Predictive Modeling of Longitudinal Data for Alzheimer's Disease Diagnosis Using RNNs

Maryamossadat Aghili[1(✉)], Solale Tabarestani[1], Malek Adjouadi[1], and Ehsan Adeli[2]

[1] Florida International University, Miami, USA
maghi001@fiu.edu
[2] Stanford University, Stanford, USA

Abstract. In this paper, we study the application of Recurrent Neural Networks (RNNs) to discriminate Alzheimer's disease patients from healthy control individuals using longitudinal neuroimaging data. Distinctions between Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and healthy subjects in a multi-modal heterogeneous longitudinal dataset is a challenging problem due to high similarity between brain patterns, high portions of missing data from different modalities and time points, and inconsistent number of test intervals between different subjects. Due to these challenges, to distinguish AD patients from healthy subjects, conventionally researchers use cross-sectional data when applying deep learning methods in neuroimaging applications. Whereas we propose a method based on RNNS to analyze the longitudinal data. After carefully preprocessing the data to alleviate the inconsistency due to different data sources and various protocols of capturing modalities, we arrange the data and feed it into variations of RNNs, i.e., vanilla Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). The accuracy, *F*-score, sensitivity, and specificity of our models are reported and are compared with the most immediate baseline method, multi-layer perceptron (MLP).

Keywords: Long Short Term Memory (LSTM) · Gated Recurrent Unit (GRU)
Recurrent Neural Networks (RNNs) · Alzheimer's Disease (AD)
Longitudinal data · Prognosis · Diagnosis

1 Introduction

Alzheimer's disease (AD) is one of the most frequent types of dementia, which leads to memory loss and other cognitive disabilities. As the majority cases of dementia fall in the Alzheimer's category, diagnosis and prognosis of this disease, especially in the early stage, has exceptional importance [1–3]. Early diagnosis, before the occurrence of the irreversible brain deformation, enables early treatment and plays a significant role in patient care, prediction of the progression risks, and severity recognition [3–5]. However, regardless of enormous efforts, pinpointing the prodromal stage of mild cognitive impairment is remained an open research field. Having incomplete samples in the longitudinal medical studies is a common phenomenon, as many patients may miss some of the tests and modalities in a time step or miss a complete visit within the

I. Rekik et al. (Eds.): PRIME 2018, LNCS 11121, pp. 112–119, 2018.
https://doi.org/10.1007/978-3-030-00320-3_14

study's lifespan. Generally, missing values occur for a variety of reasons including drop out of subjects from the study, insufficient resolution, image corruption, budget limitation, etc. [5–7]. Many algorithms simply discard subjects with missing modalities from further experiments, which indeed results in a considerable loss of valuable information. Disease diagnosis accuracy might be improved if the missing parameters could be estimated correctly from the rest of the available data or modalities. Furthermore, to have a better understanding of the disease progression and to correctly label a subject as Normal Control (NC), Mild Cognitive Impairment (MCI), or dementia (i.e., AD), data from every visit should not be scrutinized independently from the earlier steps. Currently, a majority of the classification algorithms focus on the cross-sectional data and only analyze a specific interval's biomarkers for the diagnosis and disregard the former patient's status for the decision making process. To address this shortcoming, recent studies moved toward longitudinal data analysis and proposed new methods to leverage valuable temporal data by considering the inherent correlations of such data [6–8].

Effectively mining AD longitudinal data is a challenging task, owing to its heterogeneous measurements, varying length of samples, missing modalities and tests, and small sample size. In this study, for the first time (to the best of our knowledge), we employ two RNN models, namely the Long Short Term Memory (LSTM) and the Gated Recurrent Unit (GRU), to discover the regression patterns of the subjects from the longitudinal data with missing variables and intervals, especially for the task of classifying AD/MCI *vs.* NC, which is a challenging task only depending on the cross-sectional dataset. The progression of the patients during time should be studied carefully to capture the correct status of the patient through the passage of time. Accordingly, in this study, we conduct several experiments to investigate the effectiveness of the RNNs in AD diagnosis. We compare the outcomes of the LSTM and GRU model with Multi-Layer Perceptron (MLP) to evaluate the efficacy of the sequential models.

