Expert Systems With Applications 169 (2021) 114338

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

DTI based Alzheimer's disease classification with rank modulated fusion of CNNs and random forest

Arijit De, Ananda S. Chowdhury *

Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, 700032, West Bengal, India

ARTICLE INFO

Keywords:
DTI
Alzheimer's disease classification
VoxCNN
Random forest
Decision fusion

ABSTRACT

Automated classification of Alzheimer's disease (AD) plays a key role in the diagnosis of dementia. In this paper, we solve for the first time a direct four-class classification problem, namely, AD, Normal Control (CN), Early Mild Cognitive Impairment (EMCI) and Late Mild Cognitive Impairment (LMCI) by processing Diffusion Tensor Imaging (DTI) in 3D. DTI provides information on brain anatomy in form of Fractional Anisotropy (FA) and Mean Diffusivity (MD) along with Echo Planar Imaging (EPI) intensities. We separately train CNNs, more specifically, VoxCNNs on FA values, MD values, and EPI intensities on 3D DTI scan volumes. In addition, we feed average FA and MD values for each brain region, derived according to the Colin27 brain atlas, into a random forest classifier (RFC). These four (three separately trained VoxCNNs and one RFC) models are first applied in isolation for the above four-class classification problem. Individual classification results are then fused at the decision level using a modulated rank averaging strategy leading to a classification accuracy of 92.6%. Comprehensive experimentation on publicly available ADNI database clearly demonstrates the effectiveness of the proposed solution.

1. Introduction

Alzheimer's disease (AD) is the most common type of dementia accounting for about 60% to 70% of the total number of dementia cases in the World. This deadly disease is caused by the damage and destruction of nerve cells in the brain regions related to memory and its most common symptoms are memory loss and cognitive decline. Worldwide, around 50 million people have dementia, and there are nearly 10 million new cases every year. Globally, the total deaths due to Alzheimer's and other minor types of dementia is 2 382 129 and the total number of prevailing cases are as large as 43 835 665 as reported in the year 2016 (Nichols, Szoeke, Vollset, & Abbasi, 2019). According to data from World population prospects (2019), the number of older persons – those aged 65 years or over – is expected to more than double by 2050 and to more than triple by 2100, rising from about 700 million globally in 2019 to 1.8 billion in 2050 and 2.45 billion in 2100 as shown in Fig. 1. Currently, there is no treatment available to cure AD or to alter its progressive course. However, in order to support and improve the quality of lives of AD patients, the treatment process has a big scope of improvement. Early detection and diagnosis is a major goal for dementia and AD care. AD has generally three major stages of progression-

1. Early Mild Cognitive Impairment (EMCI)
2. Late Mild Cognitive Impairment (LMCI)
3. Alzheimer's Disease (AD)

Along with the above three AD stages, we have one more class — Normal Control (CN), i.e., those who have no symptoms of AD, LMCI or EMCI. So the main task is to classify a brain scan into one of the four major AD classes. Previous efforts of classification was mainly focused on binary classification, i.e., to develop algorithms to classify between any two classes, for example, between AD and CN, between LMCI and CN, between EMCI and CN, and so on. Although it was easy to differentiate among only two types of data, it was still time-consuming to determine the actual class as one would have to eliminate each class by making multiple binary comparisons until the final class is reached. In this work, we for the first time address a direct four-class classification problem related to AD. Clearly, this task is significantly more challenging than single or multiple binary classification(s) as the differences among all four classes is not very distinct.

For the diagnosis of dementia, use of Magnetic Resonance Imaging (MRI) has been quite common. MRI can create a 3D representation of the internal brain structure through magnetic fields and radio waves. It offers the possibility of in-vivo study of pathological brain changes associated with AD. Currently, Diffusion weighted MRI (DWI or DW-MRI) is the standard version of MRI which is used in clinical diagnosis

* Corresponding author.
E-mail addresses: arijitde.etce.rs@jadavpuruniversity.in (A. De), as.chowdhury@jadavpuruniversity.in (A.S. Chowdhury).

https://doi.org/10.1016/j.eswa.2020.114338
Received 20 June 2020; Received in revised form 11 November 2020; Accepted 16 November 2020
Available online 18 November 2020

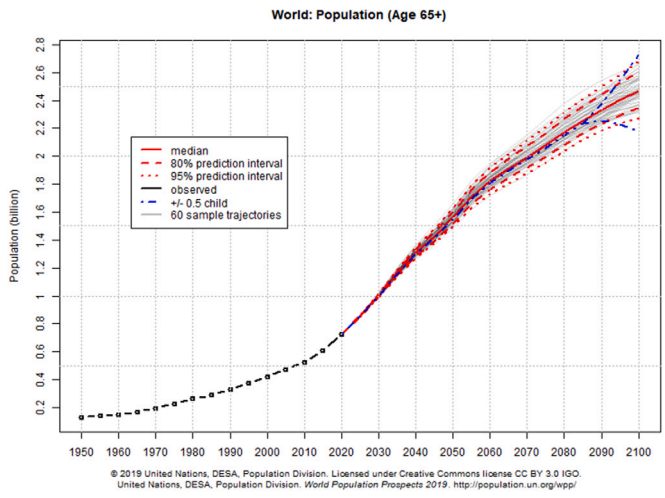

Fig. 1. Estimates and probabilistic projections of population of people aged 65 years and above in the world.

worldwide. It uses the diffusion of water molecules to generate contrast in MR images. Building upon DWI, DTI has gained popularity as it is able to capture the directions of water molecule diffusion thereby providing a lot more data about brain tissue structure than conventional MRI (Le Bihan et al., 2001). This extra information from DTI scans have opened avenues for a wide range of neurological applications, detecting AD is one of them.

Over the past decades, neuroimaging data have been used to characterize AD by exploiting machine learning (ML) methods, offering valuable tools for diagnosis and prognosis (Falahati, Westman, & Simmons, 2014; Rathore, Habes, Iftikhar, Shacklett, & Davatzikos, 2017). Many studies have proposed the use of predefined features (including regional and voxel-based measurements) obtained from image preprocessing pipelines along with different types of classifiers like support vector machines (SVM) and random forests. In recent times, deep learning (DL), a more advanced and complex ML methodology, has created a compelling effect in the domain of medical imaging. The main advantage of DL over ML is that it allows the automatic abstraction of low-to-high level latent feature representations (e.g. lines, dots or edges for low level features, and objects or complex shapes for high level features). Compared with 2D convolutions on slices, 3D convolutions on a whole MRI can capture potential 3D structural information which may be essential for discrimination and has proved to be advantageous on AD vs. Mild Cognitive Impairment (MCI) classification (Gao, Hui, & Tian, 2017). Also, it is important to note that datasets collected in neuroimaging studies are generally very small, compared to the large number of images available in datasets for image classification which are currently used to train neural networks for object classification and detection in 2D image analysis. This leads to relatively lower accuracies as there are not enough training examples for the network to learn the required features. This is mitigated by various techniques like data augmentation, feature fusion and decision fusion.

In this paper, we present a solution for the above four-class AD classification problem by combining DL and ML models on 3D DTI scan volumes. The DTI data is publicly available at the Alzheimer's Disease Neuroimaging Initiative (ADNI) website which has been discussed elaborately in Section 5.1. As we demonstrate in the paper, different types of data existing in DTI can be efficiently utilized when separate DL and ML models are applied. We use 3D-CNN, specifically VoxCNN, as the DL model in this work as it has shown good performance on MRI data (Korolev, Safiullin, Belyaev, & Dodonova, 2017). We employ three VoxCNNs to train three types of 3D volumetric data, namely, Echo Planar Imaging, Fractional Anisotropy and Mean Diffusivity in each DTI scan. For the ML part, we apply a random forest classifier to classify derived metadata in form of region-averaged Fractional Anisotropy and Mean Diffusivity values. Outputs from all the four models, i.e., three VoxCNNs and the random forest classifier are finally combined with a modulated rank averaging decision fusion approach. Our main contributions are now summarized below:

1. We address for the first time a direct four-class classification problem in AD, which is significantly more challenging than the pre-existing single or multiple binary classification(s).
2. We efficiently harness the full potential of DTI scans by applying appropriate DL and ML models in 3D for different types of information existing in them. This detailed exploration of DTI data and that too in 3D is largely absent in the prevalent literature.
3. Finally, outputs of each learning model are combined at the decision level using a modulated rank averaging technique thereby achieving state-of-the-art classification accuracy.

The rest of the paper is organized as follows: in Section 2, we discuss the related works. In Section 3, we provide some basic theoretical foundations regarding DTI. This is followed by detailed description of our proposed method in Section 4. In Section 5, we present the experimental results with detailed analysis. Finally, the paper is concluded in Section 6 with an outline of directions for future research.

