Computers in Biology and Medicine 151 (2022) 106201

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Multimodal ensemble model for Alzheimer's disease conversion prediction from Early Mild Cognitive Impairment subjects

Matthew Velazquez, Yugyung Lee *

Computer Science, School of Science Engineering, University of Missouri, Kansas City, MO, USA

ARTICLE INFO

Keywords:
Random forest
Convolutional neural network
Ensemble model with grid search
Diffusion tensor imaging
Electronic health records
Machine learning
Explainability
Alzheimer's disease
Mild cognitive impairment
Longitudinal studies

ABSTRACT

Alzheimer's Disease (AD) is the most common type of dementia. Predicting the conversion to Alzheimer's from the mild cognitive impairment (MCI) stage is a complex problem that has been studied extensively. This study centers on individualized EMCI (the earliest MCI subset) to AD conversion prediction on multimodal data such as diffusion tensor imaging (DTI) scans and electronic health records (EHR) for their patients using the combination of both a balanced random forest model alongside a convolutional neural network (CNN) model. Our random forest model leverages EHR's patient biometric and neuropsychiatric test score features, while our CNN model uses the patient's diffusion tensor imaging (DTI) scans for conversion prediction. To accomplish this, 383 Early Mild Cognitive Impairment (EMCI) patients were collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Within this set, 49 patients would eventually convert to AD (EMCI_C), whereas the remaining 335 did not convert (EMCI_NC). For the EHR-based classifier, 288 patients were used to train the random forest model, with 95 set aside for testing. For the CNN classifier, 405 DTI images were collected across 90 distinct patients. Nine clinical features were selected to be combined with the visual predictor. Due to the imbalanced classes, oversampling was performed for the clinical features and augmentation for the DTI images. A grid search algorithm is also used to determine the ideal weighting between our two models. Our results indicate that an ensemble model was effective (98.81% accuracy) at EMCI to AD conversion prediction. Additionally, our ensemble model provides explainability as feature importance can be assessed at both the model and individual prediction levels. Therefore, this ensemble model could serve as a diagnostic support tool or a means for identifying clinical trial candidates.

1. Introduction

Alzheimer's Disease (AD) is a progressive neurological disorder that causes the brain to diminish and leads to nerve cell death. Preventing the progression of AD is difficult as there are no effective treatment plans. To combat this, the focus has been placed on AD prediction from an earlier stage with the idea that treatment could be more significant when provided as early as possible or that this could benefit clinical trial enrollment [1]. This early stage, classified as Mild Cognitive Impairment (MCI), represents the onset of problems with memory recall, language, thinking, or judgment. Within the MCI umbrella, our study focuses on the Early Mild Cognitive Impairment (EMCI) subjects as they represent the furthest possible MCI subclass from AD. As 32% of patients diagnosed with MCI will develop Alzheimer's Disease [1], it is essential to have an accurate, explainable tool to identify which patients will convert.

In our previous works [2,3], we focused on the prediction aspect of this problem within a single modality, such as electronic health records (EHR) or medical images. This led to a machine learning model (i.e., random forest) that focused on EHR patient clinical data and a Convolutional Neural Network (CNN) model that performed predictions based on Diffusion Tensor Imaging (DTI) scans. While these models performed well, each had limitations and was not focused on being understandable. It became clear that combining these models into an ensemble multi-modality model with the added feature of explainability would be ideal for EMCI conversion prediction.

Additionally, the explainability of a model's predictions has been challenging to determine or is sometimes an afterthought. This has led to many high-performing prediction models that do not provide a clear rationale to healthcare providers. With explainable models, clinicians can be more confident in their diagnoses when leveraging a clinical decision-making tool. For our multi-modal work, explainability was a

* Corresponding author.
E-mail addresses: mvelazquez@mail.umkc.edu (M. Velazquez), leeyu@umsystem.edu (Y. Lee).

https://doi.org/10.1016/j.compbiomed.2022.106201
Received 4 June 2022; Received in revised form 17 September 2022; Accepted 9 October 2022
Available online 30 October 2022

key objective.

Therefore, our work focused on developing a multi-modality ensemble model for AD conversion prediction that could explain the rationale behind its predictions. The first piece leverages a random forest, a supervised learning algorithm that is efficient with classification problems Qi [4]. This would focus on interpreting patient clinical features while the other side of the ensemble, the CNN model, would handle a patient's Diffusion Tensor Imaging (DTI) scans.

Diffusion tensor imaging (DTI) is a form of magnetic resonance imaging (MRI) that detects how water moves along the brain's white matter tracts. This water molecule diffusion difference can then be contrasted to show the variation between scans. Our work centered on apparent diffusion coefficient (ADC) DTI scans. ADC measures the magnitude, within the tissue, of water molecule diffusion.

As our classes are originally imbalanced, we provide determinations on how to best balance our data as well as which augmentation forms are most appropriate. In addition, a method for dynamically choosing the ideal weight of each model within the ensemble for any given prediction is also provided. Finally, complete ensemble explainability of both the visual and clinical feature inputs is provided, and analysis is performed. The main contributions in this paper are (1) building an ensemble model against an imbalanced data set; (2) determining the ideal weighting of that ensemble per patient prediction; (3) explaining model prediction rationale for both visual and clinical features; (4) determining the conversion prediction accuracy of our model. We believe that this work will provide an understandable tool that can be used to predict patient AD conversion from a prodromal stage. In addition, this work provides both global and local explainability methods for ensemble models.

2. Related work

2.1. MCI-to-AD conversion prediction

As the AD conversion problem matures, multiple studies now evaluate based on different mixes of modalities. Zhang et al. [5] used a combination of graph theory and machine learning to predict the conversion of MCI subjects to AD based on sMRI/fMRI data. Their work explored multiple feature selection methods (e.g., random subset feature selection algorithm, minimal redundancy maximal relevance, and sparse linear regression) and achieved an accuracy of 84.71%. They also explored the relationship between AD conversion and high-sensitivity brain regions to find that both structural and functional areas were relevant as predictors.

An evaluation between unimodal and multimodal models for AD conversion was performed by Minhas et al. [6]. In their work, MRI-derived biomarkers in combination with neuropsychological measures were used to determine early AD warning signs from an MCI population. They achieved an AUC of 95.7% with their multimodality data trained through a support vector machine (SVM).

Lin et al. [7] fused four modalities (MRI, positron emission tomography, cerebrospinal fluid biomarkers, and gene data) which were then individually graded using their Extreme Learning Machine (ELM) model. Their scope focused on conversion prediction within three years as they achieved an accuracy of 84.7%. In addition, their findings demonstrated a minimum 10% increase in accuracy from using multiple modalities rather than when only a single modality was used.

Focusing on a reduced set of sociodemographic, characteristics, clinical information, and neuropsychological test scores, Grassi et al. [8] developed a new machine-learning algorithm for three-year AD conversion prediction. Their work aimed to leverage data that did not derive from expensive, invasive, or otherwise difficult procedures such as lumbar puncture, genetic testing, or neuroimaging techniques. With these restrictions, they could still obtain an AUC of 88% through an SVM.

Huang et al. [9] proposed a predictive nomogram that combined AB concentration, image features, and clinical factors to predict MCI-to-AD conversion. Analysis was also performed on how features were associated with one another and the significance of each feature. To better understand the patterns of AD conversion, they focused on examining the associations at both the micro and macro levels.

Varatharajah et al. [10] focused on which markers would be most relevant for AD conversion models. Using a mix of clinical data, MRI, and FDG-PET, they could isolate large shares of variance in the pathophysiology (amyloid, tau) variables. Their work also revealed the relevance of CR1 (complacent receptor 1) as an individual predictor of AD conversion. As a result of their work, they achieved an AUC of 93% via an SVM.

Rana et al. [11] created MudNet, a CNN model which performed both MCI-to-AD conversion prediction and time-to-AD conversion. They could group patients into high-risk and low-risk categories based on whether they were predicted to convert within 24 months. Their model used a mixture of volumetric MRI and clinical data, which also consisted of neuropsychological tests (RAVLT, ADAS-11, ADAS-13, ADASQ4, MMSE). With these inputs, they achieved an accuracy of 69.8% for conversion predictions and 66.9% for risk classification.

2.2. Explainability

Explainability for AI models has seen increased demand over recent years. Historically, models were seen as black boxes, but now multiple explainability methods can be used to provide the rationale behind a model's behavior. For the medical domain, this is highly relevant as it allows physicians to understand the process that a decision support system uses to arrive at its recommendation. In one study, Viton et al. [12] focused on using heatmaps to visually explain a CNN model's predictions on in-hospital mortality. Their multivariate time series approach allowed for critical points to be identified and the most influential variables. The visual aid can help justify the model's decisions with this detailed explainability. While the purpose of their work was explainability, they were still able to achieve an AUC of .8207, predicting in-hospital mortality risk.

