Usage Of Random Forest Ensemble Classifier Based Imputation And Its Potential In The Diagnosis Of Alzheimer's Disease

Afreen Khan, Swaleha Zubair

Abstract: Objective: To evaluate and compare the performance of Random Forest (RF) ensemble classifier in imputation and non-imputation method of missing data values, and its impact to diagnose Alzheimer's disease (AD) based on longitudinal MRI data. Method: We studied 373 MRI sessions involving 150 AD subjects aged 60 to 90 years [Mean age ± SD = 77.01 ± 7.64]. T1-weighted MRI of each subject on a 1.5-T Vision scanner were used for the image acquisition. The MRI dataset was taken from OASIS (Open Access Series of Imaging Studies) database. Based upon the MRI acquitted features in the dataset, we applied missing data imputation using RF ensemble to classify the subjects as demented or non-demented. We then compared them to determine which is more precise in the AD diagnosis Result: RF model-based imputation analysis outperforms with better accuracy than RF non-imputation method.

Index Terms: alzheimer's disease, classifier, ensemble, imputation, missing data, random forest

1 INTRODUCTION

Alzheimer's Disease (AD) is a slow fatal neurodegenerative disorder [1] that affects the brain cells over time. It is the most widely recognized kind of dementia that affects memory, behaviour and thinking ability. There is a critical increment in the number of AD cases in recent years. As indicated by 2018 statistics, 44 million people worldwide experienced this sickness and related sort of dementia [2]. The "World Alzheimer Report, 2015" updates on AD as it is predicted to influence 75 million in 2030 and 131.5 million by 2050 globally, this number will practically be twofold every 20 years [3]. Numerous factors which may or may not be related to the lifestyle of a patient can trigger off a higher risk for AD. Diagnosing the disorder in its beginning periods is of incredible significance and several techniques are used to diagnose AD. Magnetic resonance imaging (MRI) is one of the imaging methods that effectively support the diagnosis of AD [4]. It creates top-notch images of the anatomical structures of the human body, particularly in the brain, and make available valuable information for the purpose of diagnosis and research [5]. Machine learning (ML) algorithms have been effectively connected in the programmed analysis of AD [6]. The accomplishment of analysis can be characterized as the capacity of the algorithm to anticipate the correct class of unseen data i.e. normal or disease-prone [7]. The prognostic power of such analysis strategies can be enhanced by using various ML classifiers on the dataset. An influential practice in ML to increase the accuracy of traditional base classifiers is to build classifier ensembles. The benefit of these techniques over visual appraisal by a medical specialist is that they are completely automatic and thus impartial towards human errors [8], [9]. Several classifiers are there in ML literature that are based on the problem domain which are used to predict and improves the accuracy of classifiers further. Random Forest is an ensemble ML algorithm. It generates a standout amongst the best accuracies and has a significant benefit over different methods in terms of competence to deal with non-linear biological data [10], [11]. In recent years, many ML algorithms have been established to create imputation techniques [12]. Researchers have proposed numerous statistical methods to impute missing data values. As opposed to statistical methods, ML algorithm builds a data model from missing attributes. This model is used to execute classification that imputes the missing attributes. As the majority of model-based imputation algorithms need experimental data without missing attributes in the incomplete dataset and the training set to give an approximation of the missing values, the result of imputation is influenced by the experimental data [13]. In the present study, we used an ensemble technique, namely Random Forest ensemble method, to jointly predict a healthy or AD inflicted patient from longitudinal MRI data, based on two different strategies in order to deal with missing data. The paper is organized as follows: Section 2 discusses the related work and the motivation behind this paper. Section 3 communicates the theoretical background used in this study. Section 4 describes the Materials and Methods. Section 5 illustrates the Experiments and Results. Comparisons are drawn in the paradigms used, and the impact and sensitivity analysis is performed in Section 6 followed by Conclusion and Future Work in Section 7.

2 RELATED WORK

In our recent AD-related analysis, we explored various quantitative clinical rules for listing demented and non-demented patients on the basis of MRI's numerical metrics provided by Washington ADRC. Furthermore, we analyzed different related features that were extremely reliant in defining the group of AD patients. We performed Exploratory Data Analysis (EDA) among different features in a given dataset employing various ML classifiers. We intended to provide quantitative proof for the best precise ML classifier that could help in the prediction of AD in follow-up patients. We showed that ML strategies were sensitive to missing values. In fact, the missing information seems to lessen the insight into the data.

2.1 Motivation of this study

The contribution of this study is twofold. In the present experimental setup, two different strategies were employed to deal with the missing data so as to find the best model for a ML framework for the early detection of AD. The introduced model uses a classifier ensemble structure, viz. Random Forest (RF) to classify given records effectively.

