Binary Classification Using Neural and Clinical Features: An Application in Fibromyalgia With Likelihood-Based Decision Level Fusion

Didem Gökçay, Aykut Eken, and Serdar Baltacı

***Abstract*—Among several features used for clinical binary classification, behavioral performance, questionnaire scores, test results, and physical exam reports can be counted. Attempts to include neuroimaging findings to support clinical diagnosis are scarce due to difficulties in collecting such data, as well as problems in integration of neuroimaging findings with other features. The binary classification method proposed here aims to merge small samples from multiple sites so that a large cohort, which better describes the features of the disease can be built. We implemented a simple and robust framework for detection of fibromyalgia, using likelihood during decision level fusion. This framework supports sharing of classifier applications across clinical sites and arrives at a decision by fusing results from multiple classifiers. If there are missing opinions from some classifiers due to inability to collect their input features, such degradation in information is tolerated. We implemented this method using functional near infrared spectroscopy (fNIRS) data collected from fibromyalgia patients across three different tasks. Functional connectivity maps are derived from these tasks as features. In addition, self-reported clinical features are also used. Five classifiers are trained using k nearest neighborhood (kNN), linear discriminant analysis (LDA), and support vector machine (SVM). Fusion of classification opinions from multiple classifiers based on likelihood ratios outperformed individual classifier performances. When 2, 3, 4, and 5 different classifiers are fused, sensitivity, and specificity figures of 100% could be obtained based on the choice of the classifier set.**

***Index Terms*—Clinical binary classification, decision level fusion, fibromyalgia, functional connectivity, functional near infrared spectroscopy (fNIRS), likelihood.**

I. Introduction

BINARY classification for detection or rejection of a specific ailment is a decision making procedure used widely in the clinic. Physicians make binary decisions based on several graphical and numeric presentation of results obtained from a multitude of data types such as behavioral performance, questionnaires completed by the patients, physical exam of the physician, physiological tests, laboratory tests, and brain activity profiles. However, utilization of all of these types of data is not possible because of limited time and resources to collect the data, as well as economical aspects of patient treatment. Hence, physicians must rely on diagnostic aids that tolerate missing data. Unfortunately, it is hard to derive a generalized approach from several suggested methods in [1], [2] due to the complexity of the clinical data which participate heterogeneously in binary decision.

Although self-report information, obtained from individuals, shows a higher accuracy while distinguishing patient and healthy participants, it is highly subjective. Hence an objective measure to solve this problem is strongly needed. Most recently, neural activity patterns are being tried at the clinic utilizing functional near infrared spectroscopy (fNIRS) [3], [4] and fMRI [5]–[7] data. Use of not only task-based, but also resting state EEG [8] and fMRI [9], [10] diagnostic applications are also acknowledged as hot-topics. However, there are important problems and barriers that need to be faced by classifiers that use neuroimaging data obtained at the clinic. First, due to restrictions such as high intolerance to head motion, possibility of large false positive activations reduce the specificity. Second, time limitations may prohibit data collection, since neuroimaging experiments take at least half an hour. Third, published accuracy of the classifiers may not be trusted, because biomarker studies in neuroimaging use only tens of subjects, not even hundreds [11]. So, it is a necessity to aggregate clinical data in an efficient fashion to obtain higher detection rates [11].

While fusion of classifiers is a topic studied since 1998 [12], use of a widely known classifier, the maximum likelihood classifier, has been reported to improve clinical binary decisions based on complex [13] as well as simple and efficient parameter settings [14] more convincingly. Quite recently, there have been attempts to fuse classifiers based on such criteria using EEG [15] and fNIRS [16] data. In these applications, classifier fusion increased the level of success in diagnosing stress and cognitive performance relatively by 10%.

In this study, we propose to augment the methodology of [13], [15], [16] by using a simple update scheme which is easy to implement at the clinic. With this scheme, it becomes possible to rearrange the individual performances of the fused classifiers before retraining for new cases, because fusion is not performed at the feature level but at the classifier level. We illustrate the

Manuscript received January 23, 2018; revised April 15, 2018 and May 26, 2018; accepted May 26, 2018. Date of publication September 28, 2018; date of current version July 1, 2019. *(Corresponding author: Didem Gökçay.)*

D. Gökçay and S. Baltacı are with the METU Informatics Institute, Ankara 06800, Turkey (e-mail: dgokcay@metu.edu.tr; serdar.baltaci@metu.edu.tr).

A. Eken is with the Düzce University, Düzce 81620, Turkey (e-mail: ekenaykut@gmail.com).

Digital Object Identifier 10.1109/JBHI.2018.2844300

proposed method with a proof of concept, in an application in fibromyalgia which is a chronic syndrome of high sensitivity to pain with a prevalence of 2–8% among populations [17].

FM syndrome is generally diagnosed according to the American College of Rheumatology (ACR) 1990 [18] or 2010 [19] criteria in clinics. These criteria mainly depend on tender point count and duration of widespread chronic pain. In addition to these parameters, Fibromyalgia Impact Questionnaire (FIQ), depression tests (Beck Depression Inventory (BDI), Hospital Anxiety and Depression Scale (HADS) etc.), pain threshold measurements and pain scores are also used as supportive information for diagnosis of FM. Automatic diagnostic classification of FM via machine learning methods by using clinical measures is possible [20]. Classification of 53 FM patients and 74 romatoid arthiris (RA) patients was performed by using AD-ABoost classifier utilizing several clinical measures such as Global Severity Index (GSI), Positive Symptom Distress Index (PSDI) and Positive Symptom Total (PST) and medico-social features such as age, Visual Analog Scale, occupation, years in school, years with disease, average income per month and their medication. Results reported a success rate of 96% for clinical measures, 89% for medico-social characteristics and 97% for combination of clinical and medico-social features.

However, underlying abnormalities in central nervous system for FM syndrome is still unknown. To discover abnormalities that cause FM, several neuroimaging studies focused on finding neuromarkers specific to FM (see reviews [21]–[23]). There are only a handful reports regarding classification of FM from neuroimaging data. While results of these are in the range of 76–96% accuracy, some of these methods employed feature level fusion, none employed decision level fusion. In this study, we developed a decision level fusion approach to combine clinical and neuroimaging data which provides flexibility of use at the clinic. Furthermore the proposed method in the following sections is generalizable to any clinical application which contains a heterogeneous set of features that employ clinical and neuroimaging data. Last but not least, the proposed framework allows merging data from remote sites, so that sample sizes of classifiers can be increased.

