A Type-2 Fuzzy Logic Based Explainable Artificial Intelligence System for Developmental Neuroscience

Mehrin Kiani, Javier Andreu-Perez, Hani Hagras
School of Computer Science and Electronic Engineering,
University of Essex

Maria Laura Filippetti, Silvia Rigato
Department of Psychology,
University of Essex

***Abstract*—Research in developmental cognitive neuroscience face challenges associated not only with their population (infants and children who might not be too willing to cooperate) but also in relation to the limited choice of neuroimaging techniques that can non-invasively record brain activity. For example, magnetic resonance imaging (MRI) studies are unsuitable for developmental cognitive studies because they require participants to stay still for a long time in a noisy environment. In this regard, functional Near-infrared spectroscopy (fNIRS) is a fast-emerging de-facto neuroimaging standard for recording brain activity of young infants. However, the absence of associated anatomical image, and a standard technical framework for fNIRS data analysis remains a significant impediment to advancement in gaining insights into the workings of developing brains. To this end, this work presents an Explainable Artificial Intelligence (XAI) system for infant's fNIRS data using a multivariate pattern analysis (MVPA) driven by a genetic algorithm (GA) type-2 Fuzzy Logic System (FLS) for classification of infant's brain activity evoked by different stimuli. This work contributes towards laying the foundation for a transparent fNIRS data analysis that holds the potential to enable researchers to map the classification result to the corresponding brain activity pattern which is of paramount significance in understanding how developing human brain functions.**

***Index Terms*—Explainable Artificial Intelligence, Type-2 Fuzzy Systems, Genetic Algorithm, Mutivariate Pattern Analysis, Developmental Cognitive Neuroscience**

I. Introduction

Developmental cognitive neuroscience studies are essential in paving the way for our understanding of how human brain develops over time. In the realm of cognitive neuroscience, functional Magnetic Resonance Imaging (fMRI) studies have played a significant role in providing profound insights into the workings of the human brain. The superior spatial location associated with anatomical image meant that fMRI studies can conclude with high confidence that the recorded brain activity is indeed specific to a certain brain region. With the adoption of multivariate techniques for fMRI data analysis such as multivariate patterns analysis (MVPA [1]) the aim of the fMRI studies have shown a significant shift from 'where'in the brain i.e. from establishing localization of brain activity (univariate analysis) to 'how'in the brain is the information represented that can decode the brain activity (multivariate analysis). Despite the great prowess of fMRI as an imaging modality, and the associated strength of computational framework for its analysis, fMRI is still constrained by its scanner environment and huge magnets. This essentially means that young infants, or people with special needs e.g. those with movement disorders, magnetic implants and/or psychological challenges such as claustrophobia (anxiety disorder for confined space) cannot be studied using fMRI. In comparison, another non-invasive brain imagining modality, functional Near-InfraRed Spectroscopy (fNIRS) offers both portability, and some degree of flexibility for movement whilst brain activity is recorded, rendering it as a far more favorable imaging modality for many key challenging populations deemed inaccessible by fMRI for cognitive neuroscience studies [2] [3] [4].

Although, fNIRS can be employed successfully to study brain activity across a whole spectrum of populations and tasks, it is marred by the lack of sophisticated algorithms that can fully decode and explain the underlying brain activity. The limitations of fNIRS data analysis arise from a lack of anatomical image, and transparent classification models. The state-of-the-art supervised learning methods such as Support Vector Machine (SVM) or Random Forest (RF) are not interpretable models for classification of fNIRS data [5].

The significance of an interpretable model is that the relationship between a data input and its corresponding consequent (or condition/stimuli for fNIRS) is retained. In the context of fNIRS data, a classification result characterized as corresponding to stimulus A would mean that activity in a channel located in a certain anatomical brain location is FLV (where FLV is a Fuzzy Linguistic Variable that can be either low, medium or high), hence the data instance is classified as evoked/caused by stimulus A.

Consequently, to incorporate a transparent classification method for fNIRS dataset, this work presents fNIRS data classification enabled by a fuzzy system. A fuzzy system classifier is inherently explainable because of the *fuzzy rules* that govern the classification such as: IF *causes* THEN *result* is a general description of a fuzzy rule. An example of a fuzzy rule for the particular case of fNIRS dataset can be: IF *activity is FLV in channel Y* THEN *it corresponds to stimulus A*. This can explain why a particular data sample is classified as *A*. The fuzzy rules can either be provided by the experts in the domain, or can also be learnt from the data directly [6].

