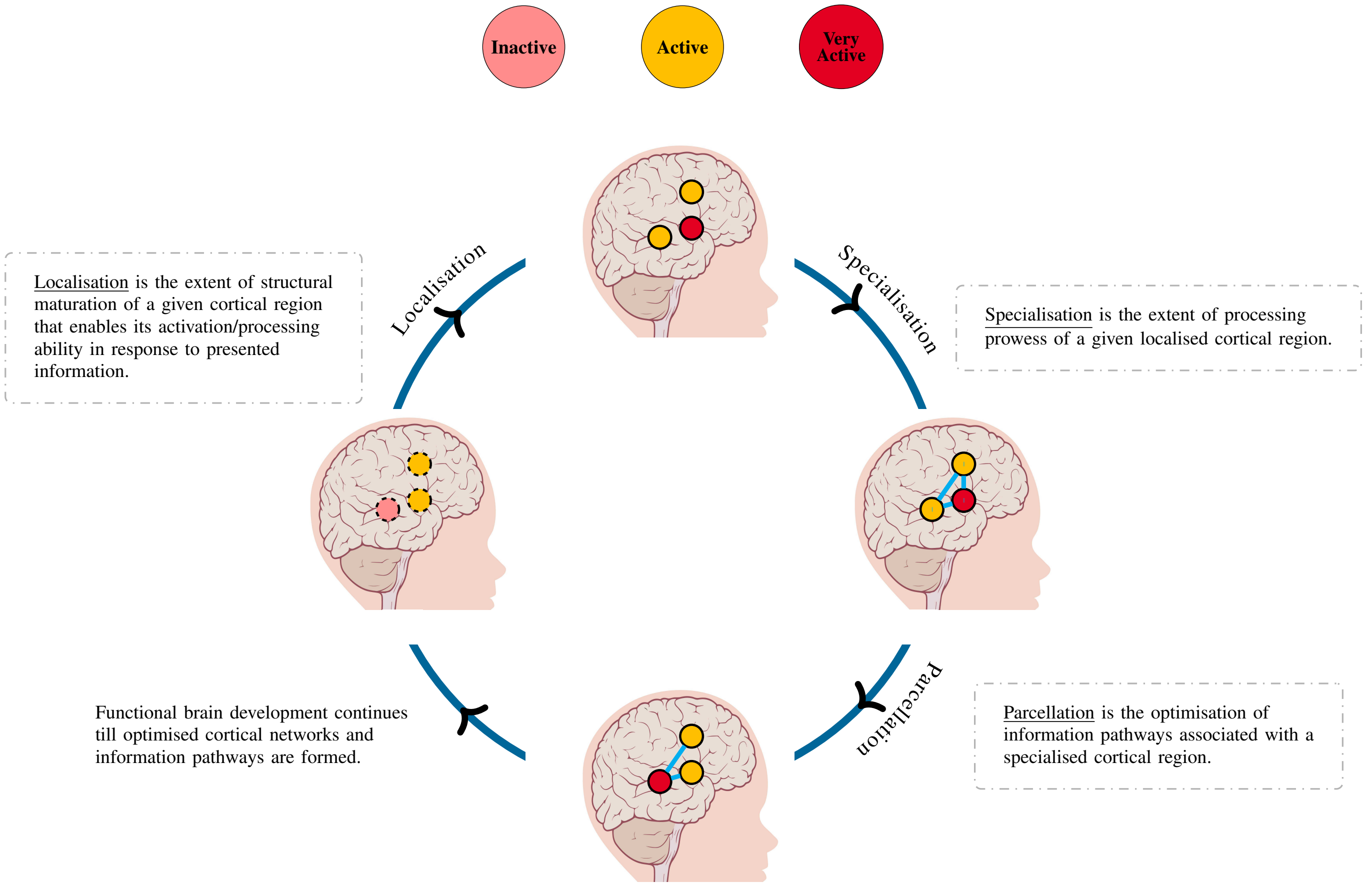
(i) *Localisation*: The extent of cortex activation for a given task.
(ii) *Specialisation*: The extent of functionality achieved by a given cortical area.
(iii) *Parcellation*: The optimisation of synaptic connections of neural circuits.

The IS framework suggests that functional brain development is a dynamic process with localisation, specialisation, and parcellation processes forming a continuous loop of development as shown in Fig. 2. As a given cortical area gains more structural maturation, its specialisation for a given task increases, which then triggers the parcellation (optimisation) of information flow in the cortical network formed to subserve that given task.

Optimisation can take place because of structural and/or functional maturation (i.e., the emergence of capabilities through growth processes) of different parts of the brain, along with more long range connections coming 'on line'. As a result of the parcellation process, not all parts of a given cortical region need to be activated nor are all connections required to transmit the information to the next level of processing. In this sense, parcellation takes place both within and between cortical regions. The increased segregation of information pathways gives rise to increased specialisation (i.e., a modular structure), thus leading to the gradual emergence of hierarchical networks.

An important consideration with regard to the hierarchical brain is that the interactions between hierarchies at multiple levels and timescales are not hard-wired, i.e., the coordination between modules is not fixed [34]. As a consequence, existing modules could subserve emerging cognitive states through a reconfiguration of the evolved circuits using *neural reuse* [35] process of brain organisation. The other two plausible processes put forward to explain functional brain organisation are *modularity* and *holism* [35]. The modular functional brain structure would imply that for each task there would be largely segregated cortical circuits with limited overlap, whereas, the holism organisation of the brain suggests that all cortical circuits may be engaged across all tasks. The idea of neural reuse seems plausible with respect to optimal usage of existent circuits evolved for a given cognitive task. In this way, while neural circuits are modular to some extent with respect to their individual functionality, neural reuse suggests that they (individual modules) have the capacity to connect with each other in numerous configurations to achieve a range of cognitive-behavioural tasks. The three aforementioned perspectives of functional structure of the brain are illustrated in Fig. 3.

Taken together, the IS framework and neural reuse perspective can shed light on functional brain development at a



Localisation is the extent of structural maturation of a given cortical region that enables its activation/processing ability in response to presented information.

Specialisation is the extent of processing prowess of a given localised cortical region.

Parcellation is the optimisation of information pathways associated with a specialised cortical region.

Functional brain development continues till optimised cortical networks and information pathways are formed.

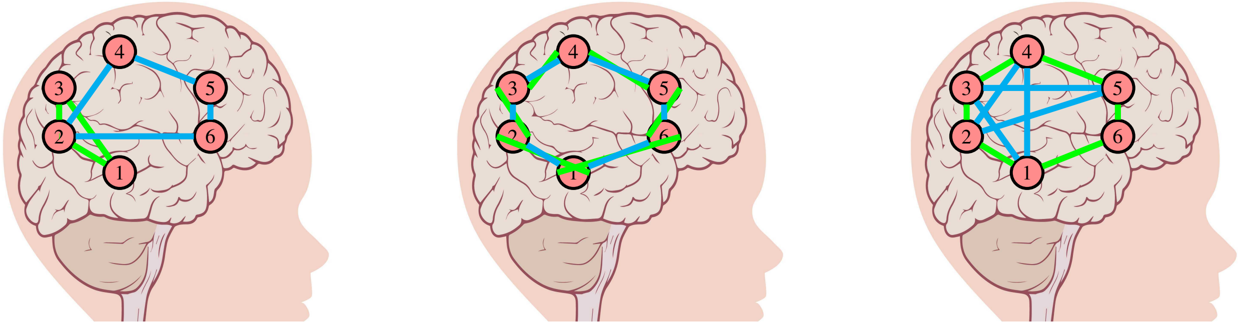
Fig. 2: The Interactive Specialisation (IS) theory focuses on explaining how different anatomical regions in a developing brain learn to cooperate to form an optimised cortical network using *localisation*, *specialisation*, and *parcellation* processes. The IS theory hypothesises that functional brain development is not a stationary, one-way process, but instead, all of its components are in a loop, giving feedback to each other as structural maturation, as well as external stimuli, pave the way for an increasingly specialised cortical network. An illustration of hypothetical developing cortical activation and interactions are shown in cyan with the circles representing different cortical regions, with varying levels of activity as denoted by their colour. Red: Very active; Amber: Active; and Pink: Inactive. The different levels of activity of a cortical region represent the amount of neurons firing and forming synaptic contacts with other neurons to process presented information.

given time point using cross-sectional DCN studies. However, a major component of functional brain development that is still to be accounted for is the associated temporal information, i.e., at what time the developmental changes are happening [34]. Clearly, all stages of functional brain development are not the same with respect to time. In this regard, investigating the temporal dimension associated with a developing brain analysis of longitudinal DCN data, i.e., neuroimaging data recorded over a certain time period, is considered imperative.

Any AI method used to shed light on functional brain development must keep in mind the aforementioned phenomena and challenges associated with DCN. To this end, we will be reviewing the extent AI methods can explain their underlying mechanisms to shed light on the processes of functional brain development.

III. Artificial Intelligence (AI) Methods in Cognitive Neuroscience

The generic field of cognitive neuroscience investigates the underlying brain functional mechanism that subserve cognitive processes such as memory, perception, understanding, and reasoning [36]. DCN is a sub-field of cognitive neuroscience that focuses on developmental population (infants and younger children) to investigate how functional brain developmental processes shape the developing brain. In principle, the AI techniques that have been applied to study the cognition states of non-developmental population (such as adults) can also be used to study the developmental population (infants and younger children). This is because all the pre-processing stages of acquired neuroimaging data (fNIRS or EEG) would be similar as well as the AI techniques that can discern the difference in brain activation patterns for adults should also be able to decode the same in infants neuroimaging data analysis as well. As opposed to cognitive neuroscience for



(a) Modular

(b) Holism

(c) Neural Reuse

Fig. 3: Three different perspectives for the functional structure of the brain are illustrated with cortical modules (denoted by circles numbered 1 to 6) and interconnections between the cortical modules (shown in green for task Y and cyan for task Z). (a) The modular structure suggests that different modules will be activated for different tasks with limited overlap of modules. This is shown by largely segregated modules activated for task Y (i.e., the modules 1, 2, and 3 activated for task Y) and task Z (i.e., the modules 2, 4, 5, and 6 activated for task Z). As such the modules engaged are largely distinct with limited overlap (of module 2 in this hypothetical case) for the two separate tasks Y and Z i.e., brain functional organisation is local. (b) In holistic organisation, all local brain regions are presumed to be involved for all tasks. Hence, all modules are activated for both task Y and task Z. (c) According to the neural reuse perspective, the brain works as a whole but functional differences can be appreciated owing to their reconfigured interactions between the same modules (use and reuse).

adult populations, due to the lack of prior assumptions or cannon models, the application of XAI in DCN helps to bring new light into a science that otherwise, with classical non-explainable or purely statistical models, would be challenging to elucidate.

