# Combining of Multiple Deep Networks via Ensemble Generalization Loss, Based on MRI Images, for Alzheimer's Disease Classification

Jae Young Choi, *Member, IEEE*, and Bumshik Lee, *Member, IEEE*

***Abstract*—This letter proposes a novel way of using an ensemble of multiple deep convolutional neural networks (DCNNs) for Alzheimer's disease classification, based on magnetic resonance imaging (MRI) images. To create this ensemble of DCNNs, we propose to combine the use of multiple MRI projections (as input) with that of different DCNN architectures to increase the deep ensemble diversity. In particular, to find the optimal fusion weights of the DCNN members, we designed a novel deep ensemble generalization loss, which accounts for interaction and cooperation during the optimal weight search. The optimization framework, equipped with our ensemble generalization loss, was formulated and solved using the sequential quadratic programming. Through this method, we achieved optimal DCNN fusion weights (i.e., a high generalization performance). The experimental results showed that our proposed DCNN ensemble outperforms current deep learning-based methods: it is able to produce state-of-the-art results on the Alzheimer's disease neuroimaging initiative (ADNI) dataset.**

***Index Terms*—Alzheimer's disease classification, ensemble deep learning, generalization loss.**

## I. Introduction

ALZHEIMER'S disease (AD), characterized by a progressive impairment of the cognitive and memory functions, is one of the greatest health threats among elderly aged 65 years or older [1]–[4]. It has been shown in [29]–[31] that computer-aided diagnosis (CAD) systems based on magnetic resonance imaging (MRI) can be very effective for the early diagnosis and monitoring of AD. For this reason, the development and improvement of CAD algorithms is of great interest. In a general CAD system for AD diagnosis, a classifier is trained to distinguish between different groups of subjects (e.g., AD, mild cognitive impairment (MCI), and normal control (NC) categories) [2]–[4]. Recent trends in the diagnosis of AD based on MRI images include the use of deep learning based methods; among these, the most promising are the deep convolutional neural networks (DCNNs) [2]–[8], [15]–[19]. However, most of the recent works on the use of DCNNs for AD classification have been limited to the use of an individual (single) DCNN model. The main limitation for AD classification is the small amount of available training data: the acquisition of data through a sufficient number of image samples is difficult and quality annotation is costly [1]–[4]. A small amount of data renders especially difficult for the achievement of a good generalized (test) AD classification performance for a single DCNN. In fact, *a DCNN based on a small dataset does not provide a high expressive power* [5], [33], *limiting the identification of highly complex AD patterns and their arbitrary appearances.*

To overcome this limitation, recent work on deep learning have suggested the use of an ensemble of DCNNs (here called "DCNN ensemble"), which performs classifications by integrating multiple DCNNs [21], [33]. However, very few research works have proposed the use of a DCNN ensemble for AD classification. The authors in [6] constructed three different DCNNs, one for each brain projection (i.e., cross-sectional planes) and combined the outputs of the DCNNs by majority voting. In [7], an ensemble of deep belief networks was composed and the final prediction for the classification of AD was determined by a voting scheme. In [8], three shallow DCNNs (trained with sagittal, axial, and coronal projections) were fused via ensemble averaging for three binary classification tasks (i.e., AD vs. NC, NC vs. MCI, and MCI vs. AD). The aforementioned work shows that AD classification can be improved by applying the voting or averaging methods to the outputs of multiple DCNNs. However, *a major drawback of this previous work is that the same weight is assigned to each of the DCNNs in the ensemble for the aggregation of the results via the average or majority voting fusion.* This naïve unweighted fusion is **not data-adaptive** and, thus, **vulnerable to "bad" DCNN ensemble members** (i.e., the weaker DCNNs in the ensemble). In fact, this approach *does not consider the interaction among the DCNN members* (in order to improve the generalized classification performance) during their fusion.

The main focus of our work is to present a novel algorithm for the weighting of each DCNN member of the ensemble, and that is optimized for AD classification. In this study, the determination of the weights was formulated as an optimization problem. The goal was to determine an optimal set of weights

Manuscript received August 8, 2019; revised December 2, 2019; accepted December 18, 2019. Date of publication January 6, 2020; date of current version January 31, 2020. This work was supported by the Hankuk University of Foreign Studies Research Fund. This workwas also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1A4A1029769). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sumohana Channappayya. *(Corresponding author: Bumshik Lee.)*

J. Y. Choi is with the Division of Computer and Electronic Systems Engineering, Hankuk University of Foreign Studies, Yongin 17305, South Korea (e-mail: jychoi@hufs.ac.kr).

