A clinical decision support system using multi-modality imaging data for disease diagnosis

Abstract

Readily available imaging technologies have made it possible to acquire multiple imaging modalities with complementary information for the same patient. These imaging modalities describe different properties about the organ of interest, providing an opportunity for better disease diagnosis, staging and treatment assessments. Extensive research has been done in multi-modality imaging data fusion and integration. However, the existing research has not yet been transformed into a clinical decision support system due to the lack of *flexibility*, *sufficient accuracy*, and *interpretability*. This paper proposes a multi-modality imaging based diagnostic decision support system (MMI-DDS) that overcomes the limitations of the existing research and integrates multi-modality imaging data for disease diagnosis. MMI-DDS includes three inter-connected components: (1) a modality-wise principal component analysis (PCA) that reduces data dimensionality and meanwhile provides the flexibility for opting out tedious and error-prone co-registration for multi-modality images; (2) a novel constrained particle swarm optimization (cPSO) based classifier that is built upon the joint set of the principal components (PCs) from all the imaging modalities and achieves nearly-optimal diagnostic accuracy; (3) a clinical utility engine that employs inverse operations to identify contributing imaging features (a.k.a. biomarkers) in diagnosing the disease. To validate MMI-DDS, we apply it to a migraine dataset with multi-modality structural and functional magnetic resonance imaging (MRI) data collected from Mayo Clinic Arizona and Washington University School of Medicine in St. Louis. MMI-DDS shows significantly improved diagnostic accuracy than using single imaging modalities alone and also identifies biomarkers that are consistent with findings in the migraine literature.

Keywords: clinical decision support, disease diagnosis, multi-modality imaging, particle swarm optimization, classification, migraine, headache

1. Introduction

Imaging has become an indispensable part of modern medicine, and is being extensively used to support diagnosis and other clinical decision making on various diseases such as brain diseases, cardiovascular diseases, and cancer. With the rapid advance of imaging technologies, it is now possible to acquire multiple modalities of imaging data for the same patient. These modalities consist of different but complementary information about the organ of interest, providing an opportunity for better clinical decision support. Taking brain diseases as an example, such as migraine and Alzheimer's disease (AD), a number of imaging modalities can be acquired, which can be broadly classified into structural imaging and functional imaging. Typical structural imaging modalities include computed tomography (CT) and magnetic resonance imaging (MRI): CT shows the gross structure of the brain based on differential absorption of X-rays. MRI produces detailed structural images of the brain using magnetic field and radio waves. Typical functional imaging modalities include functional MRI (fMRI), positron emission tomography (PET), and magnetoencephalography (MEG): fMRI measures blood oxygenation related to neural activity. PET measures physiologic functions in the brain by measuring radiation emitted from tracers injected in the bloodstream. MEG measures magnetic fields produced by the brain's electrical activity using superconducting quantum interference devices.

Recognizing the importance of combining multi-modality imaging data to support disease diagnosis, extensive research has been done, which can be generally categorized into data fusion and data integration. The former interrogates the covariation between different imaging modalities, facilitating knowledge discovery and understanding of the disease biophysiology [1-3]. However,

it does not directly support the diagnosis of each individual patient. Data integration aims at utilizing the different but complementary information contained in the multiple imaging modalities in order to assist with disease diagnosis. Methods for data integration share a common idea of building a classifier that links a combined set of features from individual imaging modalities with the diagnostic result. Commonly used classification models include linear discriminant analysis (LDA) [4, 5], quadratic discriminant analysis (QDA) [6, 7, 8], support vector machines (SVM) [9-11], and multitask learning [12, 13]. Integrating multi-modality imaging data has been shown to produce better classification accuracy than using a single modality alone in a number of diseases such as AD [4, 9, 11-13], schizophrenia [10], migraine [6, 7], and glioblastoma [5, 14].

Despite the abundance of existing research, the research has not yet been transformed into a clinical decision support system due to the lack of three important traits: *flexibility*, *sufficient accuracy*, and *interpretability*. *Flexibility* means that the system can incorporate image features defined at various aggregation levels such as voxels and regions of interest (ROIs). Both voxel-level and ROI-level features are commonly used in imaging-based studies and have their respective strengths: The former preserves the raw information in an image, which avoids information loss. The latter combines prior knowledge (e.g., the anatomical structure of an organ) to guide feature definition. Furthermore, a system with *flexibility* should be able to take image features of various types such as element (voxel or ROI)-wise features and connectivity-based features. Examples of element-wise features include cortical thickness, area, and volume using MRI and regional metabolism using PET. Examples of connectivity-based features include functional connectivity z-maps using fMRI and white matter tractography using diffusion tensor imaging (DTI). Lastly, most multi-modality imaging based studies require co-registration to ensure the images are aligned

into the same coordinate system [15, 16], which is time consuming and error-prone. A system with *flexibility* should provide an option for opting out this procedure.

Sufficient accuracy means a superior performance of the classification model which can be used for individual patient diagnosis instead of group-based analysis. Given the high-dimensionality of the joint feature set produced by multi-modality images, searching for the subset of features with the near-global optimal classification accuracy is very challenging. An exhaustive search is practically impossible. Greedy search based methods such as sequential forward selection and sequential backward selection suffer from a variety of problems such as stagnation in local optima and a high computational cost. Lately, evolutionary computation (EC) techniques such as genetic algorithms (GA) [17], genetic programming (GP) [18], differential evolution (DE) [19], and neuroevolution [20] have attracted great attention with some initial success in feature selection and classification for medical applications. A new emerging field in EC is swarm intelligence [21,22] which models the collective behavior of social swarms in nature, such as ant colonies, honeybees, and bird flocks. Although individuals in a swarm are relatively unsophisticated with limited capabilities on their own, they interact together with certain behavioral patterns to cooperatively achieve tasks necessary for their survival. This "intelligent" behavior of the swarm has inspired new algorithmic developments in solving large complex optimization problems with a wide range of application domains such as machine learning [23], bioinformatics [24], dynamical systems and operations research [25]. Particle swarm optimization (PSO) is a computational algorithm based on swarm intelligence that mimics the behavior of flying birds and their means of information exchange to solve optimization problems. Each potential solution is seen as a particle with a certain velocity, and flies through the problem space. Each particle adjusts its flight according to its own flying experience and its companions' flying experiences. The particle

swarms find optimal regions over complex search spaces through the interaction of individuals in a population of particles. PSO has been successfully applied to a number of difficult combinatorial optimization problems [26, 27]. PSO has also been shown to be computationally less expensive, converge more quickly, and find better solutions than classic EC algorithms such as GA [28, 29].

Interpretability is another important trait that a clinical decision support system should possess. In general, mathematical models can be described as black-box, white-box, or grey-box [30]. Black-box models do not convey information about their inner-workings, and only the input and output are known. White-box models convey explicit information about their internal structure, allowing the user to infer the different components and their connections. Grey-box models display partial theoretical information and use the data that is available to complete the model. In this research, white-box approaches in feature processing and model building are employed to achieve interpretability as it would allow for identification of an analytic pathway that traces back from the classification accuracy to the contributing features and their respective contributing weights. This has at least two benefits: First, it facilitates identification of biomarkers for the disease. Biomarker identification is of vital importance in medical research not only for disease diagnosis but also for understanding the biological basis and developing effective treatments. Second, practitioners tend to be reluctant to adopt black-box approaches regardless of the performance. White-box approaches allow for ready clinical adaptation and dissemination.

In this research, we develop a multi-modality imaging based diagnostic decision support system (MMI-DDS) aiming to possess the aforementioned three traits. MMI-DDS includes three key steps: First, a modality-wise principal component analysis (PCA) is applied to each imaging modality independently. Imaging features are typically high-dimensional. Some features are naturally highly correlated due to their spatial proximity or functional similarity. These pose

challenges to downstream classification model development. PCA is a well-known statistical method for dimension reduction and de-correlation. PCA is also a white-box approach because it applies a linear transformation to the imaging features, which allows for a later inverse-transformation to identify the contributing features to the classification accuracy (i.e., the biomarkers). In MMI-DDS, a modality-wise PCA is employed in order to account for the fact that different imaging modalities may measure the organ of interest from different perspectives. This also provides an option for opting out tedious and error-prone co-registration for the multi-modality images. Second, a novel constrained PSO (cPSO) based classifier is built on the joint set of principal components (PCs) across the multi-modalities. cPSO is an optimizer that searches through the joint PC set to find a small subset of PCs with near-global optimal classification accuracy. In this sense, cPSO combines feature (i.e., PCs) selection and classification in a single framework. The ability of feature selection is important for medical applications since medical data tend to contain many features. Simply training a classifier to all the available features would likely cause overfitting since many of the features are likely to be noise. In theory, the cPSO optimizer can be used for all classification models. In this paper, we choose white-box models such as LDA, QDA, and linear SVM (LSVM) to enable inverse-transformation and biomarker identification in the next step. Third, a clinical utility engine is developed to derive the analytic pathway that traces back from the classification accuracy to the contributing features (i.e., biomarkers) and their respective contributing weights. This allows for interpretation of the diagnostic result and knowledge discovery about the disease.