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

The ability of fuzzy systems to model the high imprecision, and vagueness associated with neuroimaging contexts renders them as an optimal alternative framework for effective and expressive data analysis [7]. Fuzzy systems have been used in the past to cope with neural models of high-uncertainly [8] and also for data analysis of other neuroimaging modalities such as electroencephalogram [9].

The major contribution of this work lies in using Explainable Artificial Intelligence (XAI) system to drive MVPA on neuroscience data. Here, Interval Type-2 Fuzzy Logic System (IT2FLS) is employed to power MVPA on an infant's fNIRS dataset. The paper is divided as follows: Section II presents an overview on fNIRS technology, and the basis of MVPA. Section III outlines a brief introduction to type-2 FLS used in conjunction with GA to optimize a rulebase from a given dataset. Section IV presents the experiment, and the results obtained from the proposed XAI system (i.e. IT2FLS-GA) for MVPA of infant's fNIRS data with neurophysiological discussion in Section V. Section VI delineates the conclusions, and future work.

II. An Overview on fNIRS and MVPA

A. *functional Near-InfraRed Spectroscopy (fNIRS)*

fNIRS is an optical neuroimaging modality that reads cerebral activity using Near-InfraRed (NIR) light. An fNIRS emitter shines NIR light at a specific location of interest on the surface of the head. The emitted NIR light then undergoes physiological processes like absorption, and scattering as it propagates through the brain tissue back to the surface of the head where it is recorded by a fNIRS detector. Fig 1 shows a participant wearing an fNIRS cap whilst his brain activity is recorded.

An fNIRS signal measures brain activity by recording changes in concentration of oxygen in hemoglobin molecules in the blood. The NIR light of wavelength 770-850 nm remains unaffected on encounter with skin, tissue, or bone however the hemoglobin molecules in the blood absorb NIR light significantly. Hence an increase in oxygen demand in the brain, which implies an increase in the brain activity, can be read by an fNIRS signal by measuring relative changes in hemoglobin concentration based on NIR light absorption by the hemoglobin molecules in the blood.

Fig. 1: A participant wearing an fNIRS cap during an experiment to record their brain activity.

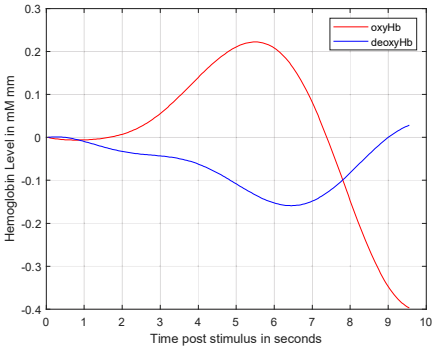


Fig. 2: A typical time-series of a bidimensional fNIRS signal post stimulus presentation.

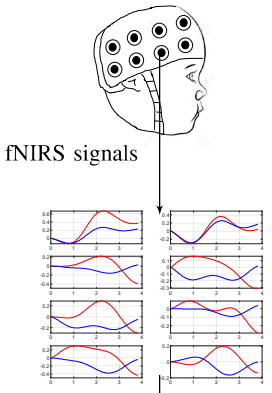
The hemoglobin molecules are characterized as oxygenated hemoglobin (oxyHb) and deoxygenated hemoglobin (deoxyHb) molecules based on their oxygen content. The relative concentrations of oxyHb and deoxyHb molecules in the blood as a function of total photon path length are calculated using the modified Beer-Lambert law (mBLL) [10]. Fig. 2 shows a characteristic fNIRS signal composed of oxyHb and deoxyHb concentration post a stimulus presentation.

As can be readily appreciated from Fig. 2, the changes in concentration of oxyHb and deoxyHb vary significantly with time. Hence accurate data analysis of any fNIRS based study is hinged on the identification of time window parameters (that is the start time, and the duration of the time window) that would allow for the best classification of the fNIRS signals for that particular study. The fNIRS signal is also dependent on the age of the participants, and the particular stimuli/task they are performing [11][12].

A guideline from previous studies of infants fNIRS data using simple audio and visual stimuli indicates that time window from 5-9 seconds is optimal for detecting hemodynamic response in infants under a year old [13]. It is also acceptable, given that the hemodynamic response in infants doesn't always follow a canonical shape, in developmental fNIRS studies to analyze the oxyHb rather than deoxyHb or a combination of the two such as total hemoglobin [14]. But it remains an open research question as to which hemodynamic dimension best represents the underlying neural activity [15] [16] [17].

