Received: 24 July 2024 | Accepted: 24 August 2024

DOI: 10.1049/htl2.12091

Healthcare Technology Letters

The Institution of Engineering and Technology WILEY

LETTER

Autoencoder imputation of missing heterogeneous data for Alzheimer's disease classification

Namitha Thalekkara Haridas[1] | **Jose M. Sanchez-Bornot**[1] | **Paula L. McClean**[2] | **KongFatt Wong-Lin**[1] | **Alzheimer's Disease Neuroimaging Initiative (ADNI)**

[1]Intelligent Systems Research Centre, School of Computing, Engineering and Intelligent Systems, Ulster University, Magee campus, Derry~Londonderry, Northern Ireland, UK

[2]Personalised Medicine Centre, School of Medicine, Ulster University, Magee campus, Derry~Londonderry, Northern Ireland, UK

Correspondence
KongFatt Wong-Lin, Intelligent Systems Research Centre, School of Computing, Engineering and Intelligent Systems, Ulster University, Magee campus, Derry~Londonderry, Northern Ireland, UK.
Email: k.wong-lin@ulster.ac.uk

Abstract

Missing Alzheimer's disease (AD) data is prevalent and poses significant challenges for AD diagnosis. Previous studies have explored various data imputation approaches on AD data, but the systematic evaluation of deep learning algorithms for imputing heterogeneous and comprehensive AD data is limited. This study investigates the efficacy of denoising autoencoder-based imputation of missing key features of heterogeneous data that comprised tau-PET, MRI, cognitive and functional assessments, genotype, sociodemographic, and medical history. The authors focused on extreme (≥40%) missing at random of key features which depend on AD progression; identified as the history of a mother having AD, APoE ε4 alleles, and clinical dementia rating. Along with features selected using traditional feature selection methods, latent features extracted from the denoising autoencoder are incorporated for subsequent classification. Using random forest classification with 10-fold cross-validation, robust AD predictive performance of imputed datasets (accuracy: 79%–85%; precision: 71%–85%) across missingness levels, and high recall values with 40% missingness are found. Further, the feature-selected dataset using feature selection methods, including autoencoder, demonstrated higher classification score than that of the original complete dataset. These results highlight the effectiveness and robustness of autoencoder in imputing crucial information for reliable AD prediction in AI-based clinical decision support systems.

1 | INTRODUCTION

Alzheimer's disease (AD), the most common cause of dementia, is a progressive brain disorder associated with memory loss, affecting day-to-day activities and cognitive decline [1]. Detecting AD and its severity level at early stage can enable better disease management and reduced care costs [2]. Further, adopting the right measures in clinical diagnosis is essential for timely treatment, care and disease management. Several assessment strategies and markers are currently available, including brain/blood-based biological assessment, medical and family history, and neuropsychological assessments [3]. Due to the variety of assessments and with symptoms overlapping with normal ageing and other types of dementia [4], diagnosis of AD remains challenging.

The implementation of technology-aided decision support system, especially involving machine learning, for diagnosis may offer a promising path [5]. However, when clinical data contains missing values, it significantly impacts early diagnosis and treatment, underscoring the importance of implementing effective measures to uphold data quality and integrity [6]. In particular, missing data, if not handled appropriately, can potentially delay treatment and lead to biased diagnostic results, including in machine learning decisions.

Clinical data may be incomplete in different scenarios. For instance, patients may not arrive for medical appointments or be unable to complete surveys [6]. Incomplete data is highly prevalent in cohort studies, especially within dementia and AD studies, due to factors such as longer study requirements, increased risk of mortality and cognitive decline among older adults. These factors impede their ability to participate in studies requiring multiple visits, leading to missing data [7]. More generally, missing data can be classified into three types: missing at random (MAR) which is dependent on observed variables, such

as when the proportion of missingness increases with dementia severity; missing completely at random (MCAR), in which the missingness is independent of any variables; and missing not at random (MNAR), in which missingness is dependent on unrecorded variables [8]. The trivial solution is to ignore the missing portion, but that may lead to low statistical power for the machine learning models. Therefore, it is imperative to employ appropriate strategies for data imputation [6].

Previous studies have made use of different imputation methods on AD datasets in different contexts [6, 9–15]. However, the data used were not sufficiently comprehensive [11, 12, 16], did not involve brain tau pathology with superior diagnostic utility [13, 14], or that traditional statistical or machine learning imputation methods were used [6, 12, 13]. In the only study that has used deep learning (autoencoders [15]) for AD data imputation [16], the personal information did not include a family history of AD and there was only one cognitive and functional assessment (mini-mental state examination, MMSE); data was not sufficiently detailed and comprehensive. Further, the type of autoencoders used was also unspecified. Importantly, missing data was not systematically evaluated in that study.

