IDEEA: Information diffusion model for integrating gene expression and EEG data in identifying Alzheimer's Disease markers

Enes Ozelbas[a,*], Tuba Sevimoglu[b] and Tamer Kahveci[c]

[a]Yildiz Technical University, Computer Engineering, Istanbul, Türkiye
[b]University of Health Sciences, Bioengineering, Türkiye
[c]University of Florida, Computer and Information Sciences and Engineering, Gainesville, USA

ARTICLE INFO

Keywords:
Alzheimer's disease
Gene expression
Electroencephalogram
Integrative machine learning

ABSTRACT

Background and Objective: Understanding the genetic components of Alzheimer's disease (AD) via transcriptome analysis often necessitates the use of invasive methods. This work focuses on overcoming the difficulties associated with the invasive process of collecting brain tissue samples in order to measure and investigate the transcriptome behavior of AD.
Methods: Our approach called IDEEA (**I**nformation **D**iffusion model for integrating gene **E**xpression and **E**EG data in identifying **A**lzheimer's disease markers) involves systematically linking two different but complementary modalities: transcriptomics and EEG data. We preprocess these two data types by calculating the spectral and transcriptional sample distances, over 11 brain regions encompassing 6 distinct frequency bands. Subsequently, we employ a genetic algorithm approach to integrate the distinct features of the preprocessed data.
Results: Our experimental results show that IDEEA converges rapidly to local optima gene subsets, in fewer than 250 iterations. Our algorithm identifies novel genes along with genes that have previously been linked to AD. It is also capable of detecting genes with transcription patterns specific to individual EEG bands as well as those with common patterns among bands. Particularly the experiments in the alpha2 (10-13 Hz) frequency range yields a notable number of AD-associated genes with a p-value of 0.05. We evaluated various aspects of our approach, including the genetic algorithm performance and band-pair association.
Conclusions: Our approach reveals AD-relevant genes with transcription patterns inferred from EEG alone, across various frequency bands, avoiding the risky brain tissue collection process. This is a significant advancement toward the early identification of AD using non-invasive EEG recordings.

1. Introduction

Alzheimer's Disease (AD) is the most prevalent form of dementia, where patients gradually become unable to properly control their actions, lose the ability to react to their surroundings, interact, and carry on a conversation. The complex disease process begins earlier than its symptoms appear. At this time, most treatments may not be very effective. Hence, currently, there is no cure, and the treatments are merely used to help reduce the symptoms [3]. As reported by the World Health Organization, there are currently 55 million people with dementia worldwide, and AD accounts for at least 60 % of the cases. Every year about 10 million new cases will be added to this number [1]. Due to the increase in life expectancy, this overload is expected to increase over time, and have significant social, economic, and health system effects [5, 2]. Early diagnosis is of utmost importance to address the challenges imposed by AD.

The neurobiological process underlying AD is significantly impacted by variations in gene expression and regulation [9]. A multitude of transcriptomics studies have been carried out to elucidate this mechanism [27, 4, 18]. To date, 101 independent genomic variants across 81 loci have been identified [6]. There is a continuous effort to discover additional genes associated with the disease and AD-related pathways. However, our understanding of AD is hampered by the fact that the disease affects many brain regions, including the hippocampus, cerebellum, and frontal cortex, and that each region has different transcriptomics abnormalities [32, 20, 17]. Existing in vitro studies so far yield a dismal track record of success when they are applied in clinical studies of AD. This high failure rate has been partially attributed to the extrapolation of encouraging findings from animal models that only partially reflect human AD pathology [14]. Lack of physiologically relevant in vitro models that accurately represent the patient's genome in the target cell type and lack of knowledge of expression levels of the target genes in corresponding brain regions has been another major obstacle to our understanding of the molecular pathways behind AD [7]. The majority of the information we have on the genetic risk factors for AD comes from the analysis of blood samples, yet the genome is translated and transcribed differently in different organs in response to various transcription factors, metabolic signals, and environmental factors resulting in the inadequacy of blood samples to provide a complete picture of the mechanism [23]. The brain tissue samples from both affected and unaffected individuals are solely needed in order to conduct a thorough and comprehensive analysis of the gene transcriptional activity of the disease. This is however not possible with existing technologies as it requires invasive techniques to gather these samples.

enes.ozelbas@yildiz.edu.tr (E. Ozelbas); tuba.sevimoglu@sbu.edu.tr (T. Sevimoglu); tkahveci@ufl.edu (T. Kahveci)
ORCID(s): 0000-0003-4665-1952 (E. Ozelbas); 0000-0003-4563-3154 (T. Sevimoglu); 0000-0002-4403-8612 (T. Kahveci)

