EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR FUNCTIONAL BRAIN DEVELOPMENT ANALYSIS: METHODS AND APPLICATIONS

By

MEHRIN KIANI

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF ESSEX
School of Computer Science and Electrical Engineering

JANUARY 2022

To the Faculty of University of Essex:

The members of the Committee appointed to examine the dissertation of MEHRIN KIANI find it satisfactory and recommend that it be accepted.

Prof. Ian Maynard, Ph.D., Chair

Mahdi Mahfouf, Ph.D.

Carina De Klerk, Ph.D.

DECLARATION

Some work and illustrations in the thesis have been quoted verbatim from the following articles:

- **Chapter 2.1 and Chapter 4:** M. Kiani, J. Andreu-Perez, H. Hagras, S. Rigato, and M. L. Filippetti, ‘Towards Understanding Human Functional Brain Development with Explainable Artificial Intelligence: Challenges and Perspectives’, in *IEEE Computational Intelligence Magazine*, (2021), In Press.

- **Chapter 5.1 and Chapter 6.1:** 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,’ in *IEEE Transactions on Cognitive and Developmental Systems*, (2019).

- **Chapter 5.2 and Chapter 6.2:** J. Andreu-Perez and L. L. Emberson and M. Kiani and M. L. Filipepetti and H. Hagras and S. Rigato, ‘Explainable artificial intelligence based analysis for interpreting infant fNIRS data in developmental cognitive neuroscience’, *Communications Biology*, vol. 4 (2021).

- **Chapter 5.3 and Chapter 6.3:** M. Kiani, J. Andreu-Perez, and H. Hagras, ‘A Temporal Type-2 Fuzzy System Based Approach for Time-dependent Explainable Artificial Intelligence’, Under Review at IEEE Transactions on Artificial Intelligence.

The **datasets** used in this work have been published in earlier studies:

- **Chapter 6.1:** J. Andreu-Perez, D. R. Leff, K. Shetty, A. Darzi, G. Z. Yang, ‘Disparity in Frontal Lobe Connectivity on a Complex Bimanual Motor Task Aids in Classification of Operator Skill Level’, *Brain Connect.*, vol. 6, (2016).

- **Chapter 6.2:** L. L. Emberson and B. D. Zinszer and R.D. Raizada and R. N. Aslin, ‘Decoding the infant mind: Multivariate pattern analysis (MVPA) using fNIRS’, *PloS one*, vol. 12, (2017).

- **Chapter 6.3:** L. M. Candanedo and V. Feldheim,‘Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models’, *Energy and Buildings*, vol. 112 (2016).

Abstract

In the last decades, non-invasive and portable neuroimaging techniques, such as functional Near-Infrared Spectroscopy (fNIRS), have allowed researchers to study the mechanisms underlying the functional development of the human brain, thus furthering the potential of Developmental Cognitive Neuroscience (DCN). However, the traditional methods used for the analysis of infant fNIRS data are still quite limited. Here, I introduce new Fuzzy Cognitive Maps, called EFCMs, for Effective Connectivity (EC) analysis of infants' fNIRS data. EFCMs can outline the interconnections between the cortical areas as well as specify the direction of EC. In contrast, to shed light on the activation level of the cortical regions, I developed a Multivariate Pattern Analysis (MVPA). The proposed MVPA is powered by eXplainable Artificial Intelligence (XAI), named eXplainable MVPA (xMVPA). The xMVPA is exemplified in a DCN study that investigates visual and auditory processing in six-month-old infants with a classification accuracy of 67.69 %. The xMVPA can identify patterns of cortical interactions formed in response to presented stimuli as hypothesised by the DCN frameworks. However, xMVPA can only analyse cross-sectional DCN studies, i.e. it is not able to analyse the temporal dynamics associated with a longitudinal DCN study. To this end, I developed a novel time-dependent XAI (TXAI) system based on Temporal Type-2 Fuzzy Sets (TT2FS). The TXAI system is exemplified on an empirical study using a real-life intelligent environments dataset to solve a time-dependent classification problem and attained a classification accuracy of 94.08%. The proposed TXAI system has the potential to inform the evolution of a process (such as functional brain development) using temporal trajectories which in turn may assist in the delineation of brain developmental trajectories.

TABLE OF CONTENTS

Page

CHAPTERS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to thank my lead PhD supervisor Dr. Javier Andreu-Perez for his relentless efforts and continued support through out the course of my PhD. Dr. Javier Andreu-Perez has been mentoring me since my MRes days, and I owe all my research training, and accomplishments to him.

I would also like to thank earnestly Prof. Dr. Hani Hagras for his expert guidance in Fuzzy Logic Systems. I am also very grateful to Dr. Silvia Rigato and Dr. Maria Laura Filippetti for their instrumental role in developing my comprehension of Developmental Cognitive Neuroscience.

This work was also not possible without the love, and care of my family and friends. I am forever grateful to my parents, my *Rhett Butler*, and my three musketeers for always seeing me through.

IEEE Copyright Reserved

Chapter One

Introduction

In this chapter, I outline the motivation of my PhD research, and an overview of the state-of-the-art research prior to my PhD work is also presented. This is followed by the objectives of my PhD research, and the resultant contribution to science. Finally, the structure for the rest of the thesis is outlined at the end of the chapter.

1.1 Motivation

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 from the structural development as well as functional 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.

The DCN frameworks (outlined in Chapter 2.1) of Interactive Specialisation (IS) and Neural Reuse provide account for functional brain development during postnatal development. However, the fundamental question in DCN of how cognitive development is mediated by structural maturation and optimised interactions among brain regions remains open. To this end, for my PhD research, I developed new eXplainable Artificial Intelligence (XAI) methods (presented in Chapter 5) that can describe functional brain development as hypothesised by the DCN frameworks. More specifically, the first proposed XAI methods, named EFCM, is based on Fuzzy Cognitive Maps (FCMs) for the analysis of Effective Connectivity (EC) in neuroimaging data to implement the neural reuse framework. The second XAI method is based on Interval Type-2 (IT2) Fuzzy Logic Systems (FLS) and delineates cortical networks, as hypothesised by the IS framework, using explainable *If-Then* patterns. The third method is based on new Temporal Type-2 Fuzzy Sets (TT2FS) based Time-dependent XAI system that has the potential to inform brain developmental trajectories.

In the next section, I summarise the state-of-the-art research methods as applied to neuroimaging data and their potential to inform functional brain development.

1.2 Overview of the state-of-the-art prior to this thesis

The key limitations in developmental research are associated with the limited choice of neuroimaging techniques that can record brain activity non-invasively, and the difference in opinion surrounding the use of standardised and explainable analysis of the data. More recently, functional Near-Infrared Spectroscopy (fNIRS) has emerged as a popular choice for investigating infant brain development, and its association with cognition and behaviour. fNIRS is a non-invasive, portable, optical neuroimaging method that allows the measurement of cerebral activity using near-infrared (NIR) light with both good temporal (around 3-5s) and spatial resolution (within 2cm) [5]. fNIRS has enabled scientists to study asleep and awake infants alike both inside the laboratory settings and in natural environments [6].

The existent inference frameworks in adult fNIRS analysis involve the use of modelling techniques which assume that signal data coming from all subjects share standard attributes. Typically, these models are based on the assumption that a canonical Haemodynamic Response Function (HRF) generated in response to a specific stimulus can be represented as a linear combination of several sources (regressors) [7]. However, for DCN studies, it is unlikely for the HRF to be known *a priori* [8], and hence General Linear Models (GLMs) are not particularly suitable for the analysis of infants' fNIRS studies. Nevertheless, GLMs have been pivotal in outlining the relative contribution of different sources using regression techniques [9].

As they stand, the current analysis frameworks are designed for static modelling (for example, models which require *a priori* information such as a known HRF) and therefore cannot be extended to studying brain processes undergoing continuous changes and development. Therefore, as also highlighted in a recent review article [10], it is necessary to investigate new analytical perspectives in DCN, as models based on adult work are not adequate to study the developing brain. In line with the aim of the present study, Rosenberg and colleagues [10] encouraged the use of data-driven (that learn from the data directly without relying on a priori information) predictive models to shed light on the neural circuits that give rise to the development of cognition and behaviour.

Another common approach for the analysis of fNIRS data is using Functional Connectivity (FC) analysis. In general, FC analysis can be directed (using Granger Causality (GC)) as well as undirected (using correlation) [11] and can shed light on the functional organisation of the infant brain [12].The directed FC describes the direction of the FC for example from motor cortex to basal ganglia whereas undirected FC only identifies functionally connected areas for example motor cortex and basal ganglia. In contrast, Effective Connectivity (EC) analysis describes a representative circuit diagram (cortical circuit) explaining observed signals (infant fNIRS signals) [13]. In this regard, EC analysis has the potential to shed light on the cortical networks formed in infants for the processing of presented information. EC analysis in infants studies is usually carried out using Dynamic Causal Modelling (DCM) [14, 15], and requires estimation of the DCM model such as sensitivity matrix computations using Monte Carlo simulation [14]. In addition, Bayesian inference is used to test the specific hypothesis for selecting a DCM model that best explains the input fNIRS data [14]. Since, DCM relies on estimated values (for sensitivity matrix) and *a priori* information for Bayesian inference, it is not particularly suitable for infant fNIRS data analysis.

In contrast to connectivity analysis, state-of-the-art machine learning algorithms (e.g. Support Vector Machines (SVM), Random Forest (RF), and neural-network based approaches) are used for predictive analysis of neuroimaging data [16], and are specifically employed to distinguish between classes (stimuli) based on input data (brain responses). The advantage of using machine learning algorithms is that they are data driven i.e., independent of *a priori* information. However, these paradigms do not explain *what* particular relations of brain activity are prototypical for different stimuli [17, 18, 19]. To this end, FLS based XAI methods for the classification of fNIRS data have the potential to outline the prototypical patterns in the form of *If-Then* statements. Hence, in this work, the second XAI method (explainable multivariate pattern analysis (xMVPA)) is based on IT2-FLS for the implementation of IS framework for functional brain development.

Although xMVPA can shed light on the cortical networks activated in response to presented stimuli, it is ill-suited for the analysis of longitudinal fNIRS studies. Specifically, the limitation of xMVPA to only analyse cross-sectional fNIRS studies is because standard fuzzy sets (such as IT2) cannot integrate temporal information in their Membership Function (MF). This is a critical limitation in FLS, as without integration of the temporal information in the MF of standard fuzzy sets, FLS cannot inform about the evolution of a real-life process (such as brain developmental trajectories).

In order to overcome the aforementioned limitations in DCN research, in the next section, I outline the objectives of my PhD research.

1.3 Objective

The objective of my PhD work is to develop innovative XAI methods in neuroscience that can provide an explanation for their classification process as hypothesised by the DCN frameworks for the processing of presented stimuli. To this end, the XAI methods I developed are:

1. Implementation of Neural Reuse framework using Effective Fuzzy Cognitive Maps (EFCMs) - presented in Chapter 5.1.
2. Implementation of IS framework using Explainable Multivariate Pattern Analysis (xMVPA) - presented in Chapter 5.2.
3. Evolution of real-life temporal process (such as functional brain development) using Time-dependent XAI (TXAI) systems - presented in Chapter 5.3.

The EFCMs (presented in Chapter 5.1) offer a more powerful methodology for the analysis of Effective Connectivity (EC), than standard Fuzzy Cognitive Maps (FCMs), in fNIRS data. Importantly, EFCMs learn the EC values directly from the fNIRS data without any underlying model assumptions. The EC values, discerned by EFCM, are in the range of $[-1, 1]$ and reflect the type (positive or negative) of interaction as well as the direction (for example, from Prefrontal Cortex (PFC) to motor cortex). In particular, the EFCMs can shed light on how EC between cortical regions gets optimised/rewired upon acquisition of a skill (such as facial processing) in infants. In this regard, EFCM has the potential to inform functional brain development as hypothesised by the neural reuse framework. However, EFCMs offer a limited account of the DCN frameworks since they can only decipher the interactions between different cortical regions, and cannot describe the level of activation of the cortical regions.

To overcome the partial explainability of the EFCMs, with respect to its implementation within the DCN frameworks, I developed the second XAI method (xMVPA presented in Chapter 5.2) based on Interval Type-2 (IT2) fuzzy sets. The xMVPA can analyse the neural substrate of infant neuroimaging data in response to presented stimuli, and delineates cortical networks prototypical for the processing of the presented stimuli in the form of *If-Then* patterns. In this work, the xMVPA is exemplified (Chapter 6.2) in a fNIRS dataset with six-month-old infants, recorded whilst the infants are presented a visual and auditory stimuli. The prototypical patterns of cortical networks delineated by the proposed xMVPA suggests a specialised network for processing of visual stimuli and a non-specialised network for the processing of auditory stimuli in six-month-old infants. The implications for delineating the cortical networks formed in response to presented stimuli are profound informing functional brain development (more details in Chapter 6.2.3).

The xMVPA can inform functional brain development in terms of the activation and interaction of the cortical networks; however, xMVPA cannot analyse longitudinal data. More specifically, the xMVPA cannot analyse the temporal information associated with the longitudinal data, i.e. at what postnatal age the developmental changes are happening [20]. This is because the IT2 fuzzy sets, on which xMVPA is based, cannot account for temporal information in its MF. In this regard, to investigate the temporal dimension associated with brain development, the analysis of longitudinal DCN data, i.e. neuroimaging data recorded over a certain time period, is considered imperative. To this end, for the processing of longitudinal fNIRS studies, I developed Temporal Type-2 Fuzzy Sets (TT2FS) based TXAI system (presented in Chapter 5.3).

By analysing neuroimaging data acquired from infants at different ages, the proposed TXAI system has the potential to describe the prototypical patterns representing the cortical networks for each age/time point shedding light on DCN frameworks. Indeed, an explainable analysis of longitudinal data is fundamental to our understanding of functional brain developmental trajectories. The insights into brain developmental trajectories can inform our educational, and social policies. Importantly, this work may also assist in early identification of developmental disorders.

In the next section, I outline my PhD works' contributions to science.

1.4 Contribution to Science

My PhD research focused on the development of explainable methods to inform our understanding of functional brain development. To this end, following is a list of the major accomplishments of my PhD research:

1. Development of Effective Fuzzy Cognitive Maps (EFCMs)

    - Elucidates the Effective Connectivity (EC) between cortical areas (functional Near-InfraRed Spectroscopy (fNIRS) channels), presented in Chapter 5.1.
    - Exemplified on surgeons' fNIRS data where participating surgeons had varying levels of expertise for performing a motor task, presented in Chapter 6.1.
    - 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, Early Access
    - Link: https://ieeexplore.ieee.org/document/8929015

2. Review of eXplainable Artificial Intelligence (XAI) methods for understanding functional brain development

- Identifies the limitations in current Artificial Intelligence (AI) methods for understanding functional brain development, outlined in Chapter 4.
- Proposes XAI methods for informing functional brain development in line with the Developmental Cognitive Neuroscience (DCN) frameworks.
- M. Kiani, J. Andreu-Perez, H. Hagras, S. Rigato, and M. L. Filippetti, “Towards Understanding Human Functional Brain Development With Explainable Artificial Intelligence: Challenges and Perspectives,” *IEEE Computational Intelligence Magazine*, 2021, In Press
- Published (pre-print): https://arxiv.org/pdf/2112.12910.pdf

3. A new method for the development of eXplainable Multivariate Pattern Analysis (xMVPA) based on Type-2 fuzzy systems
    - Outlines functional brain development in line with the Interactive Specialisation (IS) framework, presented in Chapter 5.2.
    - Exemplified on six-month-old infants’ fNIRS data, presented in Chapter 6.2.
    - Javier Andreu-Perez, Lauren L. Emberson, Mehrin Kiani, et al. (2021), "Explainable Artificial Intelligence Based Analysis for Developmental Cognitive Neuroscience". In: *Commun Biol* 4, p.1077
    - Link: https://www.nature.com/articles/s42003-021-02534-y
4. Development of new General Type-2 based Time-dependent XAI (TXAI) System
    - Temporal Type-2 Fuzzy Sets (TT2FS).
    - Applied on a real-life temporal dataset.

- Mehrin Kiani, Javier Andreu-Perez, Hani Hagras, "A Temporal Type-2 Fuzzy System Based Approach for Time-dependent Explainable Artificial Intelligence." In: *IEEE Transactions on Artificial Intelligence*, Under Review.

In the next section, I outline the structure of my thesis.

1.5 Structure of the thesis

The structure of my PhD thesis is as follows:

- In Chapter 2, I expand on the different frameworks of DCN, neuroimaging modalities, and in particular fNIRS and its data analysis approaches in the context of DCN studies.
- In Chapter 3, I provide some preliminary background that illustrates the Fuzzy Logic Systems (FLS) based XAI methods proposed for DCN studies.
- In Chapter 4 a literature review of state-of-the-art AI methods applied to brain data (infants and/or adults) is presented.
- In Chapter 5 the three new XAI methods developed in the work are presented in detail.
- In Chapter 6 the applications of the proposed XAI methods are presented with a discussion of their results, and implications for functional brain development.
- In Chapter 7 a conclusion of the proposed XAI models as well as future works is presented.

Chapter Two

Background on Neuroscience

In this chapter[1], I provide an overview of the different frameworks in Developmental Cognitive Neuroscience (DCN) for the analysis of functional brain development. As will become evident in the current chapter, the most comprehensive account for functional brain development is provided by the DCN frameworks of Interactive Specialisation (IS) and neural reuse; hence, the development of the eXplainable Artificial Intelligence (XAI) models (in Chapter 5), in this work, also focus on delineating their inference mechanisms in terms of IS and neural reuse.

In addition to DCN frameworks, this chapter also provides an overview on functional Near-InfraRed Spectroscopy (fNIRS). fNIRS is a neuroimaging modality that can be readily applied for recording infants' brain activity. This chapter also outlines the fNIRS data analysis techniques as applied to DCN studies.

[1] Some parts of the text in this chapter have been published here: M. Kiani, J. Andreu-Perez, H. Hagras, S. Rigato, and M. L. Filippetti, "Towards Understanding Human Functional Brain Development With Explainable Artificial Intelligence: Challenges and Perspectives," *IEEE Computational Intelligence Magazine*, 2021, In Press © 2022 IEEE.

2.1 Developmental Cognitive Neuroscience (DCN)

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 [23]. 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 [24].

Broadly speaking, historically the theoretical approaches towards understanding human brain development have contrasted the nature versus nurture approach [25]. On the one hand, the nature approach claims that there exist a set of innate characteristics (such as genetics) and non-learned knowledge (also termed core knowledge for representing objects, numbers, actions, and space [26]) which are universally invariant across development [27, 28]. In this sense, the nature approach proposes that the brain development of an infant unfolds as a result of their genetic or biological makeup and mostly independent of their experiences.

On the other hand, the nurture approach claims that brain development transpires mainly because of acquired knowledge based on the interaction of the infant with its environment [29]. In this sense, most of the information required for brain development is experience-dependent with some elements from the environment (such as gravity) being common for all infants. An infant's interaction with the environment is deemed multi-faceted, spanning from an infant's early childhood experiences, its social relationships, and cultural influences.

Overall, nature-nurture interaction aims to identify the relative contributions of ones' core knowledge as well as innate characteristic and those attributed from the environment in shaping the human cognitive abilities. However, with the advent of more research in the developing brain, there is now significant evidence [3] suggesting that it is the interplay between an individual's inherited genetics, core knowledge and their environment that shapes infants' development.

The interplay between nature and nurture for cognitive development is referred to as the 'constructivist' approach [30]. Both the nature and nurture approaches share the assumption that the information necessary for development to unfold already exists prior and independently from the ontogenic trajectory of the individual. Constructivist stands in an intermediate position between these two approaches, by suggesting a non-deterministic, dynamic process involving both the innate characteristics and the external factors for explaining the perceptual and cognitive development [31]. Compared to previous approaches to development, constructivism places a significant focus on the mechanisms of change and on the interactive processes (dynamic interaction between brain and behavior) that lead to the emergence of optimised brain structures (that mediate social and cognitive functions). In this regard, DCN studies have increasingly investigated the developing brain by taking the constructivist approach [3, 32].

The three main DCN frameworks that have been proposed to explain human brain development, namely 1) Maturational perspective, 2) Skill learning, and 3) Interactive Specialisation (IS) aim to answer the question of how specialised neural networks emerge during postnatal development. The rest of the chapter expands on the aforementioned DCN frameworks.

2.1.1 Maturational Perspective

The maturational perspective on human brain development undertakes a simplistic approach by suggesting that brain development occurs as a consequence of brain maturation. Broadly speaking, maturational perspective is a unidirectional approach by relating maturation of anatomical regions of brain with their respective brain functions such as perception, language, motor etc. In this sense, the maturational approach assumes a deterministic pathway from structural development of a brain region to its corresponding brain function [33]. Hence, the maturational perspective suggests a one-to-one mapping between brain structure and brain function.

Most of the research in DCN, until early 2000, has been conducted from a maturational perspective. The maturational perspective deems brain development as a maturational process. The success of this perspective comes from the identification of brain regions responsible for processing/carrying out a certain task such as, for example, the maturation of the DorsoLateral Prefrontal Cortex (DLPFC) being linked with the successful performance in object retrieval task [34]. More specifically, the A-not-B task [35] is used to investigate object retrieval in infants. In this task, the object is hidden in one of two locations (that differ with respect to their right or left location only) as the young participant watches. The infant is then expected to hold the information in mind (working memory), and successfully retrieve the object from the correct location, i.e. either left or right. By 7 to 12 months, infants increasingly appear to perform the object retrieval task successfully [36].The conclusion that the DLPFC is involved in the successful retrieval of an object is based on a study with adult non-human primates where a damage to the DLPFC impaired the participants from successful object retrieval [37]. Further, in the work by [38], a higher activation is observed in DLPFC during object retrieval task using an optical neuroimaging modality (functional Near-Infrared Spectroscopy (fNIRS)). Indeed, the mapping of DLPFC and A-not-B response task is one of the most established example of brain-behaviour relations studied in DCN [39].

Overall, the aforementioned studies support the maturational perspective, i.e., once a brain region (such as DLPFC) is structurally mature, infants are able to perform the corresponding brain function successfully (in the case of DLPFC the object retrieval task). However, there are also studies that propose that experience or learning impact on the brain functions, such as the training of a working memory task in adults by [40] which can not be explained using the maturational perspective alone.

More specifically, in the work by Olesen et al. [40], the researchers have shown that adults' working memory demonstrate a higher capacity *after* undergoing a five week training. Since the participants are adults, with mature brains, [40] conclusions can not be accounted for using the maturational perspective alone. To summarise, the ability of the human brain's structure and function to change beyond its formative years, as a result of learning, (also termed as the brain plasticity [41]) can not be explained using the maturational perspective alone.

With regards to brain plasticity, the maturational perspective suggests that this is an innate ability of the human brain to recover, however brain plasticity is only brought into action when the brain suffers an injury or stroke [31]. Other than in cases of lesion, brain development is perceived to be a static process with different cognitive functions emerging at different time points, as a consequence of brain structures developing according to their pre-determined maturational timetables. To summarise, beyond the identification of primary functions of various brain regions, the maturational perspective is not able to explain all aspects of the functional brain development such as the optimisation of inter-connectivity between brain regions post learning.

2.1.2 Skill Learning

The skill learning perspective supports a life-long learning process as a way to explain functional brain development. As such, for the skill learning perspective, human brain development has no specific earmarked developmental phases, i.e. the human brain is in a continuous developmental phase throughout the life span of an individual. The skill learning approach suggests that the cortical regions active in infants during the onset of perceptual or cognitive abilities are similar to those that are involved in complex skill acquisition in mature/adult brains. These specific cortical regions are referred to as the 'skill learning circuitry' that is activated during the acquisition of a skill either in infants or adults. An illustration of a skill learning circuitry for a sensorimotor task is shown in Fig.

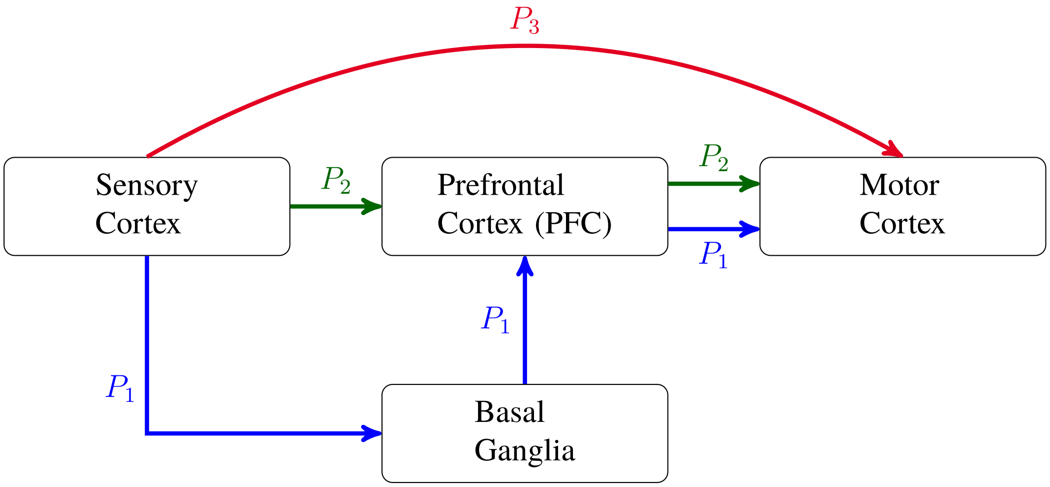


Figure 2.1 An illustration of a skill learning circuitry invoked for sensorimotor learning [42] with P1, P2, and P3 referring to possible pathways.

2.1 as suggested by [42]. The cortical regions involved in sensorimotor tasks are the sensory cortex, Basal Ganglia, Prefrontal Cortex (PFC), and Motor Cortex.

A notable DCN study that supports the skill learning framework is the work by Gauthier et al. [43]. They have shown that the fusiform gyri area of the brain is activated when adults are trained with *greebles*- objects which have spatial characteristics like the human face but haven't been seen earlier by the participants, i.e. they are novel, artificial objects. After an extensive training with the *greebles*, the same area of the brain is activated as those in infants after attaining face processing skills, i.e. the fusiform gyri. Although the work by [43] corroborates the skill learning perspective for an adult brain, it is unable to account for the dependence on structural maturation of the cortical areas involved that precedes the acquisition of skills in infants such as successful object retrieval only from 7 month onwards [36].

In addition, the notion of plasticity in the skill learning approach is deemed as a result of the continuous learning of the human brain. Hence, plasticity is a lifelong attribute with no particular age deemed more special for brain development. Referring back to Fig. 2.1 a total of three paths are shown depending on the skill acquisition level achieved for a generic sensorimotor task. The blue path, labelled P_1, represents the initial pathway that is involved for carrying out a sensorimotor task investigated in [42] and involves the sensory cortex, basal ganglia, PFC and motor cortex. The second pathway, colored green and labelled P_2, emerges when basal ganglia output to PFC has strengthened the PFC's sensory input synapses, hence the optimised sensorimotor circuit now bypasses the basal ganglia. Upon further optimisation, a new pathway P_3 emerges directly linking the sensory cortex with the motor cortex. As can be seen from the Fig. 2.1, plasticity or change in the connectivity of the human brain can be explained using the skill learning approach.

Although the skill learning approach is intuitive since it offers a dynamic mechanism of learning, the fundamentals of dynamic mechanisms or the dependency of a skill acquisition on the structural maturation of a skill learning circuitry are not elucidated. In contrast to the skill learning approach, the IS approach, presented in the next section, provides a more holistic view on functional brain development.

2.1.3 Interactive Specialisation (IS) Theory

The IS framework suggests a more probabilistic account based on a bidirectional relationship between structural and functional brain development. In particular, it assumes dynamic changes during brain development. The IS framework proposes that both feed-forward and feedback connections between different cortical regions affect the functional specialisation of cortical regions [44]. 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.

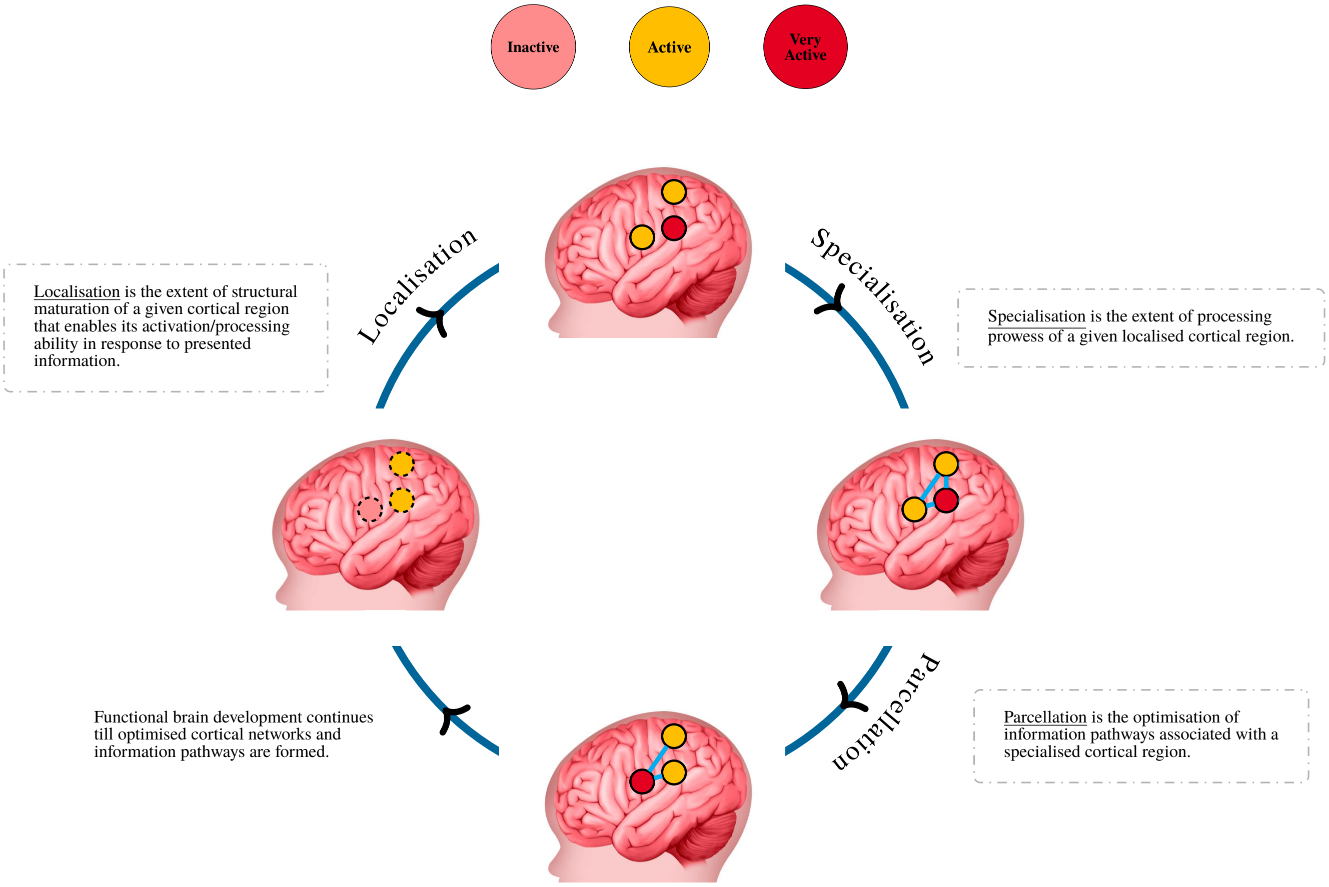

Inactive
Active
Very Active
Localisation
Specialisation
Parcellation
Localisation is the extent of structural maturation of a given cortical region that enables its activation/processing ability in response to presented information.
Specialisation is the extent of processing prowess of a given localised 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.

Figure 2.2 An illustration of the Interactive Specialisation (IS) theory, © 2022 IEEE.

(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. 2.2. 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 of different parts of the brain, along with more long range connections

Framework	Principle	Brain structure-function mapping
Maturational	Structural development	A static one-to-one mapping.
Skill Learning	Skill acquisition	A dynamic mapping that changes in response to acquisition of skill.
Interactive Specialisation	Interlinked structural and functional development	Dynamic cortical networks with inter- and intra-cortical connections.

Table 2.1 A summary of three frameworks in developmental cognitive neuroscience (DCN) [44].

coming ‘on line’. As a result of the parcellation process, not all parts of a given cortical region need to be activated nor all connections may be 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.

A summary of the three DCN frameworks is provided in Table 2.1. Clearly, the DCN frameworks are not mutually exclusive but rather focus on different aspects of brain development. Please note, for the purpose of this work, I will focus on the IS perspective, which is largely supported by DCN studies on the basis of being the most comprehensive account [33]. In addition to the IS framework, I will also be referring to the *Neural Reuse* phenomenon (explained next in Section 2.1.4) for understanding how brain structures get re-wired for the pursuit of a highly specialised, hierarchical adult brain.

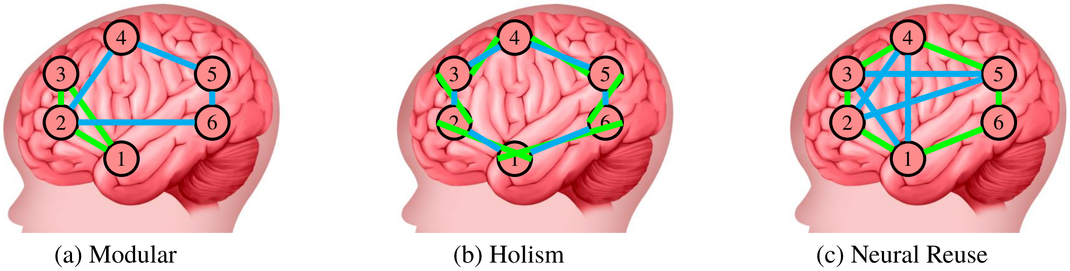
2.1.4 The Hierarchical Human Brain

The developed adult human brain, both in terms of structure and function, is a ‘small world’ network [45]. 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 to become increasingly proficient and fine-tuned to process that specific information [46]. 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 [46], as illustrated in Fig. 2.2.

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 PFC) where the information processed by lower hierarchical levels is integrated [47, 48]. For top-down modulation, the higher-order cognitive influences (such as from PFC) interact with information coming from primary area of visual processing (such as occipital cortex). The cognitive influences may change the information conveyed by occipital cortex such as attention, object expectation, and scene segmentation [49].

An important consideration with regards 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 [20]. As a consequence, existing modules could subserve emerging cognitive states by a reconfiguration of their interconnections using a *neural reuse* [50] process of brain organisation. The other two plausible processes put forward to explain functional brain organisation are *modularity* and *holism* [50].

With regards to explaining the hierarchical structure of the human brain, 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 neural reuse perspective suggests that 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



(a) Modular (b) Holism (c) Neural Reuse

Figure 2.3 The perspectives for the hierarchical structure of the brain, © 2022 IEEE.

illustrated in Fig. 2.3.

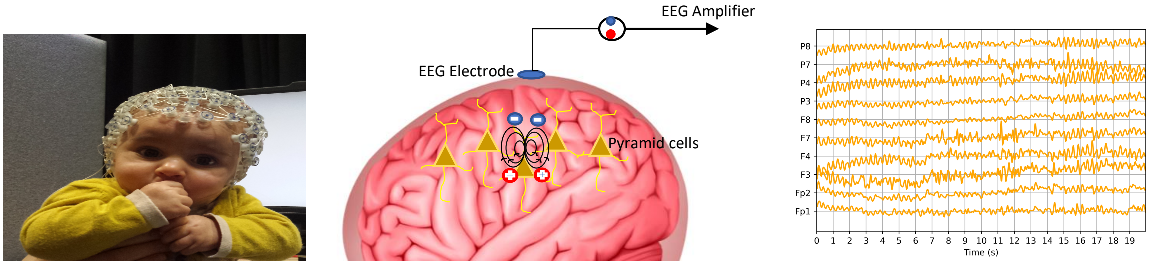
The idea of neural reuse seems plausible with respect to optimal usage of existent circuits evolved for a given cognitive task. In this way, while neural circuits are modular to some extent with respect to their individual functionality, neural reuse suggests that neural circuits can continue to acquire new uses after an primary function is established [50]. The neural reuse perspective is supported by a range of higher cognition functions such as the reuse of motor control circuits for language by Pulvermuller et al. [51]. In the study by Pulvermuller et al. [51], the authors reported that listening to words which involve an action such as ‘pick’ and ‘kick’ activates primary motor cortex. The finding that listening to or comprehension of verbs also activates motor cortex entails the neural reuse of motor cortex for language comprehension. Another example of the neural reuse of motor cortex for memory retrieval is by Casasantro et al. [52]. In their work, the authors concluded that participants were able to retrieve more (positive or negative) memories if their movement (moving marbles up or down from one container to another) was matched with the memory valence (i.e. positive memory with moving marbles up and negative memory with moving marbles downward from one container to another). Their study suggests that activation (neural reuse) of motor cortex helps with the memory retrieval of the participants.

In the next section, I outline the fNIRS neuroimaging modality for its application in DCN studies.

2.2 Functional Near Infra-Red Spectroscopy (fNIRS)

The aim of this work is to develop XAI (explainable artificial intelligence) methods that can describe functional brain development in accordance with the IS (interactive specialisation) and neural reuse framework (previously outlined in Chapter 2.1). Before I present the proposed XAI methods, in this section, I outline the different neuroimaging modalities that are common for recording infants' brain activity. For DCN studies in particular, in order to be able to investigate multifaceted aspects of cognition in development, the experimental paradigms require that infants' brain activity can be recorded during the execution of a wide range of cognitive tasks. Owing to the requirements to study cognition in infants, some key features should be taken into account when designing experimental studies with this population. While it is essential for the designated neuroimaging modality to be safe and non-invasive, the methodology should also have good temporal resolution and spatial localisation, and allow for infants to be sitting upright (to be able to watch the stimulus/task) with the freedom to perform bodily movements. Keeping in mind the aforementioned constraints, the following review of neuroimaging modalities aims to assess their applicability for DCN studies.

In order to examine the neural underpinnings of cognitive processes and their changes across development, functional Near-Infrared Spectroscopy (fNIRS) [53, 6], and Electroencephalogram (EEG) [54] have been widely used in DCN studies with infants and children. These neuroimaging modalities are both safe, non-invasive, portable, wearable, and relatively inexpensive - compared to Magnetic Resonance Imaging (MRI), 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 even in ecologically valid settings such as their homes if needed [6]. Nevertheless, fMRI has been successfully used in developmental studies with asleep infants [55, 56] and, more recently, also with awake infants [57, 58].



(a) EEG net. **(b)** EEG principle. **(c)** 10 EEG signals.

Figure 2.4 The Electroencephalogram (EEG) neuroimaging modality for DCN studies, © 2022 IEEE.

EEG measures brain electrical activity, with the electrodes placed on the scalp, that reflect the summated postsynaptic potentials of cortical neurons (also known as the ‘EEG generators’ [59]) in response to changing cognitive or perceptual states [60]. EEG activity is mainly generated by pyramidal neurons in the cerebral cortex that are perpendicular to the brain’s surface/electrode on the scalp [61]. The EEG principle is represented in Fig. 2.4b. EEG records electrical changes continuously on the scalp allowing the measurement of rapid cognitive processes [62] with high temporal accuracy, in the order of milliseconds [63]. The EEG net has electrodes fitted on it that covers the whole head (see Fig. 2.4a). Since EEG can record activity on the time scale of underlying neuronal activity, EEG signals (see Fig. 2.4c) are better suited than fNIRS signals for connectivity analysis. The connectivity analysis using EEG signals has proved pivotal in understanding the neural correlates of cognitive functions in typical infants [64] as well as those with varying underlying conditions such as low weight birth [65], those born preterm [66], and also those at risk for autism spectrum disorder [67].

In contrast to the continuous EEG signals, the Event Related Potential (ERPs) are derived from the EEG signals by averaging the time-locked segments to the presentation of an event or stimulus [68]. For example, in adults during a face processing task, a P100 component (a positive ERP) is observed at 100ms and a N170 component (a negative ERP) is observed at 170ms over occipitotemporal electrodes post stimulus presentation [69]. Likewise, in 3- to 12-month old infants, EEG based studies investigating the development of face processing have observed two ERP components over the occipitotemporal electrodes thought to be the precursors of the adult N170, i.e. the N290 and the P400 components [70]. The difference in the morphology of the ERPs (latency and amplitude) between adults and infants is expected, and arises because of the physiological differences between infants and adults (such as the head, skull, and brain tissue [71]), as well as the developmental differences between the two populations (such as the difference in local functional structure of cerebral cortex [72]).

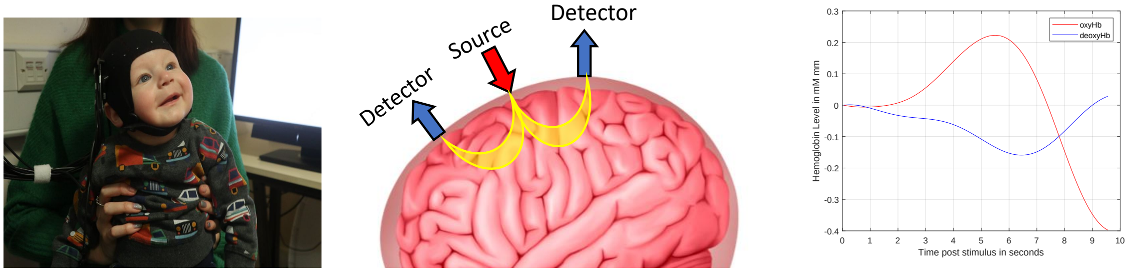
The study of cognition in infants with ERP based investigations has enabled a time based analysis that allows the identification of the sequence of cognitive processes based on the time information embedded in the ERP waveform [70]. Further, based on the difference in the morphology of the ERPs, it is also possible to investigate the difference in response to different stimuli. For example in the processing of facial expressions, the P400 amplitude was larger for fearful vs neutral or happy faces at semi-medial electrodes in infants [70]. Likewise, ERPs have been used to investigate other areas in cognition such as auditory [73] and language learning impairment [74] as well as attention [75].

The challenges associated with infant ERP-based studies include significant variation in the elicited ERP both within and between infants. In addition, the number of trials per infant are usually smaller in comparison to adults (due, for example, to higher motion artifacts [76]) and consequently the averaging of the ERP does not have as much statistical power as those of adult ERP studies.

The critical limitation of EEG for investigating functional brain development as hypothesised by the DCN frameworks (Chapter 2.1) is its limited spatial resolution [77], which makes it difficult to map brain electric activity to its corresponding anatomical regions in the brain. Although a few EEG based studies have attempted source localisation in adults (for example, Brain Electrical Source Analysis [78]) and infants (for example, [79]); the proposed source localisation algorithms are dependent on source generator assumption upon which dipole locations are determined and subsequently potential distribution is built [71]. Since EEG can not identify the source (corresponding brain region) of brain activity, without significant assumptions, I will be focusing on fNIRS modality for investigating functional brain development.

fNIRS is an optical neuroimaging modality that uses Near-Infrared (NIR) light on the scalp to record relative changes in blood haemoglobin (Hb) concentration, based on NIR light absorption by the Hb molecules, which is inferred as a measure of the cortical brain activity [53]. The NIR light absorption is minimum as it travels through the brain tissue except for the Hb chromophore. Hence, NIR light can detect changes in concentration of Hb based on the absorption of the NIR light. Though, the NIR absorption spectra of the oxygenated Hb (oxy-Hb) and deoxygenated Hb (deoxy-Hb) are different. A greater absorption of the NIR light means larger presence of Hb which is reflective of greater brain activity. This is because the part of the brain involved in the processing/execution of a task experiences a greater metabolic demand for oxygen and glucose. The increase in metabolic demand is met by an increased cerebral blood flow to that region of the brain leading to an increase in oxy-Hb and a decrease in deoxy-Hb.

An illustration of the fNIRS principle (Fig. 2.5b) along with a representative signal (oxy-Hb in red, and deoxy-Hb in blue) is shown in Fig. 2.5c. The fNIRS cap comprises of pairs of sources and detectors, and the cortical area between a source and detector is deemed as a fNIRS channel, see for example Fig. 2.5a). The source and detector probes on the fNIRS cap can be placed in various configurations such as the 10-20 or 10-10 standard montages as well as a fully customised montage where the researchers place source and detector probes against cortical areas of investigation in their study. More recently, a wearable fNIRS cap with high density, fixed distance probes for infants has



(a) fNIRS cap. **(b)** fNIRS principle. **(c)** A fNIRS signal.

Figure 2.5 The functional Near-InfraRed Spectrocopy (fNIRS) neuroimaging modality for DCN, © 2022 IEEE.

been developed to further improve the spatial localisation of fNIRS [80]. The fNIRS cap can also be customised to add EEG electrodes for a simultaneous recording (fNIRS and EEG) of the underlying brain activity [81].

Over the past 50 years, fNIRS has been used to study a range of populations whilst investigating a multitude of brain-behaviour tasks. The technological advances in fNIRS coupled with more sophisticated tools for its signal analysis have seen a rapid growth in studies investigating neural correlates using fNIRS. In particular, the portability of fNIRS to study participants both inside and outside the lab environment, as well as the freedom to perform motor tasks makes it an ideal modality for studying brain activity in naturalistic environments. Apart from portability, and low sensitivity to motion, fNIRS also offers good spatial localisation (within 2cm [82]) which allows for conclusions to be drawn about the localised cortical activity from different anatomical locations of the cortical structures. The aforementioned attributes are quintessential for gaining insight into the developing brain as these allow for greater flexibility in the experimental paradigm. As a result, more sophisticated questions regarding brain-behaviour relation can be answered.

Notwithstanding the strengths of fNIRS technology for the study of human brain development, there are several limitations that should be taken into account when employing fNIRS in cognitive neuroscience research. First, brain activity using fNIRS can only be recorded as a relative measure compared to a baseline measurement (usually the first reading of the fNIRS signal), and not absolute measurements [83]. The cause for the calculation of relative measurements of brain activity using fNIRS are further discussed in Section 2.2.1. Second and unlike functional MRI (fMRI), the fNIRS measurements do not include an anatomical image to which underlying brain regions can be referenced. This means that fNIRS channels will need to be anatomically registered to corresponding brain areas [84]. This is an important consideration when undertaking a neuroimaging study with fNIRS, since any conclusions about the localised brain activity are hinged on the identification of the corresponding anatomical locations of the fNIRS channels. The fNIRS channel registration is discussed in more detail in Section 2.2.2. Another important consideration with fNIRS modality is that it has limited penetration depth (only reads the brain activity as far as the superficial cortex [85]) and hence fNIRS based studies can not shed light on brain activity beyond the superficial cortex. Further, because there is a substantial time delay between the activation of the brain region and the observed haemodynamic response, fNIRS is not well-suited for real-time investigations of brain activity. Although these are all significant limitations, however, they do not impact the suitability of fNIRS in DCN research as such because relative measurements of brain activity, even after a significant lag, are still warranted to shed light on the activated cortical regions.