1. For missing value data, we demonstrate the prediction of AD by removing all of the missing attributes from the dataset to produce an accuracy using RF ensemble classifier.

2. In the second method, for imputation of missing values using median, we show that it can be merged with the imputation practice for an incomplete clinical dataset; which turns out to be the best imputation approach for the detection of AD using RF classifier. Our study achieves a higher accuracy for missing value imputation and gives a better choice for huge population diagnosis.

3 THEORETICAL BACKGROUND

3.1 Classifier

With ML algorithms, each dataset instance was denoted with various binary, continuous, or categorical features. As ML learning can be both supervised and unsupervised, the supervised learning illustrates situations in which the machine learns the function, with known labels and desired outputs. Whereas unsupervised learning did not possess labelled outputs. As ML require supervised tasks, so we concentrated on the requisite method in order to achieve this labelling [14].

3.2 Classifier Ensemble with Random Forest

In ML, ensemble are algorithms that combine different single classifiers to create a model which is mostly more accurate than the single classifier alone [15]. Several ensemble classifiers such as bagging and boosting are used in ML literature [16]. Diversity in bagging is provided by creating classifiers employing a randomized heuristic that are independent with each other [17]. Further randomization to bagging yields Random Forest (RF) ensemble model that improves the overall performance by de-correlating the prediction of each tree. RF ensemble is a type of classifier that learns decision trees randomly. Decision Tree is comprised of nodes that act as information related to the respective attributes. The applied information follows a decision route for a specified set of input features, contingent upon one of the nodes viz. categorical node (case of categorical data) or threshold node (case of continuous variable) [18]. Despite being faster, the decision trees are susceptible to being excessively adjusted to the training set of data or to a loss in performance through tree pruning for generalization [19]. In contrast to this, RF improves performance at each node by randomization of features while building trees. RF is used for both classification and regression and has been used in the context of missing data successfully. Each and every tree in the RF grow thereby, forming an autonomous member of the forest. Thus, these properties make the RF algorithm a suitable choice for this missing data study.

3.3 Missing Data

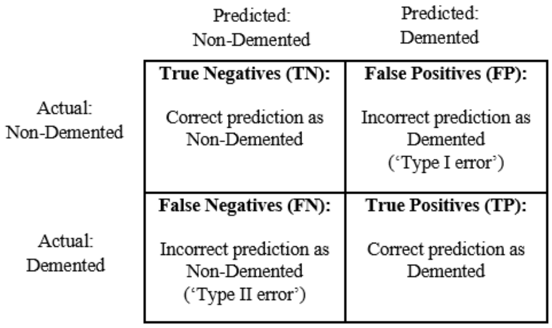
The problem of missing data is a very common issue, particularly when dealing with real-world and large data sets. It is a matter of concern in almost all longitudinal studies which experience the major attrition, thereby producing concerns for the dropouts' characteristics when compared to the rest of the subjects. The impact of missing data on the ML model has demonstrated that the outcome gets degraded by just allocating a random value to the missing data components. There are many reasons for the occurrence of missing data, likewise, inadequately described surveys, partial variable accumulation from subjects, data being deleted confidentiality purposes, or non-response from individuals refusing to give information [20], [21]. Despite the fact that handling missing data is generally not the objective of any substantive study but at the same time failing to do so correctly generates significant issues. In the first place, missing data can present possible prejudice in parameter approximation and deteriorate the generalizability of the outcome [13]. Secondly, discarding instances with missing data leads to the information loss which thereby reduces the statistical ability and augments errors [13]. Above all, the statistical techniques are created for a complete set of data. A data set should be modified and corrected into a complete data set before analyzing this dataset with missing values.

3.4 Handling Missing Data

Various techniques have been formulated for handling missing data. The easiest method removes the occurrences of missing data. This results in ineffective and biased observations to deduce the conclusions. It is not a practical approach if a considerable size of data are missing. Many other procedures exist to deal with the missing set of data, likewise, weighting techniques available-case techniques, and imputation-based techniques [22]. The latter technique is employed and discussed further in this study. The Imputation method includes prediction of those data values which are missing. It encompasses substituting the missing values by appropriate approximations such as mean or median and then applying standard complete-data methods to the filled-in data. The primary purpose behind imputation is to decrease biasness due to missing values which thereby results in higher efficiency of the model [14].