Maweu et al. [13] also worked to provide an explainable framework for their CNN model. In this case, they targeted ECG signals (one-dimensional time-series data). An interesting aspect of their approach is that they leveraged 1D-CNN models rather than the standard 2D-CNN ones. This allowed them to display descriptive statistics, feature visualization, detection, and mapping for each module of their proposed framework. With this knowledge, they could further explore the relationships between their features and how they might contribute to misclassification. This idea of identifying the rationale behind misclassification was a focus of our work and will be analyzed within our results.

2.3. Ensemble classification

Ensemble classifiers have proven to be highly efficient in recent years. The majority of these ensembles are typically similar algorithms which are then stacked. This usually consists of stacked ML algorithms (RF, SVM, XGBoost, etc.) or a chain of deep learning models. Some ensembles represent the combination of an ML model and a deep learning model, which is the approach we took for this work. We seek to add to this domain by including an explainability layer by combining a random forest classifier with a CNN model.

Mostafiz et al. [14] performed Covid-19 detection via chest x-rays by combining a random forest with a CNN model. After scanning, the initial x-ray is then enhanced and segmented before the key features are extracted. The random forest is then used for detection once the key features have been passed. Their work achieved an accuracy of 98.5% when both sides of their ensemble were engaged. However, when only one side was leveraged, their accuracy dropped to 84%, showing the advantage of ensemble classification.

Another study by Priyadarshini and Puri [15] combined multiple

machine learning algorithms (SVM, Random Forest, MLP, etc.) on top of a CNN, which served to draw out comparisons between the different methods. Their goal was exoplanet detection, and they were able to achieve 99.62% accuracy with their Ensemble-CNN model.

3. Methods

3.1. Data collection

All data used for this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and included patients from their ADNI-1, ADNI-2, and ADNI-GO studies [16]. "ADNI is a global research study that actively supports the investigation and development of treatments that slow or stop the progression of AD" [16]. ADNI aims to track AD progression using biomarkers and clinical measures to assess the brain over each stage of the disease.

The selection criteria for our work focused specifically on the EMCI subset with patients that had follow-up exams for more than a year. EMCI patients represent the stage typically 5–7 years before a potential AD diagnosis. The Wechsler Memory Scale Logical Memory II test determines this earlier subset compared to the more general MCI stage. For our classification problem, the EMCI patients were divided into two classes (EMCI_C, EMCI_NC). EMCI_C represents patients that would eventually convert to an AD diagnosis, whereas EMCI_NC represents patients that would not convert. This distinction was provided by the Clinical Dementia Rating ADNI variable of the patient's last exam diagnosis.

For the clinical feature model, 1806 exam visits were used pre-augmentation. 1608 belonged to the EMCI_NC class, while 198 were from the EMCI_C conversion class. For the DTI model, 405 DTI images were gathered, which, after our pre-processing methods, represented a singular central slice of each scan. These were then grouped into 90 unique EMCI patients, where 16 would convert to AD (EMCI_C) and 74 would not (EMCI_NC). In total, our study consisted of 383 EMCI patients (shown in Fig. 2), 49 of these within the EMCI_C class and 335 within the EMCI_NC class. Stratified by age, our largest demographic was ages 70–74, followed by 65–69. Our training/test split for this work was 75% (288 patients) to 25% (95 patients).

3.2. Clinical features selection

For the random forest component of our ensemble model, nine ADNI features were chosen, as seen in Table 1. These features contained physical biomarkers (ventricular and hippocampal volume), genetic biomarkers (APOE4), neuropsychological scale scores (FAQ, MMSE, ADAS13, ADAS11), and demographic variables (age, race). Initially, starting with over 90 features, we could eliminate many variables with a combination of SHAP analysis and Gini importance until an ideal fit had been obtained.

3.3. Ensemble classification model

We assemble an ensemble model that combines Random Forest clinical feature prediction alongside a Convolutional Neural Network (CNN) that performs predictions based on diffusion tensor imaging (DTI) scans to take advantage of our multimodality data. This allows each model's limitations to be mitigated by engaging the other model for its prediction confidence.

Random Forest, our first classifier, uses a method that constructs a multitude of decision trees which then outputs the majority vote as the prediction. As subsets of features are randomly selected for each decision tree, this provides enhanced tolerance for overfitting. For our work, this classifier can either output EMCI_C (conversion class) or EMCI_NC (stable class). Each decision tree, made up of a random assortment of our nine clinical features, gets to cast a vote. Overall prediction confidence can be determined by observing how many trees voted for the majority class. As we can assess each node's importance in a given tree, we can evaluate each feature's importance for both the model and individual predictions. This allows us a measure of explainability for the clinical feature aspect of our overall ensemble model. Other classifiers were evaluated per Table 2, but the Random Forest algorithm provided the best performance. Additional specifics on this random forest model can be seen in our prior work [2].

Our second classifier consists of a Convolutional Neural Network (CNN) with a NASNet architecture [17] as its backbone. CNN models have previously demonstrated accuracy with MRI scans as they are built to process pixel data [3]. These capture spatial and temporal dependencies within an image, making them ideal for image classification. We had initially built our model with the Inception v3 architecture [18] but found better performance with NASNet. Zoph et al. [17] integrate reinforcement learning with a controller RNN to construct a cell or layer for the NASNet network, which delivers cutting-edge ImageNet accuracy. Our CNN model combines a NASNet architecture with an RNN controller to recursively search for the best structure as it trains. Creating a network with NASNet makes the search strategy significantly more successful for PNASNet [19].

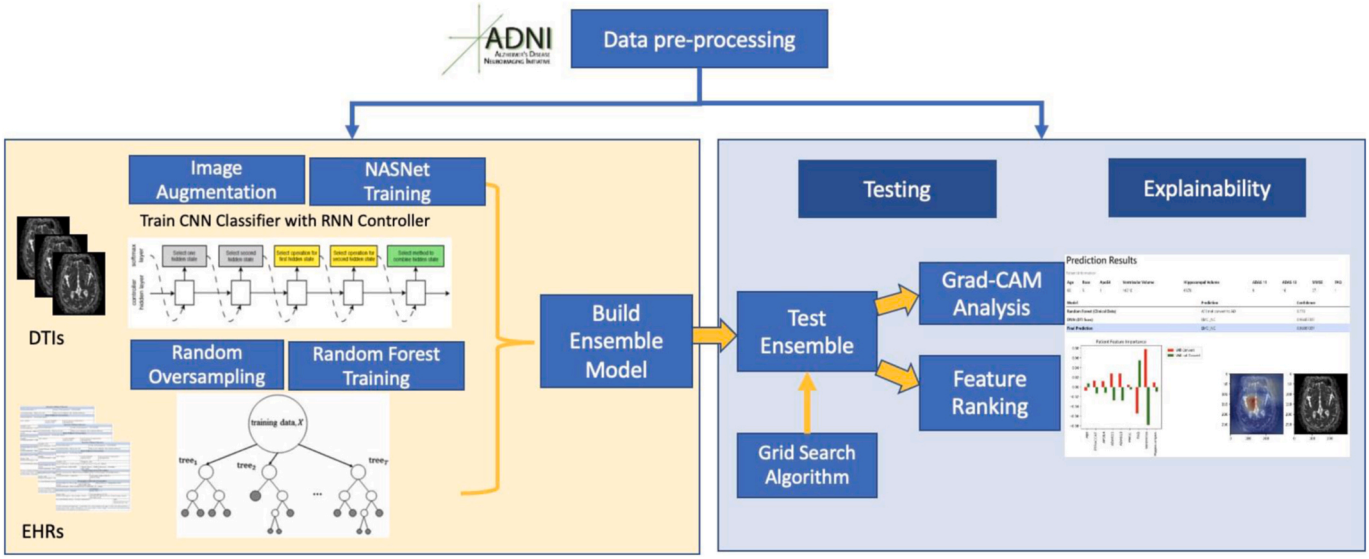

Fig. 1. Ensemble model workflow.