B. Lee is with the Department of Information and Communications Engineering, Chosun University, Gwangju 61452, South Korea (e-mail: bslee@chosun.ac.kr).

Digital Object Identifier 10.1109/LSP.2020.2964161

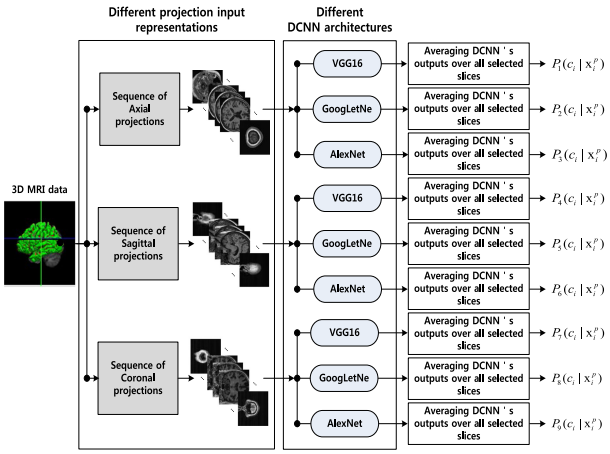


Fig. 1. Proposed DCNN ensemble creation via combined use of different MRI projections (as inputs to DCNNs) and different DCNN architectures.

that would attain *the best generalized performance for the whole ensemble*. A novel *deep ensemble generalization loss* (EGL) was devised to facilitate the interaction and the cooperation among the DCNNs during the weight optimization, in order to improve the generalized classification. The EGL can be very useful for the *combination of the complementary strengths of individual DCNN members in data-adaptive way*. To our knowledge, methods for the determination of optimal weights for the DCNN ensemble have not been studied in detail previously, nor exploited for AD classification. Our approach also differs from previous approaches [27], [28] that determines the weight of each DCNN independently and separately, based on the accuracy of each DCNN's decision.

## II. Creation of a Deep Ensemble Based on Multiple MRI Projections and Network Architectures

Individual DCNN architectures have different capabilities in terms of image data recognition, which can help increase the diversity [22] of the proposed DCNN ensemble. In light of this fact, we used three popular DCNN architectures: VGG-16 [24], GoogLeNet [25], and AlexNet [14]. Moreover, we used three different kinds of DCNN architectures and projections (i.e., the sagittal, coronal, and axial projections for each MRI data) as input representations, obtaining an ensemble of nine DCNN members. Each trained DCNN member was trained using one of the three projections, as shown in Fig. 1. The goal of the proposed ensemble construction was to increase the diversity [9] of the DCNN members via a *combined use of different MRI projections and deep network architectures*, which could provide complementary information and enhance the classification.

Without loss of generality, a set of *M* individual DCNNs were available in the ensemble. The last stage of each DCNN included a softmax layer, followed by a negative log-likelihood loss defined as:

$$-\frac{1}{N}\sum_{i=1}^{N}\log P_k(c_i|\mathrm{x}_i^p), p \in \{\text{axial, coronal, sagittal}\} \quad (1)$$

where *N* is the total number of training data, $\mathrm{x}_i^p$ is the *p*-th projection of the *i*-th training MRI data $\mathrm{x}_i$, $c_i$ is the class label of $\mathrm{x}_i^p$, $P_k(c_i|\mathrm{x}_i^p)$ is the posterior probability on the class label $c_i$ for a given $\mathrm{x}_i^p$. Note that a particular projection *p* was used as input for the *k*-th DCNN of the ensemble.