To address the limitations of previous studies on missing data imputation for AD classification, in the current study, we focused on the imputation of the missingness of the most important AD data features with respect to predicting different progressive stages of AD. These key data features were common features identified from multiple feature selection methods, including with an autoencoder to elucidate latent features. MAR missing data were then systematically generated over different proportions of missingness from a complete, comprehensive and open dataset, with the latter acting as ground truth. This was followed by the implementation of a denoising autoencoder on the missing key features. Finally, three-class (control normal CN, mild cognitive impairment MCI (a mixed group which includes prodromal stage of AD), and AD) classification of the imputed datasets by the random forest classifier was employed, and the results were compared with those based on the classification of the original complete datasets and the subset of selected features.

2 | METHODS

2.1 | Data description

The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, more specifically the ADNIMERGE-3 open repository, was employed as the primary data source for this study. The data was obtained from the https://adni.loni.usc.edu portal as per request and was accessed after approval by the Data Sharing and Publication Committee of Image and Data Archive (IDA). The dataset comprises clinical, neuropsychological assessments, neuroimaging measures, and genetic markers collected from participants diagnosed as healthy (CN), having mild cognitive impairment (MCI), or Alzheimer's disease (AD). The processed neuroimaging (tau-PET and structural MRI) is available from a previous study [17]. The final processed dataset consists of 559 samples (participants) encompassing a total of 224 features.

The processed dataset comprised 7 sociodemographic and medical history features, 40 cognitive and functional assessment (CFA) scores including their sub-assessments, and a relatively extensive 177 neuroimaging features (from active tau-PET brain regions via co-registering with structural MRI [17]). The class labels for training the classification model were based on clinician diagnosis, comprising 363 CN, 137 MCI and 59 AD cases.

The sociodemographic and medical or family background features comprised age, gender, years of education, maternal and paternal family history of AD, and the number of copies of the APoE ε4 alleles (hereafter referred simply as APoE4). The CFA scores were derived from various measures, including the Alzheimer's Disease Assessment Scale (ADAS), Cognitive Battery Assessment, Clinical Dementia Rating (CDR), Mini-Mental State Exam (MMSE), Modified Hachinski Ischemia Scale, Neuropsychological Battery Test, logical memory immediate recall test (LMIT), logical memory delayed recall test (LMDT), the Neuropsychological Inventory (NPI), and the Geriatric Depression Scale (GDS). The dataset included individual question scores from ADAS and individual subscales from NPI, while other CFAs and individual CFA subscales from ADNI were excluded due to significant missing data. The tau-PET neuroimaging data utilised the [^{18}F]AV-1451 tracer for detecting tau deposition [18].

2.2 | Data preparation and preprocessing

As part of the preprocessing pipeline, the participant identification (ID) column ('RID') was first removed. The gender column ('PTGENDER') was standardised by subtracting 1 from its values to ensure consistency in representation [19].

Certain columns in the dataset contained negative values, which required adjustment to ensure compatibility with subsequent processing steps. The columns with negative values were identified; all of them were CFA columns including COMP_MEM_SCORE, PHC_MEM, PHC_EXF, PHC_LAN, and COMP_EXEC_FUNC_SCORE, and they were transformed by adding the absolute value of their minimum to all values, effectively shifting the distribution to non-negative values.

Afterwards, normalisation [20] of features was performed as a part of preprocessing using min–max scaling method [21]. Features with a maximum value exceeding 1 were identified, excluding the target variable (AD_LABEL) and clinical dementia rating (CDR). These features were then normalised by subtracting their minimum value and dividing by the range between the maximum and minimum values. The preprocessed dataset was then used for feature selection.

2.3 | Feature selection

2.3.1 | Boruta (Method 1)

The Boruta algorithm [22] is a feature selection technique based on random forests, designed to distinguish relevant features by comparing their importance with that of random features. It iteratively evaluates the significance of each feature and selects

those that demonstrate significance above a predefined threshold. This study selected the Boruta algorithm for its ability to handle complex datasets with high-dimensional features. Moreover, Boruta harnesses the collective learning strengths of random forests, ensuring robust feature selection by considering feature interdependencies [22].

2.3.2 | Logistic regression with L1 regularisation (Method 2)

Logistic regression with L1 regularisation [23] is a linear model that penalises the absolute magnitude of the coefficients and produces sparse solutions, with certain coefficients being set to zero. This property makes logistic regression with L1 regularisation suitable for feature selection by identifying and prioritising the most relevant features while disregarding irrelevant ones [24]. In this study, logistic with L1 regularisation regression was employed to complement the Boruta algorithm and provide additional insights into feature importance based on the model coefficients.

2.3.3 | Autoencoder (Method 3)

Autoencoder-based feature selection implements deep learning techniques to learn the compact representations of high-dimensional data. By training an autoencoder model to reconstruct the input features, the encoder layer learns to capture the essential information from the data, effectively performing feature extraction [25]. In this approach, a three-layer autoencoder was utilized, comprising an input layer, an encoding layer, and a decoding layer. The input layer defined the shape of the input data, while the encoding layer, implemented as a dense layer with encoding neurons and ReLU activation, learns to encode the input features into a lower-dimensional representation, capturing the most important features. Subsequently, the decoding layer, is composed of a dense layer with the same number of neurons as the input layer. This autoencoder-based feature selection was employed to unveil the latent features that contribute significantly to the variability in the ADNI dataset. Notably, this approach complements traditional feature selection techniques by capturing nonlinear relationships and revealing hidden patterns in the data.