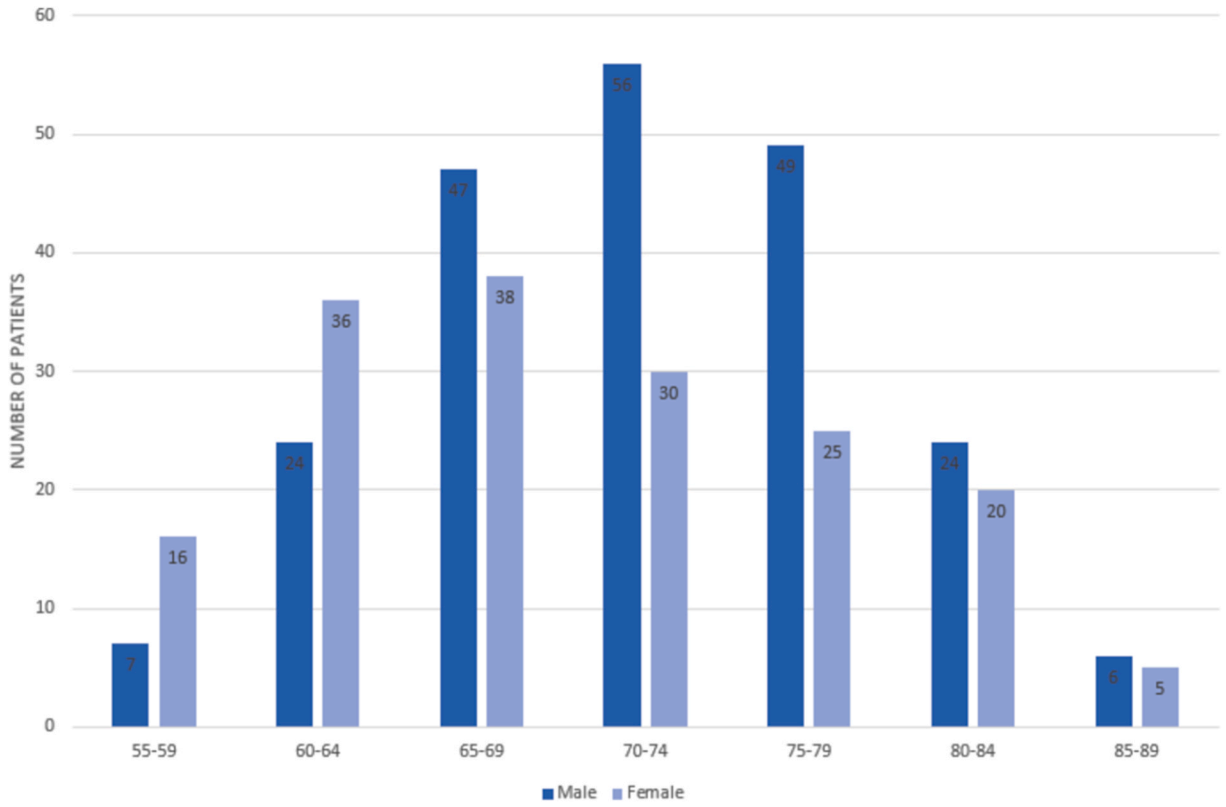

Fig. 2. Subjects' age and gender distribution.

Table 1
Clinical feature characteristics.

ADNI Feature	Subject#	EMCI_C		EMCI_NC	
		49		335	
	Description	Mean	SD	Mean	SD
DX	Diagnosis	–	–	–	–
Demographic information					
PTRACCAT	Patient Race	–	–	–	–
AGE	Patient Age	73.5	6.47	71.1	7.49
Genetic Biomarkers					
APOE4	The number of e4 alleles of APOE	0.9	0.71	0.4	0.46
Physical Biomarkers					
Hippocampus	Hippocampal volume	6875.2	947.45	7334.1	910.20
Ventricles	Ventricular volume	39282.7	21031.66	34504.6	21394.49
Neuropsychological scales					
ADAS13	13-item AD Assessment Scale	15.8	6.02	13.3	5.41
ADAS11	11-item AD Assessment Scale	9.7	4.12	8.5	3.29
FAQ	Functional Activities Questionnaire	4.1	4.38	1.82	2.50
MMSE	Mini-Mental State Examination	28.1	1.58	28.3	1.71

EMCI_C the converter group, *EMCI_NC* the stable group.

As shown in Fig. 3, these blocks consist of both standard and reduction cells. Normal cells represent convolutional cells that return a feature map of the same dimension. In contrast, reduction cells produce similarly but with the height and width reduced by a factor of two [20]. These are the only structures that the RNN controller subsequently searches. As seen in Table 3, other architectures were evaluated, but NASNet was the leading performer. As a result, this became our ideal architecture despite its computationally intensive approach. As NASNet was trained on ImageNet's 1.2 million images, our ADNI data was used to retrain the final classification layer using TensorFlow.

We then combine these classifiers to form our ensemble model. This allows us to intake either clinical data, DTI scans, or both to accurately predict AD conversion while mitigating each classifier's weaknesses. A grid search algorithm is then performed to exhaustively determine the ideal weight that each classifier should carry within the ensemble. This optimization (Table 5) resulted in a 0.55 CNN vs. 0.45 RF weighting as the ideal balance for AD conversion prediction.

3.4. Data balancing

Given the nature of our imbalanced data set, with 12.8% of patients belonging to the minority class (EMCI_C), we implement different augmentation methods for our ensemble to have better representation in our training/test data. We perform random over-sampling for our Random Forest classifier to make our two classes equivalent in size. This is done by taking random samples from the EMCI_C class with

Table 2
Clinical data classifier performance comparison.

Model/ Feature	Accuracy	Precision	Recall	F1 Score	AUC	p-value
Random Forest						
6-Features	0.892	0.907	0.980	0.942	0.88	0.91
9-Features	**0.936**	**0.952**	0.978	**0.965**	**0.96**	**0.71**
13-Features	0.916	0.916	**0.998**	0.955	0.93	0.82
Support Vector						
6-Features	0.900	0.900	1	0.948	0.52	–
9-Features	0.900	0.900	1	0.948	0.54	–
13-Features	0.900	0.900	1	0.948	**0.55**	–
Logistic Regression						
6-Features	0.894	0.902	0.990	0.944	**0.76**	–
9-Features	0.892	0.903	0.985	0.942	0.75	–
13-Features	**0.896**	**0.904**	**0.990**	**0.945**	0.75	–
XGBoost						
6-Features	0.898	0.904	**0.993**	0.946	0.87	–
9-Features	**0.920**	**0.930**	0.985	**0.957**	**0.89**	–
13-Features	0.907	0.921	0.980	0.950	0.88	–

replacement until the size matches that of the majority class. This provides 2,412 total exam visits for training rather than the original 1,354 pre-augmentation visits. Our over-sampling method was compared against both under-sampling methods and class weight modifications but continued to perform best.

For the CNN classifier, multiple augmentation methods were performed against our initial 405 EMCI images to increase the overall training size. The most effective augmentation methods were to flip the scans horizontally and to add randomization to an image's brightness. As these scans come in at different brightness levels, augmenting this allowed our model to learn at a far better rate. Variations of cropping or scaling the images did not increase our accuracy. A comparison of our visual augmentation methods sorted by accuracy can be seen in Table 3. Additionally, compared to our augmented data set, our original data set can be observed in Table 4. This table demonstrates our train/test data split and the initial class imbalance.

3.5. Grid search algorithm

We perform a grid search to exhaust possible weight combinations to find the ideal weighting for our ensemble model. First, we define our possible weight values for each model as 0.0 to 1.0 and then iterate through the process in steps of 0.1. After each weight vector is generated, they are normalized to ensure that they sum to one. Once the grid search has been completed, the weights of the highest accuracy run (0.55 CNN, 0.45 RF) are captured and used for the final ensemble model. Other weighting combinations can be observed in both Table 5. As each model running independently is also contained within this table (1, 0 and 0, 1), the advantage of using both the ensemble approach and dynamic weighting can be easily compared.

When individual patients are submitted to our model, each classifier (RF and CNN) generates a prediction and its confidence in that prediction. Our grid search-derived weighting is then factored into this prediction confidence (PC) to determine the overall ensemble prediction. As there can be disagreements between the different modalities, this weighting allows us to slightly prefer the more accurate classifier (CNN).

