Applying Topological Data Analysis to Alzheimer's Disease Diagnosis from MRI

by
Hugo Jal,
Ravi Shah,
Parth Parik,
Oliyide Basit

Introduction

As the fifth-leading cause of death in adults over the age of 65 in the United States, Alzheimer's Disease (AD) represents one of the most severe and detrimental illnesses affecting cognitive and mental health (National Center for Chronic Disease Prevention and Health Promotion, C. D. C., 2020). Characterized by progressive cognitive decline, memory loss, and personality changes, AD is primarily driven by the abnormal accumulation of beta-amyloid plaques and tau protein tangles, alongside neuroinflammation and oxidative stress, leading to synaptic dysfunction and brain atrophy (Breijyeh & Karaman, 2020). With projections estimating an increase in the number of individuals affected by AD from 5.8 million Americans to 14 million by 2060, it is imperative to invest in advanced imaging and organizational systems to accurately classify the brain's morphological changes and facilitate early mitigative action (National Center for Chronic Disease Prevention and Health Promotion, C. D. C., 2020).

The relationship between cognitive impairment and the pathological lesions observed in AD is complex, often overlapping with normal aging processes (Wyss-Coray, 2016). Structural markers such as atrophy of medial temporal and hippocampal regions, identifiable via magnetic resonance imaging (MRI), are crucial in the progression of AD (Rao, 2022). Groundbreaking research in recognizing age-related degenerative diseases has laid the foundations for continually advanced modes of identification and qualification to improve patient outcomes and therapeutic interventions. Despite existing methods to identify and categorize neurodegenerative age-related diseases, high-dimensional MRI analysis combined with topological data analysis (TDA) remains an underexplored area in neuroscientific research (Singh et al., 2023; Chazal, et al., 2021).

TDA is a mathematical framework that extracts geometric and topological features from complex datasets, allowing researchers to discern granular changes in brain morphology associated with AD progression. This approach captures both global and local features simultaneously, complementing traditional methods like Convolutional Neural Networks (CNNs). TDA holds promise for earlier and more accurate diagnosis, potentially slowing disease progression and improving patient outcomes through timely interventions tailored to individual needs.

To investigate the potential of TDA in enhancing the categorization of high-dimensional MRIs in AD patients, we have developed several deep-learning algorithms, including a Random Forest classifier, a Long Short-Term Memory network, a linear TDA model, and a stochastic TDA model. This research paper will examine our findings and explore the role TDA can play in advancing Alzheimer's Disease MRI categorization.

Background/Previous Research

Since the introduction of deep learning algorithms in medicine in the early 2000s, such algorithms have grown in use in Alzheimer's Disease research in a variety of facets, including prognosis, predictive analysis, and identification of biomarkers. Deep learning algorithms, including artificial neural networks, convolutional neural networks, and recurrent neural networks, are exceptional at extracting information from high-dimensional data such as MRIs, thus making such tools crucial for the development of research in Alzheimer's Disease.

Methods

In this study, we conducted a comprehensive comparative analysis of five distinct deep learning models to gain deeper insights into their respective performances. The built models encompassed a Random Forest Classifier, a Long Short-Term Memory (LSTM) network, a Linear Topological Data Analysis (TDA) model implemented with Gudhi's Python library, and a Stochastic TDA model incorporating persistence image.

The dataset utilized for our investigation was sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI), specifically the ADNI1: Complete 1Yr 1.5T dataset. The ADNI dataset is renowned for its comprehensive and longitudinal neuroimaging data, making it an ideal choice for studying Alzheimer's disease progression. Said dataset contains subjects who have screening, 6 and 12-month scans.

Linear Models

The linear model utilizes Lasso regression, a form of linear regression that is regularized with an L1 penalty, for predictive analysis.

$$\text{w-hat} = \text{argmin}_w \, \text{MSE}(W) + ||w||_1$$

Lasso regression happens to be particularly useful in machine learning so as to handle high dimensional data since it allows for a facilitated automatic feature selection.

Specifically, the model undergoes training and evaluation with two separate datasets: one created from persistence diagrams converted into persistence images, and the other containing original input data.