2 Dataset

The data used in this study is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether structural magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Recently the largest longitudinal dataset, which is a subset of ADNI 1/Go/2 cohorts, has been extracted from ADNI by Bruno M. Jedynak and Michael Donohue to make a baseline for researchers in the field to propose and apply quantitative templates for the progression of Alzheimer's disease. This is an invaluable baseline for accurate evaluation of the proposed algorithms.

The database has 1721 distinct subjects (521 NC, 864 MCI, and 336 AD) examined every 6 months during 11 years' period making 23 time points for a patient in the case of performing all the test regularly every six month (i.e., baseline, 6 months, 12 months, …, 132 months). For every visit multiple outcomes provided including

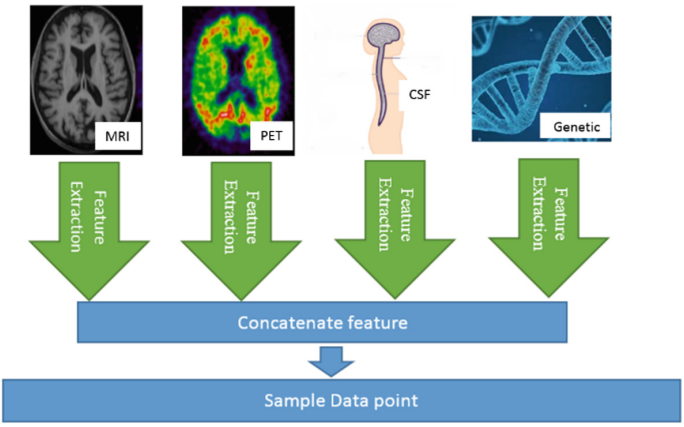

Fig. 1. Sample data point curation

ADAS13, CDRSB, RAVLT.learning MMSE, FAQ, FDG PET, Amyloid PET, CSF, ABETA, CSF TAU, CSF PTAU, FS WholeBrain, FS Hippocampus, FS Entorhinal, FS Ventricles, FS MidTemp, FS Fusiform and the covariates: age, APOE4 (yes/no), Gender, Education. The primary phenotype is the diagnostic group and Mini-Mental State Examination (MMSE). Sample data-point curation pipeline in our work is presented in Fig. 1. This figure shows that the samples are composed of features extracted from volumetric magnetic resonance imaging (MRI) including cortical thickness, hippocampal volume and shape along with fluoro-2-Deoxy-D-glucose, florbetapir F18, and PIB (which is radiotracer capable of highlighting deposits of beta-amyloid) from PET imaging, and some other Cerebrospinal fluid (CSF) features, such as TAU, PTAU and ABETA. Around 12 functional and behavioral assessment results such as Rey's Auditory Verbal Learning Test and Montreal Cognitive Assessment (MoCA) scores are also measured and used as features in this dataset.

The volumetric MRI measurements provide the cortical thickness, volume and shape of hippocampal or voxel-wise tissue probability [1–4] to measure the brain atrophy; 18-Fluoro-DeoxyGlucose PET imaging (FDG-PET) estimates the glucose hypometabolism in bilateral temporal, temporal, occipital areas or posterior cingulated brain regions [5–7]. Furthermore, global cognitive impairment tests are used by clinicians for screening and measuring individuals who are at the risk of AD; or cerebrospinal fluid (CSF) to measure the increase in t-tau, p-tau, or the decrease of amyloid-β, which is a sign of cognitive declination. Therefore, in total 47 features are used to represent each subject at each time point.

3 Models

In this section, we briefly overview the LSTM and GRU models used in our model and then explain our model design using these architectures for classifying the subjects into one of the AD, MCI, or NC categories from longitudinal data.

3.1 Long Short Term Memory Unit (LSTM)

RNNs with internal memory and feedback loop have previously been adopted mostly for processing arbitrary input sequences, like in handwriting recognition, speech recognition, natural language processing, and time series prediction applications. One of the main challenges in applying RNNs to long sequential data is that the gradient of some learnable weights become too small or too large if the network is unfolded for too many time steps. These phenomena are called the exploding and vanishing gradients problem [9]. LSTM was, hence, proposed by Hochreiter et al. for the first time in 1997 to solve the vanishing gradient problem through a gating mechanism [10]. An LSTM has three gates. The first gate determines whether the information should be forgotten or not. The second gate decides about updating the cell state, and the last gate is responsible for the cell output. Since then, several variations of LSTM architecture have been implemented especially with the utilization of Graphics Processing Units (GPUs).