II. Method

On a group of fibromyalgia (FM) patients and healthy controls (HC), three different experiments are conducted in a clinical neuroimaging lab [24], [25]. In addition, patients filled questionnaires. Clinical measurements regarding pain thresholds are collected just before the start of the neuroimaging data acquisition. Then several classifiers are trained with extracted features and a quality check is done. Among all of the classifiers tested, 5 are chosen for fusion, to improve the performance of binary detection. This provided adequate materials for us to carry on with the proof of concept for decision level fusion. The framework for binary classification based on decision level fusion with maximum likelihood is introduced in Fig. 1.

In a more general setting, for binary classification of other clinical ailments, data collection can include a wider range of data as follows. Behavioral performance could be collected through neuropsychological tests (for ex. Stroop). Self-report

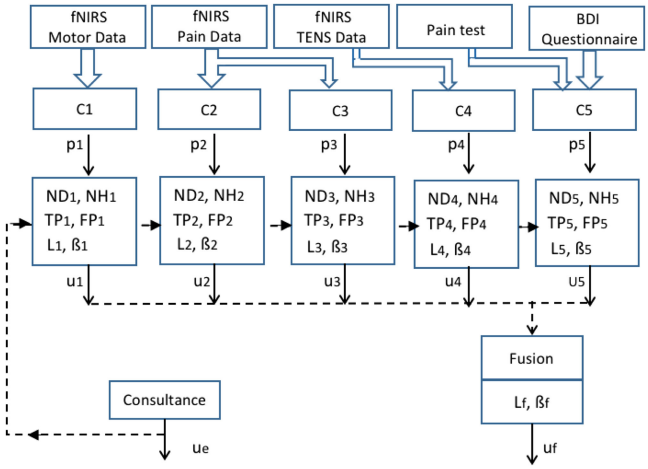


Fig. 1. Binary classification based on decision level fusion with likelihood.

based data could be collected when the patients fill standard questionnaires (for ex. BDI). Verbal report could be obtained during physical exam from the records of the physician. Physiological data could be acquired as time series using sensors that measure heart rate, skin conductance, EMG or skin temperature during an ongoing task such as a stress test. Brain activity could be acquired as time series through either resting state or task-based fMRI, EEG, MEG. All of these data can be added to the proposed framework for detection of FM, using new classifiers that operate on features obtained from the behavioral performance such as reaction time, accuracy or features from the self-reports in terms of survey scores or features obtained from the physician's report consisting of pulse, weight and other physical exam details or features obtained from physiological/brain activity based on normalized t values or p values after data analysis, and absolute peak values or onset latencies of the signals of each trial, and mean signal values in each trial.

A. Data Collection

16 healthy controls (HC) and 19 fibromyalgia (FM) patients were admitted. All participants were required to fill Beck Depression Inventory (BDI) questionnaire. Pain threshold values for both thumbs were collected by using electronic Von Frey (eVF) anesthesiometer (Ugo Basile Co., Italy). More details regarding participants are provided in [25]. The first experiment involved motor data collection in a self-paced finger tapping task (FTT). The second experiment involved painful stimulation with and without TENS device (PAIN). The third experiment, was median nerve stimulation using TENS device (MNS). For all of these experiments, fNIRS data was collected using Hitachi ETG-4000 continuous wave functional near infrared spectroscopy system. Data was collected from both hands in separate episodes. More details are provided in [24], [25].

B. Feature Extraction

After obtaining channel positions and corresponding cortical regions, data was preprocessed to remove artifacts [24], [25]. In

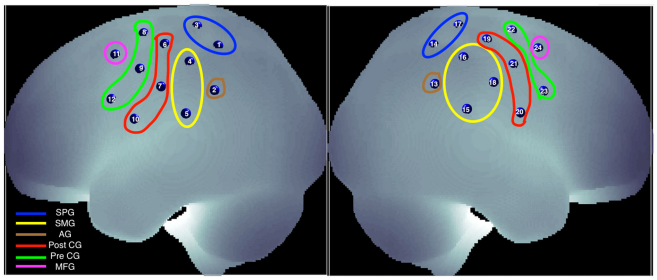


Fig. 2. Regions of Interest (ROI) of the connectivity features.

order to generate features based on different regions of interest (ROI), time series data from multiple channels were averaged. Out of 12 channels per hemisphere, 6 ROIs were created. These were middle frontal gyrus (MFG), pre central gyrus (Pre CG), post central gyrus (Post CG), superior parietal gyrus (SPG), and parts of inferior parietal lobe; supramarginal gyrus (SMG) and angular gyrus (AG) known in the literature as active regions in painful stimulation studies of FM [26]–[31] and healthy controls [32]. The channels that correspond to left (L) and right (R) hemisphere ROIs are illustrated in Fig. 2.

We created functional connectivity matrices using three metrics: cross-correlation analysis with maximal lag (CCA), seed-based correlation analysis (SCA), and Dynamic Time Warping (DTW).

1) Cross Correlation Analysis (CCA) With Maximum Lag: For two different time series $x(n)$ and $y(n)$ represent two different fNIRS channels. If we assume l as shift lag for time series y that is -20 sec to $+20$ sec for this study, maximum cross correlation coefficient for l is

$$\arg\max_{l} r_{xy}(l) = \frac{\sum_{n=-\infty}^{\infty} x(n)\, y(n-l)}{r_{xx}(0)\; r_{yy}(0)} \tag{1}$$

That $r_{xx}(0)$ and $r_{yy}(0)$ are zero-lag autocorrelation estimations of $x(n)$ and $y(n)$.

$$r_{xx}(0) = \sum_{n=-\infty}^{\infty} x(n)\, x(n) \tag{2}$$

$$r_{yy}(0) = \sum_{n=-\infty}^{\infty} y(n)\, y(n) \tag{3}$$

We used CCA with maximum lag from -20 sec to $+20$ sec with 10 Hz sampling rate (-200 $+200$ time points). Among these lag values, we used the one which returned the maximum correlation coefficient as window size in DTW.