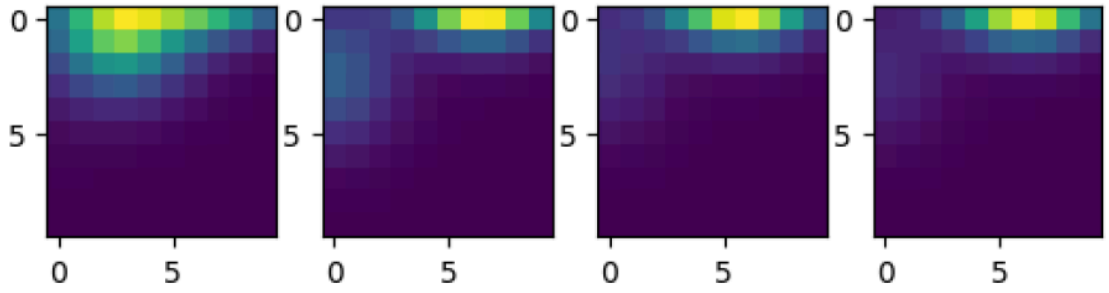

Figure 1: Persistence images for the first four samples

At first, the dataset is separated into training and testing sets with a 70-30 split ratio and a specified random state to guarantee reproducibility. Afterward, the training data is used to instantiate and fit the Lasso regression model. Lasso regression penalizes the absolute magnitude of regression coefficients, encouraging sparsity in the model and potentially enhancing interpretability by choosing important features. Moreover, the model undergoes training with defined hyperparameters such as the regularization parameter (λ) and the maximum iteration limit for optimization.

Ultimately, the model is assessed for performance using the mean squared error and the coefficient of determination. The results yielded a 93.2% accuracy.

```
MSE train : 0.173, test : 0.194
R2 train : 0.920, test : 0.932
```