3.2 Gated Recurrent Unit (GRU)

To adaptively capture dependencies of different time scales in each recurrent unit, Cho et al. [11] introduced a gated recurrent unit (GRU). Similar but not the same as LSTM design, GRU has two gates, a reset gate r, and an update gate z. Intuitively, the reset gate determines how to combine the new input with the previous memory, and the update gate defines how much of the past memory to keep around. Having simpler architecture than LSTM with a smaller number of parameters, GRU provides better results in some applications [12] and is less prone to overfitting, especially in cases that there are not enough training data.

3.3 Our Model

RNN models have achieved popularity due to their power in pattern recognition for the time series and sequential data. While there are plenty of research papers on regression and classification modeling of AD data with well-established and novel machine learning techniques, along with many deep convolutional neural networks for 2D and 3D brain MRI classification, number of research works exploiting RNNs for finding the patterns in the AD longitudinal data sets is limited [13–18]. Only a few papers recently adopted them for regression analysis on the clinical medical data [19]. Here, we employ RNN deep learning techniques for the classification of the subjects. All features are normalized by subtracting the mean value of each feature and dividing the result by the standard deviation of that feature in all samples (i.e., using their z-scores), before the analysis. To deal with missing modalities, we simply replace them with zero values. Since our goal is to showcase the usage of RNNs for longitudinal predictive analysis, we leave extensive data imputation experiments for future works. A recent work also models AD progression with RNN models [20]; however our work is different from that in multiple aspects. We use not only MRI features but also PET, Cognitive tests, and genetic features for modeling the disease. We also propose multiple approaches for handling the missing intervals and compare the potential RNN models with each other.

As described in Sect. 2, the dataset contains $N = 1721$ subjects each scanned in 24 different time points. Data from each time point is represented by $n = 47$ features. Figure 2 overviews the data arrangement. A challenge in analyzing longitudinal data sets is dealing with missing data at different time steps for some of the subjects. To address this inconsistency in the data points and to be able to input the data to RNNs, we define three settings: (1) In our first attempt, we fill the missing intervals with zero to create a same input size data for all the subjects and compose a stack of 1721, 2D matrices that all have a set of 47 biomarkers in the columns as features and all the possible time steps in the rows as time steps. We refer to this arrangement as ***zero fill***. (2) In the second attempt, we buffer the data at every time point and replicate it in its next missing interval. This scheme is named as ***replicate fill***. (3) In the last configuration we change the orientation of the input data and stack all the available intervals on top of each other, disregarding the missing intervals and pad them to the maximum size of the possible time steps, this is called ***padding***.

One LSTM and GRU model with the memory of the maximum size of the available time steps, which is 24, are designed to process this stack of data. Each subject's time point data is fed to the corresponding cell along with its final diagnosis label (i.e., AD, MCI, or NC) allowing the model to learn the pattern of the change in the features for each subject. Figure 3 represents this pipeline. In two different sets of experiments, we replace the cells in this figure with LSTM and GRU sets and report the results.

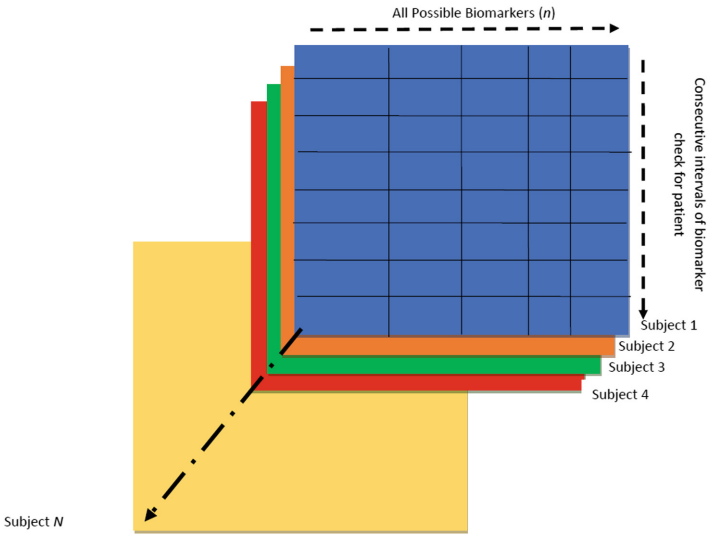

Fig. 2. Data arrangement for the RRN model

4 Experiments

In all the experiments, we train and tune the RNN model with different configurations of the hidden layers, percentages of drop out, various activation functions, loss, optimizers and different combination of other hyperparameters to find the best setting of the model through a grid search. We knowingly made the models as small as possible to avoid overfitting, which can easily mislead the comparison. Data has been split into 70% training, 15% validation set and the rest for the testing set. The best configuration of the LSTM and GRU is represented in Table 1.