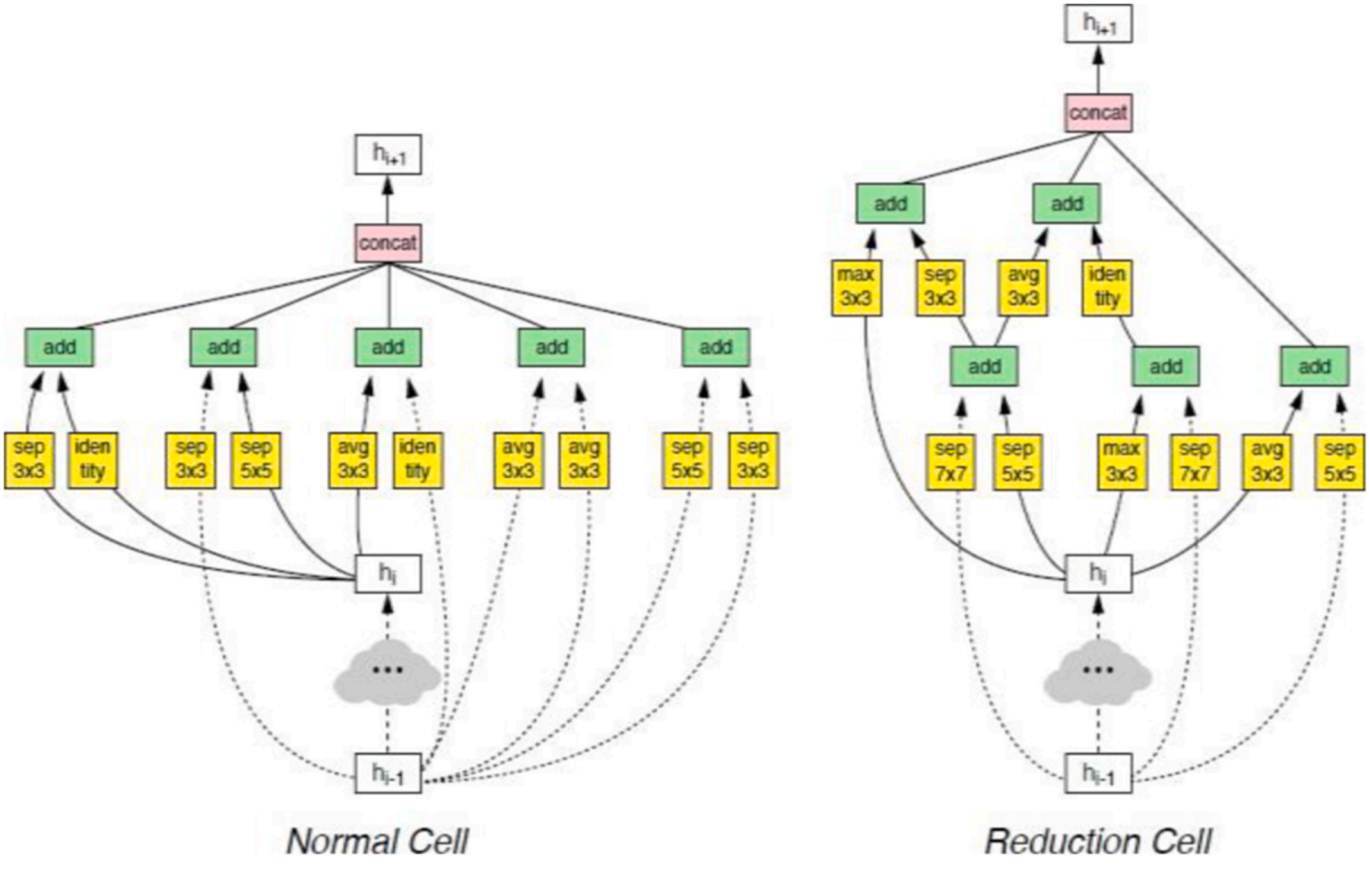

Fig. 3. NASNet architecture.

Table 3
Comparative evaluation with CNN architectures and augmentation methods.

Architecture	Accuracy	Model	Training Steps (#)	Scale Distortion	Brightness Distortion	Crop Distortion	Flipped Images	Learning Rate
NASNet	**96.4%**	$NASNet_{M1}$	8000	0	30%	0	True	.005
	89.6%	$NASNet_{M2}$	8000	0	0	0	True	.005
	79%	$NASNet_{M3}$	10000	30%	30%	30%	True	.005
	77.1%	$NASNet_{M4}$	8000	30%	30%	30%	False	.005
	75%	$NASNet_{M5}$	10000	50%	50%	50%	True	.003
Inception	94.7%	$Inception_{M1}$	8000	0	30%	0	True	.005
	73.7%	$Inception_{M2}$	8000	50%	50%	50%	True	.005
PNASNet	91.2%	$PNASNet_{M1}$	8000	0	30%	0	True	.005
	67.7%	$PNASNet_{M1}$	8000	50%	50%	50%	True	.01

Table 4
Data set by modality and class.

Data	Clinical Data		DTI	
	EMCI_C	EMCI_NC	EMCI_C	EMCI_NC
Subject#	49 Subjects	335 Subjects	16 Subjects	74 Subjects
Original Record#	198	1608	72 images	333 images
Record# after Over-sampling/ Augmentation	1608	1608	576,000 images	2,664,000 images
Training Data	1206	1206	432,000 images	1,998,000 images
Testing Data	402	402	144,000 images	666,000 images

Table 5
Weighted average classifier accuracy compared.

Iteration	CNN Weight	RF Weight	Accuracy
1	.55	.45	**98.81%**
2	.60	.40	97.62%
3	.40	.60	94.05%
4	.70	.30	96.43%
5	.65	.35	96.43%
6	.50	.50	95.24%
7	0	1	92.86%
8	1	0	96.43%

4. Results

4.1. Random forest feature characteristics

For our clinical data, the average age of the subjects was 71.4.55.6% of these were men, and there was a statistically significant age difference between the two groups ($P < .05$). Regarding the genetic and physical biomarkers, APOE4 and hippocampal volume showed substantial differences between the EMCI_C and EMCI_NC classes. Ventricular volume was consistent across both classes. With the neuropsychological scale scores, ADAS13 and FAQ showed significant differences ($P < .05$), whereas ADAS11 and MMSE did not. The correlation matrix in Fig. 4 demonstrates the totality of our clinical data feature relationships.

4.2. Ensemble model performance

Our ensemble model workflow can be observed in Fig. 1. This demonstrates how the random forest and CNN models work together or independently to output an explainable prediction for our EMCI subjects. Our random forest model is trained with 1000 trees against 2412 exam visits, while our CNN model leverages Tensorflow to retrain the final classification layer of NASNet for DTI analysis. Additionally, we pass a max_depth of 40 with nine max_features as further hyperparameters to the random forest model. We arrive at this tuning by implementing Grid Search to derive the ideal hyperparameters. Next, we

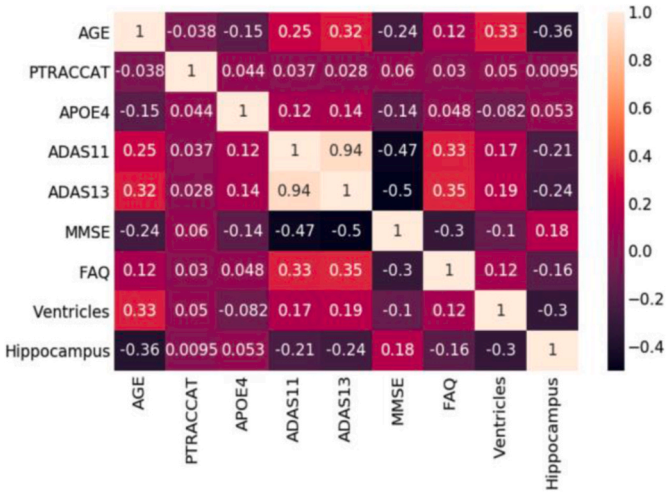

Fig. 4. Random forest model correlation matrix.

optimize our CNN model with the previously discussed augmentation distortions and then train for 8000 steps at a learning rate of 0.005. During random forest training, 25% of our clinical data are reserved for testing, while the remaining 75% account for the training data. For CNN training, 10% of our images are reserved for testing, 10% for validation, and the remaining 80% for training. These models are then combined to form our ensemble model, after which our Grid Search algorithm is applied to determine the ideal weight distribution. Once the weighting has been applied, our model explainability occurs via the combination of feature ranking and Grad-Cam analysis. This ensures that each outputted prediction has accompanying explainability.

One of the advantages of our ensemble approach is that it allows either modality to be passed absent of the other, and a prediction is still generated. When both modalities (clinical data and DTI scans) are provided, our weighted ensemble model achieves an EMCI-to-AD conversion prediction accuracy of 98.81%. With only clinical data being supplied, our model maintains an accuracy of 92.86%. When only DTI scans are provided, our model performs at 96.43% accuracy. This flexibility ensures that accurate conversion prediction can be obtained even if one is missing certain features. We also measure the model differences in Fig. 6 with the polygon area metric (PAM) proposed by Ref. [21]. The individual PAM metrics per model are shown in Table 6. As the model spent significant time being fine-tuned, the experiment was repeated

Table 6
Individual polygon area metrics: Classification accuracy (CA), sensitivity (SE), specificity (SP), Jacard Index (JI), F-score (F), area under curve (AUC).

	CA	SE	SP	JI	F	AUC
RF	.929	.615	.986	.571	.727	.960
CNN	.964	1	.958	.813	.897	.973
Ensemble	**.988**	**1**	**.986**	**.929**	**.963**	**.992**

several dozen times per model. After each repeat, performance metrics were assessed to see how to tune the model further. Additionally, cross-validation was performed to ensure different bagging combinations performed well.

The difference between each model's confusion matrix can be seen in Fig. 5. While the individual RF model struggled with false negatives, this weakness is removed when transitioning to the ensemble approach. Similarly, while the individual CNN model had three false positives, this was mitigated when predicting as part of the ensemble model (see Fig. 7).

4.3. Ensemble explainability

A key contribution of this work was to provide accurate conversion prediction and be capable of explaining the rationale behind individual predictions and the overall model. As our ensemble model weighs visual prediction alongside clinical data prediction, it is essential to know the prediction confidence related to each modality. Additionally, understanding the features or pixels that led to the overall decision within each classifier can help instill confidence in a clinical setting. This can be distinguished by providing context around global (model-level) explainability vs. local (individual-level) explainability.