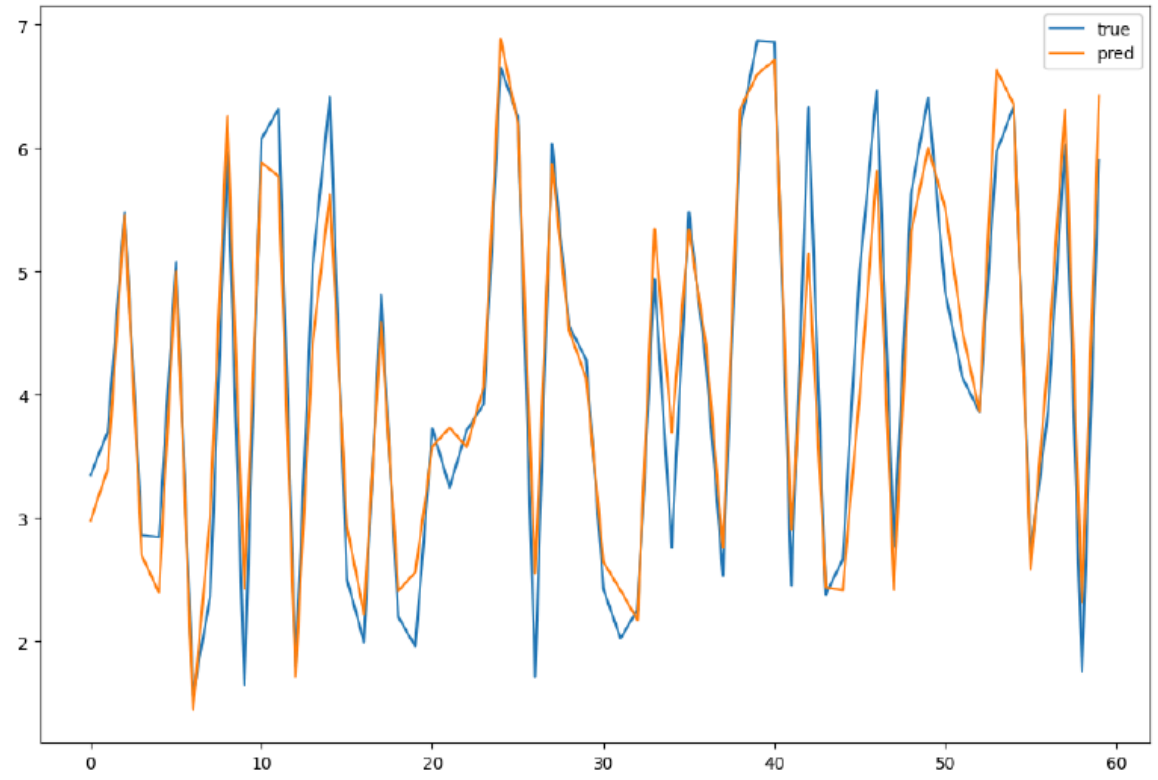

Figure 2: Graph showcasing the Lasso Regression model with persistence homology features inputted.

In contrast, the model trained without persistence homology features performed significantly worse; yielding a 69% accuracy rate.

Linear TDA with Gudhi Library

Linear Topological Data Analysis (TDA) has emerged as a powerful tool in the field of medical imaging analysis, particularly in the context of Alzheimer's disease diagnosis. In this study, we implemented a model using the Gudhi library to conduct linear TDA on MRI images for Alzheimer's diagnosis. This methodology was chosen for its ability to extract topological features from complex data structures, such as MRI images, and provide valuable insights into the underlying geometric and topological properties of the data.

By leveraging the Gudhi library, we built simplicial complexes from the MRI data, allowing us to capture the intrinsic geometric and topological information embedded in the images ("Identification of Onset and Progression of Alzheimer's Disease Using Topological Data Analysis," 2024, 196). The latter is a crucial step for identifying key topological features, such as persistent homology, which are indicative of structural changes associated with Alzheimer's disease progression.

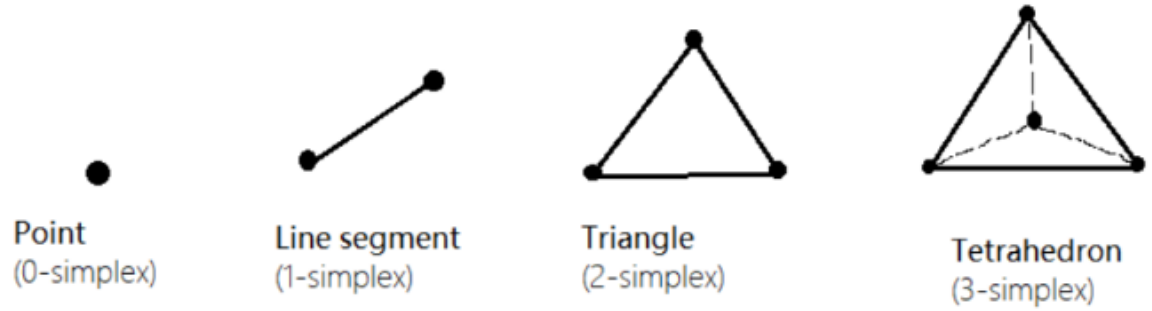

Figure 3: illustration of the first few types of simplicial complexes.

Regarding the data split, we again opted for a 70-30 split and applied a Lasso Regression model to the MRI images preprocessed with persistence images. The accuracy score yielded was 83.9%.

Random Forest Classifier

Random forest classifiers are a technique for classification and regression that involve systematically choosing subsets of features from the feature vector to construct trees in random subspaces: $\{h(x, \theta_k), k = 1,\ldots\}$ where the $\{\theta_K\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x (Ho, 1998)(Rammal et al., 2022). Given that these types of classifiers consist of a collection of decision trees that are trained with different data subsets and then averaged, they are capable of being tolerant of the problem of overfitting (Velazquez et al., 2021).

The utilized pipeline incorporates persistence diagrams to address the topological complexity of analyzing biomedical data, like MRI scans. Every data point in the dataset is processed to generate persistence diagrams; the birth and death values in the diagrams are adjusted to a specific range for data standardization. Afterward, the persistence diagrams are split into two categories depending on their class labels, then the persistence diagrams within each group are converted into persistence images, which act as characteristic attributes of the data.