2.4 | Generating missing data of selected features

MAR is a common issue in AD data, influencing the reliability of analyses and the effectiveness of predictive models [26]. To simulate MAR, we intentionally introduce varying degrees of missingness (40%, 50%, 60% and 70%) individually in the three pivotal variables of our dataset based on the selected key features that were more likely to be missing with later progressive stage of AD [6]. For example, we first identified the maternal dementia history (MOTHDEM) column as a key feature for MAR generation. Then, to introduce a specific level of missingness in the MOTHDEM column, we devised a method that relies on the relationship between MOTHDEM and another pivotal variable, such as CDR, that is objectively and strongly associated with AD diagnosis. Specifically, this method introduces missingness in MOTHDEM based on CDR values. For example, if CDR indicates MCI (CDR = 0.5), we randomly select a fraction of observations and make the corresponding MOTHDEM values missing; higher missingness fraction for AD (CDR = 1). This approach allows us to simulate missingness patterns in MOTHDEM that depend on the progressive stage of AD indicated by CDR, thereby creating a more realistic representation of MAR scenarios [6]. For missing CDR, we tagged missingness levels with AD diagnosis.

2.5 | Imputation using denoising autoencoder

Unlike traditional imputation techniques that depend solely on statistical methods or simple interpolations, denoising autoencoders harness the capabilities of deep learning architectures to learn the intricate patterns and relationships present in the data [27]. The architecture of a denoising autoencoder comprises an encoder, a decoder and a corruption process. The encoder reduces the dimensionality of the input data, mapping it to a latent space representation. The corruption process introduces noise or distortions to the input data to make the autoencoder learn robust features. The decoder then reconstructs the original input from the compressed representation [28].

During training, denoising autoencoders strive to minimise the reconstruction error between the input and its corresponding reconstruction, typically employing mean squared error as the optimisation metric. This iterative process enables the model to adeptly learn meaningful representations of the data by denoising and reconstructing inputs, rendering them invaluable for data-denoising tasks. By training the autoencoder on the observed data while deliberately introducing noise or corruption, the model becomes proficient at reconstructing the original, noise-free data. Consequently, denoising autoencoders excel in effectively filling in missing values through the learned representations [29].

This study, after synthetically generating varying proportion of missingness intentionally in the columns of the selected key features of the dataset, employs a denoising autoencoder for all imputations. The architecture of the denoising autoencoder comprised an input layer with dimensions equivalent to the number of input features in the dataset, followed by two hidden layers with 126 and 63 neurons, respectively. Dropout layers with a dropout rate of 0.2 were strategically incorporated after each hidden layer to mitigate overfitting and enhance the model's robustness. Rectified linear unit (ReLU) activation functions [30] were utilised in the hidden layers to capture nonlinear relationships within the data effectively [31]. Gaussian noise with a standard deviation of 0.1 was introduced to the input layer to augment the denoising capability of the autoencoder, facilitating the accurate reconstruction of missing values while

mitigating noise. The output layer employed linear activation to ensure continuous output values.

The autoencoder was trained using the mean squared error (MSE) loss function and the Adam optimiser [32] over 50 epochs, with a batch size of 50 and a validation split of 20%. Then the autoencoder was able to predict and impute missing values in the respective columns, thus contributing to the restoration of data integrity and enhancing the reliability of subsequent analyses.

Following the imputation process using the denoising autoencoder, the performance of the imputation was evaluated using two key metrics: root mean squared error (RMSE) [33] and imputation accuracy [34]. RMSE was computed to quantify the discrepancy between the actual values of the missing column and the imputed values generated by the autoencoder. This metric measures the average difference between the observed and predicted values, thereby assessing the imputation accuracy on a continuous scale. Additionally, imputation accuracy was calculated to assess the correctness of the imputed values produced by the autoencoder. The imputation accuracy was computed as the ratio of the number of correctly imputed values to the total number of imputed values, providing a percentage value indicative of the accuracy of the imputation process.

Following the imputation process, a visual comparison between the actual and imputed values of the selected columns was conducted using a scatter plot. Additionally, the feature importance of the denoising autoencoder was examined by analysing the weights of the first dense layer. This analysis identified the most influential features for accurate imputation, contributing to a better understanding of the model's performance.

2.6 | Classification and model performance evaluation

In the classification phase of the work, a random forest classifier [35] was employed due to its capability to handle complex data and mitigate overfitting. It also performs especially well in heterogeneous AD data [36]. This ensemble learning technique combines multiple decision trees, providing robustness to noisy data and reducing the risk of overfitting.

We employed three-class classification for classifying the three classes of the target variable AD-LABEL, namely, the CN, MCI and AD groups. Additionally, imbalance observed in the classes was handled by utilising the synthetic minority over-sampling technique (SMOTE) [37] and ensured that the classifier could effectively learn from all classes.