3.5 Metrics for Performance Evaluation

A classifier's performance is determined by testing it with a certain metric for evaluation of its outcome. In our study, to assess the effect of the imputation and non-imputation methods on the RF ensemble model, we applied five extensively used metrics, namely, accuracy, sensitivity, specificity, Receiver Operating Characteristic (ROC) curve, and Area under the ROC curve (AUC). The results with respect to this classifier system are characterized with four states, as true positive (TP), false positive (FP), true negative (TN) and false negative (FN) [23]. The results have been correlated to each other by a table, called a confusion matrix. A confusion matrix (CM) was used to evaluate the above-stated metrics. The terminology associated with the CM had been described in Figure 1 respectively. We interpreted this with our AD problem accordingly.



	Predicted: Non-Demented	Predicted: Demented
Actual: Non-Demented	**True Negatives (TN):** Correct prediction as Non-Demented	**False Positives (FP):** Incorrect prediction as Demented ('Type I error')
Actual: Demented	**False Negatives (FN):** Incorrect prediction as Non-Demented ('Type II error')	**True Positives (TP):** Correct prediction as Demented

Fig. 1 *Confusion Matrix Distribution*

Following are the metrics that are calculated from the confusion matrix:

1. Accuracy (Acc): It is the ability to calculate correct predictions.

$$Acc = \frac{(TP+TN)}{(TP+FP+FN+TN)} \qquad (1)$$

2. Sensitivity/Recall (Sn): It measures how sensitive the model is for identifying positive instances.

$$Sn = \frac{TP}{(TP+FP)} \qquad (2)$$

3. Specificity (Sp): It measures how specific the model is for identifying negative instances.

$$Sp = \frac{TN}{(TN+FP)} \qquad (3)$$

4. ROC Curve: It is useful in visualizing the complete performance of the model. In this, the curve is plotted between True Positive i.e. Recall and False Positive i.e. Specificity.

5. AUC Curve: Its value is calculated from the area below the ROC curve. It indicates the test performance at each threshold point.

4 METHODS AND MATERIALS

4.1 Dataset Description

Dataset used in this study was obtained from the longitudinal pool of OASIS (Open Access Series of Imaging Studies) MRI data in demented and non-demented older adults [1]. Each record comprises of 15 attributes including age, gender, and other MRI measurements.

4.2 Details of MRI data

1.5 Tesla Vision scanner was used for the T1-weighted MR image acquisition.

4.3 Subjects

The data set consists of 150 subjects aged between 60-96 years, for a total of 373 MRI sessions. This longitudinal collection of MRI data includes subjects that have been clinically identified with very mild to moderate AD. The subjects include 62 male and 88 female with Mean ± SD = 77.01 ± 7.64 and Median = 77.0 and are all classified as right-handed. Only the subjects of the first visit are being considered throughout the study, which makes it a total of 150 subjects.

4.4 Exploratory Data Analysis (EDA)

Prior to this work, EDA was carried out using Python programming language to understand the correlation among the various features present in the data set.

4.5 Data Pre-processing

By using the given features, two sets of variables-target/dependent variable and independent variable were classified into one of two sets of patients- those with AD and those without AD i.e. healthy patient. This dataset comprised of many missing values because not all MRI measurements were acquired. The steps of data pre-processing include 6 steps: importing the required libraries, importing the data set, handling the missing values, encoding the categorical data, splitting the data set into train set and test set, and feature scaling. After handling the missing data using two strategies-removal of missing data and missing data imputation, we used the procedure to apply the RF ensemble classifier. Later, we did hyperparameter tuning of the results got so far so as to improve the model performance.

5 COMPARISONS AND RESULTS

The diverse set of ML classifiers and their accuracy usually conflicts with each other. In spite of this, if the base classifiers are precise, diversity comes out to be low among them. Moreover, if there is no diversity amongst the base classifiers, then their combination will not yield an efficient result. Therefore, the optimal output can only be achieved by an ensemble technique comprising of correct and precise classifiers that differ as much as possible. The ensemble method, RF, can avoid this problem efficiently. In this section, we studied the early diagnosis of AD, according to the results of the ML techniques that were explained above. We then compare them to discern which is more accurate in the diagnosis of AD. The diversity is provided by two separate techniques in order to create an ensemble consisting of classifiers that disagree on their predictions.

5.1 Dropping Missing Data (Non-Imputation method)

Firstly, the dataset was trained by removing the missing set of values. Next, it was tested for the remaining set of values. After removal of rows pertaining to 8 missing values, the dataset for training and testing remains equal to 142 subjects.