B. *Multivariate Pattern Analysis (MVPA)*

In general, a MVPA is more accurate, and sensitive to cognitive states than a corresponding univariate pattern analysis since a greater number of features (or dimensions) are considered simultaneously in a MVPA [18]. A higher dimensional MVPA extends sophisticated insights into key challenges on reading the human brain such as how is



Trial No	Ch1	Ch2	Ch3	Ch4	Ch5	Ch6	Ch7	Ch8	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 MVP matrix.

Fig. 3: A MVP matrix with hypothetical data for 8 fNIRS channels associated with two stimulus conditions for ten trials.

the information represented and processed in different brain structures which were previously out of realm for a univariate analysis.

A MVP from neuroimaging data is formed by cascading data from different brain locations to be considered simultaneously in the subsequent analysis. A schematic for the MVP matrix formation from an fNIRS study is shown in Fig. 3. In Fig. 3, an infant's brain activity is recorded using 8 fNIRS channels. The activity in the channels is then encoded in a MVP matrix appended with the associated stimuli. The exact value in the MVP matrix can be any associated feature, or a combination of features for different channels, for the signal in a certain time period e.g. mean of the signal post stimulus for a certain time length on a per subject or a per trial basis.

Earlier MVPA based fMRI studies focused on establishing functional relationships between brain regions, but with an increase in the understanding of the human brain, the focus of the neuroscience studies have shifted towards pattern classification [19]. There are numerous machine learning methods which can conduct MVPA for a given fMRI dataset. For example, an auditory fMRI study used SVM based MVPA to gain insight into the role of superior temporal gyrus [20], correlation based classifier were used to investigate how visual objects are represented in temporal cortex [21], whereas memory search in humans has been investigated using neural networks [22], and linear discriminant analysis to predict the orientation of invisible stimuli from activity in visual cortex [23].

For fNIRS studies, MVPA is mostly used in conjunction with SVM, for example SVM has been used for classifying personal preference single-trial fNIRS, task engagement using attentional state [24]. Inherently, a SVM is a supervised machine learning model that aims to find a hyper plane that can distinguish between a set of classes using labeled input training data instances [19]. A more recent study by Emberson *et al.* has used Pearson correlation coefficients on infants fNIRS data [25] to explore the similarities in neural responses to similar stimuli.

III. The Proposed Type-2 Fuzzy Logic Based Explainable Artificial Intelligence (XAI) System for Developmental Neuroscience

The choice of which type of FLS to model and analyse crisp input data largely depends on the amount of uncertainty in the input data. Also, in the event an input dataset is also used to train a supervised learning model, for example a rule-based FLS, it is imperative to account for the uncertainty in the input data which can be achieved by using type-2 FLS. Type-2 FLSs have previously been used in numerous applications such as data preprocessing [26], controllers for mobile robots [27], and time series forecasting [28]. A special case of a type-2 FLS is an IT2FLS in which the third dimension (of uncertainty) values are always equal to 1 [29]. In this work, IT2 based XAI system have been used to perform MVPA on infant's fNIRS data.

In mathematical notation, interval type-2 fuzzy set ($\tilde{A}$) can be written as follows in (1) [30].

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

where $\mu_{\tilde{A}}$ represent the membership function of interval type-2 fuzzy set $\tilde{A}$.

In the illustrative interval type-2 fuzzy sets shown in Fig. 4, the membership value for temperature of 12 °C falls in the interval type-2 fuzzy sets of $\tilde{Low}$ with lower and upper membership function values as: $\underline{\mu}_{\tilde{\text{Low}}}(12) = 0$ and $\overline{\mu}_{\tilde{\text{Low}}}(12) = 1$, whereas the membership value for temperature of 12 °C falls in the interval type-2 fuzzy sets of $\tilde{Medium}$ with lower and upper membership function values as: $\underline{\mu}_{\tilde{\text{Medium}}}(12) = 0$ and $\overline{\mu}_{\tilde{\text{Medium}}}(12) = 0.4$.

The following subsections delineate the major components of the proposed type-2 fuzzy logic based XAI system for developmental neuroscience.

A. Type-2 Fuzzy Rule Inference

The rules for a FLS, that map fuzzy input sets to fuzzy output sets using fuzzy inference engine, are defined as follows:

$$\begin{aligned} &Rule\ R_q : IF\ x_1\ is\ A_{q1}\ AND\ ...\ AND\ x_n\ is\ A_{qn} \\ &THEN\ Class\ C_{j_q}\ with\ RW_q \end{aligned} \tag{2}$$

where q is the rule number, x_i is the crisp input for variable i, A_i is the input fuzzy set (also called antecedent (ANT)) for the ith variable, n is the total number of variables or dimensions of the input dataset, C_{j_q} is the consequent class from a predefined set $\{C_1, ..., C_K\}$ of classes, and RW_q is the rule weight associated with the qth rule.