The continuous advancements in fNIRS are further enabling the technology to be at the forefront for answering questions in an ever expanding field of neuroscience. Arguably, the most exciting areas of recent advancement is fNIRS hyper-scanning [86] where the brain signals are recorded from more than one person simultaneously. In this sense, fNIRS hyperscanning allows for brain activity to be recorded whilst allowing for natural social interaction between the participants and investigations usually aim to analyse inter-brain synchrony using wavelet coherence or Granger Causality (GC). Further, with wireless fNIRS [87] now it is also possible to investigate brain activity of participants with minimal constraints (no wires) which is particularly advantageous during neurorehabilitative training. Another recent advancement in neuroscience research entailing fNIRS, is the combined use of neuroimaging techniques such as fNIRS and EEG (e.g. [88]). The multimodal imaging can provide a wider picture of functional brain activity by benefiting from the advantages of different neural measures. These are all promising avenues in neuroscience research with fNIRS which have the potential to further inform brain-behaviour relations.

A flowchart of a typical neuroscience experiment with fNIRS is illustrated in Fig. 2.6. In the next sections, an outline of the fNIRS typical pre-processing stages and the standard procedure for fNIRS channels' co-registration is presented. The prevalent analysis of fNIRS signals are outlined in section 2.2.4.

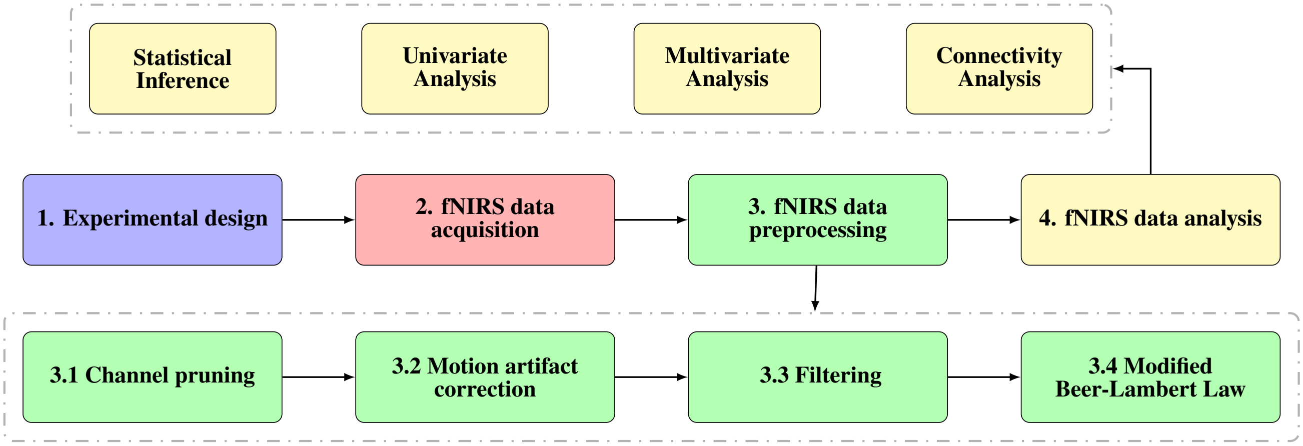


Figure 2.6 A schematic of typical neuroscience study with fNIRS.

2.2.1 Preprocessing

The noise sources in fNIRS signals are usually characterised as measurement/system (such as electronic noise or light source instability), physiological (associated with heart rate/breathing) or motion artifacts (head or body movements). The aforementioned sources of noise significantly affect the fNIRS signal quality, and therefore it is important to preprocess such signals to ensure subsequent inferences made are based on neural correlates and not noise. The standard preprocessing steps for fNIRS signals, based on the recent work of [89] and [84] are listed below.

1. Channel pruning
2. Motion artifact correction
3. Filtering
4. Modified Beer-Lambert law

Although the preprocessing steps are similar across studies, there is no established standard with respect to performing these steps [89]. For example, there is little consensus on the criterion for channel pruning such as the signal-to-noise ratio (SNR) threshold or the attrition rate in infant studies [90]. Likewise, various types of filters (low pass, high pass, band pass) have been used in fNIRS based studies with varying filter orders, and cut-off frequencies [89]. Moreover, there is also no gold-standard for the order in which the above listed steps are carried out [84]. A lack of standardisation of the above steps makes it difficult to not only draw comparison between studies but also makes it hard to replicate the results. In this regard, there is a growing need for the standardisation of fNIRS preprocessing pipelines [84] to ensure robust results and replications. Recently, different viable preprocessing pipelines for infant fNIRS studies have been proposed by [8].

Apart from the variation in the pre-processing steps, there is also no standard toolbox/program for performing such analyses. Some notable toolboxs include: 1) Statistical Parametric Mapping (SPM) for fNIRS [91], 2) HomER [92], and a more recent one: 3) NIRS-KIT [93]. They are all based in Matlab and offer a graphical user interface to perform the fNIRS data pre-processing steps. More recently, however, there has been a greater uptake of the HomER program (for example [6, 89, 84]) since it allows for custom processing scripts at subject and group levels as well as the continuous advancement in their toolbox; they have recently released a 3rd version of their toolbox [94].

In spite of the choice of the pre-processing toolbox, the first step in the preprocessing of fNIRS data usually involves the channel pruning stage. In channel pruning stage the quality of the signal from fNIRS channels is checked against a set SNR threshold. The fNIRS channels that pass the SNR threshold are then selected for subsequent analysis. In this regard, it is important to note that not the same fNIRS channels might be selected across all subjects.

The next preprocessing step involves removing or correcting for the motion artifacts. The motion artifacts, caused by head or limb movements, result in spikes or baseline shifts in the signals. Motion artifacts are particularly predominant in DCN studies because infants tend to move often during experiments, and the experimenters have little control over infants' behaviour or mood. In addition, to account for the prevalent physiological noise in the fNIRS signals, filtering of fNIRS signals is also undertaken. By filtering the fNIRS signals, frequency bands which are associated with the physiological noise are removed. The particular components removed are typically heart rate (~1Hz), Mayer waves (~0.1Hz), and breathing rate (~0.3Hz).

After removing noise and motion artifacts, the signals (from selected fNIRS channels) are converted to changes in Hb concentration. The fNIRS estimates the attenuation (ϖ), or loss, in the NIR light shone from the source by comparing the input NIR light, I_{IN}, with the detected NIR light, I_{OUT}, using the modified Beer-Lambert Law (mBBL) (2.1a) [95]. The first attenuation measure at time (t) =0, i.e. $\varpi(at\ t = 0)$, is subtracted from all subsequent attenuation measures to give differential attenuation measures, denoted as $\Delta\varpi(t)$ in (2.1b). The use of differential attenuation measures helps to minimise the NIR loss from sources other than the Hb absorption such as the NIR light scattering and absorption in the brain tissue. The concentration changes of oxy-Hb and deoxy-Hb, $\Delta\zeta$, are then derived from differential attenuation measures as a function of the source-detector distance (d) and the Differential Pathlength Factor (DPF) in (2.1c). The extinction coefficient of the chromophore at a certain wavelength λ is denoted by ϵ in (2.1c). Using a dual wavelength system, measurements for $\Delta\zeta_{oxy-Hb}$ and $\Delta\zeta_{deoxy-Hb}$ can be solved using matrix notation as outlined in (2.1c).

$$\varpi(t) = -log_{10}\frac{I_{IN}(t)}{I_{OUT}(t)} \tag{2.1a}$$

$$\Delta\varpi(t) = \varpi(t) - \varpi(t = 0) \tag{2.1b}$$

$$\Delta\zeta = \frac{\Delta\varpi(t)}{\epsilon(\lambda) * d * DPF(\lambda)} \tag{2.1c}$$

$$\begin{bmatrix} \Delta\varpi_{\lambda_1}(t) \\ \Delta\varpi_{\lambda_2}(t) \end{bmatrix} = \begin{bmatrix} \epsilon_{\lambda_1}^{oxy-Hb} & \epsilon_{\lambda_1}^{deoxy-Hb} \\ \epsilon_{\lambda_2}^{oxy-Hb} & \epsilon_{\lambda_2}^{deoxy-Hb} \end{bmatrix} \begin{bmatrix} \Delta\zeta^{oxy-Hb} \\ \Delta\zeta^{deoxy-Hb} \end{bmatrix} \tag{2.1d}$$

2.2.2 Channels' Registration

As noted earlier in Section 2.2, fNIRS has good spatial localisation, compared to EEG, however unlike fMRI, fNIRS signal measurements do not include corresponding brain anatomical map. Hence, fNIRS channels need to be registered to correctly identify corresponding anatomical brain regions from which brain activity is recorded. Channel registration is an important part of an fNIRS study because it directly affects the subsequent inference about the brain activity of cortical regions under investigation.

The most prevalent approach for fNIRS channels' registration is to use a probabilistic registration method that allows for using a database of reference MRI images in place of individual's/participant's own MRI image [96]. The probabilistic registration method maps the location of fNIRS channels to a standard brain template. The most commonly used brain template for adults is the Montreal

Neurological Institute (MNI) template which was created from 152 brains coregistered to the Talairach brain [97]. The probabilistic registration method improves the fNIRS channels location specificity when individual MRI images of participants is not possible. However, when access to MRI is possible and MRI image of the participant can be obtained, individual MRI scans are a preferred way to do spatial registration of fNIRS channels [98].

In particular for DCN studies, given that infants' head sizes can vary significantly within and across ages, channel registration becomes critical. Some examples of magnetic resonance (MR) coregistration with fNIRS channels in infant studies are by [99, 100, 101].

2.2.3 Oxy-Hb or Deoxy-Hb

As previously discussed in Section 2.2, the expected haemodynamic response reflective of brain activity is a simultaneous increase in oxy-Hb concentration and a decrease in deoxy-Hb concentration [89], also referred to as the canonical haemodynamic response (see Fig. 2.5c). However, scientific works in DCN have pointed to inconsistency in deoxy-Hb response in human infants [6] and as such the canonical haemodynamic response is rarely observed/reported in DCN studies [8]. Therefore, most DCN studies with fNIRS report data analysis with only oxy-Hb chromophore, for example [101]. This is because, in comparison to deoxy-Hb, the oxy-Hb chromophore has been reported to pertain more information than the deoxy-Hb about the underlying neural correlates [102]. However, there is now growing consensus to report both oxy-Hb and deoxy-Hb signals analysis as a step towards standardisation of fNIRS studies [84] even if no significant results are obtained from any/either of the Hb biomarkers. Some examples of DCN studies that report data analysis from both oxy-Hb and deoxy-Hb are [103, 104, 105].

In the next section, an overview of the fNIRS data analysis is presented.

2.2.4 fNIRS Data Analysis

The prevalent data analysis of fNIRS signals, post pre-processing and conversion to ΔHb concentrations, involves estimation of the HRF by 1) simple block averaging, 2) Convolution, or 3) General linear models (GLMs). The block averaging method is independent of any *a priori* assumptions about the shape of the HRF and is the de-facto choice for studies when the HRF response for a novel experimental paradigm or population is not yet established such as DCN studies [84], further discussed in Section 2.2.4. Block averaging typically involves selecting a time window of the fNIRS signals, post stimulus presentation such as 4 - 7 seconds [101], and taking the average of that time window. A range of statistical tests can then be applied on the average activation for different stimuli or events.

An important caveat with respect to block averaging is the identification of the time window also known as activation hacking. In activation hacking, the start and stop time points post stimulus presentation of the fNIRS signal oxy-Hb or deoxy-Hb) are identified that would result in best classification results. As such, there is no set procedure for activation hacking nor is there any consensus on the parameters of the time window (start and stop time points post stimulus presentation) i.e. different DCN studies have reported different time windows. For example, the work by [106] with 7-month-old infants used 6s post stimulus until 2s post offset, in six-month-old infants [101] used 4-7 s time window, whereas [107] used 8 to 10 s.

In contrast to activation hacking for block averaging, some studies make use of the whole signal by dividing the full length of the fNIRS signals into blocks, and doing statistical analysis on the averages of the blocks. For example, in the study by Pfeifer and et al. [108] the fNIRS signals is segmented into 10 blocks and the averages of these 10 blocks are used for statistical analysis (Friedman test was used in [108]).

Although block averaging is a powerful method, independent of any underlying assumptions, it takes away the temporal information embedded in the fNIRS signal. In this regard, the GLM method builds upon the temporal information stored within the fNIRS signal, and is outlined next.

General Linear Model (GLM)

The GLM method seeks to find the influence of each of the explanatory variables (often referred to as regressors in fNIRS) of the fNIRS signal using their (regressors') weighted linear combination. The weight attached to each of the regressors is an indication of the contribution of a given regressor for a fNIRS signal. The GLM can be mathematically expressed as (2.2).

$$\kappa(t) = \beta\xi(t) + \Omega \tag{2.2}$$

The β values are the weights that quantify the contribution of each regressor (denoted by ξ) to the estimated HRF (denoted by $\kappa(t)$). The noise or unexplained contributions to the fNIRS signal are referred collectively as the error, denoted by Ω. The GLM estimates the value of the β for each regressor by comparing the estimated haemodynamic response function (HRF) ($\kappa(t)$) with a predefined HRF. In this way, by reducing the difference between the estimated and the predefined HRF the β values are optimised.

The strength of the GLM method arises from the incorporation of the full time series of the fNIRS signals. This evidently gives GLM higher statistical power than block averaging method. In addition, in the GLM method it is easy to quantify the contribution of each of the predictors (such as fNIRS channel, noise etc.) using regressor (β) values. However, the GLM presumes that a specific, predefined HRF is known for the task and the underlying cortical region being investigated [89]. This is rarely the case for DCN studies i.e. no established HRF is known for infants [8]. As a consequence, such lack of known HRF for infants limits the applicability of GLM in DCN studies (As a consequence, such lack of known HRF for infants limits the applicability of GLM in DCN studies (though see [109] for an example of HRF estimation in infants, and [110] for an example of HRF estimation in adults .

Haemodynamic Response Function (HRF) in infants

Although some notable works for the identification of HRF in full-term infants have been carried out (such as [109]) there are many caveats associated with pre-determination of HRF in infants. For example, the HRF is dependent on the task/experimental paradigm as well as the cortical area under investigation i.e. the HRF for visual stimulus may not necessarily be the same as that for auditory stimulus. Likewise, the HRF in occipital cortex would be different from that in temporal cortex. The HRF would also be dependent on how much an infant engages with the stimulus. Owing to the lack of an established HRF in infants, most DCN studies do not use HRF based analysis but rather block-averaging of oxy-Hb signal for the most pronounced time window is often used (for example [101], [111]).

In the next sections, an overview of fNIRS data analysis based on block-averaging is presented using univariate and multivariate methods.

Univariate Analysis

The classical statistical inference models for fNIRS data analysis are predominantly univariate, i.e. they investigate one variable's or cortical region's data at a time (for example [112, 111]). These methods have focused on *where* in the brain the activity is more globally pronounced in response to a certain stimulus. This is possibly due to the relatively limited datasets that can be experimentally collected, and the need of cognitive neuroscientists to decode and interpret the complex multivariate patterns of neuroimaging data using straightforward approaches. This limitation is amplified in DCN, where data collection poses significant additional challenges, such as dealing with infant participants' compliance with the experiment, and sample sizes are, as a consequence, relatively smaller compared to neuroimaging studies with adults.

However, the univariate analysis tends to implement the maturation perspective of the DCN frameworks, i.e. mapping one brain region to one brain function. As outlined earlier in Chapter 2.1, the DCN frameworks of IS and neural reuse support the interaction of various brain regions for carrying out various cognitive, perceptual, or motor tasks. In this regard, a method that can analyse multiple brain regions' activity at the same time (termed multivariate analysis) is required to elucidate the cortical networks thus formed in response to presented stimuli.

In the next section, multivariate analysis is presented.

Multivariate Analysis

As discussed in the last section, the univariate methods only partially responds to the main research question for DCN studies of how emerging patterns of interaction between brain regions are associated to new cognitive functions [3]. However, by using M number of dimensions, arising from M number of fNIRS channels, Multivariate Pattern Analysis (MVPA) methods have the potential to identify associations between brain regions, and the corresponding activation levels in terms of distributed patterns, rather than just as measurements of a single source. MVPA is a classification method that aims to differentiate between classes (or stimuli) by finding patterns in the multivariate matrix that are prototypical of the classes. The MVPA has two integral components: 1) computing a Multivariate Matrix (MVM) and the 2) selection of the AI technique that will analyse the MVM. The aforementioned components of MVPA are explained in detail next.

In most multivariate analysis, the feature set is crafted by hand. That is the statistic characteristic (such as mean, amplitude etc.) 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 $\mathbf{M}$ channels (for fNIRS) and $\mathbf{E}$ number of data trials is then given as input to an Artificial Intelligence (AI) method, and is hereby referred to as a Multivariate Matrix (MVM).

Although it requires considerable subject-matter expertise to select 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 validity of the selected features to represent the dynamics of the underlying cortical networks. In this regards, 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 AI paradigms seek to find patterns within the input data that are characteristic of each of the classes (or stimuli in DCN studies). The classification of unlabelled input data is done on the basis of the prototypical patterns found by the AI paradigms. A greater classification accuracy of an AI method is a testament that it has discerned the underlying patterns to the level of the classification accuracy. The MVM can be readily analysed using any state of the art AI methods. Most AI methods such as Support Vector Machine (SVM), Random Forest (RF) etc. usually give very robust classification results with MVM.

The AI methods' inference mechanism, driving the MVPA, identifies patterns within it (MVPA) that are prototypical of the presented stimuli. However, the inference mechanism of the AI method must be able to describe the cortical networks' representations within the MVPA to shed light on the brain development. In this regard, it is critical for the AI methods to have a transparent, explainable, human-understandable inference mechanism so that the patterns found in MVPA can shed light on the cortical networks formed in response to the presented stimuli. Indeed, a deeper understanding of cortical brain networks for the processing of presented stimuli in the developing brain would shed light on the interplay between the physical growth of the activated brain regions and the emergence of new behavioural abilities during brain development [3].

In the next section, fNIRS connectivity analysis is presented.

Connectivity Analysis

Brain connectivity analysis can shed some 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 [113] that can inform about the organisation and the working of the developing human brain are: 1) Structural Connectivity 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. Whereas, FC is measured as temporal correlation between spatially remote neurophysiological events [13]. In contrast, EC measures the influence that one neural system exerts over another which can be both activity, and/or time dependent [114].

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). This is because the extent to which the AI technique can inform about the underlying cortical networks depends on the level of explainability of the AI technique.

An example of a recent work that analyses EC in infants' fNIRS data is by Bulgarelli et al. [15]. In the work by Bulgarelli et. al, the authors undertook a multimodal study (fMRI and fNIRS) to estimate the EC in a six month old asleep infant using Dynamic Causal Modelling (DCM). Although they were able to choose the most suitable DCM model that best explains the data using Bayesian inference, the DCM model is hinged on the correct estimation of the sensitivity matrix. In addition, their DCM model was able to only outline the (presence or absence of an) interconnection between different brain regions, i.e. their analysis did not include corresponding (numeric/quantitative) EC values. In this work, I have developed an explainable EC connectivity analysis (called EFCM presented in Chapter 5.1) which does not rely on *a priori* information or estimated values. Further, the ECFM method is able to delineate the cortical networks' reconfiguration using EC values as hypothesised by the neural reuse framework. In the next chapter, I outline fuzzy logic systems (FLS) which form an integral part of the xMVPA (presented in Chapter 5.2) developed for the implementation of the IS framework (Chapter 2.1).

Chapter Three

Background on Fuzzy Logic Systems (FLS) based explainable AI (XAI)

In this chapter, I present an overview of Fuzzy Logic Systems (FLS). FLS forms the basis of new explainable methods (presented in Chapter 5), that I developed in this work, for the analysis of infants' fNIRS data in line with the DCN (Developmental Cognitive Neuroscience) frameworks (previously outlined in Chapter 2.1).

Over the last few decades, the widespread application of AI (artificial intelligence) systems have enhanced many aspects of everyday life from risk management [115], sky shepherding of sheep [116], medical image segmentation [117], recognition of expertise level [21], mobile applications [118] to Covid-19 detection based on cough samples [119]. Although opaque AI systems offer remarkable prediction accuracy, they are limited by a lack of explanation behind their predictions. A lack of explanation renders the AI systems untrustworthy, and particularly inapplicable where users want to understand the decision process of the AI system. To this end, there is a growing need for transparent, human-understandable AI systems called XAI systems [120]. Several approaches taken towards the development of XAI systems include: 1) Intrinsic: a method in which model inference structure is fully transparent, and 2) Post-hoc: a model-agnostic meta-model is used to decipher the

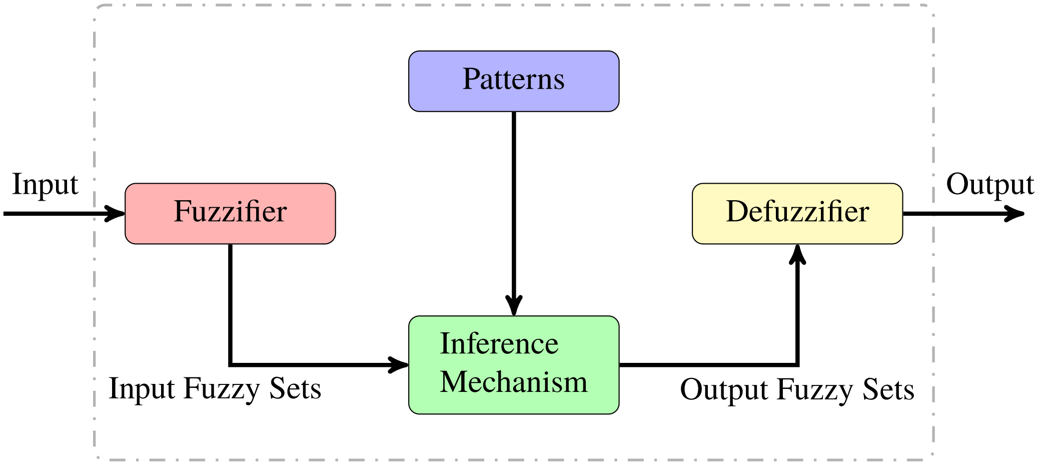


Figure 3.1 An illustration of the main components of a Fuzzy Logic System (FLS).

inference rationale of a black-box model. Within post-hoc methods attempts to understand a black-box model using an intrinsic model have also been undertaken. A particular category of these are the anchor-based models.

Although anchor-based approach provides a step towards implementing human-understandable explanations [121], explanatory patterns rest on hard thresholds and are constrained by Boolean logic. They are not suitable for complex, real-life processes which are characterised with uncertainty. In this regard, another approach to implement XAI systems is FLS [120, 122]. The FLS based XAI systems are well-suited for explainable modelling of real-life processes because of FLS capability to handle uncertainty in the input data, and subsequently improve the process model and performance. Further, the use of conceptual labels that model uncertainty and axioms of FLS based XAI systems pave way for human-understandable models for describing complex, real-life processes.

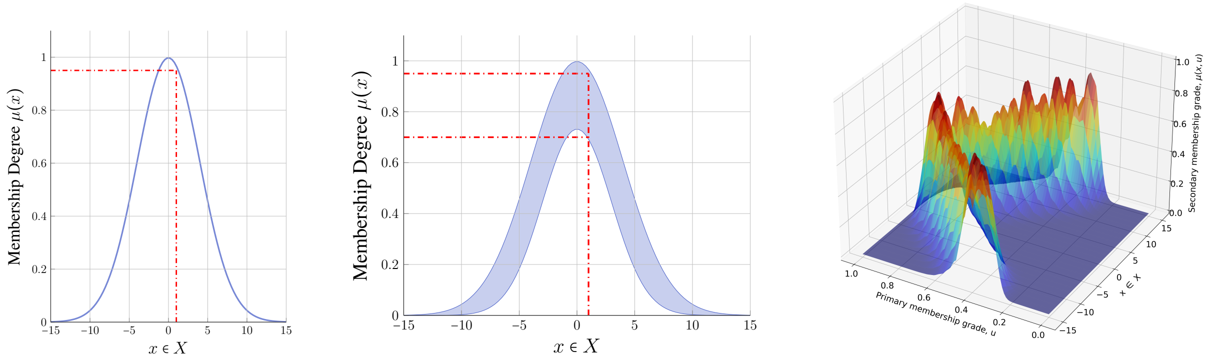
In addition to providing explainability, FLS can also be used for analysing both classification (predict one output class for example car or house) and regression (predict a quantitative number for example annual sales prediction) problems. In general, the main components of a FLS are a fuzzifier, patterns (or rules), inference mechanism, and a defuzzifier, as illustrated in Fig. 3.1. Also, since in the context of this work, ‘rules’ and ‘patterns’ can be used interchangeably hence from here on-wards only the word ‘patterns’ is used for keeping the text consistent.The aforementioned components of the FLS are described in detail below.

The fuzzifier (of a FLS) handle uncertainty in the input data using fuzzy sets that convert crisp numbers (viz. uncertain observations) to conceptual labels characterised with membership degrees [122, 123]. The fuzzy sets are defined by Membership Functions (MFs) and represent a given conceptual label. The membership degrees are usually in the range [0,1] and is a soft measure of degree of association the associated fuzzy set has for a given crisp measurement to belong to the conceptual label represented by the fuzzy set [123]. For example, an XAI system modelling the heights of people in a community using type-1 fuzzy sets may represent height using conceptual labels of *Tall, Medium,* and *Short*. The MF associated with each conceptual label’s MF will assign a crisp number for the height of a person with a membership degree; for example, a height of 6ft may get assigned membership degrees of $0.8, 0.5, 0.1$ to represent conceptual labels of *Tall, Medium,* and *Short* respectively.

While fuzzy sets transform measurements from input features into conceptual labels, the patterns in a FLS outline the relationship between the input features (antecedents) and the output (consequent) using conceptual labels and propositions. The patterns, for a given process, can be provided by the experts in the relevant field [124] or can also be learnt using input data using evolutionary algorithms [125]. In general, the patterns of a FLS are formed of two parts: the antecedent part, and the consequent part, as outlined in (3.1).

$$\text{Pattern} : \text{IF } Antecedents \text{ THEN } Consequent \text{ with dominance score} \tag{3.1}$$

An example of a pattern for a classification problem in an fNIRS neuroimaging study can be *IF activity in Channel 1 is Low AND activity in Channel 3 is High THEN stimulus is Auditory with dominance score = 0.35*. In this illustrative pattern for a classification problem, Channel 1 and



(a) T1 Fuzzy Set

(b) IT2 Fuzzy Set

(c) GT2 Fuzzy Set

Figure 3.2 The three types of fuzzy sets: (a) Type-1 (T1) (b) Interval type-2 (IT2) and (c) General type-2 (GT2) fuzzy sets.

Channel 3 are antecedents (input), Low and High refer to the conceptual labels for the activity (input feature) of Channel 1 and Channel 3 respectively whereas Auditory refers to the class (stimulus). The dominance score is a measure of the prowess of a pattern to correctly predict the class for a given input data instance.

The inference mechanism of a FLS reads from the patterns and input fuzzy sets to generate output fuzzy sets. More specifically, the inference mechanism quantifies the matching between a given data instance and the patterns. The defuzzifier, for a regression problem, defuzzifies the output fuzzy set, i.e. a quantitative prediction is made. Whereas for a classification problem, the defuzzifier, predicts a class for the output post the inference mechanism. In this work, the FLS are used for classification problems (predict the stimulus) therefore in Chapter 3.4 I outline the method for solving a classification problem using the different types of FLS.

In general, there are three main types of FLS based on the composition of the innate fuzzy sets. The three types of fuzzy sets are: 1) Type-1 (T1), 2) Interval Type-2 (IT2), and 3) General Type-2 (GT2) fuzzy sets. As can be seen in Fig. 3.2, all fuzzy sets model uncertainty in the feature domain but to different extent; T1 fuzzy sets model the least uncertainty and GT2 fuzzy sets model the most uncertainty amongst all fuzzy sets. Also, in Fig. 3.2, Gaussian function is used to illustrate the different types of fuzzy sets; however, other types of MFs (such as triangular, trapezium) can also be used. Evidently, the shape of MF impacts the inference made in a FLS [126]. Hence, it is important to choose the shape of the MF, and determine optimum parameters for the shape of MFs such as through particle swarm optimisation [127].

In addition to modelling uncertainty, all fuzzy sets share the following properties [124]:

1. A given number, from the feature domain, can belong to more than one fuzzy set simultaneously (an important distinction from classical sets).

2. Fuzzy sets are convex, i.e. the MF of fuzzy set is first monotonically non-decreasing and then monotonically non-increasing.

3. A normal fuzzy set has the maximum value of membership as 1, i.e. $\mu_A(x) = 1$ where A is a normal fuzzy set.

In the next subsections, I outline the three most common types of fuzzy sets namely: T1, IT2, and GT2 fuzzy sets.

3.1 Type-1 (T1) Fuzzy Sets

The T1 fuzzy sets where each crisp measurement, $x \in X$, gets assigned a membership degree, $\mu_{T1}(x) \subseteq [0, 1]$, but there is no ambiguity in the membership degree, for example as shown by the red dashed line in Fig. 3.2 (a): $\mu_{T1}(x = 1) = 0.95$. More specifically, the membership degree, of a crisp measurement (such as $x = 1$), using T1 fuzzy set is also a crisp number (for $x = 1$, it is 0.95).

3.1.1 Definition

A Type-1 (T1) fuzzy set, denoted by A, is defined on universe X such that $\mu_A(x) \rightarrow [0,1]$ where $\mu_A(x)$ is the associated MF of A. T1 fuzzy sets can be written mathematically as follows [124]:

$$A = \{(x, \mu_A(x) \mid x \in X)\} \tag{3.2}$$

In set notation, T1 fuzzy sets can be written for continuous universe X as:

$$A = \int_{x \in X} \mu_A(x)/x \tag{3.3}$$

where $\int$ represents union over all admissible values of $x \in X$. For discrete universe, the set notation for T1 fuzzy sets can be written as follows:

$$A = \sum_{x \in X_d} \mu_A(x)/x \tag{3.4}$$

where $\sum$ represents union over all admissible values of $x \in X_d$ in the discrete universe X_d. Also, please note, the slash '/' in the equations (3.3 and 3.4) links the values of $x \in X$ with their corresponding values of membership degree $\mu_A(x) > 0$.

In the next sections, I outline some of most common operations on T1 fuzzy sets such as the fuzzification, union, intersection, and defuzzification.

3.1.2 Fuzzification using T1 Fuzzy Sets

As mentioned earlier in the chapter, fuzzy sets transform a numerical value/measurement into a conceptual label with a membership degree (for that numerical value to belong to the conceptual label represented by the fuzzy set). This is also termed as fuzzification of a numerical value. There are two main types of fuzzification [124]: 1) Singleton and 2) Non-Singleton fuzzification. The two fuzzifications only differ in the membership degree's value. In particular, singleton fuzzification is

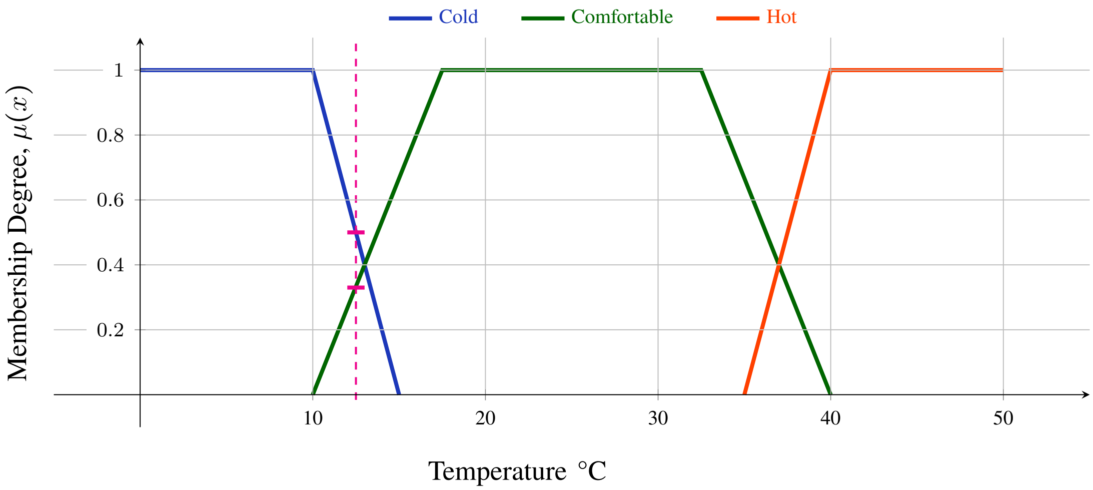
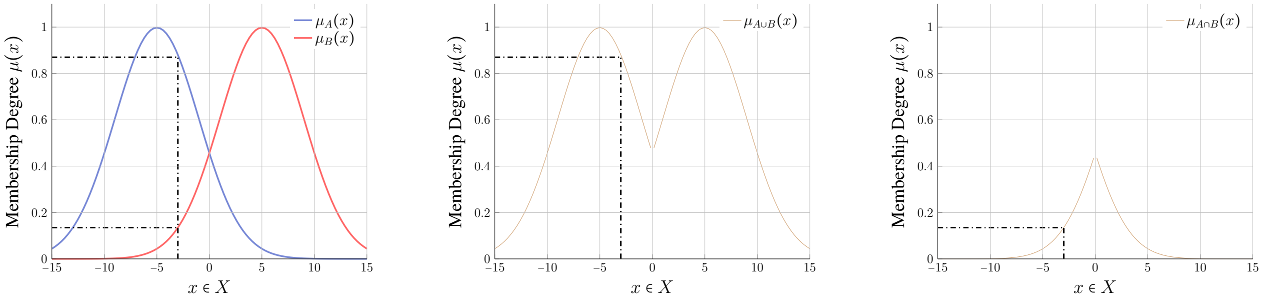


Figure 3.3 An illustrative plot to exemplify fuzzification using type-1 (T1) fuzzy sets.

only dependent on the MF whereas non-singletion fuzzification also accounts for input uncertainties [128]. However, for ease of computation, in this work, singleton fuzzification is implemented throughout, and is explained next.

An illustrative plot to exemplify how a conceptual label is characterised with uncertainty handling using T1 fuzzy sets is shown in Fig. 3.3 with reference to thermal concepts. Thermal comfort can be expressed with the conceptual labels *cold*, *comfortable* and *hot* with approximate membership degree values, μ, obtained from [129]. As can be seen in the Fig. 3.3, the definition of conceptual labels is not necessarily mutually exclusive, i.e. a certain temperature can be represented using more than one conceptual label with varying membership degrees. For example, the temperature of 12 °C has membership degree as: $\mu_{cold}(12°C) = 0.5$ and $\mu_{comfortable}(12°C) = 0.35$ and $\mu_{hot}(12°C) = 0$. The derived ambiguity in the membership degree ensures that uncertainty in a numeric value is well retained upon transformation into a conceptual label.



(a) T1 Fuzzy Sets. (b) Union. (c) Intersection.

Figure 3.4 The union and intersection of two Type-1 (T1) fuzzy sets A and B.

3.1.3 Union of T1 Fuzzy Sets

Let A and B be two T1 fuzzy sets defined on the universe X and let their respective MFs be as $\mu_A(x)$ and $\mu_B(x)$. The union of two T1 fuzzy sets is another T1 fuzzy set such that:

$$\mu_{A\cup B}(x) = \max[\mu_A(x), \mu_B(x)] \tag{3.5}$$

where max t-conorm takes the maximum of the two membership degrees corresponding to each $x \in X$. The union of T1 fuzzy sets is also illustrated in Fig. 3.4 (b). For example, for $x = -3$, the membership degrees of $\mu_A(x = -3) = 0.87$ and $\mu_B(x = -3) = 0.135$, therefore the union will be:

$$\begin{aligned}\mu_{A\cup B}(x = -3) &= \max[0.87, 0.135] \\ &= 0.87 \end{aligned} \tag{3.6}$$

3.1.4 Intersection of T1 Fuzzy Sets

The intersection of two T1 fuzzy sets A and B is another T1 fuzzy set such that:

$$\mu_{A\cap B}(x) = \min[\mu_A(x), \mu_B(x)] \tag{3.7}$$

where min t-norm takes the minimum of the two membership degrees corresponding to each $x \in X$. For example, for $x = -3$, the membership degrees of $\mu_A(x = -3) = 0.87$ and $\mu_B(x = -3) = 0.135$, therefore the intersection will be:

$$\begin{aligned} \mu_{A \cap B}(x = -3) &= \text{min}[0.87, 0.135] \\ &= 0.135 \end{aligned} \tag{3.8}$$

Please note for the computation of intersection of T1 fuzzy sets, either min t-norm or product of the membership degrees can be used. The intersection of T1 fuzzy sets is also illustrated in Fig. 3.4 (c).

3.1.5 Defuzzification of T1 Fuzzy Sets

The defuzzification of a T1 fuzzy set, A, transforms the fuzzy set into an equivalent number, and can be thought of as an inverse of fuzzification. For T1 fuzzy sets, the defuzzification usually involves computing the centroid of the fuzzy set as shown in (3.9):

$$x^* = \frac{\sum_{i=1}^{I} x_i \mu_A(x_i)}{\sum_{i=1}^{I} \mu_A(x_i)} \tag{3.9}$$

where x^* is the centroid (or defuzzified value) of the T1 MF defined on the domain $x \in X$. Here, the summation sign is used as in typical mathematical equations, i.e., for the case of the numerator, it is summing the product of x values and their corresponding membership degrees whereas for the denominator it is summing the membership degrees corresponding to all x_i values $\forall i \in [1, ..., I]$. As an example, the defuzzification of the T1 fuzzy set shown in Fig. 3.5 can be undertaken using

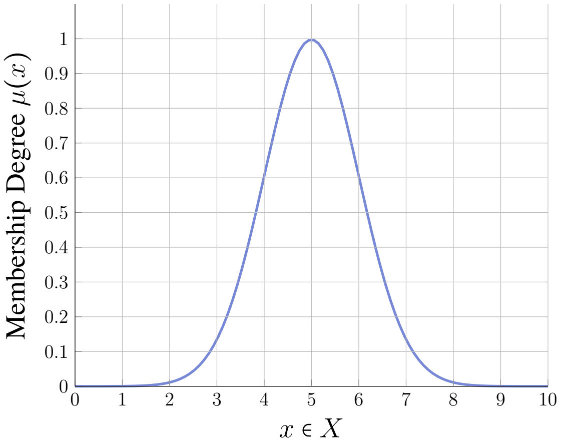


Figure 3.5 The defuzzification of a T1 Fuzzy Set using (3.9).

(3.9) as shown below:

$$\begin{aligned} x^* =& \frac{\sum_{i=1}^{I} x_i \mu_A(x_i)}{\sum_{i=1}^{I} \mu_A(x_i)} \\ =& \frac{1 * 0 + \ldots + 2 * 0.009 + \ldots + 5 * 1 + \ldots + 8 * 0.009 + \ldots + 9 * 0}{0 + \ldots + 0.009 + \ldots + 1 + \ldots + 0.009 + \ldots + 0} \\ =& 5 \end{aligned} \tag{3.10}$$

Apart from centroid calculation, other most common methods of defuzzification of T1 fuzzy sets are the centre-of-set and height defuzzification [124].

3.2 Interval Type-2 (IT2) Fuzzy Sets

The Interval Type-2 (IT2) fuzzy sets are three-dimensional (3D) and characterised by a primary membership degree, and a secondary membership degree which is always 1 [124]. The MF associated with an IT2 fuzzy set assigns each numerical value, $x \in X$, a range of primary membership degrees such that $\mu_{IT2}(x) \subseteq [0, 1]$. An example of an IT2 fuzzy set is shown in Fig. 3.2 (b) with red dashed line denoting a particular numerical value's (x=1) IT2 membership degree: $\mu_{IT2}(x = 1) = [0.7, 0.95]$.

More specifically, the membership degree, of a numerical value (such as $x = 1$), has lower (such as 0.7) and upper (such as 0.95) bounds. The area between the lower and upper bounds is often termed as the footprint of uncertainty (FOU).

3.2.1 Definition

An Interval Type-2 (IT2) fuzzy set, denoted by $\tilde{A}$, is defined on universe X such that $\mu_{\tilde{A}}(x) \rightarrow [0,1]$ where $\mu_{\tilde{A}}(x)$ is the associated MF of $\tilde{A}$. IT2 fuzzy sets can be written mathematically as follows [124]:

$$\begin{aligned} \tilde{A} = \{(x, u, 1) | \forall x \in X, \\ \forall u \in [\underline{\mu}_{\tilde{A}}(x), \overline{\mu}_{\tilde{A}}(x)] \subseteq [0,1]\} \end{aligned} \tag{3.11}$$

where $\underline{\mu}_{\tilde{A}}(x)$ is the <u>lower</u> membership degree and $\overline{\mu}_{\tilde{A}}(x)$ is the $\overline{\text{upper}}$ membership degree. In set notation, IT2 fuzzy sets can be written for continuous universe X as [130]:

$$\begin{aligned} \tilde{A} &= \int_{x \in X} \int_{u \in J_x} 1/(u, x) \\ &= \int_{x \in X} [\int_{u \in J_x} 1/u]/x \end{aligned}$$

where $\int$ represents union over all admissible values of $x \in X$ and $u \in J_x$, $J_x \subseteq [0,1]$ is the primary membership of x. The FOU of $\tilde{A}$ is the area between the lower and upper MF and is illustrated as the shaded area in Fig. 3.2(b). The FOU can also be expressed as:

$$\text{FOU}(\tilde{A}) = \bigcup_{x \in X} J_x \tag{3.12}$$

Also, the slash '/' in the equation (3.12) links the values of $x \in X$ with their corresponding values of membership degree $\mu_A(x) > 0$.

In the next sections, some of most common operations on IT2 fuzzy sets such as the fuzzification, union, intersection, and defuzzification are outlined.

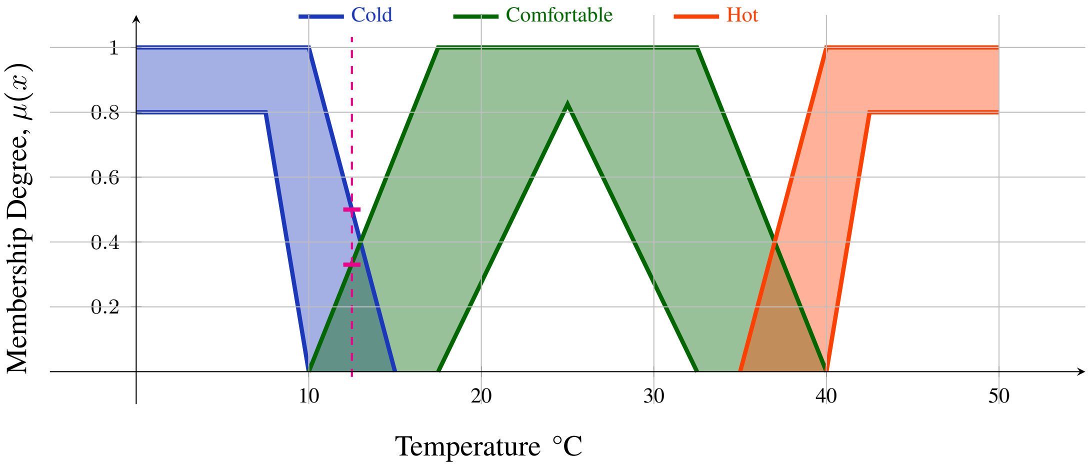
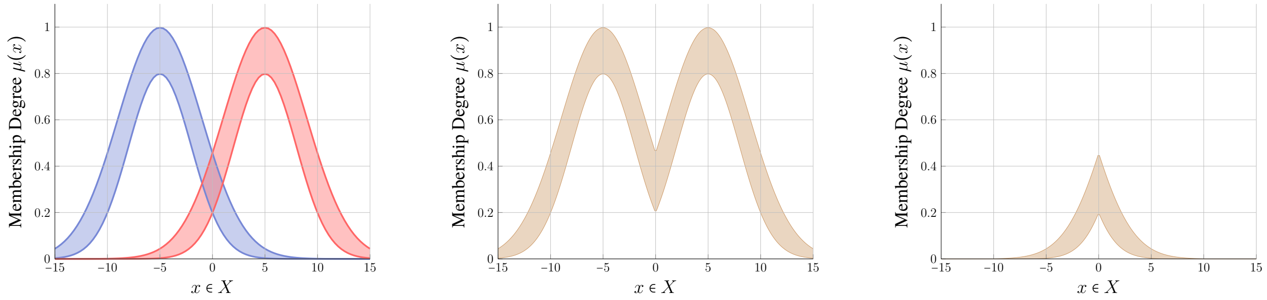


Figure 3.6 An illustrative plot to exemplify fuzzification using Interval Type-2 (IT2) fuzzy sets.

3.2.2 Fuzzification using IT2 Fuzzy Sets

The fuzzification of a measurement/observation using an IT2 fuzzy set results in an upper and lower membership degree, i.e. $\mu_{\tilde{A}}(x) = [\underline{\mu}_{\tilde{A}}(x), \overline{\mu}_{\tilde{A}}(x)]$.

An illustrative plot to exemplify how a conceptual label is characterised with uncertainty handling using IT2 fuzzy sets is shown in Fig. 3.6 with reference to thermal concepts. Thermal comfort can be expressed with the conceptual labels *cold*, *comfortable* and *hot* with approximate degree of membership values, μ, obtained from [129]. As can be seen in the Fig. 3.6, the definition of conceptual labels is not necessarily mutually exclusive, i.e. a certain temperature can be represented using more than one conceptual label with varying membership degrees. For example, the temperature of 12 °C has membership degree, μ, in the range of (0, 0.5) for *cold* and (0, 0.33) for *comfortable*. The derived ambiguity in the membership degree ensures that uncertainty in the numerical value (or neuroimaging reading from fNIRS) is well retained upon transformation into a conceptual label.



(a) IT2 Fuzzy Sets.

(b) Union.

(c) Intersection.

Figure 3.7 The union and intersection of two Interval Type-2 (IT2) fuzzy sets $\tilde{A}$ and $\tilde{B}$.