5.2 Missing Data Imputation

In this, the dataset was trained by performing imputation (by median) on the missing set of values. The data is tested for the 150 subjects, as the missing values are imputed by their median values. The Random Forest ensemble classifier was constructed next and the results on testing the data prediction ability of the classifier along with their best parameters are predicted as indicated in Table I. The results were generated by predicting data of the indicated variables, and calculating the percentage of values accurately predicted within the specified ranges. The test accuracy was calculated by using equation (1) while the error rate is estimated by: (1-Acc).

TABLE 1

RANDOM FOREST ENSEMBLE ACCURACY

Method applied	No of estimators	No of features	Depth of trees	Test Accuracy	Error
Dropping Missing Data	10	4	1	83.33%	16.67%
Missing Data Imputation	6	5	7	86.84%	13.16%

It can be deduced from the results of Table I that the imputation method outperforms the other method. A significant advancement in the accuracy of RF ensemble classifier using imputation technique can be comprehended. An average percentage increase is observed. This improvement from dropping missing data to missing data imputation is noteworthy realization.

6 ASSESSMENT OF OTHER PERFORMANCE EVALUATION METRICS

The impact of estimating the data from both the methods applied were evaluated within this section. This evaluation provided a complete performance representation, as it offered insight into the effects of imputation and non-imputation within the data, and on the effects of imputation and non-imputation on the classifier. While the classification of data consisted of two classes- demented and non-demented, we applied cross-validation (CV) of 5-fold value. CV process divides the data into two parts so as to compare them statistically and evaluate algorithms that are being learnt. One part trains the classifier i.e. training set while the other part tests the classifier i.e. testing set. These sets were then staggered over 5 (K-fold value) rounds successively. A CV with K-fold value provides the repetitive rounds of cross-validation or basis altered in exceptional cases. Therefore, we demonstrate the performance of RF ensemble classifier to diagnose AD using the evaluation metrics described in Section 3.5 above.

6.1 Sensitivity and Specificity

Using equations (2) and (3), sensitivity and specificity were calculated, as illustrated in Table II.

TABLE 2.

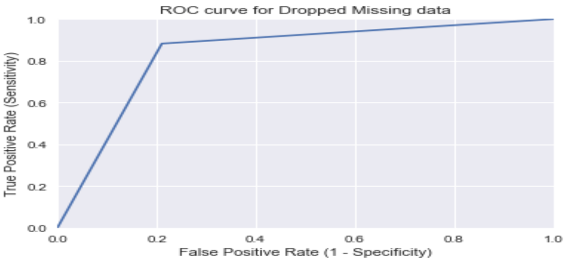
PERFORMANCE EVALUATION METRICS

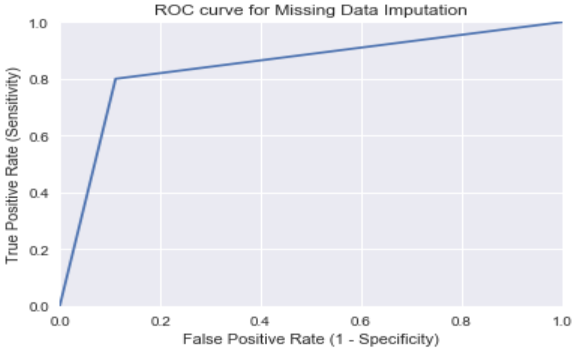
Method applied	Sensitivity/Recall	Specificity	AUC
Dropping Missing Data	88.23%	78.94%	83.59%
Missing Data Imputation	80.00%	88.00%	87.22%

Recall (true positive rate), employing the first method was more as compared to the second method. For actual positive value, the correct prediction was more when the first method was used. While the true negative rate was greater for imputation employed method as can be seen from Table II. We can conclude that the present RF ensemble model for missing data imputation (second method) was highly specific and less sensitive.

6.2 ROC Curve

ROC curve helps in selecting the threshold level that balances Sn and Sp for a better understanding of a given problem domain. The thresholds that are used to produce the curve, cannot be seen on the curve itself. Figure 2 depicts the ROC curve for both the methods applied.





(a)

(b)

Fig. 2 *ROC Curve*

6.3 AUC Curve

It can be seen from Table II that the imputation method leads to a higher percentage of AUC curve as compared to the other method. AUC is beneficial in a way such as a classifier's performance can be obtained in a single number summary form. The higher the AUC value, the better the classifier. It can be inferred from the above results that missing data imputation method proposed in the present study outperforms the dropping missing data method using the RF ensemble classifier.