For our clinical global explainability, we perform feature ranking of our nine features based on the following formula:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (1)$$

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k\in\ all\ nodes} ni_k} \quad (2)$$

$$normfi_i = \frac{fi_i}{\sum_{j\in\ all\ features} fi_j} \quad (3)$$

$$RFfi_i = \frac{\sum_{j\in\ all\ trees} norm\ fi_{ij}}{T} \quad (4)$$

Initially, in Equation (1), we determine the importance of each node per tree (ni). ni_j represents node j's importance, with C_j being a node's impurity value. Additionally, the weighted samples that reach node j are represented as w_j. From this, feature importance (fi) per tree can be calculated as seen in Equation (2). This result is then normalized between 0 and 1 (Equation (3)). This process is then averaged out to the entire forest and divided by the number of trees within the forest per Equation (4) [22].

For global explainability, our ensemble model's feature ranking per the above function can be seen in Fig. 8. We also demonstrate the permutation importance ranking seen in Fig. 9 as well as the Shapley plot (Fig. 10). Permutation rankings can reduce high cardinality bias as the features are permuted against a held-out test set. A baseline metric is established for this to occur, which has each feature permuted against it—the difference between this feature permutation and the baseline metric results in the overall permutation importance. For our primary feature ranking, age, hippocampal volume, and ventricular volume stood out as our model's most important features. With permutation ranking, age and FAQ continued to show strength, with APOE4 gaining in importance compared to its feature ranking. Finally, the Shapley plot shows how strongly each feature contributes to a positive (EMCI_C) versus a negative (EMCI_NC) prediction. The color of each value denotes whether it is high (red) or low (blue) relative to other values for that feature. The combination of these ranking systems aided us in reducing the original feature map to the final model.

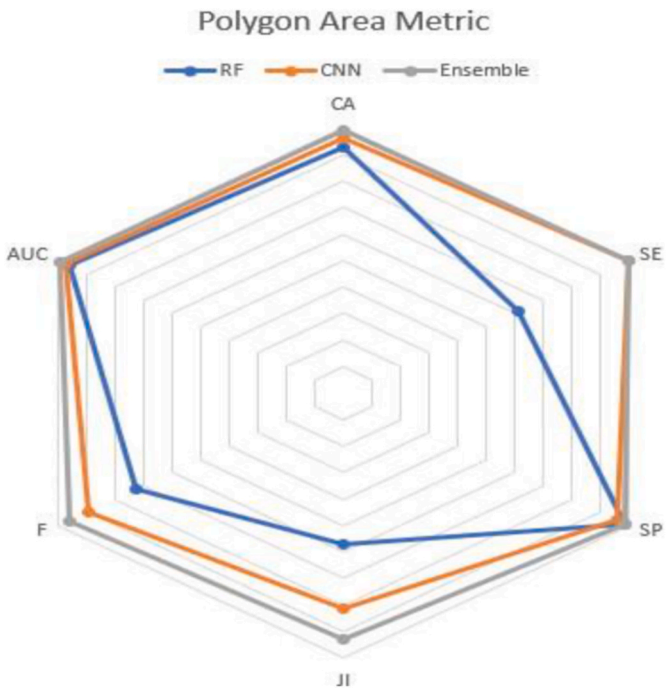

Fig. 6. Polygon area metric.

For DTI local explainability, we perform Gradient-weighted Class Activation Mapping (Grad-CAM) [23] to generate pixel heat maps. These are then superimposed on the existing image to display the most important regions for the resulting prediction. In this sense, Grad-CAM allows us to understand what our CNN model focuses on by using the gradients that flow into the final convolutional layer. These gradients then use global average pooling to obtain the necessary weights as seen in Equation (5). Examples of these heat maps for both an EMCI_C and

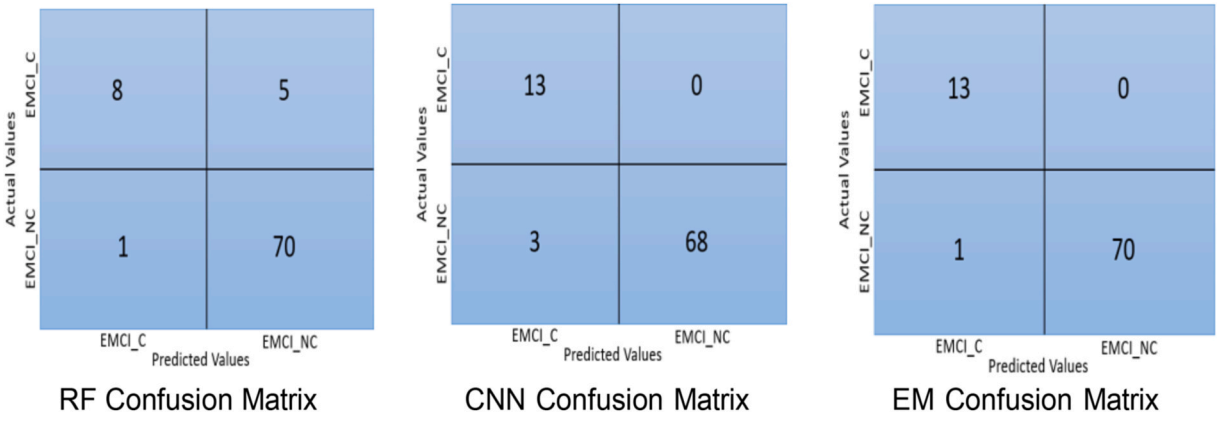

Fig. 5. Confusion matrix: (a) Random forest (RF) for EHRs, (b) convolutional neural network (CNN) for fMRI, and (c) ensemble model (EM).

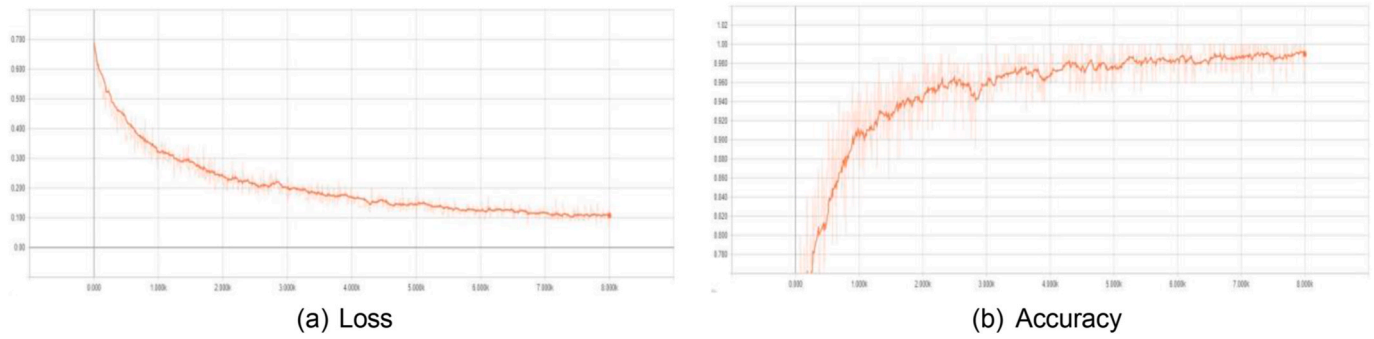

Fig. 7. Tensorboard graphs displaying the loss and accuracy during the CNN training period in steps. The X-axis is representing the steps and the Y-axis is displaying the accuracy.

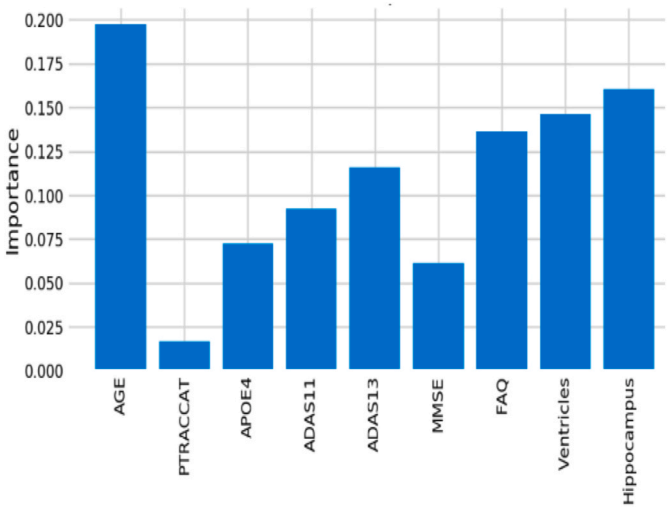

Fig. 8. Ensemble model feature importance.

EMCI_NC patient are shown in Fig. 11. For our output, the black and white image represents the initial input before Grad-Cam applies the heatmap. As Grad-Cam assesses which pixels are most relevant, it colors them red at varying intensities to demonstrate that pixel's importance to the prediction. Similarly, darker shades of blue occur when the pixel is deemed not to have a strong contribution to the prediction. Future work will explore aligning these heat maps to segmented brain regions to establish more in-depth global explainability.

