A Comparison of Deep Learning and Traditional Machine Learning Approaches in Detecting Cognitive Impairment Using MRI Scans

Wei Liu, Jiarui Zhang, and Yijun Zhao
Computer and Information Sciences Department, Fordham University, New York, NY 10023

***Abstract*—Deep learning has attracted a great amount of interest in recent years and has become a rapidly emerging field in artificial intelligence. In medical image analysis, deep learning methods have produced promising results comparable to and, in some cases, superior to human experts. Nevertheless, researchers have also noted the limitations and challenges of the deep learning approaches, especially in model selection and interpretability. This paper compares the efficacy of deep learning and traditional machine learning techniques in detecting cognitive impairment (CI) associated with Alzheimer's disease (AD) using brain MRI scans. We base our study on 894 brain MRI scans provided by the open access OASIS platform. In particular, we explore two deep learning approaches: 1) a 3D convolutional neural network (3D-CNN) and 2) a hybrid model with a CNN plus LSTM (CNN-LSTM) architecture. We further examine the performance of five traditional machine learning algorithms based on features extracted from the MRI images using the FreeSurfer software. Our experimental results demonstrate that the deep learning models achieve higher Precision and Recall, while the traditional machine learning methods deliver more stability and better performance in Specificity and overall accuracy. Our findings could serve as a case study to highlight the challenges in adopting deep learning-based approaches.**

***Index Terms*—machine learning, deep learning, Alzheimer's disease, MRI, brain imaging**

I. Introduction

Alzheimer's disease (AD) is a progressive neurological disorder that results in degenerated or dead brain cells. It is the most common type of dementia, affecting approximately 6.2 million Americans in 2021 [1]. In its early stage, patients suffer from symptoms such as forgetting recent events or conversations. Early detection of cognitive impairment (CI) is critical in identifying individuals at high risk for conversion to AD, and consequently, providing them with proper care, management, and potential interventions.

The MRI scan, which reveals the brain's anatomic structure, is widely used to identify the loss of brain mass associated with Alzheimer's disease and other dementias. However, accurate interpretation of brain MRI scans requires years of training and experience because other conditions such as tumors, hemorrhage, stroke, and hydrocephalus can masquerade as Alzheimer's disease. With the latest advances in computer vision, researchers have resorted to automatic models to detect brain abnormalities associated with AD. We give a brief survey of work in this domain in Section II.

In recent years, deep learning has attracted a great amount of interest. In medical image analysis, deep neural networks have been successfully applied to various clinical applications, including in-scanner head motion detection [2], brain MRI image artifacts reduction [3], and bone X-ray abnormality detections [4]. Nevertheless, researchers are starting to assess the limitations and challenges of the deep learning approaches [5]–[8]. Traditional machine learning approaches typically need to first perform feature engineering to obtain effective and robust features before building predictive models. On the other hand, deep learning models rely on their model structures to simultaneously perform feature extraction and model training, which could lead to inferior results compared to models leveraging information generated by some mature feature extraction methodologies. Furthermore, the exceedingly large hypothesis space arising from a deep model's expressive power makes the model selection a challenging task and likely to result in only substandard solutions.

Our study compares the efficacy of deep learning models and traditional machine learning algorithms in detecting cognitive impairment (CI) based on brain MRI scans. Our objective is to build effective binary classification models to classify subjects as cognitively normal (CN) or cognitively impaired, according to the clinical dementia ratings (CDR) provided by the domain experts. For the deep learning models, we explore a 3D convolutional neural network (CNN) model and a hybrid CNN plus LSTM model, in which the convolutional layers serve the purpose of feature extraction and the LSTM layers exploit the temporal dependencies of the MRI slices. The 3D CNN model operates directly on the 3D MRI scans, and the CNN-LSTM model operates on 2D slices of three anatomical planes. For the traditional machine learning approaches, we investigate five established algorithms, i.e., Decision Trees (DT), Random Forest (RF), Neural Network (NN), Support Vector Machine (SVM), and Naive Bayes (NB). These algorithms operate on 195 volumetric features extracted from individual MRI scans using the FreeSurfer software.

The contribution of our research is twofold: first, we investigate effective algorithms to automatically classify cognitively impaired patients from healthy controls using brain MRI scans. Second, we compare the performance of deep learning and traditional machine models and highlight some challenges associated with adopting deep learning-based approaches.