AI techniques [37] have been both inspired by, and used for the study of the learning processes in the human brain. A major component of functional brain development is attributed to unsupervised learning [38] owing to the massive amounts of unlabeled sensory data infants receive, although supervised and reinforcement learning faculties are also hypothesised to account for some facets of human brain development [39]. There is also considerable debate about how much of the functional brain development is a result of postnatal learning, and to what extent is the genome (an organism's complete set of hereditary material) responsible for shaping brain development [38].

Considering the aforementioned learning mechanisms in AI methods, and how they can potentially shed light on the human functional brain development, in the following subsections we review the most commonly used AI methods as applied to fNIRS and EEG neuroimaging data. Most studies have not necessarily used these algorithms for the analysis of infants' neuroimaging data, however, their application to infants data would be similar in principle. The overarching aim of the review is to investigate the potential and limitations of these algorithms, as applied to infants neuroimaging data analysis, to explain their learnt inference mechanism in terms of developmental brain processes of localisation, specialisation, parcellation, and neural reuse as outlined by the DCN frameworks.

For this reason, here we review AI methods with application(s) to fNIRS and EEG data, as well as some recent promising works for their applicability to DCN research and data analysis. Please note this is not meant to be an exhaustive review of all the AI methods used in cognitive neuroscience studies, nor is it designed to be used as a reference for implementing the reviewed AI methods. The aim of this review is to understand the underlying inference mechanism of the explored AI methods, and what their understating can inform us about the underlying developing brain processes.

In this regard, depending on how much can be inferred (or explained by the learnt inference mechanism(s)), the AI methods can be generally categorised as explainable (inferred output can be interpreted with linguistic concepts and propositions), partially explainable (inferred output can shed light on the feature importance or rank) or non-explainable (no insight can be obtained from the inferred output). However, with respect to their implications in DCN research, there is not much of a distinction between non-explainable and partially explainable methods; hence, we will cover the partially explainable methods within the non-explainable methods as well.

The most commonly involved AI methods explainable or not for the analysis of neuroimaging data can also be broadly classified into the following three analysis paradigms: 1) *Connectivity Analysis (CA)*; 2) *Representation Learning (RepL)*; 3) *Multivariate Pattern Analysis (MVPA)*. We briefly summarise these analysis paradigms before reviewing the explainable and non-explainable AI methods as applied to adults' and infants' neuroimaging data.

1) Connectivity Analysis:

Brain connectivity analysis can shed light on the segregation and integration of the isolated cortical networks formed to mediate coherent cognitive and behavioral states.

The three modes of brain connectivity analysis [40] that can inform us about the organisation and the workings of the developing human brain are: 1) structural connectivity (SC) 2) functional connectivity (FC) and 3) effective connectivity (EC) analysis. SC is generally associated with respect to the anatomical wiring in the brain and is typically measured in vivo using diffusion weighted imaging. FC is measured as the temporal correlation between spatially remote neurophysiological events [41]. In contrast, EC measures the influence that one neural system exerts over another which can be both activity, and/or time dependent [42].

In most cognitive studies, to understand the underlying connectivity of cortical regions for processing presented information, the analysis of FC (to investigate which spatially distinct cortical areas of the brain are engaged simultaneously) and/or EC (to investigate the extent of influence one cortical region exerts on another) is undertaken. Indeed, the analysis of FC and EC can potentially inform about brain architecture; however, to what extent the connectivity analysis effectively contributes to the understanding of brain processes is dependent on the choice of the AI technique (used for the connectivity analysis).

2) Representation Learning:

Many recent works in neuroscience are increasingly using deep leaning paradigms to investigate the underlying brain activity in response to a presented task [43]. Amongst Deep Neural Networks (DNNs), convolutional neural networks (CNNs) have gained particular interest because of their remarkable performance in unsupervised automatic feature extraction and classification of objects in challenging image classification problems [43]. Owing to the capability of CNNs to compose higher level features using lower level features, CNNs can learn representations of input data automatically overcoming the long standing challenge to handcraft a feature set in conventional AI methods [43]. In CNNs, a small matrix of numbers (called a filter) is passed over (convoluted with) the raw data, to extract features from the raw data, such as edges in images, also called a feature map. The convolution layer is followed by a pooling layer which downsamples the input to reduce both the spatial size of the input data and the number of hyperparameters in the network. A typical CNN architecture consists of the following stages:

(i) Feature Learning Blocks
- Convolution (C) + Rectified Linear Unit (ReLU).
- Pooling (P).

(ii) Classification/Regression Blocks
- Fully Connected Layers.
- Softmax, Logistic regression layer, regression loss (Root Mean Square Error (RMSE) etc.)

The performance of CNNs is critically dependent on the optimisation of hyperparameters, and owing to the large number of hyperparameters that need optimisation, most DNNs, including CNNs, require large datasets to converge. The hyperparameters of a CNN include the size of the filter(s), stride, number of hidden layers, and the learning rate.

3) Multivariate Pattern Analysis:

In most multivariate analysis, the feature set is crafted by hand i.e., the statistic characteristic (such as the mean or amplitude) of a neuroimaging signal which would best capture the neural underpinnings, corresponding to the task at hand, is chosen manually. The two dimensional matrix formed by collating together the features from $\mathbf{N}$ channels (for fNIRS) or electrodes (for EEG) and $\mathbf{J}$ number of data trials is then given as input to an AI method, and is hereby referred to as a multivariate matrix (MVM).

Although it requires considerable subject-matter expertise to curate a feature set for MVM that best represents the underlying neural activity, the classification results based on the analysis of MVM would reflect on the representational dynamics of the underlying cortical networks (as read from fNIRS channels or EEG electrodes). In this regard the classification results obtained from the analysis of MVM can be at least partially attributed to the cortical networks activation as represented by the statistical feature used for constructing the MVM.

The MVM can be readily analysed using any state-of-the-art AI methods. Most AI methods such as Support Vector Machine (SVM) and Random Forest (RF) usually give very robust classification results with MVM. This analysis approach is termed multivariate pattern analysis (MVPA) [44] and was first used for neuroimaging analysis on adult multi-voxel fMRI data [45].

In the following subsections, we review the explainable and non-explainable AI methods used on the aforementioned analysis paradigms on both non-developmental (adults) and developmental (infants) population.

A. AI in Cognitive Neuroscience for Adult Brains

In Cognitive Neuroscience, AI methods are frequently used with adult populations (mature brains). Some approaches can provide no explanation or simply partial information, and others can derive some explainable structure.

1) Non-Explainable AI Methods

A review of the non-explainable AI methods for investigating cognitive processes in adults' cognitive neuroscience studies is presented next.

a) FC with EEG using SVM

The EEG studies by Moezzi *et al.* [46] and Klados *et al.* [47] used SVM with radial basis function (RBF) as the underlying kernel to investigate EC. In particular, the work by Moezzi *et al.* is of interest with respect to DCN research as it investigated the difference in FC to recognise young (mean age 24 years) from old adult brains (mean age 71 years). The FC was studied in the standard frequency bands of delta (1–4Hz), theta (4–8Hz), alpha (8–13Hz), beta (13–30Hz) and gamma (30–45Hz). The aim was to study oscillations in standards frequency bands to uncover coordinated activity in large-scale brain networks which facilitate information flow between spatially distributed brain regions. The calculation of FC matrices was done using imaginary coherence in an attempt to account for the poor spatial localisation of the EEG signals.

Cross-validation was performed to optimise the hyperparameters (`C`: regularisation factor, and gamma kernel coefficient) of SVM, improve accuracy, and identify the most significant features. To map the FC to brain regions, a grouping approach was used to spatially localise the observed connectivity patterns. In addition, consensus features were obtained using Euclidian distance between electrode pairs to investigate FC patterns based on age. They concluded that consensus features belonging to delta, theta, alpha and gamma frequency bands had positive weights showing significantly higher FC in younger adults than in older adults. Features of the beta band had negative weights showing significantly higher functional connectivity in older adults than younger adults. However, as is also acknowledged in the original study [46], the limitation to map the consensus features to anatomical regions of the brain could not facilitate further discussion on FC pattern differences with respect to brain regions. Hence, despite the prowess of SVM to differentiate between the FC patterns of old and young brains with 93% classification accuracy, the SVM's inference mechanism could not shed light on the temporal correlations of the different cortical regions.

The non-explainability of SVM inference mechanism is because the learnt support vectors, which form the inference mechanism of SVM for distinguishing between data instances belonging to distinct classes, are defining a hyperplane that optimally separates the data instances in a high dimensional space. Hence, the support vectors as such can not be expressed in terms of the underlying brain activity patterns. The only relevant information that can be obtained from the support-vector of a linear SVM is a feature relevance/significance score but that too can not shed light on the association between the inputs to uncover the cortical networks formed.

b) FC with fNIRS using Ridge Regression (RR)

The connectivity analysis with fNIRS does not require additional spatial localisation of the measured cortical activity owing to the relatively good spatial resolution that can be achieved with fNIRS instruments [7]. Two complementary, non-explainable AI methods, namely ridge regression (RR) and interpolated functional manifold (IMF), used with fNIRS connectivity measures are reviewed next.

A fNIRS study investigating intrinsic FC of cortical networks to predict anxiety states using linear ridge regression (RR) models is done by Duan *et al.* [48]. The resting state FC was calculated using Pearson correlation coefficient for 1035 edges between 46 nodes (fNIRS channels). The RR was able to predict the anxiety score with statistical significance using the connectivity of cortical networks. The mean square error (MSE) of their model was 122.04 with correlation coefficient of `r = 0.36`.

The prowess to predict states of anxiety using FC has profound implications for the diagnosis of anxiety and related disorders. However, the ability for the regression model to explain its 1035 optimal values of β (also called regressors) in terms of FC is significantly limited. Therefore, despite getting statistically significant results, the FC analysis could not shed light on the resting state cortical networks.

c) FC with fNIRS using Interpolated Functional Manifold (IFM)

A recent study that puts forth a solution for group-wise explorative analysis using manifolds is presented by Avila-Sansores *et al.* [49]. In this work, fNIRS values are projected to an ambient space. Since there can be infinite surfaces that can cross the projected fNIRS values, the aforementioned study proposes Interpolated Functional Manifold (IFM) to select a surface. In particular, an explicit model for the surface is chosen by interpolating between the projected fNIRS values using RBF.