In a type-2 FLS, each rule will have a a lower rule weight, $\underline{RW}_q$, and an upper rule weight, $\overline{RW}_q$, as defined in (3) [31].

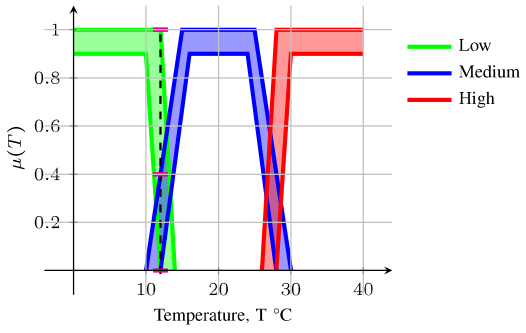
$$\begin{aligned} \overline{RW}_q &= \overline{c}_q^s \cdot \overline{s}_q^s \\ \underline{RW}_q &= \underline{c}_q^s \cdot \underline{s}_q^s \end{aligned} \tag{3}$$

where q is the rule number, c_q^s is the scaled confidence (4), and s_q^s is the scaled support (7) of the rule R_q on a training dataset.

The upper and lower confidence of a rule, R_q, is computed on a training dataset as outline in (4) [31]. It can be viewed as a conditional probability that given a particular data instance has matching antecedents as the rule R_q then what is the likelihood for that data instance to have the same consequent class as the consequent C_{j_q} of the rule R_q.

$$\begin{aligned} \overline{c}_q^s(A_q \Rightarrow C_{j_q}) &= \frac{\sum_{x_t \epsilon (A_q \Rightarrow C_{j_q})} \overline{w}_q^s(x_t)}{\sum_{q=1, x_t \epsilon A_q}^{Q} \overline{w}_q^s(x_t)} \\ \underline{c}_q^s(A_q \Rightarrow C_{j_q}) &= \frac{\sum_{x_t \epsilon (A_q \Rightarrow C_{j_q})} \underline{w}_q^s(x_t)}{\sum_{q=1, x_t \epsilon A_q}^{Q} \underline{w}_q^s(x_t)} \end{aligned} \tag{4}$$

where $\overline{w}_q^s(x_t)$ and $\underline{w}_q^s(x_t)$ are the scaled upper and lower strengths of activation for rule R_q on a data instance x_t in a training dataset.



Interval Type-2 Fuzzy Sets

Fig. 4: An illustrative plot of membership function values for Temperature, $\mu(T)$, using interval type-2 fuzzy sets named: *Low*, *Medium*, and *High*.

The strength of activation of rule R_q for a data instance x_t, $w_q(x_t)$, is a measure of the degree of match between the rule and the data instance. It is computed as outlined in (5).

$$\begin{aligned} \overline{w}_q(x_t) &= \prod_{i=1}^{N} \overline{\mu}_{\tilde{A}_i}(x_{t,i}) \\ \underline{w}_q(x_t) &= \prod_{i=1}^{N} \underline{\mu}_{\tilde{A}_i}(x_{t,i}) \end{aligned} \tag{5}$$

The strength of activations are scaled to ensure any bias incorporated due to non-equal size of the consequent classes C_j can be accounted by computing scaled strength of activations as outlined in (6). The strength of activation of a given rule R_q for a data instance x_t is scaled by the sum of activations of all the rules R_l, which have the same consequent class as the rule R_q, for the data instance x_t.

$$\begin{aligned} \overline{w}_q^s(x_t) &= \frac{\overline{w}_q(x_t)}{\sum_{l, R \epsilon C_{j_q}} \overline{w}_l(x_t)} \\ \underline{w}_q^s(x_t) &= \frac{\underline{w}_q(x_t)}{\sum_{l, R \epsilon C_{j_q}} \underline{w}_l(x_t)} \end{aligned} \tag{6}$$

The support of a rule is an indication of the coverage of training dataset by the rule. It is computed using (7).

$$\begin{aligned} \overline{s}_q^s(A_q \Rightarrow C_{j_q}) &= \frac{\sum_{x_t \epsilon (A_q \Rightarrow C_{j_q})} \overline{w}_q^s(x_t)}{Q} \\ \underline{s}_q^s(A_q \Rightarrow C_{j_q}) &= \frac{\sum_{x_t \epsilon (A_q \Rightarrow C_{j_q})} \underline{w}_q^s(x_t)}{Q} \end{aligned} \tag{7}$$