TABLE I
CDR STATISTICS OF MRI SCANS

CDR	Count	Max Age	Min Age	Mean Age
0.0	607	97	42	70.08
0.5	209	91	46	70.18
1.0	72	91	43	67.03
2.0	6	76	50	66.63

II. RELATED WORK

Cognitive impairment (CI) leads to an increased risk of developing Alzheimer's disease (AD). Automated detection of brain atrophy through MRI is an active research area. In earlier studies, Plant et al. developed a novel data mining framework in combination with three different classifiers (SVM, Bayes statistics, and VFI) to predict the conversion from CI to AD based on MRIs [9]. Their best model achieved 75% accuracy for the prediction of the conversion from mild CI to AD. Trambaiolli et al. used Support Vector Machines (SVM) to search patterns in electroencephalography (EEG) epochs to differentiate AD patients from healthy controls. Their results obtained from analysis of EEG epochs were accuracy 79.9% and sensitivity 83.2%. The analysis considering the diagnosis of each individual patient reached 87.0% accuracy and 91.7% sensitivity. Zhang et al. proposed a novel classification system to distinguish among elderly subjects with Alzheimer's disease (AD), mild cognitive impairment (MCI), and normal controls (NC) [10]. In particular, they constructed a kernel support vector machine decision tree (kSVM-DT), which achieves 80% classification accuracy.

In recent years, there has been a rapid growth of deep learning-based approaches in AD-related studies. These models can be further classified into three categories depending on the type of backbone neural network adopted in the model architecture. The first category consists of approaches based on vanilla convolutional neural networks (CNN) [11]). CNN is the most popular underlying structure adopted by researchers, attributing to its proficiency in extracting image features. The other two categories are recurrent neural network (RNN) based and Generative Adversarial Network (GAN) based approaches. For example, Lin et al. designed a deep learning approach based on convolutional neural networks (CNN) to predict mild CI to AD conversion with magnetic resonance imaging (MRI) data [12]. Their approach achieved an accuracy of 86.1% in leave-one-out cross-validations while keeping a good balance between the sensitivity and specificity. Cui et al. proposed an RNN-based longitudinal analysis for diagnosis of Alzheimer's disease [13]. Their results achieved a classification accuracy of 91.33% in detecting AD against normal controls and 71.71% for progressive mild CI vs. stable mild CI. Han et al. proposed a two-step method using GAN-based multiple adjacent brain MRI slice reconstruction to detect AD at various stages [14]. Their approach is fully unsupervised, which also discover and alert any anomalies including rare disease.

Although deep learning models have brought unprecedented potential in automatic disease diagnosis using MRI images, researchers have noted the limitations and challenges associated with these approaches [5]–[8]. As Bhatt et al. pointed out, from the literature study, it has been found that most of the DL models are application- or equipment-dependent. Most of the authors did not address the reason for the selection of DL models such as CNN or RNN. Furthermore, there are many parameters and framework tuning in the model training process, making the model selection challenging even for experienced researchers.

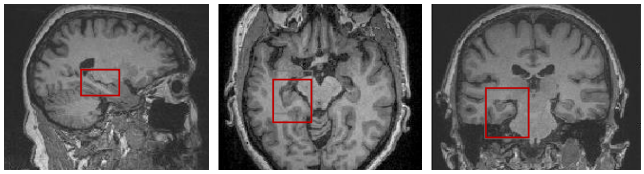

Fig. 1. Sample MRI Images from the OASIS dataset. Red boxes indicated extracted hippocampus regions for our study, viewed on sagittal (left), transverse (middle), and coronal (right) planes.

III. MATERIALS

A. Data

As mentioned in Section I, our study leverages the OASIS-3 dataset, which is the latest release in the Open Access Series of Imaging Studies (OASIS) hosted by ADRC [15]. The dataset contains over 2000 MR sessions from 1098 participants collected over the course of 15 years. To conduct our comparison study, we selected scans that have accompanying volumetric segmentation files produced through FreeSurfer processing. Furthermore, to reduce data noise, we filtered out scans that are not of standard size (i.e., 256x256x176). As a result, we retained a total of 897 scans for our study.

