Automated Feature Selection of Brain Anatomy Data in Alzheimer's Patients Using Genetic Algorithms and Hidden Markov Models

Jan Lucca Patzelt

July 30, 2023

Abstract

Alzheimer's disease is a neurodegenerative disorder with public health implications. Early detection is vital for effective management and care. This study explores, using genetic algorithms and hidden Markov models to enhance feature selection in modeling disease progression. Brain anatomy data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database is used. The genetic algorithm identifies optimal features to develop a more accurate and robust model. Results highlight opportunities for parameter optimization, improved efficiency, and validation of selected brain regions. This research sets the stage for advanced Alzheimer's studies using genetic algorithms, offering the potential for better understanding and clinical outcomes.

Introduction

Alzheimer's disease is a progressive neurodegenerative disorder characterized by a decline in cognitive functions, most notably memory, and eventually impairs the ability to carry out daily activities [1]. It is the most common form of dementia and constitutes a public health challenge as the global population ages [1]. Early detection of Alzheimer's disease is critical. It can lead to better management of the symptoms, may slow the progression of the disease, and allows patients to make informed decisions regarding their care [8]. However, predicting the progression of Alzheimer's is challenging due to the heterogeneity and complexity of the disease.

To model the progression of Alzheimer's disease over time, one approach is to make use of Hidden Markov Models (HMMs). HMMs are statistical models representing evolving systems with hidden or unobservable states [9]. Recent work has shown that HMMs can generate good models for Alzheimer's progression [5]. Specifically, the research by Hollenbenders et al. (2023) used 14 features, including various biomarkers and cognitive scores, to model the progression of Alzheimer's disease using HMMs.

In this paper, we seek to build upon the foundation laid by Hollenbenders et al. by exploring the possibility of incorporating additional features to enhance the model's predictive capability. For this purpose, we employ a genetic algorithm, a search heuristic inspired by natural selection, to identify a potentially optimal set of features [3]. Genetic algorithms have proven effective in solving optimization problems by evolving a population of candidate solutions over multiple generations. For a genetic algorithm to work a fitness function is needed. Therefore, such a function to evaluate HHMs for their fitness is needed [6] and will be developed.

The overarching goal of our study is to develop a more robust and accurate model for the progression of Alzheimer's disease, which can contribute to better clinical outcomes for patients through early detection and timely interventions.

Methods

Data

This study uses the Alzheimer's Disease Neuroimaging Initiative (ADNI) database from 2018 [7]. The data was provided by proxy via the "Center for Machine Learning" (ZML) at Heilbronn University. ADNI maintains various types of data that produce observable outputs over the progression of the disease. These include Magnetic Resonance Imaging (MRI) [10]. Here the imaging data was used to calculate a normalized value for a given brain region's volume or surface area. The data was further filtered to exclude areas like blood vessels, brain fluids, and similar features. The symmetry of the two brain hemispheres was preserved in this selection. After preparing the data in this manner it was used in the genetic representation for the genetic algorithm.

Genetic Representation

In our genetic representation model, a two-column approach is utilized (see Table 1). This approach integrates data from both hemispheres of a brain region into a single entity, while retaining the unique original values from each hemisphere. Each brain region is transformed into a distinct feature within this representation. Prior to their use in the Hidden Markov Model (HHM), these genetic features are

decoded back into their original format, ensuring compatibility with the model's inputs. A notable element of this representation is the addition of a Boolean column for every brain region. This column operates as a binary indicator, specifying whether a given brain region is active within the DNA used by the genetic algorithm. "Active" denotes presence, "inactive" signifies absence. This method enables us to achieve an efficient and precise data representation. It facilitates the genetic algorithm's analysis process and yields critical information regarding the activity status of each brain region within the algorithm's DNA.

Table 1: Example of Genetic Representation

Features	**Selected**
Pallidum	False
Paracentral	True
ParsOpercularis	False
ParsOrbitalis	True
...	
InferiorParietal	False
IsthmusCingulate	False
LateralOccipital	False
LateralOrbitofrontal	True

This approach allows for a dynamic selection of features by switching them on and off. Further, this approach allows for easy logging of the genetic representation for later analysis of the selected features. Lastly, this approach also allows for an easy mutation of the genetic representation during the mutation phase of the genetic algorithm.