For evaluations, we calculate the Accuracy, Sensitivity, Specificity, and F-score of all models. The results of LSTM and GRU models for all arrangements of the data are compared in Table 2, along with the results of their counterpart from non-recurrent networks, i.e., Multi-Layer Perceptron (MLP). The data is flattened to a 1D long vector and fed into the MLP once for each patient.

According to Table 2, LSTM and GRU models are superior to the MLP network in most of the cases as they result in the highest accuracy and F-score. Our LSTM model yields nearly 1% accuracy improvements over MLP in classifying AD patients from NC subjects. Interestingly, the RNN models with the ***zero fill*** data arrangement for the missing data yields consistently better results. The superiority is not significant, which can be mainly due to the limited amount of data in this domain, besides the high portion of the missing time points and modalities. These challenges prevented the vanilla RNNs to find the appropriate patterns despite various input data arrangement. Second, RNNs, especially the LSTM models, have a large number of trainable parameters, which necessitate the model to be trained in a great corpse of sequential data and despite having drop out layers in the architecture, they are still prone to overfitting to the training data in this relatively small dataset. The third is the limited hand engineered and structured feature set, used in this experiment. One of the main superiority of the RNNs is their power in automatic feature learning from the raw data, which can be further explored in the future.

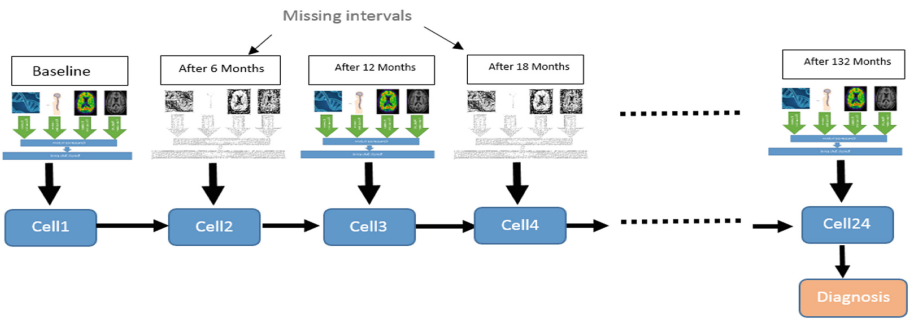

Fig. 3. RNN model used in this study.

Table 1. Model hyperparameters

	Hidden units	Activation function	Layers	Drop out
GRU	32	Softmax	1	0.3
LSTM	30	Softmax	1	0.4
MLP	20	Softmax	2	0.3

Table 2. Performance of the proposed models with three different data arrangements in classification of ADNI subjects. Best results for each data arrangement are underlined, and the best overall results of each column are in bold.

		AD - NC				AD - MCI				NC - MCI			
	Method	Accuracy	F-score	Sensitivity	Specificity	Accuracy	F-score	Sensitivity	Specificity	Accuracy	F-score	Sensitivity	Specificity
ZERO FILL	MLP	0.9467	0.9581	0.9626	0.9194	0.8474	0.8449	0.9405	0.7736	0.7729	0.7539	0.6207	**0.9670**
	LSTM	0.9526	0.9622	0.9532	**0.9516**	**0.8579**	0.8492	0.9048	0.8208	0.7729	**0.7793**	**0.7155**	0.8462
	GRU	**0.9527**	**0.9630**	**0.9720**	0.9194	0.8368	0.8360	0.9405	0.7547	0.7536	0.7536	0.6724	0.8571
REPLICATE FILL	MLP	0.9467	0.9577	0.9533	0.9345	0.8529	0.8492	0.9048	0.8208	0.7005	0.6667	0.5345	0.9121
	LSTM	**0.9586**	**0.9674**	**0.9720**	0.9355	0.8576	**0.8498**	0.9286	**0.8225**	0.7681	0.7757	**0.7155**	0.9352
	GRU	0.9527	0.9626	0.9626	0.9345	0.8211	0.8211	0.9286	0.7358	0.7101	0.7000	0.6034	0.8462
PADDING	MLP	0.9467	0.9577	0.9531	0.9355	0.8421	0.8295	0.8690	0.8208	0.7101	0.7609	0.6877	0.8423
	LSTM	0.9527	0.9623	0.9533	0.9516	0.8468	0.8298	0.8810	0.8219	0.7585	0.7619	0.6897	0.8462
	GRU	0.9408	0.9528	0.9439	0.9355	0.8158	0.8108	0.8929	0.7547	0.7101	0.7000	0.6034	0.8462