The proposed IFM method is used on subjects with varying levels of surgical expertise (knot-tying). The fNIRS values are projected onto a two-dimensional manifold and the distribution of the fNIRS values is based on pairwise distances i.e., points that are close together in the manifold have similar characteristics. For this particular study, the medical students' fNIRS responses got projected to the edges of the manifold, whereas more experienced participants' (trainees and consultants) fNIRS responses accumulated in the conceptual centre of the manifold. The graphs were validated against mixed effect models (with regressors encoding group variances) and psychophysiological interaction (PPI). Since IFM analysis may contain infinite graphs, they visualised the FC with IFM graphs by thresholding them to obtain maximum similarity of Jaccard Index(JI). The maximum JI values reported with group level analysis are 0.89 ± 0.01 and those with PPI are 0.83 ± 0.07.

The advantage of IFM approach is that an explicit analytical expression is obtained that can be used to quantitatively study the group based differences, as in the case of participants with varying level of expertise for certain motor skill. In addition, the IFM approach can facilitate fNIRS data analysis in hyper-scanning studies, i.e. blue, reading neuroimaging data from more than one person at a given time. However, it is a complimentary analysis for measuring FC since the graph of FC measures is selected by thresholding it against established group level models to obtain maximum values of JI.

d) RepL with EEG using EEGNet

In this section, we review the works that learn representations of input data with multiple levels of abstraction, using CNNs for brain-computer interface (BCI) applications. The aim of BCI is to translate brain signals into control signals for a computer (or device) to perform the desired action [50]. The advancements in BCI have enabled people with neuromuscular disorders to restore or replace some of their motor functions such as limb movements [51]. For a successful BCI, a user typically has to undergo training for generating brain signals that can encode their intention for communicating with the connected device. Likewise, an AI technique powering BCI also needs to be trained to decode the intention based brain signals, from the user, to command signals for successful control of the device.

The relevance of BCI for DCN studies come from gaining insights into the neural reuse of already evolved cortical circuits for performing a given function such as limb movement. Hence this re-learning of a user to control their limb via BCI instead of normal output pathways of peripheral nerves and muscles would be a key mechanism for successful BCI. In this regard, the decoding of the composition of the 'control' signal, based on lower level features using multiple processing layers of CNN, can have profound implications for shedding light on the consequences of neural commitment (perceptual narrowing) for defining neural reuse. Consequently, the CNNs powering BCI can shed light on the neural reuse and perceptual narrowing to perform BCI.

In the following subsections, we review the most promising studies for both EEG-based [52] and fNIRS-based [53] BCI applications.

In general, EEG-based BCI paradigms can be categorised as 1) event-related potential (ERP) and 2) oscillation based BCI paradigms. The classic ERP based brain-computer interface (BCIs) aim to recognise a relatively high amplitude characterised with low frequency in the EEG signal evoked in response to, and time-locked with, an external event/stimulus. In contrast, the oscillation based BCI paradigms make use of the signal power pertaining to specific frequency bands for classification. A general-purpose architecture for CNN developed for the classification of EEG-based BCI paradigms, called EEGNet, is proposed by Lawhern *et al.* [52]. The strength of the EEGNet lies in its successful classification for both even-related and oscillatory BCIs, as validated in the study over 4 different BCI paradigms: 3 ERP-based BCI and 1 oscillatory-based BCI.

The proposed architecture of EEGNet is illustrated in Fig. 4. In reference to Fig. 4, EEGNet undertakes the following convolutions to learn respective lower level features from the input EEG data:

(i) *C1*: Temporal convolution to learn frequency filters.
(ii) *C2*: Depth-wise convolution to learn frequency-specific spatial filters (i.e., a specific spatial filter for each frequency filter).
(iii) *C3*: Separable convolution to optimally aggregate the features maps together.

Since EEGNet is tested on 4 different EEG datasets, Ch is used to denote the number of electrodes and T represents the time samples for a particular dataset. The number of hyperparameters, per BCI paradigm for a total of 4 paradigms investigated (P1-P4), to be learnt by EEGNet with 4 temporal filters and 2 spatial filters per temporal filter, are: [P1: 1066, P2: 1082, P3: 1098, and P4: 796]. In general, the performance of EEGNet was superior for ERP-based BCI paradigms in comparison to oscillatory-based BCI with an average of ~ 80% classification accuracy across the 4 paradigms.

In an attempt to validate that their proposed EEGNet model learning is based on relevant features depicting brain activity, the authors investigated three different approaches for enabling feature explainability:

(i) *Hidden unit activations*: This was done after depth wise convolution: i.e., C2 in Fig. 4 sheds light on the spatial localisation of the activations corresponding to a particular frequency.
(ii) *Filter weights:* The visualisation of filter weights was possible because of EEGNets architecture that limits the connectivity between two convolution layers: i.e., direct visualisation of the narrow band filter frequency filters weight for C1, and the frequency-specific spatial filter weights for C2 in Fig. 4 sheds light on the relevant frequency components, and frequency specific spatial localisation.
(iii) *Feature relevance:* The relevance of individual features for classification performance of EEGNet was calculated on a per trial basis using DeepLIFT algorithm.

The validity of the features, on whose basis is the inference mechanism learnt, is of significance to establish the robustness of the CNN architecture. However, the almost ~ 950 learnt hyperparameters are not explainable since a given optimised value of a hyperparameter can not be matched to a particular representation of brain activity. In essence, the optimal values of the hyperparameters of EEGNet are a filter that can not shed light on the interconnections or EC of cortical regions.

e) RepL with fNIRS using CNN

The classic analysis paradigms for fNIRS signals are based on the statistical features most representative of the underlying activity. For representation learning on fNIRS signals using CNNs, the fNIRS signals are first transformed to equivalent image time-frequency representations known as spectrograms.

In the work by Janani *et al.* [53] the authors investigated the possibility of classification of four different motor imagery tasks, i.e., participants *imagined* moving their limbs instead of physically moving their limbs, using CNNs. More specifically, the four different motor imagery tasks were: right- and left-fist clenching, right- and left-foot tapping. The fNIRS channels were placed on top of the left and right hemispheres to record brain activity from respective cortical regions.

The spectrogram method was used to transform the fNIRS signal into a time-frequency image. The architecture of the CNN feature extraction stage had two convolution (C) layers with 1 pooling (P) layer in between the following sizes- C1: 3×23×16; P: 2×12×16 and C2: 3×3×32. The fully connected layer, had 288 nodes which connected through hidden layers, classified the input fNIRS image into 4 motor imagery tasks.

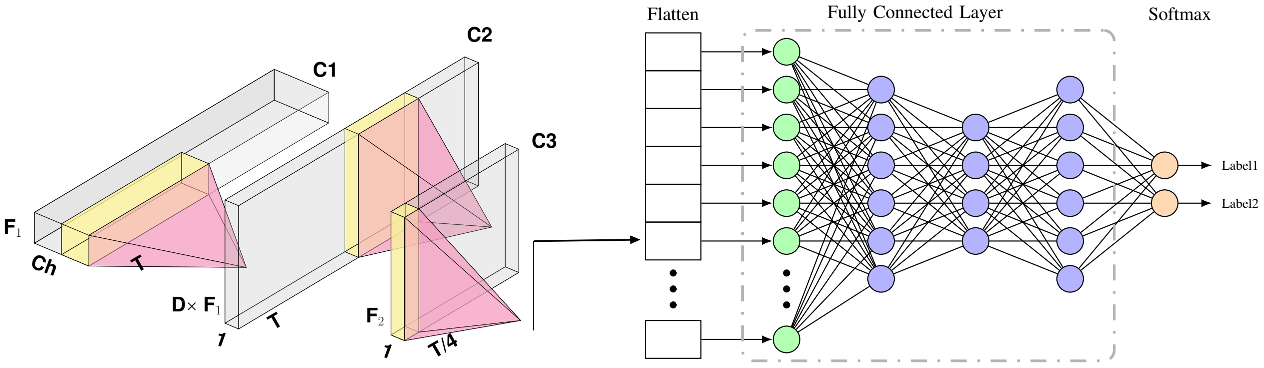
The average classification accuracy obtained over all four tasks was 72.35%. Although the CNN preformed the best amongst other standard AI methods (SVM and multi-layer perceptron), the classification accuracy was not at par with the usual high performing CNNs. The modest performance of CNN could be attributed to the large input fNIRS image dimensions (660×22).

f) MVPA with EEG using SVM

The MVM for EEG signals can be built using ERPs, wavelet coefficients or using component analysis. For ERPs, taking the average is beneficial for reducing some noise though single ERPs are more representative. In addition, if the MVPA investigation with respect to the cortical regions of the brain is critical, then source localisation of the electrodes is important.

A toolbox that has been designed in particular to make the MVPA more accessible is the Amsterdam Decoding and Modeling Toolbox (ADAM) [54]. It takes as input EEG data in standard formats of FileldTrip or EEGLAB and is able to pre-process i.e., increase signal to noise ratio, remove motion artefacts. The first level MVPA can compute a performance metric, whereas group level MVPA can compute statistical significance for patterns.

A successful application of MVPA with EEG data for the detection of a face in the wild (natural settings) is done by Cauchoix *et al.* [55]. The stimulus images were grayscale photographs of human faces presented in their natural contexts.



(a) Feature Extraction

(b) Classification Mechanism

Fig. 4: A schematic of the EEGNet [52] architecture that gave the best classification results over 4 different brain computer interface (BCI) EEG based datasets. (a) The first stage consists of a series of convolutional (C) and pooling (P) layers of varying dimensions where F_1 and F_2 denote respective filter sizes, Ch is the number of electrode channels and T stands for the number of time points. D represents the number of spatial filters. F_1 is a temporal filter, whereas F_2 is a pointwise filter. A feature map is constructed from the input matrix by convoluting it with a filter (or kernel). By using different filters on the same input, different features from the input image can be detected such as an edge. To account for nonlinearities, the feature map is then passed through an activation function. The pooling layer is used to merge semantically similar features found from convolution layers into one. The most common pooling function used is the *max* pooling but in this study [52] average pooling is used. (b) The feature set thus computed, through convolution and pooling layers, is then flattened and input to a fully connected layer (a neural network). The softmax layer, which is also an activation function, is the last layer that finally predicts the label for the class.