The rest of the paper is structured as follows: Section 2 provides a literature review. Section 3 presents development of the MMI-DDS. Section 4 presents an application of MMI-DDS for

migraine diagnosis using multi-modality structural and functional imaging data. Section 5 is the conclusion.

2. Literature review

As mentioned in the Introduction, research on combining multi-modality imaging data falls into two categories: data fusion and data integration. This paper belongs to the latter category, but we will review the existing work in both categories in this section due to their relevance.

For data fusion, multivariate statistical methods such as canonical correlation analysis (CCA), partial least squares (PLS), and independent component analysis (ICA) provide viable approaches. CCA finds linear combinations of two sets of variables, called canonical variables, with the maximum correlation between each other. The original CCA can only model two datasets. It was later extended to a multiset-CCA (M-CCA) that finds canonical variables from multiple datasets to achieve the maximum overall correlation [31]. M-CCA was used to perform data fusion of concurrently acquired fMRI and EEG in an auditory task to find covarying amplitude modulations in both modalities and the corresponding spatial activations [32]. It was also used to fuse fMRI, EEG, and MRI to make group inference for schizophrenia patients compared with healthy controls [33].

PLS is a statistical model that finds the multidimensional direction in the space of the independent variables that explains the maximum multidimensional variance direction in the space of the dependent variables. Multiway PLS, as an extension to PLS, was developed for fusion of EEG and fMRI by decomposing EEG and fMRI each as a sum of "atoms" [34]. Each EEG atom was the outer product of spatial, spectral, and temporal signatures and each fMRI atom the product of spatial and temporal signatures. The decomposition was constrained to maximize the covariance

between corresponding temporal signatures of the EEG and fMRI. This fusion aimed at identifying the coherent systems of neural oscillators that contribute to the spontaneous EEG.

ICA is a generative model that assumes the observed multivariate data to be weighted sums of unobserved independent components. ICA is a popular approach in image analysis. Earlier work focused on single imaging modalities such as fMRI and EEG with the purpose of separating the imaging data into meaningful constituent components correlated with subjects' experimental task performance. Recently, ICA has been extended in a number of ways for multi-modality data fusion. Joint ICA (jICA) assumes that the data from multiple imaging modalities share a common demixing matrix [35]. Several studies demonstrated the use of jICA in fusion of fMRIs from multiple tasks, MRI and fMRI, fMRI and EEG, and MRI and DTI for identifying group difference between patients with schizophrenia and controls [3, 35, 36]. Parallel ICA (paraICA) [2, 35, 37] was developed to relax the strong "common demixing matrix" assumption posed by jICA and provided a more flexible approach by creating the mixing matrices for different modalities separately with the goal of maximizing the independence of components within each modality while maximizing the correlation between the mixing matrices. paraICA was used to fuse fMRI and SNP (a genetic modality) in studying schizophrenia [37] and to fuse fMRI and DTI in comparing schizophrenia with bipolar disorder [2]. Tensor ICA [38] was developed to fuse three-way (spatial, temporal, and cross-subject) fMRI data by decomposing the data into a set of independent spatial maps together with associated time courses and estimated subject modes. It was applied to fMRI data collected under a visual, cognitive, and motor paradigm and was able to extract plausible activation maps, time courses, and session/subject modes as well as provide a rich description of additional processes of interest such as image artifacts and secondary activation patterns. Link ICA adopted a Bayesian framework for simultaneously modeling and discovering

common features across multiple modalities [1]. It enjoyed the flexibility of fusing imaging modalities with completely different units, signal- and contrast-to-noise ratios, voxel counts, spatial smoothness and intensity distributions by using a Bayesian formulation to automatically weigh the modalities appropriately.

While being a popular research area, multi-modality imaging data fusion does not directly support diagnosis of each individual patient, but instead provides an exploratory tool for knowledge discovery and group inference. The former is the objective of multi-modality imaging data integration. Research on data integration shares a common idea of building a classifier from a training dataset, which links a combined set of features from individual imaging modalities with a diagnostic result. This classifier can then be used to produce a probability of having the target disease for each new patient, thus providing decision support for clinical diagnosis. In theory, such a classifier can be built using any statistical classification method. Typical methods that have been used for integrating multi-modality imaging data include LDA [4, 5], QDA [6, 7], SVM [9-11], and multitask learning [12, 13]. Integrating multi-modality imaging data has been shown to produce better classification accuracy than using a single modality alone in a number of brain diseases such as AD [4, 9, 11-13], schizophrenia [10], migraine [6, 7], and glioblastoma [5, 14]. Despite the abundance of existing literature, the research is still limited in clinical usability due to lack of flexibility (e.g., only applicable to certain imaging modalities or requiring co-registration), insufficient accuracy (e.g., using off-the-shelf software to build a classification model without exploiting advanced optimizers to improve the performance), and insufficient interpretability (e.g., black-box methods prohibiting rigorous identification of contributing features or biomarkers).

3. Proposed MMI-DDS for disease diagnosis

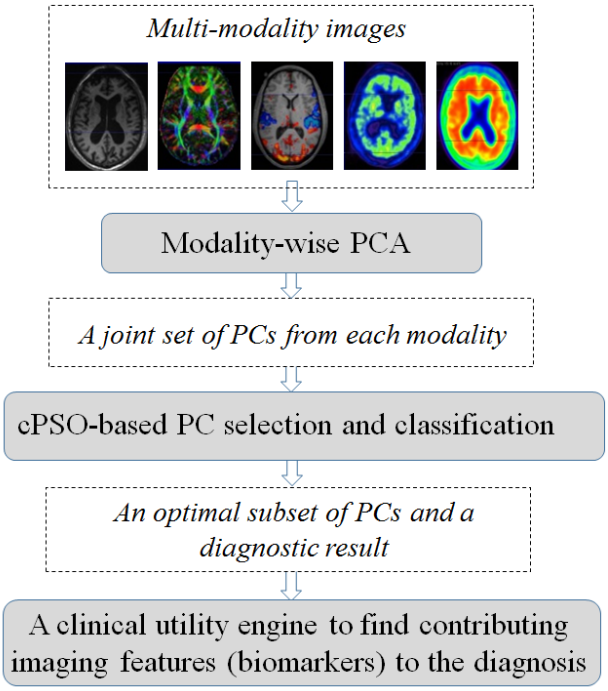

Figure 1: Layout of MMI-DDS

As shown in Figure 1, MMI-DDS includes the following main components: (1) a modality-wise PCA, (2) a cPSO-based classifier for diagnosis, and (3) a clinical utility engine for biomarker identification.

3.1 Modality-wise PCA

PCA is a statistical method that transforms the imaging features that are potentially high-dimensional and correlated into a small number of uncorrelated PCs. Each PC is a linear combination of the imaging features. The transformation is performed in such a way that the first PC has the largest possible variance and each succeeding PC has the highest variance possible under the constraint that it is uncorrelated with all the preceding PCs. We propose to perform PCA on each imaging modality separately. This is to account for the fact that different imaging modalities measure the organ of interest from different perspectives and therefore combining their features in a single PCA is inappropriate. This also provides the flexibility for opting out co-registration of the multi-modality images. Specifically, suppose there are M imaging modalities.