Persistence images are used as the primary data for classification purposes. The data is split into training and testing sets in the following manner: 70-30. The random forest is set with predetermined parameters such as a maximum depth of 2 and a constant random state of 0. The confusion matrix was implemented, and the accuracy was comprised as 100%.

```
Accuracy: 1.0
Confusion matrix
 [[30  0]
 [ 0 30]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        30
           1       1.00      1.00      1.00        30

    accuracy                           1.00        60
   macro avg       1.00      1.00      1.00        60
weighted avg       1.00      1.00      1.00        60
```

Figure 4: Random Forest Classifier Results

LSTM

The implementation of an LSTM (Long Short-Term Memory) model represents a powerful approach for Alzheimer's diagnosis from MRI images. LSTM, a type of recurrent neural network (RNN), is particularly well-suited for processing sequential data, such as the temporal information present in MRI scans.

The model we implemented is defined with a single hidden layer of 128 units. The input shape is set to match the length of the input sequence and a single feature dimension. The model is then compiled with the mean squared error loss function and the Adam optimizer, which is a popular choice for its adaptive learning rate and momentum-based updates.

During the training process, the model is fit to the training data using a batch size of 20 and a maximum of 100 epochs. To prevent overfitting, an EarlyStopping callback is used, which monitors the validation loss and stops the training if the validation loss does not improve for 50 consecutive epochs.

The results show that the LSTM model achieves excellent performance, with low MSE and high R2 score: 99.6%.

Stochastic TDA with Persistence Image

We applied Stochastic Topological Data Analysis with the aim of seeing how topological features in the MRI scans would behave across visits and whether TDA can discriminate between mild cognitive impairment and Alzheimer's.

Principal Components Analysis (PCA) is applied to TDA so as to find a lower-dimensional subspace onto which to project the data and minimize information loss (Abdesselam, 2021). This is done via an unsupervised approach where a set of linearly uncorrelated variables (eigenvectors) and their corresponding values (eigenvalues) are computed by using an orthogonal transformation (Singular Value Decomposition). In our Stochastic TDA with Persistence Image, we applied PCA with 3 components.

Moreover, we applied t-SNE in order to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. We opted for a perplexity value of 20 and 2 components. Following t-SNE, Uniform Manifold Approximation and Projection (UMAP) was used. UMAP is another powerful dimensionality reduction technique that projects the data into two dimensions, thus preserving more of the global structure compared to t-SNE.

At this point, KepplerMapper reveals the shape and connectivity of the data just before a Random Forest is trained on the persistence images. As a result, we obtained a 94.85% accuracy score.

Results & Discussion

Results

The Random Forest classifier model yielded an accuracy of 100% when evaluated on the test set. This model utilized persistence diagrams converted into persistence images to assist in the classification. The high accuracy indicates that the model effectively captures the complex topological features present in the MRI data. However, an 100% accuracy is not common and could be a sign of overfitting or a bad test set quality. Through our calculations, we determined that our model most likely did not overfit and something else may have occurred.

The Lasso regression model trained on persistence images achieved an accuracy of 93.2%, whereas the model trained without persistence homology features only reached 69% accuracy. This difference shows the value of incorporating topological features derived from TDA and persistence homology in enhancing our specific model performance.

The Linear TDA model, implemented using the Gudhi library, achieved an accuracy of 83.9%.

The LSTM model exhibited an excellent performance with an accuracy of 99.6%. We used the temporal information present in the MRI scans and the LSTM's job of handling sequential data to accurately diagnose AD.

The Stochastic TDA model yielded an accuracy of 94.85%. The use of this model allowed for the effective visualization and analysis of the data.

It is clear that Stochastic TDA performed the greatest out of the other topological models. The LSTM model performed the greatest without considering other problems. The Random Forest classifier achieved perfect accuracy, however it is unclear if any problems occurred.