We created our class labels using the clinical dementia rating (CDR) scores [16]. CDR is a 5-point scale used to characterize the following six domains of cognitive and functional performance applicable to Alzheimer's disease and related dementias: Memory, Orientation, Judgment and Problem Solving, Community Affairs, Home and Hobbies, and Personal Care. Specifically, values 0, 0.5, 1, 2, 3 are used to indicate no, very mild, mild, moderate, and severe dementia, respectively. Table I presents the CDR statistics of our dataset. In our study, we divide the patients into two categories for a binary classification task: cognitively normal (CN) with CDR=0 and cognitively impaired (CI) with CDR>0.

B. Data Preprocessing for the Deep Learning Models

Since 3D convolution is computationally expensive and CI is known to be affecting the hippocampus area [17], we followed a similar process as Lin et al. [12] and focused only on the hippocampus region. To this end, we first aligned the scans using the Flirt software provided with FSL [18] and then retained only the temporal lobe regions by removing the rest sections to reduce the noise and computational cost for the 3D-CNN model. Consequently, the 3D input to our model is extracted from coordinates (50:80, 90:130, 110:125) in each scan. Figure 1 presents sample images viewed on the three anatomical planes. The red boxes indicate the regions selected for our study.

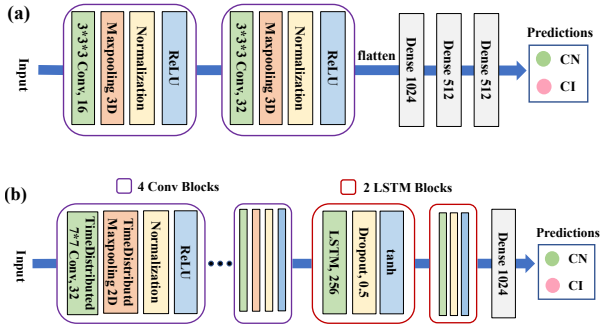

Fig. 2. Deep learning Architectures. (a) 3D-CNN (b) CNN-LSTM

C. Data Preprocessing for the Traditional Machine Learning Models

Using data processing scripts provided by OASIS, we obtained five tables for our dataset, i.e., a table of white matter parcellation volumes, two tables of cortical parcellation for each of the left and right hemispheres, and two tables of the average thickness of each cortical parcellation. We join the five tables using unique MRI session IDs and extracted 205 features. Lastly, we cross referenced the 205 features with the FreeSurfer variables provided by the OASIS-3 Data Dictionary and retained 188 brain volumetric attributes as the predictive features for the traditional machine learning models.

IV. Methods

We explored five established traditional ML methods for our classification task: SVM, Decision Trees, Random Forest, Neural Netork, and Naive Bayes. For the deep learning models, we investigated a 3D and a 2D hybrid models as illustrated in Sections IV-A and IV-B.

A. 3D-CNN Model

Figure 2(a) presents the architecture of our 3D-CNN model. It is worth noting that, due to expensive computational cost and high expressive power, 3D convolutional models typically do not need as deep structure as 2D CNNs. In our model, we employed two Conv3D layers with 16 and 32 filters, respectively. Both convolutional layers utilized a 3x3x3 kernel and the ReLU activation function. A Maxpooling3D layer with a 2x2x2 kernel follows each of the convolutional layers. Batch normalization is applied after each Maxpooling3D. The final flattened layer contains 1,280 units, which are connected to a sequence of dense layers with 1,024, 512, and 128 units, respectively. A 50% dropout layers is added after each dense layer to regularize overfitting. The final output layer has two units activated using the softmax function, representing our two classes.

B. CNN-LSTM Model

Our CNN-LSTM model is a hybrid approach to capture the spacial and temporal characteristics of sequential MRI images along the three anatomical planes (i.e., sagittal, transverse, and coronal). We converted a 3D MRI scan into sequential 2D slices along the three anatomical planes and trained a sub model for images from the same plane. The final CNN-LSTM model's output is an ensemble of the predictions of three sub-models.

Figure 2(b) presents the architecture of our CNN-LSTM approach. The model starts with 4 TimeDistributed Conv2D layers, and each is followed by a max pooling layer and a batch normalization layer. These convolutional layers serve the purpose of extracting high-level predictive features from the MRI images. The final flatten layer from the CNN block is fed as the input to the LSTM model, which consists of two LSTM blocks followed by two dense layers.