After pre-processing the EEG signals, ERPs were computed separately for correct human face target trials and correct animal face non-target trials. For face processing in adult EEG data, MVPA has been successfully used for face detection in wild settings for ERPs: P100 (positive potential observed after around 100msec of stimulus presentation) at four bilateral occipital electrodes (O1, O2, PO3, and PO4) and for the N170, (negative potential observed after around 170msec of stimulus presentation) at four right hemisphere occipitotemporal electrodes (PO10, PO8, P8, and TP8).

Their MVPA results with SVM achieved a classification accuracy of 94.8%. In addition, based on their results, they suggest that neural dynamics of face detection could be readout very early, starting ~95 ms following stimulus onset.

Although the above reviewed works involving SVM, i.e. Moezzi *et al.* [46] (section III-A1(a)) and Cauchoix *et al.* [55] (section III-A1(f)) are non-explainable methods, other works involving SVMs in particular have investigated gaining an insight into the inference mechanism by using logic programming [56], and decision trees [57]. Given the aforementioned limitations of non-explainable AI methods reviewed to inform the underlying brain mechanisms, we review the XAI methods in the next section.

2) XAI Methods

In this section, we review the XAI methods, i.e., AI methods whose inference mechanism can be explained in terms of the brain activity patterns. In addition, the insights obtained owing to the explainability of the applied XAI method for the given task are also investigated.

a) EC with fNIRS using Effective Fuzzy Cognitive Maps (EFCMs)

An fNIRS study that estimated EC amongst fNIRS channels (corresponding anatomical regions in the cortex) based on fuzzy cognitive maps (FCMs) is proposed by Kiani *et al.* [58]. A FCM is a cognitive mapping technique based on graph theory, with a formal mathematical definition given as follows in (1).

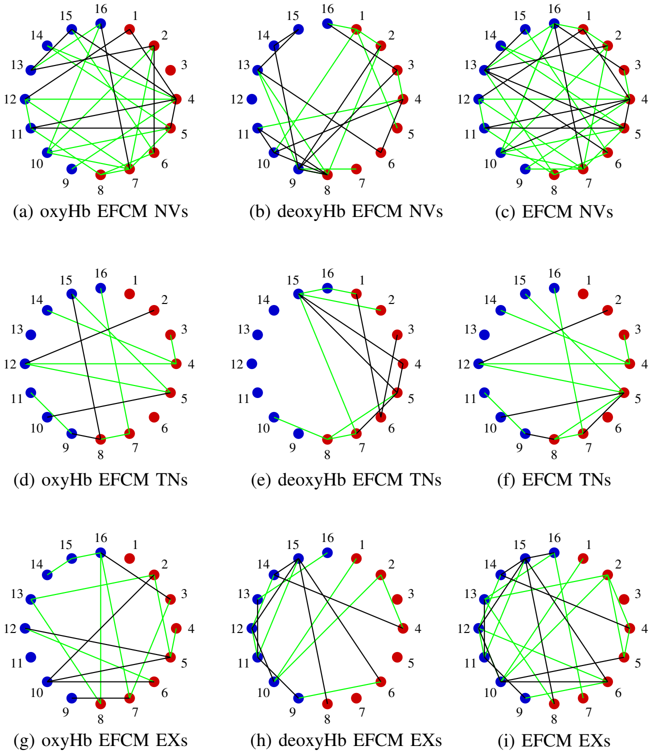
$$C_{j(t+1)} = f(\sum_{i=1}^{N} e_{ij} C_i(t)) \tag{1}$$

where N is the number of concepts (or fNIRS channels) in a given system, $C_j(t)$ is the value of a given concept C_j at iteration t, e_{ij} are the fuzzy weights or EC that concept C_i exerts on concept C_j and f is typically a sigmoid function that scales the weights to [-1,1] for comparative analysis such that a value of 1 means fully interconnected, a value of -1 means fully interconnected in the opposite direction, a value of 0 means disconnected, and a value between 0 and 1 (or -1) means interconnected to a certain extent. The optimal values of EC (e_{ij}) are typically found using an evolutionary algorithm such as Genetic algorithm (GA) (GA has also been used in the aforementioned study [58]).

The error between the estimated signal and real signal using the learnt EC weights by EFCM is computed using eq (2)

$$error = \sum_{t}^{T} \sum_{i}^{N} |C_i(t) - \hat{C}_i(t)| \tag{2}$$

The proposed FCM in Kiani *et al.* [58] is an enhanced FCM, called effective FCM (EFCM), that optimises the strength (scalar magnitude without direction) and direction separately,



(a) oxyHb EFCM NVs (b) deoxyHb EFCM NVs (c) EFCM NVs

(d) oxyHb EFCM TNs (e) deoxyHb EFCM TNs (f) EFCM TNs

(g) oxyHb EFCM EXs (h) deoxyHb EFCM EXs (i) EFCM EXs

Fig. 5: (a) The Effective Connectivity (EC) networks, as delineated by the work of Kiani *et al.* [58] for oxyHb and deoxyHb fNIRS signals recorded from prefrontal cortex (PFC), denoted in red, and motor cortex (MC), denoted in blue, of surgeons with varying levels of expertise in performing a complex visual-spatial task (more specifically laparoscopic surgery(LS)). The expertise level of the participating subjects was categorised into three levels of Novices (NVs), Trainees (TNs), and Experts (EXs). The aim of the study was to discern the difference in EC networks formed with varying levels of proficiency for carrying out a task that requires active planning and visual-motor coordination. A green line signifies the presence of a positive (reinforcement) EC between the connecting cortical regions and the presence of a black line denotes a negative (weaken) EC between the connecting cortical regions.

rendering the EFCM with more degrees of freedom to find the optimum values of EC i.e., e_{ij} in (1). In addition, they also propose tuning of the transformation (sigmoid) function, f in (1), to optimise how fast the non-normalised fuzzy degrees of relationship are squeezed into the normalised range for the fuzzy degrees of relationship.

They applied their proposed EFCM on a neuroergonomics study in which brain activity of subjects, with three varying levels of expertise in performing a surgical task, was recorded using fNIRS channel placed on the prefrontal cortex (PFC) and motor cortex (MC). The EC networks found using EFCM are shown in Fig. 5. They reported an error of 120.7, as defined in (2), averaged for all three levels of expertise in regressing the EC.

The EFCMs propose a partial explainable model in terms of estimating the EC as fuzzy weights (i.e., e_{ij}) between its concepts (fNIRS channels) which can be readily mapped to anatomical locations in the brain. In the original study of EFCMs [58] EC was estimated separately for subjects with varying levels of expertise, as shown in Fig. 5. Hence the derived EC could shed light on how the cortical networks differ in their influence on each other to subserve the complex visual-motor task on skills acquisition. In this regard EFCMs, when applied to DCN studies for estimating EC, can shed light on how developing cortical networks change in terms of their influence (EC) on account of specialisation and neural reuse processes performing a certain task.

In addition to estimating EC with statistical significance, EFCM work [58] also demonstrated the prowess to analyse the difference between estimating EC from oxyHb and deoxyHb dimensions of fNIRS signals for representing the EC in the cortical networks. Although it remains to be established which dimension of fNIRS is more representative for a certain task or specialisation level, they proposed that EC estimated using deoxyHb is more representative of the underlying EC as an individual gains experience in a certain motor task.

b) RepL with EEG using independent component analysis and Fuzzy Neural Networks (ICA-FNNs)

In the work by Lin *et al.* [59], the cognitive state of individuals while driving in a virtual-reality based driving environment, is measured using an EEG-based XAI method. In particular, their adaptive method for recognition of drowsiness of an individual is based on a combination of independent component analysis (ICA) of the EEG signals, and fuzzy neural networks (FNNs) called ICA-FNN.

The significance of trying to decode the cognition state of alertness of an individual, based on the correlation between the information obtained from their brain signals, i.e., power spectra of ICA components of EEG signals, and the individual's driving performance, i.e., the difference between the centre of the vehicle and the cruising lane, is critical in alerting the driver before a potential car accident happens. This is of relevance to shed light on the brain development processes, as ICA-FNN can potentially be applied on infants' brain data to decode their cognitive states as defined by the (un)successful execution of the task at hand.

The ICA-FNN architecture is defined over five layers, as shown in Fig. 6. Taken together, the fuzzy inference system of ICA-FNN takes the following form as shown in (3):

$$\begin{aligned} &\text{Rule : IF } antecedents \ \text{ THEN } consequent(s) \\ &\text{Rule } i \text{ : IF } x_1 \text{ is } A_1^i \ ...and\ x_j \text{ is } A_j^i ...and\ x_n \text{ is } A_n^i \\ &\qquad \text{THEN } y_i \text{ is } m_{0i} \ + \ a_{1i}x_1 + \ ... + \ a_{ji}x_j \ + ... \ + \ a_{ni}x_n \end{aligned} \tag{3}$$

where i is the rule number and $[x_1, ..., x_j, ..., x_n]$ are inputs, with conceptual labels defining the inputs as $[A_1^i, ..., A_j^i, ..., A_n^i]$ correspondingly becoming the antecedent part of the rule i. The centre of a symmetric function is m_{0i}, y_i is the consequent set, a_{ji} is a consequent parameter for the jth antecedent of the ith rule.

The antecedent part of the rules are latent variables obtained from ICA analysis of the EEG signals. Although the inference

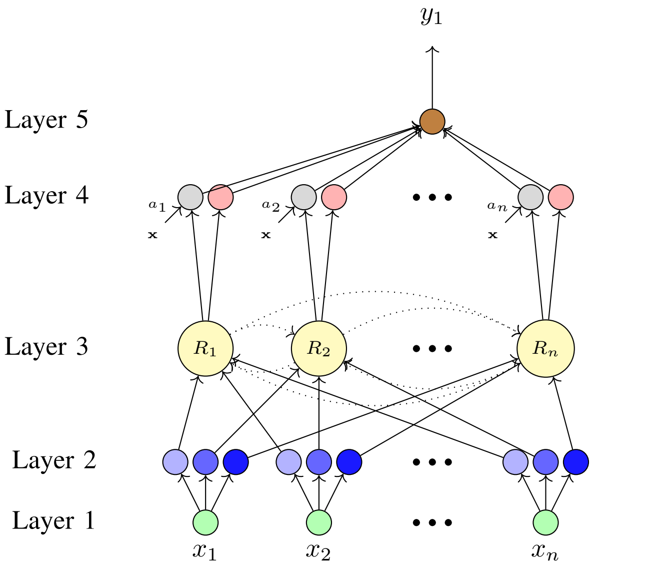


Fig. 6: The architecture of ICA-FNN by Lin *et al.* [59] consists of five layers. The second layer transforms the inputs x_1 to x_n, obtained from the input layer 1, to latent variables computed using ICA on the input data. Each node in layer 3 is a rule (R_1 to R_n), and calculates the antecedent match by computing the firing strength of each rule. Layer 4 is the consequent layer with defuzzification performed in layer 5 where a crisp output is produced.

mechanism is expressible in terms of the rules acquired, see eq. (3), the inputs are abstract features derived from EEG signals. In this regard, the rules can not shed light on the underlying cortical networks formed owing to the non-explainability of the abstract antecedents of the rules.