After establishing a lower and an upper rule weight, $\underline{RW}_q$, $\overline{RW}_q$, of each rule on a training dataset, the classification accuracy of a given rulebase is determined on the testing dataset. The consequent class, C_j, for each data instance in the testing dataset, x_s, is determined using the metric association degree. The association degree, h_q, of each rule with each data instance in the testing dataset, x_s, is computed as outlined in (8).

$$\begin{aligned} \overline{h}_q(x_s) &= \overline{w}_q^s(x_s) \cdot \overline{RW}_q \\ \underline{h}_q(x_s) &= \underline{w}_q^s(x_s) \cdot \underline{RW}_q \\ h_q(x_s) &= \frac{\overline{h}_q(x_s) + \underline{h}_q(x_s)}{2} \end{aligned} \tag{8}$$

A given testing data instance, x_s, is classified into the consequent class, C_j, corresponding to the rule with the maximum association degree with x_s.

B. Evolutionary Type-2 Fuzzy Rule Learning

The rules for a FLS can be furnished by experts in the field, or can also be learnt from the input dataset using an optimization algorithm such as Genetic Algorithm (GA) [6] [32]. The advantage of using an optimizer to learn fuzzy rules from input data in comparison to the rules been provided by experts in the field is that in the case of former the rulebase

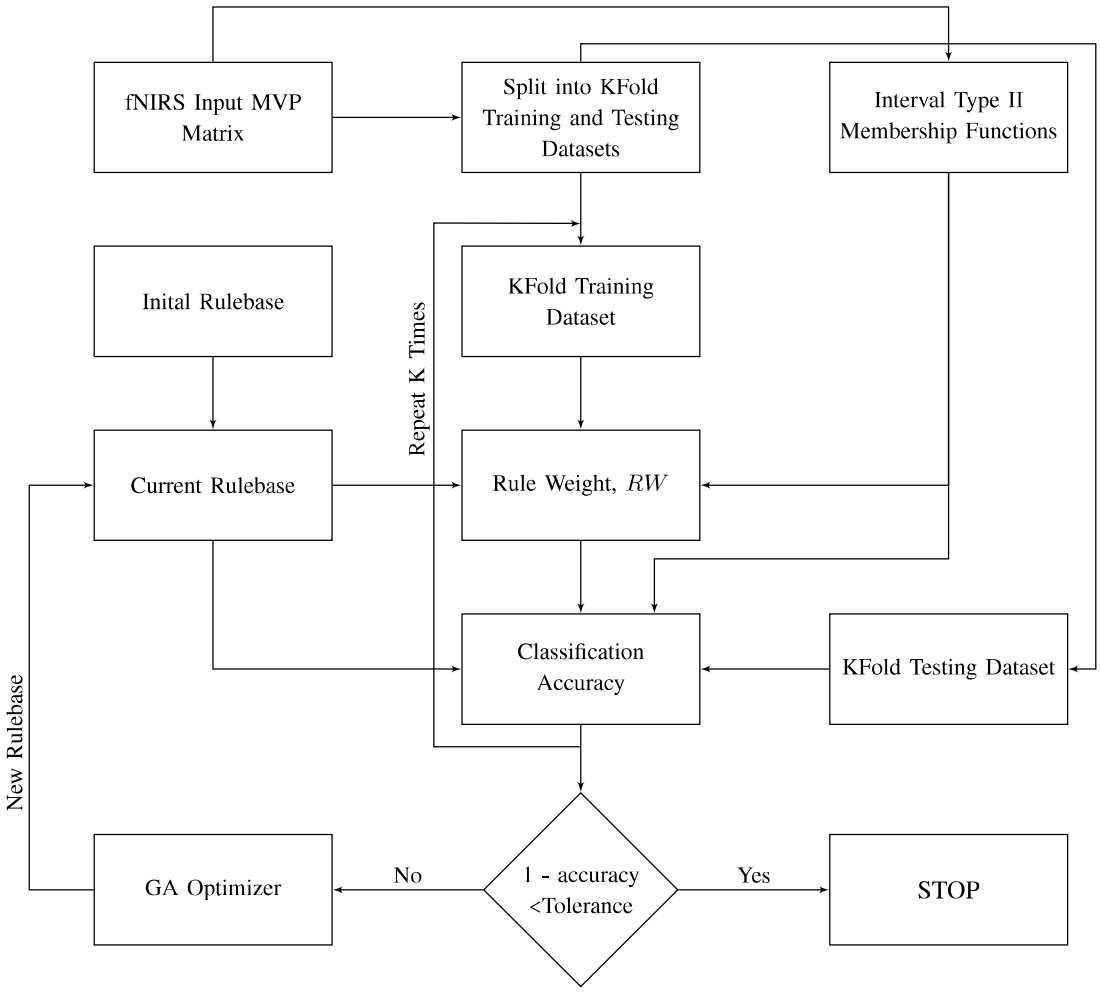


Fig. 5: A flowchart of supervised learning of a rule-based IT2 based XAI system using input data and GA.

is bias-free. Consequently, GA has been used in this work to learn rules directly from infant's fNIRS data to circumvent probable expert's prejudices.