2. Related work

Use of medical imaging for solving classification problems has been popular for a long time. Classification is critical for clinicians to ensure proper diagnosis of a disease. Neuro-science has many such applications of classification where neuro-imaging has been extensively used. Some prominent examples include brain tumor classification (Swati et al., 2019), texture classification in ALS disease (Elahi, Kalra, Zinman, Genge, Korngut, & Yang, 2020) and AD classification (Billones, Demetria, Hostallero, & Naval, 2016; Duc et al., 2020).

Extracted bio-markers and processed medical descriptors, together with statistical and conventional ML methods, have been widely used to aid the classification of AD. In Liu et al. (2013), the authors have extracted 83 Regions of Interest (ROI)s from 3D brain MRI and Positron-emission tomography (PET) scans, and proposed a Multifold Bayesian Kernelization (MBK) to diagnose AD. A Support Vector Regression (SVR) based decision support system was developed by Bucholc et al. (2019) which achieved a good accuracy. We use classical ML approach like Random Forest Classifier for only a part of our proposed method as it has limited feature extraction capabilities. However, a part of the DTI data is best handled by the Random Forest Classifier as the data contains numerical features with overlapping and very close values (linearly inseparable) making it less effective for Support Vector Machines. Further, since this part of the data has no spatial correlation, DL methods are not suitable for handling them. Now, we discuss certain DL methods which have been widely used in computer-aided diagnosis literature. Ramaniharan, Manoharan, and Swaminathan (2016) used 2D slices of MRI scans to perform AD classification based on eigen values, parametric and non-parametric classifiers. Raza et al. (2019) used a mix of ML and DL approaches to classify AD. Although, ROI-based and 2D slice-based methods can efficiently extract relevant features and partly reduce the feature dimension, they are too empirical to capture the critical features which are associated with AD classification. In contrast, 3D-CNN can capture more complete spatial features through its space association capability. Cheng, Liu, Fu, and Wang (2017a) extracted a number of 3D patches from the whole MRI and transformed those patches into features by 3D-CNN. Finally, multiple 3D-CNNs were used to combine the features yielding better results for AD classification which inspired us to use DTI data in a similar way. We also took inspiration from the work of Lebedev et al. (2014) where they used Random forest ensembles to detect AD and predict progression from MCI to AD. We used Random forest for training a type of derived data obtained from the DTI scans. Although supervised DL methods work

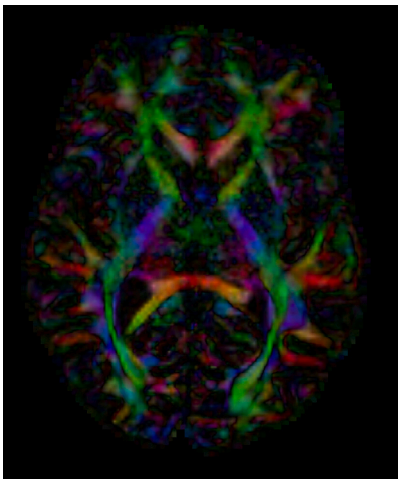

Fig. 2. Axial FA map of a DTI scan from the ADNI dataset.

better than the unsupervised methods, they are highly data dependent. The performance of the model depends on the number of training data available and in most neuroimaging domains, number of data is found to be insufficient to make a model highly accurate and effective. To address this challenge we found DTI data to be useful as each DTI scan produces 3 3D volumetric images leading to a three-fold increase in the data. Furthermore, additional information about each brain region is available which can also be utilized using machine learning methods for improving the classification performance.

As a summary, we can say that AD classification works till date deal with binary classification problems and are mostly restricted to using MRI or PET data. Due to less number of available data and less information in MRI or PET data, satisfactory accuracy for multi-class classification could not be achieved. By exploiting multiple types of data available in DTI, creating appropriate DL and ML models for them and finally fusing the individual outputs at decision level, we have obtained in this work state-of-the-art performance for a four-class AD classification problem.

3. Basics of diffusion tensor imaging

DTI is able to capture white-matter (WM) tracts in the brain while MRI is only limited to gray-matter (GM) visualizations. Alterations in WM diffusivity on DTI are known to be associated with clinical disease severity starting from the pre-clinical stages of AD. The WM integrity on DTI and its importance has been discussed in details in Kantarci et al. (2017). Hence DTI is a preferred choice for studying AD as it can capture both GM and WM information.

We now explain briefly about its mathematical motivation. DTI is a sensitive probe of cellular structure that works by measuring the diffusion of water molecules (Basser, Mattiello, & LeBihan, 1994) inside living tissues. Since the diffusion tensor is a symmetric 3×3 matrix, it can be described by its eigenvalues $(\lambda_1, \lambda_2, \lambda_3)$ and eigenvectors (e_1, e_2, e_3). The eigenvalues and eigenvectors are then used to process scalar indices and, in some studies, tractography analysis (Chun, Li, Xuan, Xun, & Qin, 2005). At each voxel, the eigenvalues represent the magnitude of diffusion and the corresponding eigenvectors reflect the directions of maximal and minimal diffusion.

Generally, each DTI scan contains Echo Planar Imaging (EPI) volume in 3D, and some diffusion tensor information at each voxel which can be used to generate Fractional Anisotropy (FA) and Mean Diffusivity (MD) volumes which are also in 3D. In the dataset that we used, FA and MD volumes were already available along with the EPI volume. The voxel intensities acquired from the MRI scan are stored in the EPI volume. The two main diffusion indices, FA and MD, are based on the eigenvalues, which represent the magnitude of the diffusion process.

3.1. Mean diffusivity

MD is a summary measure of the average diffusion properties of a voxel and is equivalent to the estimated Apparent Diffusion Coefficient (ADC) over three orthogonal directions (Soares, Marques, Alves, & Sousa, 2013). In other words, it is a measure of the mean water diffusion rate. MD values of the voxels differ for brain scans belonging to different classes of Alzheimer's and also differ in normal healthy brains. An increase in MD indicates decreased myelination and loss of axons (Mayo, Mazerolle, Ritchie, Fisk, & Gawryluk, 2017). It can be mathematically represented as-

$$MD = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} = \frac{D_{xx} + D_{yy} + D_{zz}}{3} = \frac{Trace}{3} \tag{1}$$

where D_{xx}, D_{yy}, D_{zz} are the diagonal terms of the diffusion tensor.

3.2. Fractional anisotropy

FA is a normalized measure of the fraction of the tensor's magnitude due to anisotropic diffusion, corresponding to the degree of anisotropic diffusion or directionality and ranges from 0 (isotropic diffusion) to 1 (anisotropic diffusion). Just like MD, decreased FA values are indicative of dementia and Alzheimer's. FA values are rotationally invariant, i.e. they do not have any orientation information. It can be mathematically expressed as-

$$FA = \sqrt{\frac{3}{2}} \sqrt{\frac{(\lambda_1 - D)^2 + (\lambda_2 - D)^2 + (\lambda_3 - D)^2}{{\lambda_1}^2 + {\lambda_2}^2 + {\lambda_3}^2}} \tag{2}$$

where $D = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3}$. FA maps are color coded where a certain direction is represented by a color. In these maps, red color represents left-to-right orientation, green posterior-to-anterior and blue inferior-to-superior diffusion as shown in Fig. 2 which has been generated with the 3D Slicer tool (version 4.10.2) (Fedorov et al., 2012; Pieper, Halle, & Kikinis, 2004).

3.3. Echo planar imaging

EPI is the fastest imaging sequence currently available and has the potential to revolutionize many aspects of MRI technology. It is a rapid MRI technique that is capable of producing tomographic images at video rates. It is a single-shot method having imaging times ~ 100 ms for a 128×128 matrix.

In a single-shot echo planar sequence, the entire range of phase encoding steps, usually up to 128, are acquired in one shot. In multi-shot echo planar imaging, the range of phase steps is equally divided into several shots. Each subsequent echo results in a progressively T2-weighted signal. We use DTI images with this EPI standard in our experiments.

3.4. Significance of fractional anisotropy and mean diffusivity

AD results in the loss of neurons in the brain and this neuronal degeneration can be seen as a loss of both GM and WM. Loss of neurons in certain areas of the brain results in GM atrophy that can be measured on conventional MR images. Increased MD was consistently found in the areas such as the hippocampus, the entorhinal cortex, the parahippocampal gyrus, the temporo-parietal association cortex, and the posterior cingulate gyrus (Oishi, Mielke, Albert, Lyketsos, & Mori, 2011).