V. Experimental Results

A. Model Training

Due to the high computational cost associated with training deep learning models, we trained our 3D-CNN and CNN-LSTM models using an 8:1:1 split for the training, validation, and test, respectively. For the traditional machine learning models, we conducted our experiments using a 10-fold cross-validation and reported the average model performance on the test folds. All models were regularized using standard techniques, including batch normalization, L_2 regularization, and dropout, to reduce overfitting.

We observe in Table I that our data is imbalanced with a class 1 to class 0 ratio of 607:287. To address this issue, we will incorporate cost-sensitive learning in our model training. Specifically, misclassification of the minority instances will incur a more significant penalty than that of the majority ones in the model training process. The optimal class weights were selected as hyper-parameters using either the validation set (i.e., for DL models) or a nested 10-fold cross-validation on the training data (i.e., for traditional ML models).

We trained our deep learning models on a PowerEdge R740 Linux machine with two Xeon 2.60GHz CPUs (12 cores), 192GB of memory, and a 32GB NVIDIA Tesla V100 GPU. The training converged in approximately 24 hours for each model with a learning rate of 0.0005.

B. Performance Evaluation

Table II presents the main results of our experiments. We observe that the CNN-LSTM model significantly outperformed the 3D-CNN model in overall accuracy (3D-CNN:0.53; CNN-LSTM:0.63), Recall (3D-CNN:0.54; CNN-LSTM:0.77), Specificity (3D-CNN:0.50; CNN-LSTM:0.56), and F1 score (3D-CNN:0.62; CNN-LSTM:0.74). 3D-CNN has slightly better performance in Precision (3D-CNN: 0.74; CNN-LSTM:0.71). The experimental results suggest that the LSTM component could be effective in capturing additional predictive information for our classification task.

Of the five traditional machine learning models, DT, NN, RF, and SVM exhibit similar and robust performance across all five evaluation metrics, with overall accuracies in [0.75, 0.76], Precisions in [0.60, 0.63], Recalls in [0.64, 0.68], Specificities in [0.79, 0.82], and F1 scores in [0.62, 0.64]. The NB model has a noticeably worse performance than the other four models, especially in Precision and Specificity. One

TABLE II
PERFORMANCE COMPARISON OF DEEP LEARNING AND TRADITIONAL MACHINE LEARNING APPROACHES

Model	Precision	Recall	Specificity	F1	Accuracy
	Deep Learning Models				
3D-CNN	0.74	0.54	0.50	0.62	0.53
CNN-LSTM	0.71	0.77	0.56	0.74	0.63
	Traditional Machine Learning Models				
DT	0.62	0.65	0.81	0.63	0.76
NN	0.63	0.64	0.82	0.63	0.76
NB	0.56	0.67	0.75	0.60	0.72
RF	0.61	0.68	0.79	0.64	0.76
SVM	0.60	0.64	0.80	0.62	0.75

explanation is that the volumetric features are not mutually independent, which violates the Naive Bayes assumption

Comparing the deep learning and traditional machine learning approaches, we observe that the CNN-LSTM model demonstrates higher Precision and Recall, but its performance is unsatisfactory in Specificity (0.56), resulting in a low overall accuracy (0.63). It is also interesting to note that the CNN-LSTM model is more proficient at predicting class 1 (i.e., CI, Recall=0.77) than class 0 (i.e., CN, Specificity=0.56), while the traditional models showed the opposite quality.

Our experimental results suggest that, for our task, the traditional machine learning methods exhibit more robust, stable, and better overall performance than deep learning-based approaches. Our findings may not be conclusive due to the limitations in our selection of the deep learning models. A more exhaustive search may yield more fruitful results, which highlights the challenge of general deep learning-based approaches, i.e., the choice of model, both in type and architecture, is more of an art than science. Given the highly flexible structure of deep networks and their high training costs, searching for a suitable model can be time consuming and expensive. On the other hand, for our particular task, traditional machine learning models demonstrate overall superior performance leveraging a mature feature extraction technique. Although our models' performance is not perfect, we feel that the comparison results we have obtained merit sharing with a wider audience.