Fig. 5 outlines the steps undertaken to arrive at an optimized rulebase using a given dataset. The input dataset is split into k-fold cross validation training and testing datasets to offset any bias incorporated from a particular selection of training and testing dataset [33]. All rulebases are learnt from k-fold training datasets with k-fold testing datasets used to establish the efficacy of a given rulebase.

Using an initial random rulebase with a total of Q rules, the classification accuracy of the rulebase is computed as the ratio of the correct predicted labels to the total number of data instances in the test dataset. The cost of the rulebase is computed as 1 - the mean of the classification accuracy on all k-fold testing datasets. The GA then compares the cost of the rulebase with a pre-defined tolerance criterion. If the cost is greater than the tolerance of GA, the GA then populates a new rulebase and the cycle is repeated till the tolerance criterion of the GA is met as outlined in Fig. (5).

In this work, the total number of rules to be learnt by XAI system is set at 20 rules, with maximum of 3 ANTs in a given rule. The low number of rules with maximum of 3 ANTs in any rule is set to ensure XAI system generate fewer rules with rational rule interpretability. This renders the total number of variables to be learnt by GA to be: total number of rules (20) * maximum number of antecedents (3) and FLV for each chosen antecedent (3: Low, Medium or High) and the corresponding consequent class, C_j, for each rule (1) = 20*(3+3+1) = 140 variables.

The structure of each phenotype is delineated in (9). 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 $1 * 10^{-5}$.

$$P^b = \{\phi_1^1, \phi_2^1, \phi_3^1, \lambda_1^1, \lambda_2^1, \lambda_3^1, \gamma_j^1, ..., \phi_1^Q, \phi_2^Q, \phi_3^Q, \lambda_1^Q, \lambda_2^Q, \lambda_3^Q, \gamma_j^Q\} \tag{9}$$

where P^b is the phenotype of an individual for the GA. Each ϕ denotes a particular input (channel), each λ represents the corresponding FLV for each input. These chromosomes form

the antecedent of a rule. The consequent of this rule is denoted as γ.

IV. Experiments and Results

A. fNIRS Data

In this work, infant's fNIRS datasets made available online by Emberson *et al.* [25] is analysed for obtaining an insight into the workings of the developing human brain. The infant's fNIRS dataset is unisensory i.e. infants either watched a red smiley face (visual stimulus) for 1 second or heard a toy sound (audio stimulus) for 1 second. The reader is referred to the original study by Emberson *et al.* [25] for more details on data collection, and the subsequent preprocessing stages.

The fNIRS data is obtained from a total of 10 fNIRS channels (Ch) with corresponding anatomical locations as shown in Fig. 6 [25]. The Ch1- Ch3 are located in the occipital cortex, which is associated with visual processing; Ch4 - Ch8 are in the temporal cortex which is associated with auditory processing; and Ch9 - Ch10 are in the prefrontal cortex which is mainly responsible for planning, and attention [34].

B. XAI System based MVPA

The mean of fNIRS signals from time 4-7 seconds post stimulus, from the ten fNIRS channels illustrated in Fig. 6, is used to build the MVP matrix as shown in Fig. 3. The MVP matrix is split into 6-fold cross-validated training and testing datasets, and input into IT2 based XAI system, as outlined in Fig. 5. The fuzzy rules are learnt from each training fold separately using the proposed XAI system, and the accuracy established on testing folds. Please note there is no information flow from one fold to another i.e. the proposed XAI system has no memory of the choices made and the results obtained from any given fold.

The mean classification metrics are obtained by gauging the final optimized rules obtained from XAI system on all 6 testing datasets, and reported in Table I. The classification metrics obtained from the proposed XAI system are also compared with Adaptive-Network-Based Fuzzy Inference System (ANFIS), SVM, and RF, as reported in Table I.

ANFIS uses a hybrid learning procedure by integrating both neural networks and fuzzy logic principles [35]. In this work, ANFIS is implemented using Mathworks central file exchange [36].

The main advantage of the proposed XAI system is that it provides fuzzy rules in the context of brain data whilst also offering comparative classification metrics with respect to other state-of-the-art classifiers as reported in Table I.