The distribution of disease across the brain and regional susceptibility to neurodegeneration may be influenced by spatial patterns of gene expression [8]. One way to quantify these patterns using non-invasive techniques is to study anatomical changes in both the entire brain and its particular regions. One of these techniques is the electroencephalogram (EEG), a mostly noninvasive test that detects aberrant brain activity in specific brain regions by detecting electrical signals produced by the brain using tiny sensors attached to the scalp. EEG measures changes in the brain signals that indicate unusual neuronal activity in various stages of AD as a time series [24]. A reduction in the complexity of EEG signals and changes in EEG synchrony leading to modifications in EEG recordings can be used for AD diagnosis [21]. Nonetheless making use of EEG data in AD is still a work in progress. The development of computational methods and programming tools has greatly aided our knowledge of the disease mechanism and the discovery of various genetic biomarkers [15]. For instance, Chedid and co-workers developed a fully automated EEG assessment process to detect AD in clinical settings [13]. Wu and colleagues developed a federated model to discover transcriptome and genetic impacts on brain sMRI measures in AD [40]. In another work by Sadegh Zadeh and colleagues, EEG data was used in a machine learning context to classify two and three-class configurations of AD and healthy control groups [33]. To the best of our knowledge, within the scope of Alzheimer's Disease research, no study has integrated transcriptomics and EEG data to date.

Our contributions. In this study, we address the challenges arising from the fact that gathering brain tissue samples to study the transcriptome behavior during the development of Alzheimer's disease is an invasive procedure and has a high risk of mortality. To do that, we develop an algorithm which systematically associates two distinct yet complementary modalities, namely transcriptomics and EEG data, thereby allowing us to observe EEG patterns via non-invasive, cheap, and fast data collection techniques and use these measurements as a surrogate from transcription patterns of AD genes. Our solution, named IDEEA (**I**nformation **D**iffusion model for integrating gene **E**xpression and **EE**G data in identifying **A**lzheimers disease markers), leverages a genetic algorithm strategy. It identifies a subset of genes of a given size, for which the differential transcription values between AD and healthy patients correlate with the differential EEG measurements for each EEG frequency band. By doing that, it reveals the set of AD-related genes whose transcription patterns can be estimated via EEG measurements alone for different frequency bands. Our method solves this problem by starting with a population of initial randomly generated subsets of genes. It then iteratively generates better gene subsets, by creating new gene subsets from the existing gene subsets through crossover, selection, and mutation operations. Our experimental results demonstrate that IDEEA converges quickly to local optima gene subsets in less than 250 iterations. It can find genes which are already verified as AD-associated in different studies among other novel genes as well. Finally, our results suggest that it can also find the genes whose transcription patterns are unique to EEG bands as well as those which are common across different bands. In summary, this study takes an important step toward early diagnosis of AD via non-invasive EEG measurements.

2. Methods

Here, we describe our algorithm, named IDEEA (Information Diffusion model for integrating gene Expression and EEG data in identifying Alzheimer's disease markers). Our algorithm employs a machine learning strategy to integrate transcription and EEG data from AD patients and healthy controls, and builds a model which rely only on EEG measurements to estimate anomalies on transcription values of AD patients. Figure 1 presents an overview of our IDEEA algorithm and Table 1 lists the variables and definitions used in our study. We elaborate on different steps of this algorithm from Section 2.1 to 2.4.

2.1. EEG preprocessing

We use a dataset containing resting state EEG recordings of 65 samples [28]. Among these, 36 are diagnosed as AD and the remaining 29 are in the healthy control (CN) group. We utilize the power spectral density (PSD) to help us compare the frequency domain characteristics of EEG signals of AD and CN groups, potentially revealing significant biomarkers for AD. We consider the EEG signals of each sample in six frequency bands: delta (δ, 0.5-4 Hz), theta (θ, 4-8 Hz), alpha1 ($\alpha 1$, 8-10 Hz), alpha2 ($\alpha 2$, 10-13 Hz), beta (β, 13-30 Hz), and gamma (γ, 30-45 Hz), represented by the set $\Omega = \{\delta, \theta, \alpha 1, \alpha 2, \beta, \gamma\}$. For any given band $\psi \in \Omega$, we denote its corresponding frequency range with the set $B = \{B_\delta, B_\theta, B_{\alpha 1}, B_{\alpha 2}, B_\beta, B_\gamma\}$.

To estimate the PSD of each sample across different frequency bands, we employ the Welch's method [39]. Welch's method reduces the variance of the estimated PSD by averaging multiple periodograms over a sequence of overlapping windows on the given PSD. This mitigates the effects of noise and provides a smoother spectrum.