Let $\mathbf{X}_m = [X_{1,m}, \dots, X_{n_m,m}]^T$ be the set of features corresponding to the m-th modality, $m = 1, \dots, M$. n_m is the number features for the m-th modality. Let $\mathbf{Z}_m = [Z_{1,m}, \dots, Z_{p_m,m}]^T$ be the set of PCs. Each PC is a linear combination of the features, i.e., $Z_{i,m} = \mathbf{w}_{i,m}^T \mathbf{X}_m$. $\boldsymbol{w}_{i,m}$ consists of the combination coefficients and is called the loading vector. To obtain the loading vectors for all the PCs, a dataset on the features $\mathbf{X}_m$ needs to be collected, which consists of measurements on $\mathbf{X}_m$ from N samples (i.e., patients). Using the dataset, a sample correlation matrix of $\mathbf{X}_m$, $\mathbf{S}_m$, can be computed and an eigen-decomposition is further performed on $\mathbf{S}_m$. The eigenvalues will be ordered from the largest to the smallest, $\lambda_{1,m}, \dots, \lambda_{p_m,m}$, and the corresponding eigenvectors are the loading vectors for the first through the last PC. Note that not all the PCs need to be kept for subsequent analysis, since the PCs corresponding to small eigenvalues are likely to capture noise in the data but not useful information. To determine the number of PCs to keep, a typical approach is to keep track of the cumulative percentage of variance explained by adding more PCs until a pre-specified threshold is reached. Setting the threshold to be a number between 80%-90% has been found to be adequate for most applications [5, 6, 7].

3.2 cPSO-based feature selection and classification

PSO was originally developed as a population-based stochastic optimization technique, and then extended for feature selection in classification. In this section, we first briefly introduce how generic PSO works for solving an optimization problem and for feature selection. Then, we propose a modified PSO algorithm that can honor a pre-specified maximum number of features to better avoid overfitting, called cPSO.

Consider an optimization problem with decision variables $x_1, \dots, x_D$ and an objective function $f(x_1, \dots, x_D)$ to optimize. PSO is initialized with a population of random solutions called particles. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{iD})$ represent the i-th particle, $i = 1, \dots, I$. Each particle adjusts its

position according to its own experience and the positions of other particles. Specifically, at the t-th iteration, let $\mathbf{p}_i^t$ be the best previous position of the i-th particle (i.e., the position giving the best value for the objective function) and $\mathbf{p}_g^t$ be the best position among all the particles. Then, the position adjustment, called velocity, of the i-th particle along the d-th dimension is given by:

$$v_{id}^t = \omega^t v_{id}^{t-1} + c_1 r_1 (p_{id}^t - x_{id}^t) + c_2 r_2 \left(p_{gd}^t - x_{id}^t\right), \quad (1)$$

$d = 1, \dots, D$. Here, ω^t, c_1, and c_2 are called the inertia weight, cognitive learning factor, and social learning factor, respectively. A proper choice for ω^t provides a balance between global and local exploration, and results in fewer iterations on average to find a sufficiently optimal solution. c_1 and c_2 represent the weighting of the stochastic acceleration terms that pull each particle toward $\mathbf{p}_i^t$ and $\mathbf{p}_g^t$ [28]. ω^t, c_1, and c_2 can be treated as tuning parameters of the PSO algorithm. Alternatively, they can be set by users. A number of appropriate values for the three parameters have been suggested [39]. r_1 and r_2 are sampled from a uniform distribution $U[0,1]$. Furthermore, according to the velocity in (1), the i-th particle can move to a new position, i.e.,

$$x_{id}^{t+1} = x_{id}^t + v_{id}^t. \quad (2)$$

Kennedy and Eberhardt proposed modifications on the afore-described generic PSO, so that the resulting algorithm can be used for feature selection in classification [40]. Suppose there are D features, $Z_1, \dots, Z_D$. Each feature Z_d is associated with a binary decision variable x_d. $x_d = 1$ if Z_d is selected and $x_d = 0$ otherwise. The objective function $f(x_1, \dots, x_D)$ is a cross-validated classification error that is computed using the selected features on a training dataset. Because of the binary nature of the decision variables, (2) is changed to (3) while (1) remains the same.

$$x_{id}^{t+1} = \begin{cases} 1, & if \ \ S(v_{id}^t) > r \\ 0, & otherwise \end{cases}, \quad (3)$$

where $S(v_{id}^t)$ is a sigmoid function used to map v_{id}^t to [0,1], i.e., $S(v_{id}^t) = \frac{1}{1+e^{-v_{id}^t}}$. r is sampled from $U[0,1]$.

In this paper, we propose a cPSO algorithm that can honor a pre-specified maximum number of features to avoid overfitting. Specifically, we modify (3) as follows: Let K denote the maximum number of features allowed in the classification model. For each particle, we order its velocities along all the dimensions from the largest to the smallest. Without loss of generality, we denote the ordered velocities of the i-th particle by $v_{i1}^t, \dots, v_{iD}^t$. Keep the first K largest velocities, $v_{i1}^t, \dots, v_{iK}^t$. A simple modification on (3) could be to make $x_{id}^{t+1} = 1$ if $d \leq K$ and $x_{id}^{t+1} = 0$ otherwise. Although this approach guarantees K features to be selected, the selected features may have poor quality. Here, we consider a feature to have poor quality if it has a negative velocity, $v_{id}^t < 0$, which leads to the sigmoid function $S(v_{id}^t) < 0.5$. Therefore, (3) is modified into (4) in cPSO:

$$x_{id}^{t+1} = \begin{cases} 1, & if \;\; d \leq K \; and \; S(v_{id}^t) > 0.5 \\ 0, & otherwise \end{cases}. \tag{4}$$

Using (4), only the K largest features that have good quality, i.e., have a higher probability of being selected than not being selected, will be kept. Therefore, the number of selected features can be less than or equal to K.

Next, we present the detailed steps of the cPSO algorithm. The input to cPSO includes a training dataset on the joint set of PCs by pooling together the PCs from each imaging modality, denoted by $Z_1, \dots, Z_D$, and a diagnostic result Y. The input also includes several user-specified parameters: the maximum number of PCs, K; the number of particles, I; the number of iterations, T; the maximum velocity used to limit further exploration after convergence to an optimal value, V_{max}. Set $\omega^t = 0.9 - t \cdot 0.5/T$, $c_1 = 2$, and $c_2 = 2$, which are recommended values by the literature [39]. In addition, a classification model needs to be specified. In theory, cPSO can work

with any classification model. In this paper, we focus on white-box models such as LDA, QDA, and LSVM. This is to facilitate identification of the contribution features (i.e., biomarkers) and their respective contributing weights to the classification accuracy in a mathematically and computationally tractable way.

The proposed cPSO algorithm:

Step 1 (initialization): Set the initial position of the i-th particle, $\mathbf{x}_i^0$, by randomly choosing K elements in $\mathbf{x}_i^0$ to be one while making other elements to be zero. Use the PCs corresponding to the non-zero elements in $\mathbf{x}_i^0$ to compute a cross validated (CV) classification error on the training dataset, $f(\mathbf{x}_i^0)$. Set the initial velocity, $\mathbf{v}_i^0$, by sampling each element in $\mathbf{v}_i^0$ from $U[-V_{max}, V_{max}]$. Use (4) to update the initial position of each particle and get $\mathbf{x}_i^1$. Iterate Steps 2-3 with $t = 1,2, \dots T$.

Step 2 (velocity updating): Examine all previous positions of the i-th particle, $f(\mathbf{x}_i^0), \dots, f(\mathbf{x}_i^{t-1})$, and find the position giving the smallest CV classification error, $\mathbf{p}_i^t$. Examine the current positions of all the particles, $f(\mathbf{x}_1^t), \dots, f(\mathbf{x}_I^t)$ and find the position giving the smallest CV classification error, $\mathbf{p}_g^t$. Sample r_1 and r_2 from $U[0,1]$. Use (1) to compute the velocity $\mathbf{v}_i^t$. If $v_{id}^t > V_{max}$, set $v_{id}^t = V_{max}$; if $v_{id}^t < -V_{max}$, set $v_{id}^t = -V_{max}$.

Step 3 (position updating): Order the elements in $\mathbf{v}_i^t$ from the largest to the smallest. Use (4) to compute the new position $\mathbf{x}_i^{t+1}$. If the maximum number of iterations has been reached, i.e., $t + 1 = T$, examine the current positions of all the particles, $f(\mathbf{x}_1^{t+1}), \dots, f(\mathbf{x}_I^{t+1})$, and output the position giving the smallest CV classification error as the optimal solution, together with the corresponding CV error and the PCs that are selected. Otherwise, go back to Step 2.