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

To assess our ensemble model's local explainability and demonstrate its potential as a clinical decision support tool, we've built a Flask Python application to host our model and allow for patient intake. Our application allows for patient clinical data to be entered in addition to attaching DTI scans. Partial patient information can also be provided as the application will understand if it has been provided with limited features. For example, only the Random Forest classifier will be engaged if only clinical data is provided. Likewise, the CNN model will serve as the sole predictor if a DTI scan is the only patient data provided. The application also accounts for blanks by substituting the empty field with that feature's mean average. Once the data has been submitted, our

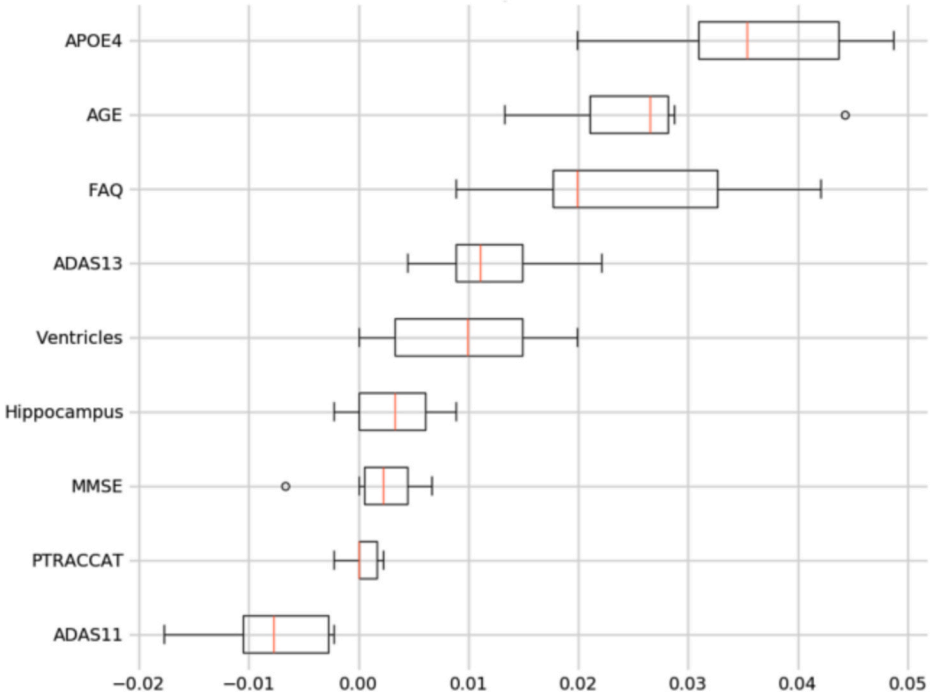

Fig. 9. Ensemble model permutation importance.

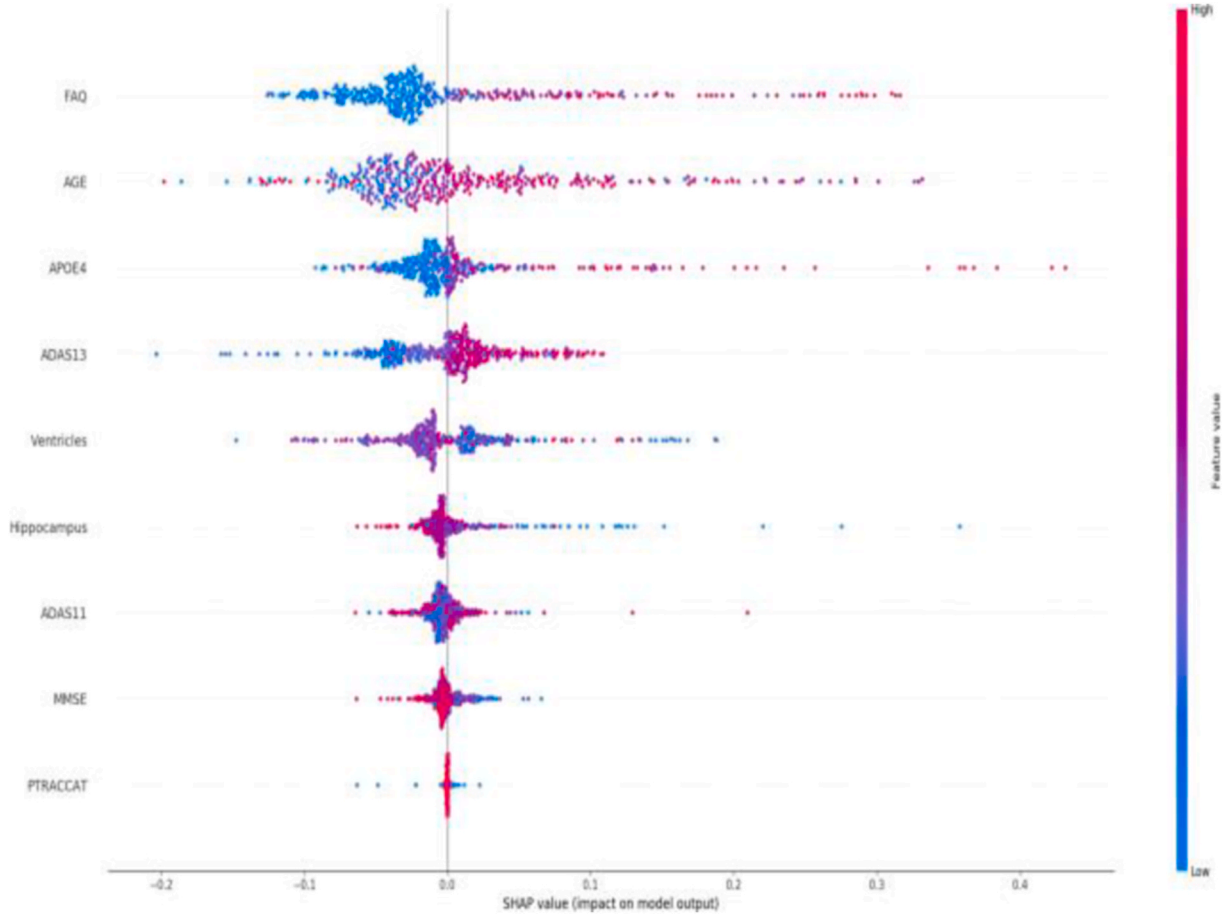

Fig. 10. Ensemble model shap summary.

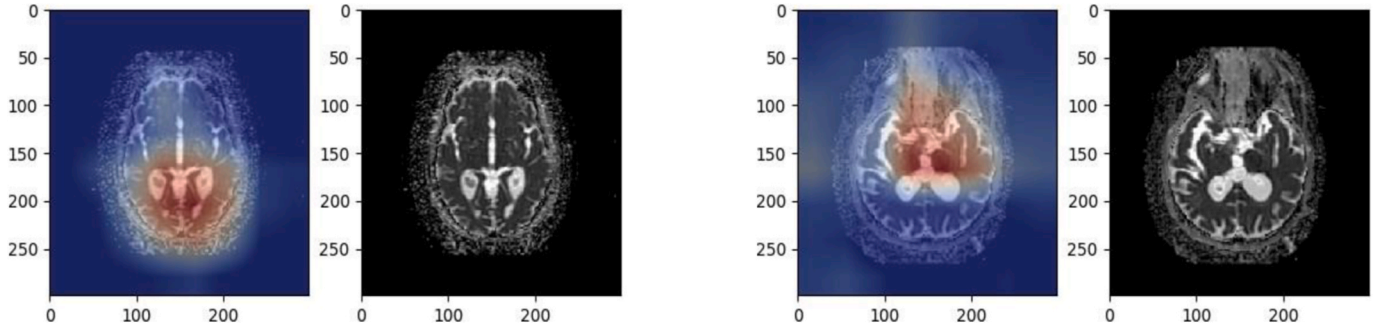

Fig. 11. Grad-CAM explainability: Grad-CAM and DTI images for EMCI_C patient (left) & EMCI_NC patient (right).

ensemble model is engaged, which outputs its prediction and explainability. From our application's output, we can see the importance of the clinical data feature importance alongside the Grad-Cam analysis. We also see the prediction confidence of each independent classifier and the overall ensemble confidence. This informs the user which modality contributed the most to the prediction, highlighting key regions/features of interest. An example of the intake form and a sample prediction can be seen in Fig. 12.

Table 7 and Fig. 13 detail three unique, individual predictions with local explainability that demonstrate our ensemble model's strength in contrast to a singular model. Within this table, prediction contributions (PC) are also shown. This represents the amount of each clinical feature's contribution to the overall RF prediction. A positive value indicates the contribution towards the ground truth class, whereas a negative value represents the contribution to the incorrect class. As an example, patient 2106 was eventually diagnosed with AD. However, based on their clinical data, our Random Forest component predicted with 52% confidence that they wouldn't convert. In contrast, after assessing the patient's DTI scan, our CNN model predicted that they would convert with 79% confidence. With the CNN model being more confident and having more weight in the overall prediction, this resulted in an ensemble confidence of 65% that the patient would convert to AD (EMCI_C). In this case, a singular RF model would have been predicted inaccurately, but with added visual analysis, our ensemble was capable of avoiding the mistake.