2) Seed-Based Correlation Analysis (SCA): Seed-Based Correlation Analysis was performed by estimating the zero-lag correlation coefficient for time series $x(n)$ and $y(n)$ by following formula.

$$r_{xy}(0) = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y} \tag{4}$$

3) Dynamic Time Warping (DTW): DTW is a similarity measurement algorithm that is generally used for two-time series that vary in time or speed [33]. It finds an optimal match between two-time series that might include stretched or compressed parts. To achieve this, it minimizes the total distance between these two-time series. For time series, $x(n) = [x_0, x_1, x_2 \ldots\ldots x_{N-1}]$ and $y(n) = [y_0, y_1, y_2 \ldots\ldots y_{N-1}]$ that have the same length N, a distance matrix A is constructed that has NxN dimensions. We can assume $x(n)$ and $y(n)$ as hemodynamic responses obtained from different channels. For every index of A;

$$A_{ij} = d(x(i), y(j)) \tag{5}$$

show the Euclidean distance between $x(i)$ and $y(j)$. Primary objective of this method is to find the path that minimizes between two time series that starts from the index (0, 0) to (N-1, N-1). This is called warping path.

$$W = [w_0, w_1, \ldots\ldots\ldots w_k] \tag{6}$$

If we assume time indices a and c for one time series and b and d for other time series, kth and k-1th point in our warping path can be identified as $w_k = (a, b)$ and $w_{k-1} = (c, d)$ such that this warping path provides the following conditions;

Monotonic condition: In warping path indices does not go back in time domain. This provides that time points are not repeated in warping path. Indices can either stay same or increase:

$$a - c \geq 0 \,\&\; b - d \geq 0 \tag{7}$$

Continuity condition: Warping path advances one step at a time. Index change between a-c and b-d can be less or equal to one.

$$a - c \leq 1 \,\&\; b - d \leq 1 \tag{8}$$

Boundary condition: The path should start from A(0, 0) and finish A(N-1, N-1).

Warping window condition: An ideal alignment path onto distance matrix A cannot be too far from the diagonal of this matrix. For warping window length r,

$$|a - b| > r \,\&\, |c - d| > r \tag{9}$$

In DTW, warping matrix is created by using linear programming. To find the minimum distance between two time series, first Euclidean distance $d(x(i), y(j))$ should be found and defined as cost value. Then, to proceed in the warping matrix, minimum value of the neighboring cells $(D(i-1, j-1), D(i-1, j), D(i, j-1))$ is chosen and the $D(i, \mathrm{j})$ can be found by the following formula.

$$D(i,j) = d(x(i), y(j)) + \min(D(i-1, j-1), D(i-1, j), D(i, j-1)) \tag{10}$$

We used maximum lag of time points obtained from CCA as window length (DTW-ML) and infinite window length in DTW (DTW-INF) to find DTW distance as a FC measure. The first usage of DTW as a FC metric for classification of time series is reported in [34]. For DTW-INF and DTW-ML, we first normalized it to $[-1, 1]$ scale. -1 value corresponds to the maximum distance in DTW distance matrix and 1 corresponds to the DTW distance equals to zero that represents the maximum similarity. Then, we directly applied Fisher's Z-transformation

to both DTW matrices to decrease the skewness and normalize the distribution.

4) Features Driven From Self-Reported Clinical Data: We used BDI score, right hand and left hand pain threshold values as features for this classification. After normalizing features by using Z-score normalization, we used Hampel filter [35]–[37]. For any feature $D = [d_1, d_2, d_3, \ldots .d_n]$ and a sliding window of length l, local median and standard-deviation is defined as;

$$m_i = \text{median}\,(d_{i-l}, d_{i-l+1}, d_{i-l+2}, \ldots .d_i, \ldots, d_{i+l}, d_{i+l}, d_{i+l}) \tag{11}$$

$$\sigma_i = h\,\text{median}\,(|d_{i-l} - m_i|, .., |d_i - m_i|, \ldots, |d_{i+l} - m_i|) \tag{12}$$

where h is the constant, defined as $h = 1/(\sqrt{2}\,erfc^{-1}(0,5)) = 1,4826$ that is used to make the median standard deviation estimate unbiased. Filter's response for i^{th} element d_i' is given by

$$d_i' = \begin{cases} m_i, & |d_i - m_i| > t\sigma_i \\ d_i, & |d_i - m_i| \leq t\sigma_i \end{cases} \tag{13}$$

In this equation, t represents the threshold value. If it is set to zero, a standard median filter was obtained. For our study, we used $l = 5$ and $t = 1$.

5) Features Driven From Individual Channels: An ideal hemodynamic response function (HRF) is created with 40 sec duration and 20 sec onset latency. Based on our previous work [24], [25], the channels that differed significantly between HC and FM populations are selected to extract features (i.e., channels 6,18 for PAIN, 3,17 for FTT and TENS). Features are defined as the DTW distance between the observed signal and ideal HRF. Features are computed separately for right and left hands, channels and PAIN, PAIN+TENS, TENS and FTT conditions. After extracting features, we normalized them by using z-score normalization.

C. Feature Reduction

After obtaining symmetric FC matrices that all have size 12 × 12, we applied t-test to compare the HC and FM groups for every experiment (PAIN, FTT, MNS, TENS), hand (Left, Right) and connectivity metric (SCA, CCA, DTW-INF and DTW-ML) separately. For significant connections, we chose significance threshold for p = 0.025 for every map. Having determined the significant connections, we saved only these connections as features for every experiment. The FC maps showing significant t-test values of the connections between the HC and FM groups are shown in Table I. The color map indicates t-test values of connectivity of each region in the rows with each region in the columns. Due to high dimensionality, we had to reduce the dimensions of the feature vector aggregated from all experiments by using Principal Component Analysis (PCA) by keeping %95 of variance.