3.2.3 Union of IT2 Fuzzy Sets

Let $\tilde{A}$ and $\tilde{B}$ be two IT2 fuzzy sets defined on the universe X and let their respective MFs be as $\mu_{\tilde{A}}(x)$ and $\mu_{\tilde{B}}(x)$. The union of two IT2 fuzzy sets, such as $\tilde{A}$ and $\tilde{B}$, is another IT2 fuzzy set such that the area enclosed in the FOU of $\tilde{A}\bigcup\tilde{B}$ is a union of the FOU's of $\tilde{A}$ and $\tilde{B}$. In mathematical notation, the union operation on two IT2 fuzzy sets can be written as follows:

$$\begin{aligned}\tilde{A}\bigcup\tilde{B} &= 1/\mathrm{FOU}(\tilde{A}\bigcup\tilde{B})\\ &= 1/[\max(\underline{\mu}_{\tilde{A}}(x),\underline{\mu}_{\tilde{B}}(x)),\ \max(\overline{\mu}_{\tilde{A}}(x),\overline{\mu}_{\tilde{B}}(x))]\end{aligned} \tag{3.13}$$

Also, the slash '/' in (3.13) links the primary membership degrees ($\mu_{\tilde{A}}(x) > 0$ and $\mu_{\tilde{B}}(x) > 0$) with their corresponding secondary membership degrees (which is always 1 for an IT2 fuzzy set). The union of two IT2 fuzzy sets is also illustrated in Fig. 3.7 (b).

3.2.4 Intersection of IT2 Fuzzy Sets

The intersection of two IT2 fuzzy sets, such as $\tilde{A}$ and $\tilde{B}$, is another IT2 fuzzy set such that the area enclosed in the FOU of $\tilde{A}\bigcap\tilde{B}$ is an intersection of the FOU's of $\tilde{A}$ and $\tilde{B}$. In mathematical notation, the intersection operation on two IT2 fuzzy sets can be written as follows:

$$\begin{aligned}\tilde{A}\bigcap\tilde{B} &= 1/\text{FOU}(\tilde{A}\bigcap\tilde{B})\\ &= 1/[\min(\underline{\mu}_{\tilde{A}}(x),\underline{\mu}_{\tilde{B}}(x)),\ \min(\overline{\mu}_{\tilde{A}}(x),\overline{\mu}_{\tilde{B}}(x))]\end{aligned} \tag{3.14}$$

Also, the slash '/' in (3.14) links the primary membership degrees ($\mu_{\tilde{A}}(x) > 0$ and $\mu_{\tilde{B}}(x) > 0$) with their corresponding secondary membership degrees *which is always 1 for an IT2 fuzzy set.* The intersection of two IT2 fuzzy sets is illustrated in Fig. 3.7 (c).

3.2.5 Defuzzification of IT2 Fuzzy Sets

The defuzzification of an IT2 fuzzy set can be done by either 1) type reduction followed by defuzzification, or using 2) direct defuzzification. In type reduction, IT2 fuzzy set is first type-reduced to a T1 fuzzy set and then the centroid of the reduced T1 fuzzy set is computed. Whereas in direct defuzzification, one crisp (or defuzzified) value representative of the IT2 fuzzy set is found.

The centroid, $C_{\tilde{A}}$, of an IT2 fuzzy set, $\tilde{A}$, is defined as the union of all the centroids of T1 fuzzy sets embedded (within the IT2 fuzzy set), denoted by $\tilde{A}_e$ [130]. In other words, $C_{\tilde{A}}$ denotes an interval enclosed by left and right centroids. In mathematical terms, the centroid of an IT2 fuzzy set can be written as follows [130]:

$$C_{\tilde{A}} \equiv \bigcup_{\forall A_e} c(A_e) = [c_l(\tilde{A}), c_r(\tilde{A})] \tag{3.15}$$

where $\bigcup$ represents the union operation. The left and right centroids of $\tilde{A}$ are the minimum and maximum of the centroids of the embedded T1 fuzzy sets, A_e i.e.:

$$c_l(\tilde{A}) = \text{min}_{\forall A_e} c(A_e) \tag{3.16}$$

$$c_r(\tilde{A}) = \text{max}_{\forall A_e} c(A_e) \tag{3.17}$$

where c_l and c_r can also be expressed as follows [131]:

$$c_l(\tilde{A}) = \frac{\sum_{i=1}^{L} x_i \overline{\mu}_{\tilde{A}(x_i)} + \sum_{i=L+1}^{J} x_i \underline{\mu}_{\tilde{A}(x_i)}}{\sum_{i=1}^{L} \overline{\mu}_{\tilde{A}(x_i)} + \sum_{i=L+1}^{J} \underline{\mu}_{\tilde{A}(x_i)}} \tag{3.18}$$

$$c_r(\tilde{A}) = \frac{\sum_{i=1}^{R} x_i \underline{\mu}_{\tilde{A}(x_i)} + \sum_{i=R+1}^{J} x_i \overline{\mu}_{\tilde{A}(x_i)}}{\sum_{i=1}^{R} \underline{\mu}_{\tilde{A}(x_i)} + \sum_{i=R+1}^{J} \overline{\mu}_{\tilde{A}(x_i)}} \tag{3.19}$$

where L and R represent the indices corresponding to the switch points x_L and x_R. The switch points are most commonly found using the iterative Karnik-Mendel (KM) algorithm [131], outlined in Appendix A. Once $c_l(\tilde{A})$ and $c_r\tilde{A}$ are computed, the final defuzzified value (or the centroid in (3.15)) can be obtained by taking their average, i.e.

$$C(\tilde{A}) = \frac{c_l(\tilde{A}) + c_r(\tilde{A})}{2} \tag{3.20}$$

In contrast to the KM algorithm which require iterations to find the switch points, another method for centroid calculation of an IT2 fuzzy set is the Nie-Tan approach [132]. The advantage of the Nie-Tan approach, over KM algorithm, is that it gives a closed form mathematical equation for the computation of a given IT2 fuzzy set's centroid. The centroid equation using Nie-Tan approach is as follows in (3.21) [132]:

$$C(\tilde{A}) = \frac{(\sum_{j=1}^{J} x_j * \overline{\mu}_{\tilde{A}}(x_j) + \sum_{j=1}^{J} x_j * \underline{\mu}_{\tilde{A}}(x_j))}{(\sum_{j=1}^{J} \overline{\mu}_{\tilde{A}}(x_j) + \sum_{j=1}^{J} \underline{\mu}_{\tilde{A}}(x_j))} \tag{3.21}$$

3.3 General Type-2 (GT2) Fuzzy Sets

General Type-2 (GT2) fuzzy systems are 3D like IT2 fuzzy sets, discussed earlier in the section 3.2, with primary and secondary membership degrees. However, unlike IT2 fuzzy sets whose secondary membership degree is always 1, the secondary membership degree of a GT2 fuzzy set can take any value $\subseteq [0,1]$. More specifically, GT2 fuzzy sets have T1 fuzzy sets as membership degree for a given numerical value, for example: $\mu_{T2}(x = 1) = \{u, \mu_{T1}(u) | \forall u \in [0,1], \forall \mu_{T1} \in [0,1]\}$ where u is called the primary membership degree and μ is called the secondary membership degree as illustrated in in Fig. 3.2 (c). The primary and secondary membership degrees enable GT2 fuzzy set to model uncertainty in the input data to greater extent however it also renders them complex, and computationally expensive.

3.3.1 Definition

A General Type-2 (GT2) fuzzy set, denoted $\tilde{A}$ is characterised with a bivariate MF, $\mu_{\tilde{A}(x,u)}$. In mathematical notation, $\tilde{A}$ is defined as [124]:

$$\tilde{A} = \{((x,u), \mu_{\tilde{A}}(x,u)) | x \in X, u \in U \equiv [0,1]\} \tag{3.22}$$

where X is universe for primary variable of $\tilde{A}$, i.e. x, and U is the universe for secondary variable of $\tilde{A}$, i.e. u. In set notation, $\tilde{A}$ can be expressed as follows:

$$\tilde{A} = \int_{x \in X} \int_{u \in [0,1]} \mu_{\tilde{A}}(x,u)/(x,u) \tag{3.23}$$

where $\int_{x \in X} \int_{u \in [0,1]}$ denotes the union over all admissible values of x and u.

The secondary degree of x, denoted $\mu_{\tilde{A}}(x,u)$ in (3.23), is also represented as $\ell_x(u)$ i.e. $\mu_{\tilde{A}}(x,u) \equiv$

$\ell_x(u) \subseteq [0,1]$. Another notation for (3.23) is:

$$\tilde{A} = \int_{x\in X}\int_{u\in[0,1]} \mu_{\tilde{A}}(x,u)/(x,u) = \int_{u\in[0,1]} \ell_x(u)/u \tag{3.24}$$

3.3.2 Fuzzification using GT2 Fuzzy Sets

The fuzzification of a numerical value in the universe of discourse, i.e. $x \in X$, using GT2 fuzzy sets results in a T1 MF. As a result of a T1 MF at each value $x \in X$ the corresponding computations are significantly complex in GT2 fuzzy sets than IT2 fuzzy sets (fuzzification results in an interval for membership degree; see section 3.2) and T1 fuzzy sets (fuzzification results in one membership degree; see section 3.1).

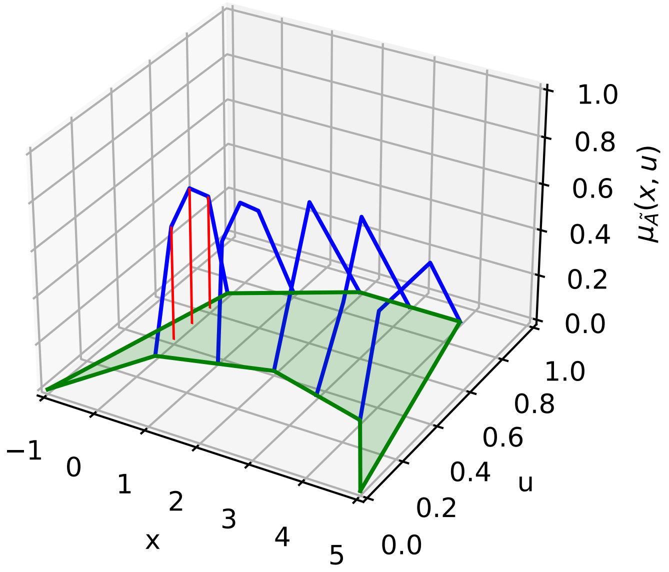


Figure 3.8 An illustrative plot to exemplify fuzzification using General Type-2 (GT2) fuzzy sets.

An over-simplified illustration of fuzzification with GT2 fuzzy sets is shown in Fig. 3.8. In the figure, five secondary MFs, $\mu_{\tilde{A}}(x = [0,1,2,3,4], u)$ corresponding to five values of $x = [0,1,2,3,4]$ are plotted in blue line. Whereas, the red lines in Fig. 3.8 link the primary degrees at the y-axis $u = [0.3, 0.4, 0.5]$ with the non-zero secondary degrees at z-axis μ. The secondary MFs, $\mu_{\tilde{A}}(x = [0, 1, 2, 3, 4], u)$ (represented in blue in Fig. 3.8) are also referred to as vertical slice representations of GT2 fuzzy sets at given values of x (in this case at $x = [0, 1, 2, 3, 4]$). In general, the vertical slice representation expresses the GT2 fuzzy set as a union of all its secondary T1 MFs.

In set notation, the GT2 MF at $x = 0$ at $u_{x=0} = [0.4, 0.5, 0.6]$ can be written as follows (see (3.24)):

$$\mu_{\tilde{A}(x=0,u)} = \ell_{x=0}(u) = 0.5/0.4 \ + \ 0.6/0.5 \ + \ 0.5/0.6 \tag{3.25}$$

where the '/' links the primary membership degrees u with their corresponding secondary membership degrees μ.

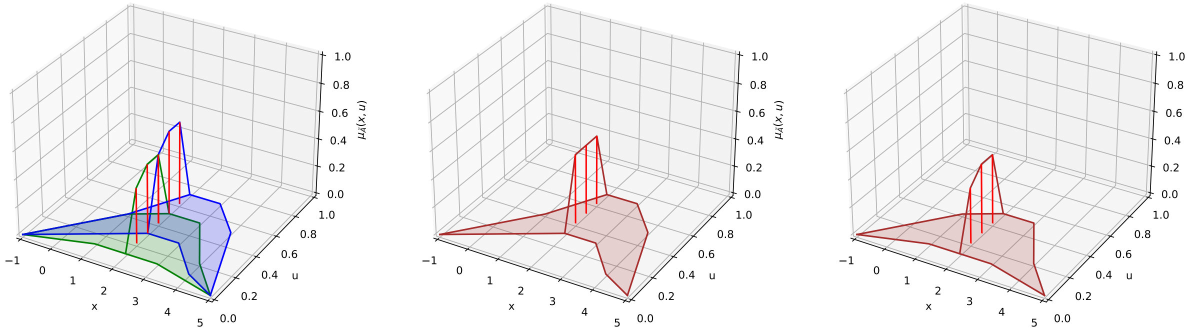
3.3.3 Union of GT2 Fuzzy Sets

The union of two GT2 fuzzy sets results in another GT2 fuzzy sets. However, as stated earlier, computations involving GT2 fuzzy sets are complex since for each value in the universe of discourse the membership degree is a function (T1 MF). A common approach for computing the union of GT2 fuzzy sets is the 'Extension Principle' [128, 133]. The extension principle undertakes vertical slices of the GT2 fuzzy set at each value of $x \in X$.

Let $\tilde{A}$ and $\tilde{B}$ be two GT2 fuzzy sets defined on the universe X and let their respective MFs be as $\mu_{\tilde{A}}(x, u)$ and $\mu_{\tilde{B}}(x, u)$. Using the extension principle, the union of two GT2 fuzzy sets $\tilde{A}$ and $\tilde{B}$, can be expressed in mathematical notation as follows [124]:

$$\mu_{(\tilde{A} \cup \tilde{B})_{(x)}} \equiv \mu_{\tilde{A}(x)} \sqcup \mu_{\tilde{B}(x)} = \int_{v \in [0,1]} \int_{w \in [0,1]} \ell_x(v) \star \hbar_x(w) / (u = v \vee \omega) \tag{3.26}$$

where $\sqcup$ denotes the union operation, $\star$ indicates the product or minimum (also called t-norm) operation, and $\vee$ indicates the maximum (also called t-conorm). The union of primary membership



(a) GT2 fuzzy sets.

(b) Union.

(c) Intersection.

Figure 3.9 An illustration of the (b) union and (c) intersection of (a) two General Type-2 (GT2) fuzzy sets using vertical-slice representation at $x = 2$.

degree (represented by v and ω for GT2 fuzzy sets $\tilde{A}$ and $\tilde{B}$ respectively) is computed for all values of $x \in X$.

In simple terms, (3.26) outlines the union of two GT2 fuzzy sets using the extension principle such that for each $x \in X$:

1. $u = v \vee \omega$ entails that for all possible combinations of the primary membership degrees (v and ω) the secondary degree $\mu_{(\tilde{A} \cup \tilde{B})_x}$ is computed by the minimum operation (or t-norm, denoted as $\star$ in (3.26)) between the corresponding secondary degrees i.e. $\ell_x(v) \star \hbar_x(\omega)$.

2. In case of a duplicate point(s) for primary membership degree after the calculation of $v \vee \omega$, then the point (primary membership degree) with the largest secondary membership degree is selected.

In Fig. 3.9 (b) an illustration for the union of two GT2 fuzzy sets $\tilde{A}$ and $\tilde{B}$ using extension principle (with vertical slice at $x = 2$) is presented. The corresponding calculations are outlined as follows:

$$\begin{aligned}
\mu_{(\tilde{A}\cup\tilde{B})_{x=2}}(u) =& \mu_{\tilde{A}(x=2)} \sqcup \mu_{\tilde{B}(x=2)} \\
=&(0.4/0.2 + 0.5/0.3 + 0.5/0.4) + (0.5/0.4 + 0.6/0.5 + 0.6/0.6) \\
=&\frac{\min(0.4,0.5)}{0.2 \vee 0.4} + \frac{\min(0.4,0.6)}{0.2 \vee 0.5} + \frac{\min(0.4,0.6)}{0.2 \vee 0.6} + \frac{\min(0.5,0.5)}{0.3 \vee 0.4} + \frac{\min(0.5,0.6)}{0.3 \vee 0.5} + \\
&\frac{\min(0.5,0.6)}{0.3 \vee 0.6} + \frac{\min(0.5,0.5)}{0.4 \vee 0.4} + \frac{\min(0.5,0.6)}{0.4 \vee 0.5} + \frac{\min(0.5,0.6)}{0.4 \vee 0.6} \\
=&0.4/0.4 + 0.4/0.5 + 0.4/0.6 + 0.5/0.4 + 0.5/0.5 + 0.5/0.6 + 0.5/0.4 + 0.5/0.5 + 0.5/0.6 \\
=&\max(0.4,0.5,0.5)/0.4 + \max(0.4,0.5,0.5)/0.5 + \max(0.4,0.5,0.5)/0.6 \\
=&0.5/0.4 + 0.5/0.5 + 0.5/0.6 \qquad (3.27)
\end{aligned}$$

3.3.4 Intersection of GT2 Fuzzy Sets

Using the extension principle, the intersection of two GT2 fuzzy sets $\tilde{A}$ and $\tilde{B}$, can be expressed in mathematical notation as follows [124]:

$$\mu_{(\tilde{A}\cap\tilde{B})_{(x)}} \equiv \mu_{\tilde{A}(x)} \sqcap \mu_{\tilde{B}(x)} = \int_{v\in[0,1]} \int_{\omega\in[0,1]} \ell_x(v) \star \hbar_x(\omega)/(u = v \wedge \omega) \qquad (3.28)$$

where $\sqcap$ denotes the intersection operation, $\star$ indicates the product or minimum (also called t-norm) operation, and $\wedge$ indicates the minimum (also called t-norm). The intersection of primary membership degrees (represented by v and ω for GT2 fuzzy sets $\tilde{A}$ and $\tilde{B}$ respectively) is computed for all values of $x \in X$.

In simple terms, (3.28) outlines the intersection of two GT2 fuzzy sets using the extension principle such that for each $x \in X$:

1. $u = v \wedge \omega$ entails that for all possible combinations of the primary membership degrees (v and ω) the secondary degree $\mu_{(\tilde{A}\cap\tilde{B})_x}$ is computed by the minimum operation (or t-norm, denoted

as $\star$ in (3.28)) between the corresponding secondary degrees i.e. $\ell_x(v) \star \hbar_x(\omega)$.

2. In case of a duplicate point(s) for primary membership degree after the calculation of $v \wedge \omega$, then the point (primary membership degree) with the largest secondary membership degree is selected (just like the union operation outlined previously).

In Fig. 3.9 (c) an illustration for the intersection of two GT2 fuzzy sets $\tilde{A}$ and $\tilde{B}$ using extension principle (with vertical slice at $x = 2$) is presented. The corresponding calculations are outlined as follows:

$$\begin{aligned}
\mu_{(\tilde{A}\cap\tilde{B})_{x=2}}(u) =& \mu_{\tilde{A}(x=2)} \sqcap \mu_{\tilde{B}(x=2)} \\
=&(0.4/0.2 + 0.5/0.3 + 0.5/0.4) + (0.5/0.4 + 0.6/0.5 + 0.6/0.6) \\
=&\frac{\text{min}(0.4, 0.5)}{0.2 \wedge 0.4} + \frac{\text{min}(0.4, 0.6)}{0.2 \wedge 0.5} + \frac{\text{min}(0.4, 0.6)}{0.2 \wedge 0.6} + \frac{\text{min}(0.5, 0.5)}{0.3 \wedge 0.4} + \frac{\text{min}(0.5, 0.6)}{0.3 \wedge 0.5} + \\
&\frac{\text{min}(0.5, 0.6)}{0.3 \wedge 0.6} + \frac{\text{min}(0.5, 0.5)}{0.4 \wedge 0.4} + \frac{\text{min}(0.5, 0.6)}{0.4 \wedge 0.5} + \frac{\text{min}(0.5, 0.6)}{0.4 \wedge 0.6} \\
=&0.4/0.2 + 0.4/0.2 + 0.4/0.2 + 0.5/0.3 + 0.5/0.3 + 0.5/0.3 + 0.5/0.4 + 0.5/0.4 + 0.5/0.4 \\
=&\text{max}(0.4, 0.4, 0.4)/0.2 + \text{max}(0.5, 0.5, 0.5)/0.3 + \text{max}(0.5, 0.5, 0.5)/0.4 \\
=&0.4/0.2 + 0.5/0.3 + 0.5/0.4
\end{aligned} \tag{3.29}$$

3.3.5 Defuzzification of GT2 Fuzzy Sets

The defuzzification of a GT2 fuzzy set, $\tilde{A}$, is usually undertaken using a z-slice representation [134] (also termed as horizontal-slice representation or an α-cut representation)). Fig. 3.10 illustrates the z-slice representation of a GT2 fuzzy set at z-slice locations of $z = [0, 0.3, 0.6, 1]$. Since a z-slice of a GT2 fuzzy set gives an IT2 fuzzy set at that z-location, the centroid of a GT2 fuzzy set can be written as the union of the centroids of the IT2 fuzzy sets. In mathematical notation, the centroid of a GT2 fuzzy set can be written as follows [124]:

$$C_{\tilde{A}}(x) = \bigcup_{z \in [0,1]} z/[c_l(z), c_r(z)] \tag{3.30}$$

where $C_{\tilde{A}}(x)$ is a T1 fuzzy set composed of the centroids of the IT2 fuzzy sets at z-levels $z \in [0,1]$. The defuzzification of a GT2 fuzzy set using z-slices is outlined in Algorithm 1.

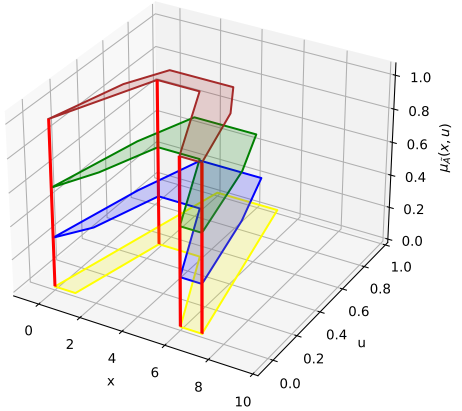


Figure 3.10 A z-slice representation of a General Type-2 (GT2) fuzzy set $\tilde{A}$.

3.4 Pattern-based Models as Classifiers

FLS can be used to solve classification problems which aim to predict a class for a given data instance. For FLS, the class is predicted using the patterns that define the relation between the input (antecedents) and output (consequent) variables of a given process. A generic nomenclature for the patterns is outlined earlier in (3.1). The motivation for solving a classification problem using FLS is that FLS provide explainable classification mechanism (in the form of patterns) whilst giving statistically significant classification accuracy ([135]).

```
Algorithm 1: The defuzzification of a GT2 fuzzy set, $\tilde{A}$.
  Result: The centroid, $x^*$.
    1. For z-slices of a GT2 fuzzy set $\tilde{A}$ at $[z_i, z_2, ...., z_I]$ for a total of $I$ slices such that $z_i \in [0, 1]$
    2. Compute the left and right centroids of the IT2 fuzzy set at each z-slice, i.e. $c_l(z_i)\ and\ c_r(z_i)$,
       using the KM algorithm outlined in Appendix A.
    3. Using centroid average compute one representative left and right centroid, i.e.
       $c_l(x)\ and\ c_r(x)$, as shown below:
         • $c_l(x) = \frac{z_1 * c_l(z_1) + ... + z_I * c_l(z_I)}{z_1 + ... + z_I}$
         • $c_r(x) = \frac{z_1 * c_r(z_1) + ... + z_I * c_r(z_I)}{z_1 + ... + z_I}$
    4. The centroid, $x^*$, can now be calculated by finding the mean of $c_l(x)\ and\ c_r(x)$ :
       $x^* = \frac{c_l(x) + c_r(x)}{2}$
```

In this work, the application of FLS is for solving classification problem for functional brain studies in line with the DCN frameworks (Chapter 2.1). Hence, the rest of the chapter outlines the preliminaries and the method for solving a classification problem using FLS. In general, for implementing a classification problem using FLS, the following information is required:

1. Input features.
   - The total number of inputs such as 10 fNIRS channels.
   - The statistical feature for each input such as, for example, mean of fNIRS signals for channels 1-3 and amplitude of fNIRS signals for channels 4-10.
2. Number of conceptual labels per feature.
   - For example the mean of an fNIRS signal can be defined using three conceptual labels such as Low, Medium, and High.
3. Membership Function (MF) per conceptual label.
   - The MF shape and parameters for each conceptual label such as triangular MF with coordinates on x-axis and y-axis at (0,0), (1,2) and (2,0).
4. Patterns.
   - The total number of patterns.
   - The maximum number of antecedents per pattern.
   - The antecedents and consequent of each pattern.
   - The dominance score of each pattern.

Some of the aforementioned information can be optimised using the input data (such as selection of input features, pattern learning, and MF parameters optimisation). For the proposed xMVPA (presented in Chapter 5.2) the learning of the aforementioned components for FLS is presented in detail in Chapter 5.2.4.

In the next sections, I outline FLS based on the different type of fuzzy sets i.e. T1-FLS, IT2-FLS, and GT2-FLS.

3.4.1 Classification using T1 Fuzzy Logic System (T1-FLS)

The patterns in an T1 FLS take the following form (3.31):

$$\begin{gathered} Pattern\ P_q : IF\ x_1\ is\ A^q_{1,l}\ ...\ AND\ x_\Psi\ is\ A^q_{\Psi,l}\ ...\ AND\ x_\Phi\ is\ A^q_{\Phi,l} \\ THEN\ class\ is\ O_k\ with\ DS_q \end{gathered} \tag{3.31}$$

where x_s are the numeric values (oxy-Hb or deoxy-Hb values) for the variable (or antecedent) Ψ (such as fNIRS channels) with a total of Φ variables i.e. $\Psi \in [1, ..., \Phi]$, $A^q_{\Psi,l}$ is the antecedent T1 fuzzy set for the Ψ^{th} variable representing the conceptual label l, where $l \in [1, ..., W]$ for a total of W conceptual labels, for the pattern number q. The output class (such as stimulus) of pattern number q is denoted by O_k where $k \in [1, ..., K]$ with K as the total number of output classes. The dominance score (DS) for the pattern number q is denoted by DS_q.

The rest of the section outlines the method for solving a classification problem using a T1-FLS. In a T1-FLS, all input and output fuzzy sets are T1, see Fig. 3.1.

1. **Degree of Membership Function (MF):** The membership degree, μ_A, for a given data instance, x, using T1 MFs, A, is as outlined in (3.2) and pictorially illustrated in Fig. 3.3.

2. **Firing Strength:** The firing strength, denoted $w_q(x)$, determines the similarity between a given data instance x and pattern P_q where q is the pattern number. It can be defined mathematically as follows:

$$w_q(x) = \prod_{\Psi=1}^{\Phi} \mu_{\mathrm{A}^{\Psi}}(x^{\Psi}) \tag{3.32}$$

where Φ is the total number of antecedents in the pattern P_q.

3. **Dominance Score:** The dominance score (DS) is a measure of the strength of a pattern, P_q, in correctly predicting a label for a given data instance. It is the product of pattern confidence, $conf_q$, and support sup_q:

$$DS_q = conf_q * sup_q \tag{3.33}$$

The pattern confidence and support are defined next.

   (a) **Pattern Confidence:** The confidence, $conf_q$, of a pattern, is an indication of the likelihood of a pattern for correctly classifying a data instance. It is computed using

$$conf_q(\Psi_q \Rightarrow Y_q) = \frac{\sum_{x \in (\Psi_q \Rightarrow Y_q)} w_q(x)}{\sum_{q=1, x \in \Psi_q}^{Q} w_q(x)} \tag{3.34}$$

   where $w_q(x)$ is the firing strength, as outlined in (3.32) for a data instance x, Ψ_q is the antecedent(s), and Y_q is the consequent of the pattern P_q with Q as the total number of patterns.

   (b) **Pattern Support:** The support, sup_q, of a pattern, P_q, is an indication of the coverage of training dataset by the pattern. It is computed using

$$sup_q(\Psi_q \Rightarrow Y_q) = \frac{\sum_{x \in (\Psi_q \Rightarrow Y_q)} w_q(x)}{Q} \tag{3.35}$$

where $w_q(x)$ is firing strength of pattern P_q for a data instance x, Ψ_q is the antecedent(s), and Y_q is the consequent of the pattern P_q with Q as the total number of patterns.

4. **Association Degree:** The class for a given data instance is determined using the metric of association degree. The pattern with the highest association degree predicts the class for a given data instance. The association degree, h_q, of a pattern P_q with a given data instance, x, is computed as outlined in (3.36).

$$h_q(x) = w_q(x) \cdot DS_q \tag{3.36}$$

where $w_q(x)$ is the firing strength as calculated in (3.32) and DS_q is the dominance score as calculated in (3.33).

The method for solving a classification problem using T1-FLS is presented above, however, the proposed method xMVPA (Chapter 5.2) is implemented using IT2 FLS which are presented next. In the next section, classification problem using IT2 FLS is outlined.

3.4.2 Classification using IT2 Fuzzy Logic System (IT2-FLS)

In this section, the method for solving a classification problem using IT2 FLS is outlined. The proposed XAI method (xMVPA presented in Chapter 5.2) for the classification of infants' fNIRS data is based on IT2 FLS. The IT2 fuzzy sets are a trade-off between the simplistic T1 fuzzy sets and the computationally expensive GT2 fuzzy sets. IT2 offer greater uncertainty handling than T1 fuzzy sets whilst being computationally simpler than the 3D GT2 fuzzy sets. The rules in an IT2 FLS take the following form (3.37):

$$\begin{gathered} Pattern\ P_q : IF\ x_1\ is\ \tilde{A}^q_{1,l}\ ...\ AND\ x_\Psi\ is\ \tilde{A}^q_{\Psi,l}\ ...\ AND\ x_\Phi\ is\ \tilde{A}^q_{\Phi,l} \\ THEN\ class\ is\ O_k\ with\ DS_q \end{gathered} \tag{3.37}$$

where x_s are the crisp brain activity values for the variable Ψ (fNIRS channel) with a total of Φ variables i.e. $\Psi \in [1, ..., \Phi]$, $\tilde{A}^{q}_{\Psi,l}$ is the IT2 fuzzy antecedent set for the Ψ^{th} variable representing the conceptual label l, where $l \in [1, ..., W]$ for a total of W conceptual labels, for the pattern number q. The output class of pattern P_q is denoted by O_k where $k \in [1, ..., K]$ with K as the total number of output classes.

1. **Degree of Membership Function (MF):** The conceptual labels, defined using IT2 fuzzy concepts Ã [120] can be written in mathematical notation as follows in eq. (3.38).

$$\begin{aligned} \tilde{\mathrm{A}} = \{(x, u, 1) | \forall x \in X, \\ \forall u \in [\underline{\mu}_{\tilde{\mathrm{A}}}(x), \overline{\mu}_{\tilde{\mathrm{A}}}(x)] \subseteq [0, 1]\} \end{aligned} \tag{3.38}$$

   where $\mu_{\tilde{\mathrm{A}}}$ represent the membership degree function of IT2 fuzzy concept Ã.

2. **Firing Strength:** The firing strength, $w_q(x)$, of pattern, P_q, for a data instance, x, is a measure of the degree of match between the pattern and the data instance. It is computed as outlined in (3.39).

$$\begin{aligned} \overline{w}_q(x) = \prod_{\Psi=1}^{\Phi} \overline{\mu}_{\tilde{\mathrm{A}}^{\Psi}}(x^{\Psi}) \\ \underline{w}_q(x) = \prod_{\Psi=1}^{\Phi} \underline{\mu}_{\tilde{\mathrm{A}}^{\Psi}}(x^{\Psi}) \end{aligned} \tag{3.39}$$

   where $\Psi \subseteq \{1, ..., \Phi\}$, where Φ is the total number of antecedents.

3. **Dominance Score:** The dominance score (DS) is a measure of a given pattern's strength and is computed as shown in (3.40).

$$\overline{DS}_q = \overline{conf}_q \times \overline{sup}_q \tag{3.40}$$

$$\underline{DS}_p = \underline{conf}_q \times \underline{sup}_q$$

where $conf$ is the confidence and sup is the support of the q^{th} pattern.

(a) **Pattern confidence:** The confidence, $conf_q$, of a pattern, P_q, is an indication of the likelihood of a pattern for correctly classifying a data instance. It is computed using

$$\begin{aligned} \overline{conf}_q(\Psi_q \Rightarrow Y_q) &= \frac{\sum_{x \in (\Psi_q \Rightarrow Y_q)} \overline{w}_q(x)}{\sum_{q=1, x \in \Psi}^{Q} \overline{w}_q(x)} \\ \underline{conf}_q(\Psi_q \Rightarrow Y_q) &= \frac{\sum_{x \in (\Psi_q \Rightarrow Y_q)} \underline{w}_q(x)}{\sum_{q=1, x \in \Psi}^{Q} \underline{w}_q(x)} \end{aligned} \tag{3.41}$$

where $\overline{w}_q(x)$ and $\underline{w}_q(x)$ are the upper and lower firing strength, as outlined in eq. (3.39), for pattern P_q on a data instance x, Ψ_q is the antecedent(s), and Y_q is the consequent of the pattern P_q with Q as the total number of patterns.

(b) **Pattern support:** The support, sup_q, of a pattern, P_q, is an indication of the coverage of training dataset by the pattern. It is computed using (3.42).

$$\begin{aligned} \overline{sup}_q(\Psi_q \Rightarrow Y_q) &= \frac{\sum_{x \in (\Psi_q \Rightarrow Y_q)} \overline{w}_q(x)}{Q} \\ \underline{sup}_q(\Psi_q \Rightarrow Y_q) &= \frac{\sum_{x \in (\Psi_q \Rightarrow Y_q)} \underline{w}_q(x)}{Q} \end{aligned} \tag{3.42}$$

where $\overline{w}_q(x_{i,t})$ and $\underline{w}_q(x)$ are the upper and lower strengths of activation for pattern P_q on a data instance x, Ψ_q is the antecedent(s), and Y_q is the consequent of the pattern P_q with Q as the total number of patterns.

4. **Association Degree:** The association degree, h_q, of pattern P_q with each data instance is computed as outlined in 3.43).

$$\begin{aligned}\overline{h}_q(x) &= \overline{w}_q(x) \cdot \overline{DS}_q \\ \underline{h}_q(x) &= \underline{w}_q(x) \cdot \underline{DS}_q \\ h_q(x) &= \frac{\overline{h}_q(x) + \underline{h}_q(x)}{2}\end{aligned} \tag{3.43}$$

In the next section, General Type-2 (GT2) fuzzy sets and systems are presented. The GT2 fuzzy sets are computationally extensive (in comparison to T1 and IT2 fuzzy sets) on account of being 3D.

3.4.3 Classification using GT2 Fuzzy Logic System (GT2-FLS)

In this section, the method for solving a classification problem using GT2 FLS is outlined. The rules in an GT2 FLS (are similar to those for IT2 FLS) take the following form (3.44):

$$\begin{gathered}Pattern\ P_q : IF\ x_1\ is\ \tilde{A}^q_{1,l}\ ...\ AND\ x_\Psi\ is\ \tilde{A}^q_{\Psi,l}\ ...\ AND\ x_\Phi\ is\ \tilde{A}^q_{\Phi,l} \\ THEN\ class\ is\ O_k\ with\ DS_q\end{gathered} \tag{3.44}$$

where x_s are the crisp brain activity values for the variable Ψ (fNIRS channel) with a total of Φ variables i.e. $\Psi \in [1, ..., \Phi]$, $\tilde{A}^q_{\Psi,l}$ is the antecedent GT2 fuzzy set for the Ψ^{th} variable representing the conceptual label l, where $l \in [1, ..., W]$ for a total of W conceptual labels, for the pattern number q. The output class of pattern P_q is denoted by O_k where $k \in [1, ..., K]$ with K as the total number of output classes.

The GT2 fuzzy sets are 3D and the associated computations with GT2 fuzzy sets are more extensive since for each $x \in X$ the primary membership degree is a T1 MF. Consequently, and as previously outlined, for union and intersection of GT2 fuzzy sets extension principle involving vertical slices of GT2 is carried out (see Section 3.3.5), and for defuzzification z-slices (or horizontal slices are taken) (see Section 3.3.5).

For solving a classification problem using GT2 fuzzy set, in this work, the z-slice approach [134] is undertaken that type-reduces the GT2 fuzzy set to an IT2 fuzzy set at each z location (see Fig. 3.10). Therefore, for pre-determined z-locations, the method outlined for IT2 fuzzy sets in Section 3.4.2 can be repeated. The final stage involving the classification problem involves defuzzification of the association degree for each pattern. The defuzzification of the association degree can be carried out using Algorithm 1. The overall method for solving a classification problem using a GT2 FLS is outlined in Algorithm 2.

Algorithm 2: The classification using GT2 FLS using z-slice representation.

```
Result: The output class.
For z-slices at predefined levels $z_i \in$
 $[z_1, ..., z_I]$ and a given data instance $x$, the output class for $x$ can be determined as follows:
 for $z_i < z_I$ do
     Compute membership degree as outlined in (3.38).
     Compute firing strength as outlined in (3.39).
     Compute dominance score (DS) as outlined in (3.40).
     Compute association degree as outlined in (3.43).
Defuzzify the association degree using Algorithm 1 for each pattern.
The pattern with the highest association degree predicts the output class.
```

In the next chapter an overview of the AI methods most commonly used for fNIRS based studies (with adults and infants) is presented.

Chapter Four

Artificial Intelligence (AI) for Neuroscience

The[2] generic field of cognitive neuroscience investigates the underlying brain functional mechanisms that subserve cognitive processes such as memory, perception, understanding, and reasoning [136]. 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. This is because all the pre-processing stages of acquired neuroimaging data (fNIRS) would be similar. Moreover, AI techniques that can identify the difference in brain activation patterns for adults should also be able to decode the same in infants' neuroimaging data analysis. As opposed to cognitive neuroscience for 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 [137] 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

[2] Some parts of the text in this chapter have been published here: M. Kiani, J. Andreu-Perez, H. Hagras, S. Rigato, and M. L. Filippetti, "Towards Understanding Human Functional Brain Development With Explainable Artificial Intelligence: Challenges and Perspectives," *IEEE Computational Intelligence Magazine*, 2021, In Press © 2022 IEEE.

learning [138] 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 [139]. 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 [138].

The overarching aim of this chapter is to investigate the potential and limitations of these algorithms, as applied to infant 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 I review AI methods with application(s) to fNIRS, 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.

4.1 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 (such as deep neural networks) or simply partial information, and others can derive some explainable structure with respect to describing the underlying brain activity. Towards this end, I will be reviewing the extent AI methods can explain their underlying mechanisms to shed light on the processes of functional brain development. 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.

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

4.1.1 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 [53]. Two complementary, non-explainable AI methods namely Ridge Regression (RR), and Interpolated Functional Manifold (IFM) used with fNIRS connectivity measures are reviewed next.

A fNIRS study investigating intrinsic FC of cortical networks to predict anxiety states using linear RR models was carried out by Duan *et al.* [140]. The resting state FC was calculated using Pearson correlation coefficient for 1035 edges between 46 nodes (fNIRS channels). The RR model was able to predict the anxiety score with statistical significance using the connectivity of cortical networks. The mean square error of their model was 122.04 with correlation coefficient of 0.36.

The prowess to predict anxiety state 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 (see (2.2)) in terms of FC is significantly limited (it is hard to interpret 1035 values and relate the brain activation). Hence, despite obtaining statistically significant results, the FC analysis could not shed significant light on the resting state cortical networks.

4.1.2 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.* [141]. 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 Radial Basis Function (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 were projected to the edges of the manifold, whereas more experienced participants (trainees and consultants) fNIRS response 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 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 study quantitatively 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. 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 for JI.

4.1.3 RepL with fNIRS using Convolutional Neural Networks (CNN)

Many recent works in neuroscience are increasingly using deep leaning paradigms to investigate the underlying brain activity in response to a presented task [142]. 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 [142]. 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 [142]. 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 + Rectified Linear Unit (ReLU).
- Pooling.

(ii) Classification/Regression Blocks

- Fully Connected Layers.
- Softmax, Logistic regression layer, regression loss (Root Mean Square Error (RMSE) etc.)

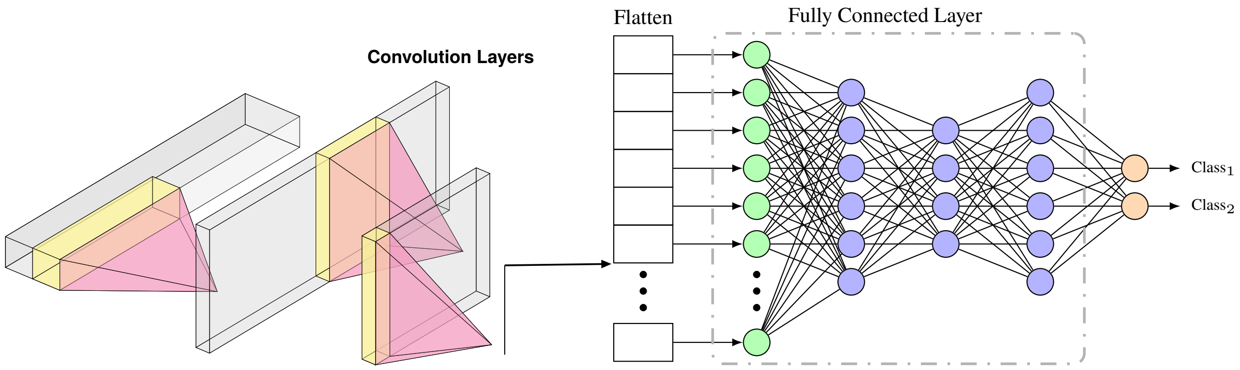
The performance of CNNs is critically dependent on the optimisation of hyperparameters, and owing to the large number of hyperparameters that need optimisation, most DNNs, including CNNs, require large datasets to converge. The hyperparameters of a CNN include the size of the filter(s), stride, number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

In this section, I 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 [143]. The advancements in BCI have enabled people with neuromuscular disorders to restore or replace some of their motor functions such as limb movements [144]. 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. Since the user would typically know how to control their limb in the normal way, they would have to re-learn the control of their limb through BCI. 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 regards, 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. Hence, the CNNs powering BCI can shed light into the neural reuse and perceptual narrowing to perform BCI.

Next, I review one of the most promising studies for fNIRS-based BCI application [145]. 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.* [145] the authors investigated the possibility of classification of four



(a) Feature Extraction **(b)** Classification Mechanism

Figure 4.1 A general schematic of a CNN for representation learning, © 2022 IEEE.

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 layers with 1 pooling layer in between of the following sizes- Convolution1: 3×23×16; Pooling: 2×12×16 and Convolution2: 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 (MLP)); 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). 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). In addition, the limited explainability of CNNs in terms of the underlying brain activity critically limits its applicability to inform functional brain development. In the next section, I review the application of AI in DCN.

4.2 AI in Developmental Cognitive Neuroscience (DCN)

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 to apply models used in adult research to DCN, however this entails making some assumptions (such as the shape of HRF, see Chapter 2.2.4). 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 [101] which is a correlation based MVPA). In fact, there is an evident scarcity for undertaking AI methods in DCN research.

In the following section, I review the study by [101] that undertakes MVPA with fNIRS using correlation.

4.2.1 MVPA with fNIRS using correlation

An fNIRS based MVPA is aimed at spatial investigations into the cortical regions' activations encoded in the MVM. The work by [101] 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 the use of MVPA that improved the decoding sensitivity in comparison to their previous work using univariate methods [146]. Also 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 (PFC) (see Fig 6.5 (a)) to be the most critical channels for decoding between visual and auditory stimuli (as noted in Table 6.3).

The identification of fNIRS channels, and their corresponding anatomical locations, sheds light on the localised activation of the cortex as described 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 of infants' fNIRS data by [101] and to overcome its limitation of partial explainability, I designed an explainable MVPA (xMVPA), outlined in Chapter 5.2. The proposed xMVPA is able to explain its inference mechanism in terms of the IS frameworks (presented in Chapter 2.1); thereby informing about the localisation and specialisation of the cortical regions for the processing of presented information. The xMVPA is designed for cross-sectional fNIRS data, i.e. it can analyse fNIRS data that is collected from infants of similar age. In this way, the explainable results provided by xMVPA can shed light on the localisation, and specialisation of the cortical networks formed for the processing of the presented information in infants of a particular age group.

In the next chapter, I present the new XAI methods developed for the analysis of functional brain development.

Chapter Five

Proposed Explainable AI (XAI) Methods

In this chapter the proposed XAI methods are outlined. The motivation for the development of the the XAI methods is to be be able to describe functional brain development as hypothesised by the DCN frameworks of IS and neural reuse (previously outlined in Chapter 2.1). The proposed methods overcome the limitations of prevalent data analysis techniques in DCN studies (previously discussed in Chapter 2.2) such that they 1) are not dependent on large datasets, 2) do not rely on a priori models, and 3) provide an explanation for their classification process.

More specifically, the XAI methods are:

1. Effective Fuzzy Cognitive Maps (EFCMs).

2. Explainable Multivariate Pattern Analysis (xMVPA).

3. Time-dependent eXplainable Artificial Intelligence (TXAI) system.

The rest of the chapter outlines the aforementioned XAI methods in detail.

5.1 Effective Fuzzy Cognitive Maps (EFCMs)

Effective Fuzzy Cognitive Maps (EFCMs)[3] are derived from Fuzzy Cognitive Maps (FCMs) and based on graph theory. EFCM is capable of representing *fuzzy connectivity* between variables (such as fNIRS channels) as fuzzy degrees of relationships (such as effective connectivity (EC)) in a complex system (such as functional brain development). More specifically, the EFCMs propose a partial explainable model in terms of estimating the EC as fuzzy weights between its concepts (fNIRS channels). In this regard EFCMs, when applied to fNIRS based DCN studies for estimating EC, can shed light on how cortical networks evolve in terms of their influence (EC) on account of emerging cognitive functions as hypothesised by the neural reuse framework (previously outlined in Chapter 2.1).