Evaluation of the random forest classifier was performed using 10-fold cross-validation, to reduce sampling bias and enhance generalisability. The classifier was trained on resampled training datasets to mitigate the impact of biased sampling, and its performance was assessed using the following metrics: accuracy, precision, recall and F1-score. Additionally, the confusion matrix and classification report were generated to provide detailed insights into the classifier's predictive performance for each class label.

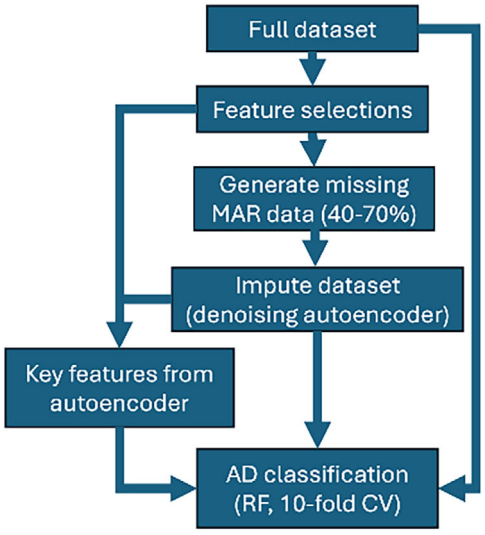

FIGURE 1 Schematic of workflow of current study. CV, cross-validation; MAR, missing at random; RF, random forest classifier.

We conducted classification using random forest classifier for all imputed datasets as well as the original complete dataset. Additionally, we evaluated the performance of classification models using features selected from three different feature selection methods: Boruta, logistic regression, and autoencoder. Furthermore, we included features extracted from the denoising autoencoder as part of our analysis. The overall process of the current study is schematically shown in Figure 1.

2.7 | Software, and data and code availability

Google Colab (https://colab.google/), a cloud-based Jupyter Notebook environment provided by Google, was utilised for conducting the study, which included data preprocessing, feature selection, missing data generation, imputation, classification and validation tasks. Codes are available at https://github.com/NamithaHaridas/Denoising_AE_ADNI.

The original ADNI dataset was not included as part of this repository. Requests to access the original datasets should be directed to ADNI (http://adni.loni.usc.edu/).

3 | RESULTS

3.1 | MAR of consistently extracted features

Feature extractions were performed using Boruta, logistic regression with L1 regularisation and autoencoder with respect to AD severity and the common top-ranked features were extracted. Amongst the data comprising demographic, CFAs, genetic, and neuroimaging markers, we identified maternal dementia history (MOTHDEM) and gender as influential factors, with maternal dementia history serving as a potential genetic predisposition and gender reflecting sex-related differences in AD risk. Further, the key neuroimaging (region-of-interest, ROI) data features identified were the left

hemispheric amygdala, white matter left hemispheric entorhinal and inferiortemporal cortices, consistent with known neurodegenerative changes associated with AD progression [38]. Additionally, the genetic biomarker APoE4 was identified, highlighting its significant role as an AD risk factor, and the clinical dementia rating (CDR), which has been shown to be closely correlated with AD severity [39].

Amongst these extracted features, MOTHDEM, APoE4 and CDR are more likely to be affected by AD severity. For instance, with more severe AD, the patients may not remember their mother had AD. Patients with more severe AD may also be less likely to undergo genetic screening, which requires specialised technical facilities. As CDR takes about 90 min to administer [40], relatively long amongst the considered CFAs, it is more probable that patients with severe AD will not have the patience or ability to undergo such a long assessment. Although it is plausible that patients with overt AD may also be less likely to undergo PET scans, the processed dataset was dominated by PET-MRI's active ROIs and it is in practice unlikely that individual imaging ROI features would be missing. Hence, we did not remove any PET-MRI features in this study. Nevertheless, even without this, our proposed columnar missingness approach could incur substantial reduction in classification accuracy (see below).

For the MOTHDEM and APoE4 variables, missing data of MAR type were generated based on the clinical dementia rating (CDR) values on the samples (rows) with CDR values of 0.5 or 1 (the higher the value, the more severe the condition) at the probability of missingness (0.4, 0.5, 0.6, 0.7). Similarly, for the CDR variable, missing data of MAR type was introduced, but this time it was based on the clinical diagnosis (AD_LABEL), where samples (rows) labelled as MCI or AD were more likely to have missing CDR values at the probability of missingness (0.4, 0.5, 0.6, 0.7). See Section 2.4 for further details. Although we focused on only the missingness of these three variables, which had a strong influence on detecting AD severity (via feature selection), we generated large proportions of missingness within these variables (from 40% to 70% missing data per variable) and classification was substantially poorer (see below).

3.2 | Imputation with denoising autoencoder

Next, denoising autoencoder was employed separately to perform the imputation of missingness introduced at 40%, 50%, 60% and 70% for the pivotal columns MOTHDEM, APoE4 and CDR. Further to visualise the accuracy and reliability of our imputation, we plotted scatter plots comparing the actual and imputed values and showed their similarities (Figure S1). To gain deeper insights, we extracted important features from the dense layer of the autoencoders by calculating the feature importance from each imputed dataset. Table 1 shows the full list of the commonly selected features using the employed traditional feature-selection methods and denoising autoencoders, consistent with previous work [17].