Majority of the published AD research has used a cross-sectional design and consistently revealed low FA and high MD in widespread WM regions including the frontal, parietal, and temporal lobes (including hippocampal regions), as well as the corpus callosum and longitudinal association tracts (Mayo et al., 2017). Thus FA and MD are clear biomarkers and their values indicate which stage of AD the patient is in. This can be seen in Fig. 3 where EPI, FA and MD slices of

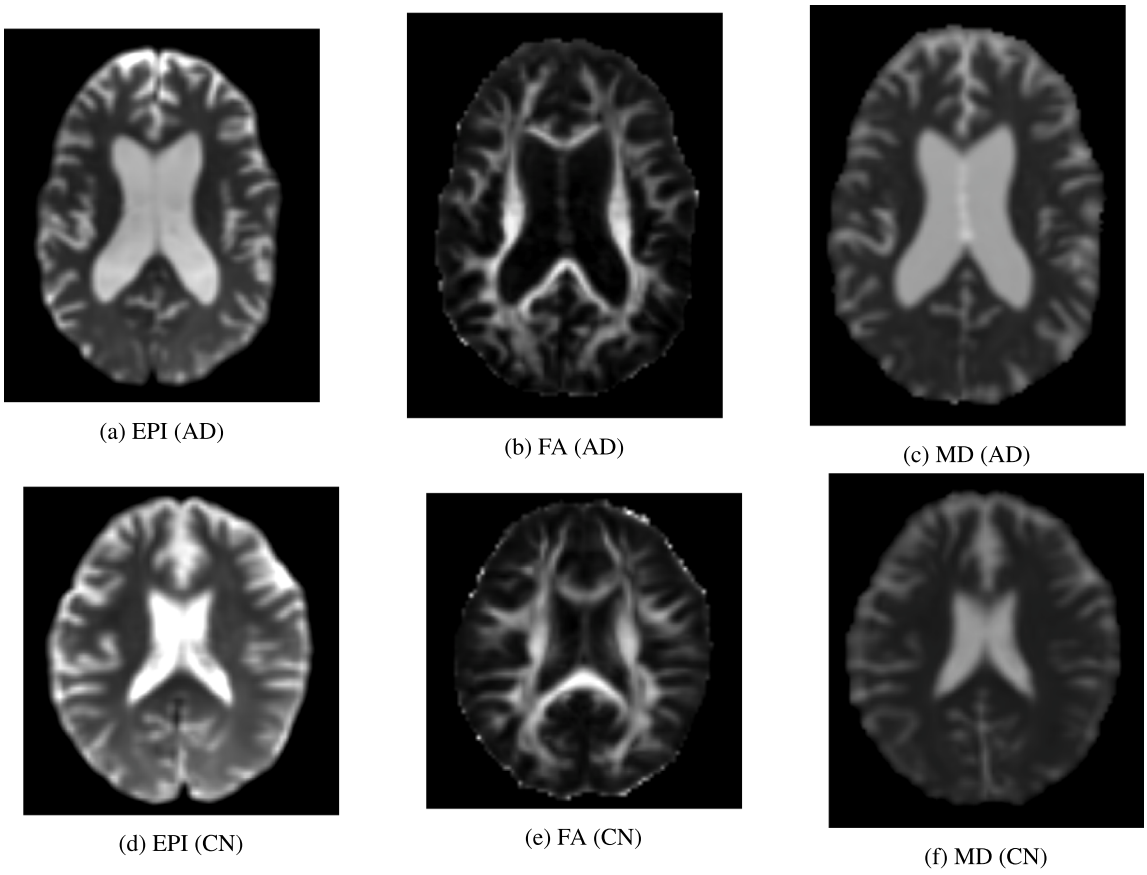

Fig. 3. Axial slices of EPI, FA and MD maps of two patients. (a), (b) (c) denote the slices of AD patient and (e), (f), (g) denotes that of healthy person. EPI and MD shows the WM in black, while FA shows the WM in white. We can see that in AD patient the total WM region in lesser than normal person which is clearly indicated by the larger size of lateral ventricle body in AD patient.

two different patients, one with AD and one healthy, are shown with visible difference in brain WM structures. These unique traits of DTI have motivated us to explore this imaging modality for a four-class classification of AD. So, we decide to apply and combine appropriate classifiers on the DTI data to achieve the above goal.

4. Proposed method

Our solution pipeline consists of three VoxCNN networks and one random forest classifier, each of which outputs a 4×1 probability vector. Each probability vector contains the probabilities for the data to be in the four classes, namely, AD, CN, EMCI and LMCI. The probability vectors are then linearly combined by ranking the models based on their accuracy. In particular, we multiply each vector with a weight which is proportional to the rank of the corresponding model and the difference in accuracy between the corresponding model and the model which is rank-wise immediately next to it. We present a schematic of our solution in Fig. 4. An algorithm showing the details of the fusion strategy is shown in Algorithm 1. We end this section with another algorithm (Algorithm 2) where all the steps of our solution are mentioned. We now provide detailed descriptions of our DL models, ML models and the modulated rank averaging technique for decision level fusion.

4.1. VoxCNN

VoxCNN architecture has four volumetric convolutional blocks for extracting features (with a number of filters increasing from layer to layer), two deconvolutional layers with batch normalization and dropout for regularization and an output with SoftMax nonlinearity for classification. We have kept the network architecture as it is defined in the article (Korolev et al., 2017). The input files are in nifti format, and they are normalized between 0 and 1 for keeping similarity among all models. This data preprocessing part has been done with the nilearn and Scikit-learn packages (Abraham et al., 2014; Pedregosa et al., 2011b). Considering the dataset size, model size and the limitations of GPU memory, we modified the batch iteration process in order to get samples of each class in every batch. The probability of having only one class represented inside a batch for infinite number of samples is $\frac{1}{c^b}$ where c denotes the number of classes and b denotes the batch size. Therefore, for large batch sizes this probability is low. However, for our problem, this value is still high enough to thwart the learning process. Balancing of the samples inside each batch was hence undertaken to obtain stable learning curves.

4.2. Random forest classifier

Random forest classifier (RFC), an ensemble of decision trees (Breiman, 2001) is chosen for its very high accuracy and capability to handle large volume of data. Random forests are well suited for multi-class classification, they do not tend to overfit, can handle outliers well and has fewer number of parameters to tune (Pedregosa et al., 2011a) as compared to other state of the art classifiers like Gradient Boosting Machines (GBM) (Nawar & Mouazen, 2017), XGBoost (Chen & Guestrin, 0000), etc. Also, they are more resilient to noisy data, a feature which can be useful for medical applications (Yang, Wang, Mi, Lin, & Cai, 2009). Each DTI scan comes with averaged FA values of 57 regions and averaged MD values of the same 57 regions, which is derived from the images and is not actually image data. Hence, we can say that we are performing classification based on meta data in tabular format.

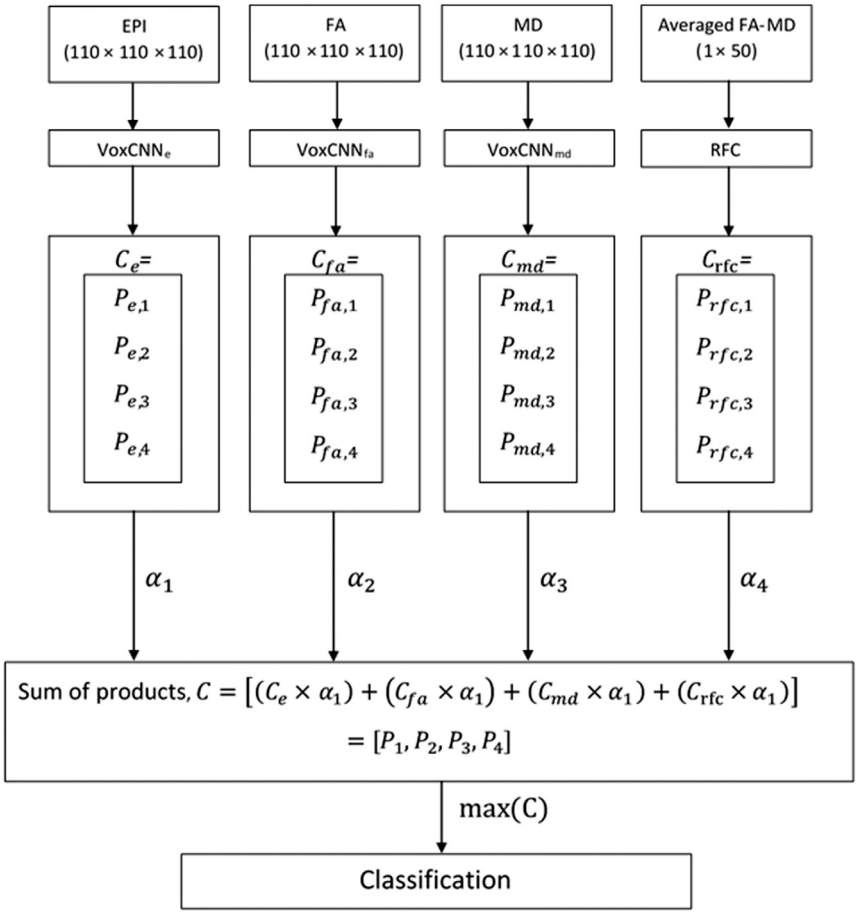

Fig. 4. Solution pipeline architecture: C_e, C_{fa}, C_{md} and C_{rfc} denote the probability vectors of EPI, FA, MD and RFC models respectively, each of which contains the probabilities of the four AD classes; Finally, $P_i s$ denote the probabilities for the four AD classes after applying modulated rank averaging.