Another example from Table 7 can be seen with EMCI_NC patient 4220. However, with this patient, the RF prediction (EMCI_NC, 99% confidence) helped overrule the incorrect CNN prediction (EMCI_C, 58% confidence). The Grad-CAM analysis in Fig. 13 shows the difficulty in assessing this patient's DTI scan as the heat map overlaid most of the

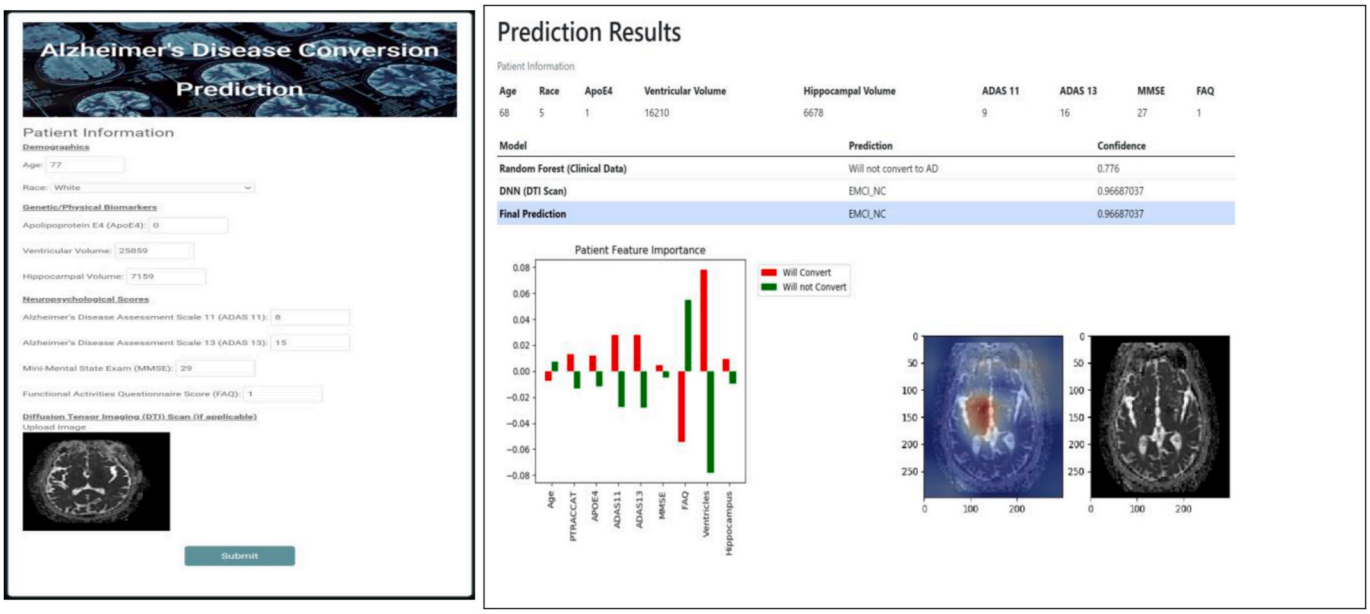

Fig. 12. Conversion prediction intake (left) and results (right).

Table 7
Example features and prediction contributions (pc) from three distinct cases
GT: Ground Truth, RF: Random Forest, CNN: Convolutional Neural Network, EN: Ensemble; EMCI_C: C, EMCI_NC: NC.

Subject 1: 2106				Subject 2: 4220				Subject 3: 4897			
GT	RF	CNN	EN	GT	RF	CNN	EN	GT	RF	CNN	EN
C	NC	C	C	NC	NC	C	NC	NC	NC	C	C
100%	52%	79%	65%	100%	99%	58%	68%	100%	89%	95%	57%
Attribute	Value		PC	Attribute	Value		PC	Attribute	Value		PC
Ventricles	25859		.163	FAQ	0		.033	FAQ	7		.147
Hippocampus	7159		.07	ADAS13	5		.022	APOE4	0		−.068
Age	77		.057	APOE4	0		.019	MMSE	29		−.028
Race	White		.054	Age	71		.014	Race	White		.026
ADAS11	8		.041	Hippocampus	7851		.01	Age	75		−.026
ADAS13	15		.035	ADAS11	2		.008	Hippocampus	6676		−.023
APOE4	0		−.015	MMSE	30		.004	Ventricles	34505		−.022
MMSE	29		.002	Ventricles	23127		.002	ADAS11	8		−.012
FAQ	1		.001	Race	White		−.002	ADAS13	19		−.002

brain. However, despite the model weighting favoring the visual analysis, the ensemble could still make the correct prediction with a final confidence level of 68%.

Patient 4897 represents an instance where our ensemble model provided an incorrect prediction (EMCI_C, 57% confidence). Despite the clinical data pointing toward an EMCI_NC classification, the DTI prediction confidence won out (95% CNN vs. 89% RF). From Fig. 13, we see that the visual model focused heavily on the ventricles, whereas the RF feature ranking placed ventricular volume as 7th in predictive power for this specific individual.

Overall, with our EMCI_C subset, 38% of the ensemble predictions had disagreements between the RF and CNN model but resulted in a correct ensemble prediction. For EMCI_NC, 4.3% of the predictions encountered disagreements between modalities. Given these findings, we see that the ensemble benefits conversion class prediction more significantly than the EMCI_NC class.

Table 8 shows that our proposed model outperforms recently published multi-modality models for AD conversion prediction. A defining difference is our usage of DTI over traditional sMRI and our ensemble classification in place of a single classifier. While many competing authors leverage multiple modalities, they typically limit their studies to a single classifier rather than an ensemble approach. Additionally, our model can predict from 5 to 7 years out due to focusing on EMCI rather than the more general MCI data set.

5. Limitations/future work

A limitation of our study is that all patients were derived from the ADNI data set. For our work, this was acceptable. However, a clinical setting implementation of our model could benefit from additional data sets to account for further feature variation. In addition, these different ADNI studies (ADNI1, ADNIGO, etc.) also leveraged scanners with varying field strengths, which could have affected our results. This will be a consideration for our future work as we aim for more robust modeling.

Another limitation is that gradient-based saliency techniques have shown some unreliability regarding medical imaging [24]. For future work, alternate mappings will be explored and evaluated against our Grad-Cam maps.

Future work will explore performing time-series analysis via an ensemble model in addition to binary classification. This would allow patient progression trajectories to be determined rather than distinguishing between EMCI_C and EMCI_NC. In addition, the generated heatmaps from our CNN model would also be aligned to segmented brain areas to explore potential findings from that relationship.

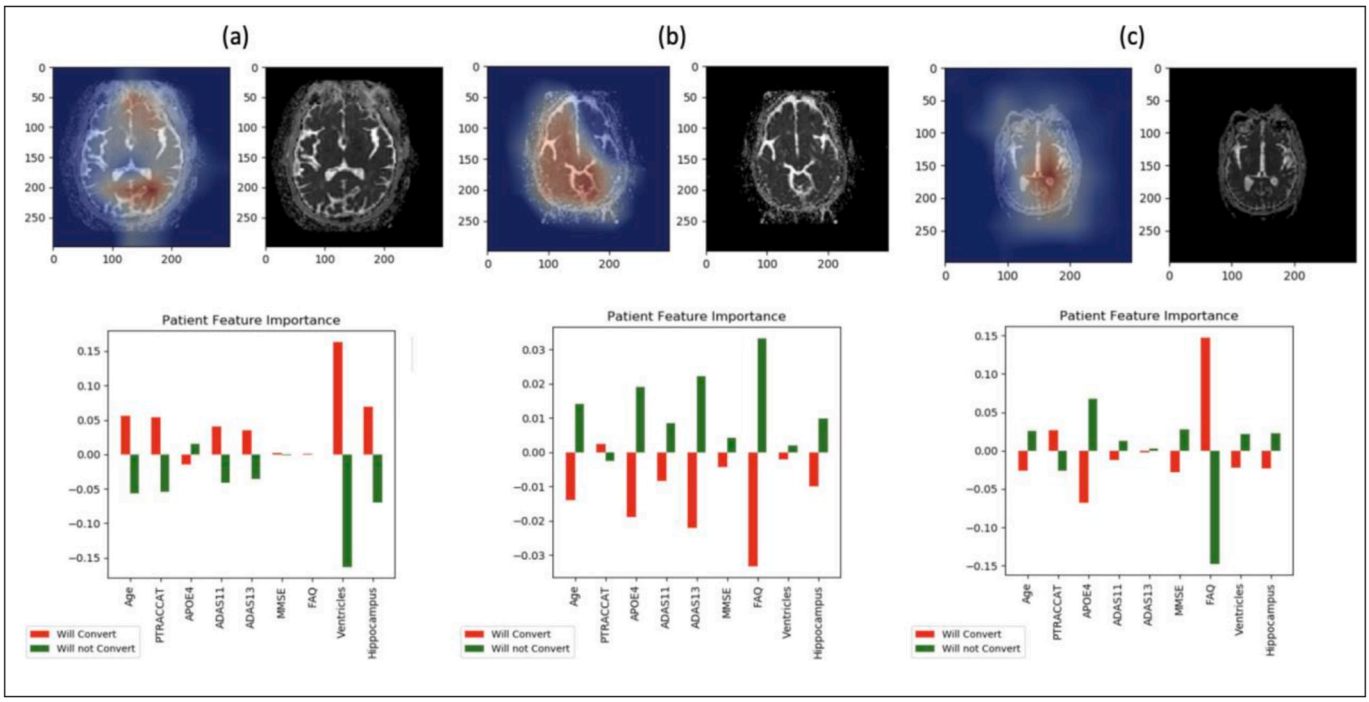

Fig. 13. Grad-CAM explainability (heat map overlay on left, intake image on right) and feature importance ranking: (a) Patient 2106, (b) patient 4220, (c) patient 4897.