## III. Determining Optimal Fusion Weights via Deep Ensemble Generalization Loss

The output of our DCNN ensemble classifier for given input MRI data was defined as:

$$P(c_i|\mathrm{x}_i) = \sum_{k=1}^{M} w_k\, P_k(c_i|\mathrm{x}_i^p) \quad (2)$$

where $w_k$ is the weight assigned to each individual DCNN, $w_k \geq 0$, and $\sum_{k=1}^{M} w_k = 1$. Note that from each 3D MRI data (scan), we usually obtain a large number of slice images from which to choose. We used an entropy-based sorting algorithm [13] to choose the most informative 32 slices from each projection of each 3D MRI data. As also noticed in our previous work [34], the MRI images with high entropy values showed more informative slices, which are usually located in the center of each projection. $P_k(c_i|\mathrm{x}_i^p)$, shown in Eq. (2), was computed by averaging 32 posterior probabilities calculated from the chosen 32 slices (associated with each projection *p*). In order to predict class labels, we considered the weighted average of the 2- or 3-dimensional vectors (in the 3-D case, each corresponded to one of the AD, MCI, and NC classes) produced by our nine DCNN member models (see Eq. (2)). The selected class was that with the highest probability in the resulting 2-D or 3-D (weighted average) prediction vector.

In Eq. (2), $w_k$ reflects the relative importance of each DCNN member. To find optimal the $w_k(k = 1, \ldots, M)$, we used a novel optimization framework: we explored different weight combinations via the proposed deep ensemble generalization loss (EGL) technique. Note that the use of the optimal weights should have minimized the generalization error of the ensemble. Kuncheva *et al.* [22] clearly showed that the generalization error of the ensemble network is determined by the right balance between its accuracy and diversity. In our optimization framework, different weight combinations are evaluated via a deep ensemble generalization loss (EGL), which is defined as:

$$\begin{aligned} f(\mathbf{w}) = & -\log\sum_{k=1}^{M} w_k P_k(c_i|\mathrm{x}_i^p) \\ & -\gamma\sum_{k=1}^{M} w_k \sqrt{\frac{1}{M}\sum\nolimits_{k\neq n}\left(P_k(c_i|\mathrm{x}_i^p) - P_n(c_i|\mathrm{x}_i^p)\right)^2} \end{aligned} \quad (3)$$

where $P_k(c_i|\mathrm{x}_i^p)$ is the *k*-th DCNN output on the class label $c_i$ (given the projection $\mathrm{x}_i^p$) and *M* is the number of DCNNs in the ensemble. In Eq. (3), the first term serves to compute the classification error of the DCNN ensemble. The second term (based on the root quartic negative correlation (RTQRT-NCL) [23]) is a measure of the difference among individual DCNN's outputs (on data sample): it quantifies the variability (i.e., disagreement) among the DCNN members, or the ensemble diversity. The $\gamma$ parameter serves to control the trade-off between the two terms in Eq. (3). For $\gamma = 0$, the ensemble weight $w_k$ is optimized

for each DCNN member, by considering only their ensemble accuracy. However, as the value of $\gamma$ increases, higher weights would be imposed on the DCNN members if they are very diverse. A good compromise was found by setting $\gamma = 0.25$.

The $f(\mathbf{w})$ expresses the goodness of $w_k$; smaller $f(\mathbf{w})$ values indicate better $w_k$. Moreover, it was parameterized based on a set of ensemble weights: $\mathbf{w} = \{w_k : k = 1, \ldots, M\}$. Hence, the objective of the proposed ensemble weight optimization is to minimize $f(\mathbf{w})$. We solved the following optimization problem:

$$\mathbf{w}_{\text{opt}} = \arg\min_{\mathbf{w}} \quad f(\mathbf{w}|\mathbf{V}) = \arg\min_{\mathbf{w}} \frac{1}{|\mathbf{V}|} \sum_{\mathbf{x}_i^p \in \mathbf{V}} f(\mathbf{w})$$

$$\text{s.t.} \quad \sum_{k=1}^{M} w_k = 1 \text{ and } \quad w_k \geq 0, \quad \forall k \tag{4}$$

where $|\mathbf{V}|$ denotes a validation set (i.e., the whole dataset), which should be kept unseen by all the DCNN members during their construction (Fig. 1), and $|\cdot|$ is the cardinality of the set.