Finally, we discuss how to select the maximum number of PCs, K. A general trend is that the CV classification error will decrease as K increases. However, this does not mean that a larger K is always preferred, because the decrease in the CV error after K is beyond a certain value is so

minimal that it is neither statistically significant nor practically useful. Allowing a larger K than needed will produce an over-complicated model that likely has problems with over-fitting. Therefore, a recommended approach for choosing the optimal K, i.e., K^*, is to plot the CV errors against different values of K with K ranging from the smallest to the largest, and look for the "elbow" point as the K^*. This is a similar idea to the scree plot used to find the optimal number of PCs in PCA. Alternatively, we may adopt a more rigorous approach that uses hypothesis testing (e.g., a two-sample t test) to compare the CV errors corresponding to K and $K+1$, $K = 1,2,\ldots$. The K^* could be one whose CV error is significantly smaller than that of K^*-1 but not than K^*+1. Other methods for choosing K^* might also be adopted, such as penalizing the error with K (similar to the methods used with AIC and BIC). We acknowledge that this is an open area that no single approach dominates. In practice, these alternative approaches could be tried and the results may be cross-referenced with each other.

3.3 Clinical utility engine for clinical interpretation and biomarker identification

The goal of the clinical utility engine is to identify the contributing original features and their respective contributing weights to the model with best classification accuracy found by cPSO. These can be analytically derived for white-box classification models such as LDA, QDA, and LSVM. We first define some common notations: Let $\mathbf{z}$ be the set of PCs selected by cPSO. $\mathbf{z} = [\mathbf{z}_1^T, \ldots, \mathbf{z}_M^T]^T$, where $\mathbf{z}_m$ represents the selected PCs from the m-th modality, $m = 1, \ldots, M$.

$$\mathbf{z}_m = \mathbf{W}_m^T \mathbf{X}_m \,, \tag{5}$$

where $\mathbf{W}_m$ is the loading matrix obtained from the modality-wise PCA discussed in Section 3.1. Let $\mathbf{w}_m^{j\,T}$ be the j-th row of $\mathbf{W}_m$. Then, (5) can be written as

$$\mathbf{z}_m = \sum_{j=1}^{n_m} \mathbf{w}_m^j X_{j,m}. \tag{6}$$

Next, we will present the development of three inverse-operators for LDA, QDA, and LSVM in achieving the goal of the engine.

LDA inverse operator

The LDA model takes the following form:

$$log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}\mathbf{z} - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + log\frac{\pi}{1-\pi}, \quad (7)$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ are the means of $\mathbf{z}$ for the two classes. LDA assumes that the two classes have the same covariance matrix of $\mathbf{z}$, which is represented by $\boldsymbol{\Sigma}$. $\pi = P(Y=1)$. The classification rule of LDA is that if $log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} > 0$, assign the sample to class 1, and to class 0 otherwise.

$\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}$, and π can be estimated from training data by maximum likelihood estimation (MLE). Then, (7) can be simplified as:

$$log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} = \mathbf{v}^{\boldsymbol{T}}\mathbf{z} + v_0, \quad (8)$$

where $\mathbf{v} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)$ and $v_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + log\frac{\pi}{1-\pi}$. Letting $\mathbf{v} = [\mathbf{v}_1^T, \dots, \mathbf{v}_M^T]^T$, where $\mathbf{v}_m$ are the coefficients corresponding to $\mathbf{z}_m$, and substituting (6) into (8), we get

$$log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})} = \sum_{m=1}^{M}\sum_{j=1}^{n_m}\mathbf{v}_m^T\mathbf{w}_m^j X_{j,m} + v_0. \quad (9)$$

It is clear from (9) that the magnitude of $\mathbf{v}_m^T\mathbf{w}_m^j$ indicates the contribution of each imaging feature $X_{j,m}$ to the classification accuracy. The sign of $\mathbf{v}_m^T\mathbf{w}_m^j$ indicates the direction of the contribution.

QDA inverse operator

The QDA model takes on the following form:

$$log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} = -\frac{1}{2}\mathbf{z}^T(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_0^{-1})\mathbf{z} + (\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1})\mathbf{z} - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 +$$

$$log\frac{\pi}{1-\pi} + log\sqrt{|\boldsymbol{\Sigma}_0|/|\boldsymbol{\Sigma}_1|}, \quad (10)$$

QDA assumes that the two classes have the different covariance matrices of $\mathbf{z}$, which are represented by $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_0$. Then, (10) can be simplified as:

$$log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} = \mathbf{z}^T\boldsymbol{\Phi}\mathbf{z} + \mathbf{q}^T\mathbf{z} + q_0, \quad (11)$$

where $\boldsymbol{\Phi} = -\frac{1}{2}(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_0^{-1})$, $\mathbf{q} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$, and $q_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + log\frac{\pi}{1-\pi} + log\sqrt{|\boldsymbol{\Sigma}_0|/|\boldsymbol{\Sigma}_1|}$. $\boldsymbol{\Phi}$ is a block diagonal matrix under the assumption that the modalities are independent, i.e., $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Phi}_M \end{bmatrix}$, where $\boldsymbol{\Phi}_m$ is the matrix corresponding to the m-th modality, $m = 1, \dots, M$. Letting $\mathbf{q} = [\mathbf{q}_1^T, \dots, \mathbf{q}_M^T]^T$, where $\mathbf{q}_m$ are the coefficients corresponding to $\mathbf{z}_m$, and substituting (6) into (11), we get

$$log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})} = \sum_{m=1}^{M}\sum_{j=1}^{n_m}\mathbf{w}_m^{j\,T}\boldsymbol{\Phi}_m\mathbf{w}_m^j X_{j,m}^2 + \sum_{m=1}^{M}\sum_{j=1}^{n_m}\mathbf{q}_m^T\mathbf{w}_m^j X_{j,m} +$$

$$\sum_{m=1}^{M}\sum_{j=1}^{n_m}\sum_{\substack{k=1 \\ k\neq j}}^{n_m}\mathbf{w}_m^{j\,T}\boldsymbol{\Phi}_m\mathbf{w}_m^k X_{j,m}X_{k,m} + q_0. \quad (12)$$

It is difficult to assess the contribution of each imaging feature $X_{j,m}$ to the classification accuracy based on (12), because of the existence of the cross terms $X_{j,m}X_{k,m}, k = 1, \dots, n_m, k \neq j$. To tackle this difficulty, we propose to take the expectation of $log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}$ with respect to the $X_{k,m}$'s, or equivalently the conditional expectation of $log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}$ with respect to $\mathbf{X}_m$ given $X_{j,m}$. This would average out the contribution from each $X_{k,m}$ and leave only the $X_{j,m}$ to be linked with the classification accuracy. Specifically,

$$E_{\mathbf{X}_m|X_{j,m}}\left[log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}\right] = \mathbf{w}_m^{j\,T}\mathbf{\Phi}_m\mathbf{w}_m^j X_{j,m}^2 + \mathbf{q}_m^T\mathbf{w}_m^j X_{j,m} + \sum_{\substack{k=1\\k\neq j}}^{n_m}\mathbf{w}_m^{k\,T}\mathbf{\Phi}_m\mathbf{w}_m^k \cdot E_{\mathbf{X}_m|X_{j,m}}\left[X_{k,m}^2\right] +$$

$$\sum_{\substack{k=1\\k\neq j}}^{n_m}\mathbf{q}_m^T\mathbf{w}_m^k \cdot E_{\mathbf{X}_m|X_{j,m}}[X_{k,m}] + 2\cdot\sum_{\substack{k=1\\k\neq j}}^{n_m}\mathbf{w}_m^{j\,T}\mathbf{\Phi}_m\mathbf{w}_m^k \cdot X_{j,m}\cdot E_{\mathbf{X}_m|X_{j,m}}\left[X_{k,m}\right] +$$

$$\sum_{\substack{k=1\\k\neq j}}^{n_m}\sum_{\substack{l=1\\l\neq j\\l\neq k}}^{n_m}\mathbf{w}_m^{k\,T}\mathbf{\Phi}_m\mathbf{w}_m^l \cdot E_{\mathbf{X}_m|X_{j,m}}[X_{k,m}X_{l,m}] + q_{0,m} + f(\mathbf{X}_{-m}), \quad (13)$$

where $q_{0,m}$ denotes the portion of q_0 that is associated with the *m*-th modality. Since our purpose here is to assess the contribution of $X_{j,m}$, the imaging features from other modalities than the *m*-th modality are not relevant. Therefore, the terms involving these features are put into $f(\mathbf{X}_{-m})$. Furthermore, assume that the imaging features in each modality follows a multivariate normal distribution, i.e., $\mathbf{X}_m \sim N(\boldsymbol{\mu}_m, \mathbf{\Sigma}_m)$, where $\boldsymbol{\mu}_m = \left(\mu_{1,m} \ldots \mu_{n_m,m}\right)^T$ and $\mathbf{\Sigma}_m = \begin{pmatrix} \sigma_{1,1,m} & \cdots & \sigma_{1,n_m,m} \\ \vdots & \ddots & \vdots \\ \sigma_{n_m,1,m} & \cdots & \sigma_{n_m,n_m,m} \end{pmatrix}$. $\boldsymbol{\mu}_m$ and $\mathbf{\Sigma}_m$ can be estimated from training data. Under this distribution, the expectations in (13) can be derived as:

$$E_{\mathbf{X}_m|X_{j,m}}\left[X_{k,m}\right] = \mu_{k,m} + \frac{\sigma_{k,j,m}(X_{j,m}-\mu_{j,m})}{\sigma_{j,j,m}}, \quad (14a)$$

$$E_{\mathbf{X}_m|X_{j,m}}\left[X_{k,m}^2\right] = \sigma_{k,k,m} - \frac{\sigma_{k,j,m}^2}{\sigma_{j,j,m}} + \left(\mu_{k,m} + \frac{\sigma_{k,j,m}(X_{j,m}-\mu_{j,m})}{\sigma_{j,j,m}}\right)^2, \quad (14b)$$

$$E_{\mathbf{X}_m|X_{j,m}}[X_{k,m}X_{l,m}] = \sigma_{k,l,m} - \frac{\sigma_{k,j,m}\sigma_{l,j,m}}{\sigma_{j,j,m}} + \left(\mu_{k,m} + \frac{\sigma_{k,j,m}(X_{j,m}-\mu_{j,m})}{\sigma_{j,j,m}}\right)\left(\mu_{l,m} + \frac{\sigma_{l,j,m}(X_{j,m}-\mu_{j,m})}{\sigma_{j,j,m}}\right). \quad (14c)$$

After substituting (14a-c) into (13), (13) can be simplified to the general form of

$$E_{\mathbf{X}_m|X_{j,m}}\left[log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}\right] = Q_{j,m}\cdot X_{j,m}^2 + L_{j,m}\cdot X_{j,m} + c_{j,m},$$

where $Q_{j,m}$ and $L_{j,m}$ given by:

$$Q_{j,m} = \mathbf{w}_m^{j\,T}\mathbf{\Phi}_m\mathbf{w}_m^j + \sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\right)^2 \mathbf{w}_m^{k\,T}\mathbf{\Phi}_m\mathbf{w}_m^k + 2\cdot\sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\right)\mathbf{w}_m^{j\,T}\mathbf{\Phi}_m\mathbf{w}_m^k +$$

$$\sum_{\substack{k=1\\k\neq j}}^{n_m}\sum_{\substack{l=1\\l\neq j\\l\neq k}}^{n_m}\left(\frac{\sigma_{k,j,m}\sigma_{l,j,m}}{\sigma_{j,j,m}^2}\right)\mathbf{w}_m^{k\,T}\mathbf{\Phi}_m\mathbf{w}_m^l, \tag{15a}$$

$$L_{j,m} = \mathbf{q}_m^T\mathbf{w}_m^j + \sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\right)\mathbf{q}_m^T\mathbf{w}_m^k + 2\cdot\sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\left(\mu_{k,m} - \frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\mu_{j,m}\right)\right)\mathbf{w}_m^{k\,T}\mathbf{\Phi}_m\mathbf{w}_m^k +$$

$$2\cdot\sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\mu_{k,m} - \frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\mu_{j,m}\right)\mathbf{w}_m^{j\,T}\mathbf{\Phi}_m\mathbf{w}_m^k + \sum_{\substack{k=1\\k\neq j}}^{n_m}\sum_{\substack{l=1\\l\neq j\\l\neq k}}^{n_m}\left(\frac{\sigma_{l,j,m}}{\sigma_{j,j,m}}\left(\mu_{k,m} - \frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\mu_{j,m}\right) +\right.$$

$$\left.\frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\left(\mu_{l,m} - \frac{\sigma_{l,j,m}}{\sigma_{j,j,m}}\mu_{j,m}\right)\right)\mathbf{w}_m^{k\,T}\mathbf{\Phi}_m\mathbf{w}_m^l, \tag{15b}$$

and $c_{j,m}$ includes terms that do not have $X_{j,m}$ so there is no need to explicitly spell it out. It is clear that $Q_{j,m}$ and $L_{j,m}$ indicate the quadratic and linear contribution of each imaging feature $X_{j,m}$ to the classification accuracy, respectively.

LSVM inverse operator

The LSVM model takes the following form:

$$f(\mathbf{z}) = \mathbf{s}^T\mathbf{z} + s_0, \tag{16}$$

where $\mathbf{s}$ and s_0 are estimated from the objective function $\min_{\mathbf{s},s_0,\boldsymbol{\xi}} \frac{1}{2}\mathbf{s}^T\mathbf{s} + C\sum_i \xi_i$ subject to $y_i f(\mathbf{z}_i) \geq 1 - \xi_i$ and $\xi_i \geq 0\ \forall i$, where C is the penalty parameter, ξ_i is the slack variable for sample i in a training dataset, y_i is the class of sample i, and $f(\mathbf{z}_i)$ is the predicted value of sample i. Letting $\mathbf{s} = [\mathbf{s}_1^T, \ldots, \mathbf{s}_M^T]^T$, where $\mathbf{s}_m$ are the coefficients corresponding to $\mathbf{z}_m$, and substituting (6) into (16), we get

$$f(\mathbf{X}) = \sum_{m=1}^{M}\sum_{j=1}^{n_m}\mathbf{s}_m^T\mathbf{w}_m^j X_{j,m} + s_0. \tag{17}$$

It is clear from (17) that the magnitude of $\mathbf{s}_m^T\mathbf{w}_m^j$ indicates the contribution of each imaging feature $X_{j,m}$ to the classification accuracy. The sign of $\mathbf{s}_m^T\mathbf{w}_m^j$ indicates the direction of the contribution.

4. Clinical application: a migraine Study

Approximately 36 million Americans suffer from migraine [41]. Current clinical diagnosis is primarily symptom-based, which is prone to patient subjectivity. Imaging has shown great promise for providing objective measures of the disease and for improving the diagnostic accuracy [6, 7]. However, most existing research on migraine diagnosis focuses on single modalities. In this section, we present a study of using MMI-DDS to integrate multi-modality structural and functional imaging data for migraine diagnosis.

4.1 Subject selection and image acquisition and preprocessing

The data used for this application were obtained from Mayo Clinic Arizona and Washington University School of Medicine in St. Louis: A total of 106 subjects who had structural and functional MRI data were included in this analysis, consisting of 57 individuals with migraine (PMs) and 49 healthy controls (HCs). These 106 subjects were a subset of subjects included in prior analyses [6] [7]. PMs were diagnosed in accordance with the diagnostic criteria defined by the International Classification of Headache Disorders [42].

Structural MRI data were obtained from two Siemens 3T MRI machines. Using a cortical reconstruction and segmentation program in the FreeSurfer image analysis suite (version 5.3, http://www.surfer.nmr.mgh.harvard.edu/), cortical area, thickness and volume measurements of 68 ROIs were extracted. Additionally, resting-state functional connectivities, i.e., fMRI data, were collected for each subject. Standard Statistical Parametric Mapping (SPM) methods were used to preprocess the fMRI data. Specifically, fMRI signals were temporally filtered between 0.01 to 0.1 Hz to retain the low frequency components. Variance relating to signals of no interest was removed through linear regression. 33 ROIs were chosen based on commonly cited regions for which PMs show abnormalities [43, 44]. Among the 33 ROIs, there are 16 pairs; each pair consists of two

regions with the same name but located at the left and right sides of the brain, respectively. The remaining one ROI is located in the middle of the brain. We aggregated each pair of ROIs into one ROI by averaging their respective time courses. This reduces the number of ROIs to 16+1=17. Partial correlations between the 17 ROIs were computed, forming 136 connectivity features. Note that we also tried keeping the original 33 ROIs without pair-wise aggregation, but the result was not as good as the one with aggregation.

In summary, this study utilizes two imaging modalities in terms of the image acquisition techniques, i.e., structural MRI and fMRI. Structural MRI produces three sets of features for 68 ROIs, i.e., area features, thickness features, and volume features. Because these three sets measure different aspects of the brain structure, they are treated as three modalities in our analysis. As a result, four modalities are used in MMS-DDS, including cortical area (68 features), thickness (68 features), volume (68 features), and resting-state functional connectivity (136 features).