C. Explainable Fuzzy Rules

A total of 20 fuzzy rules with at most 3 ANTs are obtained form the proposed IT2 based XAI system. An example of a rule for both stimuli - Video and Audio, are listed in (10). As can be readily interpreted from the rules in (10), the proposed IT2 based XAI system offers a unique perspective into the workings and interconnections of the developing human brain i.e. if a data instance is classified as Video based on Rule 1 then it implies for that particular data instance the value of Ch3 is High and Ch6 is Low, and Ch9 is Low.

$$\begin{aligned} &\text{Rule } \mathrm{R}_1 \text{ :IF } Ch3 \; is \; High \text{ AND } Ch6 \; is \; Low \text{ AND} \\ &\qquad Ch9 \; is \; Low \text{ THEN } stimulus \; is \; Video \\ &\text{Rule } \mathrm{R}_2 \text{ :IF } Ch1 \; is \; High \text{ AND } Ch3 \; is \; Low \text{ AND} \\ &\qquad Ch5 \; is \; Medium \text{ THEN } stimulus \; is \; Audio \end{aligned} \tag{10}$$

D. Neuropshycological Discussion

An interpretable model for fNIRS data is of paramount importance since the relation between the data and the corresponding stimuli would enable researchers to gain insight into the elusive workings of the human brain. For example, from the rules listed in (10), it can be understood how the proposed XAI system is discriminating between data samples evoked from Video and Audio stimulus.

The rules obtained from the proposed XAI system are also in-line with the established literature. The results reported in the original study by Emberson *et al.* [25] conclude that Ch3 is the most important channel for decoding accuracy. However, the analysis could not provide further insight as to *why* Ch3 is the most significant channel. Whereas the rules obtained from the proposed XAI system clearly state that Ch3 is important because it appears as an ANT in the rules for both Video and Audio. The rules from the proposed XAI system also signify what value (i.e. FLV) of Ch3 is associated with other Chs (i.e. Ch6 and Ch9 for R_1 and Ch1 and Ch5 for R_2 in (10)).

The true potential of explainable fNIRS data analysis lies in empowering scientists to study different trajectories of brain development. This can essentially unlock key similarities and differences between typical and atypical brain development trajectories. Consequently, there is potential for further investigations that could shed light into early diagnosis and intervention.

V. Conclusions and Future Work

The motivation for this empirical study lies in harnessing explainable classification results for infant's brain data i.e. mapping the output (stimulus) back to the characteristics of the input (brain) data, which is of paramount significance in understanding the developing human brain.

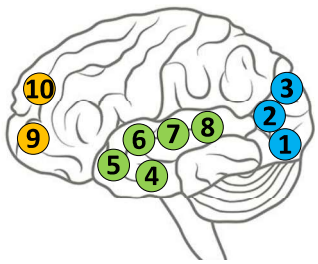


Fig. 6: The anatomical locations of the ten channels (Ch) are: Ch1 - Ch3 (blue): occipital cortex, Ch4 - Ch8 (green): temporal cortex, and Ch9 - Ch10 (orange): prefrontal cortex.

TABLE I: Classification metrics on mean infant fNIRS MVPA for time window 4-7s post stimulus driven by the proposed IT2 based XAI system, ANFIS, RF, and SVM.

Metric	**XAI system**	**ANFIS**	**RF**	**SVM**
Kappa	0.59 ± 0.04	0.36 ± 0.17	0.31 ± 0.092	0.41± 0.072
Accuracy	61.15 ± 4.09%	61.25± 5.0%	65.19 ± 4.56%	70.57 ± 3.70%
Precision	88.40 ± 4.07%	70.18 ± 6.62%	65.99 ± 4.96%	70.29± 4.90%

In this work, an explainable fNIRS data analysis is undertaken using an IT2 based XAI system in tandem with GA. The reason for choosing a type-2 FLS lies in its capability of modeling the uncertainty in the input data which is of paramount importance in this work since the input data is also being used to learn the fuzzy rules. The explainable rules offered by the proposed XAI system give a unique insight into how developing brains function in response to different stimuli.

The future work aims at validating the rules obtained from the proposed XAI system by conducting a new study. A special emphasis will be placed in the new study on the anatomical location of the fNIRS channels, especially in terms of co-registration between infants, to ensure error in identification of anatomical locations arising from the curvature of the infant's heads can be minimized.

REFERENCES

[1] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 3863–3868, 2006.

[2] 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," *Hamlyn Symposium*, 2014.

[3] 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 (In Press)*, 2019.

[4] 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.

[5] 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.

[6] N. N. Karnik, J. M. Mendel, and Q. Liang, "Type-2 Fuzzy Logic Systems," *IEEE Transactions on Fuzzy Systems*, vol. 7, pp. 643–658, 1999.