Although neural networks can perform classification on tabular data, it is more complex in terms of setting up the model and hyper-parameter tuning, and also it is computationally expensive, whereas RFCs do not need a lot of memory resources and the training can be parallelized in a multi-core processor that greatly speeds up the training process which is a very crucial factor in the field medical imaging where online and in-situ measurements are indispensable (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014). In order to use all available information from each DTI scan and considering the suitability of DL for volumetric image data and RFC for tabular meta data, we have used RFC along with VoxCNN to arrive at the best possible outcome. Each feature value in the meta data is a real number in $[0, 1]$. Thus, RFC is employed (code available at Pedregosa et al., 2011b) to randomly select a subset of features from the total feature set to arrive at a suitable classification decision. RFC votes for the most popular class among the individual trees. The information gain I for the jth node in a decision tree is given by-

$$I = H(S_j) - \sum_{i=L,R} \frac{|S_j^i|}{|S_j|} H(S_j^i) \tag{3}$$

where $H(S_j)$ denotes the entropy of the jth node S_j. The entropy of a node for a discrete set of K class labels $c = 1, 2, \ldots, K$ is given by:

$$H(S_j) = -\sum_{c \in K} p(c) log_2(p(c)) \tag{4}$$

Further, $|S_j|$ denotes the number of training images in the node S_j. So, $|S_j^L|$ and $|S_j^R|$ respectively represent the number of data in the left child and the right child of the node S_j. However, before the data is fed into the RFC, it goes through a set of pre-processing steps described below.

4.2.1. Synthetic minority oversampling technique

Since the data is class imbalanced, Synthetic Minority Oversampling Technique (SMOTE) sampling (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) has been used which generates excellent synthetic samples from the data and re-samples all classes to match the number of samples in the majority class.

Let x_{ij} be the value of the jth variable $(j = 1, \ldots, p)$ for the ith sample $(i = 1, \ldots, n)$ that belongs to class $c, (c = 1, \ldots, K)$. In the present problem, number of classes $K = 4$. Let, $k_c = \frac{n_c}{n}$ is the proportion of samples from class c, n_c is the number of samples in class c and n is the total number of samples. Further, let the sample size of the minority class be denoted by n_{min}. We use capital letters (as X) to denote random variables, lowercase letters (as x) to denote observations and bold letters ($\mathbf{x}$) to indicate a set of variables. The Gaussian distribution with mean μ and standard deviation σ is indicated with $N(\mu, \sigma)$ and the uniform distribution defined on $[0, 1]$ with $U(0, 1)$.

For each sample from the minority class (x), 40 (or $n_{min} - 1$ if $n_{min} \leq 5$) samples from the minority class with the smallest Euclidean distance from the original sample were identified (nearest neighbors) and one of them was randomly chosen (x^R). The new synthetic SMOTE sample was defined as-

$$\boldsymbol{S} = X + u \cdot (\boldsymbol{x}^R - \boldsymbol{x}) \tag{5}$$

where u was randomly chosen from $U(0, 1)$. Note that u is same for all variables but different for each SMOTE sample.

4.2.2. Spatially uniform ReliefF

Spatially Uniform ReliefF (SURF) identifies the nearest neighbors (both hits and misses) based on a distance threshold from the target

instance defined by the average distance between all pairs of instances in the training data. Since there are 114 continuous valued features in the data which are very similar to each other having values between 0 and 1, it is quite challenging to classify the data. Hence, SURF (Greene, Penrod, Kiralis, & Moore, 2009) is used to select the top 50 best features among the 114 features which improves the accuracy of classification.

4.3. Modulated rank averaging based decision level fusion

In traditional majority voting method, the prediction results of majority of the classifiers are used as the final prediction labels. Since, each classifier is independent and the error rates between different classifiers are irrelevant, such strategy can be useful. However, for multi-class classification tasks, this method may not be very effective. Single classifiers perform well on most subjects; but for some subjects which are difficult to classify, the error rates will be increased due to the uncertainty among multiple categories. Let us take the following example with three classifiers. The four-class output probabilities from the SoftMax layer of these three classifiers for {AD, EMCI, LMCI, CN} are respectively given by I: {0.7, 0.05, 0.2, 0.05}, II: {0.3, 0.5, 0.1, 0.1}, III: {0.2, 0.4, 0.3, 0.1}. Based on the majority voting method, the prediction result is EMCI. But this inference is not completely correct, since the prediction result of classifier I is more credible (the winner is predominant) while that of classifiers II and III have more uncertainties (differences between the winner class and other classes are not that high).

In our approach, we use a weight adjusted probability vector fusion technique along with ranking of the classification models based on their individual accuracy. Our approach deals with applying a different weight for each network. Networks that had a lower classification error in the training set will have a larger weight when combining the results for each image from the test set. The algorithm for our proposed method can be seen in Algorithm 1.

Let $\mathcal{R}^n$ be the n-dimensional feature space. Suppose, $\mathcal{X} = [x_1, x_2, \dots, x_n]^T$ be the n-dimensional feature vector, $\mathcal{X} \in \mathcal{R}^n$, $\Omega = [\omega_1, \omega_2, \dots, \omega_K]$ be the set of potential class labels and $C = [C_1, C_2, \dots, C_l]$ be the set of trained models for decision fusion. Given the input pattern $\mathcal{X}$, the output of the ith model is denoted as-

$$C_i(\mathcal{X}) = \left[P_{i,1}(\mathcal{X}), P_{i,2}(\mathcal{X}), \dots, P_{i,m}(\mathcal{X})\right]^T \tag{6}$$

where $C_{i,j}(\mathcal{X})$, $i = 1, 2, \dots, l$ $j = 1, 2, \dots, m$ represents the probability that $\mathcal{X}$ belongs to class ω_j. Basically, $C_i(\mathcal{X})$ denotes the probability vector of model C_i. In our case $l = 4$ as there are 4 models and $K = 4$ as there are four classes. The fused output of l models is constructed as in:

$$C(\mathcal{X}) = F\left[C_1(\mathcal{X}), C_2(\mathcal{X}), \dots ..., C_l(\mathcal{X})\right] \tag{7}$$

where F is the fusion rule described below.

Given some input data, the output probability vectors from all CNNs and one RFC are multiplied by a weight α before the prediction. So, for a given input $\mathcal{X}$, the output probability vector $C(\mathcal{X})$ is given by:

$$C(\mathcal{X}) = \sum_{j=1}^{l} \alpha_j C_j(\mathcal{X}) \tag{8}$$

where $C_j(\mathcal{X})$ is the output the network j for a given input $\mathcal{X}$.

The weight α_j is primarily chosen by rank. It is based on the order of accuracy in the validation set (the test fold in case of K-Fold cross validation) but the relative differences in accuracy of each model are taken into account. Let $R()$ be the function that gives the position of the model based on validation accuracy sorted in increasing order. For example, the model with the largest accuracy will have an $R()$ value of K where K is the number of classes ($K = 4$ in our case), the model with second largest accuracy will have $R()$ value of $(K - 1)$ and so on until the model with the lowest accuracy will have $R() = 1$. Let, $d_1, d_2, \dots, d_{l-1}$ be the differences in validation accuracies of l models which are themselves sorted in ascending order of validation accuracy.

Algorithm 1: MRA($Accuracy$)

Input: Evaluation Accuracy of each model as $Accuracy[A_1, A_2, \dots, A_l]$
Output: Weight for each model as $Weights[\alpha_1, \alpha_2, \dots, \alpha_l]$
/* Store the original accuracies in Acc_{Ranks} so that values in Weights can be in the same order as Accuracy */
$Acc_{Ranks} = Accuracy$
$sum_{factors} = l$
// $l = 4$ in our case
Sort the Accuracy array
$rank \leftarrow l$
foreach *acc in Accuracy* **do** // Iterate through the sorted accuracies
  for $i \leftarrow 1$ **to** l **do**
    if $acc = Acc_{Ranks}[i]$ **then** /* If accuracy value equals i^{th} index, then store the rank at the i^{th} index */
      $Acc_{Ranks}[i] \leftarrow rank$
      $rank = rank - 1$
    end
  end
end
for $j \leftarrow 1$ **to** $l - 1$ **do**
  $d_j \leftarrow Accuracy[l - j + 1] - Accuracy[l - j]$
  $f_j = l - j + (1 - d_j)$
  $sum_{factors} = sum_{factors} + f_j$
end
$Weights[1] \leftarrow \frac{l}{sum_{factors}}$
for $i \leftarrow 2$ **to** l **do**
  $Weights[i] \leftarrow \frac{f_{i-1}}{sum_{factors}}$
end
return $Weights$