D. Classification

We used the Matlab Classification Learner Application. The algorithms are trained and tested by 10 and 20 fold cross validation. Because of the small sample size in the FM and HC classes, we used data imputation to avoid over generalization. Data imputation is a widely used approach in the data-hungry deep learning systems [38]. As a rule of thumb, regeneration of new data should be restricted to cover at most 30% of the data pool [39]. Since the features extracted from our experiments are converted to z-scores already, data is imputed by using the mean and variance of each feature, separately for the two classes, abiding by the same methodology used in [38]. We generated imputed data for 6 FM samples and 9 HC samples, so that a total of 25 samples are reached at both groups.

Several classifiers are trained using k nearest neighborhood (kNN), support vector machine (SVM), linear discriminant analysis (LDA) and Logistic Regression and Bagged Trees ensemble algorithms embodied in the Matlab toolkit.

As illustrated in Fig 1, each classifier C_i has public executables and stored characteristics such as $N_{H_i}, N_{D_i}, T_{P_i}, F_{P_i}, L_i$ such that N_{H_i}, N_{D_i}, are the number of healthy controls and the number of patients with disease. T_{P_i} is the ratio of true positives, obtained by dividing the correctly classified patients, N_{C_i}, to N_{D_i}. F_{P_i} is the ratio of falsely classified healthy controls, obtained by dividing the falsely classified controls, N_{F_i}, to N_{H_i} (i.e., false positives). L_i is the likelihood ratio of the classifier to detect the disease [14].

$$L_i = \frac{T_{P_i}}{F_{P_i}} \tag{14}$$

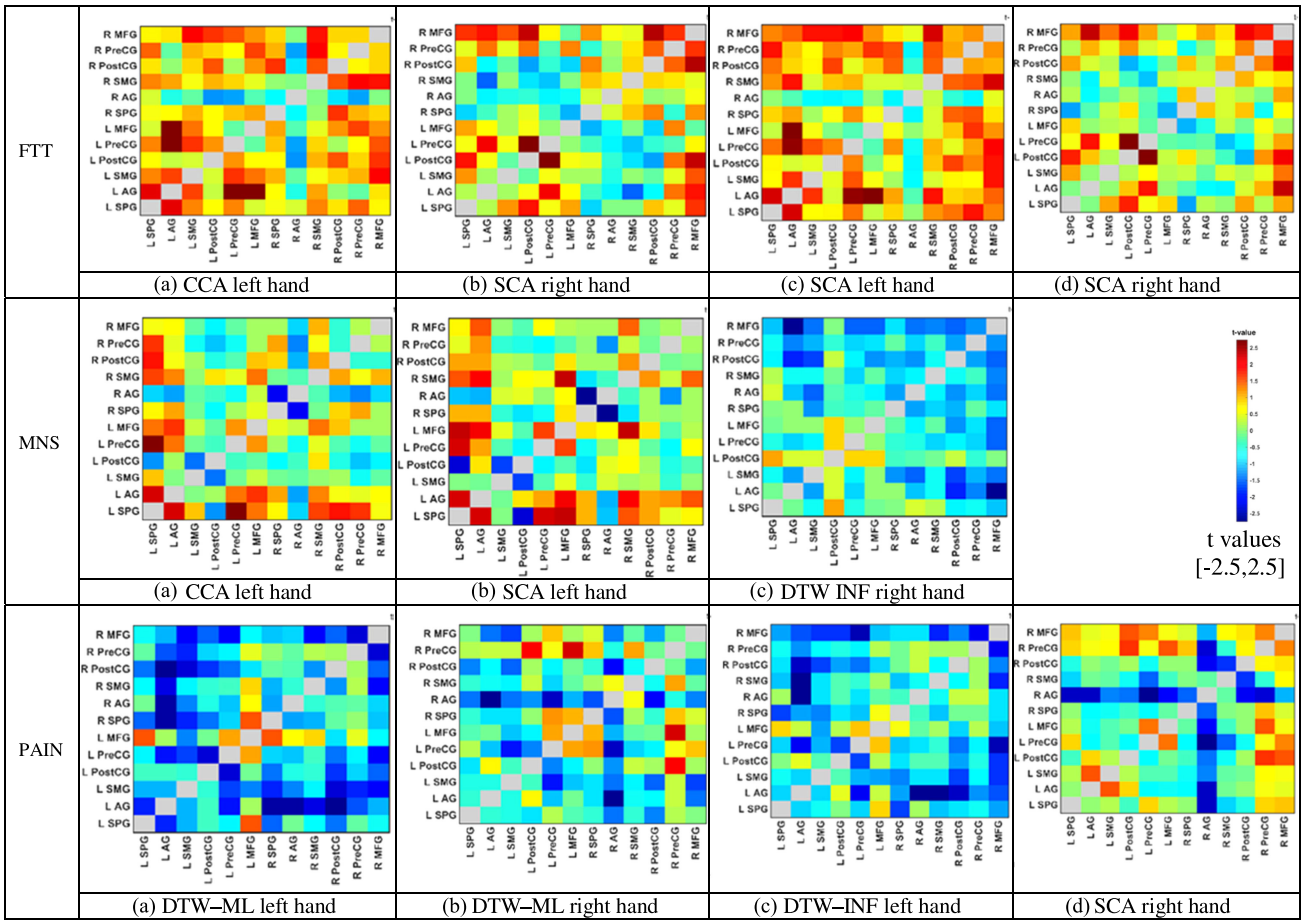
Having a likelihood ratio equal to 1 is non-specific, but having a likelihood ratio smaller than 1 indicates that classifier C_i has negative opinion regarding existence of the disease. On the other hand, if the likelihood ratio is larger than 1, than the detection rate of the classifier increases approximately 5% for each 1 point increase in likelihood [14]. The higher the likelihood L_i, the better the chance of C_i to detect the disease. Now assume that a classifier C_i can be used because its features are measured from the patient. Then using a threshold value, β_i, (which ranges between 0–1) the patient is decided to test positive based on a-posteriori probability, when $p_\text{i} > \beta_\text{i}$. This decision returns a boolean opinion about the existence of disease as u_i from classifier C_i.

$$\begin{array}{ll} \text{if} & p_i > ß_i \\ \text{then} & u_i = 1 \\ \text{else} & u_i = 0 \end{array} \tag{15}$$