4.2 Classification accuracy by multi-modality imaging data integration

In this experiment, we show the performance of our system in integrating all the imaging modalities. Specifically, we first apply modality-wise PCA to each modality and keep the PCs that explain 85% of the variance in the data of the respective modality. Then, cPSO takes as input the data on the combined PC set across all the modalities. The optimal parameter K for cPSO is found to be $K^* = 8$, 6, and 9, respectively. K* was chosen as the value at the "elbow" of the plot of CV errors against different values of K. Table 1 (last column) shows the CV classification errors corresponding to LDA, QDA, and LSVM under their respective K^*. For comparison, we also apply our system to integrating the three sets of features from structural MRI, i.e., cortical area, thickness, and volume, and the result is shown in the first column of Table 1. Furthermore, we report the

result on using resting-state functional connectivity from fMRI alone. These analyses aim to show the benefit of integrating structural and functional imaging data.

Table 1: CV classification errors (avg +/- std error) of the proposed MMI-DDS applied to MRI alone, fMRI alone, and MRI+fMRI combined

	MRI (area+thickness+volume)	**fMRI**	**MRI+fMRI**
LDA	24.43% +/- 0.79%	27.17% +/- 0.74%	21.79% +/- 0.50%
QDA	26.32% +/- 0.53%	29.72% +/- 0.75%	22.45% +/- 0.48%
LSVM	20.38% +/- 0.63%	25.38% +/- 0.89%	17.17% +/- 0.19%

In all three classifiers, our system's ability for integrating data from structural and functional imaging modalities is evident. Using a two-sample t-test, the CV error of MRI+fMRI is significantly lower than MRI alone with p values of 0.0062, 2.2×10^{-5} and 2.8×10^{-4} for LDA, QDA, and LSVM, respectively. Because the CV errors of MRI are lower than fMRI, there is no need to compare MRI+fMRI with fMRI. We conclude the integration of multi-modality imaging can significantly improve the diagnosis accuracy. Furthermore, among the three classifiers, LSVM achieves the lowest error, i.e., highest accuracy of 83%, using MRI+fMRI.

Please note in the single modality migraine study [6] where structural MR data were analyzed, the classification accuracy was 68%; and the single modality migraine study using fMRI data had 81% classification accuracy [7]. One may argue that the 83% accuracy reported in this study is a marginal improvement compared to 81% accuracy. We contend that first, Table 1 indicates the statistical differences between the two approaches (fMRI+MRI vs. fMRI) using the same features sets. Second, a voxel-by-voxel connectivity approach was adopted in [7] while 136 features measuring the correlations among 17 ROIs were used in this research. Since one of the key traits of the proposed MMI-DDS is *interpretability*, the use of a ROI based approach may have easy adoption in clinical practice. It is certainly the interest of the team to explore the use of a

voxel-by-voxel approach to investigate whether a better accuracy may be achieved from this dataset.

4.3 Biomarker identification

For each classification model in the last column of Table 1, we apply the proposed clinical utility engine to find the contribution of each feature in the respective imaging modality. Because LSVM gives the highest accuracy, next we examine the result for LSVM more closely. Specifically, we would like to focus on the features that have large positive or negative contributions to the classification accuracy, i.e., features whose contribution weights are large in magnitude. These features have higher likelihood of being potential migraine biomarkers. To this end, we pool the weights from all the modalities together and rank them from the largest to the smallest in terms of their magnitudes. This would give us a rank for the features. Table 2 lists the features that rank in the top 5%. These roughly correspond to features that are significant at 0.05 significance level, a common choice for assessing statistical significance. Figures 2 highlights the ROIs corresponding to the area features in Table 2 on the brain surface. Figure 3 shows the resting-state functional connectivity in Table 2 on the brain surface.

Table 2: Imaging features that rank in the top 5% in terms of the magnitudes of contribution weights for LSVM (L: left hemisphere of the brain; R: right hemisphere of the brain)

Feature set	Features
Area (MRI)	Frontal pole (L), Inferior temporal (L), Middle temporal (L), Transverse temporal (L), Transverse temporal (R), Banks of the superior temporal (R), Precentral (R), Paracentral (R), Entorhinal (R)
Thickness (MRI)	Insula (R)
Volume (MRI)	None
Resting-state functional Connectivity (fMRI)	< Posterior cingulate, Dorsolateral prefrontal > < Anterior cingulate, Amygdala> < Inferior lateral parietal, Supplementary motor > < Primary somatosensory, Temporal pole > < Temporal pole, Caudate> < Middle cingulate, Secondary somatosensory > < Inferior lateral parietal, Temporal pole >

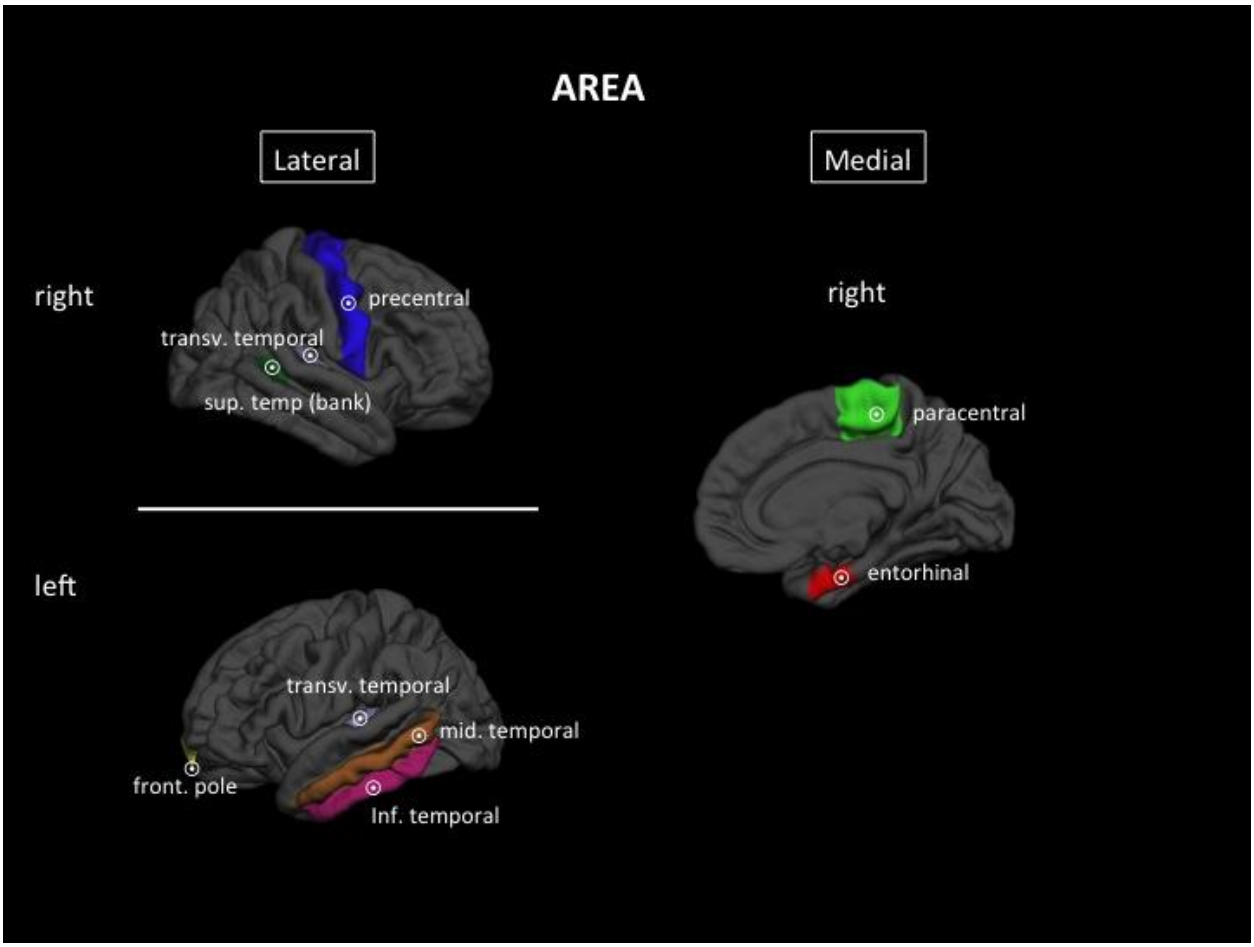

Figure 2: ROIs corresponding to the area features in Table 2 shown on an average inflated brain surface. front. pole=frontal pole; inf. temporal=inferior temporal; mid. temporal=middle temporal; sup. temp (bank)= bank of the superior temporal; transv. temporal=transverse temporal