To find $\mathbf{w}_{\text{opt}}$, we employed a constrained nonlinear optimization algorithm [10], in which different weight combinations were evaluated via our $f(\mathbf{w})$ based on the validation set. The widely used sequential quadratic programming (SQP) [12] was adopted to find $\mathbf{w}_{\text{opt}}$. The primary process of the SQP was to compute the Hessian of the Lagrangian function, by using a quasi-Newton updating method, and then to generate a quadratic programming (QP) sub-problem. The solution of this sub-problem was used to define the search direction for $\mathbf{w}_{\text{opt}}$. Using $f(\mathbf{w})$ from Eq. (3), we formulated the Lagrangian function for the SQP:

$$L(\mathbf{w}, \lambda) = f(\mathbf{w}) + \lambda \left( \sum_{m=1}^{M} w_m - 1 \right) + \sum_{m=1}^{M} u_m w_m \tag{5}$$

where $\lambda$ and $u_k (k = 1, \ldots, M)$ are the Lagrangian multipliers for the equality and inequality constraints (as defined in Eq. (4)), respectively. At each *k*-th iteration of the SQP method, the following QP sub-problem was solved [12]

$$\min_{\mathbf{d} \in \Re^M} \frac{1}{2} \mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{d}_k + \nabla f(\mathbf{w}_k)^{\mathrm{T}} \mathbf{d}_k$$

$$\nabla g(\mathbf{w}_k)^{\mathrm{T}} \mathbf{d}_k + g(\mathbf{w}_k) = 0$$

$$\nabla c(\mathbf{w}_k)^{\mathrm{T}} \mathbf{d}_k + c(\mathbf{w}_k) \leq 0 \tag{6}$$

where $g(\mathbf{w}_k) = \sum_{m=1}^{M} w_m - 1$, $c(\mathbf{w}_k) = \sum_{m=1}^{M} u_m w_m$ , and $\nabla$ is the gradient operator with respect to $\mathbf{w}_k$, and $\mathbf{H}_k$ is a positive definite approximation of the Hessian matrix of the Lagrangian function (shown in Eq. (5)). In our method, $\mathbf{H}_k$ can be updated using the most popular quasi-Newton method: the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [10]. The solution to the QP sub-problem (Eq. (6)) produces the vector $\mathbf{d}_k$ (i.e., the search direction), which was employed to formulate a new iterate:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha_k \, \mathbf{d}_k. \tag{7}$$

The step length parameter $\alpha_k$ was determined in order to produce a sufficient decrease in the merit function [12]. Note that, in the SQP, global convergence can be achieved by using an appropriate merit function. Here, we used the merit function proposed by [11]. Based on [11], [12], $\mathbf{w}_{k+1}$ was assumed to be converging to $\mathbf{w}_{\text{opt}}$ when $\|\mathbf{w}_{k+1} - \mathbf{w}_k\| \leq \varepsilon$ for all $k$ (e.g., $\varepsilon = 0.01$) or when the maximum number of iterations was reached. At the time of testing, the final $\mathbf{w}_{\text{opt}}$ is used to compute the ensemble test response and finally classify the MRI brain data.

TABLE I
DEMOGRAPHIC AND CLINICAL INFORMATION OF THE SUBJECTS FROM THE ADNI DB

| | Subject | Gender (M/F) | Age (mean ± STD) | MMSE (mean ± STD) |
|---|---|---|---|---|
| AD | 175 | 100/75 | 75.78 ± 7.78 | 23.28 ± 2.10 |
| MCI | 305 | 172/133 | 75.65 ± 7.89 | 27.14 ± 1.72 |
| NC | 335 | 195/140 | 75.33 ± 6.65 | 29.31 ± 1.13 |

STD: standard deviation.
MMSE: mini mental state examination [19].

## IV. RESULTS

In this work, we analyzed a baseline 3D structural MRI dataset obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu). We selected the T1-weighted MRI data (typically 240 × 256 × 176 voxels, for a voxel size of 1.5 $mm^3$ and a slice thickness of 1 mm) from the ADNI 1 phase. This phase included 815 subjects: 175 of them were selected from the AD group, 305 from the MCI group, and the remaining 335 from the NC group (see Table I for the demographic and clinical information). Statistical parametric mapping (SPM) [31] was used to normalize the image data, following a template of the International Consortium for Brain Mapping. The whole dataset was randomly split as follows: 40% for the training of the DCNN ensemble, 20% for its validation, and 40% for the test. This random partition was repeated ten times to avoid any bias possibly introduced by the random partition of the dataset. In this study, we reported the mean (average) and the 95% confidence interval of the classification accuracy.