In our case, we have four models. Let their validation accuracies, after being trained on full dataset, be denoted by A_1, A_2, A_3 and A_4. The jth difference d_j can be written as-

$$d_j = A_j - A_{j+1} \tag{9}$$

where $j = 1, 2, \dots, l-1$. The main reason for calculating the differences in accuracy of the model is because the differences among the four models (in terms of accuracy) is not uniform. Hence, the differences in weight values for each model should also not be uniform. But, in normal rank based weighting method, each model gets a weight that is just 1 less than the previous model's weight in spite of having accuracy values which are non-uniformly different. Thus, we decide to factor in the individual contributions of each model and penalize them according to the difference in accuracy. Based on this value of d_j, we calculate for each rank value, a factor f_j which is the sum of the rank value $R(A_j)$ and 1 minus the difference d_j. We write,

$$f_j = R(A_j) + (1 - d_j) \tag{10}$$

where $j = 1, 2, \dots, l - 1$. Finally, we calculate the weight α_j by normalizing the factor f_j:

$$\alpha_j = \frac{f_j}{\sum_{j=1}^{l-1} f_j + R_{max}} \tag{11}$$

where, $j = 1, 2, \dots, l - 1$ and R_{max} being the rank of the model with highest accuracy ($R_{max} = 4$ in our case). The weights are then multiplied with the outputs of each model and hence Eq. (8) can be

Algorithm 2: Algorithm of proposed solution

Input : EPI volumes $E = [e_1, e_2, \ldots, e_n]$
FA volumes $F = [fa_1, fa_2, \ldots, fa_n]$
MD volumes $M = [md_1, md_2, \ldots, md_n]$
FA-MD regional averages for 57 brain regions
$V = [v_1, v_2, \ldots, v_n]$ where each $v_i = [p_1, p_2, \ldots, p_{57}, q_1, q_2, \ldots, q_{57}]$, p_j being the averaged FA value and q_j being the averaged MD value for a particular brain region.
n = number of training examples

Output: Probability Vector $C(\mathcal{X}) = [P_1(\mathcal{X}), P_2(\mathcal{X}), P_3(\mathcal{X}), P_4(\mathcal{X})]$ where $P_k(\mathcal{X})$ signifies the probability of input $\mathcal{X}$ belonging to class ω_k

/* Initialize an Accuracy array of size 4 to store the accuracies of each model */

$Accuracy$ = empty

/* Train EPI data in VoxCNN network and store accuracy */

$A_e \leftarrow VoxCNN_e(E)$

/* Train FA data in VoxCNN network and store accuracy */

$A_{fa} \leftarrow VoxCNN_{fa}(F)$

/* Train EPI data in VoxCNN network and store accuracy */

$A_{md} \leftarrow VoxCNN_{md}(M)$

/* Train EPI data in VoxCNN network and store accuracy */

$A_v \leftarrow RFC(V)$

$Accuracy \leftarrow [A_e, A_{fa}, A_{md}, A_v]$

// Get the weights from Algorithm 1 by giving $Accuracy$ as input

$Weights \leftarrow MRA(Accuracy)$

/* For any new input $\mathcal{X} = [e, fa, md, v]$, where e = EPI Volume, fa = FA Volume, md = MD volume and v = Averaged FA-MD values, find the prediction results from each of the four models */

$C_e(e) \leftarrow VoxCNN_e(e)$

$C_{fa}(fa) \leftarrow VoxCNN_{fa}(fa)$

$C_{md}(md) \leftarrow VoxCNN_{md}(md)$

$C_{rfc}(v) \leftarrow RFC(v)$

$C(\mathcal{X}) \leftarrow [(Weights[0] \times C_e(e)) + (Weights[1] \times C_{fa}(fa)) + (Weights[2] \times C_{md}(md)) + (Weights[3] \times C_{rfc}(v))]$

represented as:

$$C(\mathcal{X}) = \sum_{j=1}^{l-1} \alpha_j C_j(\mathcal{X}) + \alpha_L C_L(\mathcal{X}) \tag{12}$$

In the above equation, $\alpha_L = \frac{R_{max}}{\sum_{j=1}^{l-1} f_j + R_{max}}$ with L denoting the model having the highest rank after sorting.

5. Experimental results

In this section, we first discuss data preparation. This is followed by implementation details. We then extensively evaluate our solution including ablation studies and comparisons with external approaches.

5.1. Data preparation

For experimentation, publicly available ADNI database is used (Jack et al., 2008). We take a subset of ADNI DTI data that has been pre-processed with alignment and skull-stripping. Since there are patients that have multiple images taken during a period of time, to minimize possible information "leaks", only the last images were taken for each subject. Also, there were data from two other classes namely MCI and Significant Memory Concern (SMC), but the number of data for these classes were so less that we did not consider those two classes for our experiments. Resulting dataset has 655 images of four classes: 150 of AD patients, 150 of LMCI, 205 of EMCI and 150 of CN. As the number of data for each class is unequal and the total number of data is not enough for training a deep learning model, SMOTE oversampling is used to increase the number of samples for each minority class. Each of the scan contains 3 3D volumes of size $110 \times 110 \times 110$ containing EPI, FA and MD data.

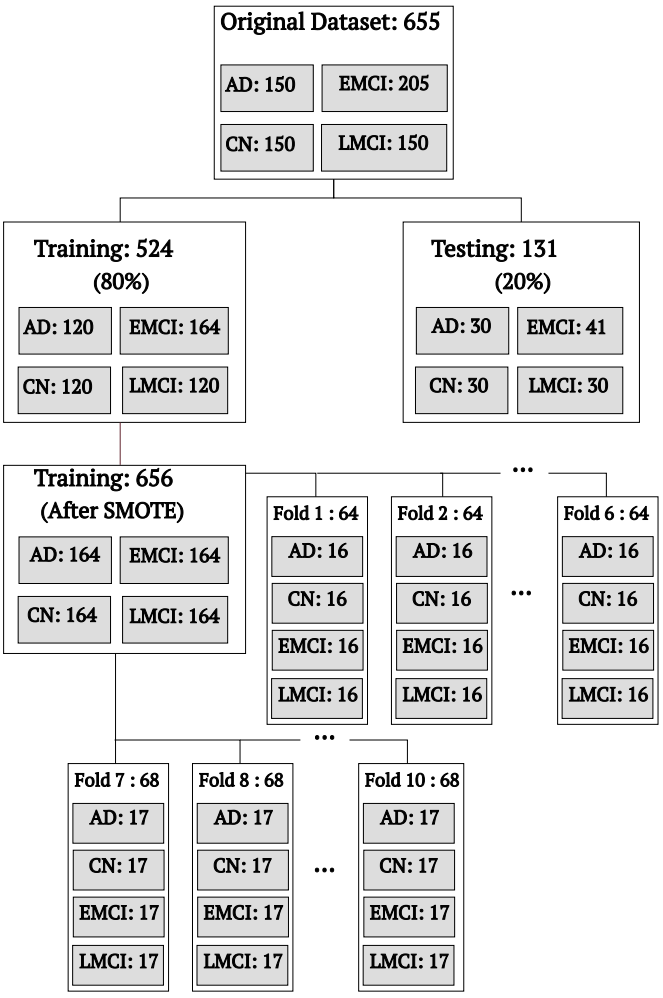

Fig. 5. Dataset division strategy.

We first divided the dataset containing 655 images into training and testing sets keeping in mind the number of images in each class. So we randomly took 80% of images each from AD, CN, EMCI and LMCI and created the training set containing total of 524 images. The remaining 20% from each class were taken to form the testing set totaling 131 images.

We then applied SMOTE oversampling technique to increase the number of samples of each minority class (AD, EMCI and LMCI) to match that of the majority class (CN). As a result each class has a total of 164 images now.

We then split the entire training set into 10 folds for 10-fold cross validation keeping the number of images from each class same. This resulted in 6 folds each containing 64 images(16 from each of AD, CN, EMCI and LMCI) and 4 folds each containing 68 images(17 from each of AD, CN, EMCI, and LMCI). The details of dataset division is explained in Fig. 5.

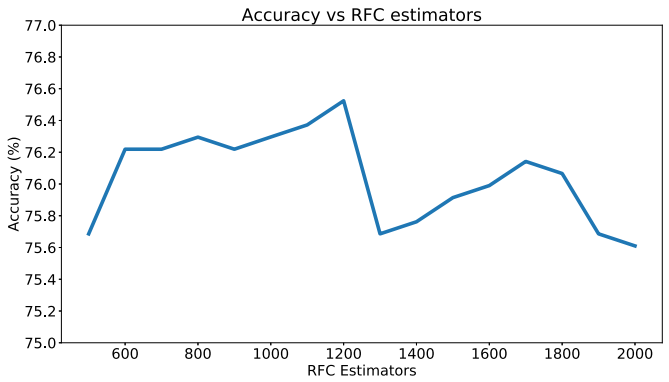

Fig. 6. Graph showing how number of estimators in Random Forest Classification affects accuracy.