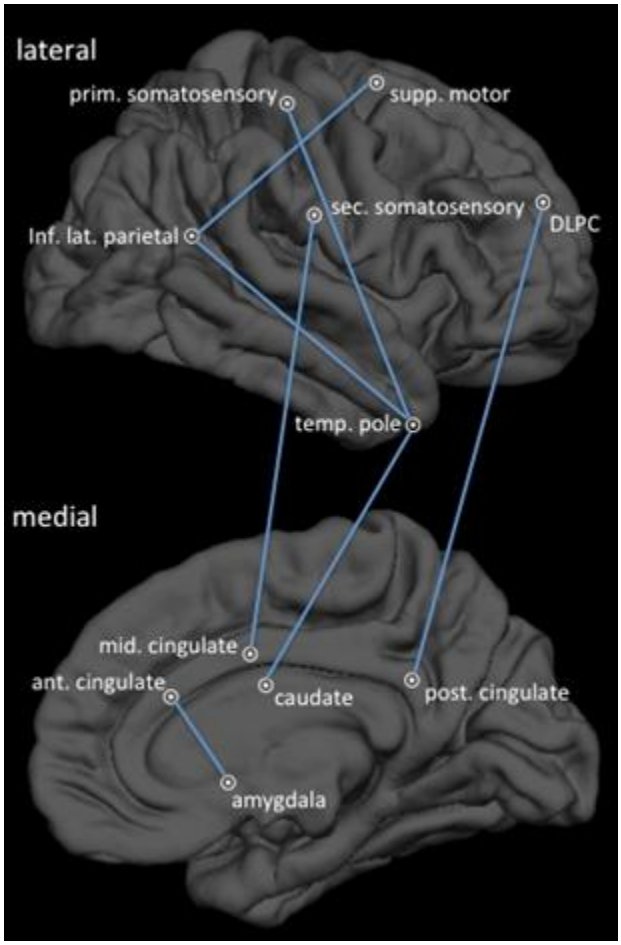

Figure 3: Resting-state functional connectivities corresponding to Table 2. For illustration purposes, functional connectivities are shown on an inflated right hemisphere average brain surface. DLPC=dorsolateral prefrontal; ant. cingulate=anterior cingulate; inf. lat. parietal=inferior lateral parietal; mid. cingulate=middle cingulate; post. cingulate=posterior cingulate; prim. somatosensory=primary somatosensory; sec. somatosensory=secondary somatosensory; supp. motor=supplementary motor; temp. pole=temporal pole

As expected, given the symptoms of migraine, brain regions most contributing to migraine classification (those listed in Table 2) play important roles in pain processing and processing of multisensory stimuli. Whereas some are regions that are predominantly responsible for sensory-

discriminative pain processing (e.g. somatosensory cortex), others are responsible for affective-emotional processing (e.g. amygdala, anterior cingulate cortex), cognitive processing (e.g. prefrontal cortex), or integration of incoming sensory information from different domains (e.g. temporal pole). Several of these regions have commonly been identified as having atypical structure or function in previous migraine studies. The temporal pole, a multisensory region that integrates somatosensory, visual, auditory, and olfactory stimuli [45], has frequently been identified to have atypical structure, function and functional connectivity in migraine studies [46, 47]. Atypical function of the temporal pole in PMs might contribute to common migraine symptoms such as the exacerbation of migraine headache intensity when exposed to lights and sounds. The anterior cingulate cortex is involved in affective components of pain processing including pain anticipation [48], and has been shown to have atypical activation, structure, and functional connectivity in PMs [44, 49, 50]. The amygdala and middle cingulate cortex are also involved with determining pain affect, with the middle cingulate cortex possibly having additional roles in the integration of other aspects of pain processing (e.g. sensory discriminative, affective, cognitive) [48, 51]. One fMRI study on heat pain processing found that interictal PMs showed stronger middle cingulate cortex activation than HCs [52]. PMs have also been demonstrated to have atypical stimulus-induced activation of the amygdala during migraine attacks and atypical functional connectivity of the amygdala compared to HCs. [50, 53]. Our findings are consistent with these previous findings.

5. Conclusion

In this paper, we developed a clinical decision support system, MMI-DDS, that integrates multi-modality imaging data for disease diagnosis. The system was designed to achieve flexibility, sufficient accuracy, and interpretability, which are three important traits required for clinical

decision support systems, but unfortunately are inadequately addressed by prior research. Specifically, our proposed system included a modality-wise PCA, a cPSO algorithm for classification, and a clinical utility engine for identifying contributing features to facilitate biomarker identification. We applied the proposed MMI-DDS to migraine diagnosis by integrating cortical thickness, area, and volume data acquired from structural MRI and resting-state functional connectivity data from fMRI. A high accuracy of 83% was achieved by integrating the structural and functional modalities together, which is significantly better than using single modalities alone. Furthermore, the clinical utility engine identified contributing features to the classification accuracy. Highly ranked features according to their respective contributing weights were found to be relevant to migraine as confirmed by existing studies. Future research includes extending the system's capability to multi-class classification that is useful for disease subtype classification, and to prediction of numerical response variables such as disease severity.

6. References

[1] A. R. Groves, C. F. Beckmann, S. M. Smith and M. W. Woolrich, "Linked independent component analysis for multimodal data fusion," *Neuroimage,* vol. 54, pp. 2198-2217, 2011.

[2] J. Sui, G. Pearlson, A. Caprihan, T. Adali, K. A. Kiehl, J. Liu, J. Yamamoto and V. D. Calhoun, "Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA joint ICA model," *Neuroimage,* vol. 57, pp. 839-855, 2011.

[3] V. D. Calhoun, T. Adali, N. Giuliani, J. Pekar, K. Kiehl and G. Pearlson, "Method for multimodal analysis of independent source differences in schizophrenia: combining gray matter structural and auditory oddball functional data," *Hum. Brain Mapp.,* vol. 27, pp. 47-62, 2006.

[4] S. Huang, J. Li, J. Ye, T. Wu, K. Chen, A. Fleisher and E. Reiman, "Identifying alzheimer's disease-related brain regions from multi-modality neuroimaging data using sparse composite linear discrimination analysis," in *Advances in Neural Information Processing Systems,* 2011, pp. 1431-1439.

[5] L. S. Hu, S. Ning, J. M. Eschbacher, N. Gaw, A. C. Dueck, K. A. Smith, P. Nakaji, J. Plasencia, S. Ranjbar and S. J. Price, "Multi-parametric MRI and texture analysis to visualize

spatial histologic heterogeneity and tumor extent in glioblastoma," *PloS One*, vol. 10, pp. e0141506, 2015.

[6] T. J. Schwedt, C. D. Chong, T. Wu, N. Gaw, Y. Fu and J. Li, "Accurate classification of chronic migraine via brain magnetic resonance imaging," *Headache: The Journal of Head and Face Pain*, vol. 55, pp. 762-777, 2015.

[7] C. D. Chong, N. Gaw, Y. Fu, J. Li, T. Wu and T. J. Schwedt, "Migraine classification using magnetic resonance imaging resting-state functional connectivity data," *Cephalalgia*, Jun 15, 2016.

[8] Zhang Q., Wu, Q., Zhang, J., He, L., Huang, J., Zhang, J., Huang H., Gong, Q., "Discriminative Analysis of Migraine without Aura: Using Functional and Structural MRI with a multi-feature classification approach", PloS One 11(9): e0163875. doi:10.1371/journal.pone.0163875,

[9] Y. Fan, S. M. Resnick, X. Wu and C. Davatzikos, "Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study," *Neuroimage*, vol. 41, pp. 277-285, 2008.

[10] H. Yang, J. Liu, J. Sui, G. Pearlson and V. D. Calhoun, "A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia," *Frontiers in Human Neuroscience*, vol. 4, pp. 192, 2010.

[11] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen and Alzheimer's Disease Neuroimaging Initiative, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 55, pp. 856-867, 2011.

[12] G. Yu, Y. Liu, K. Thung and D. Shen, "Multi-task linear programming discriminant analysis for the identification of progressive MCI individuals," *PloS One*, vol. 9, pp. e96458, 2014.

[13] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, J. Ye and Alzheimer's Disease Neuroimaging Initiative, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *Neuroimage*, vol. 61, pp. 622-632, 2012.