From each 3D MRI data, we selected the most informative 32 slices and used them in the training and testing stages. Since the size of the generated slice image was different for each projection, all the slice images were converted to a resolution of 128 × 128 before using them as input for the DCNNs in the ensemble. A data augmentation strategy was used to reduce overfitting during the training of each DCNN member in the ensemble. With this objective, we applied a shift translation [14]–[21] to each training MRI data. The max shift translation was set to 2, generating 13,400 augmented training MRI data (subjects) for each class, and resulting in 428,880 slice images per each projection. The "MRI dataset" used for testing stage was not augmented. To construct the DCNN members, we followed the training procedure described in a previous paper. For the VGG-16 [24], we used a SGD (stochastic gradient descent) with a momentum of 0.9. The training was then regularized applying an $L_2$ penalty (weight $= 10^{-3}$) and two dropout layers for the first two fully-connected layers (rate $= 0.5$). In the case of GoogLeNet [25], we set the learning rate to 0.05, the weight decay to $10^{-3}$, and the momentum to 0.9. For AlexNet [14], we applied a SGD with a weight decay 0.0005 and a momentum of 0.9; moreover, the learning rate was initialized at 0.01 and

TABLE II
EFFECTIVENESS OF THE PROPOSED DCNN ENSEMBLE FUSION METHOD IN TERMS OF CLASSIFICATION ACCURACY (%) (95% CONFIDENCE INTERVAL) FOR 3-WAY CLASSIFICATION TASKS (I.E., CLASSIFICATION OF THE MRI DATA AS AD, MCI, OR HC)

| DCNN ensemble fusion approach | AD/MCI/NC |
|---|---|
| Baseline (accuracy of best single DCNN member) | 82.50 [82.41 82.60] |
| Majority voting fusion [26] | 87.12 [87.82 87.91] |
| Unweighted average fusion [9] | 88.98 [88.05 89.26] |
| Weighted fusion using random search [27] | 91.29 [90.78 91.40] |
| Weighted fusion based on validation accuracy [28] | 90.61 [90.01 90.91] |
| Weighted fusion with our ensemble generalization loss | **93.84** [93.22 94.05] |

Bold values denote the best results.

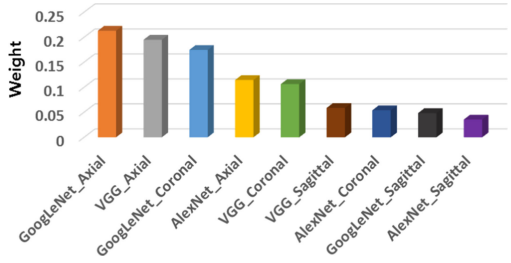


Fig. 2. Optimal fusion weights of the DCNN ensemble members determined via our method. The weights are sorted from left to right: from the most important to the least important.

reduced three times prior to termination. All the aforementioned DCNN architectures were trained with a mini-batch size of 128 for 200 epochs. Table II compares our DCNN ensemble fusion solution with other popular network ensemble fusion approaches. Our approach seems to outperform all the other ensemble fusion strategies, validating its usefulness. In addition, the classification accuracy obtained for each of the nine DCNN members was in the range of [76.84%, 83.07%], with a mean $\pm$ standard deviation (STD) of 79.92 $\pm$ 1.75 computed over 10 random partitions. Compared to the best single DCNN member (denoted as "Baseline" in Table II), the proposed ensemble significantly increased the classification accuracy (by a maximum of 11.34%). The ensemble weights $\mathbf{w}_{\text{opt}}$ determined using our method are visualized in Fig. 2. Interestingly, three of the top four ensemble members were based on axial projections (which were used as input); this indicates that the DCNNs with axial projections result in a better generalization power, promoting a correct AD classification.

Our proposed DCNN ensemble was compared with those of recent studies, which were focused on MRI-based AD diagnosis based on an ADNI cohort. Since there is a difference in dataset size and construction between our and previous studies due to the different subject selection and different data partition (subject number), we avoided a direct comparison of the methods' performances. Nevertheless, as described in [15]–[19], [30]–[32], we compared the methods' results to demonstrate the effectiveness of our method, since all performances were measured based on the same ADNI cohort. Table III shows the classification performances for three binary classification tasks. For the AD vs. NC task, our method achieved the second best classification accuracy, close to that reported in [19] (93.30%). Moreover, our method provided the best results for the AD vs. MCI and