In addition to the limited explainability of the rules obtained, the rules are also different across the subjects. This is because of the adaptive feature selection mechanism based on ICA; hence, the rules obtained can not be generalised across the subjects. Moreover, each subject would also need to undergo the training phase separately to tune the hyperparameters using the backpropagation algorithm, to learn the cognitive states of each individual. In this way, the inter-subject variabilities of cognition states are largely accounted for and yield improved performance for the recognition of drowsiness/alertness. They reported a remarkable average accuracy of 98.2 ± 1.0 % over five subjects.

The rules are also learnt on-line, i.e., during the training phase for each subject. By learning the rules on-line the inter-subject variabilities are accounted for. This also helps for better recognition of drowsiness since ICA-FNN would remember the inference mechanism is learnt for each subject to better decode that particular subject's level of drowsiness. Also, for a particular subject 2, they reported the acquired correlation between the abstract features learnt and the state of drowsiness of subject 2 as 0.93 and 0.88. Their investigations also concluded that drowsiness related regions are generally found to be in parietal and occipital lobes.

For the application of ICA-FNN in DCN studies, the on-line learning of the hyperparameters specific to each subject would need to be modified, based on the task at hand, since infants will not be able to provide feedback about their cognition state. In addition, to address the inter- and intra-subject variabilities, type-2 fuzzy frameworks can be utilised as explored in these works [60, 61]. Similarly, EFCMs estimated values of EC can be mapped generally to the specialisation and neural reuse of the DCN processes, however not much insight can be gained about the activation(s) of the individual cortical regions. This is mainly because of how EFCMs seek to find the optimal values of the EC, by trying to minimise the error between the estimated and the actual values of the fNIRS signals with the help of GA. Hence, not much could be inferred about which part of the cortex is, for example, more active from the optimised EC values.

In the next section, we review some of the AI methods as applied to DCN studies.

B. *AI in Developmental Cognitive Neuscience*

The de-facto standard for analysis of DCN studies is univariate analysis based on simple statistical tests, where the cortical regions most active in response to the presented stimulus is recognised, i.e., it is an activation based analysis. There is also a tendency of translating models used in adult research to DCN; however, this entails making some assumptions. In contrast, very few DCN studies have focused on decoding the multivariate patterns in brain activity of infants in response to the presented stimuli (such as [62] which is a correlation based MVPA). In fact, there is an evident scarcity for undertaking AI methods in DCN research.

In this subsection, we review the non-explainable and explainable AI methods as applied to DCN studies for conducting MVPA.

1) Non-Explainable AI Method

(a) MVPA with EEG using SVM

In the study by Bayet *et al.* [63], time resolved EEG based MVPA is conducted using a linear SVM. Infants aged 12 to 15 months participated in the study. The aim of the study was to investigate whether neural representations in the adult brain are different from the developing brain for the processing of visual stimuli (animals vs human body parts). The group-wise classification results of the SVM based MVPA was able to successfully decode between infants' and adults' brain activation patterns in response to the presented stimuli. However, infant multivariate representations didn't linearly separate for animal and body images.

The study was able to establish that neural representation for visual information processing, of animals vs body parts differ significantly between infants and adults. These findings were significant by suggesting that the cortical networks undergo the processes of localisation and specialisation to process the presented visual stimulus information. However, the study could not shed light on what cortical networks were activated for adults, and likewise, for infants, that could explain the underlying brain mechanism correspondingly. This is mainly because of the non-explainable inference mechanism of SVM as discussed previously in III-A1.

(b) MVPA with fNIRS using correlation

In contrast to EEG's MVPA analysis (which is temporally driven), an fNIRS based MVPA is aimed at spatial investigations into the cortical regions' activations encoded in the MVM. A hypothetical construction of a MVM using six fNIRS signals reading from six different cortical regions is depicted in Fig. 7 (a) - (b). The work by Emberson et al. [62] decoded the brain responses in 19 six-months-old infants' fNIRS signals in response to auditory and visual stimuli. They decoded the signals by undertaking a MVPA driven by correlation and reported an average classification accuracy of 66.67% for trial-level decoding.

The significance of their work lies in usage of MVPA that improved the decoding sensitivity in comparison to their previous work that used univariate methods [64]. A feature significance analysis was also undertaken to determine which features (fNIRS channels) are most significant for recognising the fNIRS signals in response to visual and auditory stimuli. Their results indicated channel 1 (occipital cortex), channel 3 (occipital cortex), and channel 8 (prefrontal cortex) to be the most critical channels for decoding between visual and auditory stimuli.

The identification of fNIRS channels and their corresponding anatomical locations sheds light on the localised activation of the cortex as delineated by the IS framework. In addition, the improved sensitivity of MVPA on account of analysing more than one variable (fNIRS channels' activity) rather than univariate analysis further corroborates that cortical networks (interaction between multiple cortical regions) are formed for the processing of perceptual stimuli. In this sense, the correlation based MVPA is able to implicitly imply the formation of cortical networks. However, what exactly entails the cortical networks is unknown because the presence and type of interaction between the fNIRS channels is unrevealed by the correlation based MVPA.

Motivated from the success of the correlation based MVPA by Emberson [62] and to overcome its limitation of partial explainability, we designed an explainable MVPA (xMVPA) which is reviewed next.

2) XAI Method: MVPA with fNIRS using eXplainable MVPA (xMVPA)

In order to retain the brain activity patterns in MVM during the learning of the classification mechanism of a given ML algorithm to drive MVPA, a previous work from our group [65] has explored using Fuzzy Logic to power MVPA, called eXplainable MVPA (xMVPA). The Fuzzy Logic System (FLS) is unique in its ability to compute with words (CWW) as well as account for the uncertainty in the input data by assigning a membership grade μ in the range $[0, 1]$ to each input value x.

In the work [65] an interval type-2 fuzzy logic system (IT2-FLS) [66] is used for powering the MVPA to analyse the brain activity patterns of six-month-old infants in response to sensory inputs. The MVM, constructed from fNIRS channels of interest on the occipital (associated with visual processing), temporal (associated with auditory processing), and PFC (associated with thinking and planning), is first converted into a conceptual linguistic label (CoL) MVM based on the definition of the membership functions (MFs) of the CoLs, as illustrated in Fig. 7 (c) - (d). A mathematical definition of IT2 membership functions is given in eq. (4).

$$\begin{aligned}\tilde{A} = \{(x, \mu, 1) | \forall x \in X,\\ \forall \mu \in [\mu_{\tilde{A}}(x), \mu_{\tilde{A}}(x)] \subseteq [0, 1]\}\end{aligned} \tag{4}$$

where $\mu_{\tilde{A}}$ represent the MF of interval type-2 fuzzy set $\tilde{A}$ defined over input x.

The CoL MVM is then fed to an evolutionary algorithm (such as GA) to find optimal patterns in the CoL MVM (train dataset) such that maximum classification accuracy can be obtained on the test dataset of the CoL MVM (Fig. 7 (e)). The discerned patterns by the xMVPA are able to shed light on the activations and interconnections of the activated regions in the cortex in response to the presented stimuli. A general nomenclature of a pattern discerned by xMVPA is given in eq. (5).

$$\begin{aligned}Pattern\ P_m : IF\ x_1\ is\ A_{q_1}\ AND\ ...AND\ x_n\ is\ A_{q_n}\\ THEN\ class\ is\ C_j\ with\ PW_m\end{aligned} \tag{5}$$

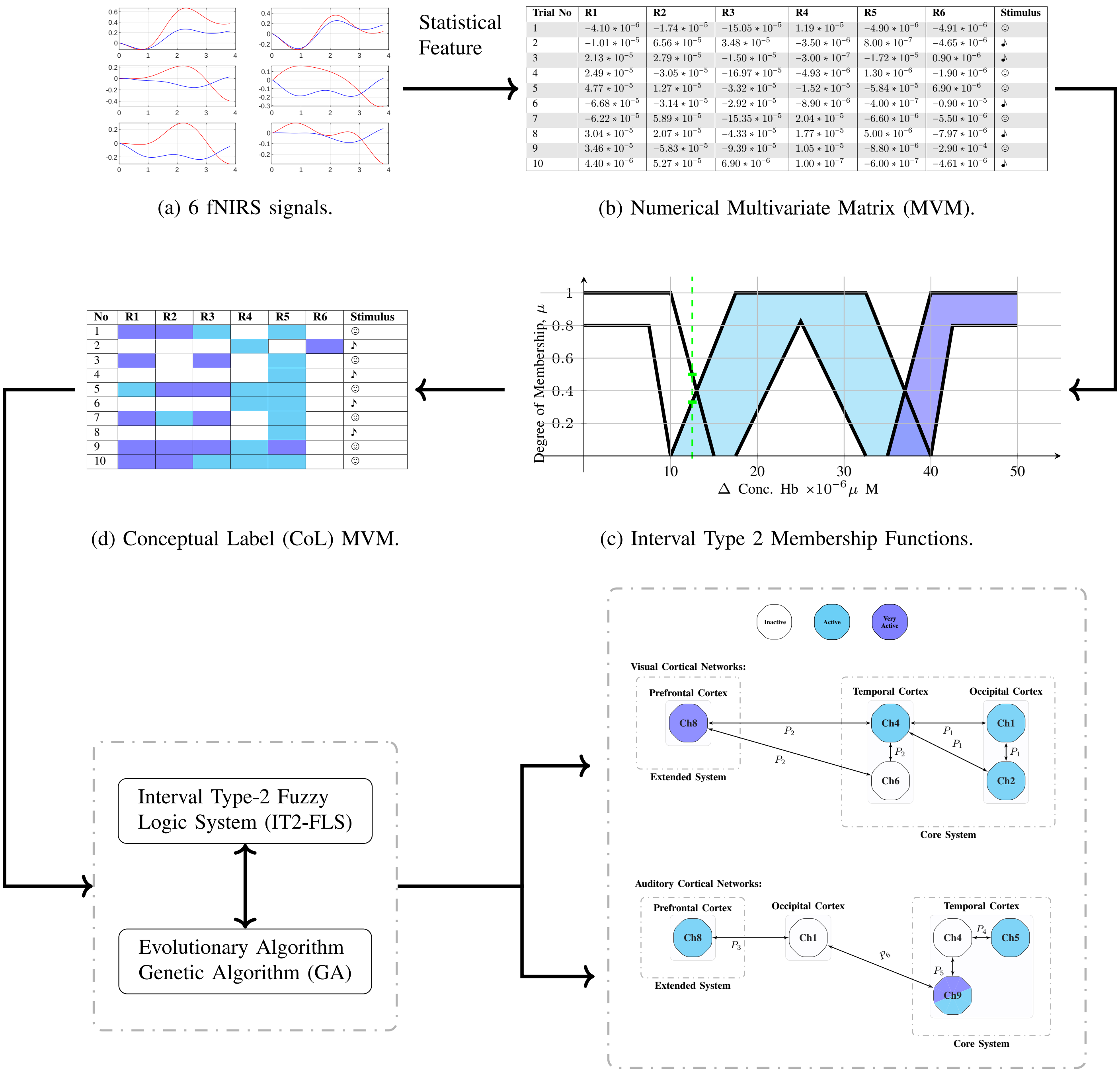
where x_i are the crisp brain activity values for the variable i (fNIRS channel), A_i is the antecedent set for the i_{th} variable with a total of n variables and $j \in J$ number of output classes. The class predicted with the pattern is C_j with the pattern weight given by PW_m. The pattern weight, PW_m, is a measure of the relevance of a given pattern as deemed by xMVPA based on its confidence, c, and support, s, values. The confidence can be viewed as the conditional probability that, given the antecedent(s) of P_m, how likely the predicted class will be C_j, whereas support measures the coverage (based on the number of matching data instances) of training dataset by the pattern P_m. The $\overline{upper}$ and $\underline{lower}$ PW_m are defined as outlined in eq. (6).