A mathematical formulation for EFCMs is outlined as follows in (5.1):

$$\chi_j(\tau+1) = \varrho(\sum_{k=0,k\leq\tau}^{\wp}(y_j(k)\sum_{i=1}^{\mathbf{M}}(\nu_{ij}\odot\vartheta_{ij})\chi_i(\tau-\wp))) \tag{5.1}$$

where $\chi_j(\tau)$ is the fuzzy value of the concept or node (or signal in fNIRS paradigm) j at time τ based on the strength and direction of the interaction (EC value) in the range $[-1,1]$, $\nu_{ij}\odot\vartheta_{ij}$, where ($\odot$ represent element wise multiplication), of concept j with other concepts in the system, and the past values of the concept j, $\chi_j(\tau-\wp)$. The total number of concepts (fNIRS channels) in a given system is represented by $\mathbf{M}$. The number of past fuzzy values of a given concept to be considered by the EFCM when computing the new fuzzy value for that concept is dictated by the order of the EFCM, represented by $\wp$. The EFCM also allows for tuning the strength of influence of the preceding fuzzy values on the current fuzzy value for a given concept by optimizing values of $y_j(k)$, $\vartheta_{ij}\in\mathbf{Z}$, and

[3] Some parts of the text in this section have been published here: 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, Early Access © 2019 IEEE.

$\nu_{ij} \in \mathbf{D}$ using Genetic Algorithm (GA).

The transformation function, denoted ϱ in (5.1), is based on the sigmoid function as outlined in (5.2).

$$\varrho(x) = 2sigmoid(2x) - 1 \tag{5.2}$$

$$sigmoid(x) = \frac{1}{1 + exp^{-\iota x}} \tag{5.3}$$

where ι is a parameter used to define a particular shape of the sigmoid function. The most common value of ι found in literature is 5 [147, 148].

The proposed regularisation of EFCM is achieved by straining the weights of the connection matrix $\mathbf{D}$, using a soft regulariser, $\mathbf{Z}$. The soft regulariser $\mathbf{Z}$ is a generalisation of a dropout mask since the elements of $\mathbf{Z}$ can attain any value in the range of $[0, 1]$. The elements $\nu_{ij} \in \mathbf{D}$ and $\vartheta_{ij} \in \mathbf{Z}$ are both optimised using GA, however, only elements of $\mathbf{D}$ can model the direction of the EC whereas $\mathbf{Z}$ can only optimise the strength of the EC, and hence cannot affect the direction of EC unlike $\mathbf{D}$.

A quantitative assessment of the predicted states, $\chi_j(\tau + 1)$ in (5.1), generated by the EFCM can be done using the two standard FCM error functions $error_1$ and $error_2$ outlined in (5.4) and (5.5) [147], respectively. The error is estimated by comparing the reconstructed signals from the resultant EFCM model with the original signals. In this work, $error_2$ (5.5), is computed for quantifying the performance of the proposed EFCM technique for evaluating the EC in Chapter 6.1 since it also accounts for the rate at which the state of a concept is changing.

$$error_1 = \sum_{\tau=1}^{T} \sum_{i=1}^{N} |\chi_i(\tau) - \hat{\chi}_i(\tau)| \tag{5.4}$$

$$error_2 = \sum_{\tau=1}^{T} \sum_{i=1}^{\mathbf{M}} |(C_i(\tau) - \hat{\chi}_i(\tau))^2 + (\chi'_i(\tau) - \hat{\chi'}_i(\tau))^2| \tag{5.5}$$

where $\hat{\chi}$ is the predicted state of the concept of resultant FCM, and χ' is the rate of change in the state of a concept.

In comparison with standard FCM, EFCMs optimises the strength (scalar magnitude without direction) and direction separately rendering it (EFCM) with more degrees of freedom to find the optimum values of EC, i.e $\nu_{ij} \odot \vartheta_{ij}$ in eq (5.1). In addition, EFCM also undertakes tuning of the transformation (sigmoid) function, ϱ in eq (5.1), to optimise how fast the non-normalised fuzzy degrees of relationship are squeezed into the normalised range for the fuzzy degrees of relationship. Further, EFCM establishes that for deciphering EC in a complex process such as brain cortical networks, a higher order of EFCM, $\wp$, is more suited to discern the underlying EC. In the next sections, the aforementioned revision in EFCM namely: 1) Transformation function tuning, 2) EFCM order, and 3) EFCM learning are outlined in detail.

5.1.1 Transformation Function Parameter Tuning

A transformation function is responsible for normalizing the fuzzy degrees of relationships (or EC values) in a specified range. The particular transformation function used in this work is the sigmoid function, see (5.2). Essentially, the gradient of the transformation function determines how fast the non-normalized fuzzy degrees of relationship are squeezed into the normalized range for the fuzzy degrees of relationship.

Fig. 5.1 shows the transformation function plot for two values of the parameter, ι, that defines the gradient of the function. The value for ι is assumed to be 5 for most practical applications [147, 148]. However, with $\iota = 5$, the gradient of the transformation function proved too steep to suit the needs for this work- the fuzzy degrees of relationship were being squashed to the normalised range too fast, and hence was not able to discern the underlying EC.

In contrast, the empirically found value of $\iota = 1$, is more inclusive of the non-normalized fuzzy degrees of relationship to be translated to the non-extreme values in the normalized range for the fuzzy degrees of relationship by having a less steep gradient. This can also be seen in Fig. 5.1. The transformation function with $\iota = 1$ (red) is including values from approximately $(-4, 4)$ to be converted to non-extreme normalized equivalents whereas with $\iota = 5$, values from the approximate range of $(-1, 1)$ only are being translated to non-extreme counterparts in the normalized fuzzy degrees of relationship.

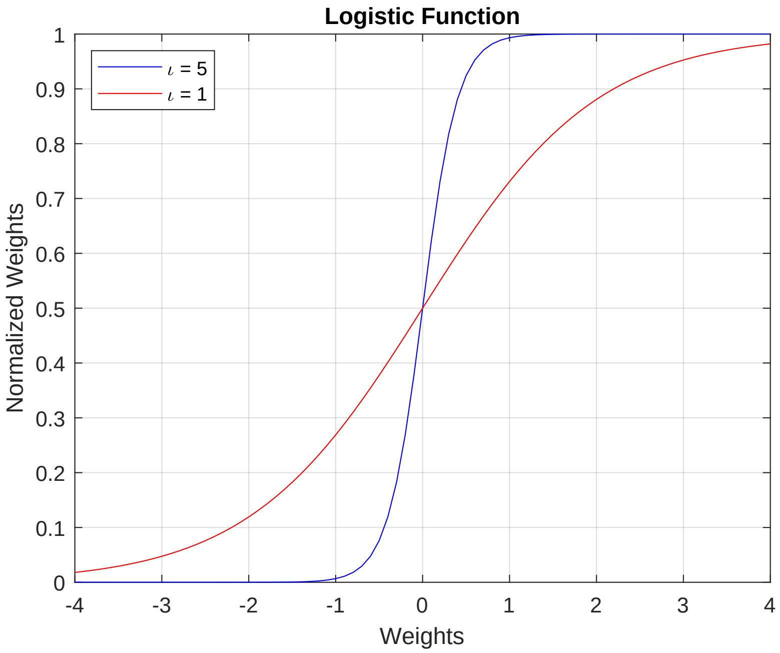


Figure 5.1 Transformation function with $\iota = 5$ (blue) and $\iota = 1$ (red), © 2019 IEEE.

5.1.2 EFCM Order

The EFCM order dictates how many past fuzzy values, $\wp$, in a set of observations will be considered by the EFCM when trying to discern a fuzzy effective connection between them. The implications of the choice of the order of the EFCM are largely dependent on the particular application. In (5.1), the parameter $\wp$ defines the order of the EFCM. As is also evident in (5.1), the impact of the preceding state(s) on the current state can also be scaled by tuning the value of the parameter y.

The motivation for employing a higher order EFCM for fNIRS signals lies in a more accurate deciphering of the fuzzy EC between the fNIRS signals. This is owing to complex, higher order fuzzy effective connections amongst the fNIRS signals which first order EFCM dynamics cannot comprehend very well.

5.1.3 EFCM Learning

The EFCM learning can be achieved either manually using the information provided by experts or by employing an automated process that can use historical information to develop FCMs [149]. There is an increasing trend to use computerised techniques to uncover the fuzzy relations between concepts for an FCM simulation [150] since using an automated approach to learn the inherent model of a given system does not introduce a bias in the FCM simulation that may be incorporated into the results had the FCM evolution been governed by human knowledge.

In accordance with the greater advantages of automated FCM learning, in this work, EFCM learning is achieved by utilising the Genetic Algorithm (GA). GA is an optimisation algorithm that is based on evolutionary ideas of natural selection and genetics and is capable of solving both constrained and unconstrained optimisation problems. Owing to its robust, and heuristic nature, GA can be applied to learn the inherent model of a given system using the historical data of the system [147].

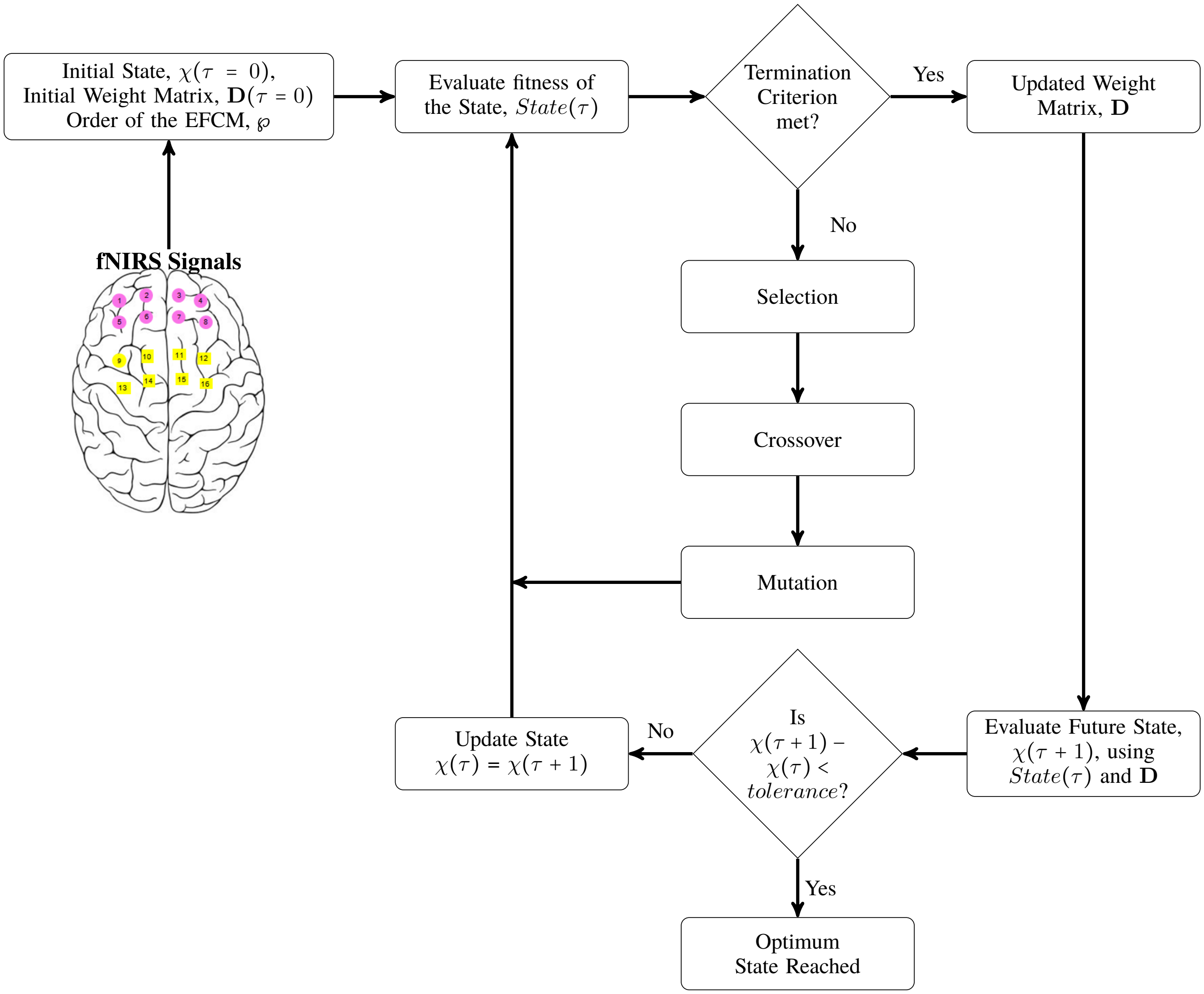

Initial State, $\chi(\tau = 0)$,
Initial Weight Matrix, $\mathbf{D}(\tau = 0)$
Order of the EFCM, $\wp$
Evaluate fitness of the State, $State(\tau)$
Termination Criterion met?
Yes
Updated Weight Matrix, $\mathbf{D}$
No
fNIRS Signals
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Selection
Crossover
Mutation
Update State $\chi(\tau) = \chi(\tau + 1)$
No
Is $\chi(\tau+1) - \chi(\tau) < tolerance$?
Evaluate Future State, $\chi(\tau + 1)$, using $State(\tau)$ and $\mathbf{D}$
Yes
Optimum State Reached

Figure 5.2 A flowchart of the algorithm for predicting EC in fNIRS signals using the proposed EFCM model, © 2019 IEEE.

The flowchart in Fig. 5.2 outlines the steps for the generation of a resultant connection matrix, $\mathbf{D} \odot \mathbf{Z}$, by GA using historical data to be incorporated in the EFCM simulation. The EFCM builds the next state vector, $\chi(\tau + 1)$, of a given system using the resultant connection matrix, $\mathbf{D} \odot \mathbf{Z}$, and the historical data, $\chi(\tau)$, till a chosen tolerance criterion is achieved.

$$fitness = \frac{1}{10(T-1)\mathbf{M}} error \tag{5.6}$$

where T is the length of each of the total $\mathbf{M}$ signals and the $error$ is computed using (5.5).

The fitness of each predicted state $\hat{\chi}(\tau)$ is gauged against an a priori termination criterion, and if achieved, the weight matrix $D(\tau)) \odot Z(\tau)$ is updated accordingly, see Fig 5.2. In case the termination criterion is not met, the GA will try to look for new offsprings using selection, crossover, and mutation, to generate a new predicted state $\hat{\chi}(\tau)$ which will then again be compared with the termination criterion, and the process will continue to repeat itself till the chosen termination criterion is satisfied.

In the next section, I present the second XAI mechanism based on interval type-2 (IT2) fuzzy logic system (FLS).

5.2 Explainable Multivariate Pattern Analysis (xMVPA)

In this section[4], an XAI inference mechanism for infants' fNIRS data based on IT2-FLS (previously outlined in Chapter 3.2) is presented. The proposed XAI inference mechanism will drive the MVPA (multivariate pattern analysis)[16], a technique first introduced for functional MRI data analysis with adults, and recently used for study the infant mind with fNIRS [101]. MVPA deciphers multiple

[4] Some parts of the text in this section have been published here: J. Andreu-Perez, L. L. Emberson, M. Kiani, M. L. Filippetti, H. Hagras, and S. Rigato, "Explainable Artificial Intelligence Based Analysis for Developmental Cognitive Neuroscience," *Commun Biol*, vol. 4, p. 1077, 2021 © 2021 Nature.

fNIRS channels activity simultaneously to identify informative differences in brain regions' activation in response to stimuli.

The proposed **eXplainable MVPA (xMVPA)** is an XAI inference mechanism for brain haemodynamics data that uses evolutionary learning procedure [151] to train the model that drives the MVPA. A generic nomenclature for the functional patterns learnt by the xMVPA, as defined in (5.7). The patterns are captured (or learnt) directly from the input fNIRS measurements. By identifying cortical networks activated for the processing of perceptual information, these patterns can pinpoint the emergence of the specialisation of different brain regions and their interactions, critically contributing to the existent literature of neurodevelopmental trajectories. A generic nomenclature of xMVPA patterns that map interactions among brain regions (antecedents) to particular stimuli (consequent) is as follows:

$$\begin{aligned} &\text{IF activity is } CoL \text{ in Ch. } 1 \text{ AND activity is } CoL \text{ in Ch. } 2 \ldots \\ &\qquad\qquad\qquad\qquad \text{THEN it corresponds to } Stimulus_A \end{aligned} \tag{5.7}$$

where CoL stands for a conceptual label that denotes the level of activity in a given channel (Ch.), such as *inactive*, *active*, or *very active*. More specifically, since the xMVPA is based on IT2 fuzzy sets, the xMVPA patterns can be written in terms of IT2 fuzzy sets as in (5.8):

$$\begin{aligned} &\text{Pattern } P_q: \text{ IF } x_1 \text{ is } \tilde{A}^q_{1,l} \text{ AND} \ldots AND\, x_\Psi\, is\, \tilde{A}^q_{\Psi,l} \ldots \text{ AND } x_\Phi \text{ is } \tilde{A}^q_{1,\Phi} \\ &\qquad \text{THEN stimulus is } Y_q \text{ with } DS_q \end{aligned} \tag{5.8}$$

where q is the pattern number, x_Ψ is the numeric brain activity value of fNIRS channel Ψ i.e. $\Psi \in [1, ..., \Phi]$, $\tilde{A}^q_{\Psi,l}$ is the antecedent IT2 fuzzy set for the Ψ^{th} variable representing the conceptual label l, where $l \in [1, ..., W]$ for a total of W conceptual labels with Φ as the total number of antecedents, Y_q is the consequent stimulus class for the pattern, and PS_q is the pattern score associated with the qth pattern.

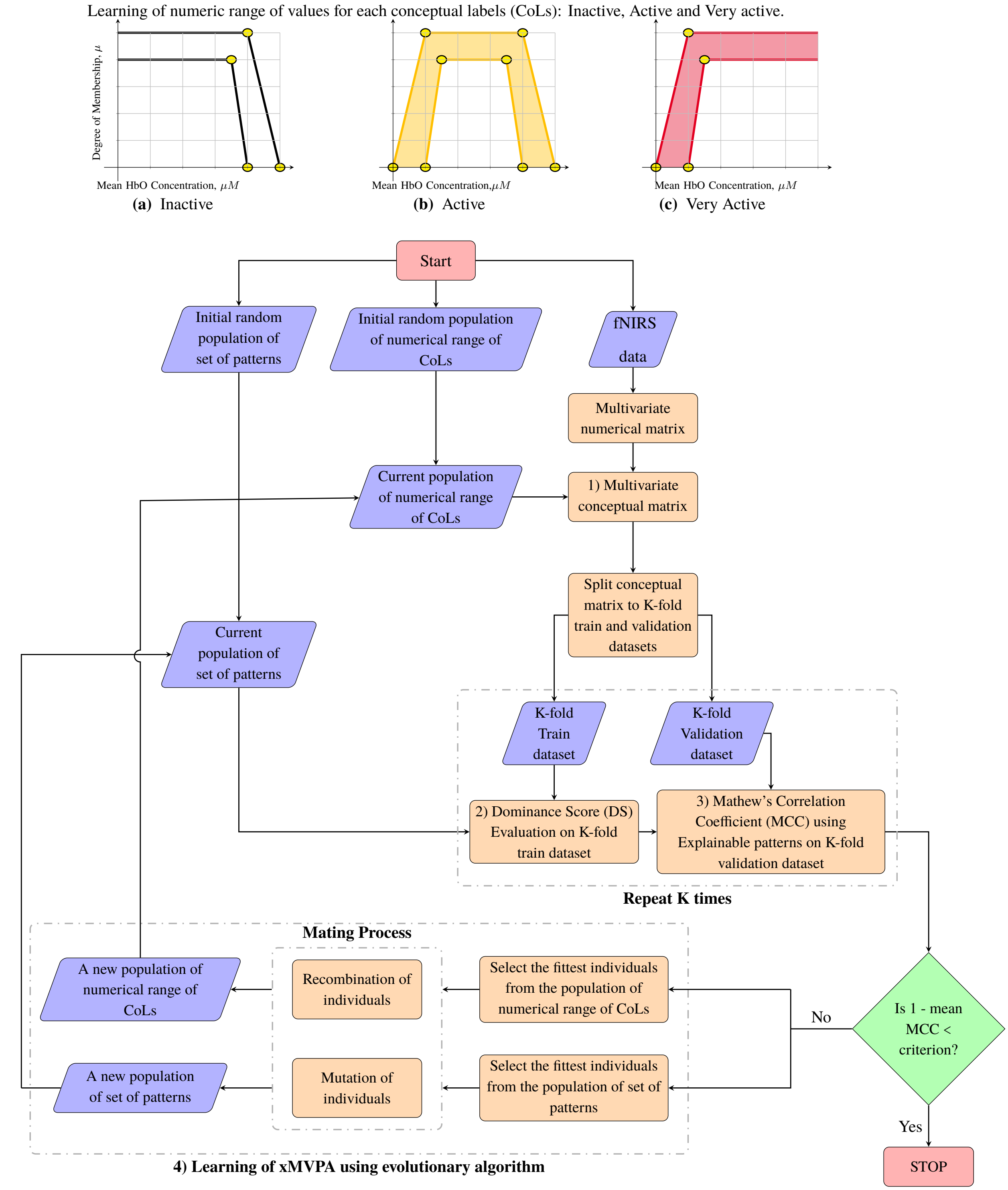
In general, xMVPA inference mechanism consists of the following integral processes:

1. Brain activation concept definition;
2. Pattern dominance score evaluation;
3. Matching of data with the stimulus by the explainable pattern;
4. Learning of xMVPA:
   (a) Learning of conceptual labels;
   (b) Learning of patterns.

In this work, the xMVPA learns directly from the data by splitting the the conceptual multivariate data into 5-fold disjoint training and validation datasets [152]. To ensure there is no bias in the selection of the training and validation datasets, the conceptual multivariate data is split into 5-fold disjoint training and validation datasets [152]. Also, please note that in the xMVPA inference mechanism there is no information flow from the learning of patterns from one training fold to another training fold. The interlinks between the different processes of the xMVPA inference mechanism are delineated in a flowchart in Fig. 5.3 (d). A description of each of these processes is provided next.

5.2.1 Conceptualisation of Brain Activation Levels

The xMVPA works on a MVM that has elements characterised by conceptual labels. For this, the numerical MVM formed by combining the data from all channels of interest is converted into a *conceptual* MVM. In the present work, the conceptual labels of *inactive*, *active*, and *very active* are used to represent the level of brain activity measured by a fNIRS channel. The conversion of numeric data into conceptual labels is based on the *numerical range of values* represented by each of the conceptual labels. The shape of the MFs for the conceptual labels is as outlined in Fig. 5.3 (a) -


Learning of numeric range of values for each conceptual labels (CoLs): Inactive, Active and Very active.
Degree of Membership, μ
Mean HbO Concentration, μM
Mean HbO Concentration,μM
Mean HbO Concentration, μM
Start
Initial random population of set of patterns
Initial random population of numerical range of CoLs
fNIRS data
Multivariate numerical matrix
Current population of numerical range of CoLs
1) Multivariate conceptual matrix
Split conceptual matrix to K-fold train and validation datasets
Current population of set of patterns
K-fold Train dataset
K-fold Validation dataset
2) Dominance Score (DS) Evaluation on K-fold train dataset
3) Mathew's Correlation Coefficient (MCC) using Explainable patterns on K-fold validation dataset
Repeat K times
Mating Process
A new population of numerical range of CoLs
Recombination of individuals
Select the fittest individuals from the population of numerical range of CoLs
No
Is 1 - mean MCC < criterion?
A new population of set of patterns
Mutation of individuals
Select the fittest individuals from the population of set of patterns
Yes
STOP
4) Learning of xMVPA using evolutionary algorithm

(a) Inactive **(b)** Active **(c)** Very Active

(d) A flowchart outlining the steps for the construction of xMVPA.

Figure 5.3 A schematic delineating the steps for the construction of xMVPA, © 2021 Nature.

(c). The numeric values to be learnt for the definition of *inactive* and *very active* conceptual labels are 4 each, while 8 numeric values need to be optimised for the trapezium shaped MF for *active* and marked with yellow circles in Fig. 5.3 (a) - (c). The range of numeric values for each conceptual label are learnt using an evolutionary algorithm with more details as outlined in the section 5.2.4.

5.2.2 Explainable Patterns' Dominance Score Evaluation

Starting with an initial random set of patterns, the $\overline{upper}$ and $\underline{lower}$ bounds of the dominance score, $\overline{DS}_q$ and $\underline{DS}_q$ respectively, for each of the patterns P_q in the set are determined on a given k-fold training dataset as shown in eq. (5.9) [151].

$$\begin{aligned} \overline{DS}_q &= \overline{conf}_q \cdot \overline{sup}_q \\ \underline{DS}_q &= \underline{conf}_q \cdot \underline{sup}_q \end{aligned} \tag{5.9}$$

where q is the pattern number, $\overline{conf}_q$ and $\underline{conf}_q$ is the upper and lower confidence score of the pattern P_q respectively, and $\overline{sup}_q$ and $\underline{sup}_q$ is the upper and lower support of the pattern P_q, on a training dataset.

The confidence score, $conf_q$, of a pattern, P_q, can be viewed as the possibility that a given data instance is an observation of this pattern, i.e. $conf_q$ is the likelihood of a given data instance to be a representative observation for the same stimulus as the pattern stimulus (consequent) Y_q, given the data instance has matching interactions of brain regions as the rule P_q, i.e. the same antecedents. The mathematical formulation for the computation of a pattern confidence is as provided in (3.41). The support, sup_q, of a given pattern is an indication of the coverage of training dataset by the pattern. The mathematical formulation for the computation of a pattern support is as provided in (3.42).

5.2.3 Stimulus Prediction

A set of optimal patterns, with corresponding dominance scores, DS_q, are obtained using an evolutionary search (section 5.2.4) guided by the results of a k-fold cross-validation (`k=5`) procedure. The most likely stimulus for a given data instance, where a data instance is a row (*i*) in the validation dataset, is achieved by evaluating the association of the data instance with all the patterns (pattern-based explanations). The stimulus response of a data instance is predicted as the consequent of the pattern with the highest association degree, i.e. visual or auditory stimulus.

The stimulus for each data instance in the validation dataset, x_i, is determined using the metric of association degree. The association degree, h_q, of pattern P_q with each data instance in the validation dataset, x_i, is computed as outlined in eq. (5.10).

$$\begin{aligned} \overline{h}_q(x_i) &= \overline{w}_q(x_i) \cdot \overline{DS}_q \\ \underline{h}_q(x_i) &= \underline{w}_q(x_i) \cdot \underline{DS}_q \\ h_q(x_i) &= \frac{\overline{h}_q(x_i) + \underline{h}_q(x_i)}{2} \end{aligned} \tag{5.10}$$

where $\overline{w}_q$ and $\underline{w}_q$ are the upper and lower firing strengths of a pattern P_q for a data instance of the validation set x_i. More information on the firing strengths is provided in (3.39). To summarise, a given validation data instance, x_i, is classified as a response to the stimulus, Y_q, corresponding to the pattern P_q with the maximum association degree with x_i.

5.2.4 xMVPA learning from data

Search of explainable patterns

The initial set of patterns used in the proposed xMVPA inference mechanism is randomly generated to ensure that there is no bias introduced in the learning of the set of patterns. An evolutionary GA is integrated in the xMVPA inference mechanism to identify those set of patterns that together give the best classification results. Fig. 5.3 (d) outlines the steps undertaken to reveal an optimised set of patterns using a given dataset.

All sets of patterns are learnt using k-fold cross-validation to establish the general ability of a given set of patterns. Using an initial random set of patterns with a total of Q patterns, the Mathew's Correlation Coefficient (MCC) of the set of patterns is computed as MCC gives a more balanced measure of the quality of binary (two-class) classifications. It is computed as shown in eq. (5.11) [19]:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.11}$$

where True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are as defined in the confusion matrix in Fig. 6.7 (a).

The cost of the set of patterns is computed as 1 - the mean of the MCCs of all k-fold validation datasets. The GA then compares the cost of the set of patterns with a pre-defined tolerance criterion. If the cost is greater than the tolerance of GA, the GA then populates a new set of patterns and the cycle is repeated till the tolerance criterion of the GA is met as outlined in Fig. 5.3 (d). More details on the GA are provided in Chapter 5.2.4.

To maximise the model interpretability, the total number of patterns to be learnt by xMVPA system is set at 20 patterns, with maximum of 3 channels interactions in a given pattern (as three-point messages are the recommended standard for science communications [153]). The evolutionary system [125] will aim to maximise the accuracy of prediction while using a maximum of 20 patterns, where each pattern consist of a maximum 3 antecedents. This renders the total number of variables, to be optimised for pattern learning, by GA to be: total number of patterns (20) * maximum number of channels (3) and conceptual label for each chosen channel (3: Inactive, Active or Very Active) and the corresponding stimulus class for each pattern (1) = 20*(3+3+1) = 140 variables.

Learning of conceptual labels

The number of parameters to be learnt for conceptual labels definition are the lower and upper numeric range of values for each concept. For a given channel, the number of variables that need to be learnt for the channel's equivalent conceptual labels numeric range are 16 (4 for inactive as shown in Fig. 5.3a, 8 for active as shown in Fig. 5.3b, and 4 for very active as shown in Fig. 5.3c). Hence, in this work, for ten channels the total number of variables to be optimized for conceptual labels numeric range are 16*10 = 160.

Hence, the grand total of variables to be learnt by the GA are 140+160 = 300 variables. The structure of each phenotype is described in eq. (5.12). The population size of GA, i.e. the number of feasible solutions is set at 200, with selection done using *tournament* and the GA tolerance is set at $1 * 10^{-5}$.

$$\begin{aligned}\rho^b = \{&\phi_1^1, \phi_2^1, \phi_3^1, \Upsilon_1^1, \Upsilon_2^1, \Upsilon_3^1, Y^1, ...,\\ &\phi_1^Q, \phi_2^Q, \phi_3^Q, \Upsilon_1^Q, \Upsilon_2^Q, \Upsilon_3^Q, Y^Q, ..., \qquad \text{‘}\\ &\gamma_{IA^j}^1, ..., \gamma_{IA^j}^4, \gamma_{Ac^j}^1, ..., \gamma_{Ac^j}^8, \gamma_{VA^j}^1, ..., \gamma_{VA^j}^4, ...,\\ &\gamma_{IA^n}^1, ..., \gamma_{IA^n}^4, \gamma_{Ac^n}^1, ..., \gamma_{Ac^N}^8, \gamma_{VA^n}^1, ..., \gamma_{VA^n}^4\}\end{aligned} \tag{5.12}$$

where ρ^b is the phenotype of an individual b (a potential solution) for the GA for a total of Q patterns. Each ϕ denotes a particular channel, each Υ represents the corresponding conceptual labels associated with each channel. These chromosomes form the antecedent of a pattern. The consequent of the patterns is denoted as Y. The γ represent the numeric values for the range of each of the conceptual labels of all the **F** fNIRS channels. In particular, $\gamma_{CoL^j}^{NV}$, subscript CoL (conceptual label) denotes the value of concept that can be *inactive*: IA, *active*: Ac, and *very active*: VA, along with the associated channel number j and the numeric value (NV) in the superscript: 4 NVs for *Inactive* and *Very active*, and 8 NVs for *Active*.

Evolutionary Learning of xMVPA

As previously outlined in Chapter 5.1.3, Genetic Algorithms (GA) are a type of evolutionary algorithm [125] that are based on the *survival of the fittest* phenomenon from Darwin's evolutionary theory. The 'survival of fittest' idea states that given limited resources for a population of individuals within some environment, a competition for those resources causes a natural selection of the individuals in that population and eventually only the *fittest* individuals *survive*, i.e. gain access to the limited resources. Consequently, the population of the individuals that survive are the best of the possible individuals.

In the present work, individual solution's (which comprise of the set of patterns, and the numeric range of conceptual labels, see (5.12)) cost (1 - the mean of the resultant MCCs from cross-validation (5.11)) is compared against a set tolerance criterion (see Fig. 5.3). The solutions that have the best fitness values become the 'parents' for the next generation (offsprings) of the 'solutions'. The next generation of solutions is found by incorporating novelty using recombination and/or mutation in the parent solutions. Recombination is an operator that is applied on two or more parents to produce the offsprings whereas mutation is applied to one parent and results in a new altered/mutated offspring. In this way, the application of recombination and mutation generates a new generation of the solutions. In turn, these newly generated solutions are evaluated against the tolerance criterion and given fitness score. If the fitness score is less than the tolerance criterion the search for the optimal solution is stopped, else the process is iterated until a solution with sufficient quality (i.e. meets the tolerance criterion) is found or an iteration limit is reached.

Despite extending novel insights into the functional brain development, a critical limitation of xMVPA is its inability to analyse longitudinal data. Infants' longitudinal neuroimaging data holds key information for delineating brain developmental trajectories. In particular, the infants' longitudinal data analysis can shed light on how neural networks optimise over time to pave way for the hierarchical structure of the human brain. In this regard, there is a need to carry out infants' longitudinal brain data analysis to implement the neural reuse framework, as first discussed in Chapter 2.1. To this end, a Time-dependent XAI (TXAI) system that is capable of analysing longitudinal fNIRS data is developed. The proposed TXAI system integrate the time information using new Temporal Type-2 Fuzzy Sets (TT2FS). Based on the temporal information embedded in the TXAI system, it (TXAI system) is able to inform on the temporal dynamics within the longitudinal data. The proposed TXAI system, for the analysis of infants' longitudinal data, is presented next (in Chapter 5.3).

5.3 Time-dependent XAI (TXAI) Systems

Owing to the inability of standard FLS to integrate temporal information, the proposed xMVPA (presented in Chapter 5.2) can only inform about the workings of the developing brain associated with a given time point, i.e. at the age of the participating infants. More specifically, xMVPA is not able to shed light on how the cortical networks would rewire as the infants' brain develop (both structural and functional) with time as hypothesised by the neural reuse framework (outlined in Chapter 2.1). Further, the incorporation of temporal information associated with infants' brain data could allow to describe brain developmental trajectories which can in turn inform social, clinical, and educational policies. To this end, in this chapter, to model time-dependent real-life processes more effectively, **the theory of a new *Temporal Type-2 Fuzzy Set (TT2FS)* based approach for Time-dependent XAI (TXAI) system** is presented.

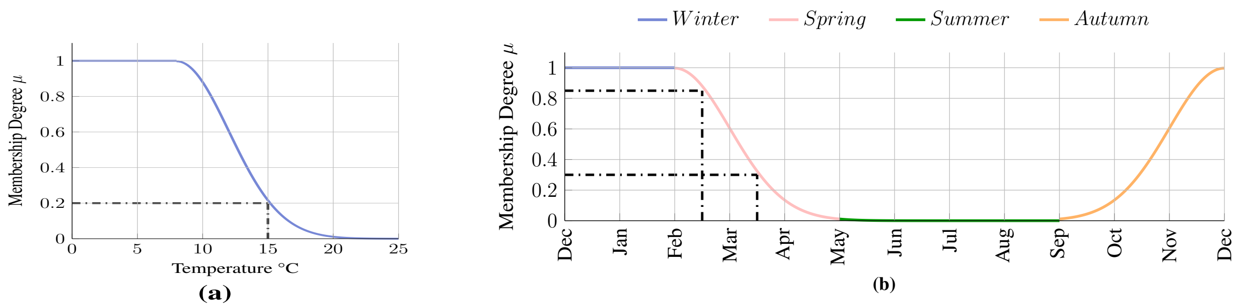
There have been few notable attempts in the literature to model time in the MFs. The work by Garibaldi *et al.* [154] on *non-stationary* fuzzy sets proposed that variation within a MF can be incorporated by perturbing the parameters of the MF. Their work aims to develop non-deterministic fuzzy reason as a way to model the variability in fuzzy decision making to mimic the variability in expert opinions. However, their work does not incorporate the variation within a fuzzy concept with respect to time, which is the aim of the present work, to represent the time-variant transformation of a same fuzzy linguistic variable.

Similarly, the work by Kostikova *et al.* [155] proposes *dynamic* fuzzy sets by extending the classical fuzzy set to include a time dimension for representing MF at different time points. They propose four different types of dynamic MFs depending on how many parameters are changed in the definition of the dynamic MF. However, the dynamic MF is essentially a set of functions determined at different time points with no bearing on the temporal variation in the fuzzy concept.

In another work, Maeda *et al.* [156] propose *dynamic fuzzy reason* to deal with the notion of *time delay* between premise and consequent. An example of where a time delay between premise and consequent assumes critical importance is: 'If it starts snowing, the traffic on road will increase about 30 minutes later'. They propose the use of fuzzy relations between a fuzzy concept and its fuzzy time interval to assign a credit degree to the concept. The temporal fuzzy reasoning provides a framework for modelling delay in fuzzy reasoning and the temporal dynamics of a fuzzy concept.

To the best of my knowledge, there is no work in the literature on fuzzy sets that delineates the incorporation of time-based variation in a fuzzy concept to compute the membership degree for the crisp values of the fuzzy concept. The prowess of TXAI system for incorporating time information for modelling time-variant processes is of paramount significance since the insights provided by a TXAI system can shed light on both spatial (feature domain) and temporal behaviour of the time-dependent process. In addition, no previous work has aimed at delineating the trajectories of a time-variant process with respect to time. To this end, in this work, I propose TXAI systems that can integrate information from both the feature domain and time domain.

The TXAI system is based on TT2FS (Temporal Type-2 Fuzzy Sets) that incorporate information from not only the uncertainty in the input domain of the fuzzy linguistic term, but also from its time of occurrence. In particular, the information from the time of occurrence is integrated into



(a)

(b)

Figure 5.4 An illustrative Type-1 (T1) Membership Function (MF) for the fuzzy concept of 'Cold' in the (a) universe of temperature in °C and (b) in the universe of time: months of a year.

the membership degree of the TT2FS using fuzzy relations such that it varies with respect to time (time-dependent).

Next, I present the most common fuzzy relations and outline how they can be used for implementing TT2FS.

5.3.1 Fuzzy relations between fuzzy linguistic variables and time related measures

In this work, fuzzy relations are used to interrelate the information with respect to the degree of truth of a determined linguistic term or conceptual label, A, within the domain X, and time, T, to form TT2FSs such that the likelihood of occurrence of A in $x \in X$, i.e. the primary membership degree $\mu_A(x)$, is credited by a measure that is dependent on time such as frequency. The application of fuzzy relation, for constructing TT2FSs, is motivated by the work on *dynamic fuzzy reasoning models* in [156]. They outline fuzzy relations that can be used to model time dependencies, as noted in Table 5.1.

Before reviewing the different relations that can be applied to construct a TT2FS, the conditions that need to be fulfilled by the associated temporal MF (TMF) are listed below:

Name	Definition of the relation
Godel	$\Gamma_G(t,x) = \begin{cases} 1 & \text{if } \mu_{T_A}(t) \leq \mu_A(x) \\ \mu_A(x) & \text{if } \mu_{T_A}(t) > \mu_A(x) \end{cases}$
Lukasiewicz	$\Gamma_L(t,x) = 1 \wedge (1 - \mu_{T_A}(t) + \mu_A(x))$
Gaines-Rescher	$\Gamma_{GR}(t,x) = \begin{cases} 1 & \text{if } \mu_{T_A}(t) \leq \mu_A(x) \\ 0 & \text{if } \mu_{T_A}(t) > \mu_A(x) \end{cases}$
Mamdani	$\Gamma_M(t,x) = \mu_{T_A}(t) \wedge \mu_A(x)$

Table 5.1 Fuzzy relations between the universe of concept X and time domain T.

(i) The TMF should be continuous.

(ii) The TMF should be convex.

(iii) The range of the TMF $\subseteq [0,1]$.

(iv) The TMF should reflect in the value of membership degree the intrinsic magnitudes of membership degree in feature domain and in frequency of occurrence domain, i.e. they should be directly proportional. For example, if $\mu_A(x)$ is high and the time representation is also high then the result from the relation between them should also be high and vice versa.

An illustrative comparison of the TT2FSs formed for the conceptual label ‘Cold’ of feature thermal concept using the fuzzy relations listed in Table 5.1 is shown in Fig. 5.5. The fuzzy relations are applied on hypothetical primary MF of ‘Cold’ in feature domain (temperature) and time domain (months of a year). As can be seen in Fig. 5.5, the different fuzzy relations are encapsulating distinct inter-dependencies between time and feature domain. All relations meet the criteria (i) - (iii) listed above however, only the Mamdani relation meets the criterion (iv) as well since it gives credit to μ_{Cold} based on the variable frequency of occurrence of ‘Cold’ as observed in different months of the year. Hence, in this work, the Mamdani relation is used to construct the TT2FSs.

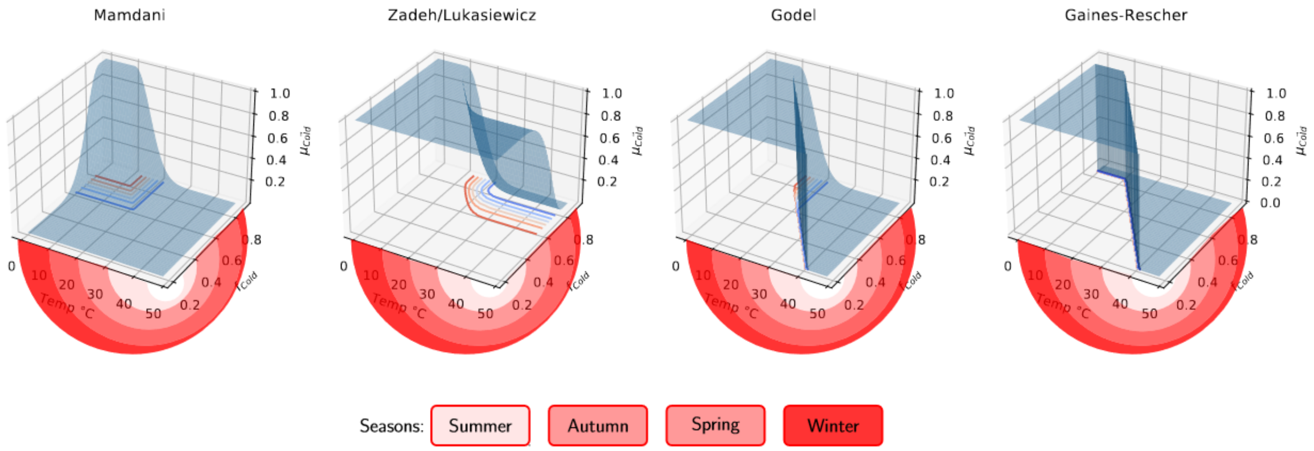


Figure 5.5 A comparison of TT2FSs for the conceptual label 'Cold' for feature thermal concept constructed with the most commonly used fuzzy relations.

5.3.2 Conditional relative frequency distribution of a fuzzy linguistic term

In the proposed TT2FS, a measure of conditional relative frequency between time and the occurrence of a linguistic term is employed. I denote as A an instance of a linguistic term from a set of conceptual labels (CoL), $CoLs := [CoL_1, CoL_2, ..., CoL_W]$ of a specific linguistic variable or input.

Definition 5.3.1 (Discrete conditional relative frequency with respect to time)**.** *The discretized conditional relative frequency is defined as the likelihood of observing a linguistic term A based on its membership degree, across time. This is denoted as $g_A(t_n, \mu_A(x))$ with time t discretised over N time points (t_n) such as $t_n \in [t_1, ..., t_N]$, and is given by:*

$$g_A(t_n, \mu_A(x)) = \frac{\sum\limits_{x \in X, t_n} \delta_{nl}}{\max\limits_{[t_1,...,t_N]} \left(\sum\limits_{x \in X, t_n} \delta_{nl} \right)} \tag{5.13}$$

δ_{nl} is a Kronecker delta function [157] (e.g. δ_{ab} = 0 if a ≠ b, δ_{ab} = 1 if a=b) that takes the value of 1 when the following condition applies, $\exists\ argmax_l(\mu_{CoL_j}(x^{t_n})) : Col_l = A$, $\forall l \in [1, ..., W]$, and 0 otherwise. Note x^{t_n} is a realisation of x at time t_n.

The numerator in (5.13) finds the count of occurrences of a given A for a determined time point t_n across all data instances, whereas the denominator is finding the maximum value of the count of occurrences of A across all N time points and all data instances. The resultant discrete conditional relative frequency $g_A(t_n, \mu_A(x))$ is interpolated to form a conditional distribution $f_A(t, \mu_A(x))$. For the sake of notational simplicity, I denote the later distribution as f_A and the discrete conditional relative frequency as g_A from here on-wards.

Let us assume that the linguistic variable is thermal sensation defined on the input domain ($x \in X$) of temperature in °C and the associated conceptual labels be: [Cold, Comfortable, Hot]. For a given crisp input of temperature such as 15°C, the associated primary membership degree for all three conceptual labels of Cold, Comfortable, and Hot be $\mu_{Cold}(15,°C) = [0.4],,\mu_{comf.}(15°C) = [0.3]$, $\mu_{hot}(15°C) = [0]$ respectively. In this illustrative case, the temperature of 15 °C has a maximum membership grade, amongst all conceptual labels, for *Cold* and hence 15 °C is assigned with the conceptual label of *Cold*. Referring back to (5.13), for computing the conditional relative frequency for *Cold* the numerator is going to sum all the data instances where the crisp inputs are assigned with *Cold* for a given time point t_n such as a particular month of a year. The denominator finds the mode of occurrence of *Cold* across all months. The result of the division will scale the g_{Cold} values to [0,1].

An illustration for calculating the g_{Cold} values using (5.13), with a total of 12 time points as the months of a year is shown in Fig. 5.6 (b) with continuous values of f_{Cold}, found using interpolation of g_{Cold}, plotted in Fig. 5.6 (c). Please note the associated time intervals, (as listed in the illustration in Fig. 5.6 are seasons in a year such as Winter, Spring, Summer, and Autumn), are for easing the computational complexity of the four-dimensional (4D) TT2FSs as will be explained later in Section 5.3.4 by taking time interval based slice of the TT2FS.

5.3.3 Temporal Type-2 Fuzzy Sets (TT2FS)

In this section, a formal definition of Temporal Type-2 Fuzzy Sets (TT2FS) is presented. TT2FS are 4D as they incorporate information from the input domain (X), time domain (T), frequency of occurrence domain (F) and are characterised by a temporal MF (TMF).

The computation of TMF, hereby termed as *temporal fuzzification*, involves two stages: 1) fuzzification of crisp input values of A from feature domain X to form T1 $\mu_A(x)$, as undertaken in standard T1 fuzzy sets; and 2) computation of the conditional distribution of A, f_A. The temporal fuzzification is illustrated in Fig. 5.6 (a) and defined next.

Definition 5.3.2 (Temporal membership function)**.** *The Temporal MF (TMF) can be defined as follows:*

$$\mu_{\tilde{A}}(x,t,f_A) = \mu_A(x) \otimes f_A \tag{5.14}$$

where $\otimes$ is a relation operator, $\mu_A(x)$ is the primary membership of A in feature domain credited by the conditional distribution of A, denoted f_A, using the Mamdani relation (outlined earlier in Sec 5.3.1).