TABLE III
CLASSIFICATION ACCURACIES (%) OF STATE-OF-THE-ART METHODS FOR THREE BINARY CLASSIFICATION TASKS AND PROPORTION OF CORRECT CLASSIFICATIONS

| Method \ Classification Task | AD vs. NC | AD vs. MCI | NC vs. MCI |
|---|---|---|---|
| CNN with 2D+ $\varepsilon$ fusion [15] | 91.41 | 69.53 | 66.25 |
| Deep Ensemble Sparse Regression Network [16] | 87.85 | N/A | 69.19 |
| Multiple Kernel Learning (MKL) [17] | 76.63 | 90.20 | 79.42 |
| Combining CNN and RNN [18] | N/A | 91.19 | 78.86 |
| Cross-Modal Transfer Learning [8] | 92.50 | 85.00 | 80.00 |
| 3D Inception-based CNN [19] | **93.30** | 86.70 | 73.30 |
| Complex Brain Networks [29] | 91.00 | 87.70 | 85.50 |
| Proposed DCNN Ensemble | 93.15 | **94.71** | **93.39** |

TABLE IV
CLASSIFICATION ACCURACIES (%) OF STATE-OF-THE-ART METHODS FOR 3-WAY CLASSIFICATIONS (AD VS. MCI VS. NC) ON THE ADNI COHORT

| Method | Accuracy (%) |
|---|---|
| Deep Sparse Multi-Task Learning [30] | 57.70 |
| 3D CNN [31] | 89.47 |
| Joint Regression [32] | 72.90 |
| Hidden Cues [3] | 48.79 |
| 3D-ACNN [2] | 89.10 |
| Multiple Kernel Learning (MKL) [17] | 90.20 |
| Proposed DCNN Ensemble | **93.84** |

NC vs. MCI tasks (they were improved by 4.52% and 8.89%, respectively, compared to the best previously published results), which represent more challenging classification tasks from a clinical point of view [31], [32]. Moreover, we compared the accuracy of our DCNN ensemble method with those of other state-of-the-art methods on the ADNI cohort for the 3-way classification task. Our method exhibited a superior classification performance (Table IV), confirming its usefulness. The training of our proposed ensemble was computationally time-consuming compared to those of the single DCNN-based approaches. However, the average testing time of our DCNN ensemble was low ($\sim$4.64 s for the MRI data) when using an Intel Xeon E5-2620 v4 CPU with 512 GB of RAM. This runtime performance validates the efficiency of the proposed DCNN ensemble algorithm when it is applied to Alzheimer's disease CAD systems [35].

## V. CONCLUSION

We proposed a new DCNN ensemble method for the classification of Alzheimer's disease (AD) using structural MRI data. A novel optimization framework was introduced to determine the optimal ensemble weights from a generalization perspective. In addition, we proposed the combined use of different MRI projections and DCNN architectures for the construction of the DCNN ensemble. The proposed method achieved competitive performances compared to other state-of-the-art deep networks based on AD classification approaches. The proposed algorithm could be incorporated into a clinical decision support system, reducing diagnostic errors and highlighting early predictors of AD.

## ACKNOWLEDGMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI).

## REFERENCES


[1] M. Zhao, R. H. M. Chan, T. W. S. Chow, and P. Tang, "Compact graph based semi-supervised learning for medical diagnosis in Alzheimer's disease," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1192–1196, Oct. 2014.

[2] E. Hosseini-Asl, R. Keynton, and A. El-Baz, "Alzheimer's disease diagnostics by adaptation of 3D convolutional network," in *Proc. 2016 IEEE Int. Conf. Image Proc.*, 2016, pp. 126–130.

[3] T. Glozman, and O. Liba, "Hidden cues: Deep learning for Alzheimer's disease classification CS331B project final report," 2016, pp. 1–8.

[4] M. Hon, and N. M. Khan, "Towards Alzheimer's disease classification through transfer learning," in *Proc. 2017 IEEE Int. Conf. on Bioinformat. Biomedicine*, 2017, pp. 1166–1169.

[5] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[6] K. Aderghal, J. Benois-Pineau, K. Afdel, and C. Gwenaëlle, "FuseMe: Classification of sMRI images by fusion of Deep CNNs in 2D+ $\epsilon$ projections," in *Proc. 15th Int. Workshop Content-Based Multimedia Indexing*, 2017, pp. 34:1–34:7.

[7] A. Ortiz, J. Munilla, J. M. Gorriz, and J. Ramirez, "Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease," *Int. J. Neural Syst.*, vol. 26, no. 7, p. 1650025, 2016.