E. Fusion

The proposed method herein relies on the assumption that method sharing at the classifier level can be put to practice in the clinics, provided that achieving interoperability at classifier level can easily be guaranteed using apps that operate on mobile or remote platforms. Hence we assume that the clinics are already populated with several classifiers that utilize features driven from the resources available at their disposal. Based on the availability of multiple features obtained from multiple tests or data resources, multiple classifiers may return opinions. Each classifier contributes to the overall opinion based on the decision it supports. If the decision of a classifier is positive, the detection rate of the classifier is taken into account (i.e., number of true

TABLE I
T-MAPS THAT INDICATE SIGNIFICANT CONNECTIVITY DIFFERENCES BETWEEN HEALTHY CONTROLS AND FM PATIENTS FOR 3 DIFFERENT TASKS



positives divided by the number of type 1 errors). On the other hand, if the decision is negative, the rate of erroneous rejection of the classifier is taken into account (i.e., number of type 2 errors divided by the number of true negatives). More specifically, the fused likelihood, L_{f} is defined as follows.

$$L_f = \prod_{i=1}^{K} D_i \tag{16}$$

where D_i represents individual decision from classifier C_{i} such that:

$$\begin{array}{ll} \text{if} & u_i = 1 \\ \text{then} & D_i = \frac{T_{P_i}}{F_{P_i}} \\ \text{else} & D_i = \frac{1-T_{P_i}}{1-F_{P_i}} \end{array} \tag{17}$$

The fused likelihood L_{f}, ranges from 0 to infinity. Values of L_{f}, that are less than 0 indicate that the presence of disease is negative. Values of L_{f}, greater than 1 indicate positive opinion about the presence of disease, such that the larger the L_{f}, the higher the chance for having it. Therefore, a threshold β_{f} which range between 1 to infinity is applied to make the final decision u_{f} based on K classifiers such that,

$$\begin{array}{ll} \text{if} & L_f > ß_f \\ \text{then} & u_f = 1 \\ \text{else} & u_f = 0 \end{array} \tag{18}$$

where u_{f} is the final decision of the fused classifiers. Determining the threshold value β_{f} can be done intuitively, based on calculations regarding the change of detection percentages [14], or based on more sophisticated algorithms that optimize the decision [13], or based on log likelihood which is widely used at the clinics. Referring back to Fig. 1, if it is possible to obtain opinion from an expert physician after classification, then this opinion, u_{e}, -which is either 1 or 0- can be used to update the parameters of each classifier C_{i} that participated in the decision using the following formulae (This is shown as the dashed line in Fig. 1).

$$N_{Hi} = N_{Hi} + (1 - u_e) \tag{19}$$

$$N_{Di} = N_{Di} + u_e \tag{20}$$

$$N_{Ci} = N_{Ci} + u_i \,.\, u_e \tag{21}$$

$$N_{Fi} = N_{Fi} + u_i \,.\, (1 - u_e) \tag{22}$$

True positive and false positive ratios should also be updated to reflect new detection rates. It is important to mention that,

TABLE II
PERFORMANCE OF CLASSIFIERS IN DETECTION OF FM (UPPER ROW: AFTER INDIVIDUAL TRAINING, LOWER ROW: AFTER FUSION)

	Type	Features	Sensitivity	Specificity
C1	Cubic kNN	connectivity	0.64 0.69	0.79 0.80
C2	Fine kNN	channel	0.84 0.74	0.79 0.80
C3	Quadratic SVM	connectivity	0.85 0.86	0.73 0.79
C4	LDA	connectivity	0.83 0.81	0.74 0.80
C5	Cubic kNN	clinical test and survey	0.73 0.79	0.92 0.90

while this update changes the inferred efficiency of classifier C_i in detecting the disease, it does not require retraining the remotely based classifier with the locally collected data, as far as fusion is concerned. Because, the likelihood ratio is dependent only to T_p and F_p ratios. If the features are saved at the local site, along with the decisions u_i, u_e, then these can be sent to the site of the remote classifier for retraining procedures to be started in batch mode at a later time. Thereby, this method allows to merge decisions from multiple sites, which may be operating remotely, by simple multiplication of the D_i values obtained from the classifiers running at those sites.

III. RESULTS

After training several classifiers with 10-fold and 20-fold cross over for 35 subjects, we discarded the classifiers with accuracy less than 60%. As seen in Fig. 1, classifiers C1, C3, C4 operated on connectivity data of the FTT, PAIN and MNS experiments, C2 operated on the channel data of the PAIN experiment and C5 operated on the pain tests and surveys reported in [25]. Parameters of the selected classifiers C_1–C_5 were as follows. In cubic kNN, k: 10, distance metric: Minkowski, Exponent: 3. In fine kNN, k: 1, distance metric: Euclidean. In SVM, kernel: quadratic, polynomial: 2, box constraint:1. We tested multiple β_i values, and computed associated L_i. For each classifer C_i, we chose the β_i which returned the highest L_i. The best values were 0.5 for β_1, β_2 β_3, β_5 and 0.7 for β_4. We presume that with more data available at the clinic, β_i can be trained in batch mode as suggested by [12].

Table II reports the sensitivity and specificity scores of the individual classifiers, after testing for 15 subjects, by arranging a disjoint test subset from the training subset. As seen from Table II, sensitivity ranges between 0.64–0.85 at the end of the training round, while specificity of individual classifiers ranges between 0.73–0.92. At the end of fusion, the T_P rates of individual classifiers are updated with new data, but only a subtle improvement is obtained in sensitivity. When the F_P rates of individual classifiers are updated at the end of fusion, the specificity range of individual classifiers improved slightly to the 0.79–0.90 range.

Fusion is performed by applying (16-18) and then deciding on u_f based on the logarithm of L_f. We set β_f as 0.6, but optimally this should be arranged via some training, during retraining of the classifiers based on expert opinion, u_e. Table III reports the sensitivity and specificity scores of our decision level fusion, by assuming availability of 2, 3, 4 or 5 classifiers. As seen from Table III, likelihood based fusion returns a sensitivity of 1.00 for several combinations of different classifiers. In terms of specificity, the upper range is 0.89–1.00 when there are less than 5 classifiers. When the sensitivity values of the sets of classifiers presented in Table III are aggregated so that average sensitivity S_i is calculated for all fusion sets that contained C_i, the highest values were obtained for S_3 and S_4 as 0.90. This is expected since C_3 represents connectivity in the pain matrix and C_4 represents authentic activity of FM patients due to allodynia [24], [25]. Hence C3 and C4 lead the decision in fusion more strongly than other classifiers.