By quantifying the RMSE of the actual versus imputed value differences, Figure 2 shows the RMSE scores and imputation accuracy (ratio of the number of correctly imputed values to the total number of imputed values) at different missingness levels of the three key AD features. This indicated that our methods of implementing missing values and autoencoder-based imputation were conforming to expectation.

TABLE 1 Common features selected from traditional feature selection methods and denoising autoencoder's latent features.

Feature category	Commonly selected features
Sociodemographic	MOTHDEM (history of mother with AD) FATHDEM (history of father with AD) PTGENDER (gender)
Genetic	APoE4
Cognitive and functional assessment	ADAS_Q10SCORE (ADAS sub-assessment score), CDR (clinical dementia rating)
Brain imaging (active tau-PET ROI)	left amygdala, left entorhinal (white matter), left inferiortemporal (white matter), right middletemporal cortex, corpus callosum (central, mid-anterior, mid-posterior, right white matter), right precentral (white matter), right medial orbitofrontal (white matter), right posterior cingulate cortex, brain.stem, right ventral diencephalon, right supramarginal cortex, right transverse temporal (white matter), right cerebellum cortex, left precentral (white matter), right lingual (white matter), left inferior temporal (white matter), right amygdala, left thalamus proper, right lateral orbitofrontal cortex

3.3 | Robust AD classification of CN MCI and AD of imputed data

After the imputation process, random forest classification with 10-fold cross validation was performed on the original dataset, and imputed datasets with 40%, 50%, 60% and 70% missingness in MOTHDEM, CDR and APoE4. We also performed the same classification for a dataset with only the common features selected using the three feature selection methods and included the important features that were performed during each imputation process from the denoising autoencoder in it (see Table S1 for details).

These results revealed the accuracy, precision, recall, and F1 score as 0.85, 0.85, 0.85, and 0.84, respectively, for the original dataset, indicating robust classification performance. These metrics underscore the efficacy of the random forest classifier in accurately predicting AD classes using the original dataset.

Subsequently, the imputed datasets showed consistent classification accuracies, ranging from 0.71 to 0.83 across different missingness levels (Table S1), indicating that the denoising autoencoder effectively imputed missing values without compromising classification performance. Additionally, classifiers trained on feature-selected datasets exhibited higher classification performance (0.89), suggesting that feature-selection techniques retained relevant information for AD prediction, contributing to the classifier's robustness.

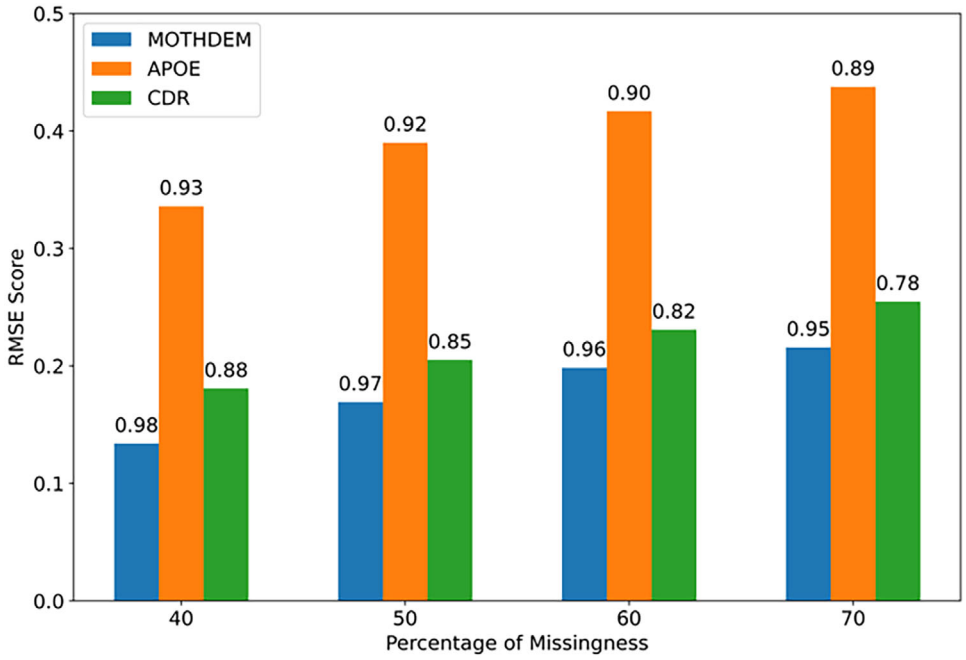

FIGURE 2 Root mean squared error (RMSE) scores and imputation accuracy of the key features maternal AD history (MOTHDEM), APoE4 and clinical dementia rating (CDR) after imputation.

We further evaluated the random forest classification's precision, recall and F1 score of each of the three classes of the target variable AD_LABEL (with classes CN, MCI and AD). Figure 3 summarises the results for each dataset highlighting the F1 scores of the three classes (see Table S2 for details, including precision and recall).