[8] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, and G. Catheline, "Classification of Alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning," in *Proc. 2018 IEEE 31st Int. Symp. Comput.-Based Medical Syst.*, 2018, pp. 345–350.

[9] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1-2, pp. 1–39, 2010.

[10] J. Nocedal and S. Wright, "Numerical optimization," *Springer Sci. Bus. Media*, 2006.

[11] S. P. Han, "A globally convergent method for nonlinear programming," *J. Optim. Theory Appl.*, vol. 22, no. 3, pp. 297–309, 1977.

[12] P. T. Boggs and J. W. Tolle, "Sequential quadratic programming," *Acta Numerica*, vol. 4, pp. 1–51, 1995.

[13] C. Studholme, D. L. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognit.*, vol. 32, no. 1, pp. 71–86, 1999.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[15] K. Aderghal, J. Benois-Pineau, and K. Afdel, "Classification of sMRI for Alzheimer's disease diagnosis with CNN: Single Siamese networks with 2D+? approach and fusion on ADNI," in *Proc. 2017 ACM Int. Conf. Multimedia Retrieval*, 2017, pp. 494–498.

[16] H. I. Suk, S. W. Lee, D. Shen, and Alzheimer's Disease Neuroimaging Initiative, "Deep ensemble learning of sparse regression models for brain disease diagnosis," *Med. Image Analysis*, vol. 37, pp. 101–113, 2017.

[17] O. B. Ahmed, J. Benois-Pineau, M. Allard, G. Catheline, C. B. Amar, and Alzheimer's Disease Neuroimaging Initiative, "Recognition of Alzheimer's disease and mild cognitive Impairment with multimodal image-derived biomarkers and multiple kernel learning," *Neurocomputing*, vol. 220, pp. 98–110, 2017.

[18] D. Cheng and M. Liu, "Combining convolutional and recurrent neural networks for Alzheimer's disease diagnosis using PET images," in *Proc. 2017 IEEE Int. Conf. Imag. Syst. Techn.*, 2017, pp. 1–5.

[19] A. Khvostikov, K. Aderghal, A. Krylov, G. Catheline, and J. Benois-Pineau, "3D inception-based CNN with sMRI and MD-DTI data fusion for Alzheimer's disease diagnostics," 2018, *arXiv:1809.03972*.

[20] P. Drot´ar, J. Mekyska, I. Rektorov´a, L. Masarov´a, Z. Smekal, and M. Faundez-Zanuy, "A new modality for quantitative evaluation of Parkinson's disease: In-air movement," in *Proc. IEEE 13th Int. Conf. Bioinformat. Bioeng.*, 2013, pp. 1–4.

[21] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," 2019, *arXiv:1901.06032*.

[22] L. I. Kuncheva and C. J. Whitaker, "Measure of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, 2003.

[23] R. McKay and H. Abbass, "Anticorrelation measures in genetic programming," in *Proc. Australasia-Japan Workshop Intell. Evol. Syst.*, 2001, pp. 45–51.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, pp. 398–406, 2015.

[25] C. Szegedy *et al.*, "Going deeper with convolutions," 2014, *arXiv:1409.4842*.

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

[27] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.

[28] Z. H. Zhou, *Ensemble Methods: Foundation and Algorithms,* Boca Raton, FL, USA: CRC Press, 2012.

[29] X. Li, Y. Li, and X. Li, "Predicting clinical outcomes of Alzheimers disease from complex brain networks," in *Proc. Int. Conf. Adv. Data Mining Appl.*, 2017, pp. 519–525.

[30] H. l. Suk, S. W. Lee, and D. Shen. "Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis," *Brain Structure Function*, vol. 221, no. 5, pp. 2569–2587, 2016.

[31] A. Payan and G. Montana, "Predicting Alzheimer's disease: A neuroimaging study with 3D convolutional neural networks," 2015, *arXiv:1502.02506*.

[32] X. Zhu, H. I. Suk, S. W. Lee, and D. Shen, "Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis," *Brain Imag. Behav.*, vol. 10, no. 3, pp. 818–828, 2016.

[33] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[34] B. Lee, W. Ellahi, and J. Y. Choi, "Using CNN with data permutation scheme for classification of Alzheimer's disease in structural Magnetic Resonance Imaging (sMRI)," *IEICE Tr. Inf. Syst.*, vol. E102-D, no. 7, pp. 1384–1395, Jul. 2019.

[35] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," 2019, *arXiv:1811.10052*.