Genetic Algorithm

The genetic algorithm uses a set of starting DNA to generate the first generation. Here two approaches were chosen one completely random set of starting genetic representations and one that used the features from the previous research of Hollenbenders et al. [5].

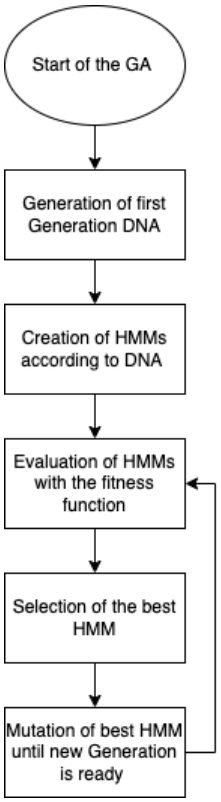

Figure 1: Structure of the Genetic Algorithm

For a starting point, a generational size of 30 individuals was chosen. These individuals were then evaluated for their fitness. This evaluation was done using a fitness function on the generated HHMs Then the fittest individual was selected. The fittest individual was then mutated until a new generation of 30 new individuals was created. For this a mutation rate of 10 % was chosen as a starting point

[4]. The genetic representation of the fittest individual was logged and saved for later evaluation.

To summarize, the genetic algorithm creates multiple HMMs using the generated genetic representations and evaluates them for their fitness. Then the genetic algorithm selects the fittest individual of a generation and mutates this individual until there is a new generation to evaluate.

Fitness Function

The fitness function operationalizes the state transition probabilities found within Hidden Markov Models (HMMs). In this model, the HMM is conceptualized as a directed graph. The individual states within this graph represent the sequential stages of Alzheimer's disease, whereas the edges symbolize the transition probabilities between these states.

The steps of this function are as follows:

1. Transitions between states are quantified using a numerical assignment. Specifically, a transition from a less severe stage (lower state) to a more severe stage (higher state) is attributed a positive value. Conversely, a transition from a more severe to a less severe stage is accorded a negative value.

2. The fitness function aggregates these transition values and divides the sum by the total number of states in the HMM. The resultant quotient, serves as the fitness score for the model.

3. Self-transitions (state-to-self) are excluded from the fitness calculations, implying that a model's stagnation at the same disease stage does not influence its fitness score.

4. This model is premised on the assumption that a higher fitness score should be associated with models demonstrating progression to higher disease stages and penalized for regression to lower stages.

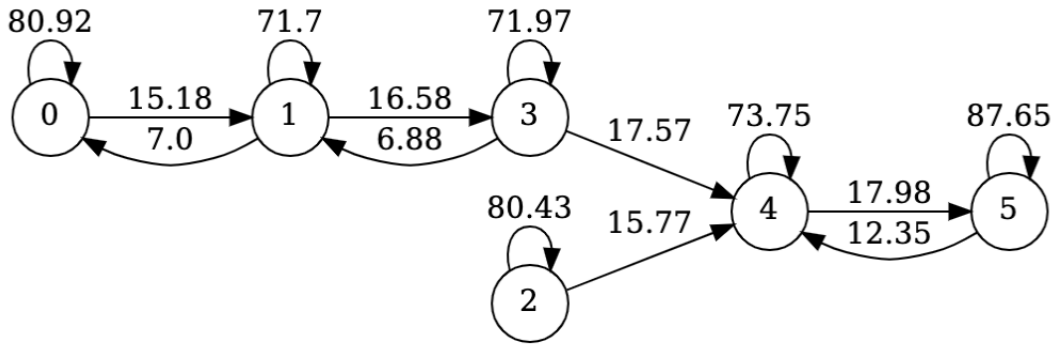

Figure 2: Diagram of a HMM for Fitness evaluation. The circles form zero to five represent the internal states of the HMM. The arrows represent the direction of the transition and the transition probapility.

The fitness score in Figure 2 can be calculated as follows:

$$\frac{(15.18 + 16.58 + 17.57 + 15.77 + 17.98) - (7.0 + 6.88 + 12.35)}{6} = 9.47$$

This calculation can be generalized by using the following formula:

$$F = \frac{\left(\sum_{i=1}^{n} P_i\right) - \left(\sum_{j=1}^{m} N_j\right)}{nS}$$

Here F is the fitness value of the function. The parameter P_i is the sum of all positive transitions. The parameter N_j is the sum of all negative transitions. The parameter nS is the number of states in the model.