Focusing on the F1 scores of each class (CN, MCI, and AD), we examined the classifier's ability to accurately classify individuals with different disease progressive stages. In the original dataset, the random forest classifier demonstrated strong performance, with F1 scores of 0.91 for CN, 0.69 for MCI, and 0.85 for AD. This indicates robust classification across all progressive stages, which demonstrate the effectiveness of the classifier on the complete dataset (see Table S2 for further details). Noticeably, there was a higher detriment to the F1 scores with missing CDR than MOTHDEM and APoE4.

4 | DISCUSSION

Missing data is a common issue in healthcare datasets, and its presence can significantly impact the performance of predictive models, particularly in the context of AD diagnosis where accurate identification of relevant features is crucial. This study focused on missing at random (MAR) patterns of key AD data features. Previous studies did not consider such systematically selected features for data imputation evaluation. Specifically, we employed a denoising autoencoder, a powerful nonlinear unsupervised learning technique, for imputing missing values in critical AD features: demographics (MOTHDEM), genotype (APoE4), and CFA (CDR).

For the imputed datasets, particularly those with 40% missingness, exhibited robust performance in terms of recall, precision, and F1 score, suggesting that denoising autoencoder effectively imputed missing values without significantly compromising classification performance. In a previous study [16], an accuracy of 78.67% in AD classification of the imputed AD dataset was obtained, but they did not specify the missing percentage of the imputation performed. In comparison, we achieved an overall accuracy of 79%–85% in AD classification for our MAR datasets.

The random forest classifier used here for the three-class classification provided key insights into the performance of different datasets in identifying individuals with AD. The original dataset exhibited a high F1 score (0.85) for the AD class, indicating its ability to effectively identify and accurately predict AD cases. Similarly, the feature-selected dataset demonstrated a higher F1 score (0.89), suggesting that feature-selection techniques retained crucial information for AD prediction. Interestingly, the performances were slightly higher than that for the original dataset, due to a reduction in model complexity with only the relevant features.

Generally, we observed a trend of decreasing F1 scores with increasing levels of missingness in the imputed datasets. In the case of MOTHDEM, the F1 scores for CN and MCI exhibited gradual decrease as the percentage of missingness increased, while the F1 score for AD remained relatively stable. This suggests that the imputation process had a more pronounced effect on the classification of CN and MCI individuals compared to those with AD. Similarly, for APoE and CDR, we observed fluctuations in F1 scores across different classes with increasing levels of missingness. These variations highlight the importance