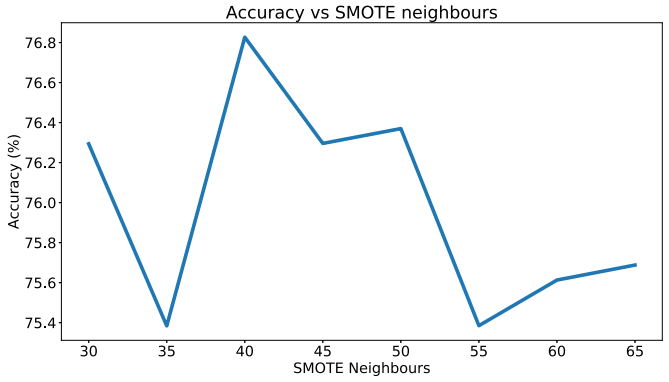

Fig. 7. Graph showing how number of SMOTE neighbors affects accuracy.

Each of the FA and MD data has a corresponding metadata file containing the averaged FA, MD values for each brain region which is extracted to create the 4th dataset containing real numbered values which was fed into RFC after going through the same dataset division process as explained above. Each image file is stored as a 3D tensor of shape $110 \times 110 \times 110$ in Nifti file format (.nii). The EPI image file contains voxel intensity values and the FA and MD files have real numbered values within 0 and 1. For each model, a 10-fold cross validation has been employed to get better approximation of the prediction performance before applying the decision level fusion.

As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this article. A complete listing of ADNI investigators can be found at the ADNI website (Jack et al., 2008).

5.2. Implementation details

After creating training and testing sets for each type of data i.e., EPI, FA, MD and regional FA-MD values as explained in Section 5.1, we trained the training set with 10-fold cross validation. It performed better than 5-fold cross validation as the number of training data in each fold significantly increased. We fed the accuracy value of each model (three CNNs for EPI, FA, MD and one RFC for FA-MD averaged values) to compute the modulated weights which were used in our novel modulated rank averaging method. Then, for each data in the testing set, classification output of the individual models was multiplied with the corresponding weights previously calculated and a weighted averaging was done to arrive at the final classification.

We first furnish the details of VoxCNN training parameters. This is followed by elaborate discussions on setting of various parameters pertaining to RFC.

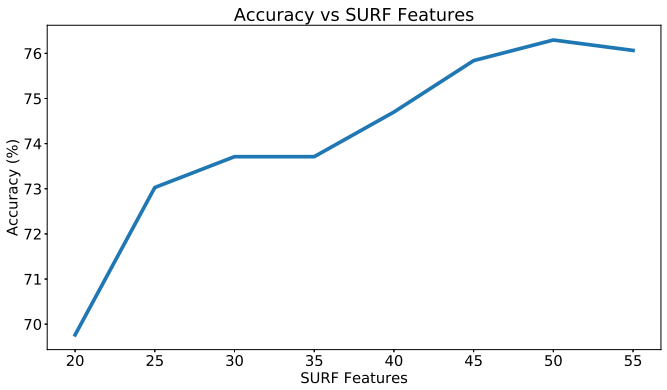

Fig. 8. Graph showing how number of SURF features affects accuracy.

Table 1
Individual evaluation accuracy of EPI, FA, MD and RFC models.

Model	Accuracy (%)
EPI	87.4
FA	76.4
MD	84.5
RFC	69.4

5.2.1. VoxCNN training parameters

We train the network using AdaM optimizer with learning rate of 27×10^{-6} and batch size of 32 for 80 epochs for each fold to get the perfect class separation on the fold training set and stabilize the performance metrics on fold validation set. The final accuracy is calculated on the separately kept test set containing 131 samples. We keep the training parameters same for each model (EPI, FA, and MD) in order to avoid conflict while combining the outputs of the models so that the decision level fusion is unbiased.

5.2.2. Random forest classifier parameters

The classifier is also tuned on the training set by experimenting with the number of trees/estimators. From Fig. 6, it is evident that setting the number of trees to 1200 yields best results. The function to measure split quality is set to 'entropy' as 'gini' function often leads to poorer results than 'entropy'.

5.2.3. Synthetic minority oversampling technique parameters

SMOTE oversampling (see Section 4.2.1) is used to create synthetic training samples which helps in balancing the number of training data for each class. After experimenting with the number of neighbors in SMOTE on the training set, keeping other parameters constant, we have found that the best results are obtained when the number of neighbors is set to 40. This is evident in the Accuracy vs. SMOTE neighbors graph in Fig. 7.

5.2.4. Spatially uniform ReliefF parameters

SURF selects the best features from numerous features to find a balance between computation time and accuracy. In our case, we have selected 50 features. By experimenting with the number of features in a small dataset keeping number of RFC estimators and SMOTE neighbors constant, it was found that 50 features gives the best accuracy value as evident in Fig. 8.

5.3. Individual model performance

The evaluation accuracy and Area under curve (AUC) of Receiver Operating Characteristics (ROC) for the individual models evaluated with the testing set are shown in Table 1. It can be seen that the RFC has performed worst while the EPI VoxCNN model has performed best. The models in decreasing order of performance are EPI, MD, FA and RFC.

Table 2
Comparison of evaluation accuracy among different feature fusion methods along with the accuracy (in %) for five types of classifications (AD vs. CN, AD vs. MCI, MCI vs. CN, AD vs. MCI vs. CN, AD vs. EMCI vs. LMCI vs. CN).

Approach	Accuracy (%) with Classes				
	AD vs. CN	MCI vs. CN	AD vs. MCI	AD vs. CN vs. MCI	AD vs. CN vs. EMCI vs. LMCI
Liu et al. (2015)	91.4	82.1			
Shi, Chen, Zhang, Smith, and Liu (2017)	91.95	83.72			
Lei, Chen, Ni, and Wang (2016)	96.93	82.75			
Madusanka, Choi, So, and Choi (2019)	86.61	82.05	78.95		
Xiao et al. (2017)	85.71	86.11	79.44	75	
Our feature fusion method					79.34
Our decision fusion method					**92.6**

Table 3
Differences in evaluation accuracy at 95% confidence level.

Difference	Confidence Interval (95%)	
	Min.	Max.
Ours vs. Rank averaging	3.298×10^{-4}	167.854×10^{-4}
Ours vs. Majority voting	294.871×10^{-4}	528.495×10^{-4}

5.4. Combined model performance

The models in various combinations are fused at decision level using three different fusion techniques as mentioned below.

5.4.1. Majority voting method

Since each model gives a single class decision, we take that as a vote for that particular class and consider the final output as the class that gets the maximum number of votes. In this way we found that the accuracy has significantly increased from that of the individual models.

5.4.2. Rank averaging method

In this scheme, the output probabilities of each model are simply multiplied by the rank of that model sorted according to descending accuracy (Frazão & Alexandre, 2014). This method is better than the majority voting method by 3%.

5.4.3. Modulated rank averaging method

Our proposed model performed even better than the Rank Averaging method. As the accuracy of each model sorted in increasing order is not uniform (see Table 1), i.e., their relative differences are not same, the weight assigned to them should also not be uniform. We modified the weights of Rank Averaging method to factor in the differences of the model accuracies (see Eq. (10)). The ROC curves for each of the combined models are plotted in Fig. 9 where we can see that modulated rank average performs slightly better than the rank averaging model. This better performance is quantitatively corroborated by the accuracy and AUC values in Table 4. The per-class metrics are also shown in Table 5.

To demonstrate statistical significance of the improvement in results, we have computed confidence interval (with a confidence of 95%) for the difference between evaluation accuracy values of the proposed modulated rank averaging method and other two decision fusion approaches, i.e., rank averaging and majority voting. If the confidence interval includes zero, the difference is not significant at that confidence level. If the confidence interval does not include zero, then the sign of the differences in the accuracy values indicates which alternative is better (Raz, 1992). Since, the confidence intervals (with a confidence of 95%) do not include zero in either of the cases, we can say that the results presented in Table 3 confirm that the proposed decision level fusion approach yields statistically significant improvements over the other two existing decision fusion strategies.

Table 4
Evaluation accuracy and Area under Curve (AUC) of the fusion methods with all four model combinations.

Fusion approach	Accuracy (%)	AUC
Majority voting	88.7	0.923
Rank averaging	91.8	0.957
Modulated rank averaging	92.6	0.962

Table 5
Per class metrics containing the precision, recall and f1-score for each disease class.

Class	Precision	Recall	f1-score
AD	0.96	0.92	0.94
CN	1.00	0.97	0.98
EMCI	0.88	0.88	0.88
LMCI	0.85	0.92	0.88

Table 6
Accuracy and Area Under Curve (AUC) for all combinations using Modulated Rank Averaging method.