[7] J. Andreu-Perez, "Fuzzy learning and its applications in neural-engineering," *Neurocomputing*, vol. 389, pp. 196–197, 2020.

[8] J. Andreu-Perez, F. Cao, H. Hagras, , and G.-Z. Yang, "Self-adaptive online brain–machine interface of a humanoid robot through a general type-2 fuzzy inference system." *IEEE Transactions on Fuzzy Systems*, vol. 26, pp. 101–116, 2016.

[9] D. Achanccaray, C. Flores, C. Fonseca, and J. Andreu-Perez, "A Fuzzy Genetic Algorithm for Optimal Spatial Filter Selection for P300-Based Brain Computer Interfaces," *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018.

[10] 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, p. 4053–4075, 2014.

[11] C. Issard and J. Gervain, "Variability of the hemodynamic response in infants: Influence of experimental design and stimulus complexity," *Developmental cognitive neuroscience*, vol. 33, pp. 182–193, 2018.

[12] S. Lloyd-Fox, A. Blasi, and C. E. Elwell, "Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy," *Neuroscience and Biobehavioral Reviews*, vol. 34, pp. 269–284, 2010.

[13] 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, no. 4, pp. 1246–1252, 2007.

[14] J. Gervain, J. Mehler, J. F. Werker, C. A. Nelson, G. Csibra, S. Lloyd-Fox, M. Shukla, and R. N. Aslin, "Near-infrared spectroscopy: a report from the McDonnell infant methodology consortium," *Dev Cogn Neurosci.*, vol. 1, no. 1, pp. 22–46, 2011.

[15] 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.

[16] 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.

[17] 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.

[18] M. Hanke, Y. O. Halchenko, P. B. Sederberg, S. J. Hanson, J. V. Haxby, and S. Pollmann, "Pymvpa: a python toolbox for multivariate pattern analysis of fmri data," *Neuroinformatics*, vol. 7, pp. 37–53, 2009.

[19] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fmri data," *Trends in Cognitive Sciences*, vol. 10, pp. 424–430, 2006.

[20] F. D. Martino, G. Valente, N. Staeren, JohnAshburner, R. Goebel, and E. Formisano, "Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns," *Neuroimage*, vol. 43, pp. 44–58, 2008.

[21] M. Spiridon and N. Kanwisher, "How distributed is visual category information in human occipito-temporal cortex? an fmri study," *Neuron*, vol. 35, pp. 1157–1165, 2002.

[22] S. M. Polyn, V. S. Natu, J. D. Cohen, and K. A. Norman, "Category-specific cortical activity precedes retrieval during memory search," *Science*, vol. 310, pp. 1963–1966, 2005.

[23] J.-D. Haynes and G. Rees, "Predicting the orientation of invisible stimuli from activity in human primary visual cortex," *Nature Neuroscience*, vol. 8, p. 686–691, 2005.

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

[25] 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, 2017.

[26] R. John, P. Innocent, and M. Barnes, "Type 2 fuzzy sets and neuro-fuzzy clustering of radiographic tibia images," *Proceedings of 6th International Fuzzy Systems Conference*, 1997.

[27] K. C. Wu, "Fuzzy interval control of mobile robots," *Computers and Electrical Engineering*, vol. 22, 1996.

[28] N. N. Karnik and J. M. Mendel, "Applications of Type-2 Fuzzy Logic Systems to Forecasting of Time-Series," *Information Sciences*, vol. 120, pp. 89–111, 1999.

[29] Q. Liang and J. M. Mendel, "Interval type-2 fuzzy logic systems: theory and design," *IEEE Transactions on Fuzzy Systems*, vol. 8, pp. 535–550, 2000.

[30] J. R. Castro, O. Castillo, and P. Melin, "An Interval Type-2 Fuzzy Logic Toolbox for Control Applications," *IEEE International Fuzzy Systems Conference*, 2007.

[31] 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, 2016.
[32] J. R. Koza, "Genetic programming," 1997.
[33] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *International Joint Conference on Arti cial Intelligence*, 1995.
[34] M. I. Posner, S. E. Petersen, P. Fox, and M. E. Raichle, "Localization of Cognitive Operations in the Human Brain," *Science*, vol. 240, pp. 1627–1631, 1998.
[35] J.-S. R. Jang, "ANFIS: Adaptive-Network-based Fuzzy Inference System," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 23, pp. 665–685, 1993.
[36] S. A. Manthiri. MATLAB Central File Exchange. Accessed: 2020-01-30. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/62245-anfis-classifier