5 Conclusion

In this paper, we introduced the applications of LSTM and GRUs to model prediction tasks over the longitudinal data from the ADNI dataset. The proposed models can be used for the diagnosis of Alzheimer's disease. We also incorporated three different strategies to deal with the incomplete and missing data (from time points and modalities). Trying different variations of RNNs (i.e., LSTM and GRU), we found slightly better performance using the LSTM model. Our model can classify AD *vs.* NC with an accuracy of 95.9%, even with simple replicate and zero filling of the missing data. It also performs better classification of AD *vs.* MCI and NC *vs.* MCI patients. As a direction for future works, designing an end-to-end convolutional and LSTM model for this longitudinal dataset can be of great interest, to accurately learn powerful image features (from MRI and PET) and simultaneously learn the classifier parameters.

References

1. Glenner, G.G., Wong, C.W.: Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. Biochem. Biophys. Res. Commun. **120**(3), 885–890 (1984)
2. McKhann, G., Drachman, D., Folstein, M., Katzman, R.: Views & reviews clinical diagnosis of Alzheimer's disease. Neurology **34**(7), 939 (1984)

3. Cuingnet, R., et al.: Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage **56**(2), 766–781 (2011)
4. Petersen, R.C.: Mild cognitive impairment as a clinical entity and treatment target. Arch. Neurol. **62**(7), 1160–1163 (2004). Discussion 1167
5. Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J.: Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage **104**, 398–412 (2015)
6. Nie, L., Zhang, L., Meng, L., Song, X., Chang, X., Li, X.: Modeling disease progression via multisource multitask learners: a case study with Alzheimer's disease. IEEE Trans. Neural Netw. Learn. Syst. **28**(7), 1508–1519 (2017)
7. Zhou, J., Yuan, L., Liu, J., Ye, J.: A multi-task learning formulation for predicting disease progression. In: Proceedings of the 17th ACM SIGKDD KDD, p. 814 (2011)
8. Zhang, D., Shen, D.: Multi modal multi task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage **59**(2), 895–907 (2013)
9. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Nets **5**(2), 157–166 (1994)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
11. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078 (2014)
12. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values, pp. 1–14 (2016)
13. Chen, Y., Shi, B., Smith, C.D., Liu, J.: Nonlinear feature transformation and deep fusion for Alzheimer's disease staging analysis. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (eds.) MLMI 2015. LNCS, vol. 9352, pp. 304–312. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24888-2_37
14. Fang, C., Li, C., Cabrerizo, M., Barreto, A., Andrian, J., Loewenstein, D.: A novel Gaussian discriminant analysis-based computer aided diagnosis system for screening different stages of Alzheimer's Disease. In: BIBE, pp. 279–284 (2017)
15. Shi, J., Zheng, X., Li, Y., Zhang, Q., Ying, S.: Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J. Biomed. Heal. Inform. **2194** (2017)
16. Chaves, R., et al.: SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting. Neurosci. Lett. **461**(3), 293–297 (2009)
17. Zhu, X., Il Suk, H., Wang, L., Lee, S.W., Shen, D.: A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. Med. Image Anal. **38**, 205–214 (2017)
18. Lebedev, A.V., et al.: Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. Neuroimage (Amst) **6**, 115–125 (2014)
19. Bange, S.-J., Wange, Y., Yange, Y.: Phased-LSTM based predictive model for longitudinal EHR data with missing values (2016)
20. Cui, R., Liu, M., Li, G.: Longitudinal analysis for Alzheimer's Disease diagnosis using RNN. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, pp. 1398–1401 (2018)