Combinations	Accuracy (%)	AUC
All	92.6	0.962
EPI + FA + MD	91.2	0.955
EPI + FA + RFC	86.48	0.913
EPI + MD + RFC	88.87	0.934
FA + MD + RFC	85.32	0.908
EPI + FA	85.26	0.901
EPI + MD	90.42	0.943
EPI + RFC	84.3	0.886
FA + MD	89.37	0.928
FA + RFC	72.74	0.815
MD + RFC	85.29	0.894

Table 7
Comparison of evaluation accuracy with other state-of-the-art approaches.

Approach	Modalities	Number of classes	Accuracy (%)
Bi et al. (2019)	MRI	3	92.5
Billones et al. (2016)	MRI	3	91.85
Vu, Ho, Yang, Kim, and Song (2018)	MRI + PET	3	91.13
Cheng, Liu, Fu, and Wang (2017b)	MRI	3	87.15
Duc et al. (2020)	MRI	3	87.15
Gunawardena, Rajapakse, and Kodikara (2017)	MRI	3	84.4
Our Method	**DTI**	**4**	**92.6**

5.5. Ablation study of the models

Keeping the best fusion strategy, i.e., modulated rank averaging based decision fusion, we now show the utility of the four models. An extensive ablation study is carried out in that regard. Within that study, we compare the performance with all 4 models (EPI, FA, MD and RFC), all possible combinations of any 3 models at a time as well as any 2 models at a time. The accuracy and AUC for each combination are shown in Table 6. This table shows that all four models are necessary

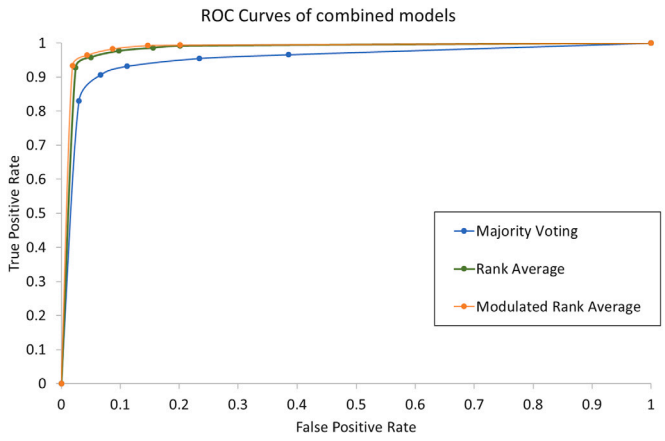

Fig. 9. ROC curves of the combined models (Majority Voting, Rank Averaging and Modulated Rank Averaging).

to yield best result. It is also interesting to note from a comparison of Tables 1 and 6 that inclusion of RFC, a classical ML tool, can improve the overall accuracy when combined with DL models on EPI or MD values. Furthermore, when clubbed with FA + MD, RFC can once again increase the overall accuracy. So, we show how effective can be the combination of DL and ML methods for a given problem.

5.6. Comparison with feature fusion strategies

We also ran an experiment to apply feature fusion instead of decision level fusion. Features were extracted from the individual DL models (EPI, FA and MD), from the penultimate Fully Connected (FC) layer (dense_1). It consisted of 64 dimensions for each of the EPI, FA and MD models. Concatenating all the features from the DL models along with the input features of the RFC model 4.2, a total of 242 features were obtained (64 + 64 + 64 + 50). All these features were then trained in a RFC with 1200 trees. This approach of classification gave a mean accuracy of 79.34% after doing 10-fold cross-validation. We also compare our proposed modulated rank averaging method with some state-of-the-art feature fusion approaches in Table 2. All the above approaches perform either a binary or a three-way classification. It is interesting to note that our proposed approached has achieved 92.6% accuracy in a direct 4-class classification problem. Hence, we confirm that for this classification problem, decision level fusion works better than feature fusion.

5.7. Comparison with other approaches

We compare our method with seven state-of-the-art methods but no results are available for 4-class AD classification on ADNI dataset using DTI data. So, we show comparisons from the problem perspective and compare our results with the papers which have addressed the same AD classification problem but from differing modalities and fewer classes. All the seven methods showed classification among AD, MCI and CN, i.e., three classes. But we have segregated the MCI class into more detailed EMCI and LMCI classes and thus made a four class classification. Classification between EMCI and LMCI is particularly difficult as their differences are not very significant. Also, none of them used DTI modality. They used MRI or a combination of MRI and PET to achieve their goal. The comparisons, shown in Table 7, clearly points to the supremacy of our approach.

6. Conclusion

Classification of AD is an important part of dementia diagnosis, especially for the aging population. This classification is still mostly done manually by the neurologists with the help of brain scans. Existing methods for automated classification are restricted to either a two-class or a three-class problem from MRI. In this paper, we for the first time introduced an automated solution for four class classification of AD using an efficient processing of 3D DTI data. We first trained separate deep learning (VoxCNN) and machine learning (Random Forest) models on different pieces of information in DTI scan volumes. Using a modulated rank averaging decision fusion strategy, we then combined the individual classification results. Comprehensive experimentation, including comparisons and ablation studies, on the publicly available ADNI database clearly demonstrate the effectiveness of the formulation.

In future, we very much look forward to apply our model in the actual clinical practice. Note that during the training, the proposed model (both DL and ML) is tuned with the weights required for the task at hand and importantly, these weights can be saved for future prediction purposes. The model with saved weights does not require a lot of space (approx. 100–200 MBs) and can hence be loaded in even mobile devices which can predict and classify new and previously unseen data. In clinical practice, a software can be developed which takes in input in the form of images and loads the saved model from permanent storage and gives the prediction results without requiring to train again with lots of data. For instance, Yao et al. (2016) have developed a screening system that can detect individuals infected with influenza from three clinical features (heart rate, respiration rate, and facial temperature). Their system is especially interesting as it uses contactless technologies that make it particularly suitable for clinical application with contagious patients (Yao et al., 2016). More recently, Dagdanpurev et al. (2019) have developed a similar screening system using the same three clinical features. Their system uses a random tree algorithm to predict the patient's infection status and is easily understood by physicians because it can be expressed as a flow chart (Dagdanpurev et al., 2019). So, as discussed above, we strongly believe that it should certainly be possible to apply the proposed approach, which couples DL and ML in the actual clinical practice. We also plan to extend the present model to accurately solve a six class AD classification problem by incorporating two more classes, namely, Mild Cognitive Impairment and Significant Memory Concern. Another direction of future research will be to predict the next stage of AD progression with significant accuracy.

CRediT authorship contribution statement

Arijit De: Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper. **Ananda S. Chowdhury:** Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Arijit De was supported through the Tata Consultancy Services (TCS) Research Scholar Program (RSP) of Tata Consultancy Services Pvt. Ltd.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14. http://dx.doi.org/10.3389/fninf.2014.00014.

Basser, P. J., Mattiello, J., & LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, *66*(1), 259–267. http://dx.doi.org/10.1016/S0006-3495(94)80775-1.

Bi, X., Li, S., Xiao, B., Li, Y., Wang, G., & Ma, X. (2019). Computer aided Alzheimer's disease diagnosis by an unsupervised deep learning technology. *Neurocomputing*, http://dx.doi.org/10.1016/j.neucom.2018.11.111.

Billones, C. D., Demetria, O. J. L. D., Hostallero, D. E. D., & Naval, P. C. (2016). DemNet: A convolutional neural network for the detection of Alzheimer's disease and mild cognitive impairment. In *2016 IEEE region 10 conference (TENCON)* (pp. 3724–3727).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., et al. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert Systems with Applications*, *130*, 157–171. http://dx.doi.org/10.1016/j.eswa.2019.04.022.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. http://dx.doi.org/10.1613/jair.953.

Chen, T., & Guestrin, C. XGBoost: A scalable tree boosting system. arXiv 2016. arXiv preprint arXiv:1603.02754.

Cheng, D., Liu, M., Fu, J., & Wang, Y. (2017a). Classification of MR brain images by combination of multi-CNNs for AD diagnosis. In *Proc. SPIE, Vol. 10420*. http://dx.doi.org/10.1117/12.2281808.

Cheng, D., Liu, M., Fu, J., & Wang, Y. (2017b). Classification of MR brain images by combination of multi-CNNs for AD diagnosis. In *Ninth international conference on digital image processing (ICDIP 2017), Vol. 10420*. International Society for Optics and Photonics, Article 1042042. http://dx.doi.org/10.1117/12.2281808.

Chun, S. Y., Li, K. C., Xuan, Y., Xun, M. J., & Qin, W. (2005). Diffusion tensor tractography in patients with cerebral tumors: A helpful technique for neurosurgical planning and postoperative assessment. *European Journal of Radiology*, *56*(2), 197–204. http://dx.doi.org/10.1016/j.ejrad.2005.04.010.