VI. CONCLUSION

In this paper, we investigated two deep learning and five traditional machine learning approaches in detecting cognitive impairment using brain MRI scans. For the deep learning models, we explored a 3D-CNN model and a hybrid CNN-LSTM model. The CNN-LSTM significantly outperformed the 3D-CNN model in four out of the five evaluation metrics, suggesting its effectiveness in capturing the temporal dependencies in the MRI slices. We built our machine learning models using 188 volumetric features extracted using the FreeSurfer software. Although the machine learning models exhibit lower Precision and Recall than the deep learning models, they demonstrated superiority in overall accuracy, Specificity, and reliable stability. The success of the traditional machine learning algorithms in our study could be due to our problem's idiosyncratic nature that makes it particularly amenable to handling with such methods. The presence of adequate feature engineering techniques (e.g., FreeSurfer) could have further helped these methods in tackling the task. Nevertheless, our findings are consistent with the recognized challenges in developing deep learning models and cast a positive light on the value of traditional techniques.

REFERENCES

[1] A. Association *et al.*, "2010 alzheimer's disease facts and figures," *Alzheimer's & dementia*, vol. 6, no. 2, pp. 158–194, 2010.

[2] H. R. Pardoe, S. P. Martin, Y. Zhao, A. George, H. Yuan, J. Zhou, W. Liu, and O. Devinsky, "Estimation of in-scanner head pose changes during structural mri using a convolutional neural network trained on eye tracker video," *bioRxiv*, 2021.

[3] Y. Zhao, J. Ossowski, X. Wang, S. Li, O. Devinsky, S. P. Martin, and H. R. Pardoe, "Localized motion artifact reduction on brain mri using deep learning with effective data augmentation techniques," *arXiv preprint arXiv:2007.05149*, 2020.

[4] M. He, X. Wang, and Y. Zhao, "A calibrated deep learning ensemble for abnormality detection in musculoskeletal radiographs," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[5] C. Bhatt, I. Kumar, V. Vijayakumar, K. U. Singh, and A. Kumar, "The state of the art of deep learning models in medical science and their challenges," *Multimedia Systems*, pp. 1–15, 2020.

[6] M. P. Véstias, "Deep learning on edge: Challenges and trends," *Smart Systems Design, Applications, and Challenges*, pp. 23–42, 2020.

[7] G. Futia and A. Vetrò, "On the integration of knowledge graphs into deep learning models for a more comprehensible ai—three challenges for future research," *Information*, vol. 11, no. 2, p. 122, 2020.

[8] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[9] C. Plant, S. J. Teipel, A. Oswald, C. Böhm, T. Meindl, J. Mourao-Miranda, A. W. Bokde, H. Hampel, and M. Ewers, "Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease," *Neuroimage*, vol. 50, no. 1, pp. 162–174, 2010.

[10] Y.-D. Zhang, S. Wang, and Z. Dong, "Classification of alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree," *Progress In Electromagnetics Research*, vol. 144, pp. 171–184, 2014.

[11] M. Pereira, I. Fantini, R. Lotufo, and L. Rittner, "An extended-2d cnn for multiclass alzheimer's disease diagnosis through structural mri," in *Medical Imaging 2020: Computer-Aided Diagnosis*, vol. 11314. International Society for Optics and Photonics, 2020, p. 113141V.

[12] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu *et al.*, "Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment," *Frontiers in neuroscience*, vol. 12, p. 777, 2018.

[13] R. Cui, M. Liu, A. D. N. Initiative *et al.*, "Rnn-based longitudinal analysis for diagnosis of alzheimer's disease," *Computerized Medical Imaging and Graphics*, vol. 73, pp. 1–10, 2019.

[14] C. Han, L. Rundo, K. Murao, Z. Á. Milacski, K. Umemoto, E. Sala, H. Nakayama, and S. Satoh, "Gan-based multiple adjacent brain mri slice reconstruction for unsupervised alzheimer's disease diagnosis," in *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer, 2019, pp. 44–54.

[15] P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. Vlassenko *et al.*, "Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease," *MedRxiv*, 2019.

[16] J. C. Morris, "The clinical dementia rating (cdr): Current version and," *Young*, vol. 41, pp. 1588–1592, 1991.

[17] Y. Mu and F. H. Gage, "Adult hippocampal neurogenesis and its role in alzheimer's disease," *Molecular neurodegeneration*, vol. 6, no. 1, pp. 1–9, 2011.

[18] S. Smith, P. R. Bannister, C. Beckmann, M. Brady, S. Clare, D. Flitney, P. Hansen, M. Jenkinson, D. Leibovici, B. Ripley *et al.*, "Fsl: New tools for functional and structural brain image analysis," *NeuroImage*, vol. 13, no. 6, p. 249, 2001.