[14] L. S. Hu, S. Ning, J. M. Eschbacher, L. C. Baxter, N. Gaw, S. Ranjbar, J. Plasencia, A. C. Dueck, S. Peng, K. A. Smith, P. Nakaji, J. P. Karis, C. Quarles, T. Wu, J. Loftus, R. Jenkins, B. P. O'Neill, W. Elmquist, J. M. Hoxworth, D. Frakes, J. Sarkaria, K. R. Swanson, N. Tran, J. Li and J. R. Mitchell, "Radiogenomics to Characterize Regional Genetic Heterogeneity in Glioblastoma," *Neuro-Oncology*, In Press.

[15] J. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Med. Image Anal.*, vol. 2, pp. 1-36, 1998.

[16] D. L. Hill, P. G. Batchelor, M. Holden and D. J. Hawkes, "Medical image registration," *Phys. Med. Biol.*, vol. 46, pp. R1, 2001.

[17] A. Fraser and D. Burnell, "Computer models in genetics." *Computer Models in Genetics.,* 1970.

[18] J. R. Koza, *Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems.* Stanford University, Department of Computer Science, 1990.

[19] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optimiz.,* vol. 11, pp. 341-359, 1997.

[20] D. Floreano, P. Dürr and C. Mattiussi, "Neuroevolution: from architectures to learning," *Evolutionary Intelligence,* vol. 1, pp. 47-62, 2008.

[21] M. R. Bonyadi and Z. Michalewicz, "Particle swarm optimization for single objective continuous space problems: a review," *Evol. Comput.,* 2016.

[22] J. Kennedy, J. F. Kennedy, R. C. Eberhart and Y. Shi, *Swarm Intelligence.* Morgan Kaufmann, 2001.

[23] S. Das, B. Panigrahi and S. Pattnaik, "Nature-inspired algorithms for multi-objective optimization," *Handbook of Research on Machine Learning Applications and Trends: Algorithms Methods and Techniques, Hershey, New York,* vol. 1, pp. 95-108, 2009.

[24] S. Das, A. Abraham and A. Konar, "Swarm intelligence algorithms in bioinformatics," in *Computational Intelligence in Bioinformatics*Anonymous Springer, 2008, pp. 113-147.

[25] K. E. Parsopoulos, *Particle Swarm Optimization and Intelligence: Advances and Applications: Advances and Applications.* IGI Global, 2010.

[26] B. Jarboui, N. Damak, P. Siarry and A. Rebai, "A combinatorial particle swarm optimization for solving multi-mode resource-constrained project scheduling problems," *Applied Mathematics and Computation,* vol. 195, pp. 299-308, 2008.

[27] X. Chu, Q. Lu, B. Niu and T. Wu, "Solving the distribution center location problem based on multi-swarm cooperative particle swarm optimizer," in *International Conference on Intelligent Computing,* 2012, pp. 626-633.

[28] X. Wang, J. Yang, X. Teng, W. Xia and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recog. Lett.,* vol. 28, pp. 459-471, 2007.

[29] B. Jarboui, M. Cheikh, P. Siarry and A. Rebai, "Combinatorial particle swarm optimization (CPSO) for partitional clustering problem," *Applied Mathematics and Computation,* vol. 192, pp. 337-345, 2007.

[30] M. E. Khan and F. Khan, "A comparative study of white box, black box and grey box testing techniques," *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 3, 2012.

[31] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika,* vol. 58, pp. 433-451, 1971.

[32] N. M. Correa, T. Eichele, T. Adalı, Y. Li and V. D. Calhoun, "Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI," *Neuroimage,* vol. 50, pp. 1438-1445, 2010.

[33] N. M. Correa, Y. Li, T. Adali and V. D. Calhoun, "Fusion of fMRI, sMRI, and EEG data using canonical correlation analysis," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing,* 2009, pp. 385-388.

[34] E. Martınez-Montes, P. A. Valdés-Sosa, F. Miwakeichi, R. I. Goldman and M. S. Cohen, "Concurrent EEG/fMRI analysis by multiway partial least squares," *Neuroimage,* vol. 22, pp. 1023-1034, 2004.

[35] V. D. Calhoun and T. Adali, "Feature-based fusion of medical imaging data," *IEEE Transactions on Information Technology in Biomedicine,* vol. 13, pp. 711-720, 2009.

[36] L. Xu, G. Pearlson and V. D. Calhoun, "Joint source based morphometry identifies linked gray and white matter group differences," *Neuroimage,* vol. 44, pp. 777-789, 2009.

[37] J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N. I. Perrone-Bizzozero and V. Calhoun, "Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA," *Hum. Brain Mapp.,* vol. 30, pp. 241-255, 2009.

[38] C. F. Beckmann and S. M. Smith, "Tensorial extensions of independent component analysis for multisubject FMRI analysis," *Neuroimage,* vol. 25, pp. 294-311, 2005.

[39] R. Poli, J. Kennedy and T. Blackwell, "Particle swarm optimization," *Swarm Intelligence,* vol. 1, pp. 33-57, 2007.

[40] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference On,* 1997, pp. 4104-4108.

[41] O. Daniel and A. Mauskop, "Nutraceuticals in Acute and Prophylactic Treatment of Migraine," *Current Treatment Options in Neurology,* vol. 18, pp. 1-8, 2016.

[42] Headache Classification Committee of the International Headache Society (IHS), "The International Classification of Headache Disorders, 3rd edition (beta version)," *Cephalalgia,* vol. 33, pp. 629-808, Jul, 2013.

[43] C. Mainero, J. Boshyan and N. Hadjikhani, "Altered functional magnetic resonance imaging resting-state connectivity in periaqueductal gray networks in migraine," *Ann. Neurol.,* vol. 70, pp. 838-845, 2011.

[44] A. Russo, A. Tessitore, A. Giordano, D. Corbo, L. Marcuccio, M. De Stefano, F. Salemi, R. Conforti, F. Esposito and G. Tedeschi, "Executive resting-state network connectivity in migraine without aura," *Cephalalgia,* vol. 32, pp. 1041-1048, Oct, 2012.

[45] T. J. Schwedt, "Multisensory integration in migraine," *Curr. Opin. Neurol.,* vol. 26, pp. 248-253, Jun, 2013.

[46] T. J. Schwedt, L. Larson-Prior, R. S. Coalson, T. Nolan, S. Mar, B. M. Ances, T. Benzinger and B. L. Schlaggar, "Allodynia and descending pain modulation in migraine: a resting state functional connectivity analysis," *Pain Medicine,* vol. 15, pp. 154-165, 2014.

[47] M. A. Rocca, A. Ceccarelli, A. Falini, B. Colombo, P. Tortorella, L. Bernasconi, G. Comi, G. Scotti and M. Filippi, "Brain gray matter changes in migraine patients with T2-visible lesions: a 3-T MRI study," *Stroke,* vol. 37, pp. 1765-1770, Jul, 2006.

[48] S. Palermo, F. Benedetti, T. Costa and M. Amanzio, "Pain anticipation: An activation likelihood estimation meta-analysis of brain imaging studies." *Hum. Brain Mapp.,* vol. 36, pp. 1648-1661, 2015.

[49] C. Jin, K. Yuan, L. Zhao, L. Zhao, D. Yu, K. M. Deneen, M. Zhang, W. Qin, W. Sun and J. Tian, "Structural and functional abnormalities in migraine patients without aura," *NMR Biomed.,* vol. 26, pp. 58-64, 2013.

[50] T. J. Schwedt, B. L. Schlaggar, S. Mar, T. Nolan, R. S. Coalson, B. Nardos, T. Benzinger and L. J. Larson-Prior, "Atypical Resting-State Functional Connectivity of Affective Pain Regions in Chronic Migraine," *Headache: The Journal of Head and Face Pain,* vol. 53, pp. 737-751, 2013.

[51] L. E. Simons, E. A. Moulton, C. Linnman, E. Carpino, L. Becerra and D. Borsook, "The human amygdala and pain: evidence from neuroimaging," *Hum. Brain Mapp.,* vol. 35, pp. 527-538, 2014.

[52] T. J. Schwedt, C. D. Chong, C. C. Chiang, L. Baxter, B. L. Schlaggar and D. W. Dodick, "Enhanced pain-induced activity of pain-processing regions in a case-control study of episodic migraine," *Cephalalgia,* vol. 34, pp. 947-958, Oct, 2014.

[53] A. Stankewitz and A. May, "Increased limbic and brainstem activity during migraine attacks following olfactory stimulation," *Neurology,* vol. 77, pp. 476-482, Aug 2, 2011.