Dagdanpurev, S., Abe, S., Sun, G., Nishimura, H., Choimaa, L., Hakozaki, Y., et al. (2019). A novel machine-learning-based infection screening system via 2013–2017 seasonal influenza patients' vital signs as training datasets. *Journal of Infection*, *78*(5), 409–421.

Duc, N. T., Ryu, S., Qureshi, M. N. I., Choi, M., Lee, K. H., & Lee, B. (2020). 3D-deep learning based automatic diagnosis of alzheimer's disease with joint MMSE prediction using resting-state fMRI. *Neuroinformatics*, *18*(1), 71–86. http://dx.doi.org/10.1007/s12021-019-09419-w.

Elahi, G. M. E., Kalra, S., Zinman, L., Genge, A., Korngut, L., & Yang, Y.-H. (2020). Texture classification of MR images of the brain in ALS using M-cohog: A multi-center study. *Computerized Medical Imaging and Graphics*, *79*, Article 101659. http://dx.doi.org/10.1016/j.compmedimag.2019.101659.

Falahati, F., Westman, E., & Simmons, A. (2014). Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. *Journal of Alzheimer's Disease*, *41*, 685–708. http://dx.doi.org/10.3233/JAD-131928.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., et al. (2012). 3D slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, *30*(9), 1323–1341. http://dx.doi.org/10.1016/j.mri.2012.05.001.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*(90), 3133–3181.

Frazão, X., & Alexandre, L. A. (2014). Weighted convolutional neural network ensemble. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8827*, 674–681. http://dx.doi.org/10.1007/978-3-319-12568-8_82.

Gao, X. W., Hui, R., & Tian, Z. (2017). Classification of CT brain images based on deep learning networks. *Computer Methods and Programs in Biomedicine*, *138*, 49–56. http://dx.doi.org/10.1016/j.cmpb.2016.10.007.

Greene, C. S., Penrod, N. M., Kiralis, J., & Moore, J. H. (2009). Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Mining*, *2*(1), 5. http://dx.doi.org/10.1186/1756-0381-2-5.

Gunawardena, K., Rajapakse, R., & Kodikara, N. (2017). Applying convolutional neural networks for pre-detection of alzheimer's disease from structural MRI data. In *2017 24th international conference on mechatronics and machine vision in practice (M2VIP)* (pp. 1–7). IEEE, http://dx.doi.org/10.1109/M2VIP.2017.8211486.

Jack, C. R., Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. http://dx.doi.org/10.1002/jmri.21049.

Kantarci, K., Murray, M. E., Schwarz, C. G., Reid, R. I., Przybelski, S. A., Lesnick, T., et al. (2017). White-matter integrity on DTI and the pathologic staging of Alzheimer's disease. *Neurobiology of Aging*, *56*, 172–179. http://dx.doi.org/10.1016/j.neurobiolaging.2017.04.024.

Korolev, S., Safiullin, A., Belyaev, M., & Dodonova, Y. (2017). Residual and plain convolutional neural networks for 3D brain MRI classification. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)* (pp. 835–838). http://dx.doi.org/10.1109/ISBI.2017.7950647.

Le Bihan, D., Mangin, J. F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., et al. (2001). Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, *13*(4), 534–546. http://dx.doi.org/10.1002/jmri.1076.

Lebedev, A., Westman, E., Van Westen, G., Kramberger, M., Lundervold, A., Aarsland, D., et al. (2014). Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical*, *6*, 115–125. http://dx.doi.org/10.1016/j.nicl.2014.08.023.

Lei, B., Chen, S., Ni, D., & Wang, T. (2016). Discriminative learning for Alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion. *Frontiers in Aging Neuroscience*, *8*, 77. http://dx.doi.org/10.3389/fnagi.2016.00077.

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., et al. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, *62*(4), 1132–1140. http://dx.doi.org/10.1109/TBME.2014.2372011.

Liu, S., Song, Y., Cai, W., Pujol, S., Kikinis, R., Wang, X., et al. (2013). Multifold Bayesian kernelization in Alzheimer's diagnosis. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, N. Navab (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2013, 16(Pt 2)* (pp. 303–310). Berlin, Heidelberg: Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-642-40763-5_38.

Madusanka, N., Choi, H.-K., So, J.-H., & Choi, B.-K. (2019). Alzheimer's disease classification based on multi-feature fusion. *Current Medical Imaging*, *15*(2), 161–169. http://dx.doi.org/10.2174/1573405614666181012102626.

Mayo, C. D., Mazerolle, E. L., Ritchie, L., Fisk, J. D., & Gawryluk, J. R. (2017). Longitudinal changes in microstructural white matter metrics in Alzheimer's disease. *NeuroImage: Clinical*, *13*, 330–338. http://dx.doi.org/10.1016/j.nicl.2016.12.012.

Nawar, S., & Mouazen, A. M. (2017). Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. *Sensors*, *17*(10), http://dx.doi.org/10.3390/s17102428.

Nichols, E., Szoeke, C. E., Vollset, S. E., & Abbasi (2019). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *18*(1), 88–106. http://dx.doi.org/10.1016/S1474-4422(18)30403-4.

Oishi, K., Mielke, M. M., Albert, M., Lyketsos, C. G., & Mori, S. (2011). DTI analyses and clinical applications in Alzheimer's disease. *Journal of Alzheimer's Disease*, *26*(s3), 287–296. http://dx.doi.org/10.3233/JAD-2011-0007.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011a). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830, URL: https://dl.acm.org/doi/10.5555/1953048.2078195.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011b). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pieper, S., Halle, M., & Kikinis, R. (2004). 3D slicer. In *2004 2nd IEEE international symposium on biomedical imaging: Nano to macro (IEEE Cat No. 04EX821), Vol. 1* (pp. 632–635). http://dx.doi.org/10.1109/ISBI.2004.1398617.

Ramaniharan, A. K., Manoharan, S. C., & Swaminathan, R. (2016). Laplace Beltrami eigen value based classification of normal and Alzheimer MR images using parametric and non-parametric classifiers. *Expert Systems with Applications*, *59*, 208–216. http://dx.doi.org/10.1016/j.eswa.2016.04.029.

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., & Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, *155*(July), 530–548. http://dx.doi.org/10.1016/j.neuroimage.2017.03.057.

Raz, T. (1992). The art of computer systems performance analysis: Techniques for experimental design, measurement, simulation, and modeling (Raj Jain). *SIAM Review*, *34*(3), 518–519. http://dx.doi.org/10.1137/1034111.

Raza, M., Awais, M., Ellahi, W., Aslam, N., Nguyen, H., & Le-Minh, H. (2019). Diagnosis and monitoring of Alzheimer's patients using classical and deep learning techniques. *Expert Systems with Applications*, *136*, 353–364. http://dx.doi.org/10.1016/j.eswa.2019.06.038.

Shi, B., Chen, Y., Zhang, P., Smith, C. D., & Liu, J. (2017). Nonlinear feature transformation and deep fusion for Alzheimer's disease staging analysis. *Pattern Recognition*, *63*, 487–498. http://dx.doi.org/10.1016/j.patcog.2016.09.032.

Soares, J. M., Marques, P., Alves, V., & Sousa, N. (2013). A hitchhiker's guide to diffusion tensor imaging. *Frontiers in Neuroscience*, *7*(7 MAR), 1–14. http://dx.doi.org/10.3389/fnins.2013.00031.

Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S., et al. (2019). Brain tumor classification for MR images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, *75*, 34–46. http://dx.doi.org/10.1016/j.compmedimag.2019.05.001.

Vu, T.-D., Ho, N.-H., Yang, H.-J., Kim, J., & Song, H.-C. (2018). Non-white matter tissue extraction and deep convolutional neural network for Alzheimer's disease detection. *Soft Computing*, *22*(20), 6825–6833. http://dx.doi.org/10.1007/s00500-018-3421-5.

World population prospects 2019. (2019). URL: https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf Accessed: 2020-05-16.

Xiao, Z., Ding, Y., Lan, T., Zhang, C., Luo, C., & Qin, Z. (2017). Brain MR image classification for alzheimer's disease diagnosis based on multifeature fusion. *Computational and Mathematical Methods in Medicine*, *2017*, http://dx.doi.org/10.1155/2017/1952373.

Yang, F., Wang, H.-z., Mi, H., Lin, C.-d., & Cai, W.-w. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinformatics*, *10 Suppl 1*(Suppl 1), S22. http://dx.doi.org/10.1186/1471-2105-10-S1-S22.

Yao, Y., Sun, G., Matsui, T., Hakozaki, Y., van Waasen, S., & Schiek, M. (2016). Multiple vital-sign-based infection screening outperforms thermography independent of the classification algorithm. *IEEE Transactions on Biomedical Engineering*, *63*(5), 1025–1033. http://dx.doi.org/10.1109/TBME.2015.2479716.