TABLE III
PERFORMANCE OF FUSED CLASSIFIERS IN FM DETECTION

C1	C2	C3	C4	C5	sens	spec
1	1				0.67	0.89
1		1			0.83	1.00
1			1		0.83	1.00
1				1	0.83	0.89
	1	1			0.83	0.78
	1		1		**1.00**	**0.89**
	1			1	0.67	1.00
		1	1		0.67	0.89
		1		1	0.83	1.00
			1	1	0.83	1.00
1	1	1			0.83	0.89
1	1		1		0.83	0.78
1	1			1	0.67	0.89
1		1	1		**1.00**	**0.89**
1		1		1	0.83	1.00
1			1	1	0.83	1.00
	1	1	1		**1.00**	**0.78**
	1	1		1	0.83	0.89
	1		1	1	0.67	0.89
		1	1	1	0.83	1.00
1	1	1	1		**1.00**	**0.89**
1	1	1		1	**1.00**	**1.00**
1	1		1	1	**1.00**	**1.00**
	1	1	1	1	**1.00**	**0.89**
1		1	1	1	**1.00**	**1.00**
1	1	1	1	1	**1.00**	**1.00**

IV. DISCUSSION

The fusion approach proposed in this study is similar to the method proposed in [13] for fusion of clinical data to detect breast cancer, fusion of EEG and pupil data to detect increased

human performance at target detection task [15] and fusion of EEG and fNIRS data to detect stress of a human operator [16]. Our proposal was different from these with respect to the classifier detection rates, because we proposed to update likelihood values whenever expert opinion about the final decision is available, without waiting for retraining of individual classifiers.

A. FM Classification

To our best knowledge, this is the first study that uses functional connectivity metrics derived from fNIRs data as well as self-reports for diagnosis of FM.

The performance figures we reported in detection of FM are superior to the ones reported in the literature. In [40] resting state functional connectivity data obtained from 50 participants is used to diagnose 17 FM, 16 RA, 17 HC participants. MVPA was used for extracting models to discriminate salience network and default mode network. Maximum accuracy to classify FM was obtained using SVM with linear kernel with regularization as %76. When painful and non-painful fMRI tasks were used along with logistic regression and SVM, patients and healthy controls were correctly classified with 93% accuracy, 92% sensitivity and 94% specificity by using a combined feature set [41].

There exists one other study in FM which includes comparison of classification performance using clinical and neuroimaging data [42]. In that study, several different classifiers including ANN, SVM, Naïve bayes, J48 decision tree, logistic regression and kNN were used to classify 14 FM, and 12 HC subjects using volumes of 56 different brain regions via structural MRI and self-report data including mood and pain intensity. While maximum accuracy of structural MRI based classification was limited to 76% using the J48 classifier, maximum accuracy of self-report based classification was found to be 96%.

In comparison to the aforementioned studies, we developed and applied a different concept to obtain a generalized decision for classification of FM. In our study, we utilized different FC metrics including cross correlation (CCA), seed-based correlation (SCA) and dynamic time warping (DTW) to obtain significantly different connections between the two subject groups, FM and HC. We trained four different classifiers for three different neuroimaging tasks (FTT, MNS, PAIN), three with connectivity data and one with channel data. In addition, we trained a separate classifier for self-reported data. The fusion method we used is more powerful than the other methods, not only because it outperforms them, but because it allows for making quick decisions based on the availability of features that drive individual classifiers. When the maximum likelihood based classifier is applied to FM as a proof of concept, it detected the disease with a 100% sensitivity and 100% specificity when 4 and 5 classifiers are fused.

B. Data Imputation

Sometimes data can be missing at random (for example, a sensor may fail due to power outage) [2]. Data loss is encountered more at the clinic, due to shortage of time in collecting a test, or due to inability of the patient to cooperate, or lack of availability of the diagnostic device for collecting data. In such cases, missing data can be an overbearing factor, reducing the performance of the classifier, during training or test. Several methods based on statistics, models and machine learning are proposed in [2] to reproduce missing data from the existing distribution of features. However these methods are more suitable for feature level fusion. The decision level fusion approach proposed in our framework is robust in its treatment of missing data, because when some classifiers turn off because of missing features, the classification process continues with the remaining classifiers. On the other hand, some of the features shared by multiple classifiers may be more important for prediction. Therefore multiple imputation can be used by replacing each missing value with a set of plausible ones that represent the uncertainty about the right value to impute [2]. We adopted this approach to FM to generate more samples.

C. Machine Learning Effect Size

Regardless of the better sensitivity and specificity figures we obtained in comparison to other methods in FM classification, we are aware of the limitations of the small sample size. In studies that contain small sample size, within-sample heterogeneity is minimized by recruiting patients with common characteristics, excluding nuisance factors that relate to age, education and other biological variability. Hence, studies with small sample size –like ours- report accuracies above 90% [11]. However, when tests are performed on independent samples in a larger cohort, for binary classification with linear classifiers, accuracies drop proportional to the logarithm of the proportion of shared attributes [11]. This is due to heterogeneity between samples among the independent training and test sets. Hence studies with smaller sample size may reach high prediction accuracy at the cost of lower generalizability. On the other hand, studies with large sample size have more generalization power, at the cost of lower accuracy.

Features that are significantly different between groups may have less distinctive contribution to disease prediction. For a normally distributed feature among the groups to be classified, a large effect size (eg. 0.8) which is found to be highly significant may only lead to a modest detection rate just above 60%. Thereby, a new scale should be defined for interpretation of effect sizes in machine learning –or decision level fusion. For instance, accuracies that range from 60–70%, 70–80% and above 80% can be identified as modest, medium and large respectively. As reported in [11], the effect size in machine learning is large in small cohorts with 15–60 samples, and it drops exponentially as sample size increases.