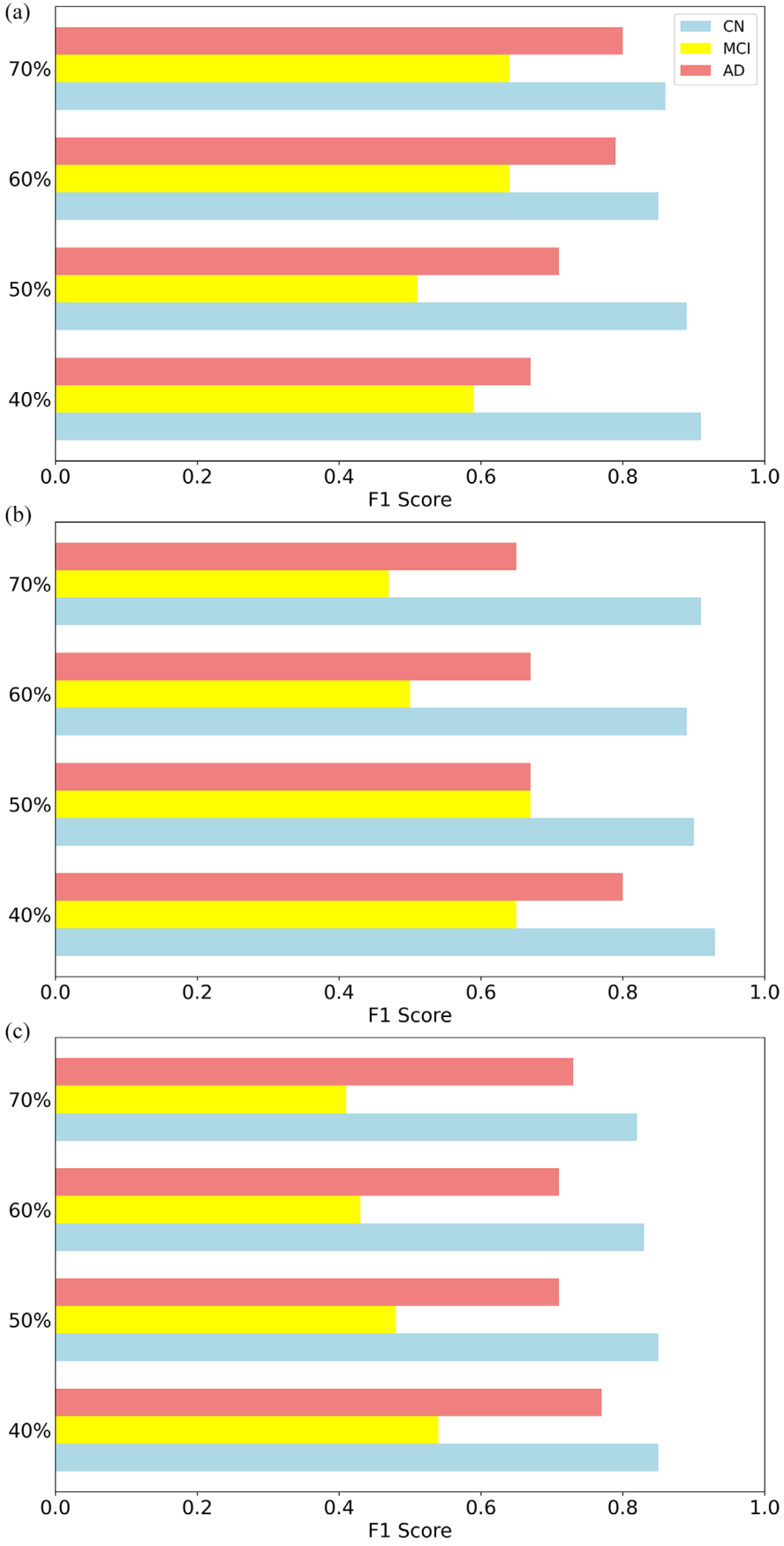

FIGURE 3 F1 scores after three-class classification of the imputed datasets. (a) Maternal AD history (MOTHDEM); (b) APoE4; (c) clinical dementia rating (CDR). AD, Alzheimer's disease; CN, control normal; MCI, mild cognitive impairment.

of carefully considering imputation strategy for heterogeneous and types of missing data.

There are notable limitations in the current study. Firstly, our analysis was confined to a single dataset, and the generalisability of our findings to other datasets may be limited. This should be addressed in future studies. Future studies should also explore the integration of multiple imputation methods to enhance imputation accuracy and mitigate potential biases.

In conclusion, our study highlights the importance of using an autoencoder for robust data imputation for high-performance machine classification across different disease progressive stages, underscoring its potential diagnostic utility in clinical decision-making and disease management.

AUTHOR CONTRIBUTIONS

Namitha Thalekkara Haridas: Conceptualization; data curation; formal analysis; investigation; methodology; software; validation; visualization; writing—original draft; writing—review and editing. **Jose M. Sanchez-Bornot**: Data curation; formal analysis; validation; writing—review and editing. **Paula L. McClean**: Validation; writing—review and editing. **KongFatt Wong-Lin**: Conceptualization; data curation; investigation; methodology; project administration; resources; supervision; validation; writing—original draft; writing—review and editing.

ACKNOWLEDGEMENTS

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; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical 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 (www.fnih.org). 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.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The original data that support the findings of this study are available from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, more specifically the ADNIMERGE-3 open repository at https://adni.loni.usc.edu portal as per request and was accessed after approval by the Data Sharing and Publication Committee of Image and Data Archive (IDA). Processed data are available upon reasonable request. Codes for the current study are available at https://github.com/NamithaHaridas/Denoising_AE_ADNI

ORCID

KongFatt Wong-Lin https://orcid.org/0000-0001-8724-4398

REFERENCES

1. Barnett, R.: Alzheimer's disease. Lancet 393(10181), 1589 (2019)
2. Arvanitakis, Z., Shah, R.C., Bennett, D.A.: Diagnosis and management of dementia. JAMA, J. Am. Med. Assoc. 322(16), 1589–1599 (2019)
3. National Institute for Health and Care Excellence (UK): Dementia: Assessment, Management and Support for People Living with Dementia and Their Carers. National Institute for Health and Care Excellence (UK), London (2018)
4. Weller, J., Budson, A.: Current understanding of Alzheimer's disease diagnosis and treatment. F1000Res. 7, 1161 (2018). https://doi.org/10.12688/f1000research.14506.1
5. Wong-Lin, K., et al.: Shaping a data-driven era in dementia care pathway through computational neurology approaches. BMC Med. 18(1), 1–10 (2020)
6. McCombe, N., et al.: Practical strategies for extreme missing data imputation in dementia diagnosis. IEEE J. Biomed. Health Inf. 26(2), 818–827 (2022)
7. Hardy, S.E., Allore, H., Studenski, S.A.: Missing data: A special challenge in aging research. J. Am. Geriatr. Soc. 57(4), 722–729 (2009)
8. Rubin, D.B.: Inference and missing data. Biometrika 63(3), 581–592 (1976)
9. Stekhoven, D.J., Bühlmann, P.: MissForest: Non-parametric missing value imputation for mixed-type data. Bioinformatics 28(1), 112–118 (2012)
10. Baneshi, M.R., Talei, A.R.: Does the missing data imputation method affect the composition and performance of prognostic models? Iran. Red Crescent Med. J. 14(1), 31–36 (2012)
11. Chandrasekaran, G., Xie, S.X.: Improving regression analysis with imputation in a longitudinal study of Alzheimer's Disease. J. Alzheimers Dis. 99(1), 263–277 (2024)
12. Campos, S., et al.: Evaluating imputation techniques for missing data in ADNI: A patient classification study. In: Pardo, A., Kittler, J., (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 3–10. Springer International Publishing, Cham (2015)
13. Aghili, M., Tabarestani, S., Adjouadi, M.: Addressing the missing data challenge in multi-modal datasets for the diagnosis of Alzheimer's disease. J. Neurosci. Methods 375, 109582 (2022)
14. Makin, S.: The amyloid hypothesis on trial. Nature 559(7715), S4–S7 (2018)
15. Beaulieu-Jones, B.K., Moore, J.H.: Missing data imputation in the electronic health record using deeply learned autoencoders. Pac. Symp. Biocomput. 22, 207–218 (2017)
16. Akramifard, H., Balafar, M.A., Razavi, S.N.: Early detection of Alzheimer's disease based on clinical trials, three-dimensional imaging data, and personal information using autoencoders. J. Med. Signals Sens. 11(2), 120 (2021)
17. McCombe, N., et al.: Alzheimer's disease classification using cluster-based labelling for graph neural network on heterogeneous data. Healthc. Technol. Lett. 9(6), 102–109 (2022)
18. Ossenkoppele, R., et al.: Accuracy of tau positron emission tomography as a prognostic marker in preclinical and prodromal Alzheimer disease: A head-to-head comparison against amyloid positron emission tomography and magnetic resonance imaging. JAMA Neurol. 78(8), 961–971 (2021)