Theorem 5.3.1. *The TMF of A, constructed using Mamdani relation (5.14), $\mu_{\tilde{A}}(x,t,f_A)$ is $\subseteq [0,1]$.*

Proof. The range of $\mu_{\tilde{A}}(x,t,f_A)$ follows directly from the range of primary MF of A: $\mu_A(x) \subseteq [0,1]$, and the conditional distribution of A: $f_A \subseteq [0,1]$. Hence, by crediting $\mu_A(x)$ with f_A using Mamdani relation (taking the min or product), it follows that the range of $\mu_{\tilde{A}}(x,t,f_A) \subseteq [0,1]$. ∎

Proposition 5.3.1.1. *If the primary membership of TMF is normal and the conditional distribution f is normal, according to (5.13), then the resultant TMF membership after applying the Mamdani relation yields a normal TMF, therefore we can imply that*

$$\underset{x \in X}{sup}\, \mu_{\tilde{A}}(x,t,f_A) = 1 \tag{5.15}$$

Proof. Given a $f_A \subseteq [0,1]$ and a $\mu_A(x) \subseteq [0,1]$ both with $sup = 1$, $\forall x \in X$ by deduction, $\exists x :$ $f_A \times \mu_A(x) \vee min(f_A, \mu_A(x)) = 1$ ■

Next, we define the TT2FS which are characterised by a TMF.

Definition 5.3.3 (Temporal Type-2 Fuzzy Sets (TT2FS))**.** *A TT2FS $\vec{A}$ of the universe of discourse $X \times T \times F$ is characterised by a credited TMF $\mu_{\vec{A}}(x,t,f_A) : X \times T \times F \to [0,1]$ where X is the feature domain of A characterised by a T1 MF $\mu_A(x)$, T is the time domain of A, F is the frequency of occurrence domain of A characterised by conditional frequency distribution with respect to time f_A. In mathematical set notation, $\vec{A}$ can be written as (5.16):*

$$\begin{aligned} \vec{A} =&\{(x,t,f_A,\mu_{\vec{A}}(x,t,f_A)) \mid \\ &\forall x \in X, \forall t \in T, \forall \mu_A(x) \subseteq [0,1], \\ &\forall f_A \in F \subseteq [0,1]\} \end{aligned} \tag{5.16}$$

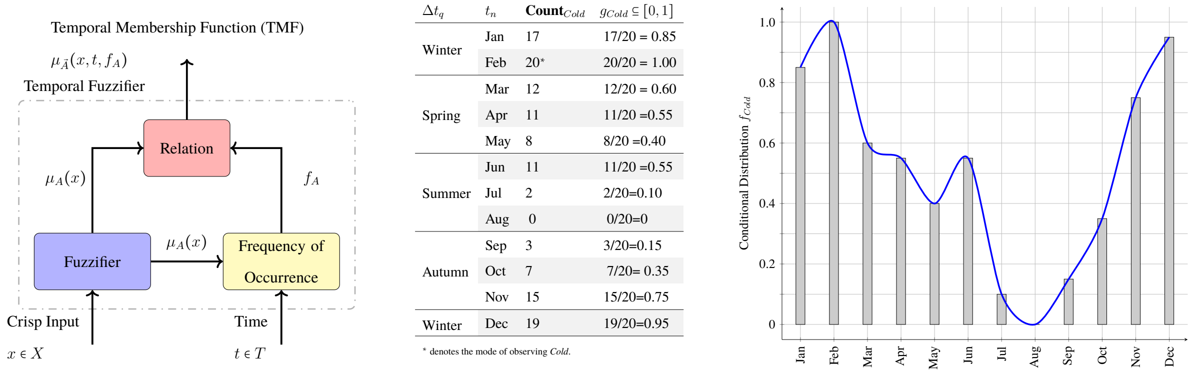
where $\mu_{\vec{A}}(x,t,f_A) \subseteq [0,1]$. Please note the conditional distribution, f_A, is a continuous distribution interpolated from discrete conditional relative frequency, g_A, and is defined mathematically earlier in (5.13). $\vec{A}$ can also be expressed as:

$$\vec{A} = \int_{x \in X} \int_{t \in T} \int_{f_A \in F} \mu_{\vec{A}}(x,t,f_A)/f_A/t/x \tag{5.17}$$

where $\int \int \int$ denotes the aggregation over all admissible values of x, t, and f_A. The associated TMF, $\mu_{\vec{A}}(x,t,f_A)) \subseteq [0,1]$, scales the $\mu_A(x)$ based on its conditional distribution f_A as defined in (5.14).

5.3.4 Time Interval slice followed by Z-slice (TS-ZS) approach for TT2FS

A popular approach for minimising the computational demand of 3D GT2 fuzzy sets is to use z-slice based framework [134], previously outlined in Chapter 3. Motivated from the effectiveness of z-slice



Δt_q	t_n	**Count**$_{Cold}$	$g_{Cold} \subseteq [0,1]$
Winter	Jan	17	17/20 = 0.85
	Feb	20*	20/20 = 1.00
Spring	Mar	12	12/20 = 0.60
	Apr	11	11/20 =0.55
	May	8	8/20 =0.40
Summer	Jun	11	11/20 =0.55
	Jul	2	2/20=0.10
	Aug	0	0/20=0
Autumn	Sep	3	3/20=0.15
	Oct	7	7/20= 0.35
	Nov	15	15/20=0.75
Winter	Dec	19	19/20=0.95

* denotes the mode of observing *Cold*.

(a) Temporal Fuzzifier.

(b) g_{Cold}

(c) f_{Cold}

Figure 5.6 (a) A schematic of temporal fuzzification for constructing Temporal Membership Function (TMF).

based framework for simplifying the computations for GT2 fuzzy sets, in this work, the approach of taking time interval slice followed by z-slice (TS-ZS) is taken for performing operations on TT2FSs. The TS-ZS approach is explained in more detail as follows:

(i) TS: Time interval based slice to convert 4D TT2FSs into 3D. The 3D time interval based TT2FS is similar to 3D GT2 fuzzy set, with both sharing the feature domain on $x-$axis. On $y-$axis is the frequency of occurrence domain, for that time interval, for time interval based TT2FS, while for GT2 fuzzy sets, primary membership grade is on $y-$axis. And on $z-$axis is the temporal membership degree for time interval based TT2FS while for GT2 fuzzy set secondary membership degree is on $z-$axis.

(ii) ZS: z-Slice based approach for the time interval based 3D TT2FS as utilised for GT2 fuzzy sets. The z-slices at specific z-levels render a given 3D fuzzy set to an equivalent IT2 fuzzy set with lower and upper primary membership degrees. For the case of TS-ZS based TT2FSs, the primary membership degrees are the conditional distribution values for that time interval at a given z-level.

In the following sections, a formal definition for the operations on TT2FSs is given with $\vec{A}$ and $\vec{B}$ denoting two TT2FSs characterised by TMFs $\mu_{\vec{A}}(x,t,f_A)$ and $\mu_{\vec{B}}(x,t,f_B)$ respectively as outlined in (5.18):

$$\begin{aligned}\vec{A} &= \int_{x\in X}\int_{t\in T}\int_{f\in F}\mu_{\vec{A}}(x,t,f_A)/f_A/t/x\\ \vec{B} &= \int_{x\in X}\int_{t\in T}\int_{f\in F}\mu_{\vec{B}}(x,t,f_B)/f_B/t/x\end{aligned} \tag{5.18}$$

where X is the feature domain, T is the time domain, and F is the frequency of the occurrence domain.

5.3.5 Union of TT2FS

A general procedure for undertaking the union (and intersection) operations on the 4D TMFs is outlined in Algorithm 3. The union of two TT2FSs $\vec{A}$ and $\vec{B}$ is a TT2FS defined as $\vec{A}\cup\vec{B}$ in (5.19):

$$\vec{A}\cup\vec{B} = \int_{x\in X}\int_{t\in T}\int_{f\in F}\mu_{\vec{A}\cup\vec{B}}(x,t,f)/f/t/x \tag{5.19}$$

where $\mu_{\vec{A}\cup\vec{B}}$ can be calculated by discretising the T domain, and taking z-slices on $\mu_{\vec{A}\cup\vec{B},\Delta t_\alpha}(x,t,f)$ values as outlined in (Alg 3.1) of Algorithm 3. In particular, for union operation, at time interval Δt_α (Alg 3.1) takes the form of (5.20) when using the max t-conorm:

$$\mu_{\vec{A}\cup\vec{B},\Delta t_\alpha}(x,f_{\Delta t_\alpha}) = \sum_x \sum_{f_{\Delta t_\alpha}\in[max(l_A,l_B),max(u_A,u_B)]} z_i/f_{\Delta t_\alpha} \tag{5.20}$$

5.3.6 Intersection of TT2FS

Likewise, the intersection of TT2FSs can be written as shown in (5.21)

Algorithm 3: Union and Intersection Operations on TT2FSs

```
Result: Resultant Temporal Membership Function (TMF) $\mu_{\vec{A}\oslash\vec{B}}(x, t, f_{A\oslash B})$ where $\oslash$ denotes
        the operation of either union or intersection.
Let concepts A and B on feature domain X have TMFs denoted by $\mu_{\vec{A}}(x, t, f_A(t, \mu_A(x)))$ and
  $\mu_{\vec{B}}(x, t, f_B(t, \mu_B(x)))$ respectively with time intervals $\Delta t_\alpha \in [\Delta t_1, ..., \Delta t_V]$ and zslices
  discretised at $z_i \in [z_1, z_2, ..., z_I]$
For each time interval $\Delta t_\alpha$ the operation (union or intersection) on 3D time interval based
  TMF is computed independently by first taking the z-slices at $z_i \in [z_1, z_2, ..., z_I]$ which
  renders the 3D time interval based TMF into Interval Type-2 (IT2) TMFs
For each IT2 TMF, the operation is done as shown below in eq. (Alg 3.1)
for $x \in X$ do
    for $z_i < z_I$ do
        $\mu_{\vec{A}\oslash\vec{B},\Delta t_\alpha}(x, f_{\Delta t_\alpha}) = \sum_{x} \sum_{f_{\Delta t_\alpha} \in [\odot(l_A, l_B), \odot(u_A, u_B)]} z_i / f_{\Delta t_\alpha}$    (Alg 3.1)
    end
end
where the summation signs in eq. (Alg 3.1) denotes the aggregation in set theoretic operation, $l$ and $u$
  are the lower and upper conditional distribution values respectively of set $\vec{A}$ and $\vec{B}$ on z-slice $z_i$ and
  time interval $\Delta t_q$. For union operation, in eq. (Alg 3.1), the $\odot$ denotes $max$ and for intersection
  operation $\odot$ denotes $min$.
```

$$\vec{A} \cap \vec{B} = \int_{x \in X} \int_{t \in T} \int_{f \in F} \mu_{\vec{A} \cap \vec{B}}(x, t, f)/f/t/x \tag{5.21}$$

where $\mu_{\vec{A} \cap \vec{B}}$ can be calculated by discretising the T domain, and taking z-slices on $\mu_{\vec{A} \cap \vec{B}, \Delta t_\alpha}(x, t, f)$ values as outlined in (Alg 3.1) of Algorithm 3. In particular, for intersection operation, at time interval Δt_α (Alg 3.1) takes the form of (5.22) when using the min t-norm. However, either product or min can be applied.

$$\mu_{\vec{A} \cap \vec{B}, \Delta t_\alpha}(x, f_{\Delta t_\alpha}) = \sum_x \sum_{f_{\Delta t_\alpha} \in [min(l_A, l_B), min(u_A, u_B)]} z_i / f_{\Delta t_\alpha} \tag{5.22}$$

5.3.7 Defuzzification of TT2FS

In general, defuzzification converts a fuzzy set to an equivalent crisp number, and can be thought of as the inverse of fuzzification. As outlined earlier in Chapter 3, for T1 fuzzy sets, defuzzification usually involves computing the centroid of the T1 fuzzy set [158] to compute a representative crisp number, as shown in (5.23).

$$x^* = \frac{\sum_{i=1}^{I} x_i \mu(x_i)}{\sum_{i=1}^{I} \mu(x_i)} \tag{5.23}$$

where x^* is the centroid of the T1 MF defined on the domain $x \in X$. Here, the summation sign is used as in typical mathematical equations, i.e., for the case of the numerator, it is summing the product of x values and their corresponding membership degrees whereas for the denominator it is summing the membership degrees corresponding to all x_i values $\forall i \in [1, ..., I]$.

For a 3D GT2 fuzzy set, defuzzification usually involves three steps, outlined as follows:

(i) Transforming a 3D GT2 fuzzy set to IT2 fuzzy sets by slicing the GT2 fuzzy set at given z-levels such as $z_i \in [z_1, ..., z_I]$.

```
Algorithm 4: Defuzzification of TT2FSs for a given time interval $\Delta t_\alpha$
  Result: Crisp value for a given time interval, denoted by $crisp_{\Delta t_\alpha}$, where $\Delta t_\alpha$ is the $\alpha^{th}$ time interval.
  Let feature A on feature domain X have temporal membership function (TMF) denoted by
    $\mu_{\tilde{A}}(x, t, f_A(t, \mu_A(x)))$ with time intervals $\Delta t_\alpha \in [\Delta t_1, ..., \Delta t_V]$ and z-slices discretised at
    $z_i \in [z_1, z_2, ..., z_I]$
  For each 3D time interval based TMF, the defuzzification can be done independently, by first
    taking the z-slices at $z_i \in [z_1, z_2, ..., z_I]$ which renders the 3D time interval based TMF into
    Interval Type-2 (IT2) MFs
  The left and right centroid for each IT2 TMF at z-location $z_i$, denoted by $C_{z_i,\Delta t_\alpha}$, can be
    computed using Karnik-Mendel (KM) method [159] to give $[c_l, c_r]$ at that z-slice $z_i$ and time
    interval $\Delta t_\alpha$ as outlined in eq. (Alg 4.1)
  for $z_i \le z_I$ do
      $C_{z_i,\Delta t_\alpha} = [c_{l_{z_i,\Delta t_\alpha}},\ c_{r_{z_i,\Delta t_\alpha}}]$    (Alg 4.1)
  end
  Defuzzifcation of the type reduced T1 fuzzy sets, using centroid average, to find equivalent
    $c_{l_{\Delta t_\alpha}}$ and $c_{r_{\Delta t_\alpha}}$
    $c_{l_{\Delta t_\alpha}}(x) = \frac{(z_1 * c_{l_{z_1,\Delta t_\alpha}}) + (z_2 * c_{l_{z_2,\Delta t_\alpha}}) + ... + (z_I * y_{l_{z_I,\Delta t_\alpha}})}{z_1 + z_2 + ... + z_I}$    (Alg 4.2)
    $c_{r_{\Delta t_\alpha}}(x) = \frac{(z_1 * c_{r_{z_1,\Delta t_\alpha}}) + (z_2 * c_{r_{z_2,\Delta t_\alpha}}) + ... + (z_I * y_{r_{z_I,\Delta t_\alpha}})}{z_1 + z_2 + ... + z_I}$    (Alg 4.3)
  A crisp value, $crisp_{\Delta t_\alpha}$, can now be computed by applying Nie-Tan method [132] on $c_{l_{\Delta t_\alpha}}$ and
    $c_{r_{\Delta t_\alpha}}$.
```

(ii) Type reducing the z-level based IT2 fuzzy sets results in two T1 fuzzy sets using KM (Karnik Mendel) method [159] (outlined in Appendix A). The type-reduced T1 fuzzy sets are composed of the left and right centroids of the IT2 fuzzy sets. More specifically, the KM method requires iterative process to compute left and right centroids resulting in two T1 fuzzy sets: $[c_{l_{z_1}}, c_{l_{z_2}}, ..., c_{l_{z_I}}]$ and $[c_{r_{z_1}}, c_{r_{z_2}}, ..., c_{r_{z_I}}]$ where $c_{l_{z_1}}$ is the left centroid at z-level 1 and $c_{r_{z_1}}$ is the right centroid at z-level 1 and so on.

(iii) Defuzzifcation of the type reduced T1 fuzzy sets, using centroid average, to find equivalent $c_l(x)$ and $c_r(x)$.

$$c_l(x) = \frac{(z_1 * c_{l_{z_1}}) + (z_2 * c_{l_{z_2}}) + ... + (z_I * c_{l_{z_I}})}{z_1 + z_2 + ... + z_I} \tag{5.24}$$

$$c_r(x) = \frac{(z_1 * c_{r_{z_1}}) + (z_2 * c_{r_{z_2}}) + ... + (z_I * c_{r_{z_I}})}{z_1 + z_2 + ... + z_I} \tag{5.25}$$

(iv) The final type-reduced crisp value is found using the Nie-Tan method [132] on c_l and c_r.

In this work, the defuzzification of 4D TT2FS also involves TS-ZS approach (explained earlier in section 5.3.4), i.e., taking the time interval based slice followed by z-slices. The time interval based TMF is 3D, and for each of the time interval (Δt_α) based TMF, z-slices at particular z_i levels renders them as IT2 fuzzy sets. The KM procedure [159] can be applied on IT2 fuzzy sets, at each z-level, to compute T1 fuzzy sets composed of $[c_{l_{z_i,\Delta t_\alpha}}, c_{r_{z_i,\Delta t_\alpha}}]$ as outlined in (Alg 4.1). Using the centroid defuzzifier, the T1 fuzzy sets are defuzzified to give one equivalent c_l and c_r, for that time interval, as outlined in (Alg 4.2) and (Alg 4.3). The Nie-Tan method [132] is then applied to compute one crisp value for that time interval. The defuzzification of TT2FSs, for a given time interval, is summarised in Algorithm 4. The procedure outlined in Algorithm 4 can be repeated for each time interval, i.e. Δt_α where $\alpha \in [1, ..., V]$, to obtain a crisp value for all time intervals.

5.3.8 TXAI Inference System (TXAI-IS)

In this section, the TXAI Inference System (TXAI-IS) for classification problems is outlined. A general flowchart for the TXAI-IS is outlined in Fig. 5.7. The temporal fuzzifier constructs the 4D TT2FSs as outlined in Fig. 5.6 (a). To analyse a given dynamic process with respect to time, the TXAI-IS works for each time interval Δt_α where $\Delta t_\alpha \in [\Delta t_1, ..., \Delta t_V]$ independently. To this end, the 4D TT2FSs are first sliced based on the Δt_α, and inference is made on time sliced 3D TT2FSs using the temporal patterns for the same Δt_α. Each time interval would entail a unique temporal pattern base. The temporal patterns can either be furnished by experts in the field or can be learnt from the input data using evolutionary algorithms such as Genetic Algorithm (GA) [160].

In the next subsections, the classification TXAI-IS is outlined in detail as the empirical study on which TXAI system is exemplified also undertakes a classification problem, i.e., occupancy dataset [161] is analysed to determine whether or not a room is occupied.

Classification

For the classification problem, the TXAI-IS will predict one class for a given data instance for each time interval. The overall TXAI-IS for classification undertakes the following steps:

(i) Compute the membership degree for the time interval based 3D TT2FSs.

- The time interval based 3D TT2FSs are transformed into IT2 fuzzy sets by taking slices at predefined z-levels. The membership degree at each z-level, such as $z_i \in [z_1, ..., z_I]$ where I is the total number of z slices, for a given 3D TT2FS is given as follows [134]:

$$\tilde{A} = \{(x, u, z) | \forall x \in X, \qquad \forall u \in [\underline{\mu}_{\tilde{A}}(x), \overline{\mu}_{\tilde{A}}(x)] \subseteq [0, 1]\} \tag{5.26}$$

where $\mu_{\tilde{A}}$ is the membership degree of the IT2 fuzzy set $\tilde{A}$ at the predefined z level.

(ii) Compute the firing strength for each pattern, at each z-level.

- The $\overline{upper}$ and $\underline{lower}$ firing strength of a given patterns P_q, $\overline{w}_q$ and $\underline{w}_q$ respectively, is the degree of match between the pattern P_q and the data instance x. It is computed as:

$$\overline{w}_q(x) = \prod_{\Psi=1}^{\Phi} \overline{\mu}_{\tilde{A}^{\Psi}}(x^{\Psi})$$
$$\underline{w}_q(x) = \prod_{\Psi=1}^{\Phi} \underline{\mu}_{\tilde{A}^{\Psi}}(x^{\Psi}) \tag{5.27}$$

where q is the pattern number, Φ is the total number of antecedents in the pattern P_q.

(iii) Compute the dominance score (DS) for each pattern, at each z-level.

- The DS is a measure of a given pattern's dominance and is computed as shown in (5.28).

$$\overline{DS}_q = \overline{conf}_q \times \overline{sup}_q \tag{5.28}$$
$$\underline{DS}_q = \underline{conf}_q \times \underline{sup}_q$$

where $conf$ is the confidence of the pattern P_q and sup is the support of the qth pattern.

- The confidence of a pattern is a measure of the likelihood to correctly classify a given data instance. It is calculated as shown in eq. (5.29)

$$\overline{conf}_q(\Psi_q \Rightarrow Y_q) = \frac{\sum_{x\in(\Psi_q \Rightarrow Y_q)} \overline{w}_q(x)}{\sum_{q=1,x\in(\Psi_q)}^{Q} \overline{w}_q(x)} \tag{5.29}$$
$$\underline{conf}_q(\Psi_q \Rightarrow Y_q) = \frac{\sum_{x\in(\Psi_q \Rightarrow Y_q)} \underline{w}_q(x)}{\sum_{q=1,x\in(\Psi_q)}^{Q} \underline{w}_q(x)}$$

where Ψ_q and Y_q are the antecedents and consequent respectively of the pattern P_q. The numerator sums the firing strength of all the data instances that have the same antecedents and consequent as the pattern P_q. Whereas the denominator sums the firing strength of all the data instances that have the same antecedents as the pattern P_q irrespective of the consequent- for all the patterns $[1, ..., Q]$, where Q is the total number of patterns.

- The support of a pattern is calculated as shown in (5.30)

$$\overline{sup}_q(\Psi_q \Rightarrow Y_q) = \frac{\sum_{x\in(\Psi_q\Rightarrow Y_q)} \overline{w}_q(x)}{Q} \quad (5.30)$$

$$\underline{sup}_q(\Psi_q \Rightarrow Y_q) = \frac{\sum_{x\in(\Psi_q\Rightarrow Y_q)} \underline{w}_q(x)}{Q}$$

  with Q as the total number of patterns.

(iv) Compute the association degree of each pattern, with a given data instance, for each z-level.

- The association degree of a pattern P_q with a given data instance x is computed as shown in (5.31):

$$\overline{h}_q = \overline{w}_q(x) \times \overline{DS}_q \quad (5.31)$$

$$\underline{h}_q = \underline{w}_q(x) \times \underline{DS}_q$$

(v) Predict the class.

- Find a value of the association degree, h, for each pattern by using Nie-Tan [132] method on the $\underline{h}$ and $\overline{h}$ which are found using (Alg 4.2) and (Alg 4.3).
- The pattern with the highest association degree, h, predicts the class for the given data instance.

(vi) The steps outlined above (i)-(v) are repeated for each time interval to predict a class for all time intervals.

Numerical Step-wise Example

In this section, a binary classification problem using TXAI-IS is exemplified using a hypothetical dataset with two input features, *Feature1* and *Feature2*, and one output. Let time intervals be defined

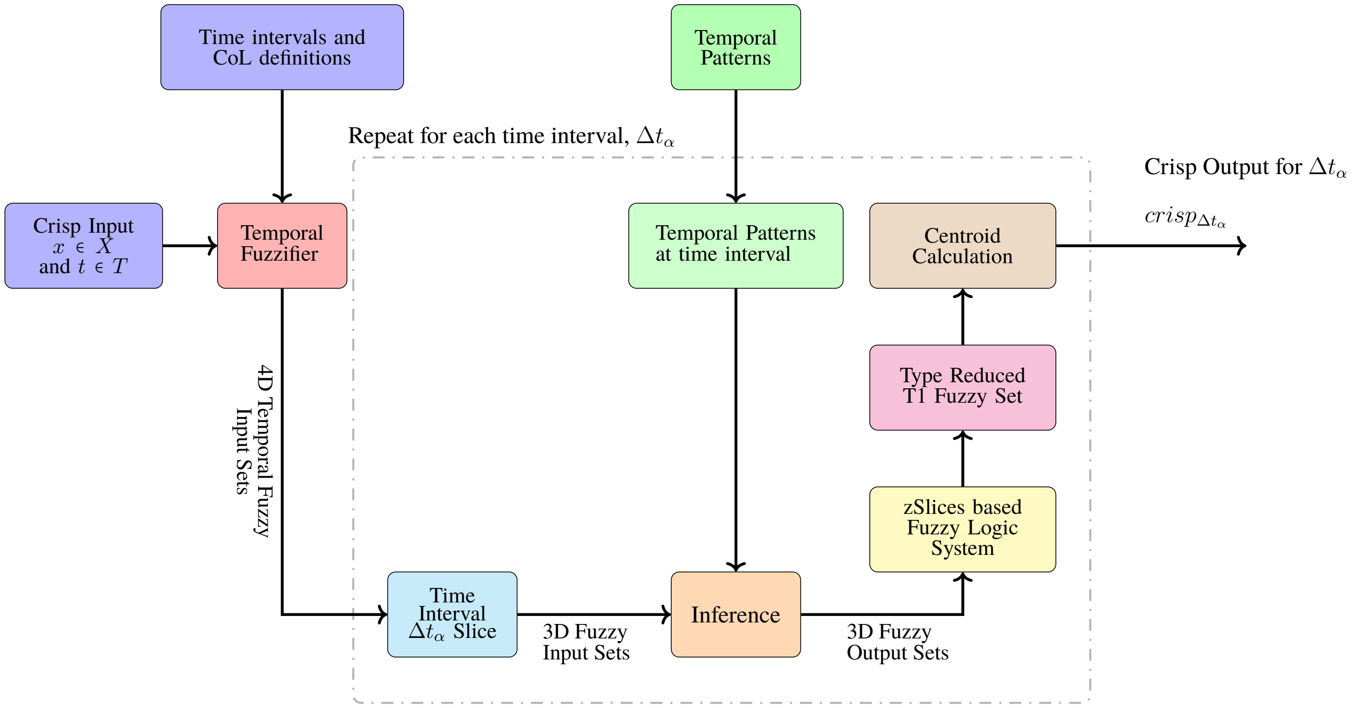

Time intervals and CoL definitions
Temporal Patterns
Repeat for each time interval, Δt_α
Crisp Output for Δt_α
$crisp_{\Delta t_\alpha}$
Crisp Input $x \in X$ and $t \in T$
Temporal Fuzzifier
Temporal Patterns at time interval
Centroid Calculation
Type Reduced T1 Fuzzy Set
4D Temporal Fuzzy Input Sets
zSlices based Fuzzy Logic System
Time Interval Δt_α Slice
3D Fuzzy Input Sets
Inference
3D Fuzzy Output Sets

Figure 5.7 A general schematic representation delineating the interlinks between salient components of a Time-dependent eXplainable Artificial Intelligence Inference System (TXAI-IS).

over a day such as Morning, Daytime, and Evening with three conceptual labels associated with the inputs (*Feature1* and *Feature2*) be: [Low, Medium, High] and output classes be *Output1* and *Output2*.

First, TT2FSs for both inputs (*Feature1* and *Feature2*) are constructed using temporal fuzzifier, as outlined in Fig. 5.6. Also, for each time interval, the patterns will be different but the overall process to determine the output class is same. In the following steps, I exemplify how the output class is predicted for one time interval, in this example, Morning.

Let the patterns (P) outlining the relation between input features and output for Morning be as listed in (5.32). The corresponding lower and upper dominance score at each z-level are as listed in Table 5.3. In the following steps i)- iv) I show how a corresponding class for $Output$ is predicted using TXAI-IS for input values of *Feature1* = 19.7 and *Feature2* be = 4.3. In this example, the z-level is discretised at $z_{0.2}$, $z_{0.4}$, $z_{0.6}$, $z_{0.8}$, and $z_{1.0}$.

$$\begin{aligned}
P_1 &: \text{IF } \textit{Feature1} \text{ is } Low \text{ and } \textit{Feature2} \text{ is } Medium \\
&\quad \text{THEN } \textit{Output} \text{ is } \textit{Output2} \\
P_2 &: \text{IF } \textit{Feature1} \text{ is } Medium \text{ and } \textit{Feature2} \text{ is } Medium \\
&\quad \text{THEN } \textit{Output} \text{ is } \textit{Output1} \\
P_3 &: \text{IF } \textit{Feature1} \text{ is } High \text{ and } \textit{Feature2} \text{ is } High \\
&\quad \text{THEN } \textit{Output} \text{ is } \textit{Output1}
\end{aligned} \tag{5.32}$$

(i) The membership degree for each conceptual label of the inputs *Feature1* and *Feature2* is determined from the time interval (Morning) based 3D TMF. The membership degree is the value of the conditional distribution at a given input value and corresponding z-level as outlined in (5.26). Let the corresponding membership degrees for each conceptual label of the inputs

Feature1 and *Feature2* be as noted in Table 5.2.

Feature	**CoLs**		$z_{0.2}$	$z_{0.4}$	$z_{0.6}$	$z_{0.8}$	$z_{1.0}$
Feature1	Low	Lower	0.50	0.52	0.54	0.52	0.51
		Upper	0.61	0.63	0.64	0.61	0.60
	Med.	Lower	0.63	0.63	0.65	0.63	0.61
		Upper	0.77	0.78	0.78	0.77	0.75
	High	Lower	0.65	0.64	0.64	0.63	0.63
		Upper	0.69	0.69	0.68	0.68	0.67
Feature2	Low	Lower	0.31	0.31	0.31	0.31	0.31
		Upper	0.32	0.32	0.32	0.32	0.32
	Med.	Lower	0.50	0.55	0.55	0.54	0.53
		Upper	0.58	0.59	0.59	0.58	0.57
	High	Lower	0.40	0.40	0.40	0.42	0.44
		Upper	0.43	0.43	0.46	0.46	0.49

Table 5.2 The hypothetical lower and upper membership degrees of the Conceptual Labels (CoLs) of *Feature1* and *Feature2*.

(ii) The firing strength of each pattern listed in (5.32) are found, using the membership degree in Table 5.2, as outlined in (5.27) and listed in Table 5.3. As an example, for P_1 the lower firing strength at $z = 0.6$, $\underline{w}_{1_{z=0.6}}$, can be calculated as follows:

$$\begin{aligned}\underline{w}_{1_{z=0.6}}(x = [19.7, 4.3]) &= \prod_{i=1}^{2} \underline{\mu}(x^i) \\ &= 0.54 \ast 0.55 = 0.297 \end{aligned} \tag{5.33}$$

(iii) The association degree of each pattern with the input data instance is determined, using the firing strength in Table 5.3, as outlined in (5.31). The upper and lower values of the association degree for the five z-levels are as listed in Table 5.4. As an example, for P_2 the upper association degree at $z = 0.2$, $\overline{h}_{2_{z=0.2}}$, can be calculated as follows:

Pattern	Firing Strength, w	z-level					Consequent	Dominance Score (DS)	z-level				
		$z_{0.2}$	$z_{0.4}$	$z_{0.6}$	$z_{0.8}$	$z_{1.0}$			$z_{0.2}$	$z_{0.4}$	$z_{0.6}$	$z_{0.8}$	$z_{1.0}$
P_1	Lower	0.25	0.286	0.297	0.281	0.27	*Output2*	Lower	0.31	0.30	0.30	0.29	0.27
	Upper	0.354	0.372	0.378	0.354	0.342		Upper	0.35	0.34	0.34	0.31	0.30
P_2	Lower	0.315	0.347	0.358	0.34	0.323	*Output1*	Lower	0.69	0.69	0.68	0.66	0.66
	Upper	0.447	0.46	0.46	0.447	0.427		Upper	0.73	0.73	0.72	0.72	0.72
P_3	Lower	0.26	0.256	0.256	0.265	0.277	*Output1*	Lower	0.22	0.21	0.21	0.21	0.21
	Upper	0.297	0.297	0.313	0.313	0.328		Upper	0.24	0.22	0.22	0.22	0.22

Table 5.3 The lower and upper firing strengths, w, for the hypothetical patterns listed in (5.32)

$$\begin{aligned} \overline{h}_{2_{z=0.2}} &= \overline{w}_{2_{z=0.2}}(x) \times \overline{DS}_{2_{z=0.2}} \\ &= 0.447 * 0.73 = 0.326 \end{aligned} \tag{5.34}$$

(iv) The consequent of the pattern with the highest association degree with the input data instance becomes the predicted class for a given time interval. The crisp value for the association degree of each pattern is found using (Alg 4.2) and (Alg 4.3). As an example, the crisp value of association degree for P_3 is found as follows:

$$\begin{aligned} h_{3_l} &= \frac{0.2 * (\underline{h}_{3_{0.2}}) + ... + 1.0 * (\underline{h}_{3_{1.0}})}{0.2 + 0.4 + 0.6 + 0.8 + 1.0} \\ &= \frac{0.2 * 0.057 + 0.4 * 0.054 + ... + 1 * 0.058}{3} = 0.056 \\ h_{3_u} &= \frac{0.2 * (\overline{h}_{3_{0.2}}) + ... + 1.0 * (\overline{h}_{3_{1.0}})}{0.2 + 0.4 + 0.6 + 0.8 + 1.0} \\ &= \frac{0.2 * 0.071 + 0.4 * 0.065 + ... + 1 * 0.072}{3} = 0.0696 \\ h_{3_{crisp}} &= \frac{0.056 + 0.0696}{2} = \frac{0.1256}{2} = 0.063 \end{aligned} \tag{5.35}$$

In this illustrative example, P_2 has the highest association degree (tabulated in Table 5.4) hence the predicted output for the input data instance (*Feature1* = 19.7 and *Feature2* be = 4.3) for time interval Morning is the consequent of P_2, i.e., *Output1*.

P	**h**	$z_{0.2}$	$z_{0.4}$	$z_{0.6}$	$z_{0.8}$	$z_{1.0}$	h_{crisp}
P_1	Lower	0.077	0.086	0.089	0.081	0.073	0.097
	Upper	0.124	0.126	0.128	0.11	0.103	
P_2	Lower	0.217	0.239	0.243	0.225	0.213	0.274
	Upper	0.326	0.336	0.331	0.322	0.308	
P_3	Lower	0.057	0.054	0.054	0.056	0.058	0.063
	Upper	0.071	0.065	0.069	0.069	0.072	

Table 5.4 The lower and upper association degrees, h, for each of the three patterns (P) listed in (5.32).

The same process can be repeated for each time interval with their respective patterns to predict a class for the output. Hence, in this numerical example, there will be three output classes for a total of three time intervals.

Estimating Temporal Trajectories from TXAI System

The temporal trajectories of a dynamic system can be outlined by the TXAI system by making use of the conditional distribution integrated into the TXAI system. The trajectories of a TXAI system is motivated by the work of Filev et al. [162] that embodies fuzzy transition events defined by joint possibility encompassing the current and future prototypical patterns. More specifically, the TXAI system can delineate a pattern transition matrix (PTM) which will entail the joint possibility of the patterns in present (Δt) and future (Δt^{+}) time intervals. In mathematical terms, for a total of U patterns in time interval Δt, and a total of V patterns in time interval Δt^{+}, the PTM can be written as follows [162]:

$$PTM(\Delta t, \Delta t^{+}) = \begin{bmatrix} \pi_{11} & \cdots & \pi_{1N} \\ \vdots & \ddots & \vdots \\ \pi_{M1} & \cdots & \pi_{GH} \end{bmatrix} \tag{5.36}$$

where π_{ij} is the Pattern Transition Possibility (PTP) for the i^{th} pattern, P_i, in time interval Δt and the j^{th} pattern, P_j, in time interval Δt^{+} as given by (5.37).

$$\pi_{ij} = \eta_{ij} \times \frac{S_{ij}}{S_{\Delta t^{+}}} \tag{5.37}$$

where η_{ij} is the joint possibility for the two patterns to be prototypical in their respective time intervals, and the ratio $\frac{S_{ij}}{S_{\Delta t^{+}}}$ entails the number of times P_i and P_j are observed in their respective time intervals with respect to all V patterns in Δt^{+}. The following equations, (5.38) - (5.40), outline how η_{ij} and the ratio $\frac{S_{ij}}{S_{\Delta t^{+}}}$ are computed.

$$\eta_{ij}(P_{i,\Delta t}, P_{j,\Delta t^{+}}) = \sigma_i(P_i, \Delta t) \times \sigma_j(P_j, \Delta t^{+}) \tag{5.38}$$

Where σ is computed by applying the t-norm operator (product or minimum type) to the conditional distribution values of the antecedents of a given pattern (P) in a given time interval (Δt or Δt^{+}); mathematically expressed as shown in equation (5.39) for pattern (P_q) in time interval (Δt). The computation of the conditional distribution, f, is previously outlined in Section 5.3.3 (in particular see (5.13)).

$$\sigma_q(P_{q,\Delta t}) = f_q(\Psi_{1,P_q}, \Delta t) \times f_q(\Psi_{2,P_q}, \Delta t) \times \ldots \times f_q(\Psi_{\Phi,P_q}, \Delta t) \tag{5.39}$$

where Φ is the total number of antecedents (Ψ) of pattern P_q. The elements for computing the ratio $\frac{S_{cd}}{S_{\Delta t^{+}}}$ are outlined in (5.40):

$$\begin{aligned} S_{ij} &= \sum P_{j,\Delta t} P_{j,\Delta t^{+}} \\ S_{\Delta t^{+}} &= \sum_{j=1}^{H} P_{j,\Delta t^{+}} \end{aligned} \tag{5.40}$$

where the numerator, S_{ij}, represents the sigma count of the number of times P_j and P_j are observed in their respective time intervals, and the denominator, $S_{\Delta t^+}$, denotes the sigma count of observing all H patterns in Δt^+.

Chapter Six

Applications of Proposed Explainable AI (XAI) Methods

In this chapter, I present the applications of the proposed XAI methods (outlined in Chapter 5) developed in this work. More specifically, the applications for each of the proposed XAI method are:

1. Application of EFCMs on adults' fNIRS based skill acquisition study for the implementation of neural reuse framework.

2. Application of xMVPA on infants' fNIRS based perceptual stimuli study for the implementation of Interactive Specialisation (IS) framework.

3. Application of TXAI on real-life temporal, room occupancy dataset for potential delineation of brain developmental trajectories.

The rest of the chapter outlines the results of the XAI methods on their respective applications, and presents their implications for delineating functional brain development as hypothesised by the DCN frameworks (previously outlined in Chapter 2.1). Also, for both EFCM and xMVPA which have been exemplified on fNIRS data, the analysis for both oxy-Hb and deoxy-Hb are presented to follow best practices in fNIRS studies (previously discussed in Chapter 2.2.3).

6.1 Application of EFCMs on skill acquisition study

In this section[5], the efficacy of the proposed EFCM for deciphering the change in EC on account of neural reuse framework is validated using an adults' fNIRS study [163]. The reason this particular dataset is chosen to exemplify the neural reuse phenomenon is because the participants can be categorised based on their expertise level for performing a complex visual-spatial task (namely laparoscopic surgery (LS)). The neural reuse phenomenon suggests that existent cortical circuits for computation of a given task have the ability to reconfigure themselves in multiple configurations as an infant gains experience. Hence, a comparison of the deciphered EC values by EFCM for participants with different level of expertise (for example novices vs. experts) would shed light on the different configurations of the cortical networks formed. Although this is an adult fNIRS study, the participants have varying level of expertise therefore an EFCM analysis of EC based on the level of expertise can still shed light on how cortical networks get rewired upon acquisition of a skill, resembling developmental processes. This is in fact similar to the reconfiguration of the cortical networks in infants on the onset of new cognitive abilities [164, 165] as hypothesised by the neural reuse framework (which suggests the rewiring of existing cortical networks upon experience) as previously outlined in Chapter 2.1.

The fNIRS dataset [163] also entails separate records for oxy-Hb and deoxy-Hb levels in the investigated brain regions. Hence, in order to follow the best practices in fNIRS studies [84], the EC within the 16 real fNIRS signals for both the oxy-Hb fNIRS signal and deoxy-Hb fNIRS signals are analysed separately. This is also important because no direct correlation of the two Hb (haemoglobin) dimensions of fNIRS has been established in the literature yet [166, 167, 168]. However, the

[5] Some parts of the text in this section have been published here: 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, Early Access © 2019 IEEE.

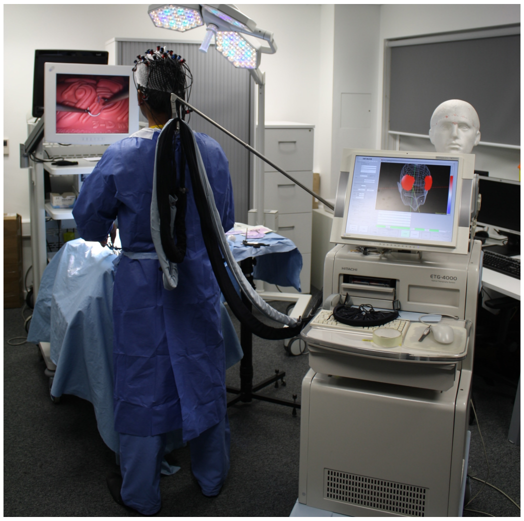


Figure 6.1 Brain activity being recorded via fNIRS whilst participants perform LS task, © 2019 IEEE.

discrepancies between the oxy-Hb and deoxy-Hb signal may be meaningful to elucidate transient neural activity as revealed in [169].

For the adult fNIRS study, a local research ethics committee approval was obtained (project number: 05/Q0403/142). More specifically, the study involved 27 right-handed male surgeons affiliated with National Health Service, and Imperial College London. Amongst the 27 surgeons there were 9 Novice (NVs), 10 Trainee (TNs), and 8 Expert (EXs) consultants. The participants performed the LS whilst their brain activity is recorded, as shown in Fig. 6.1. The reader is referred to [163] for a more detailed description of the participants, the task, and the pre-processing of the fNIRS signals.

A total of 16 fNIRS signals are analysed in this study. The fNIRS digitized probe positions were registered from a real-world coordinate to the MNI space. The MNI coordinates were transformed to Talairach space [170], and looked up in a brain atlas [171] to establish their relations with the Region of Interest (ROI). A detailed explanation of the fNIRS probe positions transformation is provided in the study by [163]. The measured channels locations are as shown in Fig. 6.2.

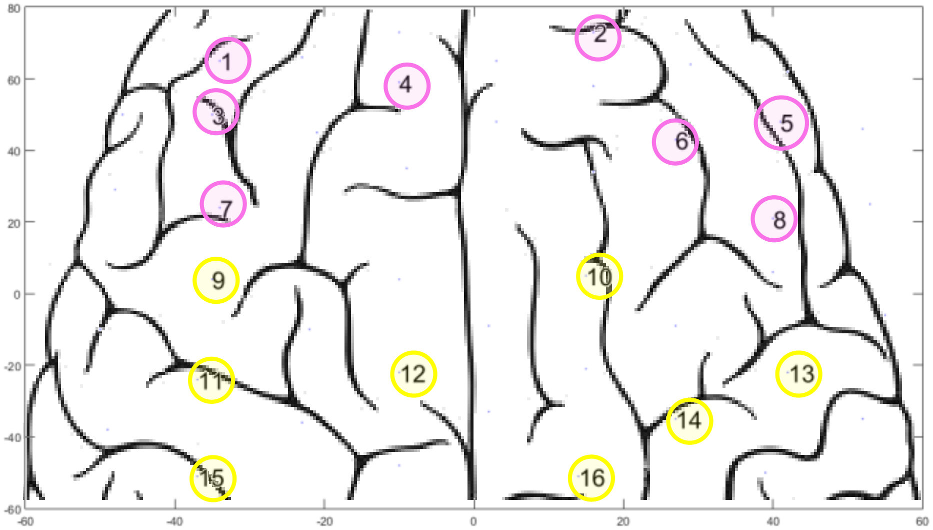


Figure 6.2 The 16 fNIRS channel position differentiated based on ROIs- Prefrontal Cortex (PFC) in pink, and Motor Cortex (MC) in yellow, © 2019 IEEE.

In this work, a third order EFCM is used, i.e. $\wp = 3$ in (5.1). The EFCM learning is undertaken using the GA such that the resultant connectivity matrix, $\mathbf{D} \odot \mathbf{Z}$, depicted the inherent EC of the 16 synthetic fNIRS signals as close as possible. The population size of the GA is set to 1000, the maximum number of generations is defined as 1000 x the population. The maximum fitness is set to 0.99 and the tolerance criterion is set to $1e^{-1}$. The genetic operator used is crossover and the selection is made by tournament. A detailed description of the GA parameters can be found in [147, 150]. Moreover, the GA learning is driven by the *fitness* of the predicted state $\hat{\chi}(\tau)$ with respect to the original state $\chi(\tau)$. The fitness of the predicted state $\hat{\chi}(\tau)$ is evaluated according to the fitness function in (5.6).

6.1.1 Results

In this section, the results of the application of the third order EFCM on an adults' fNIRS dataset are presented. The expertise level of the participating subjects was categorised into three levels of NVs, TNs, and EXs. The aim of the study was to learn the difference in EC networks formed with varying levels of proficiency for carrying out a task that requires active planning, and visual-motor coordination.

In order to facilitate the comparison of cortical network reconfiguration, as an individual gains a certain degree of expertise in doing LS, Fig. 6.3 shows a plot of the EFCMs generated from oxy-Hb and deoxy-Hb fNIRS signals along with corresponding combined EFCM for NVs, TNs, and EXs. 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. Please note only the most significant connections are shown in Fig. 6.3 which have fuzzy degrees of relationship greater than 90th percentile.