In summary, this fitness function utilizes the state transition probabilities in the

HMM to evaluate the efficacy of the model. The fitness score named Normalized Transition Score reflects the model's capacity for progression towards more severe disease stages and penalizes regression to less severe stages.

Results

This section presents the evaluation results conducted using the genetic algorithm. Two endurance runs were performed, each lasting 168 hours (7 days). The runtime of the endurance runs was limited by the performance of the chosen library for generating the HMMs [2]. The problem here is that this library can not generate the models in parallel which limits the speed of the algorithm. The purpose of these runs was to assess the algorithm's performance under different starting conditions and to analyze the impact of feature selection on the results.

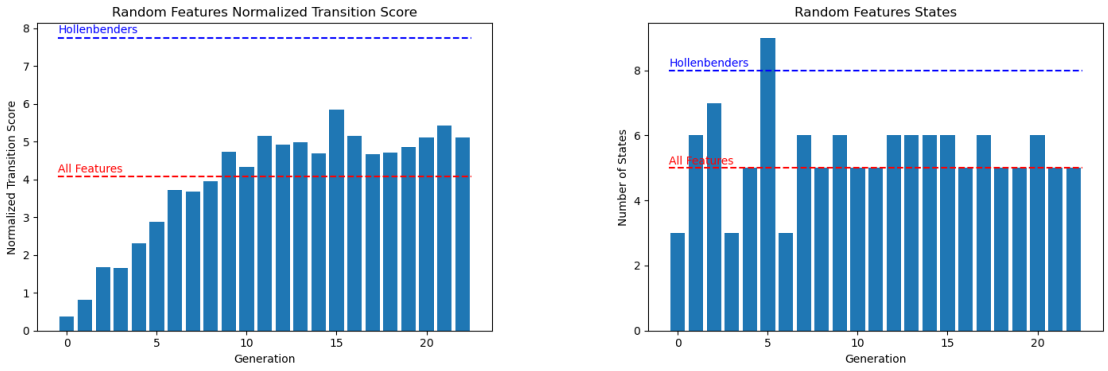

Figure 3: Side-by-Side Comparison of Fitness and State Values for Randomly Selected Features

One endurance run was initiated with pre-selected features[5], while the other run began with a completely random selection of features. By comparing the outcomes of these two runs, we aimed to understand the influence of feature initial-

ization on the performance of the genetic algorithm. Throughout the endurance runs, the genetic algorithm evolved models over multiple generations. For each generation, the best-performing model was identified and considered for further analysis. This allowed us to track the progress and improvements made by the algorithm over time. To assess the fitness of the models, we calculated the fitness value for the model proposed by Hollenbenders[5]. This served as a benchmark to evaluate the performance of the genetic algorithm against an existing state-of-the-art model. Additionally, we calculated the fitness value when all features were present in the model. This provided insights into the potential benefits of utilizing the entire feature set. The oscillation and eventual plateauing of the fitness value,

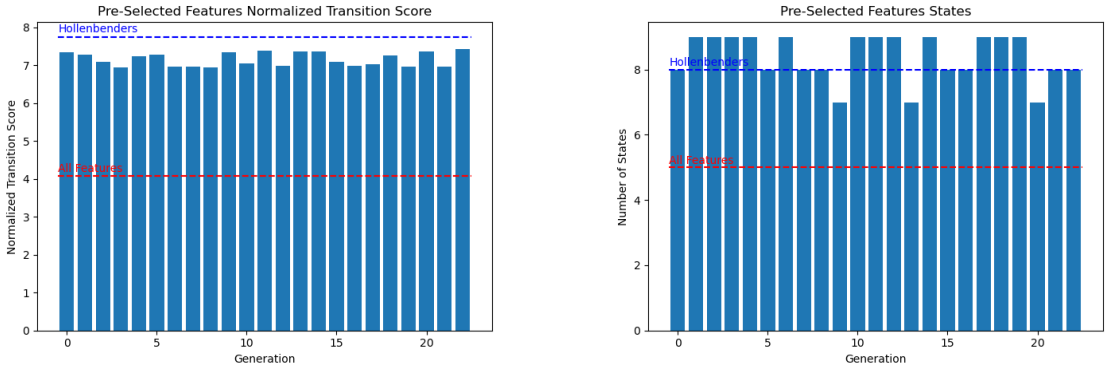

Figure 4: Side-by-Side Comparison of Fitness and State Values for Pre-selected Features

as displayed in Figure 4, were noted during the observations. The fitness value in Figure 3, on the other hand, exhibited a stark rise before moderately surpassing and plateauing beyond the all features benchmark, as was also observed. Further, the distribution of the features for each generation during the endurance run was analyzed in the following heatmaps.