$$\begin{aligned}\overline{PW_m} &= \overline{c_m} * \overline{s_m}\\ \underline{PW_m} &= \underline{c_m} * \underline{s_m}\end{aligned} \tag{6}$$

where c_m and s_m are confidence and support for the pattern P_m which can be calculated as outlined in [67].

The measurement of PW for each pattern enables a comparison to evaluate the efficacy of each learnt pattern for the given classification task. In the study by Andreu-Perez *et al.* [65] a total of six patterns were identified for the processing of sensory stimuli that were able to recognise brain activity patterns of six-month-old infants, for the visual and auditory stimuli, encoded in CoL MVM that obtained a classification accuracy of 67.69%. In [65], the patterns as well as the MF definitions (start, height and endpoints) were learnt using GA with a total number of hyperparameters to be learnt equal to 300.

The xMVPA delineated patterns for visual processing that shed light on an occipital-temporal network as a core system



Trial No	R1	R2	R3	R4	R5	R6	Stimulus
1	$-4.10 * 10^{-6}$	$-1.74 * 10^{-5}$	$-15.05 * 10^{-5}$	$1.19 * 10^{-5}$	$-4.90 * 10^{-6}$	$-4.91 * 10^{-6}$	☺
2	$-1.01 * 10^{-5}$	$6.56 * 10^{-5}$	$3.48 * 10^{-5}$	$-3.50 * 10^{-6}$	$8.00 * 10^{-7}$	$-4.65 * 10^{-6}$	♪
3	$2.13 * 10^{-5}$	$2.79 * 10^{-5}$	$-1.50 * 10^{-5}$	$-3.00 * 10^{-7}$	$-1.72 * 10^{-5}$	$0.90 * 10^{-6}$	♪
4	$2.49 * 10^{-5}$	$-3.05 * 10^{-5}$	$-16.97 * 10^{-5}$	$-4.93 * 10^{-6}$	$1.30 * 10^{-6}$	$-1.90 * 10^{-6}$	☺
5	$4.77 * 10^{-5}$	$1.27 * 10^{-5}$	$-3.32 * 10^{-5}$	$-1.52 * 10^{-5}$	$-5.84 * 10^{-5}$	$6.90 * 10^{-6}$	☺
6	$-6.68 * 10^{-5}$	$-3.14 * 10^{-5}$	$-2.92 * 10^{-5}$	$-8.90 * 10^{-6}$	$-4.00 * 10^{-7}$	$-0.90 * 10^{-5}$	♪
7	$-6.22 * 10^{-5}$	$5.89 * 10^{-5}$	$-15.35 * 10^{-5}$	$2.04 * 10^{-5}$	$-6.60 * 10^{-6}$	$-5.50 * 10^{-6}$	☺
8	$3.04 * 10^{-5}$	$2.07 * 10^{-5}$	$-4.33 * 10^{-5}$	$1.77 * 10^{-5}$	$5.00 * 10^{-6}$	$-7.97 * 10^{-6}$	♪
9	$3.46 * 10^{-5}$	$-5.83 * 10^{-5}$	$-9.39 * 10^{-5}$	$1.05 * 10^{-5}$	$-8.80 * 10^{-6}$	$-2.90 * 10^{-4}$	☺
10	$4.40 * 10^{-6}$	$5.27 * 10^{-5}$	$6.90 * 10^{-6}$	$1.00 * 10^{-7}$	$-6.00 * 10^{-7}$	$-4.61 * 10^{-6}$	♪

No	R1	R2	R3	R4	R5	R6	Stimulus
1	🟪	🟪	🟦		🟦		☺
2				🟦		🟪	♪
3	🟪		🟪		🟦		☺
4					🟦		♪
5	🟦	🟪	🟪	🟦	🟦		☺
6				🟦	🟦		♪
7	🟪	🟦	🟪		🟦		☺
8					🟦		♪
9	🟪	🟪	🟪	🟦	🟪		☺
10	🟪	🟪	🟦	🟦	🟦		☺

(a) 6 fNIRS signals.

(b) Numerical Multivariate Matrix (MVM).

(d) Conceptual Label (CoL) MVM.

(c) Interval Type 2 Membership Functions.

(e) xMVPA inference mechanism

(f) The patterns delineated by xMVPA.

Fig. 7: (a) Six hypothetical fNIRS signals (b) An illustrative multivariate matrix (MVM) of brain activity from 6 different regions (R) for a total of 10 trials. The rows are trials, whereas the columns are the different regions from where brain activity is measured.(c) An illustrative plot to exemplify how conceptual labels (CoLs) can be used to represent brain activity. Brain regions can be expressed with the CoLs *Inactive*, *Active* and *Very Active* with approximate degree of membership values, μ. The derived ambiguity in the degree of membership ensures that uncertainty in the numeric data (or neuroimaging reading from fNIRS) is well retained upon transformation into a conceptual label (CoL). (d) The corresponding CoL MVM. Here, the colours denote the level of activity in different brain regions in response to the presented stimuli. For example, violet representing high activity, cyan low activity, and white no activity. (e) The CoL MVM is given as an input to xMVPA which computes the classification accuracy of a given set of patterns on the given data by computing with words (CWW) using interval type-2 fuzzy logic system (IT2-FLS). The optimal set of patterns is found using an evolutionary algorithm (a genetic algorithm is used in the study). (f) A model for visual and auditory processing in six-month-old infants, based on the six patterns (P_1 - P_6) revealed by the xMVPA inference mechanism [65]. The colour of the channel's (Ch) octagon is based on its activity level: Inactive (white), Active (cyan), and Very Active (violet).

that undertakes the primary processing of facial features, and of the PFC as an extended system that processes the emotion associated with the visual stimulus, as shown in Fig. 7 (f). The proposed model for auditory information processing consists of the temporal cortex as a core system for processing non-speech auditory stimuli, and of the PFC as an extended system that processes the emotion associated with the auditory stimulus, also shown in Fig. 7 (f).

IV. Discussion

The implications for a greater insight into the developing brain mechanisms, both structural (physical) and functional (cognitive), are profound. According to a study in 2007, one in every four people is affected by a mental illness either directly or indirectly [74]. In this regard, the study of a developing brain can inform us about *typical and atypical brain developmental trajectories* which may in turn facilitate the early *identification of and intervention for brain disorders* during childhood. However, at present, the DCN research has limited translation into shaping brain development because of typically small datasets (owing mostly to non-cooperating infant behaviour), the non-availability of *a-priori* information about the underlying mechanisms in a developing brain, as well as a lack of explainability of the AI techniques applied for the analysis of infant's neuroimaging data.

The neuroimaging modalities of fNIRS and EEG are considered to be the most 'infant friendly', and have reached a pinnacle in their own right, where real time neuronal activity can be recorded with EEG, or can be localised to a corresponding anatomical location within 2cm using fNIRS. To benefit from their complimentary high resolutions, more recent studies have undertaken multimodal (fNIRS-EEG) brain activity analysis to gain greater insights into functional brain development. Likewise the improvement in AI methods, in particular, the revolution of ANNs into DNNs and the remarkable feature learning ability of CNNs with hierarchical networks has led to breakthroughs in many challenging image classification, and speech recognition problems [43]. However, despite the advent of advanced neuroimaging technologies and the availability of sophisticated CNNs, the DCN research has not benefited as much from the aforementioned technological and computational advances in comparison to other complex fields (such as image classification).

To this end, to bridge the gap between the information recorded by the neuroimaging technologies and the insights acquired from the analysis of the neuroimaging data, a review of the most prevalent AI techniques in different analysis paradigms is undertaken in the present work. In particular, the AI methods are investigated for their similarity with theoretical frameworks for DCN including Interactive specialisation (IS) [3] and the neuroconstructivist approach [28] (summarised in section II). The main processes in brain development include: 1) localisation 2) specialisation 3) parcellation and 4) neural reuse; and if an AI technique's learning mechanism can shed light on these DCN processes, it can then inform us about the underlying mechanisms of a developing brain in line with the aforementioned DCN processes.

In this regard, the inherent limitation of most AI methods to not be able to explain what was *observed* in terms of brain activity patterns, during learning of their inference mechanism, renders them ill-suited to shed light on the DCN processes despite obtaining remarkable classification performance. A comparison of the strength and limitations of the AI methods, reviewed in this work, is summarised in Table I. Indeed, the bottleneck is not the classification prowess of the reviewed AI methods owing to their advanced learning techniques to acquire abstract representations from the input data. The limitation of AI methods, as applied to DCN studies, is that without *explainability* of the learnt inference mechanism, not much insight can be gained on the activated cortical regions for a given task. Of the AI methods reviewed, the only XAI method in DCN, to the best of the author's knowledge, is fuzzy logic based xMVPA [65].