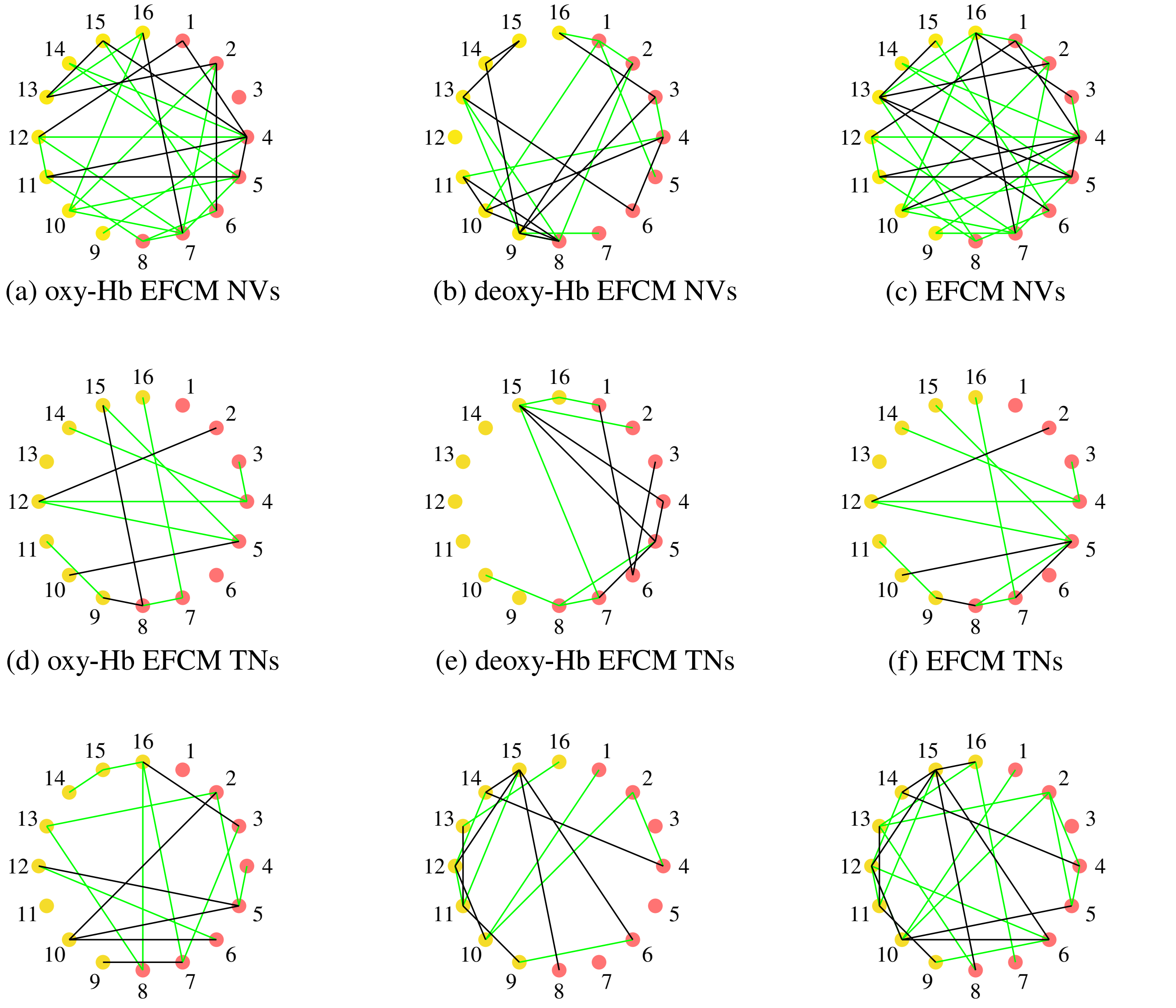
A noteworthy observation that can be made from Fig. 6.3 is that as the expertise level increases, the number of significant EC (strength of fuzzy degrees of relationship greater than 90th percentile) is more for deoxy-Hb EFCMs as compared to oxy-Hb EFCMs signifying that perhaps deoxy-Hb signals hold more latent information with regards to channels underlying EC as compared to oxy-Hb signals if the brain networks have evolved owing to more experience. The quiescent EC structure within the deoxy-Hb signals for more experienced subjects could also explain why a greater number of t-test H_0 got rejected in Table 6.2 against the corresponding oxy-Hb channels.

The error between the estimated signal and real signal using the learnt EC weights by EFCM is computed using (5.5). The average error values (5.5) along with corresponding standard deviations comparing the accuracy in estimating fuzzy effective connections between oxy-Hb and deoxy-Hb by proposed EFCM are listed in Table 6.1. Essentially, the Hb group EFCM simulation for which the average error is greater, its weights are scaled down by the corresponding percentage hence taking forward more of the accurate EC estimates to contribute to the combined EFCM. For example, for NVs, the oxy-Hb weights are scaled down by 3.1% before being mapped into the overall EFCM.

	Novices (NVs)	**Trainees (TNs)**	**Experts (EXs)**
oxy-Hb	120.25 ± 29.85	122.79 ± 32.89	156.38 ± 46.52
deoxy-Hb	116.62 ± 27.14	119.99 ± 43.44	125.49 ± 73.45
Error	Oxy-Hb by 3.1%	Oxy-Hb by 2.3%	Oxy-Hb by 19.8%

Table 6.1 Average error values for oxy-Hb and deoxy-Hb fNIRS simulation by $\wp = 3$ EFCM, © 2019 IEEE.

In order to assess the EFCM results for statistical significance, Table 6.2 lists the number of the t-test H_0 being accepted for all subjects of category NVs, TNs, and EXs. The null hypothesis is defined as H_0: The original and predicted signals connectivity values have the same mean at 95% confidence level. The results indicate that for most of the channels, across all subjects and expertise levels, the H_0 is accepted indicating the prowess of the proposed higher order EFCM for delineating the EC in fNIRS data. However, the results in Table 6.2 indicate that underlying effective connections in oxy-Hb results are better understood by the proposed third order EFCM in comparison to the deoxy-Hb for less experienced subjects. However, the accuracy of the predicted signals, for both oxy-Hb and deoxy-Hb signals, can also be seen to decline as the expertise level increases. This could perhaps be owing to changes in the memory order of the EC of the underlying channels on account of increased experience but needs further investigation since, in this work, the order of the EFCM method is not varied with the change in expertise level.



(a) oxy-Hb EFCM NVs

(b) deoxy-Hb EFCM NVs

(c) EFCM NVs

(d) oxy-Hb EFCM TNs

(e) deoxy-Hb EFCM TNs

(f) EFCM TNs

(g) oxy-Hb EFCM EXs

(h) deoxy-Hb EFCM EXs

(i) EFCM EXs

Figure 6.3 The third order EFCM networks showing the reconfiguration of cortical networks as hypothesised by the neural reuse framework, © 2019 IEEE.

	$\mathbf{H_0/Subj}$					
	Novices (NVs)		**Trainees (TNs)**		**Experts (EXs)**	
Ch.	**oxy-Hb**	**deoxy-Hb**	**oxy-Hb**	**deoxy-Hb**	**oxy-Hb**	**deoxy-Hb**
1	100.0%	88.9%	90.0%	90.0%	87.5%	75.0%
2	100.0%	100.0%	90.0%	90.0%	87.5%	100.0%
3	100.0%	100.0%	100.0%	100.0%	100.0%	87.5%
4	100.0%	88.9%	80.0%	90.0%	75.0%	87.5%
5	100.0%	77.8%	100.0%	100.0%	62.5%	87.5%
6	100.0%	100.0%	90.0%	80.0%	75.0%	100.0%
7	100.0%	77.8%	90.0%	100.0%	62.5%	75.0%
8	100.0%	100.0%	70.0%	80.0%	87.5%	100.0%
9	100.0%	100.0%	90.0%	100.0%	75.0%	87.5%
10	100.0%	100.0%	90.0%	80.0%	87.5%	87.5%
11	100.0%	88.9%	90.0%	90.0%	100.0%	100.0%
12	100.0%	77.8%	100.0%	80.0%	50.0%	62.5%
13	100.0%	88.9%	70.0%	80.0%	87.5%	75.0%
14	100.0%	88.9%	80.0%	70.0%	87.5%	87.5%
15	100.0%	88.9%	90.0%	90.0%	87.5%	87.5%
16	100.0%	88.9%	80.0%	90.0%	62.5%	87.5%

Table 6.2 Percentage ratio of No. of t-Test H_0 accepted to No. of Subjects (Subj), $H_0/Subj$, for oxy-Hb and deoxy-Hb of 9 NVs, 10 TNs, and 8 EXs at 95% confidence level, © 2019 IEEE.

6.1.2 Discussion

In this work the adults' fNIRS data [163] entailed subjects which differed in their level of expertise for performing a pre-defined LS task. The fNIRS data recorded, whilst the subjects performed LS task, is used to assess the efficacy of the innovative, regularised EFCM method presented. In this section, the results obtained for EC between ROIs from the proposed EFCM method are discussed against similar works in the literature.

The results from third order EFCM indicate network connections change from random activations to evenly spread out along with more positive influences as the expertise level increase as shown in Fig. 6.3. In addition, there are more long-range connections in both PFC and motor cortex for EXs and TNs in comparison to NVs. This is in agreement with the current findings [172, 173] as well as the results from the original study [163] that as an individual progress in learning, their brain networks evolve and optimise their connections.

In particular, the original study [163] reported statistically significant differences in correlation (Rv) coefficients between the Hb (haemoglobin) responses from different brain regions (prefrontal and motor cortices) based on the expertise levels of the participants. More specifically, the Rv coefficients are an approximation of the the squared Pearson correlation coefficient [174]. Although the results in [163] suggest different correlations between the Hb responses from prefrontal and motor cortices, the results can not inform about the underlying EC between different brain regions for carrying out the LS task based on the expertise level of the participants.

In this work, another perspective for underpinning the EC analysis in between ROIs is done by collapsing the weights according to PFC and motor cortex as shown in Fig. 6.4. This is done by first averaging the weights for all subject's oxy-Hb and deoxy-Hb of a given expertise level, and subsequently finding the mean of the average values for the individual Hb's ROIs weights. The resultant averaged values are then scaled to unit length for each connection, i.e. PFC to PFC, PFC to motor cortex, motor cortex to PFC, and motor cortex to motor cortex across the three expertise level. A similar trend can also be seen in Fig. 6.4, with NVs relying more on inter PFCs and almost non-dependent on motor cortex's connections as compared to TNs and EXs. TNs and EXs rely more on inter motor cortex connections and less on inter PFC with progression towards a balanced corroboration between ROIs as more experience is gained, hence spreading the cognitive load in contrast to NVs [172, 163]. This progression trend with increased experience is also intercepted well by EFCM as can be seen in Fig. 6.3 with more strong positive cause-effect relations between PFC and motor cortex for TNs and EXs as compared to NVs.

The shift in brain activations from PFC to motor cortex on account of increased experience is also in line with the findings from the work of [175] which observed a decrease in EC in frontal pathway encompassing the regions of inferior frontal gyrus and precentral gyrus brain as a particular visual-motor coupled task is learnt. Another distinct work by [176] - focusing on the particular role that motor cortex takes whilst a certain task is learnt - concludes that motor cortex is critical for learning of a task but not essential for execution of a previously learnt motor task. They conclude that motor cortex engages sub-cortical motor cortex circuits upon learning of a task and assigns them the execution of the learnt task. This account frees the motor cortex for any new learning activity and is in coherence with the findings of this work that brain networks evolve to a more balanced configuration, on acquisition of a certain level of expertise, without draining any particular segment of the brain.

The work by [177] also reports a similar trend, in a non-human primate study, of high dependency on PFC when the participant is inexperienced with the task undertaken.The task carried out by the participant in the study [177] was to switch on flashing lights by pressing corresponding buttons. Once the participant learnt the task, the button and flashing light sequence is changed so as to coerce it to suppress its learnt behaviours. The study concluded that PFC regions are more involved during the learning of a task whereas basal ganglia takes over when a goal directed behaviour is required.

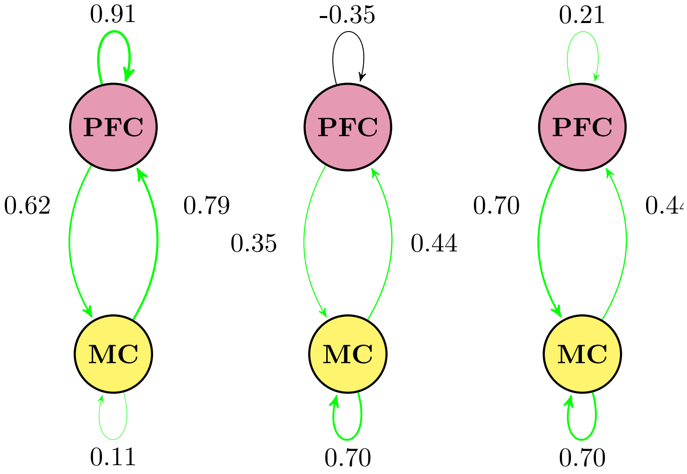


Figure 6.4 EC strength between ROIs: PFC and motor cortex for NVs (left), TNs (center), and EXs (right), © 2019 IEEE.

6.1.3 Implications for functional brain development analysis

Owing to the differences in discerned EC by the EFCM, on the basis of varying level of expertise of participants, EFCM is able to shed light on the differences in cortical networks' interconnections upon skill acquisition. When applied to DCN studies for estimating EC, EFCMs can inform us on how cortical networks rewire, in terms of their interconnections (EC), with the onset of new cognitive abilities. The elucidation of cortical network reorganisation would help understand functional brain development as hypothesised by the neural reuse framework.

The EFCMs results from adults' fNIRS skill acquisition study suggests that participants with lower expertise level tend to activate the PFC more than the participants with higher expertise level, whereas the motor cortex engagement increases upon greater dexterity in performing the motor task (LS) (Fig. 6.4). The reconfiguration of adults' cortical networks upon learning the motor skill, as suggested by EFCM, is similar to infants' literature on cortical network optimisation upon acquisition of motor skills [164]. In particular, in the fNIRS based study by [165], the authors have shown an increase in motor cortex activity and a decrease in DLPFC activity (by means of higher oxy-Hb concentration) as infants aged 5 to 13 months learn motor skills.

The parallels between the EFCM analysis of adults fNIRS data upon skill acquisition, and that of infants' literature on onset of motor functions, suggest that EFCM is a viable mechanism to study functional brain development as per the neural reuse framework. Although EFCMs estimated values of EC can be mapped generally to the neural reuse of the DCN processes, not much insight can be gained about the activation(s) of the individual cortical regions. In this sense, the EFCMs propose a partially explainable method in terms of estimating the EC as fuzzy weights between its concepts (fNIRS channels). 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 addition to estimating EC with statistical significance, EFCM results also demonstrated the prowess to analyse the difference between estimating EC from oxy-Hb and deoxy-Hb 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; EFCM results suggest that EC estimated using deoxy-Hb is more representative of the underlying EC as an individual gains experience in a certain motor task (Table 6.1 and 6.2).

6.2 Application of xMVPA in DCN

In the present work[6], the xMVPA inference mechanism is applied for the explainable classification analysis of infant fNIRS data obtained from an earlier study by [101]. In this study, fNIRS was used to record six-month-old infants' haemodynamic responses to auditory (a toy sound) and visual stimuli (a dynamic red smiley face). In between the trials, a jittered video of dimmed fireworks was displayed. A schematic representation of the auditory and visual trials is shown in Fig. 6.5 (b). The infant fNIRS data was recorded from 10 channels (see Fig. 6.5 (a) for the anatomical locations of the 10 fNIRS channels after MR co-registration).

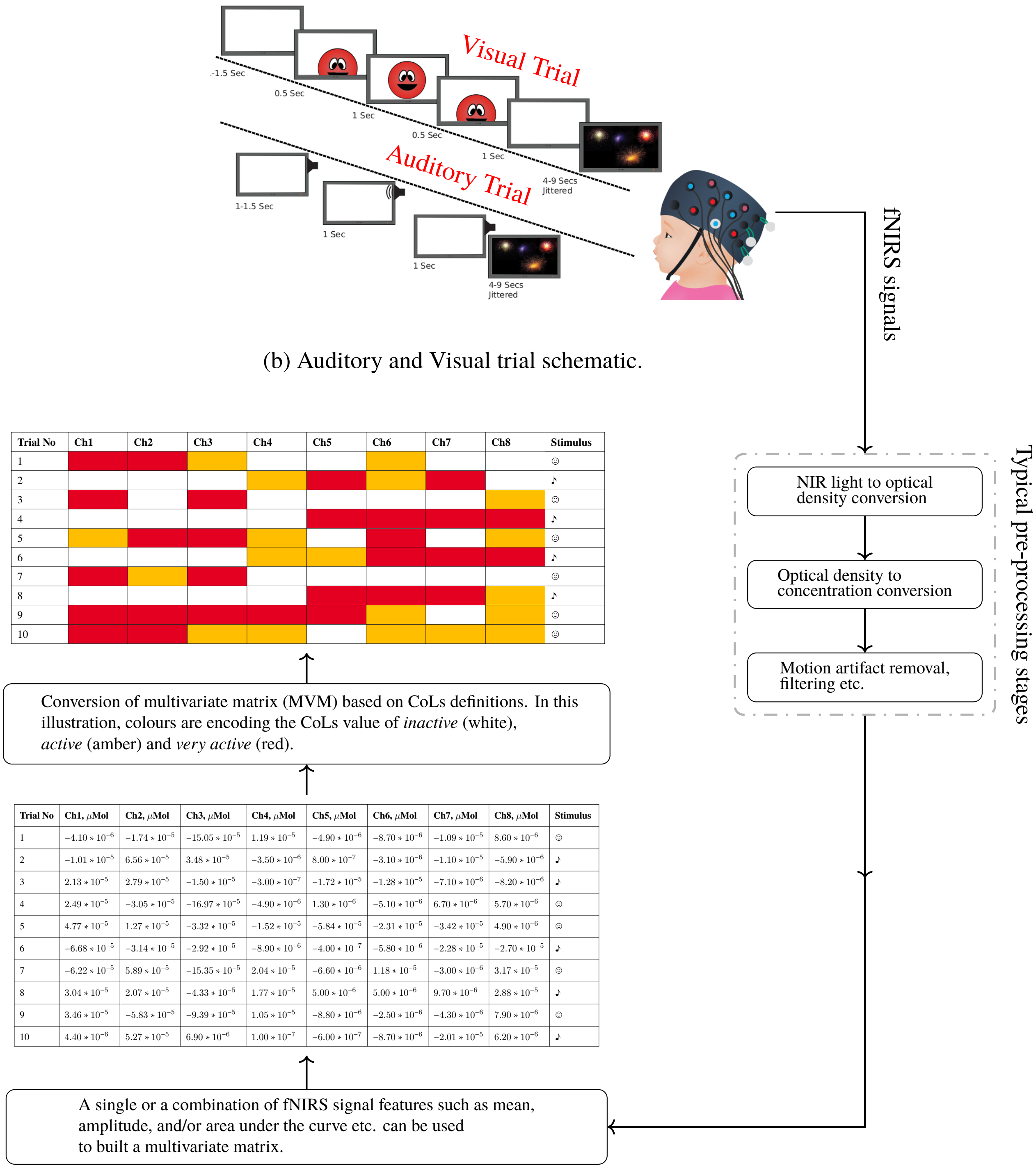
In the present work, a MVM is constructed by calculating the mean of the oxy-Hb signal for each of the 10 channels from time 4-7 seconds post stimulus presentation for each trial. The rows in the MVM consist of all the trials with each entry in the 2-dimensional matrix (for row (*i*) and column (*j*) being the average of the j^{th} channel activity from time 4-7 s post stimulus for the i^{th} trial. Please see Fig. 6.5 (c) that outlines the steps for the construction of a MVM.

The xMVPA identifies informative activation patterns by combining the input neuroimaging data from all fNIRS channels of interest into a MVM (multivariate matrix). Here, the MVM is constructed by calculating the mean of the oxy-Hb signal from each of the 10 channels in the time-window 4-7s, following stimulus presentation for each trial. A grid-search was undertaken to find the optimal time window of 4-7s. In line with previous infant fNIRS studies [178], and as reported by Emberson et al. [101], the focus is on examining the oxy-Hb signals. Nevertheless, there will be no changes in the proposed xMVPA method for using either or both of the dimensions of the fNIRS signals to construct the MVM. Moreover, the results of the proposed xMVPA on deoxy-Hb signals are provided

[6] Some parts of the text in this section have been published here: J. Andreu-Perez, L. L. Emberson, M. Kiani, M. L. Filippetti, H. Hagras, and S. Rigato, "Explainable Artificial Intelligence Based Analysis for Developmental Cognitive Neuroscience," *Commun Biol*, vol. 4, p. 1077, 2021 © 2021 Nature.

Ch. Number	**Ch1**	**Ch2**	**Ch3**	**Ch4**	**Ch5**
Lobar Atlas	Occipital	Occipital	Parietal	Temporal	Temporal
LONI Atlas	Middle occipital gyrus	Middle occipital gyrus	Middle occipital gyrus	Middle temporal gyrus	Middle temporal gyrus
Ch. Number	**Ch6**	**Ch7**	**Ch8**	**Ch9**	**Ch10**
Lobar Atlas	Temporal	Temporal	Frontal	Temporal	Frontal
LONI Atlas	Superior temporal gyrus	Middle temporal gyrus	Inferior frontal gyrus	Superior temporal gyrus	Postcentral gyrus

(a) The anatomical location of the 10 fNIRS channels (Ch) in the study by [101].


Visual Trial
-1.5 Sec
0.5 Sec
1 Sec
0.5 Sec
1 Sec
4-9 Secs
Jittered
Auditory Trial
1-1.5 Sec
1 Sec
1 Sec
4-9 Secs
Jittered
fNIRS signals
Typical pre-processing stages
NIR light to optical density conversion
Optical density to concentration conversion
Motion artifact removal, filtering etc.

(b) Auditory and Visual trial schematic.

Trial No	Ch1	Ch2	Ch3	Ch4	Ch5	Ch6	Ch7	Ch8	Stimulus
1	🟥	🟥	🟨			🟨			☺
2				🟨	🟥	🟨	🟥		♪
3	🟥		🟥					🟨	☺
4					🟥	🟥	🟥	🟥	♪
5	🟨	🟥	🟥	🟨		🟥		🟨	☺
6				🟨	🟨	🟥	🟥	🟥	♪
7	🟥	🟨	🟥						☺
8					🟥	🟥	🟥	🟨	♪
9	🟥	🟥	🟥	🟥	🟥	🟨		🟨	☺
10	🟥	🟥	🟨	🟨		🟨	🟨	🟨	☺

Conversion of multivariate matrix (MVM) based on CoLs definitions. In this illustration, colours are encoding the CoLs value of *inactive* (white), *active* (amber) and *very active* (red).

Trial No	Ch1, μMol	Ch2, μMol	Ch3, μMol	Ch4, μMol	Ch5, μMol	Ch6, μMol	Ch7, μMol	Ch8, μMol	Stimulus
1	$-4.10 * 10^{-6}$	$-1.74 * 10^{-5}$	$-15.05 * 10^{-5}$	$1.19 * 10^{-5}$	$-4.90 * 10^{-6}$	$-8.70 * 10^{-6}$	$-1.09 * 10^{-5}$	$8.60 * 10^{-6}$	☺
2	$-1.01 * 10^{-5}$	$6.56 * 10^{-5}$	$3.48 * 10^{-5}$	$-3.50 * 10^{-6}$	$8.00 * 10^{-7}$	$-3.10 * 10^{-6}$	$-1.10 * 10^{-5}$	$-5.90 * 10^{-6}$	♪
3	$2.13 * 10^{-5}$	$2.79 * 10^{-5}$	$-1.50 * 10^{-5}$	$-3.00 * 10^{-7}$	$-1.72 * 10^{-5}$	$-1.28 * 10^{-5}$	$-7.10 * 10^{-6}$	$-8.20 * 10^{-6}$	♪
4	$2.49 * 10^{-5}$	$-3.05 * 10^{-5}$	$-16.97 * 10^{-5}$	$-4.90 * 10^{-6}$	$1.30 * 10^{-6}$	$-5.10 * 10^{-6}$	$6.70 * 10^{-6}$	$5.70 * 10^{-6}$	☺
5	$4.77 * 10^{-5}$	$1.27 * 10^{-5}$	$-3.32 * 10^{-5}$	$-1.52 * 10^{-5}$	$-5.84 * 10^{-5}$	$-2.31 * 10^{-5}$	$-3.42 * 10^{-5}$	$4.90 * 10^{-6}$	☺
6	$-6.68 * 10^{-5}$	$-3.14 * 10^{-5}$	$-2.92 * 10^{-5}$	$-8.90 * 10^{-6}$	$-4.00 * 10^{-7}$	$-5.80 * 10^{-6}$	$-2.28 * 10^{-5}$	$-2.70 * 10^{-5}$	♪
7	$-6.22 * 10^{-5}$	$5.89 * 10^{-5}$	$-15.35 * 10^{-5}$	$2.04 * 10^{-5}$	$-6.60 * 10^{-6}$	$1.18 * 10^{-5}$	$-3.00 * 10^{-6}$	$3.17 * 10^{-5}$	☺
8	$3.04 * 10^{-5}$	$2.07 * 10^{-5}$	$-4.33 * 10^{-5}$	$1.77 * 10^{-5}$	$5.00 * 10^{-6}$	$5.00 * 10^{-6}$	$9.70 * 10^{-6}$	$2.88 * 10^{-5}$	♪
9	$3.46 * 10^{-5}$	$-5.83 * 10^{-5}$	$-9.39 * 10^{-5}$	$1.05 * 10^{-5}$	$-8.80 * 10^{-6}$	$-2.50 * 10^{-6}$	$-4.30 * 10^{-6}$	$7.90 * 10^{-6}$	☺
10	$4.40 * 10^{-6}$	$5.27 * 10^{-5}$	$6.90 * 10^{-6}$	$1.00 * 10^{-7}$	$-6.00 * 10^{-7}$	$-8.70 * 10^{-6}$	$-2.01 * 10^{-5}$	$6.20 * 10^{-6}$	♪

A single or a combination of fNIRS signal features such as mean, amplitude, and/or area under the curve etc. can be used to built a multivariate matrix.

(c) Multivariate pattern matrix construction.

Figure 6.5 A schematic for the construction of a multivariate pattern matrix, © 2021 Nature.

in Section 6.2.1. A total of 19 babies' data is included in the analysis, with multiple trials per baby, amounting to a total of 524 trials. Experimental control and signal assessment were performed to avoid the inclusion of any possible noise artifacts or covariates in our data [146, 101]. The reader is referred to the earlier study by [101] for more details on the experimental setup, data collection, sample, control, exclusion and the subsequent pre-processing steps.

The numerical neuroimaging data in the MVM is then translated into conceptual labels of brain activation defined as *inactive*, *active*, and *very active* to represent the average activity level of each channel for the time-window considered. A flow chart outlining the steps for generating a multivariate pattern matrix with conceptual labels is presented in Fig. 6.5 (c). The data instances in the multivariate pattern matrix characterised by the conceptual labels for each trial are subsequently used to train the xMVPA for explainable classification results of the infant data in response to the visual and auditory stimuli. More details of the proposed xMVPA inference mechanism are provided as follows.

In this work, the evaluation of xMVPA is performed by splitting the observations transformed into the conceptual MVM into 5 mutually-exclusive train and validation sets (viz. k-fold cross-validation). The patterns are initially generated at random with the maximum number of patterns in a given set to be 20, and the maximum number of channels (or antecedents) in a given pattern to be 3, i.e. a given pattern would outline interactions from a maximum of 3 channels/brain regions. The small number of patterns, with short antecedents, ensures that a given set of patterns is comprehensive and easily interpretable [120, 153].

6.2.1 Results

In the following sections, the results of the xMVPA of [101] using oxy-Hb and deoxy-Hb signals are presented.

Oxygenated Haemoglobin (oxy-Hb) Results

xMVPA revealed six functional patterns of interactions between cortical regions using the publicly available DCN dataset of auditory versus visual stimulus processing[7]. The patterns form the inference mechanism for xMVPA as they predict the stimulus (or class) for the brain activity instances (or data instances).

[7] Data available at: https://dataspace.princeton.edu/handle/88435/dsp01xs55mf543

The six patterns provided by xMVPA that outline the brain regions' activation and interaction for processing visual and auditory information are given below:

Pattern P_1 : IF Ch1 is $Active$ AND Ch2 is $Active$ AND Ch4 is $Active$
THEN stimulus is $Visual$ with dominance score 0.581

Pattern P_2 : IF Ch4 is $Active$ AND Ch6 is $Inactive$ AND Ch8 is $Very\ Active$
THEN stimulus is $Visual$ with dominance score 0.019

Pattern P_3 : IF Ch1 is $Inactive$ AND Ch8 is $Active$
THEN stimulus is $Auditory$ with dominance score 0.434

Pattern P_4 : IF Ch4 is $Inactive$ AND Ch5 is $Active$
THEN stimulus is $Auditory$ with dominance score 0.406

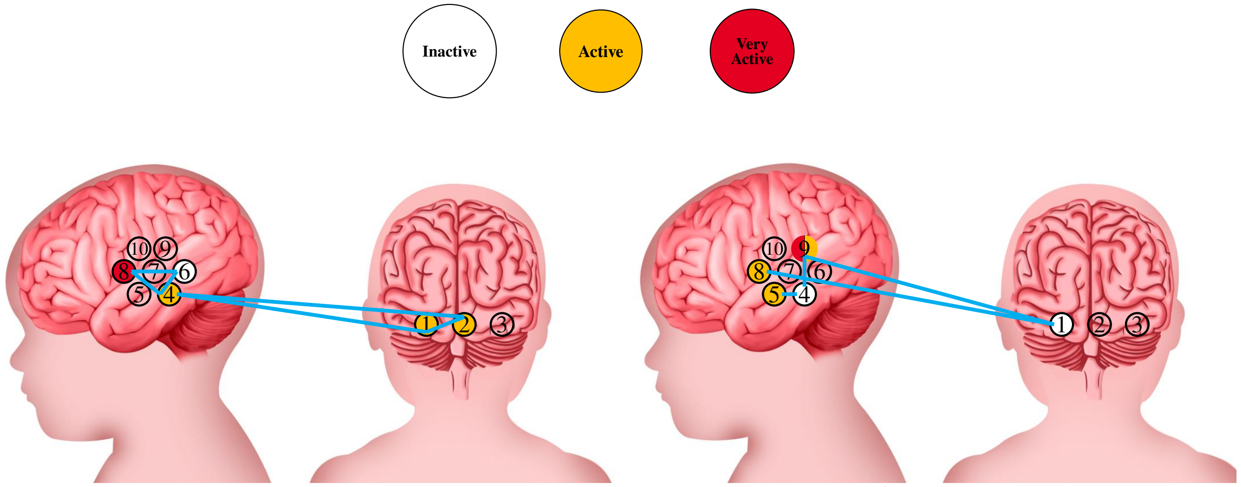
Pattern P_5 : IF Ch4 is $Inactive$ AND Ch9 is $Very\ Active$
THEN stimulus is $Auditory$ with dominance score 0.239

Pattern P_6 : IF Ch1 is $Inactive$ AND Ch9 is $Active$
THEN stimulus is $Auditory$ with dominance score 0.082

where dominance score (DS) is in the range $(0,1)$. Dominance Score of a pattern indicates the overall information prowess of a given pattern with DS=1 being the most informative pattern and DS=0 being the least informative pattern. (6.1)

Patterns P_1 and P_2 identified interactions between regions involved in the processing of the visual stimulus, as shown in Figure 6.6 (a). Firstly, P_1 showed a prominent involvement of the occipital cortex, where channel 1 and channel 2 are both classified as *active*. Secondly, both P_1 and P_2 identified an *active* status of channel 4, located in the temporal cortex (see Fig. 6.5 (a)). Finally, P_2 identified an *inactive* status of channel 6 in the temporal cortex in combination with an *active* status of channel 4 and a *very active* status of channel 8, located in the frontal cortex.

The patterns of interactions in response to the auditory stimulus are shown in Fig. 6.6 (b). Here, channels that are active correspond to the PFC (channel 8 in P_3), and temporal cortex (channel 5 in P_4 and channel 9 in P_5 and P_6). The occipital cortex is not engaged in the processing of auditory stimulus as indicated by the *inactive* status of channel 1 in combination with both PFC (channel 8 *active* in P_3) and temporal cortex (channel 9 *active*) in P_6 activation.



(a) Patterns for visual stimulus.

(b) Patterns for auditory stimulus.

Figure 6.6 The patterns (cyan) identified by the xMVPA delineate the contributions between brain regions evoked by (a) visual and (b) auditory stimuli, © 2021 Nature.

Taken altogether, the patterns identified by the xMVPA show activation over the occipital and temporal cortices for visual stimulus processing, and over the temporal cortex for auditory stimulus processing. The patterns also identified activity over the frontal cortex for the processing of both auditory and visual stimuli.

Another important observation from the patterns in P_1 - P_6 is that no individual channel in the temporal cortex with sufficient decoding accuracy stood out for processing the auditory stimulus presented to the infants in the study, i.e. no channel had stimulus-specific activation (for example

Statistical performance measures

Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$

Fscore = $2 \times \frac{TP}{2{\times}TP + FP + FN}$

Positive Predictive Value (PPV) = $\frac{TP}{TP + FP}$

Negative Predictive Value (NPV) = $\frac{TN}{TN + FN}$

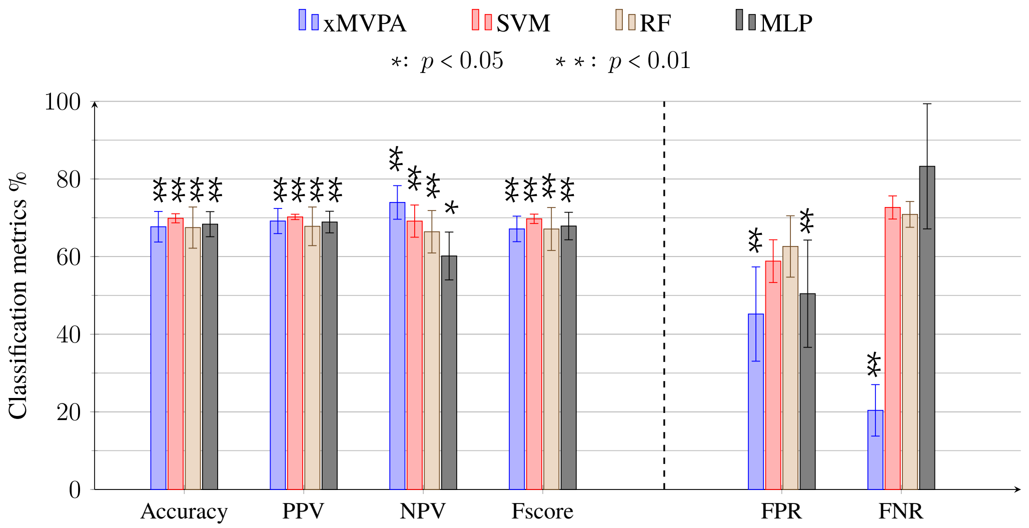
False Positive Rate (FPR) = $\frac{FP}{FP + TN}$

False Negative Rate (FNR) = $\frac{FN}{FN + TP}$

Confusion matrix elements

		Actual	
		Visual Stimulus - **Positive**	Auditory Stimulus - **Negative**
Predicted	Visual Stimulus - **Positive**	True Positive (TP)	False Positive (FP)
	Auditory Stimulus - **Negative**	False Negative (FN)	True Negative (TN)

(a) Measures used for evaluation and comparison.



(b) Decoding performance metrics and its significance levels for xMVPA and black box models.

Figure 6.7 A comparison of xMVPA performance with state-of-the-art classification methods: Support Vector Machine (SVM), Random Forest (RF), and MultiLayer Perceptron (MLP), © 2021 Nature.

active for auditory processing, and *inactive* for visual processing) as reported in Table 6.3. This might be due to a more diffuse cortical activity [179], in line with what is suggested by fMRI and fNIRS studies that report widespread activation in response to auditory stimuli, such as sounds (e.g. [56], [180]), in the infant brain.

The absence of decoding strength in the temporal cortex in response to auditory stimuli is also consistent with the correlation based MVPA analysis reported in [101]. However correlation based MVPA method was unable to specify neither the semantics of such activation difference (i.e. less active or more active) nor the channels combination yielding higher decoding, i.e. just the independent decoding strength and significant activation for each channel, as outlined in Table 6.3. More specifically, as noted in Table 6.3, the correlation based MVPA results inform which channels have decoding strength (Ch1, Ch3, and Ch8) and which channels have significant activation (Ch1, Ch7, Ch9, and Ch10). However, the correlation based MVPA results can not inform about the channel combinations (patterns/ cortical networks) made for the processing of the presented stimuli. The xMVPA overcomes this limitation as it can inform about the prototypical channel combination for the processing of stimuli (in the form of patterns), rendering it suitable for describing the underlying cortical networks as per the IS framework.

A range of statistical performance measures derived from the confusion matrix, outlined in Fig.

Anatomical Location	**Occipital Cortex**			**Temporal Cortex**					**Pre-Frontal Cortex**	
Activation Level	**Ch1**	**Ch2**	**Ch3**	**Ch4**	**Ch5**	**Ch6**	**Ch7**	**Ch9**	**Ch8**	**Ch10**
Decoding Strength	✓		✓						✓	
Significant Activation	✓						✓	✓		✓
Visual Processing	Active	Active		Active		Inactive			Very Active	
Audio Processing	Inactive			Inactive	Active			Active or Very Active	Active	

Table 6.3 A comparison of xMVPA with correlation based MVPA with [101], © 2021 Nature.

6.7 (a), are calculated to quantify the performance of the xMVPA patterns. The confusion matrix helps in assessing the robustness of a given model's inference mechanism by indicating whether or not the model is 'confusing' the classes, i.e. decoding visual stimulus when it is auditory stimulus (or vice versa). Please note, in Fig. 6.7 (a), the visual stimulus is referred to as positive class, and auditory stimulus is referred to as negative class.

The bar graph in Fig. 6.7 (b) shows a comparison of the statistical performance measures (Accuracy, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Fscore, False Positive Rate (FPR) and False Negative Rate (FNR) defined in Fig. 6.7 (a)) between the xMVPA and the state-of-the-art machine learning algorithms SVM, RF, and MLP. The statistical performance measures of accuracy, PPV, NPV, and Fscore for xMVPA are comparable to those obtained for SVM, RF, and MLP. However, the xMVPA outperforms all the other models for the metrics FPR and FNR. The lowest values of FPR and FNR for xMVPA indicate the most robust classification method (also named *decoding model* in MVPA [101]) for the input fNIRS data, i.e. the xMVPA obtains the least fNIRS instances predicted as auditory when they are in factual evoked by visual stimuli and vice versa. Altogether, this comparison confirms that the xMVPA's patterns clearly discern the differences in the fNIRS instances for the six-month-old brain in response to visual and auditory stimuli.

Deoxygenated Haemoglobin (deoxy-Hb) Results

In this section, the results of the xMVPA inference mechanism on the deoxy-Hb signals obtained from [101] are presented. The xMVPA is applied on the MVM formed by calculating the mean of the deoxy-Hb signals from each of the 10 channels in the time-window 4-7s, following stimulus presentation, for each trial. Please note that the construction of the MVM and the xMVPA parameters are identical for both oxy-Hb and deoxy-Hb signals. The xMVPA results for oxy-Hb are presented in Section 6.2.1.

The evaluation of xMVPA on the MVM from deoxy-Hb signals gives an average classification accuracy of 64.88% with a standard deviation of 4.81%. The eight patterns provided by xMVPA that outline the brain regions' activation and interaction for processing visual and auditory information for deoxy-Hb signals are given below:

Pattern P_1 : IF Ch2 is $Active$ AND Ch3 is $Very\ Active$
THEN stimulus is $Visual$ with dominance score 0.02

Pattern P_2 : IF Ch5 is $Very\ Active$ AND Ch8 is $Very\ Active$
THEN stimulus is $Visual$ with dominance score 0.01

Pattern P_3 : IF Ch2 is $Very\ Active$ AND Ch5 is $Very\ Active$
THEN stimulus is $Auditory$ with dominance score 0.67

Pattern P_4 : IF Ch4 is $Very\ Active$ AND Ch7 is $Very\ Active$
THEN stimulus is $Auditory$ with dominance score 0.08

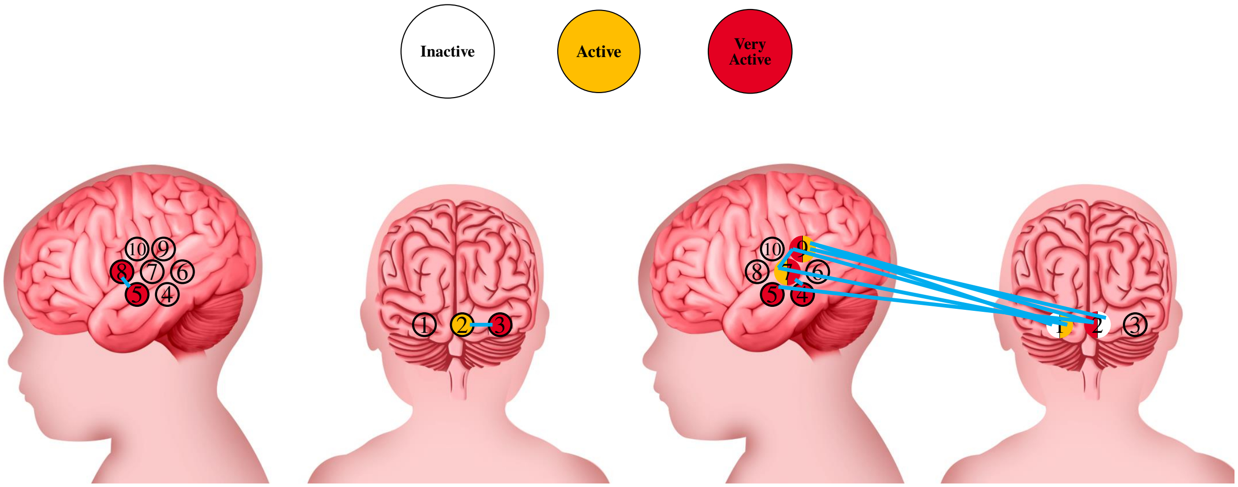
Pattern P_5 : IF Ch2 is $Very\ Active$ AND Ch9 is $Active$
THEN stimulus is $Auditory$ with dominance score 0.08

Pattern P_6 : IF Ch1 is $Inactive$ AND Ch9 is $Active$
THEN stimulus is $Auditory$ with dominance score 0.06

Pattern P_7 : IF Ch2 is $Inactive$ AND Ch9 is $Active$
THEN stimulus is $Auditory$ with dominance score 0.03

Pattern P_8 : IF Ch1 is $Active$ AND Ch7 is $Active$ AND Ch9 is $Very\ Active$
THEN stimulus is $Auditory$ with dominance score 0.02

where dominance score (DS) is in the range (0,1). The greater the value of dominance score the more informative that pattern is with DS= 0 being the least informative pattern.



(a) Patterns for visual stimulus.

(b) Patterns for auditory stimulus.

Figure 6.8 An illustration of the patterns (cyan) identified by the xMVPA, using deoxy-Hb signals,© 2021 Nature.

A total of two patterns, P_1 and P_2, have been identified by xMVPA for the processing of visual information from the deoxy-Hb signals. P_1 describes the contributions of only occipital channels, i.e. channels 2 and 3, whereas P_2 identifies the contributions between channel 5 (temporal cortex) and channel 8 (PFC). However, none of the channels from the occipital and PFC have been found engaged by the xMVPA for visual processing. Moreover, the dominance score of the patterns is almost negligible, i.e. 0.02 and 0.01 for P_1 and P_2 respectively.

For the auditory processing, the xMVPA found six patterns: P_3 - P_8. However, only P_3 is a relevant pattern with a dominance score of 0.67. P_3 uncovers the contributions of occipital and temporal channels as it outlines both channel 2 (occipital cortex) and channel 5 (temporal cortex) to be very active. The remaining patterns, P_4 - P_8, for auditory processing are not as supported as P_3, with P_4 delineating contributions within the temporal cortex (channel 4 and channel 7) and P_5 - P_8 outlining contributions between the occipital and temporal cortices. Unlike the patterns found for visual processing, none of the patterns for the auditory processing outline the contributions from the PFC.

An illustration of the patterns for both visual and auditory processing using deoxy-Hb signals is shown in Fig. 6.8.

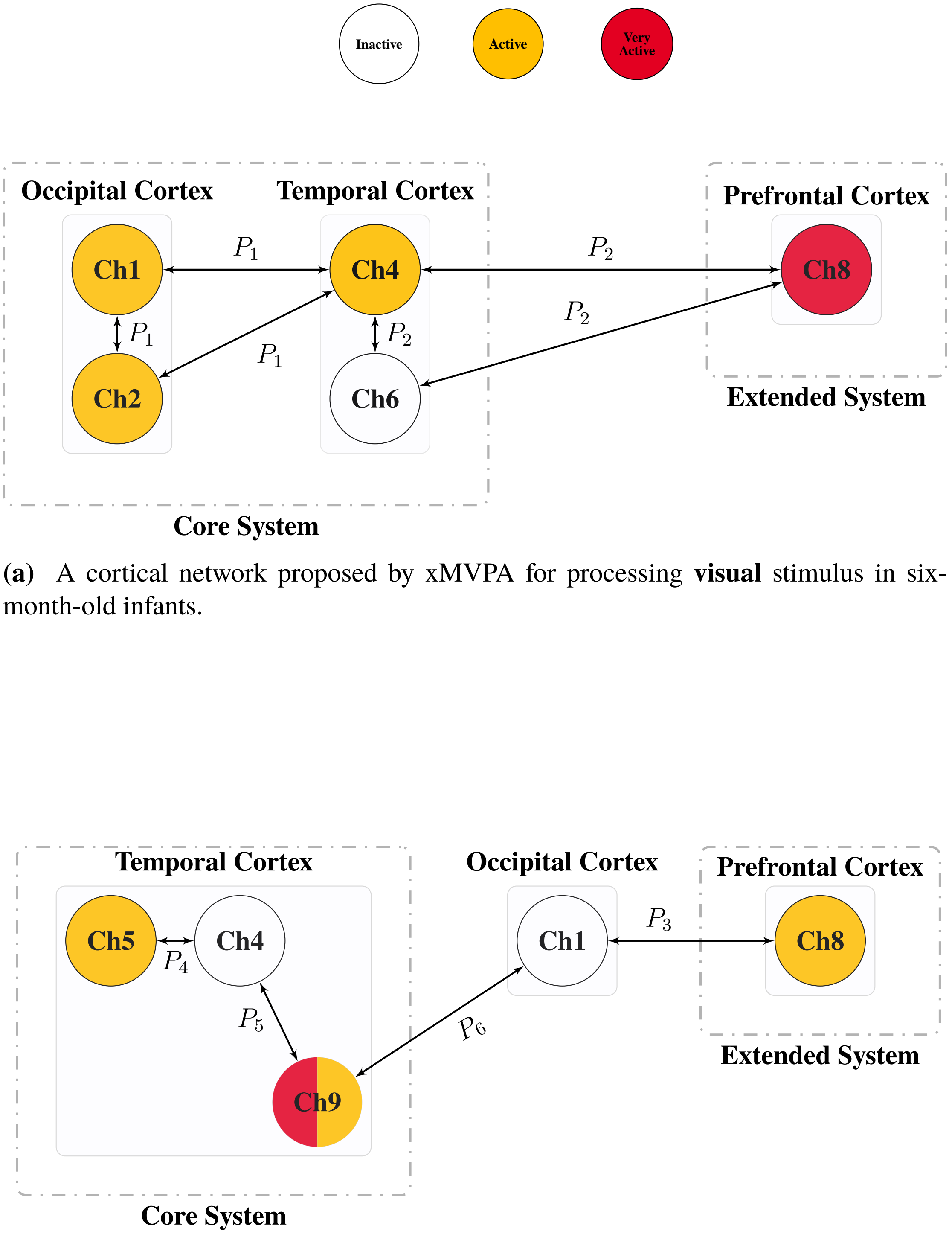
6.2.2 Discussion

In this chapter, an explainable method for analysing and interpreting infant fNIRS data is presented. The proposed xMVPA is an MVPA based on XAI that provides functional patterns characterised by conceptual labels delineating contributions between brain regions for information processing. The xMVPA is applied for the analysis of a group of six-month-old infants' brain activity in response to visual and auditory stimuli [101], and identified six patterns of cortical networks.

xMVPA Patterns found in oxy-Hb fNIRS Signals

The results showed that the classification accuracy obtained on the infant fNIRS dataset by the proposed xMVPA is comparable to the state-of-the-art machine learning algorithms frequently used for MVPA (e.g. SVM, RF, and MLP; see Fig. 6.7 (b)) using oxy-Hb fNIRS signals, thus demonstrating the validity of our model. This is of critical importance for an advancement in DCN because, in contrast to our xMVPA, the classification process of these standard machine learning algorithms is opaque [17, 18] and thus cannot inform our understanding of the developing brain.