19. Alasadi, S.A., Bhaya, W.S.: Review of data preprocessing techniques in data mining. J. Eng. Appl. Sci. 12(16), 4102–4107 (2017)
20. Manimekalai, K., Kavitha, A.: Missing value imputation and normalization techniques in myocardial infarction. ICTACT J. Soft Comput. 8(3), 8 (2018)
21. de Schutter, B., van den Boom, T.J.: Model predictive control for max-min-plus-scaling systems – efficient implementation. In: Proceedings of the 6th International Workshop on Discrete Event Systems, pp. 343–348. IEEE, Piscataway, NJ (2002)
22. Kursa, M.B., Jankowski, A., Rudnicki, W.R.: Boruta–A system for feature selection. Fundam. Inform. 101(4), 271–285 (2010)
23. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22 (2010)
24. Tay, J.K., Narasimhan, B., Hastie, T.: Elastic net regularization paths for all generalized linear models. J. Stat. Softw. 106, 1–31 (2023)
25. Xu, X., et al.: Autoencoder based feature selection method for classification of anticancer drug response. Front. Genet. 10, 233 (2019). https://doi.org/10.3389/fgene.2019.00233
26. Seaman, S., Galati, J., Jackson, D., Carlin, J.: What is meant by 'missing at random'? Stat. Sci. 28(2), 257–268 (2013)
27. Vincent, P., et al.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103. Association for Computing Machinery, New York, NY (2008)
28. Bajaj, K., Singh, D.K., Ansari, M.A.: Autoencoders based deep learner for image denoising. Procedia Comput. Sci. 171, 1535–1541 (2020)
29. Costa, A.F., et al.: Missing data imputation via denoising autoencoders: The untold story. In: Proceedings of the 17th International Symposium on Advances in Intelligent Data Analysis XVII, IDA 2018, pp. 87–98. Springer International Publishing, Cham (2018)
30. Agarap, A.F.: Deep learning using rectified linear units (ReLU). arXiv:1803.08375 (2018)
31. Baldi, P.: Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp. 37–49. Microtome Publishing, Brookline, MA (2012)
32. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
33. Wang, W., Lu, Y.: Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. IOP Conf. Ser.: Mater. Sci. Eng. 324, 012049 (2018)
34. Ramnarine, S., et al.: When does choice of accuracy measure alter imputation accuracy assessments? PLoS One 10(10), e0137601 (2015)
35. Yang, F., et al.: Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC Bioinf. 10, 1–14 (2009)
36. Bucholc, M., et al.: A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. Expert Syst. Appl. 130, 157–171 (2019)
37. Chawla, N.V., et al.: SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357 (2002)
38. Braak, H., Braak, E.: Neuropathological staging of Alzheimer-related changes. Acta Neuropathol. 82(4), 239–259 (1991)
39. Ding, X., et al.: A hybrid computational approach for efficient Alzheimer's disease classification based on heterogeneous data. Sci. Rep. 8(1), 9774 (2018)
40. Tan, J.E., Strauss, E., Sherman, E.M.S.: Clinical Dementia Rating. In: Kreutzer, J.S., DeLuca, J., Caplan, B., (eds.) Encyclopedia of Clinical Neuropsychology, pp. 1–2. Springer, New York, NY (2011)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Haridas, N.T., Sanchez-Bornot, J.M., McClean, P.L., Wong-Lin, K.: Autoencoder imputation of missing heterogeneous data for Alzheimer's disease classification. Healthc. Technol. Lett. 11, 452–460 (2024). https://doi.org/10.1049/htl2.12091