The capability of fuzzy logic in CWW (computing with words), and modelling uncertainty are particularly well-suited for neuroimaging data which is characterised by inter-subject variabilities. The xMVPA was able to discern patterns in the neural underpinnings of audio and visual processing in six-month-old infants. xMVPA is explainable since it identifies the patterns in the input data prototypical to the presented stimuli. The classification accuracy reflects the validity of the discerned patterns to represent the true brain activity patterns in correspondence to each stimulus. The learnt patterns are also explainable as they inform about the activations and interactions of the cortical areas.

The discerned patterns for the visual and auditory processing are illustrated in Fig. 7 (f). The cortical network formed for visual processing has a hierarchical structure with the processing of raw data processed in the occipital cortex, and the processed information is then passed to PFC where higher level processing is done. This pattern is widely observed in adult literature of visual processing [75], and verified the patterns found by xMVPA for processing of a visual stimulus in six-month-old infants. Based on the similarity of the cortical network formed for the visual stimulus with those of adults, the cortical network is identified as specialised.

In contrast, the cortical network formed for auditory processing is hypothesised to be a non-specialised cortical network. The authors [65] suggested the formed network to be non-specialised based on the 'inactive' activation status of a channel (Ch1) forming the link between the information pathway from the temporal cortex to PFC. This non-specialised network was hitherto unknown in DCN literature and was only discerned because of the explainable attributes of the xMVPA.

Likewise, xMVPA can shed light on interactions and activations for time resolved brain activity. To investigate the process of neural reuse, the patterns would need to be established for different time points. In this regard, an analysis of the interconnections that were present at a given time point and how these interconnections rewired to acquire a new cognitive or behavioural state at a later time point can be investigated in line with the neural reuse process of functional brain development.

TABLE I: A comparison of different artificial intelligence (AI) methods used on adult(A) and infant(I) neuroimaging data for Connectivity Analysis (CA), Representation Learning (RepL) and Multivariate Pattern Analysis (MVPA) with fNIRS and EEG data. The strengths (S) and limitations (L) of the AI methods are summarised. For CA, separate AI methods most commonly used with functional connectivity (FC), and effective connectivity (EC) are reported. The full name of the algorithms reviewed are: Support Vector Machine (SVM), Random Forest (RF), 1Dimensional Convolutional Neural Network with Long Short Term Memory (1DCNN-LSTM), Ridge Regression (RR), Integrated Functional Manifold (IFM), EEGNet based Convolutional Neural Networks (EEGNet), Symbol-Concept Association Network (SCAN), Convolutional Neural Networks (CNN), Effective Fuzzy Cognitive Maps (EFCM), Induced Type-2 Fuzzy Deep Brain Learning Network (IT2FDBN), Independent Component Analysis based Fuzzy Neural Network (ICA-FNN) and eXplainable MVPA (xMVPA).

	Analysis Paradigm	**Artificial Intelligence (AI) Methods**	**STRENGTHS (S) and LIMITATIONS (L)**	**Population**
Non-Explainable AI (non-XAI) Methods	CA with EEG	FC: SVM [46] [47], RF: [68]	**S:** Robust regression mechanism. **L:** Limited spatial localisation.	**A**
		EC: 1DCNN-LSTM [69]	**S:** Automated feature learning. **L:** Dependency on large datasets.	**A**
	CA with fNIRS	FC: RR [48]	**S:** Simple analytical model. **L:** Results are dependent on regressors (β).	**A**
		FC: IFM [49]	**S:** Groupwise exploration is possible. **L:** Manifold assumption.	**A**
	RepL with EEG	EEGNet (CNN) [52] SCAN (CNN) [70]	**S:** Automated feature learning. **L:** Dependency on large datasets.	**A**
	RepL with fNIRS	CNN [71] , [72], [53]	**S:** Automated feature learning. **L:** Dependency on large datasets.	**A**
	MVPA with EEG	SVM [55] [63]	**S:** Robust classification mechanism. **L:** Limited spatial localisation.	**A + I**
	MVPA with fNIRS	Correlation [62]	**S:** Better spatial localisation with infant level and trial level decoding results. **L:** Lack of insight into the interactions between the important fNIRS channels.	**I**
Explainable AI (XAI) Methods	CA with fNIRS	EC: EFCM [58]	**S:** Learnt parameters are EC. **L:** Performance non-scalable.	**A**
	FNNs with EEG	IT2-FDBL [73]	**S:** The empirical model proposed mimics short term memory of the brain. **L:** Dependency on large datasets.	**A**
		ICA-FNN [59]	**S:** Automatic feature selection with uncertainty handled with fuzzy models. **L:** The obtained rules are not comparable per se because of online definition of linguistic labels.	**A**
	MVPA with fNIRS	xMVPA [65]	**S:** Brain activity patterns defined by CWW. **L:** Linguistic variables are determined a priori.	**I**

V. Conclusion

Cognitive developmental delays and abnormalities are commonly associated with behavioural disorders that can become challenging conditions to treat in adulthood (*such as attention deficit hyperactivity disorder, autism spectrum or bipolar disorders*) [76]. The application of AI for cognitive neuroscience can extend novel ways of interrogating brain function by maximizing neuroimaging data. This is of particular interest for the study of developmental brains where the classical assumptions of brain function for adults cannot serve as guidance.

In this paper, the aim was to highlight the current gap in DCN research due to non-explainable AI methods. Since there is no insight obtained on the learnt classification mechanism on the basis of brain activity patterns, this critically limits the translation of DCN research to shape developing brain trajectories despite acquiring statistically significant classification results. To bridge the gap between DCN research and the translation of their insight(s), we suggest that future DCN research adopt a more explainable classification mechanism using XAI methods such as xMVPA [65].

References

[1] J. Stiles and T. L. Jernigan, “The basics of brain development,” *Neuropsychology review*, vol. 20, pp. 327–48, 2010.

[2] S. Ackerman. "Discovering the Brain." (1992), [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK234146/ (visited on 01/19/2021).

[3] M. H. Johnson, "Functional brain development in humans," *Nature Reviews Neuroscience*, vol. 2, pp. 475–483, 2001.

[4] Y. Munakata, B. J. Casey, and A. Diamond, " Developmental cognitive neuroscience: progress and potential," *Trends in Cognitive Sciences*, vol. 8, pp. 122–128, 3 2004.

[5] A. Karmiloff-Smith, "Development itself is the key to understanding developmental disorders," *Trends in Cognitive Sciences*, vol. 2, pp. 389–398, 10 1998.

[6] M. H. Johnson, "Into the minds of babes," *Science*, vol. 286, p. 247, 5438 1999.

[7] T. Wilcox and M. Biondi, "Fnirs in the developmental sciences," *WIREs Cognitive Science*, vol. 6, pp. 263–283, 2015.

[8] S. Lloyd-Fox, A. Blasi, and C. E. Elwell, "Illuminating the Developing Brain: The Past, Present and Future of Functional Near Infrared Spectroscopy," *Neuroscience and Biobehavioural Reviews*, vol. 34, pp. 269–84, 2010.

[9] A. Dereymaeker *et al.*, "Review of sleep-eeg in preterm and term neonates," *Early Human Development*, vol. 113, no. 1, pp. 87–103, 2017.

[10] P. Fransson *et al.*, "Spontaneous Brain Activity in the Newborn Brain During Natural Sleep—An fMRI Study in Infants Born at Full Term," *Pediatric Research*, vol. 66, pp. 301–305, 2009.

[11] A. Blasi *et al.*, "Early specialization for voice and emotion processing in the infant brain," *Current Biology*, vol. 21, pp. 1220–1224, 14 2011.

[12] B. Deen, H. Richardson, D. Dilks, and et al., "Organization of high-level visual cortex in human infants," *Nature Communications*, vol. 8, p. 13 995, 3 2017.

[13] C. T. Ellis, L. J. Skalaban, T. S. Yates, V. R. Bejjanki, N. I. Córdova1, and N. B. Turk-Browne, "Re-imagining fMRI for awake behaving infants," *Nature Communications*, vol. 11, 2020.

[14] I. de Roever, G. Bale, S. Mitra, J. Meek, N. J. Robertson, and I. Tachtsidis, "Investigation of the pattern of the hemodynamic response as measured by functional near-infrared spectroscopy (fNIRS) studies in newborns, less than a month old: a systematic review," *Frontiers in Human Neuroscience*, vol. 12, p. 371, 2018.

[15] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10–20 system electrode placement," *Electroencephalography and Clinical Neurophysiology*, vol. 66, pp. 376–382, 1987.

[16] J.-D. Haynes and R. Geraint, "Decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, pp. 523–534, 2006.

[17] J. W. Britton, L. C. Frey, J. L. Hopp, and et al., "Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants," in Chicago: American Epilepsy Society, 2016, ch. Introduction.

[18] P. L. Nunez and R. Srinivasan, *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, 2006.

[19] M. de Haan, M. H. Johnson, and H. Halit, "Development of face-sensitive event-related potentials during infancy: A review," *International Journal of Psychophysiology*, vol. 51, pp. 45–58, 1 2003.

[20] B. Burle, L. Spieser, C. Roger, L. Casini, T. Hasbroucq, and F. Vidal, "Spatial and temporal resolutions of EEG: Is it really black and white. A scalp current density view," *International journal of Psychophysiology*, vol. 97, pp. 210–220, 3 2015.

[21] K. T. Sweeney, T. E. Ward, and S. F. McLoone, "Artifact removal in physiological signals—practices and possibilities," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, pp. 488–500, 3 2012.

[22] M. Saadati, J. Nelson, and H. Ayaz, "Multimodal FNIRS-EEG classification using deep learning algorithms for brain-computer interfaces purposes," *In International Conference on Applied Human Factors and Ergonomics*, 2019.

[23] L. C. Chen, P. Sandmann, J. D. Thorne, and et al., "Association of Concurrent fNIRS and EEG Signatures in Response to Auditory and Visual Stimuli," *Brain Topography*, vol. 28, pp. 710–725, 1 2015.

[24] L. French and E. M. Kennedy, "Annual Research Review: Early intervention for infants and young children with, or at-risk of, autism spectrum disorder: a systematic review," *Journal of Child Psychology and psychiatry*, vol. 59, 4 2018.