By using the decision level fusion approach proposed here, sample size can be increased on the go, with the update rules and batch mode retraining of the classifiers. Furthermore, heterogeneity can be accommodated by fusing data across classifier sites. And in the meantime, accuracy can still be kept at a fair level, without sharp drops due to heterogeneous tests, because there are multiple classifiers, some of which are unaffected by the heterogeneity of the nuisance features. Our approach is in line with the call in [11], which states that '*machine learning studies with larger samples should be conducted in order to be*

of diagnostic value. The time has come to make the step toward a next "generation" of machine learning studies, using large samples and/or independent validation samples and (re)use (smaller) studies'.

V. Conclusion

We proposed a decision fusion approach based on likelihood for quick, reliable and feasible binary classification. In this framework, individual classifiers effect the final decision in an amount proportional to their rate of detection. Overall detection rate increases based on the number of classifiers that are available to participate in the decision. The system is fault tolerant, because even when some classifiers are unavailable, it can still make a feasible decision based on the existing classifiers. Decision fusion based on likelihood has another advantage. Log of likelihood can be used to assign logistic levels of disease or confidence to the final decision. Hence the proposed system is relevant clinically.

Expert opinion can be incorporated to the system we proposed. In general, the accuracy of the classifier drops in proportion to the interclass kappa (i.e., the rating consistency of the experts) [11]. If expert opinion is available on individual decisions, and if it is known that interclass kappa is high, it is possible to increase the detection rates of individual classifiers without running retraining but using the simple update rule we have suggested, to improve the false positive and false negative rates. Individual classifiers which are located at remote sites can be retrained later in batch mode, provided that the data contained at the local clinics are allowed to be uploaded elsewhere. This way, local data can be harvested globally, for building a larger cohort.

Interoperability at the classifier level can be established by using apps that operate on mobile or remote platforms. Local clinics can be populated with several classifiers that utilize features driven from the resources available at their disposal. Remotely, multiple classifiers can be trained on disjoint patient populations at several different sites in the world. As soon as public access is provided to run the codes of these classifiers in test mode, local clinics can use them for detection tests.

It is imperative that success at the clinic hinges on the widespread sharing of classifier applications. However, this scheme is contingent upon installation of patient privacy and secure cybernetic interoperability protocols. Currently there are ongoing efforts for data sharing [43] and tool sharing, but automatic classification attempts in neuroimaging are rather new [44]. If public access of remotely based classifiers is made a priority in the future, perhaps as a part of outsourcing neuroimaging data analysis [45], this simple and robust binary classification method can serve the community in clinical decision making.

References

[1] O. Dekel, O. Shamir, and L. Xiao, "Learning to classify with missing and corrupted features," *Mach. Learn., J. Article*, vol. 81, no. 2, pp. 149–178, Nov. 1, 2010.

[2] P. J. Garcia-Laencina, J. L. Sancho-Gomez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," (in English), *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, Mar. 2010.

[3] M. Mihara and I. Miyai, "Review of functional near-infrared spectroscopy in neurorehabilitation," *Neurophotonics*, vol. 3, no. 3, Jul. 2016, Art. no. 031414.

[4] M. Wolf, M. Ferrari, and V. Quaresima, "Progress of near-infrared spectroscopy and topography for brain and muscle clinical applications," *J Biomed. Opt.*, vol. 12, no. 6, Nov.-Dec. 2007, Art. no. 062104.

[5] A. S. Bick, A. Mayer, and N. Levin, "From research to clinical practice: implementation of functional magnetic imaging and white matter tractography in the clinical environment," *J. Neurol Sci.*, vol. 312, no. 1/2, pp. 158–165, Jan. 15, 2012.

[6] B. C. Dickerson, "Advances in functional magnetic resonance imaging: technology and clinical applications," *Neurotherapeutics*, vol. 4, no. 3, pp. 360–370, Jul. 2007.

[7] Y. Zhou, K. Wang, Y. Liu, M. Song, S. W. Song, and T. Jiang, "Spontaneous brain activity observed with functional magnetic resonance imaging as a potential biomarker in neuropsychiatric disorders," *Cogn. Neurodyn.*, vol. 4, no. 4, pp. 275–294, Dec. 2010.

[8] Y. Bai, X. Xia, and X. Li, "A review of Resting-State electroencephalography analysis in disorders of consciousness," *Frontiers Neurol.*, vol. 8, pp. 471–479, 2017.

[9] J. M. Billings, M. Eder, W. C. Flood, D. S. Dhami, S. Natarajan, and C. T. Whitlow, "Machine learning applications to Resting-State functional MR imaging analysis," *Neuroimag. Clin. North Amer.*, vol. 27, no. 4, pp. 609–620, Nov. 2017.

[10] J. Brakowski *et al.*, "Resting state brain network function in major depression - Depression symptomatology, antidepressant treatment effects, future research," *J. Psychiatric Res.*, vol. 92, pp. 147–159, Sep. 2017.

[11] H. G. Schnack and R. S. Kahn, "Detecting neuroimaging biomarkers for psychiatric Disorders: Sample size matters," *Frontiers Psychiatry.*, vol. 7, pp. 50–62, 2016.

[12] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[13] J. L. Jesneck, L. W. Nolte, J. A. Baker, C. E. Floyd, and J. Y. Lo, "Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis," *Med. Phys.*, vol. 33, no. 8, pp. 2945–2954, Aug. 2006.

[14] S. McGee, "Simplifying likelihood ratios," *J. Gen. Internal Med.*, vol. 17, no. 8, pp. 647–650, Aug. 2002.

[15] M. Qian *et al.*, "Decision-level fusion of EEG and pupil features for single-trial visual detection analysis," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 7, pp. 1929–1937, Jul. 2009.

[16] F. Al-Shargie, T. B. Tang, and M. Kiguchi, "Stress assessment based on decision fusion of EEG and fNIRS signals," *IEEE Access*, vol. 5, pp. 19889–19896, 2017.

[17] D. J. Clauw, "Fibromyalgia: A clinical review," *JAMA*, vol. 311, no. 15, pp. 1547–1555, Apr. 16, 2014.

[18] F. Wolfe *et al.*, "The american college of rheumatology 1990 criteria for the classification of fibromyalgia. report of the multicenter criteria committee," *Arthritis Rheumatism*, vol. 33, no. 2, pp. 160–172, Feb. 1990.