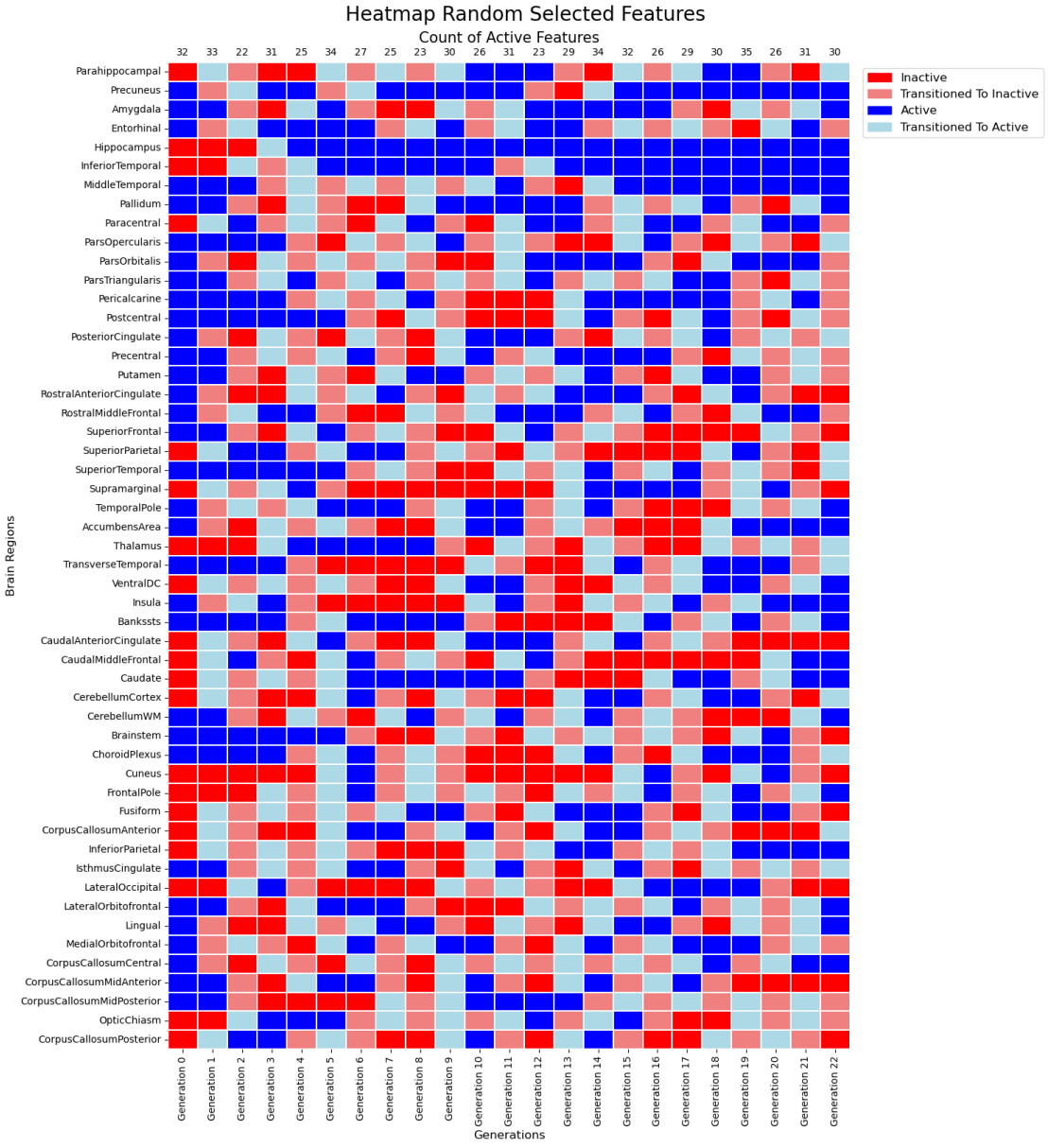

Figure 5: Heatmap Randomly Selected Features Over Generations

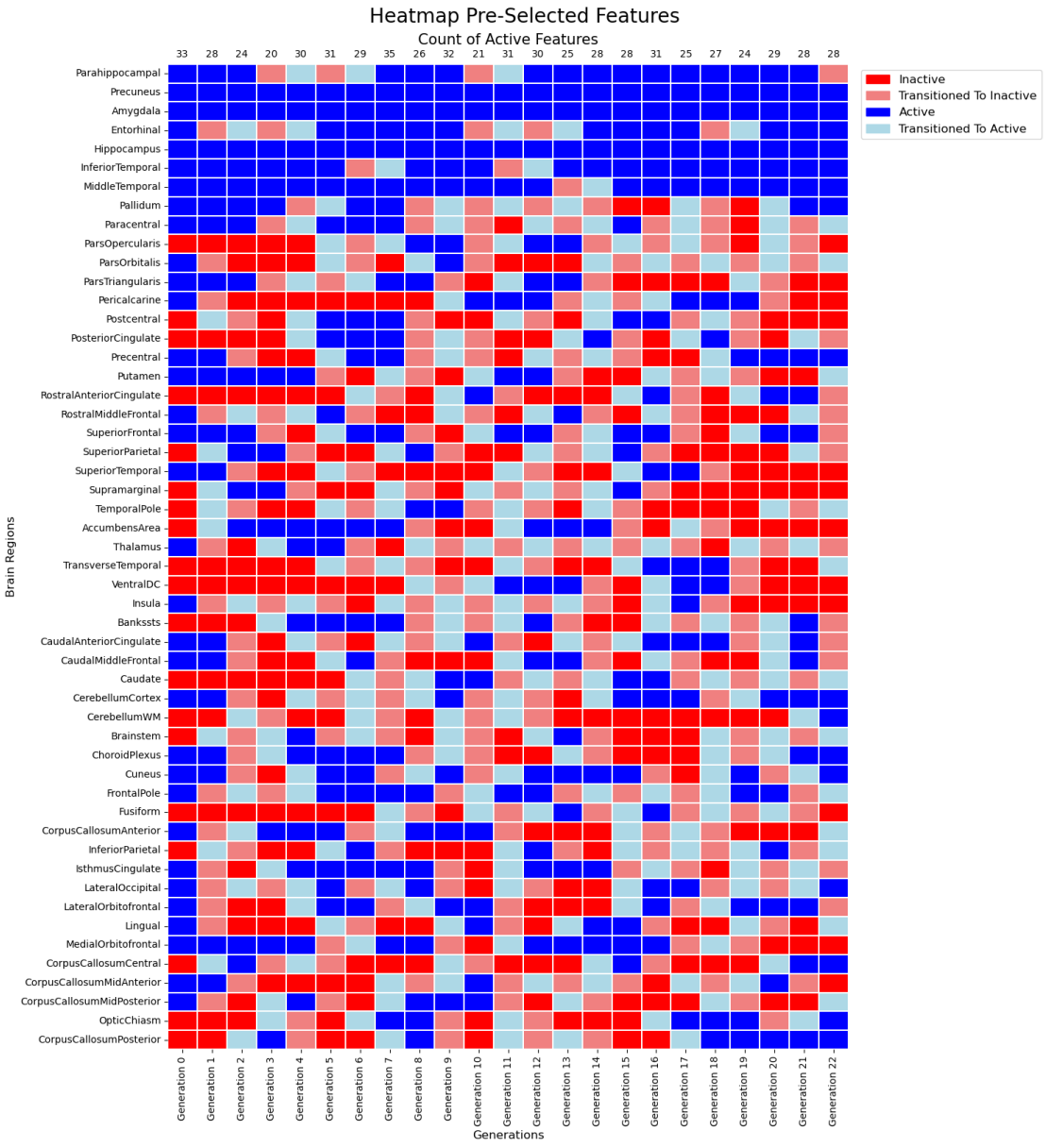

Figure 6: Heatmap Pre-Selected Features Over Generations

The report presents two heatmaps, each illustrating the progression of feature selection throughout 23 generations. Figure 5 shows the selected features over the generations for the randomized features. Figure 6 shows the selected features over the generations for the preselected features. One interesting point can be seen in Figure 5. Here all the known seven features from Hollenbenders are active, which seems to influence the Normalized Transition Score in a positive way as seen in Figure 3.