[25] N. Ganea *et al.*, "Development of adaptive communication skills in infants of blind parents," *IEEE Transactions on Fuzzy Systems*, vol. 54, p. 2265, 12 2018.

[26] M. A. Schel and T. Klingberg, "Specialization of the right intraparietal sulcus for processing mathematics during development," *Cerebral Cortex*, vol. 27, pp. 4436–4446, 9 2017.

[27] A. Karmiloff-Smith, "Beyond modularity: A developmental perspective on cognitive science," *European Journal of Disorders of Communication*, vol. 29, pp. 95–105, 1 1994.

[28] A. Karmiloff-Smith, "Preaching to the converted? From constructivism to neuroconstructivism," *Child Development Perspectives*, vol. 3, pp. 99–102, 2 2009.

[29] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, pp. 186–198, 2009.

[30] S. L. Pendl *et al.*, "Emergence of a hierarchical brain during infancy reflected by stepwise functional connectivity," *Human brain mapping*, vol. 38, pp. 2666–2682, 5 2017.

[31] H. C.Barrett, "A hierarchical model of the evolution of human brain specializations," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 10 733–10 740, 1 2012.

[32] M. H. Johnson and M. de Haan, *Developmental Cognitive Neuroscience*. Wiley, 2010.

[33] ——, "Interactive specialisation," in *Developmental Cognitive Neuroscience*, Wiley, 2010, ch. 12, pp. 204–223.

[34] D. D'Souza and A. Karmiloff-Smith, "Why a developmental perspective is critical for understanding human cognition," *Behavioral and Brain Sciences*, vol. 39, 2016.

[35] M. Anderson and M. Penner-Wilger, "Neural reuse: A fundamental organizational principle of the brain," *Behavioral and Brain Sciences*, vol. 33, pp. 245–266, 2010.

[36] P. S. Churchland and T. J.Sejnowski, "Perspectives on cognitive neuroscience," *Science*, vol. 242, pp. 741–745, 4879 1998.

[37] O. Theobald, *Machine learning for absolute beginners: a plain English introduction*. Scatterplot Press, 2017.

[38] A. M. Zador, "A critique of pure learning and what artificial neural networks can learn from animal brains," *Nature Communications*, vol. 10, 3770 2019.

[39] T. Wilcox, H. Bortfeld, R. Woods, E. Wruck, and D. A. Boas, "Reinforcement learning in the brain," *Journal of Mathematical Psychology*, vol. 53, pp. 139–154, 3 2009.

[40] A. A. Fingelkurts, A. A. Fingelkurts, and S. Kähkönen, "Functional connectivity in the brain—is it an elusive concept?" *Neuroscience and Biobehavioral Reviews*, vol. 28, pp. 827–836, 8 2005.

[41] K. J. Friston, "Functional and effective connectivity: a review," *Brain Connectivity*, vol. 1, pp. 13–36, 1 2011.

[42] K. J. Friston, C. Frith, and R. Frackowiak, "Time-dependent changes in effective connectivity measured with PET," *Human Brain Mapping*, vol. 1, pp. 69–79, 1 1993.

[43] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.

[44] J. V. Haxby, A. C. Connolly, and J. S. Guntupalli, "Decoding neural representational spaces using multivariate pattern analysis," *Annual Review of Neuroscience*, vol. 37, pp. 435–456, 2014.

[45] K. Norman, S. Polyn, G. Detre, and J. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fMRI data," *Trends in Cognitive Sciences*, vol. 10, pp. 424–30, 9 2006.

[46] B. Moezzi *et al.*, "Characterization of Young and Old Adult Brains: An EEG Functional Connectivity Analysis," *Neuroscience*, vol. 422, pp. 230–239, 2019.

[47] M. A. Klados, P. Konstantinidi, R. Dacosta-Aguayo, V.-P.Kostaridou, A. Vinciarelli, and M. Zervakis, "Automatic Recognition of Personality Profiles Using EEG Functional Connectivity during Emotional Processing," *Brain Sciences*, vol. 10, pp. 2076–3425, 5 2020.

[48] L. Duan, N. T. V. Dam, H. Ai, and P. Xu, "Intrinsic organization of cortical networks predicts state anxiety: an functional near-infrared spectroscopy (fNIRS) study," *Translational Psychiatry*, vol. 10, 402 2020.

[49] S.-M. Ávila-Sansores, G. Rodríguez-Gómez, I. Tachtsidis, and F. Orihuela-Espina, "Interpolated functional manifold for functional near-infrared spectroscopy analysis at group level," *Neurophotonics*, vol. 7, p. 045 009, 2020.

[50] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, pp. 1211–1279, 2 2012.

[51] X. Zhang, D. Cao, J. Liu, Q. Zhang, and M. Liu, "Protocol: Effectiveness and safety of brain-computer interface technology in the treatment of poststroke motor disorders: A protocol for systematic review and meta-analysis," *British Medical Journal Open*, vol. 11, 1 2021.

[52] V. J. Lawhern, A. J. Solon, N. Waytowich, S. Gordon, C. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for

EEG-based brain–computer interfaces.," *Journal of Neural Engineering*, vol. 15, p. 056 013, 2018.

[53] A. Janani, M. Sasikala, C. Harleen, N. Shajil, and G. Venkatasubramanian, "Investigation of deep convolutional neural network for classification of motor imagery fNIRS signals for BCI applications," *Biomedical Signal Processing and Control*, vol. 62, p. 102 133, 2020.

[54] J. J. Fahrenfort, J. van Driel, S. van Gaal, and C. N. L. Olivers, "From ERPs to MVPA Using the Amsterdam Decoding and Modeling Toolbox (ADAM)," *Frontiers in Neuroscience*, vol. 12, p. 368, 2018.

[55] M. Cauchoix, G. Barragan-Jason, T. Serre, and E. J. Barbeau, "The Neural Dynamics of Face Detection in the Wild Revealed by MVPA," *Journal of Neuroscience*, vol. 34, pp. 846–854, 3 2014.

[56] F. Shakerin and G. Gupta, "White-box induction from svm models: Explainable ai with logic programming," *Theory and Practice of Logic Programming*, vol. 20, pp. 656–670, 5 2020.

[57] C. P. R. Vieira and L. A. Digiampietri, "A study about explainable articial intelligence: Using decision tree to explain svm," *Revista Brasileira de Computação Aplicada*, vol. 12, pp. 113–121, 1 2020.

[58] M. Kiani, J. Andreu-Perez, H. Hagras, E. I. Papageorgiou, M. Prasad, and C.-T. Lin, "Effective Brain Connectivity for fNIRS with Fuzzy Cognitive Maps in Neuroergonomics," *IEEE Transactions on Cognitive and Developmental Systems*, 2019.

[59] C. Lin, L. W. Ko, and et al., "Adaptive EEG-based alertness estimation system by using ICA-based fuzzy neural networks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, pp. 2469–2476, 2006.

[60] A. Saha, A. Konar, and A. K. Nagar, "EEG analysis for cognitive failure detection in driving using type-2 fuzzy classifiers," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 1, pp. 437–453, 6 2017.

[61] L. Ghosh, A. Konar, P. Rakshit, and A. K. Nagar, "Hemodynamic analysis for cognitive load assessment and classification in motor learning tasks using type-2 fuzzy sets," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, pp. 245–260, 3 2018.

[62] L. L. Emberson, B. D. Zinszer, R. D. S. Raizada, and R. N. Aslin, "Decoding the infant mind: Multivariate pattern analysis (MVPA) using fNIRS," *Public Library of Science ONE*, vol. 12, e0172500, 2017.

[63] L. Bayet, B. D. Zinszer, E. Reilly, and et al., "Temporal dynamics of visual representations in the infant brain," *Developmental Cognitive Neuroscience*, vol. 45, p. 100 860, 2020.

[64] L. L. Emberson, J. E. Richards, and R. N. Aslin, "Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 9585–9590, 31 2015.

[65] J. Andreu-Perez, L. L. Emberson, M. Kiani, M. L. Filippetti, H. Hagras, and S. Rigato, "Explainable Artificial Intelligence Based Analysis for Interpreting Infant fNIRS Data in Developmental Cognitive Neuroscience," Under Review.

[66] H. Hagras, "Toward human-understandable, explainable AI," *Computer*, vol. 51, pp. 28–36, 2018.

[67] M. Antonelli, D. Bernardo, H. Hagras, and F. Marcelloni, "Multiobjective Evolutionary Optimization of Type-2 Fuzzy Rule-Based Systems for Financial Data Classification," *IEEE Transactions on Fuzzy Systems*, vol. 25, pp. 249–264, 2 2017.

[68] C. Kamarajan *et al.*, " Random Forest Classification of Alcohol Use Disorder Using EEG Source Functional Connectivity, Neuropsychological Functioning, and Impulsivity Measures," *Behavioral Sciences*, vol. 10, 3 2020.

[69] A. Saeedi, M. Saeedi, A. Maghsoudi, and et al., "Major depressive disorder diagnosis based on effective connectivity in EEG signals: a convolutional neural network and long short-term memory approach," *Cognitive Neurodynamics*, 2020.

[70] G. Honke *et al.*, "Representation learning for improved interpretability and classification accuracy of clinical factors from EEG," *arXiv preprint arXiv:2010.15274.*, vol. 7, 2020.

[71] M. Saadati, J. Nelson, and H. Ayaz, "Mental Workload Classification From Spatial Representation of FNIRS Recordings Using Convolutional Neural Networks," *IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2019.

[72] M. A. Tanveer, M. J. Khan, M. J. Qureshi, N. Naseer, and K. Hong, "Enhanced Drowsiness Detection Using Deep Learning: An fNIRS Study," *Frontiers in Psychology*, vol. 7, pp. 137 920–137 929, 2019.

[73] L. Ghosh, A. Konar, P. Rakshit, and A. K. Nagar, "Mimicking Short-Term Memory in Shape-Reconstruction Task Using an EEG-Induced Type-2 Fuzzy Deep Brain Learning Network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, pp. 571–588, 4 2020.

[74] *Information about Mental Illness and the Brain*, https://www.ncbi.nlm.nih.gov/books/NBK20369/, Accessed: 2020-12-14, 2007.

[75] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, "The distributed human neural system for face perception," *Trends in Cognitive Science*, vol. 4, pp. 223–233, 6 2000.

[76] E. G. Feil *et al.*, " Early Intervention for Preschoolers at Risk for Attention-Deficit/Hyperactivity Disorder: Preschool First Step to Success," *Behavioral Disorders*, vol. 41, pp. 95–106, 2 2016.