The validity and efficacy of the proposed xMVPA method are also demonstrated against the correlation based MVPA presented in the previous study by [101]. As reported in Table 6.3, channel 1 is the only channel to have both decoding strength in the correlation based MVPA reported by [101], and stimulus-specific activation for visual and auditory processing in the xMVPA analysis (see Table 6.3), i.e. channel 1 is specifically *active* in response to the visual stimulus, but *inactive* in response



(a) A cortical network proposed by xMVPA for processing **visual** stimulus in six-month-old infants.

(b) A cortical network proposed by xMVPA for processing **auditory** stimulus in six-month-old infants.

Figure 6.9 The cortical networks proposed by xMVPA using oxy-Hb fNIRS signals, © 2021 Nature.

to the auditory stimulus. This specific pattern of activation is also consistent with the localisation of channel 1 in the occipital cortex, responsible for the processing of visual information [181]. In addition, the xMVPA patterns also outline the interconnection of channel 1 with other channels (channel 2 and channel 4 in P_1), uncovering a network of cortical regions for visual processing.

The xMVPA has identified two brain activity patterns (P_1 and P_2) in response to the dynamic visual stimulus presented to the six-months-old infants in the study. Specifically, we found activation of the occipital cortex and the PFC, with partial activation of the temporal cortex.

The activation of the occipital cortex for processing visual information in infancy is well-established in the literature. For example, [182] reported activity over the occipital cortex when 6.5-month-old infants were presented with an occlusion event involving objects. [183] showed that 3-month-old infants' occipital cortex was activated for both dynamic (moving mobile objects) and static visual stimuli (black-and-white checkerboard pattern). Similar to our findings, in response to the dynamic stimulus, they also reported activation over temporal and prefrontal cortices. Hence, the patterns P_1 and P_2 provided by the xMVPA are in line with the existent literature, suggesting that a specific cortical network of regions involving the occipital, temporal, and prefrontal cortices is involved in the processing of dynamic visual information.

It is important to note that the dynamic visual stimulus used by [101] displayed human facial attributes. Extending previous findings of studies that investigated face processing in young infants (e.g. [184, 185]), a specific inter-regional interaction between the occipital and temporal cortices (P_1) in response to the face stimulus is found. A similar network of occipital and temporal regions for visual processing is also found in the adult literature [186]. In particular, the occipito-temporal region is identified as a 'core system' in the model of the distributed human neural system for face perception in adults [187]. Thus the interaction between occipital and temporal cortices identified in the pattern P_1 in the present study provides evidence for the existence of an equivalent 'core system' for face processing in six-month-old infants (Fig. 6.9 (a)).

In addition, pattern P_2 identified inter-regional interaction between the prefrontal and temporal cortices. This indicates that infants as young as six months of age recruit an extended neural system for processing social stimuli, such as faces, adding to the existent literature that found similar activations in older infants [188, 189]. This is also in line with the 'extended system' in the model of face perception in adults [187], which is a dedicated network over the temporal and prefrontal cortices for processing basic facial emotions.

In line with [111], no pattern is found in P_1 or P_2 suggesting direct inter-regional interactions between the occipital (channel 1 and channel 2) and the PFC (channel 8, inferior frontal gyrus) in response to the visual stimulus. However, previous studies have demonstrated the involvement of the PFC during the presentation of visual stimuli in newborns [181] and 3-month-old infants [183]. While there is evidence supporting the functional role of the PFC in the early postnatal period [190], it is possible that the functional connections between visual and frontal cortex undergo experience-dependent synaptic pruning during this time [191] leading to potential functional specialisation in the occipital cortex by 6 months of age [192]. In support to this hypothesis, a study by [193] demonstrated a decrease in connectivity between prefrontal and occipital cortices from birth to six months. Taken together, the results reported by [193] and [111], as well as the absence of interaction between occipital and PFC in the present work, suggest that the role of the PFC is not significant in the core processing of visual information at 6 months of age. However, the direct connections with the temporal cortex suggest that the PFC may play a role in the extended system for deriving meaning from the visual stimulus. This is in line with the established role of the PFC as an overall control unit that receives input from perceptual cortices and generates meaning from the received input [3, 194].

Based on the above discussion on the patterns provided by the xMVPA, a model for the cortical pathways for the processing of visual stimulus in six-month-old infants is presented in Fig. 6.9 (a).

The model for the developing brain has similar modules and interconnections as the adult neural system for face perception presented by Haxby et al. in [187] suggesting that by 6 months of age the cortical activity associated with face processing is already similar to that of mature brains.

A total of four patterns, P_3 to P_6, were identified by the xMVPA for the processing of auditory stimulus. Specifically, while patterns P_4, P_5, and P_6 describedd the involvement of the temporal cortex, the activation of the PFC is observed in pattern P_3. This evidence is in line with the literature, whereby non-speech auditory stimuli elicit consistent responses in the infant temporal [195] and PFC [111].

While activation of the prefrontal and temporal cortices were found, none of the pattern revealed an interaction between these areas. Previous studies with infants reported non-synchronised activity in temporal and prefrontal cortices in response to non-speech auditory stimuli [196, 197], whereas activation in both temporal and prefrontal cortices has been reported in response to speech-like sounds [198, 199]. Considering that in the present work, the auditory cue presented to infants was a non-speech stimulus, the xMVPA results are in line with the literature and suggest that inter-regional interactions between the temporal cortex and PFC might be specific to speech-like sounds [198, 199]. While this interpretation would fit both with the xMVPA results and with the available evidence from previous infant research, further studies should use the xMVPA model to directly test this hypothesis.

None of the patterns identified activation of the occipital cortex in response to the auditory stimulus, indeed channel 1 was found *inactive* in patterns P_3 and P_6. While this is not surprising, as the occipital cortex is usually recruited in response to visual, rather than auditory stimuli [181], it is important to point out that this finding further strengthens the validity of the proposed xMVPA method.

The xMVPA method also shows a selective pattern of activation over the temporal cortex that is specific to visual vs. auditory stimuli. Specifically, the channels of the temporal cortex which are active in response to the visual stimulus are instead inactive in response to the auditory stimulus, i.e. channel 4 is *active* in P_1 and P_2 for visual processing and *inactive* in P_4 and P_5 for auditory processing. This confirms the multifaceted role of temporal cortex in the processing of sensory stimuli thereby some areas are dedicated to visual processing (eg. [182, 183, 187, 184]) whilst others are associated with auditory processing (eg. [56, 198, 199]).

Based on this body of evidence, with this work we hypothesise a non-synchronised model for the cortical pathways engaged in the processing of non-speech auditory stimuli in six-month-old infants. This proposed model is composed of a 'Core' and an 'Extended' system as shown in Fig. 6.9 (b). The temporal cortex will form the core system for processing non-speech auditory stimuli, while the PFC will form the extended system for processing the emotion associated with the auditory stimulus. When inactive, the occipital cortex enables the occurrence of these patterns.

Taken together, the patterns P_1 to P_6 obtained by the proposed xMVPA have not only provided corroborative evidence for the existent literature for the processing of perceptual information in infants, but also revealed new brain regions activation and interactions not yet established for the developing brain. Learning new cortical pathways directly from the neuroimaging data is of fundamental significance in DCN research to shed light on functional brain development in absence of established assumptions.

xMVPA Patterns found in deoxy-Hb fNIRS Signals

An illustration of the cortical network formed using the xMVPA patterns obtained from deoxy-Hb signals of six-months-old infants for the processing of auditory stimulus is shown in Fig. 6.10. There is no prominent cortical network for visual processing using deoxy-Hb signals since the two xMVPA patterns P_1 and P_2, for visual processing, have almost negligible dominance score.

As such, no direct comparison of visual cortical networks formed from the xMVPA patterns using the oxy-Hb and deoxy-Hb is possible since no prominent cortical network is uncovered for the visual processing using deoxy-Hb signals. A comparison for the cortical network for the auditory processing, for oxy-Hb and deoxy-Hb, reveals a notable absence of PFC for the deoxy-Hb cortical network.

Although I have presented a preliminary comparison of the cortical networks formed for the processing of oxy-Hb and deoxy-Hb, it is important to note that most developmental studies with fNIRS do not investigate the deoxy-Hb signals because of their low SNR, as well as inconsistent response in infants [6, 200]. Nevertheless, a comparison of the deoxy-Hb signals' (for the six-month-old infants data) decoding accuracy and that of oxy-Hb signals using correlation based MVPA and other state-of-the-art classifiers, and the proposed method of xMVPA is reported in Table 6.4. As can be readily appreciated from the decoding accuracy, values of xMVPA reported in Table 6.4 are indeed similar to or better than the ones obtained with opaque box methods, while rendering insightful

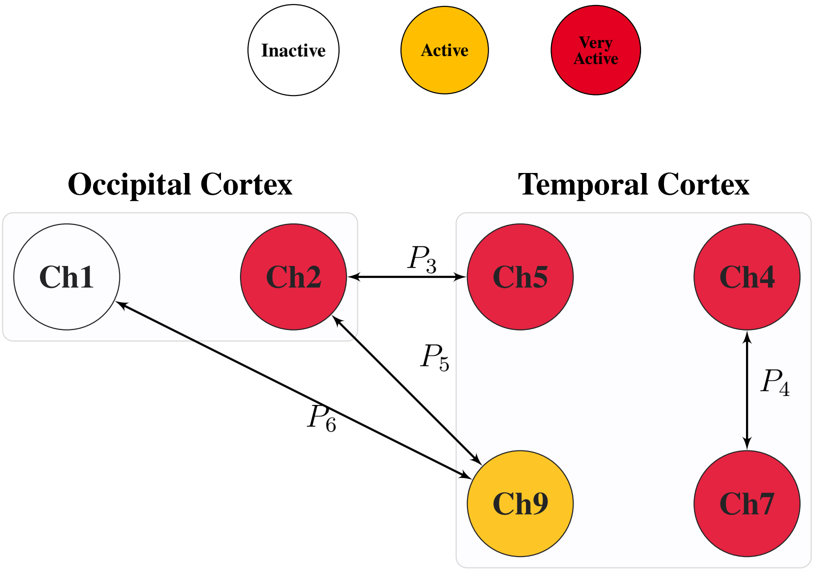


Figure 6.10 An illustration of the cortical network formed for non-speech auditory processing in six-month-old infant using deoxygenated haemoglobin (deoxy-Hb) signals, © 2021 Nature.

explainability on the decoding patterns. However, mostly low dominance patterns were present with deoxy-Hb. This corroborates what is reported in DCN literature [6, 200] that decoding of oxy-Hb signals is more consistent and informative than deoxy-Hb signals.

Given the unreliable deoxy-Hb signals, a lack of infant studies investigating deoxy-Hb signals in the literature, and mostly under-supported xMVPA patterns on account of low dominance score, the implications of the xMVPA patterns obtained using deoxy-Hb signals are not discussed.

Haemoglobin (Hb)	**MVPA**		**SVM**		**RF**		**MLP**		**xMVPA**	
	Avg.	*SD*	*Avg.*	*SD*	*Avg.*	*SD*	*Avg.*	*SD*	*Avg.*	*SD*
oxy-Hb	66.67	17.45 %	69.87	1.17	67.47	5.32	68.36	3.22	67.69	3.52%
deoxy-Hb	33.98	16.57 %	61.65	3.29	60.76	2.47	57.72	3.21	64.88	4.81%

Table 6.4 A comparison of average decoding accuracy (avg.) with standard deviation (SD) using oxy-Hb and deoxy-Hb data from the earlier work of MVPA by [101], © 2021 Nature.

6.2.3 Implications for functional brain development analysis

The proposed xMVPA overcomes the need of knowing *a prioi* the haemodynamic response function (HRF) for infants; which is rarely known [84] (as previously outlined in Section 2.2.4). In addition, xMVPA offers an explainable inference mechanism in terms of patterns which represents a stepping stone for furthering our understanding of the functional development of the human brain as hypothesised by the interactive specialisation (IS) framework (discussed in Chapter 2.1). Here, xMVPA is applied on fNIRS data obtained in response to visual and auditory stimuli in a group of six-month-old infants [101]. The xMVPA identified six patterns, for oxy-Hb signals, describing cortical activations and inter-regional interactions specific to each of the perceptual stimuli. These patterns corroborated the existing evidence in the DCN literature, while providing further insight about auditory processing

Problem	**Input/Output**	**Feature/Class**	**Conceptual Labels (CoLs)**	**N**	**Time Intervals, Δt_α**
Classification	Input	Temperature	Low, Medium, High	24	Morning, Daytime, Evening
		Light	Low, Medium, High	24	Morning, Daytime, Evening
		CO_2	Low, Medium, High	24	Morning, Daytime, Evening
	Output	Occupied Not Occupied	-	-	Morning, Daytime, Evening

Table 6.5 The classification problem is exemplified using the proposed Time-dependent eXplainable Artificial Intelligence (TXAI) system with occupancy dataset [161].

in infants.

6.3 Application of TXAI in Temporal dataset

In this section, a temporal occupancy dataset [161] is used to exemplify the proposed TXAI system modelling. The occupancy dataset entails measurements of a room along with the time of when the measurement is recorded. In particular, it includes measurements of the room temperature, light, CO_2, and a binary class of whether or not the room is occupied. There are 8,143 data instances in the dataset taken over a period of a few weeks.

In this work, the dataset [161] is used for classification problem where TXAI system predicts whether or not the room is occupied based on the room measurements. The inputs of temperature, light, and CO_2 are used to predict whether or not the room is occupied. Three conceptual labels of Low, Medium, and High are associated with inputs of temperature, light, and CO_2. The primary MF of the conceptual labels for all inputs are empirically found. The time is discretised at each hour of the day hence a total of N = 24 time points with a total of three time intervals defined at Morning, Daytime, and Evening, as also summarised in Table 6.5.

The conditional distribution for each conceptual label of every input is computed on the entire dataset. Once the conditional distributions are computed, the learning procedure focuses on the data belonging to each interval. A 10-repeated nested cross-validation procedure is adopted. The dataset is split into a disjoint stratified train, validation and the test set to ensure a random selection of the datasets (train, validation, and test) is not creating any bias in the results. Each repetition, 20% of the dataset is held out as a test set, and the remaining is used to build the train and validation sets. Train and validation sets are determined in an inner 10-fold procedure, where a fold is used for validation and the rest for training to determine the pattern weights. Balanced accuracy and other performance metrics are computed over each validation and test set. The z-slices are obtained on locations [0.2, 0.4, 0.6, 0.8, 1.0].

A pattern-base is formed for each time interval. The patterns are learned using GA (previously outlined in Chapter 5.1.3 and Chapter 5.2.4) [125] such that they (patterns) attain optimally balanced accuracy on the validation datasets. The GA parameters specification includes the number of generations, set at 20, with each generation having a population of 50. Moreover, the GA is leveraged to find the patterns that are prototypical for each time interval. The number of antecedents in each pattern can be at most 3 but not more to underpin explainability and hamper model complexity therefore precluding over-fitting. For the same reason, the maximum number of patterns in each candidate pattern-base for each time interval was limited to 30, although further pruned when its weight (eq. (5.28)) does not surpass a tolerance threshold of 0.001.

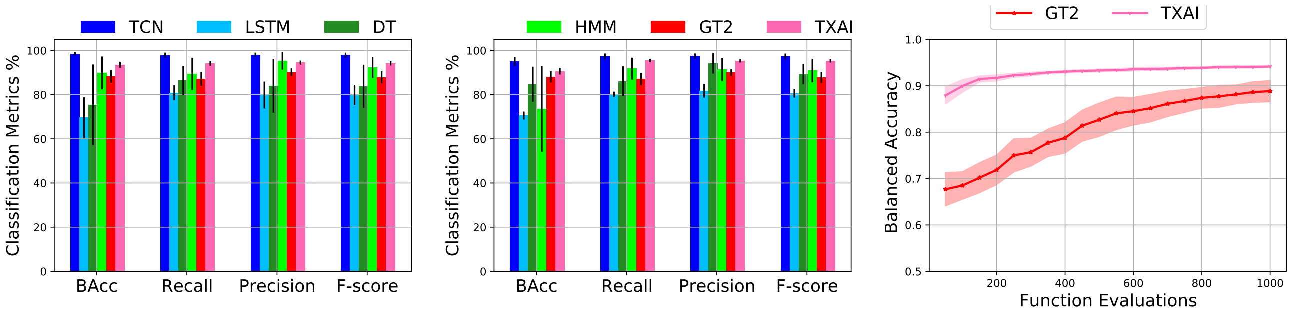
In order to compare the performance of the proposed TXAI system, numerous state-of-the-art classifiers which can both analyse time-series data and/or are explainable have been used. More specifically, for comparison with temporal analysis Long Short-Term Memory (LSTM) [201] and Hidden Markov Models (HMM) [202] are used, for comparison with explainable models the standard GT2 based XAI system is used, and for partial explainablility Decision Trees (DT) [203] is used. In addition, a comparison is also made with a temporal convolutional network (TCN) [204] for comparison with deep learning methods [205]. Parametrization and configuration was set to default mode of their respective libraries (Sklearn and Keras). For methods with no modelling with respect to a time component, time is given as an extra input feature. Moreover, the train, validation, and test dataset splits are similar across all methods and for GT2 based XAI in particular, the location of z-slices, and the GA parameters for pattern learning are also identical to those of TXAI system.

6.3.1 Results

For the classification problem undertaken, using the occupancy dataset, the proposed TXAI system and numerous state-of-the-art classification methods predict whether or not the room is occupied. The mean (and standard deviation) f-score obtained using TXAI system on the 10 test datasets is 95.30% which is the highest score on the test dataset across all classifiers except TCN. The other classification metrics investigated in this work are balanced accuracy, recall, and precision. A bar plot for the aforementioned classification metrics for both the proposed TXAI and the state-of-the-art AI methods (TCN, LSTM, DT, HMM, GT2 based XAI) on 10 times repeated 10-fold validation and test datasets is shown in Fig. 6.11 (a) and (b) respectively. In addition, a convergence graph that outlines how the GA optimisation converges with respect to balanced accuracy for both TXAI and GT2 based XAI systems is also shown in Fig. 6.11 (c).

The original study [161] applied numerous state-of-the-art AI methods for the classification of the occupancy dataset. They reported best performing methods' (RF, linear discriminant analysis, classification and regression trees) accuracies ranging from 95% to 99%. For the RF model, the results suggest that the most important features were Light, CO_2, and Temperature (in descending order). For the classification and regression trees, the models also highlight the importance of Light and Temperature for correctly predicting the occupancy status of the room. While the feature importance, and tree based models shed partial explainability (in terms of input feature importance), these models are unable to shed light on the interdependence of the input features for the classification of the room occupancy status. In contrast, the proposed TXAI method can shed light on the interdependence of the input features in the form of patterns.

The patterns outlined by TXAI and GT2 based XAI systems which are prototypical for whether or not the room is occupied are listed in Table 6.6. For the TXAI system, the patterns are found separately for each time interval (Morning, Daytime, and Evening) whereas, for GT2 based XAI system, the time intervals are one of the antecedents of the patterns. In general, for both TXAI and GT2 based XAI systems, the patterns outline that when the room measurements have higher values, the room is more likely to be occupied, and when the room measurements are on the lower end, the room is more likely to be not occupied.



a) Validation datasets b) Test datasets c) Convergence graph

Figure 6.11 A comparison of the classification prowess of the proposed time-dependent eXplainable artificial intelligence (TXAI) system with numerous state-of-the-art classification systems.

For the TXAI system, the temporal trajectories of a time-variant system can also be investigated using the pattern transition matrices (PTMs), previously outlined in section 5.3.8. The individual PTMs transitioning from one time interval (Δt) to another i.e., from Morning to Daytime, from Daytime to Evening, and from Evening to Morning, represent the joint possibilities of observing a given pattern in Δt^+ with respect to the patterns in Δt. The patterns corresponding to the highest PTPs (pattern transitioning possibilities) are also joined with lines in the column *PT* (pattern transitions) in Table 6.6 and illustrated in a schematic in Fig. 6.12.

Method	Time	P No.	Pattern	DS	PT
Time-dependent eXplainable Artificial Intelligence (TXAI)	Morning	1	IF Light is High THEN room is Occupied	0.346	
		2	IF Temperature is High THEN room is Occupied	0.079	
		3	IF CO$_2$ is Medium THEN room is Occupied	0.050	
		4	IF CO$_2$ is High THEN room is Occupied	0.046	
		5	IF Temperature is High AND Light is High THEN room is Occupied	0.018	
		6	IF Light is High AND CO$_2$ is Medium THEN room is Occupied	0.014	
		7	IF Light is High AND CO$_2$ is High THEN room is Occupied	0.012	
		8	IF Temperature is Medium THEN room is Occupied	0.012	
		9	IF Temperature is High AND CO$_2$ is High THEN room is Occupied	0.011	
		10	IF Temperature is Medium AND CO$_2$ is Medium THEN room is Occupied	0.007	
		11	IF Light is Low THEN room is Not Occupied	1.000	
		12	IF Temperature is Low THEN room is Not Occupied	0.073	0.218 0.226
	Daytime	1	IF Light is High THEN room is Occupied	0.473	
		2	IF Temperature is High THEN room is Occupied	0.277	0.335
		3	IF CO$_2$ is High THEN room is Occupied	0.110	
		4	IF Temperature is Medium AND Light is High THEN room is Occupied	0.017	
		5	IF Temperature is High AND Light is High AND CO$_2$ is High THEN room is Occupied	0.015	
		6	IF Light is Low THEN room is Not Occupied	1.000	
		7	IF CO$_2$ is Low THEN room is Not Occupied	0.50	
		8	IF Light is Medium THEN room is Not Occupied	0.147	
		9	IF Temperature is High AND Light is Low THEN room is Not Occupied	0.011	
	Evening	1	IF Light is High THEN room is Occupied	0.005	
		2	IF Light is Low THEN room is Not Occupied	1.000	
		3	IF Light is Low AND CO$_2$ is Low THEN room is Not Occupied	0.108	
		4	IF Temperature is High AND Light is Low THEN room is Not Occupied	0.041	
eXplainable Artificial Intelligence (XAI)		1	IF Light is High AND Time is Daytime THEN room is Occupied	0.580	
		2	IF Light is High AND Time is Morning THEN room is Occupied	0.425	
		3	IF Temperature is High AND Time is Daytime THEN room is Occupied	0.419	
		4	IF Light is Low AND Time is Morning THEN room is Not Occupied	1.000	
		5	IF CO$_2$ is Medium AND Time is Evening THEN room is Not Occupied	0.789	

Table 6.6 The prototypical patterns (P) were obtained by the proposed time-dependent explainable artificial intelligence (TXAI) system for the binary classification problem (room occupied or not) using the occupancy dataset.

6.3.2 Discussion

In this work, the proposed TXAI system is used to model an occupancy dataset [161] for the classification problem of whether or not the room is occupied. For comparison purposes, several state-of-the-art explainable (GT2 based XAI system), partially explainable (DT), and non-explainable methods that can analyse temporal information (LSTM and HMM) as well as TCN are also applied to the aforementioned classification problem. As can be noted from the Fig. 6.11 (a) and (b), TXAI offers greater classification performance than all classifiers (for e.g. for mean fscore TXAI performs better than LSTM by 18.19%, DT by 6.81%, HMM by 4.90% , GT2 based XAI system by 8.58% on test datasets) except TCN (for mean fscore TCN performs better than TXAI by 4.67% on test datasets). However, the TCN classification mechanism is not explainable hence unable to shed light on the prediction of the room occupancy based on input features of Temperature, Light, CO_2, and Time.

With respect to the comparison with the GT2 based XAI system, the only explainable system apart from the proposed TXAI system, a convergence graph plotted in Fig. 6.11 (c) also highlights that TXAI system converges (~500 function evaluations) twice as faster than standard GT2 based XAI system (~1000 function evaluations) whilst also yielding higher classification metrics (Fig. 6.11 (a) and (b)). Moreover, the patterns outlined by the explainable systems, TXAI and XAI systems, are listed in Table 6.6, and both systems are in agreement that when the room measurements (Temperature, Light, and CO_2) have higher values, then the room is likely to be occupied, and when the room measurements are lower, then the room is likely to be not occupied. However, the patterns for TXAI also offer greater insight into how the room measurements are interlinked with respect to predicting room occupancy. For example, for the time interval Morning, pattern no 5 (see Table 6.6) outlines that if both inputs of Temperature and Light have high values then the room is likely to be occupied. In this regard, patterns across time intervals shed light on the intertwined conceptual labels of the

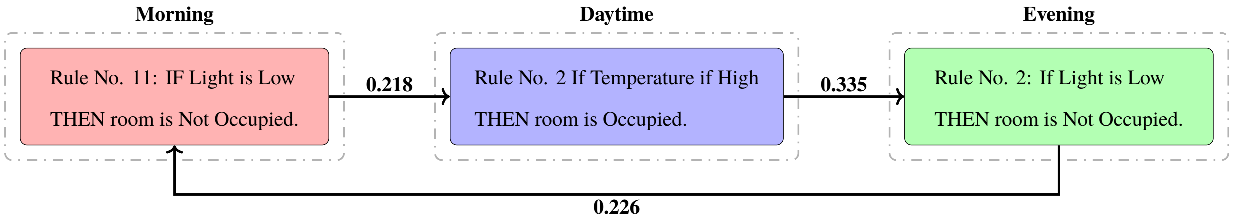

Morning
Daytime
Evening
Rule No. 11: IF Light is Low THEN room is Not Occupied.
Rule No. 2 If Temperature if High THEN room is Occupied.
Rule No. 2: If Light is Low THEN room is Not Occupied.
0.218
0.335
0.226

Figure 6.12 A schematic presenting the evolution of the occupancy system based on the rules with the highest Pattern Transition Possibilities (PTPs).

inputs prototypical for decoding the room occupancy.

Furthermore, the TXAI systems are also able to shed light on the temporal trajectories of the system being modelled using PTMs, previously outlined in Section 5.3.8, and illustrated in Fig. 6.12. The PTPs (Pattern Transition Possibilities), which are the elements of the PTMs, represent the joint likelihood of observing a pattern in one time interval (rows) and then observing another pattern in the next time interval (columns). For example, in the PTM transitioning from Morning to Daytime, the patterns with the highest PTP are pattern number 12 (for time interval Morning) and pattern number 5 (for the next time interval Daytime). For the particular case of the occupancy datasets, the PTMs and the corresponding PTPs outline the trajectory across time as the TXAI system transitions from one time interval to another. In this case, an analysis of the occupancy dataset can be leveraged for the efficient energy management of smart homes using the predictive power of the PTMs [206].

Indeed, the motivation for developing the TXAI systems is to be able to analyse time-dependent real processes across time. In this regard, conditional distribution integrated within the TXAI system can be used to obtain the PTMs. The PTMs entail the likelihood of observing the transition of a real-life process from one time point to another. The proposed TXAI system can shed light not only on which patterns are prototypical for each of the time intervals but also on the likelihood of observing the patterns across the different time points.

6.3.3 Implications for functional brain development analysis

Infants' longitudinal data analysis has the potential to describe brain development trajectories since it holds information of how the infants' brain is working across different times/ages. It is important to investigate the brain development trajectories as they can inform on typical and atypical brain development which in turn can be leveraged to inform clinical, educational and social policies. The brain developmental trajectories can inform about functional brain development in line with the DCN frameworks of IS and neural reuse (outlined in Chapter 2.1). To this end, the proposed TXAI system has been designed to integrate temporal information as well as able to outline the trajectories of a time-dependent process (i.e., brain development trajectories for the case of functional brain development analysis).

In the next chapter, I present a discussion on the three proposed XAI models' results and findings as well as the conclusion.

Chapter Seven

Conclusions and Future Work

In this chapter, I present the conclusion and future works for the proposed XAI methods.

7.1 Conclusion

Developmental Cognitive Neuroscience (DCN) is a multidisciplinary field that aims to inform typical and atypical human brain development. The brain development is a complex process as it entails a bi-directional relation between the structural maturation and emerging functional capabilities of the different regions of the brain. In this regard, the most prevalent and suitable DCN frameworks that account for the phenomena underlying brain development are the interactive specialisation (IS) [3] and neural reuse accounts [50], as discussed in Chapter 2.1. The IS account suggests that postnatal brain development emerges as a result of the optimisation of interactions between different regions of the brain. In more detail, it suggests that cortical regions interact and compete with each other to acquire their role in new computational abilities, therefore becoming more specialised with development. Critically, the onset of new behavioural abilities is associated with changes in activity over cortical networks, and not by the onset of activity in single regions. Whereas the neural reuse framework accounts for the hierarchical structure of the brain brought about by the rewiring of the cortical networks on acquisition of a skill.

In order to investigate functional brain development, a recording of infants' brain activity as they perform different motor, and cognitive tasks is quintessential. A neuroimaging modality that is safe, 'infant-friendly' and offers good spatial and temporal resolution is functional Near-InfraRed Spectroscopy (fNIRS) [6], described in Chapter 2.2. The fNIRS cap has sources and detectors fitted on it to record the underlying cortical region's activity. More specifically, fNIRS uses the relative absorption of Near InfraRed (NIR) light, by the haemoglobin (Hb) chromophores, as a measure of a cortical brain activity [53].

The fNIRS data analysis, in DCN studies in particular, has been predominantly univariate, i.e., it aims to identify the region with the most pronounced activity in comparison to other cortical regions in response to the presented information [146]. The univariate analysis has proved pivotal in the recognition of the prototypical functions of different regions of the brain. However, it (univariate analysis) can not shed light on the interactions of the different regions of the brain for the processing of presented information. In this regard, multivariate pattern analysis (MVPA) which investigates multiple regions' data simultaneously has the potential to uncover the cortical networks formed in response to the presented information. However, for MVPA to inform functional brain development, the artificial intelligence (AI) methods driving the MVPA need to elucidate the patterns/representations (classification mechanism) they learn from the data in explainable, human-understandable language [22].

Owing to the current gap in DCN research due to non-explainable AI methods there is limited insight obtained from the learnt classification mechanism on the basis of brain activity patterns, as discussed in Chapter 4. A lack of explainable classification models 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), in this work, explainable classification mechanism are developed. The proposed XAI methods are: 1) Effective Fuzzy Cognitve Maps (EFCMs), 2) Explainable Multivariate Pattern Analysis (xMVPA), and 3) Time-dependent XAI (TXAI) Systems.

EFCMs (presented in Chapter 5.1) offer enhanced formulation, than the classical fuzzy cognitive maps (FCMs), to learn the strength of interaction (effective connectivity (EC)) between cortical regions (fNIRS channels). The EC values give a quantitative measure of the interaction between the corresponding cortical regions of the brain. In this regard, EFCMs can shed light on the rewiring of the cortical networks as hypothesised by the neural reuse framework upon functional brain development. In this work, I exemplified the EFCM method on an adults' fNIRS dataset (presented in Chapter 6.1) in which the participating surgeons had varying level of expertise (NVs, TNs, and EXs) for performing a complex motor task (Laparoscopic Surgery (LS)). The significance of this dataset, for the analysis of functional brain development, lies in the varying expertise level of the participating surgeons akin to how infants expertise for performing a certain task differs at different ages. Hence an EFCM analysis of the surgeons' dataset, with varying level of expertise, can also potentially shed light on the evolution of cortical networks as hypothesised by the neural reuse framework.

The EFCM learns the EC values directly from the fNIRS data, presented in Fig. 6.3. More specifically, in Fig. 6.3, the cortical networks based on the learnt EC values are presented. Moreover, in Fig. 6.4, the cortical networks based on two ROIs (motor cortex and PFC) are presented. In both Fig. 6.3 and 6.4 the evolution of the cortical networks can be deciphered from the reconfiguration of the interconnections. Although EFCMs can shed light on the evolution of the cortical networks, they are limited in their explainability. In more detail, the EFCM cannot inform about the activations of the different cortical regions such as PFC is more active than motor cortex. This is a critical limitation of the EFCM since it cannot describe the functional brain development as hypothesised by the IS framework.

In addition to limited explainability, the EFCM prowess to decipher the EC values correctly also depend on the order of the EFCM. In this regard, it is a good practise to explore the effect of modifying the order of the EFCM model as the expertise level increases. This can be of significance because as the brain networks evolve the memory order of the underlying EC in the brain network may also increase or decrease. This needs to be further investigated but an intuitive hypothesis can be that optimisation of brain networks would have a bearing on the memory order of EC within the ROIs.

Whereas EFCMs are based on graph theory, the other two proposed XAI methods (xMVPA and TXAI) developed in this work are based on fuzzy logic systems (FLS). The two main advantages of using FLS based XAI systems are: 1) FLS are explainable, i.e. their classification mechanism consists of patterns which can be easily understood, and 2) FLS can handle uncertainty in the input data using fuzzy sets that convert uncertain observations (such as fNIRS signal measurements) to conceptual labels characterised with membership values [122, 123]. The fuzzy sets are characterised by MFs and represent a given conceptual label. The membership value are usually in the range [0,1] and is a soft measure of degree of association the associated fuzzy set has for a given observation to belong to the conceptual label represented by the fuzzy set [123]. For example, an XAI system modelling the brain activity of infants using Type-1 fuzzy sets may represent brain activity using conceptual labels of *inactive, active,* and *very active*. The MF associated with each conceptual label's MF will assign a crisp number for the brain activity of an infant with a membership grade; for example, a brain activity of 2mM may get assigned membership grades of $0.8, 0.5, 0.1$ to represent conceptual labels of *inactive, active,* and *very active* respectively.

The proposed FLS based XAI method is called explainable MVPA (xMVPA) [105] and presented in Chapter 5.2. The xMVPA is based on IT2-FLS (Interval Type-2 Fuzzy Logic System) since it has the capability to model uncertainty to a greater level (than T1 fuzzy sets) whilst not as computationally expensive as the GT2 fuzzy sets. In this regard, xMVPA based on IT2 fuzzy sets can model the uncertainty in the infants' fNIRS dataset as well as describe an explainable inference mechanism that can shed light on functional brain development.

xMVPA is here applied on fNIRS data obtained in response to visual and auditory stimuli in a group of six-month-old infants [101]. The xMVPA identified six patterns (listed in Chapter 6.2) describing cortical activations and inter-regional interactions specific to each of the perceptual stimuli. In particular, the cortical networks for the processing of visual stimulus (illustrated in Fig. 6.9 (a)) outlined the formation of a 'core system' composed of occipital and temporal cortices and an 'extended system' composed of PFC. In this regard, xMVPA suggests the formation of a specialised cortical network as formed by adults for the processing of visual information [187].

In contrast to the specialised cortical network suggested by the xMVPA for the processing of visual stimulus, xMVPA suggested a non-specialised cortical network for the processing of auditory stimulus. The non-specialised cortical network is suggested by the missing interconnection, in Fig. 6.9 (b), between the temporal cortex and the PFC. Overall, the xMVPA patterns corroborated the existing evidence in the DCN literature for visual processing, while providing further insight about auditory processing in infants.

With regards to the xMVPA implementation, an important consideration is the selection of the optimal time window for the fNIRS signals. The statistical feature (such as mean or amplitude) of the selected fNIRS signal (based on the time window) is then used to construct a multivariate matrix (MVM). The xMVPA then works on the MVM and hence the time window selection can considerably impact (increase or decrease) the classification accuracy scores. In this work, 4-7s of the fNIRS signal was found using grid search to give the highest accuracy (same time window is also reported by the original study [101]). Therefore, it is recommended to find the optimal time window (using grid search or evolutionary algorithms) so that the patterns learnt by xMVPA can yield a high accuracy score.

In addition to the selection of the optimal time window for fNIRS signals, the granularity in the activation levels for the patterns is also an important factor that can impact the classification accuracy of the xMVPA method. In this work, only three conceptual labels are chosen to represent the level of activity of the channels, i.e. Inactive, Active and Very Active. The rationale behind keeping 3 conceptual labels is to keep the patterns easily interpretable. However, the number of conceptual labels should be carefully chosen for each study depending on the question at hand, and the level of interpretability required.

The shape of the membership functions (MFs) is another important factor that can impact the performance of the xMVPA. More specifically, the shape can be either triangular, trapezium and/or Gaussian. In addition to the selection of the shape, the parameters that define the shape (such as the mean and standard deviation for the Gaussian MF) is also critical to the xMVPA's performance. In this work, I used trapezium shaped MFs, and the parameters for the MFs are found through the Genetic Algorithm (GA). The MFs shape and parameters are important because the conversion of a numeric measurement to a conceptual label depends on the MF.

Nevertheless, given its capability and reliability of identifying patterns of inter-regional interactions for information processing, the xMVPA provides a technical framework for implementing the IS account proposed by Johnson [3] for explaining functional brain development. However, a limitation of the xMVPA is that it can only analyse cross-sectional data and can not shed light on the temporal dynamics associated with a longitudinal dataset. This is because xMVPA which is based on IT2 fuzzy sets cannot integrate temporal information in their MF (Membership Function). To account for the temporal components associated with functional brain development, in this work, I present the time-dependent XAI (TXAI) systems in Chapter 5.3. The TXAI systems, based on new Temporal Type-2 Fuzzy Sets (TT2FSs), can account for the likelihood of a measurement's occurrence in the time domain using (the measurement's) frequency of occurrence. In TT2FSs, a four-dimensional

(4D) time-dependent MF is developed where fuzzy relations are used to construct the inter-relations between the elements of the universe of discourse and its frequency of occurrence. In addition, the temporal trajectories of a dynamic process such as functional brain development can be outlined by the TXAI system by making use of the conditional distribution integrated into the TXAI system.

In this work, the TXAI system is exemplified on a real-life occupancy dataset [161] to determine whether or not the room is occupied based on sensor readings of temperature, light, and carbon dioxide. The TXAI system performance is also compared with other state-of-the-art classification methods with varying levels of explainability. The TXAI system manifested better classification prowess, with 10-fold test datasets, with a mean recall of 95.40% than a standard XAI system (based on non-temporal general type-2 (GT2) fuzzy sets) that had a mean recall of 87.04%. TXAI also performed significantly better than most non-explainable AI systems between 3.95%, to 19.04% improvement gain in mean recall. Temporal convolution network (TCN) was marginally better than TXAI (by 1.98% mean recall improvement) although with a major computational complexity. In addition, TXAI can also outline the most likely time-dependent trajectories using the frequency of occurrence values embedded in the TXAI system; viz. given a pattern at a determined time interval, what will be the next most likely pattern at a subsequent time interval.

Although TXAI provides a greater insight than xMVPA with respect to the integration of the temporal information, care must be taken to select the parameters that can impact the performance of the TXAI method such as the optimal time window selection for fNIRS signal, the number of conceptual labels, and the shape of the MFs (as previously discussed for the xMVPA). In addition, for the TXAI method, the number of time intervals as well as their definition can have significant impact on the performance of the TXAI method. In this work, I chose three time intervals (Morning, Afternoon, and Evening) for the classification of the occupancy dataset using TXAI. The reason for selecting three time intervals is because of the common perception of three time intervals (Morning, Afternoon, and Evening) in a given day. In this regard, the TXAI system outlined the most likely patterns for each time interval as well as the likelihood of pattern transition from one time interval to another. Therefore the selection of the time intervals will have a bearing the temporal trajectories of the system.

In the next section, I outline the future works for the proposed explainable methods (EFCM, xMVPA, TXAI) developed in this work.

7.2 Future Works

In this chapter, I outline the future research in functional brain development that can focus on the application of the proposed XAI methods (EFCMS, xMVPA, and TXAI) on infants' fNIRS data. All three methods are developed to analyse a single dimension of Hb (i.e., either oxy-Hb or deoxy-Hb). In this regard, a future work can extend the methods to analyse a possible combination of the two Hb chromophores. In addition, all methods are built for fNIRS modality only, it is worth investigating if multi-modal data such as fNIRS and EEG can also be investigated after a possible extension of the methods.

More specifically, for EFCMs which undertake partially explainable effective connectivity (EC) analysis, future work can focus on its application in DCN studies to shed light on the reorganisation of the cortical networks with the onset of cognitive and motor abilities. More specifically, EFCMs can be applied to infants' fNIRS data before they attain a certain cognitive or motor ability and after the onset of the ability, such as grasping an object [164], to learn the optimisation of the cortical networks as hypothesised by the neural reuse framework.

For xMVPA, future applications can undertake the analysis of infants' fNIRS data recorded in response to various cognitive and perceptual stimuli. The xMVPA results would entail the cortical network formed in response to the presented stimuli and therefore have the potential to further our understanding of the workings of the developing brain in response to the presented stimuli. The xMVPA performance against more Fuzzy logic based methods such as fuzzy-SVM [207] can also be investigated. However, only cross-sectional fNIRS based DCN studies can be investigated using the xMVPA.

Future applications of the TXAI system could focus on infants' longitudinal fNIRS datasets as TXAI has been developed to provide complementary time-stamps patterns and map brain regions activation and interactions. In this regard, TXAI systems offer a promising avenue for the study of developmental brain trajectories in terms of maturation and inter-regional functional interactions [3]. It is important to investigate the brain developmental trajectories as they can inform on typical and atypical brain development, which in turn can be leveraged to inform clinical, educational and social policies. The delineation of brain developmental trajectories will further enhance the potential of the TXAI systems to critically contribute to the field of DCN.

REFERENCES

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

[2] S. Ackerman. (1992). Discovering the Brain, [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, 2004.

[5] T. Wilcox, H. Bortfeld, R. Woods, E. Wruck, and D. A. Boas, “Using near-infrared spectroscopy to assess neural activation during object processing in infants,” *Journal of Biomedical Optics*, vol. 10, p. 011 010, 2005.

[6] S. Lloyd-Fox, A. Blasi, and C. E. Elwell, “Illuminating the Developing Brain: The Past, Present and Future of Functional Near Infrared Spectroscopy,” *Neurosci Biobehav Rev.*, vol. 34, pp. 269–84, 2010.

[7] S. Tak and J. C. Ye, “Statistical analysis of fnirs data: A comprehensive review,” *NeuroImage*, vol. 85, pp. 72–91, 2014.

[8] J. Gemignani and J. J. Gervain, “Comparing different pre-processing routines for infant fNIRS data.,” *Developmental cognitive neuroscience*, vol. 48, p.100943, 2021.

[9] P. Pinti, A. Merla, C. Aichelburg, F. Lind, S. Power, E. Swingler, A. Hamilton, S. Gilbert, P. W. Burgess, and I. Tachtsidis, “A novel GLM-based method for the Automatic IDentification of functional Events (AIDE) in fNIRS data recorded in naturalistic environments,” *Neuroimage*, vol. 155, 2017.

[10] M. D. Rosenberg, B. Casey, and A. J. Holmes, “Prediction complements explanation in understanding the developing brain,” *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[11] A. K. Seth, A. B. Barrett, and L. Barnett, “Granger causality analysis in neuroscience and neuroimaging.,” *Journal of Neuroscience*, vol. 35, pp. 3293–3297, 8 2015.

[12] B. Blanco, M. Molnar, M. Carreiras, L. Collins-Jones, E. Rosas, R. Cooper, and C. Caballero-Gaudes, “Group-level cortical functional connectivity patterns using fNIRS: assessing the effect of bilingualism in young infants,” *Neurophotonics*, vol. 8, p. 025 011, 2 2021.

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

[14] S. Tak, A. Kempny, K. J. Friston, A. P. Leff, and W. D. Penny, “Dynamic causal modelling for functional near-infrared spectroscopy.,” *NeuroImage*, vol. 111, pp. 338–349, 2015.

[15] C. Bulgarelli, A. Blasi, S. Arridge, S. Powell, C. C. de Klerk, V. Southgate, S. Brigadoi, W. Penny, S. Tak, and A. Hamilton, “Dynamic causal modelling on infant fnirs data: A validation study on a simultaneously recorded fnirs-fmri dataset.,” *Neuroimage*, vol. 175, pp. 413–424, 2018.

[16] 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.

[17] J. Gemignani, E. Middell, R. L. Barbour, H. L. Graber, and B. Blankertz, “Improving the Analysis of Near-Infrared Spectroscopy Data With Multivariate Classification of Hemodynamic Patterns: A Theoretical Formulation and Validation,” *J Neural Eng.*, vol. 15, 2018.

[18] A. R. Harrivel, D. H. Weissman, D. C. Noll, and S. J. Peltier, “Monitoring attentional state with fnirs,” *Front. Hum. Neurosci*, vol. 7, p. 861, 2013.

[19] J. Andreu-Perez, D. R. Leff, K. Shetty, A. Darzi, and G.-Z. Yang, “Disparity in frontal lobe connectivity on a complex bimanual motor task aids in classification of operator skill level,” *Brain connectivity*, vol. 5, pp. 375–388, 2016.

[20] D. DSouza and A. Karmiloff-Smith, “Why a developmental perspective is critical for understanding human cognition,” *Behavioral and Brain Sciences*, vol. 39, 2016.

[21] 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, Early Access.

[22] M. Kiani, J. Andreu-Perez, H. Hagras, S. Rigato, and M. L. Filippetti, “Towards Understanding Human Functional Brain Development With Explainable Artificial Intelligence: Challenges and Perspectives,” *IEEE Computational Intelligence Magazine*, 2021, In Press.

[23] A. Karmiloff-Smith, “Development itself is the key to understanding developmental disorders,” *Trends in cognitive sciences*, vol. 2, pp. 389–398, 1998.