7 Conclusion And Future Work

Missing data affect a considerable amount of loss of information in the various studies, especially when it pertains to clinical data. In general, no insight can be achieved into the principal cause that is responsible for the missing data. The present study investigated the performance of the imputed data method over the dropped missing data strategy using the RF ensemble model. The impact assessment of both the methods depicted that when the imputation is applied to the dataset, this method provides 87% accuracy when compared to the other method. The proposed technique was competent enough to diagnose the AD patients and classify them as demented and non-demented effectively. Moreover, the present work demonstrated that RF ensemble model not only improves the performance but also increased the accuracy of the classifier. As the results are dataset dependent, the main limitation of this study is the size of the dataset, which is consisting of limited MRI features. Also, we focused on one ensemble technique only. However, we envisage that in future we shall focus on a larger dataset and using other ensemble classifiers in a different way to develop effective decision systems for the detection and prediction of AD.

REFERENCES

[1] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, “Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults,” J. Cogn. Neurosci., vol. 22, no. 12, pp. 2677–2684, 2010.

[2] E. Acuna and C. Rodriguez, “The Treatment of Missing Values and its Effect on Classifier Accuracy,” in Classification, Clustering, and Data Mining Applications, Springer, Berlin, Heidelberg, 2004, pp. 639–647.

[3] C. Patterson, “The state of the art of dementia research: New frontiers World Alzheimer Report 2018,” 2018.

[4] C. R. Jack et al., “Magnetic resonance imaging in Alzheimer ’ s Disease Neuroimaging Initiative 2,” Alzheimer's Dement., vol. 11, no. 7, pp. 740–756, 2015.

[5] E. E. Bron, P. Boudewijn, F. Lelieveldt, M. Smits, and S. Klein, “Fast parallel image registration on CPU and GPU for diagnostic classification of

Alzheimer's disease," Front. Neuroinform., vol. 7, no. January, pp. 1–15, 2014.

[6] F. Falahati, E. Westman, and A. Simmons, "Multivariate Data Analysis and Machine Learning in Alzheimer's Disease with a Focus on Structural Magnetic Resonance Imaging," J. Alzheimer's Dis., vol. 41, pp. 685–708, 2014.

[7] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," NeuroImage (Elsevier), vol. 59, no. 2, pp. 895–907, 2012.

[8] S. M. Stivaros, A. Gledson, G. Nenadic, X. Zeng, J. Keane, and A. Jackson, "Decision support systems for clinical radiological practice — towards the next generation," Br. J. Radiol., vol. 83, pp. 904–914, 2010.

[9] A. Belle, M. A. Kon, and K. Najarian, "Biomedical Informatics for Computer-Aided Decision Support Systems: A Survey," Sci. World J., 2013.

[10] B. H. Menze et al., "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," BMC Bioinformatics, vol. 16, pp. 1–16, 2009.

[11] R. Caruana, "An Empirical Comparison of Supervised Learning Algorithms," in Proceedings of the 23rd International Conference on Machine Learning, 2006.

[12] K. Lakshminarayan, S.A. Harp, T. Samad, Imputation of missing data in industrial databases, Appl. Intell. 11 (1999) 259–275.

[13] Y. Dong and C. J. Peng, "Principled missing data methods for researchers," Springerplus, vol. 222, no. 2, pp. 1–17, 2013.

[14] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," Comput. Methods Programs Biomed., vol. 104, no. 3, pp. 443–451, 2011.

[15] G. Biau, L. Devroye, G. Lugosi. "Consistency of random forests and other averaging classifiers." Journal of Machine Learning Research, to appear, 2008.

[16] R. Polikar, A. Topalis, D. Parikh, D. Green, J. Frymiare, J. Kounios, C.M. Clark, An ensemble based data fusion approach for early diagnosis of Alzheimer's disease, Inf. Fusion 9 (2008) 83–95.

[17] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[18] Y. Qi, J. Klein-Seetharaman, Z. Bar-Joseph. "Random forest similarity for protein-protein interaction prediction from multiple sources," in Pacific Symposium on Biocomputing 10, pp. 531 - 542, 2005.

[19] T. K. Ho. "Random decision forests." ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition, Vol. 1, 1995.

[20] G. Ssali, T. Marwala. "Estimation of missing data using computational intelligence and decision trees." Proceedings of IEEE International Joint Conference On Neural Networks, Hong Kong.

[21] N. J. Horton, K. P. Kleinman. "Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models," in The American Statistician, Vol. 61, No. 1, pp. 79, 2007.

[22] D. J. Fogarty. "Multiple imputation as a missing data approach to reject inference on consumer credit scoring." http://interstat.statjournals.net/Year/2006/articles/0609001.pdf.

[23] M. Pampaka, G. Hutcheson, and J. Williams, "Handling missing data: analysis of a challenging data set using multiple imputation," Int. J. Res. Method Educ., vol. 39, no. 1, pp. 19–37, 2016.