Also notable is that none of the selected features reach the fitness value of Hollenbenders, even if they were initialized with the known seven features.

Discussion

Random selection surpasses the performance of the benchmark set when all features are selected. Similarly, pre-selection techniques also outdo the all-features benchmark. Furthermore, Hollenbenders' model demonstrates superior fitness score, exceeding the other presented genetic methodologies.

Another key finding from this investigation is the behavior of the fully randomized algorithm. Interestingly, this algorithm rapidly reaches a plateau, suggesting that inherent optimization potential may be embedded. The hypothesis is that specific parameters, such as the mutation rate and generation size, may regulate this behavior. Hence, future research should focus on conducting a detailed sensitivity analysis of these parameters. This could enhance understanding of the dynamics underlying this algorithm, fine-tuning its performance for a more optimal selection of brain regions related to Alzheimer's disease.

Attention was also given to the algorithm that initiates its processes with pre-

selected features. This algorithm exhibited an oscillatory pattern around its starting point, calling for further exploration. Increasing the generation sizes and the number of generations might provide critical insights into the nuances of this behavior, potentially leading to a more precise selection of brain regions and improving the accuracy and reliability of Alzheimer's disease research.

Moreover, the fitness function evaluation and genetic algorithm optimization are considered crucial for the success of this novel approach. The fitness function, an essential evaluation metric in genetic algorithms, should be analyzed and optimized to ensure it accurately gauges the performance of the GAs in selecting relevant brain regions. Running the GA with a larger generation size and more generations could improve the performance and accuracy of the brain region selection, thus enhancing the reliability of the overall findings.

Regarding computational efficiency, transitioning the Hidden Markov Model (HMM) library from synchronous to asynchronous graphics processing unit (GPU) operations could enhance processing power and speed. This acceleration might translate into improved algorithmic performance and superior results in selected brain regions relevant to Alzheimer's disease.

Lastly, the GA outputs require rigorous examination, specifically selecting pertinent brain regions. Comprehensive statistical analyzes and plausibility checks of the selected regions are crucial. Validating the selected regions and confirming their alignment with known pathological features of Alzheimer's disease is paramount. Understanding these features could enable more strategic guidance of the GA, enhancing its effectiveness in identifying significant brain regions.

In conclusion, while the initial venture into applying GAs for selecting relevant brain regions in Alzheimer's disease research has yielded promising results,

it has also uncovered several areas for further investigation. This work serves as a stepping stone, setting the groundwork for advancing Alzheimer's disease research through the innovative application of genetic algorithms.

References

[1] 2022 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 18(4):700–789, 2022.

[2] hmmlearn. https://pypi.org/project/hmmlearn/, 4 2023.

[3] David Beasley, David R Bull, and Ralph Robert Martin. An overview of genetic algorithms: Part 1, fundamentals. *University computing*, 15(2):56–69, 1993.

[4] Ahmad Hassanat, Khalid Almohammadi, Esra'a Alkafaween, Eman Abunawas, Awni Hammouri, and VB Surya Prasath. Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach. *Information*, 10(12):390, 2019.

[5] Yasmin Hollenbenders, Monika Pobiruchin, and Alexandra Reichenbach. Two routes to alzheimer's disease based on differential structural changes in key brain regions. *Journal of Alzheimer's Disease*, (Preprint):1–14, 2023.

[6] Annu Lambora, Kunal Gupta, and Kriti Chopra. Genetic algorithm-a literature review. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 380–384. IEEE, 2019.

[7] Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, 2010.

[8] AP Porsteinsson, RS Isaacson, Sean Knox, MN Sabbagh, and I Rubino. Diagnosis of early alzheimer's disease: clinical practice in 2021. *The journal of prevention of Alzheimer's disease*, 8:371–386, 2021.

[9] Lawrence Rabiner and Biinghwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[10] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Jesse Cedarbaum, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, et al. 2014 update of the alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & dementia*, 11(6):e1–e120, 2015.