[19] F. Wolfe *et al.*, "The american college of rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity," *Arthritis Care Res. (Hoboken)*, vol. 62, no. 5, pp. 600–610, May 2010.

[20] B. Garcia-Zapirain, Y. Garcia-Chimeno, and H. Rogers, "Machine learning techniques for automatic classification of patients with fibromyalgia and arthritis," *Int. J. Comput. Trends Technol.*, vol. 25, no. 3, pp. 149–152, 2015.

[21] R. H. Gracely and K. R. Ambrose, "Neuroimaging of fibromyalgia," *Best Pract. Res. Clin. Rheumatol.*, vol. 25, no. 2, pp. 271–284, Apr. 2011.

[22] R. Staud, "Brain imaging in fibromyalgia syndrome," *Clin. Exp. Rheumatol.*, vol. 29, no. 6, pp. S109–S117, Nov.-Dec. 2011.

[23] L. Jorge and E. Amaro, Jr., "Brain Imaging in Fibromyalgia," (in English), *Current Pain Headache Reports*, vol. 16, no. 5, pp. 388–398, 2012.

[24] A. Eken, D. Gokcay, C. Yilmaz, B. Baskak, A. Baltaci, and M. Kara, "Association of fine motor loss and allodynia in fibromyalgia: An fNIRS study," *J. Motor Behav.*, pp. 1–13, Dec. 6, 2017, doi: 10.1080/00222895.2017.1400947.

[25] A. Eken, M. Kara, B. Baskak, A. Baltaci, and D. Gokcay, "Differential efficiency of transcutaneous electrical nerve stimulation in dominant versus nondominant hands in fibromyalgia: placebo-controlled functional near-infrared spectroscopy study," *Neurophotonics*, vol. 5, no. 1, Jan. 2018, Art. no. 011005.

[26] R. H. Gracely, F. Petzke, J. M. Wolf, and D. J. Clauw, "Functional magnetic resonance imaging evidence of augmented pain processing in fibromyalgia," *Arthritis Atollmatol.*, vol. 46, no. 5, pp. 1333–1343, May 2002.

[27] D. B. Cook, G. Lange, D. S. Ciccone, W. C. Liu, J. Steffener, and B. H. Natelson, "Functional imaging of pain in patients with primary fibromyalgia," *J. Rheumatol.*, vol. 31, no. 2, pp. 364–378, Feb. 2004.
[28] T. Giesecke *et al.*, "Evidence of augmented central pain processing in idiopathic chronic low back pain," *Arthritis Rheumatol.*, vol. 50, no. 2, pp. 613–623, Feb. 2004.
[29] R. H. Gracely *et al.*, "Pain catastrophizing and neural responses to pain among persons with fibromyalgia," *Brain*, vol. 127, no. Pt 4, pp. 835–643, Apr. 2004.
[30] M. Burgmer, E. Pogatzki-Zahn, M. Gaubitz, E. Wessoleck, G. Heuft, and B. Pfleiderer, "Altered brain activity during pain processing in fibromyalgia," *Neuroimage*, vol. 44, no. 2, pp. 502–508, Jan. 15, 2009.
[31] J. Pujol *et al.*, "Mapping brain response to pain in fibromyalgia patients using temporal analysis of FMRI," *PLoS One*, vol. 4, no. 4, 2009, Art. no. e5224.
[32] A. V. Apkarian, M. C. Bushnell, R. D. Treede, and J. K. Zubieta, "Human brain mechanisms of pain perception and regulation in health and disease," *Eur. J. Pain*, vol. 9, no. 4, pp. 463–484, Aug. 2005.
[33] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans.Acoust. Speech Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978.
[34] R. J. Meszlenyi, P. Hermann, K. Buza, V. Gal, and Z. Vidnyanszky, "Resting State fMRI functional connectivity analysis using dynamic time warping," *Frontiers Neurosci.*, vol. 11, pp. 75–92, 2017.
[35] F. R. Hampel, "A general qualitative definition of robustness," *Ann. Math. Stat.*, vol. 42, pp. 1887–1896, 1971.
[36] F. R. Hampel, "The influence curve and its role in robust estimation," *J. Amer. Stat. Assoc.*, vol. 69, no. 346, pp. 383–393, 1974.
[37] R. K. Pearson, Y. Neuvo, J. Astola, and M. Gabbouj, "Generalized Hampel Filters," *EURASIP J. Adv. Signal Process.*, vol. 87, no. 1, pp. 1–18, Aug. 5, 2016.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with Deep convolutional neural networks," (in English), *Commun. ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017.
[39] D. B. Rubin, "Multiple imputation after 18+ years," (in English), *J. Amer. Stat. Assoc.*, vol. 91, no. 434, pp. 473–489, Jun. 1996.
[40] B. Sundermann *et al.*, "Diagnostic classification based on functional connectivity in chronic pain: model optimization in fibromyalgia and rheumatoid arthritis," *Acad. Radiol.*, vol. 21, no. 3, pp. 369–377, Mar. 2014.
[41] M. Lopez-Sola *et al.*, "Towards a neurophysiological signature for fibromyalgia," *Pain*, vol. 158, no. 1, pp. 34–47, Jan. 2017.
[42] M. E. Robinson, A. M. O'Shea, J. G. Craggs, D. D. Price, J. E. Letzen, and R. Staud, "Comparison of machine classification algorithms for fibromyalgia: Neuroimages versus self-report," *J. Pain*, vol. 16, no. 5, pp. 472–477, May 2015.
[43] K. J. Gorgolewski *et al.*, "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments," *Sci. Data*, vol. 3, Jun. 21, 2016, Art. no. 160044.
[44] J. V. Haxby, "Multivariate pattern analysis of fMRI: the early beginnings," *Neuroimage*, vol. 62, no. 2, pp. 852–855, Aug. 15, 2012.
[45] A. S. Dick and U. Hasson, "Outsourcing neuroimaging data analysis Implications for scientific accountability and issues in the public interest," *Trends Cognitive Sci.*, vol. 14, no. 1, pp. 2–4, Jan. 2010.