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

[25] J. Stiles, “Brain development and the nature versus nurture debate,” *Progress in brain research*, vol. 189, pp. 3–22, 2011.

[26] E. S. Spelke and K. Kinzler, “Core knowledge,” *Developmental science*, vol. 10, pp. 89–96, 2007.

[27] S. Carey and E. M. Markman, “Cognitive development,” *Cognitive science*, vol. 5, pp. 201–254, 1999.

[28] E. S. Spelke, “Core knowledge,” *American psychologist*, vol. 55, p. 1233, 2000.

[29] J. Stiles, *The fundamentals of brain development: Integrating nature and nurture*. Harvard University Press, 2008.

[30] S. R. Quartz and T. J. Sejnowski, “The neural basis of cognitive development: A constructivist manifesto,” *Behavioral and brain sciences*, vol. 20, pp. 537–556, 1997.

[31] M. H. Johnson and M. de Haan, “Prefrontal cortex, working memory, and decision-making,” in *Developmental Cognitive Neuroscience*, Wiley, 2010, ch. 1, pp. 5–6.

[32] J. L. Elman, E. Bates, and M. Johnson, *Rethinking innateness: A connectionist perspective on development*. MIT Press, 1996.

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

[34] A. Diamond and P. S. Goldman-Rakic, “Comparison of human infants and rhesus monkeys on Piaget’s AB task: Evidence for dependence on dorsolateral prefrontal cortex.,” *Experimental brain research*, vol. 74, pp. 24–40, 1989.

[35] J. Piaget and M. T. Cook, *The development of object concept*, 1954.

[36] A. Diamond, *Neuropsychological insights into the meaning of object conceptdevelopment*, 1991.

[37] M. H. Johnson and M. de Haan, “Prefrontal cortex, working memory, and decision-making,” in *Developmental Cognitive Neuroscience*, Wiley, 2010, ch. 10, pp. 194–196.

[38] A. A. Baird, J. Kagan, T. Gaudette, K. A. Walz, N. Hershlag, and D. A. Boas, “Frontal lobe activation during object permanence: Data from near-infrared spectroscopy,” *NeuroImage*, vol. 16, pp. 1120–1126, 2002.

[39] A. Diamond, “Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry,” *Principles of frontal lobe function*, vol. 466, p. 503, 2002.

[40] P. J. Olesen, H. Westerberg, and T. T. Klingberg, “Increased prefrontal and parietal activity after training of working memory,” *Nature neuroscience*, vol. 7, pp. 75–79, 2004.

[41] B. Kolb and I. Q. Whishaw, “Brain plasticity and behavior,” *Annual review of psychology*, vol. 49, pp. 43–64, 1998.

[42] H. Makino, E. J. Hwang, N. G. Hedrick, and T. Komiyama, “Circuit mechanisms of sensorimotor learning,” *Neuron*, vol. 92, pp. 705–721, 2016.

[43] I. Gauthier and M. J. Tarr, “Becoming a “Greeble” expert: Exploring mechanisms for face recognition,” *Vision research*, vol. 37, pp. 1673–1682, 1997.

[44] M. H. Johnson and M. de Haan, “Interactive specialisation,” in *Developmental Cognitive Neuroscience*, Wiley, 2010, ch. 12, pp. 204–223.

[45] E. Bullmore and O. Sporns, “Complex brain networks: graph theoretical analysis of structural and functional systems,” *Nat Rev Neurosci*, vol. 10, pp. 186–198, 2009.

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

[47] S. L. Pendl, A. P. Salzwedel, B. D. Goldman, L. F. Barrett, W. Lin, J. H. Gilmore, and W. Gao, “Emergence of a hierarchical brain during infancy reflected by stepwise functional connectivity,” *Human brain mapping*, vol. 38, pp. 2666–2682, 2017.

[48] 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, 2012.

[49] C. Gilbert and W. Li, “Top-down influences on visual processing,” *Nat Rev Neurosci*, vol. 14, pp. 350–363, 2013.

[50] 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.

[51] F. Pulvermuller, “Brain mechanisms linking language and action,” *Nature reviews neuroscience*, vol. 6, 2005.

[52] D. Casasanto and K. Dijkstra, “Motor action and emotional memory,” *Cognition*, vol. 115, 2010.

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

[54] A. Dereymaeker, K. Pillay, J. Vervisch, M. D. Vos, S. V. Huffel, K. Jansen, and G. Naulaers, “Review of sleep-eeg in preterm and term neonates,” *Early Human Development*, vol. 113, no. 1, pp. 87–103, 2017.

[55] P. Fransson, B. Skiold, M. Engstrom, B. Hallberg, M. Mosskin, U. Åden, H. Lagercrantz, and M. Blennow, “Spontaneous Brain Activity in the Newborn Brain During Natural Sleep—An fMRI Study in Infants Born at Full Term,” *Pediatr Res*, vol. 66, pp. 301–305, 2009.

[56] A. Blasi, E. Mercure, S. Lloyd-Fox, A. Thomson, M. Brammer, D. Sauter, Q. Deeley, G. J. Barker, V. Renval, S. Deoni, D. Gasston, S. C. Williams, M. H.Johnson, A. Simmons, and D. G.M.Murphy, “Early specialization for voice and emotion processing in the infant brain,” *Current Biology*, vol. 21, pp. 1220–1224, 14 2011.

[57] B. Deen, H. Richardson, D. Dilks, and et al., “Organization of high-level visual cortex in human infants,” *Nat Commun*, vol. 8, p. 13 995, 3 2017.

[58] 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, pp. 1–12, 2020.

[59] L. Avitan, M. Teicher, and M. Abeles, “EEG generator–a model of potentials in a volume conductor,” *J Neurophysiol.*, vol. 102, p. 19 710 370, 2009.

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

[61] 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.

[62] 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, 2003.

[63] 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, 2015.

[64] R. Haartsen, B. van der Velde, E. Jones, M. H. Johnson, and C. Kemne, “Using multiple short epochs optimises the stability of infant EEG connectivity parameters,” *Scientific reports*, vol. 10, pp. 1–13, 2020.

[65] P. Grieve, J. Isler, A. Izraelit, B. Peterson, W. Fifer, M. Myers, and R. I. Stark, “EEG functional connectivity in term age extremely low birth weight infants.,” *Clinical Neurophysiology*, vol. 119, pp. 2712–2720, 2008.

[66] E. Meijer, K. Hermans, A. Zwanenburg, W. Jennekens, H. Niemarkt, P. Cluitmans, C. V. Pul, P. Wijn, and P. Andriessen, “Functional connectivity in preterm infants derived from EEG coherence analysis.,” *European journal of paediatric neurology*, vol. 18, pp. 780–789, 2014.

[67] G. Righi, A. L. Tierney, H. Tager-Flusberg, and C. A. Nelson, “Functional Connectivity in the First Year of Life in Infants at Risk for Autism Spectrum Disorder: An EEG Study,” *PLoS ONE*, vol. 9, e105176, 2014.

[68] M. D. Haan, *Infant eeg and event-related potentials*. Psychology Press, 2013.

[69] M. J. Taylor, M. Batty, and R. J. Itier, “The faces of development: a review of early face processing over childhood,” *Journal of cognitive neuroscience*, vol. 16, pp. 1426–1442, 2004.

[70] J. M. Leppänen, M. C. Moulson, V. K. Vogel-Farley, and C. A. Nelson, “An ERP study of emotional face processing in the adult and infant brain,” *Child development*, vol. 78, pp. 232–245, 2007.

[71] S. Hoehl and S. Wahl, “Recording infant ERP data for cognitive research,” *Developmental Neuropsychology*, vol. 37, pp. 187–209, 2012.

[72] J. Pujol, L. Blanco-Hinojo, D. Macia, G. Martínez-Vilavella, J. Deus, V. Pérez-Sola, N. Cardoner, C. Soriano-Mas, and J. Sunyer, “Differences between the child and adult brain in the local functional structure of the cerebral cortex,” *NeuroImage*, vol. 237, p.118150, 2021.

[73] G. Thierry, “The use of event-related potentials in the study of early cognitive development,” *Infant and Child Development*, vol. 14, pp. 85–94, 2005.

[74] A. A. Benasich, N. Choudhury, J. T. Friedman, T. Realpe-Bonilla, C. Chojnowska, and Z. Gou, “The infant as a prelinguistic model for language learning impairments: predicting from event-related potentials to behavior,” *Neuropsychologia*, vol. 44, pp. 396–411, 2006.

[75] J. E. Richards, “Attention affects the recognition of briefly presented visual stimuli in infants: An ERP study.,” *Developmental Science*, vol. 6, pp. 312–328, 2003.

[76] 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, 2012.

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

[78] W. Miltner, C. B. R. J. Jr, G. V. Simpson, and D. S. Ruchkin, “A test of brain electrical source analysis (BESA): A simulation study,” *Electroencephalography and clinical neurophysiology*, vol. 9, pp. 295–310, 2005.

[79] G. D. Reynolds and J. E. Richards, “Cortical source localization of infant cognition,” *Developmental Neuropsychology*, vol. 34, pp. 312–329, 2009.

[80] E. M. Frijia, A. Billing, S. Lloyd-Fox, E. Rosas, L. Collins-Jones, M. M. Crespo-Llado, M. P. Amadó, T. Austin, A. Edwards, L. Dunne, and G. Smith, “Functional imaging of the developing brain with wearable high-density diffuse optical tomography: a new benchmark for infant neuroimaging outside the scanner environment,” *NeuroImage*, vol. 225, p.117490, 2021.

[81] J. Uchitel, E. E. Vidal-Rosas, R. J. Cooper, and H. Zhao, “Wearable, Integrated EEG–fNIRS Technologies: A Review,” *Sensors*, vol. 21, p.6106, 2021.

[82] G. E. Strangman, Z. Li, and Q. Zhang, “Depth sensitivity and source-detector separations for near-infrared spectroscopy based on the Colin27 brain template,” *PLoS One*, vol. 8, e66319, 2013.

[83] T. Suto, M. Fukuda, M. Ito, T. Uehara, and M. Mikuni, “Multichannel near-infrared spectroscopy in depression and schizophrenia: cognitive brain activation study,” *Biological psychiatry*, vol. 55, pp. 501–511, 2004.

[84] M. A. Yucel, A. V. Lühmann, F. Scholkmann, J. Gervain, I. Dan, H. Ayaz, D. Boas, R. J. Cooper, J. Culver, C. E. Elwell, and A. Eggebrecht, “ Best practices for fNIRS publications.,” *Neurophotonics*, vol. 8, p.012101, 2021.

[85] V. Quaresima, S. Bisconti, and M. Ferrari, “A brief review on the use of functional near-infrared spectroscopy (fNIRS) for language imaging studies in human newborns and adults,” *Brain Lang.*, vol. 121, pp. 79–89, 2012.

[86] X. Cui, D. M. Bryant, and A. L. Reiss, “NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation,” *Neuroimage*, vol. 59, pp. 2430–2437, 2012.

[87] L. Holper, T. Muehlemann, F. Scholkmann, K. Eng, D. Kiper, and M. Wolf, “Testing the potential of a virtual reality neurorehabilitation system during performance of observation, imagery and imitation of motor actions recorded by wireless functional near-infrared spectroscopy (fNIRS),” *Journal of neuroengineering and rehabilitation*, vol. 7, pp. 1–13, 2010.

[88] M. Saadati, J. Nelson, and H. Ayaz, “Multimodal FNIRS-EEG classification using deep learning algorithms for brain-computer interfaces purposes,” 2019.

[89] P. Pinti, A. H. Felix Scholkmann, P. Burgess, and I. Tachtsidis, “Current status and issues regarding pre-processing of fnirs neuroimaging data: An investigation of diverse signal filtering methods within a general linear model framework,” *Frontiers in Human Neuroscience*, vol. 12, p. 505, 2019.

[90] S. Baek, S. Marques, K. Casey, M. Testerman, F. McGill, and L. Emberson, “Attrition Rate in Infant fNIRS Research: A Meta-Analysis,” *BioRxiv*, 2021.

[91] J. C. Ye, S. Tak, K. E. Jang, J. Jung, and J. Jang, “NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy,” *Neuroimage*, vol. 44, pp. 428–447, 2009.

[92] T. J. Huppert, S. G. Diamond, M. A. Franceschini, and D. A. Boas, “HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain,” *Appl. Opt.*, vol. 48, pp. 280–298, 2009.

[93] X. Hou, Z. Zhang, C. Zhao, L. Duan, Y. Gong, Z. Li, and C. Zhu, “NIRS-KIT: a MATLAB toolbox for both resting-state and task fNIRS data analysis,” *Neurophotonics*, vol. 8, p.010802, 2021.

[94] *OPENFNIRS-An fNIRS hardware and software ecosystem*, https://openfnirs.org/software/homer/, Accessed: 2021-12-16.

[95] W. B. Baker, A. B. Parthasarathy, D. R. Busch, R. C. Mesquita, J. H. Greenberg, and A. G. Yodh, “Modified Beer-Lambert law for blood flow,” *Biomed Opt Express*, vol. 5, no. 11, pp. 4053–4075, 2014.

[96] A. K. Singh, M. Okamot, H. D. V. Jurcak, and I. Dan, “Spatial registration of multichannel multi-subject fNIRS data to MNI space without MRI,” *Neuroimage*, vol. 27, pp. 842–851, 2005.

[97] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, “Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space,” *Journal of computer assisted tomography*, vol. 18, pp. 192–205, 1994.

[98] D. Tsuzuki and I. Dan, “Spatial registration for functional near-infrared spectroscopy: from channel position on the scalp to cortical location in individual and group analyses,” *Neuroimage*, vol. 85, pp. 92–103, 2014.

[99] S. Lloyd-Fox, J. E. Richards, A. Blasi, D. Murphy, C. Elwell, and M. Johnson, “Coregistering functional near-infrared spectroscopy with underlying cortical areas in infants.,” *Neurophotonics*, vol. 1, p.025006, 2014.

[100] T. Yamada, K. Matsuda, T. Iwano, and S. Umeyama, “Precise spatial co-registration in simultaneous fNIRS and fMRI measurements using markers coaxially fixable to the optodes,” *Optical Techniques in Neurosurgery, Neurophotonics, and Optogenetics*, vol. 8928, p. 89280S, 2014.

[101] L. L. Emberson, B. D. Zinszer, R. D. S. Raizada, and R. N. Aslin, “Decoding the infant mind: Multivariate pattern analysis (MVPA) using fNIRS,” *PLoS ONE*, vol. 12, e0172500, 2017.

[102] R. Luke, M. Shader, E. Larson, A. Gramfort, A. K. Lee, and D. McAlpine., “Oxygenated hemoglobin signal provides greater predictive performance of experimental condition than de-oxygenated,” *BioRxiv*, pp. 705–721, 2021.

[103] K. Sakatani, S. Chen, W. Lichty, H. Zuo, and Y. P. Wang, “Cerebral blood oxygenation changes induced by auditory stimulation in newborn infants measured by near infrared spectroscopy,” *Early human development*, vol. 53, pp. 229–236, 1999.

[104] C. M. Lu, Y. J. Zhang, B. B. Biswal, Y. F. Zang, D. L. Peng, and C. Z. Zhu, “Use of fNIRS to assess resting state functional connectivity,” *Journal of neuroscience methods*, vol. 186, pp. 242–249, 2010.

[105] J. Andreu-Perez, L. L. Emberson, M. Kiani, M. L. Filippetti, H. Hagras, and S. Rigato, “Explainable Artificial Intelligence Based Analysis for Developmental Cognitive Neuroscience,” *Commun Biol*, vol. 4, p. 1077, 2021.

[106] L. J. Powell, B. Deen, and R. Saxe, “Using individual functional channels of interest to study cortical development with fNIRS.,” *Developmental science*, vol. 21, p.e12595, 2018.

[107] G. Taga and K. Asakawa, “Selectivity and localization of cortical response to auditory and visual stimulation in awake infants aged 2 to 4 months,” *Neuroimage*, vol. 36, pp. 1246–1252, 2007.

[108] M. D. Pfeifer, F. Scholkmann, and R. Labruyère, “Signal processing in functional near-infrared spectroscopy (fNIRS): methodological differences lead to different statistical results,” *Frontiers in human neuroscience*, vol. 11, 2018.

[109] T. Arichi, G. Fagiolo, M. Varela, A. Melendez-Calderon, A. Allievi, N. Merchant, N. Tusor, S. J. Counsell, E. Burdet, C. Beckmann, and A. D. Edwards, "Development of BOLD signal hemodynamic responses in the human brain," *Neuroimage*, vol. 63, pp. 663–673, 2012.

[110] T. Hiroyasu, S. Yoshitake, and S. Hiwa, "Adaptive HRF and BF approaches to fNIRS activation analysis," *In Front. Neuroinform- Conference Abstract*, 2016.

[111] L. L. Emberson, G. Cannon, H. Palmeri, J. E. Richards, and R. N. Aslin, "Using fnirs to examine occipital and temporal responses to stimulus repetition in young infants: Evidence of selective frontal cortex involvement," *Developmental Cognitive Neuroscience*, vol. 23, pp. 26–38, 2017.

[112] R. N. Aslin, M. Shukla, and L. L. Emberson, "Hemodynamic correlates of cognition in human infants," *Annu Rev Psychol.*, vol. 66, pp. 349–379, 2015.

[113] 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, 2005.

[114] 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, 1993.

[115] M. S. A. Lee and L. Floridi and A. Denev, "Innovating with confidence: embedding AI governance and fairness in a financial services risk management framework," in. Springer, 2021, ch. 9, pp. 353–371.

[116] K. J. Yaxley, K. F. Joiner, and H. Abbass, "Drone approach parameters leading to lower stress sheep flocking and movement: sky shepherding," *Scientific reports*, vol. 11, pp. 1–9, 2021.

[117] R. Li, D. Auer, C. Wagner, and X. Chen, "A generic ensemble based deep convolutional neural network for semi-supervised medical image segmentation," *In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1168–1172, 2020.

[118] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain intelligence: go beyond artificial intelligence," *Mobile Networks and Applications*, vol. 23, pp. 368–375, 2018.

[119] J. Andreu-Perez, H. Pérez-Espinosa, E. Timonet, M. Kiani, and et al., *A generic deep learning based cough analysis system from clinically validated samples for point-of-need covid-19 test and severity levels*, Early Access, 2021.

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

[121] G. Alicioglu and B. Sun, “A survey of visual analytics for explainable artificial intelligence methods,” *Computers & Graphics*, 2021.

[122] L. A. Zadeh, “Outline of a New Approach to the Analysis of Complex Systems and Decision Processes,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 1, pp. 28–44, 1975.

[123] ——, “Fuzzy Sets,” *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[124] J. M. Mendel, *Uncertain rule-based fuzzy systems*. Springer, 2017.

[125] F. Herrera, “Genetic fuzzy systems: Taxonomy, current research trends and prospects,” *Evolutionary Intelligence*, vol. 1, no. 1, pp. 27–46, 2008.

[126] J. R. Boston, “Effects of the shape of fuzzy membership functions on fuzzy inference,” *Proceedings of 3rd International Symposium on Uncertainty Modeling and Analysis and Annual Conference of the North American Fuzzy Information Processing Society*, pp. 32–37, 1995.

[127] E. E. Omizegba and G. E. Adebayo, “Optimizing fuzzy membership functions using particle swarm algorithm,” *IEEE International Conference on Systems, Man and Cybernetics*, pp. 3866–3870, 2009.

[128] J. M. Mendel, “Non-singleton fuzzification made simpler,” *Information Sciences*, vol. 559, pp. 286–308, 2021.

[129] C. Li, J. Yi, G. Zhang, and M. Wang, “Modeling of thermal comfort words using interval type-2 fuzzy sets,” in *Fourth International Conference on Intelligent Control and Information Processing (ICICIP)*, 2013, pp. 626–631.

[130] D. Wu and J. M. Mendel, “Uncertainty measures for interval type-2 fuzzy sets,” *Information sciences*, vol. 177, pp. 5378–5393, 2007.

[131] N. N. Karnik and J. M. Mendel, “Centroid of a type-2 fuzzy set,” *Information Sciences*, vol. 132, pp. 195–220, 2001.

[132] M. Nie and W. W. Tan, “Towards an efficient type-reduction method for interval type-2 fuzzy logic systems,” *IEEE International Conference on Fuzzy Systems*, pp. 1425–1432, 2008.

[133] M. Mizumoto and K. Tanaka, “Some properties of fuzzy sets of type 2,” *Information and control*, vol. 31, pp. 312–340, 1976.

[134] C. Wagner and H. Hagras, “Toward general type-2 fuzzy logic systems based on zSlices,” *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 4, pp. 637–660, 2010.

[135] 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, 2017.

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

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

[138] A. M. Zador, “A critique of pure learning and what artificial neural networks can learn from animal brains,” *Nat Commun*, vol. 10, pp. 1–7, 3770 2019.

[139] 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, 2009.

[140] 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,” *Transl Psychiatry*, vol. 10, pp. 1–9, 2020.

[141] 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, 4 2020.

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

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

[144] 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,” *BMJ Open*, vol. 11, 2021.

[145] 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.

[146] 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, no. 31, pp. 9585–9590, 2015.

[147] W. Stach, L. Kurgan, W. Pedrycz, and M. Reformat, “Genetic learning of fuzzy cognitive maps,” *Fuzzy Sets and Systems*, vol. 153, pp. 371–401, 2005.

[148] S. Bueno and J. L. Salmeron, “Benchmarking main activation functions in fuzzy cognitive maps,” *Expert Systems with Applications*, vol. 36, pp. 5221–5229, 2009.

[149] W. Stach, L. Kurgan, and W. Pedrycz, “Higher-order fuzzy cognitive maps,” *NAFIPS*, 2006.

[150] E. I. Papgeorgiou, “Learning algorithms for fuzzy cognitive maps - a review study,” *IEEE Transactions of Systems, Man, and Cybernetics*, vol. 42, pp. 150–163, 2011.

[151] 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, 2017.

[152] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *International Joint Conference on Artificial Intelligence*, vol. 14, 1995, pp. 1137–1145.

[153] E. L. Aurbach, K. E. Prater, E. Cloyd, T. Emily, and L. Lindenfeld, “Foundational Skills for Science Communication: A Preliminary Framework,” American Association for the Advancement of Science (AAAS), Tech. Rep., 2019.

[154] J. M. Garibaldi, M. Jaroszewski, and S. Musikasuwan, “Nonstationary Fuzzy Sets,” *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 4, pp. 1072–1086, 2008.

[155] A. V. Kostikova, P. V. Tereliansky, A. V. Shuvaev, V. N. Parakhina, and P. N. Timoshenko4, “Expert Fuzzy Modeling of dynamic properties of complex systems,” *ARPN Journal of Engineering and Applied Sciences*, vol. 11, no. 17, pp. 10 222–10 230, 2016.

[156] H. Maeda, S. Asaoka, and S. Murakami, “Dynamical fuzzy reasoning and its application to system modeling,” *Fuzzy Sets and Systems*, vol. 80, no. 1, pp. 101–109, 1996.

[157] D. Kozen and M. Timme, “" Indefinite summation and the Kronecker delta,” *Https://hdl.handle.net/1813/83* 2007.

[158] Q. Liang and J. Mendel, “Interval type-2 fuzzy logic systems: theory and design,” *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 5, pp. 535–550, 2000.

[159] C. Chen, D. Wu, J. M. Garibaldi, R. I. John, J. Twycross, and J. M. Mendel, “A Comprehensive Study of the Efficiency of Type-Reduction Algorithms,” *IEEE Transactions on Fuzzy Systems*, vol. 29, pp. 1556–1566, 2021.

[160] S. Mirjalili, “Genetic algorithm,” *In Evolutionary algorithms and neural networks*, pp. 43–55, 2019.

[161] L. M. Candanedo and V. Feldheim, “Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models,” *Energy and Buildings*, vol. 112, pp. 28–39, 2016.

[162] D. P. Filev and I. Kolmanovsky, “Generalized markov models for real-time modeling of continuous systems,” *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 4, pp. 983–998, 2013.

[163] J. Andreu-Perez, D. R. Leff, K. Shetty, A. Darzi, and G-Z. Yang, “Disparity in frontal lobe connectivity on a complex bimanual motor task aids in classification of operator skill level,” *Brain Connectivity*, vol. 6, 2016.

[164] S. Lloyd-Fox, R. Wu, J. E. Richards, C. E. Elwell, and M. M. H. Johnson, “Cortical activation to action perception is associated with action production abilities in young infants.,” *Cerebral Cortex*, vol. 25, pp. 289–297, 2015.

[165] R. Nishiyori, M. K. Harris, K. Baur, and S. K. Meehan, “Changes in cortical hemodynamics with the emergence of skilled motor ability in infants: An fnirs study,” *Brain research*, vol. 1772, p.147666, 2021.

[166] N. D. Tam and G. Zouridakisb, “Temporal decoupling of oxy- and deoxy-hemoglobin hemodynamic responses detected by functional near-infrared spectroscopy (fnirs),” *Journal of Biomedical Engineering and Medical Imaging*, vol. 1, 2014.

[167] N. D. Tam and G. Zouridakis, “Differential temporal activation of oxy- and deoxy-hemodynamic signals in optical imaging using functional near-infrared spectroscopy (fnirs),” *BMC Neuroscience*, vol. 16, 2015.

[168] I. Tachtsidis and F. Scholkmann, “False positives and false negatives in functional near-infrared spectroscopy: Issues, challenges, and the way forward,” *Neurophotonics*, vol. 3, 2016.

[169] S. Sasai, F. Homae, H. Watanabe, and G. Taga, “Frequency-specific functional connectivity in the brain during resting state revealed by nirs,” *NeuroImage*, vol. 56, pp. 252–257, 2011.

[170] A. R. Laird, J. L. Robinson, K. M. McMillan, D. Tordesillas-Gutierrez, and S. T. Moran, “Comparison of the disparity between talairach and mni coordinates in functional neuroimaging data: Validation of the lancaster transform,” *Neuroimage*, vol. 51, pp. 677–683, 2010.

[171] J. L. Lancaster, M. G. Woldorff, L. M. Parsons, M. Liotti, C. S. S. Freitas, L. Rainey, P. V. K. D. D. Nickerson, S. A. Mikiten, and P. T. Foxm, “Automated talairach atlas labels for functional brain mapping,” *Human Brain Mapping*, vol. 10, pp. 120–131, 2000.

[172] M. Kiani, J. Andreu-Perez, D. R. Leff, A. Darzi, and G. Z. Yang, “Shedding light on surgeons’ cognitive resilience: A novel method of topological analysis for brain networks,” *The Hamlyn Symposium on Medical Robotics*, p. 55, 2014.

[173] M. Kiani, J. Andreu-Perez, and E. I. Papageorgiou, “Improved estimation of effective brain connectivity in functional neuroimaging through higher order fuzzy cognitive maps,” *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017.

[174] H. Abdi, “Rv coefficient and congruence coefficient. In: Salkind N.J (eds.),” *Encyclopedia of measurement and statistics, 1st ed. New York, NY: Sage Publications*, 207.

[175] J. N. Sanes, “Neocortical mechanisms in motor learning,” *Current Opinion in Neurobiology*, vol. 13, pp. 225–231, 2003.

[176] R. Kawai, T. Markman, R. Poddar, A. L. F. R. Ko, A. K. Dhawale, A. R. Kampff, and B. P. Olveczky, “Motor cortex is required for learning but not for executing a motor skill,” *Neuron*, vol. 86, pp. 800–812, 2015.

[177] F. Chersi, M. Mirolli, G. Pezzulo, and G. Baldassarre, “A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning,” *Neural Networks*, vol. 41, pp. 212–224, 2013.

[178] J. Gervain, J. Mehler, J. F.Werker, C. A. Nelson, G. Csibra, S. L.-F. M. Shukla, and R. N. Aslin, “Near-infrared spectroscopy: a report from the McDonnell infant methodology consortium,” *Dev Cogn Neurosci*, vol. 1, pp. 22–46, 2011.

[179] R. N. Aslin, M. Shukla, and L. L. Emberson, “Hemodynamic correlates of cognition in human infants,” *Annual Review of Psychologyn*, vol. 66, pp. 349–379, 2015.

[180] G. Taga, H. Watanabe, and F. Homae, “Spatiotemporal properties of cortical haemodynamic response to auditory stimuli in sleeping infants revealed by multi-channel near-infrared spectroscopy,” *Royal Society*, vol. 369, pp. 4495–4511, 2011.

[181] G. Taga, K. Asakawa, K. Hirasawa, and Y. Konishi, “Hemodynamic responses to visual stimulation in occipital and frontal cortex of newborn infants: A near-infrared optical topography study,” *Pathophysiology*, vol. 10, pp. 277–281, 2004.

[182] T. Wilcox, H. Bortfeld, R. Woods, E. Wruck, and D. A. Boas, “Hemodynamic response to featural changes in the occipital and inferior temporal cortex in infants: A preliminary methodological exploration,” *Developmental Science*, vol. 11, pp. 361–370, 2008.

[183] H. Watanabe, F. Homae, T. Nakano, and G. Taga, “Functional activation in diverse regions of the developing brain of human infants,” *NeuroImage*, vol. 43, pp. 346–357, 2008.

[184] H. Halit, M. de Haan, and M. Johnson, “Cortical specialisation for face processing: Face-sensitive event-related potential components in 3- and 12-month-old infants,” *NeuroImage*, vol. 19, pp. 1180–1193, 2003.

[185] N. Tzourio-Mazoyer, S. D. Schonen, F. Crivello, B. Reutter, Y. Aujard, and B. Mazoyer, “Neural correlates of woman face processing by 2-month-old infants,” *Neuroimage*, vol. 15, pp. 454–461, 2001.

[186] M. S. Beauchamp, K. E. Lee, J. V. Haxby, and A. Martin, “Fmri responses to video and point-light displays of moving humans and manipulable objects,” *Journal of Cognitive Neuroscience*, vol. 15, pp. 991–1001, 2003.

[187] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, “The distributed human neural system for face perception,” *Trends Cogn Sci*, vol. 4, pp. 223–233, 6 2000.

[188] T. Grossmann and T. S. A. D.Friederici, “Developmental changes in infants’ processing of happy and angry facial expressions: A neurobehavioral study,” *Brain and Cognition*, vol. 64, pp. 30–41, 2007.

[189] Y. Minagawa-Kawai, S. Matsuoka, I. Dan, N. Naoi, K. Nakamura, and S. Kojima, “Prefrontal activation associated with social attachment: Facial-emotion recognition in mothers and infants,” *Cerebral Cortex*, vol. 19, pp. 284–292, 2009.

[190] M. de Haan and M. H. Johnson, “Overview of prefontal development,” in *The Cognitive Neuroscience of Development*. East Sussex, UK: Psychology Press, 2005, pp. 178–186.

[191] D. Maurer, L. Gibson, and F. Spector, “Synesthesia in infants and very young children,” in *The oxford handbook of synesthesia*. Oxford, UK: Oxford University Press, 2013, vol. 12, pp. 46–53.

[192] W. Gao, J. H. Gilmore, K. S. Giovanello, J. K. Smith, D. Shen, H. Zhu, and W. Lin, “Temporal and spatial evolution of brain network topology during the first two years of life,” *PLoS ONE*, vol. 6, e25278, 2011.

[193] F. Homae, H. Watanabe, T. Otobe, T. Nakano, T. Go, Y. Konishi, and G. Taga, “Development of global cortical networks in early infancy,” *Journal of Neuroscience*, vol. 30, pp. 4877–4882, 2010.

[194] T. Grossmann, “Mapping prefrontal cortex functions in human infancy,” *Infancy*, vol. 18, pp. 303–24, 2013.

[195] G. Dehaene-Lambertz, “Cerebral specialization for speech and non-speech stimuli in infants,” *Journal of Cognitive Neuroscience*, vol. 12, pp. 449–460, 2000.

[196] T. Imada, Y. Zhang, M. Cheour, S. Taulu, A. Ahonen, and P. K. Kuhl, “Infant speech perception activates broca’s area: A developmental magnetoencephalography study,” *Brain Imaging*, vol. 17, pp. 957–962, 2006.

[197] N. R. Altman and B. Bernal, “Brain activation in sedated children: Auditory and visual functional mr imaging,” *Pediatric Imaging*, vol. 221, pp. 56–63, 2001.

[198] G. Dehaene-Lambertz, S. Dehaene, and L. Hertz-Pannier, “Functional neuroimaging of speech perception in infants,” *Science*, vol. 298, pp. 2013–2015, 2002.

[199] G. Taga, F. Homae, and H. Watanabe, “Effects of source-detector distance of near infrared spectroscopy on the measurement of the cortical hemodynamic response in infants,” *NeuroImage*, vol. 38, pp. 452–460, 2007.

[200] S. J. Hespos, A. L. Ferry, C. J. Cannistraci, J. Gore, and S. Park, “Using optical imaging to investigate functional cortical activity in human infants,” in *Imaging the Brain with Optical Methods*. New York: Springer, 2010, pp. 159–176.

[201] A. Sherstinsky, “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[202] S. Fine, Y. Singer, and N. Tishby, “The hierarchical hidden Markov model: Analysis and applications,” *Machine learning*, vol. 32, pp. 41–62, 1998.

[203] Y. Ben-Haim and E. Tom-Tov, “A Streaming Parallel Decision Tree Algorithm. Journal of Machine Learning Research,” *Journal of Machine Learning Research*, vol. 11, 2010.

[204] P. Remy, *Temporal convolutional networks for keras*, https://github.com/philipperemy/keras-tcn, 2020.

[205] I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning,” *MIT Press*, 2016.

[206] H. R. Rocha, I. H. Honorato, R. Fiorotti, W. C. Celeste, L. J. Silvestre, and J. A. Silva, “An artificial intelligence based scheduling algorithm for demand-side energy management in smart homes,” *Applied Energy*, vol. 282, p. 116 145, 2021.

[207] X. Jiang, Z. Yi, and J. C. Lv, “Fuzzy SVM with a new fuzzy membership function,” *Neural Computing and Applications*, vol. 15, 2006.

Appendix A

Karnik-Mendel (KM) Algorithm

The KM algorithm [131] for computing the left centroid of an IT2 fuzzy is outlined in Algorithm 5. For finding the **left centroid,** c_l as outlined in (3.18), the value of L needs to be determined as follows:

Algorithm 5: The KM Algorithm for computing the left centroid, c_l.

```
Result: The left centroid, $c_l$.
Initialise $\theta_i \leftarrow \frac{\underline{\mu}_A(x_i)+\overline{\mu}_A(x_i)}{2}$ where $i = 1, 2, ..., I$ and $c_2 \leftarrow 0$
Compute $c_1 \leftarrow \frac{\sum_{i=1}^{I} x_i\theta_i}{\sum_{i=1}^{I} \theta_i}$
while $c_1 \neq c_2$ do
    $c_1 \leftarrow c_2$
    Find $\mathfrak{I}$ $(1 \leq \mathfrak{I} \leq I - 1)$ such that $x_{\mathfrak{I}} \leq c_1 \leq x_{\mathfrak{I}+1}$
    Set $\theta_i \leftarrow \begin{cases} \overline{\mu}_{\tilde{A}(x_i)} & i \leq \mathfrak{I} \\ \underline{\mu}_{\tilde{A}(x_i)} & i > \mathfrak{I} \end{cases}$
    Compute $c_2 \leftarrow \frac{\sum_{i=1}^{I} x_i\theta_i}{\sum_{i=1}^{I} \theta_i}$
Set $c_l \leftarrow c_2$ and $L \leftarrow \mathfrak{I}$
```

The following steps outline the procedure for computing the **right centroid,** c_r as outlined in (3.19), using the KM algorithm outlined in Algorithm 6.

Algorithm 6: The KM Algorithm for computing the right centroid, c_r.

```
**Result:** The right centroid, $c_r$.
Initialise $\theta_i \leftarrow \frac{\underline{\mu}_A(x_i)+\overline{\mu}_A(x_i)}{2}$where $i = 1, 2, ..., I$ and $c_2 \leftarrow 0$
Compute $c_1 \leftarrow \frac{\sum_{i=1}^{I} x_i\theta_i}{\sum_{i=1}^{I}\theta_i}$
**while** $c_1 \neq c_2$ **do**
    $c_1 \leftarrow c_2$
    Find $\mathfrak{I}$ $(1 \leq \mathfrak{I} \leq I - 1)$ such that $x_{\mathfrak{I}} \leq c_1 \leq x_{\mathfrak{I}+1}$
    Set $\theta_i \leftarrow \begin{cases} \underline{\mu}_{\tilde{A}(x_i)} & i \leq \mathfrak{I} \\ \overline{\mu}_{\tilde{A}(x_i)} & i > \mathfrak{I} \end{cases}$
    Compute $c_2 \leftarrow \frac{\sum_{i=1}^{I} x_i\theta_i}{\sum_{i=1}^{I}\theta_i}$
Set $c_r \leftarrow c_2$ and $R \leftarrow \mathfrak{I}$
```

Appendix B

List of Symbols

No.	Symbol	Description
1	α	Index for time intervals of TT2FS i.e. $\alpha \in [1, ..., V]$ with V as the total number of time intervals
2	β	Weight/Contribution of regressor
3	δ	Kronecker delta function
4	Δ	Change
5	λ	Wavelength
6	ζ	Hb concentration
7	ϵ	Extinction coefficient of the chromophore
8	κ	Estimated Haemodynamic Response Function (HRF)
9	ξ	Regressors
10	Ω	Error between true and estimated Haemodynamic Response Function (HRF)
11	χ	Concept (or fNIRS channel) value
12	τ	Iteration number
13	ν	Effective Connectivity (EC) value
14	Θ	Total number of iterations
15	μ_A	Membership degree of T1 fuzzy set A

continued . . .

. . . continued

No.	Symbol	Description
16	μ_B	Membership degree of T1 fuzzy set B
17	$\overline{\mu_{\tilde{A}}}$	Upper membership degree of IT2 fuzzy set, $\tilde{A}$
18	$\overline{\mu_{\tilde{B}}}$	Upper membership degree of IT2 fuzzy set, $\tilde{B}$
19	$\underline{\mu_{\tilde{A}}}$	Lower membership degree of IT2 fuzzy set, $\tilde{A}$
20	$\underline{\mu_{\tilde{B}}}$	Lower membership degree of IT2 fuzzy set, $\tilde{B}$
21	$\mu_{\tilde{A}}$	Secondary membership degree of GT2 fuzzy set, $\tilde{A}$
22	$\mu_{\tilde{B}}$	Secondary membership degree of GT2 fuzzy set, $\tilde{B}$
23	$\ell_x(u)$	Equivalent of $\mu_{\tilde{A}}(x, u)$
24	$\hbar_x(u)$	Equivalent of $\mu_{\tilde{B}}(x, u)$
25	Φ	Total number of antecedents
26	ϕ	A particular fNIRS channel
27	Ψ_q	The antecedent(s) in pattern number q
28	ρ	Phenotype
29	$\wp$	Order of the FCM or EFCM
30	ϱ	Sigmoid function
31	ι	Parameter defining the shape of sigmoid function
32	ϑ	Optimised soft regularizer values in EFCM
33	Υ	Corresponding CoLs associated with each channel
34	Γ	Fuzzy Relation
35	$\Im$	Index in KM Algorithm
36	γ	Numeric values for the range of each of the CoLs of all the $\mathbf{F}$ fNIRS channels

continued . . .

. . . continued

No.	Symbol	Description
37	π	Pattern Transition Probability
38	ϖ	Attenuation
39	ω	Primary membership degree of GT2 fuzzy set $\tilde{B}$
40	η	Joint possibility for two patterns to occur in their respective time intervals
41	A	T1 (Type-1) Fuzzy set
42	$\tilde{A}$	Interval Type-2 (IT2) or General Type-2 (GT2) Fuzzy Set
43	$\vec{A}$	Temporal Type-2 Fuzzy Set
44	A_e	Type-1 Embedded fuzzy sets within IT2 fuzzy sets.
45	Ac	Active
46	b	Individual solution of GA
47	B	T1 (Type-1) Fuzzy set
48	$\tilde{B}$	Interval Type-2 (IT2) or General Type-2 (GT2) Fuzzy Set
49	$\vec{B}$	Temporal Type-2 Fuzzy Set
50	c	Centroid
51	c_l	Left centroid of IT2 fuzzy set, $\tilde{A}$
52	c_r	Right centroid of IT2 fuzzy set, $\tilde{A}$
53	$C_{\tilde{A}}$	The union of embedded T1 fuzzy sets' (A_e) centroids. Interval enclosed by left and right centroids of IT2 fuzzy set, $\tilde{A}$.
54	Ch	Channel
55	$conf_q$	Confidence of pattern number q
56	d	fNIRS source-detector distance

continued . . .

. . . continued

No.	Symbol	Description
57	$\mathbf{D}$	EC Connection Matrix in EFCM
58	DPF	Differential Pathlength Factor
59	DS	Dominance Score of a pattern
60	$\mathbf{E}$	Total number of data instances (or fNIRS trials)
61	f	Conditional distribution
62	F	Frequency of occurrence domain
63	g	Discrete conditional relative frequency
64	G	Total number of patterns in time interval Δt
65	h	Association degree
66	H	Total number of patterns in time interval Δt^{+}
67	i	Generic Iterator
68	I_{IN}	Input NIR light
69	I_{OUT}	Output NIR light
70	IA	Inactive
71	j	Generic Iterator
72	J_x	Universe of primary membership value
73	k	Generic Iterator
74	K	Total number of output classes
75	l	Iterator for CoLs i.e. $l \in [1, ..., W]$where W is the total number of CoLs
76	L	Left switch point in KM algorithm
77	$\mathbf{M}$	Total number of concepts (or fNIRS channels)

continued . . .

. . . continued

No.	Symbol	Description
78	N	Total number of points for discretising time.
79	O	Output class or label of a pattern
80	P	Pattern
81	q	Pattern number
82	Q	Total number of patterns
83	R	Right switch point in KM algorithm
84	sup_q	Support of pattern number q
85	t	Time
86	T	Time domain
87	u	Primary membership degree of IT2 or GT2 fuzzy set
88	U	Universe of secondary variable of IT2 or GT2 fuzzy set
89	v	Primary membership degree of GT2 fuzzy set $\tilde{A}$
90	V	Total number of time intervals
91	VA	Very Active
92	w	Firing Strength
93	W	Total number of conceptual labels (CoLs)
94	x	Input value
95	x^*	Defuzzified/crisp value of a fuzzy set or centroid of T1 fuzzy set which is also a crisp value
96	X	Universe of discourse
97	X_d	Discrete universe of discourse

continued . . .

. . . continued

No.	Symbol	Description
98	y	Strength of influence on a given concept (or fNIRS channel)
99	Y_q	Consequent fuzzy set for pattern number q
100	z	Location of z-slice
101	$\mathbf{Z}$	Soft regularizer matrix in EFCM

Appendix C

List of Abbreviations

No.	Abbreviation	Definition
1	3D	Three Dimensional
2	4D	Four Dimensional
3	AI	Artificial Intelligence
4	BCI	Brain-Computer Interface
5	CA	Connectivity Analysis
6	CNN	Convolutional Neural Networks
7	CO_2	Carbon dioxide
8	DCN	Developmental Cognitive Neuroscience
9	DCM	Dynamic Causal Modeling
10	deoxy-Hb	deoxygenated Haemoglobin
11	DLPFC	Dorsolateral Prefrontal Cortex
12	DNN	Deep Neural Networks
13	DPF	Differential Pathlength Factor
14	DT	Decision Tree
15	EC	Effective Connectivity
16	EEG	Electroencephalogram

continued . . .

. . . continued

No.	Abbreviation	Definition
17	EFCM	Effective Fuzzy Cognitive Map
18	ERP	Event Related Potential
19	EXs	Experts
20	FC	Functional Connectivity
21	FCM	Fuzzy Cognitive Map
22	FLS	Fuzzy Logic System
23	fMRI	functional Magnetic Resonance Imaging
24	fNIRS	functional Near-Infrared Spectroscopy
25	FNR	False Negative Rate
26	FOU	Footprint of Uncertainty
27	FPR	False Positive Rate
28	GA	Genetic Algorithm
29	GC	Granger Causality
30	GLM	General Linear Models
31	GT2	General Type-2
32	GT2-FLS	General Type-2 Fuzzy Logic System
33	Hb	Haemoglobin
34	HMM	Hidden Markov Models
35	HRF	Haemodynamic response function
36	IFM	Interpolated Functional Manifold
37	IS	Interactive Specialisation

continued . . .

. . . continued

No.	Abbreviation	Definition
38	IT2	Interval Type-2
39	IT2-FLS	Interval Type-2 Fuzzy Logic System
40	JI	Jaccard Index
41	KM	Karnik-Mendel
42	LS	Laparoscopic Surgery
43	LSTM	Long Short-Term Memory
44	mBBL	modified Beer-Lambert Law
45	MCC	Mathew’s Correlation Coefficient
46	MF	Membership Function
47	MLP	Multilayer Perceptron
48	MNI	Montreal Neurological Institute
49	MR	Magentic Resonance
50	MRI	Magnetic Resonance Imaging
51	MVM	Multivariate Matrix
52	MVPA	Multivariate Pattern Analysis
53	N170	A negative ERP observed at 170ms
54	NIR	Near-Infrared
55	NPV	Negative Predictive Value
56	NVs	Novices
57	oxy-Hb	oxygenated Haemoglobin
58	P100	A positive ERP observed at 100ms

continued . . .

. . . continued

No.	Abbreviation	Definition
59	PFC	Prefrontal Cortex
60	PPI	Psychophysiological Interaction
61	PPV	Positive Predictive Value
62	PTM	Pattern Transition Matrix
63	PTP	Pattern Transition Possibility
64	ReLU	Rectified Linear Unit
65	RepL	Representation Learning
66	RF	Random Forest
67	RMSE	Root Mean Square Error
68	ROI	Region of interest
69	RR	Ridge Regression
70	SD	Standard deviation
71	SNR	Signal-to-Noise Ratio
72	SPM	Statistical Parametric Mapping
73	SVM	Support Vector Machine
74	T1	Type-1
75	T1-FLS	Type-1 Fuzzy Logic System
76	TCN	Temporal Convolutional Networks
77	TMF	Temporal Membership Function
78	TN	True Negative
79	TNs	Trainees

continued . . .

. . . continued

No.	Abbreviation	Definition
80	TP	True Positive
81	TS-ZS	Time Slice followed by Z Slice
82	TT2FS	Temporal Type-2 Fuzzy Set
83	TXAI	Time-dependent eXplainable Artificial Intelligence
84	TXAI-IS	TXAI- Inference System
85	XAI	eXplainable Artificial Intelligence
86	xMVPA	eXplainable Multivariate Pattern Analysis