Discussion

Persistence images proved to be highly effective in improving the accuracy of different models. This shows that it is important to transform topological data into a format more compatible with different algorithms.

Moreover, the LSTM model's high accuracy also shows the value of using temporal information in MRI data. AD is a multifaceted disease, and LSTM's ability to use sequential data makes it suited for the diagnosis of Alzheimers.

The application of PCA and t-SNE in the Stochastic TDA model aided in visualization and analysis of high-dimensional data. In this way, they reduced the complexity of the data yet preserved essential topological features for more intuitive interpretations and insights.

Additionally, comparative analysis indicates that though TDA increases the benefit for all models, the choice of the model can be tailored according to specific research needs and characteristics of data. For instance, Random Forest and LSTM models would be preferred for their high accuracy, while the linear models with TDA offer valuable insights into the importance of features and model interpretability. Combining TDA with other data types, such as genetic or clinical data, may lead to a better accuracy and a greater understanding of AD. The development of topological diagnostic tools combined with other forms of machine learning could help in the early detection of AD progression. The exploration of further techniques in topological data analysis and the use with other machine learning alogrithms can unlock new possibilities in the diagnosis of AD and potentially other diseases.

References

Abdesselam, R. (2021). *A Topological Approach of Principal Component Analysis*. Hal Science. Retrieved February 27, 2024, from https://hal.science/hal-03205861/

Breijyeh, Z., & Karaman, R. (2020). *Comprehensive Review on Alzheimer's Disease: Causes and Treatment*. MDPI. Retrieved May 5, 2024, from https://doi.org/10.3390/molecules25245789

Chazal, F., & Michel, B. (2021, July 16). *An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists*. Frontiers. https://www.frontiersin.org/articles/10.3389/frai.2021.667963/full

Garcia, A. C. (2019, April 17). *STUDY OF BRAIN IMAGING CORRELATES OF MILD COGNITIVE IMPAIRMENT (MCI) AND ALZHEIMER S DISEASE (AD) WITH MACHINE LEARNING. A Maste*. UPCommons. Retrieved February 27, 2024, from https://upcommons.upc.edu/bitstream/handle/2117/133055/anna_canal_garcia_UPC.pdf?sequence=1&isAllowed=y

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832-844. 10.1109/34.709601

Identification of Onset and Progression of Alzheimer's Disease Using Topological Data Analysis. (2024). In S. Devismes, P. S. Mandal, V. V. Saradhi, B. Prasad, A. R. Molla, & G. Sharma (Eds.), *Distributed Computing and Intelligent Technology: 20th International Conference, ICDCIT 2024, Bhubaneswar, India, January 17–20, 2024, Proceedings* (pp. 193-205). Springer Nature Switzerland.

National Center for Chronic Disease Prevention and Health Promotion, C. D. C. (2020, October 26). *What is Alzheimer's Disease? | CDC*. Centers for Disease Control and

Prevention. Retrieved February 26, 2024, from https://www.cdc.gov/aging/aginginfo/alzheimers.htm

Rammal, A., Assaf, R., Goupil, A., Kacim, M., & Vrabie, V. (2022). Machine learning techniques on homological persistence features for prostate cancer diagnosis. *BMC Bioinformatics*, *23*(476). 10.1186/s12859-022-04992-5

Singh, Y., Farrelly, C. M., Hathaway, Q. A., Leiner, T., Jagtap, J., Carlsson, G. E., & Erickson, B. J. (2023, April 1). *Topological data analysis in medical imaging: current state of the art - Insights into Imaging*. Insights into Imaging. Retrieved February 27, 2024, from https://insightsimaging.springeropen.com/articles/10.1186/s13244-023-01413-w

Velazquez, M., Lee, Y., & Alzheimer's Disease Neuroimaging Initiative. (2021). Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects. *PLoS One*, *16*(4). https://doi.org/10.1371/journal.pone.0244773

Wyss-Coray, T. (2016). *Ageing, neurodegeneration and brain rejuvenation*. PubMed. Retrieved May 6, 2024, from https://pubmed.ncbi.nlm.nih.gov/27830812/