Table 8
MCI-to-AD conversion prediction model comparison.

Approach	Modalities	Data	Model	MCI-to-AD Pred.		Year
		(subject size)		ACC	AUC	
Proposed Model (ours)	Clinical data/DTI	ADNI (383)	Ensemble (RF/CNN)	**98.81%**	**99.2%**	5
CNN Model (2019) (ours)	DTI	ADNI (383)	CNN	96.43%	98.1%	5
RF Model (2021) (ours)	Clinical data	ADNI (383)	RF	92.86%	96%	5
Zhang [5]	sMRI/rs-fMRI	ADNI (108)	SVM	84.71%	88.8%	3
Minhas [6]	Clinical data/MRI	ADNI (85)	SVM	81%	95.7%	1
Pan et al. (2020)	MRI	ADNI (509)	Ensemble (CNN/EL)	62%	59%	3
Lin et al. [7]	Clinical data/MRI/FDG-PET	ADNI (617)	ELM	84.7%	88.8%	3
Rana et al. [11]	Clinical data/MRI	ADNI (559)	CNN	69.8%	83%	5
Grassi et al. [8]	Clinical data	ADNI (550)	SVM	–	88%	3
Huang et al. (2019)	Clinical data/MRI	ADNI (290)	SVM	80%	84.6%	5
Varatharajah et al. [10]	Clinical data/MRI/FDG-PET	ADNI (135)	SVM	93%	93%	3

6. Conclusion

An ensemble model for EMCI to AD conversion probability within five years is proposed. Either DTI scans, clinical data, or both can be used for this reason. First, our balanced random forest assesses the clinical data input before our CNN evaluates the DTI scan. Each modality generates separate prediction confidence, which is then factored into our ideal model weight (45% RF, 55% CNN). With this approach, our model achieves an accuracy of 98.8% on EMCI to AD conversion prediction within five years. In this study, we observed that DTI scans are better at AD conversion prediction (96.43%) than clinical data alone (92.86%). We also demonstrated ensemble explainability by employing clinical data feature ranking and Grad-CAM analysis for DTI heat map generation. This allows for greater confidence and understanding of the prediction rationale when framing our model as a decision-support tool.

Acknowledgement

The coauthor, Yugyung Lee, would like to acknowledge the partial support of the National Science Foundation (NSF) Grant No. 1747751, 1935076, and 1951971.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; Eurolmmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (Julie Gerberding). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic

Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] D. International, World Alzheimer Report 2011: the Benefits of Early Diagnosis and Intervention; World Alzheimer Report 2011: the Benefits of Early Diagnosis and Intervention. Technical Report, Alzheimer's Disease International, 2011. URL: www.alz.co.uk/worldreport2011.

[2] M. Velazquez, Y. Lee, Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects, PLoS One 16 (2021), e0244773, https://doi.org/10.1371/journal.pone.0244773.

[3] M. Velazquez, R. Anantharaman, S. Velazquez, Y. Lee, RNN-based alzheimer's disease prediction from prodromal stage using diffusion tensor imaging, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) RNN-Based, 2019, pp. 1665–1672, https://doi.org/10.1109/bibm47256.2019.8983391.

[4] Y. Qi, Random forest for bioinformatics, in: Ensemble Machine Learning, Springer, 2012, pp. 307–323.

[5] T. Zhang, Q. Liao, D. Zhang, C. Zhang, J. Yan, R. Ngetich, J. Zhang, Z. Jin, L. Li, Predicting MCI to AD conversation using integrated sMRI and rs-fMRI: machine learning and graph theory approach, Front. Aging Neurosci. 13 (2021) 429, https://doi.org/10.3389/FNAGI.2021.688926/BIBTEX.

[6] S. Minhas, A. Khanum, A. Alvi, F. Riaz, S.A. Khan, F. Alsolami, M. A Khan, Early MCI-To-AD Conversion Prediction Using Future Value Forecasting of Multimodal Features, Computational intelligence and neuroscience 2021, 2021, https://doi.org/10.1155/2021/6628036.

[7] W. Lin, Q. Gao, J. Yuan, Z. Chen, C. Feng, W. Chen, M. Du, T. Tong, Predicting alzheimer's disease conversion from mild cognitive impairment using an Extreme learning machine-based grading method with multimodal data, Front. Aging Neurosci. (2020) 77, https://doi.org/10.3389/FNAGI.2020.00077, 0.

[8] M. Grassi, N. Rouleaux, D. Caldirola, D. Loewenstein, K. Schruers, G. Perna, M. Dumontier, A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures, Front. Neurol. 10 (2019), https://doi.org/10.3389/fneur.2019.00756.

[9] K. Huang, Y. Lin, L. Yang, Y. Wang, S. Cai, L. Pang, X. Wu, L. Huang, A multipredictor model to predict the conversion of mild cognitive impairment to alzheimer's disease by using a predictive nomogram, Neuropsychopharmacology 45 (2020) 358–366.

[10] Y. Varatharajah, V.K. Ramanan, R. Iyer, P. Vemuri, Predicting short-term mci-to-ad progression using imaging, csf, genetic factors, cognitive resilience, and demographics, Sci. Rep. 9 (2019) 1–15.

[11] S.S. Rana, X. Ma, W. Pang, E. Wolverson, A multi-modal deep learning approach to the early prediction of mild cognitive impairment conversion to alzheimer's disease, in: 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), IEEE, 2020, pp. 9–18.

[12] F. Viton, M. Elbattah, J.L. Guerin, G. Dequen, Heatmaps for visual explainability of CNN-based predictions for multivariate time series with application to healthcare, in: 2020 IEEE International Conference on Healthcare Informatics, ICHI 2020, 2020, https://doi.org/10.1109/ICHI48887.2020.9374393.

[13] B.M. Maweu, S. Dakshit, R. Shamsuddin, B. Prabhakaran, CEFEs: a CNN explainable framework for ECG signals, Artif. Intell. Med. 115 (2021), 102059, https://doi.org/10.1016/J.ARTMED.2021.102059.

[14] R. Mostafiz, M.S. Uddin, N.A. Alam, M. Mahfuz Reza, M.M. Rahman, Covid-19 detection in chest X-ray through random forest classifier using a hybridization of deep CNN and DWT optimized features, Journal of King Saud University - Computer and Information Sciences (2020), https://doi.org/10.1016/J.JKSUCI.2020.12.010.

[15] I. Priyadarshini, V. Puri, A convolutional neural network (CNN) based ensemble model for exoplanet detection, Earth Sci. Inform. 14 (2 14) (2021) 735–747, https://doi.org/10.1007/S12145-021-00579-5, 2021.

[16] Alzheimer's Therapeutic Research Institute ADNI Team, Alzheimer's disease neuroimaging initiative (ADNI), URL: http://adni.loni.usc.edu/study-design/.

[17] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697–8710.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[19] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.J. Li, L. Fei-Fei, A. Yuille, J. Huang, K. Murphy, Progressive neural architecture search, in: Proceedings of the European Conference on Computer Vision, ECCV), 2018, pp. 19–34.

[20] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[21] O. Aydemir, A new performance evaluation metric for classifiers: polygon area metric, J. Classif. 38 (2020) 16–26, https://doi.org/10.1007/S00357-020-09362-5.

[22] W. La Cava, C. Bauer, J.H. Moore, S.A. Pendergrass, Interpretation of machine learning predictions for patient outcomes in electronic health records, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2019, p. 572.

[23] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-cam: Why Did You Say that?, 2016 arXiv preprint arXiv:1611.07450.

[24] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, J. Adebayo, M.D. Li, J. Kalpathy-Cramer, Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging, 2021, https://doi.org/10.1148/RYAI.2021200267.