Let us denote the continuous EEG signal as a function of time t, represented by $e(t)$. We split this signal into overlapping segments by sliding a window of length of five seconds. Windowing provides smooth edges in the segments, reducing spectral leakage during the time-to-frequency domain conversion. Let us denote the mth segment with $e_m(t)$. For each segment $e_m(t)$, we apply the windowing function $w(t)$ using the Hadamard product to obtain the windowed segment $w(t) \circ e_m(t)$. We then slide the window by half of the window size to get the next segment $e_{m+1}(t)$. This allows subsequent segments to share half of their data points. After windowing, we transform each segment into the frequency domain using the Fast Fourier Transform (FFT). Let us denote the number of frequency components as n. We denote the resulting vector with $\vec{F} = [F_1, F_2, \ldots, F_n]$. $\forall k, k \in \{1, 2, \ldots, n\}$, F_k corresponds to a discrete frequency value in Hertz (Hz) within the specified range of $0.5 \leq F_k \leq 45$

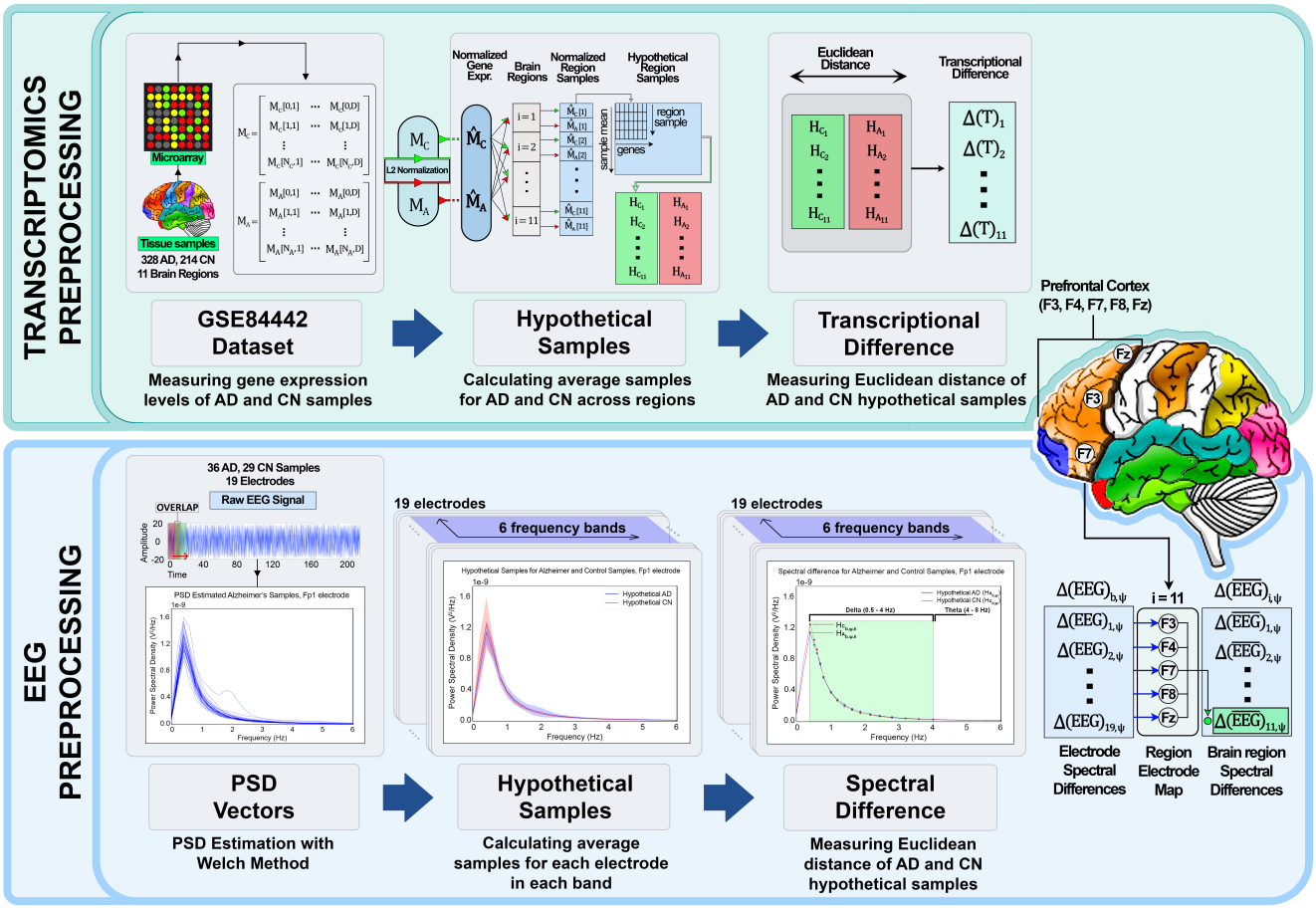

Figure 1: The flowchart depicts a parallel approach to analyze Alzheimer's Disease (AD) and control (CN) samples using transcriptomics and EEG datasets. On the top row, gene expression matrices from the tissue samples in GSE84442 Dataset are displayed, producing hypothetical samples of AD and CN across 11 brain regions. The variations in the hypothetical samples are quantified using Euclidean distance to indicate transcriptional differences. The bottom section shows the analysis of raw EEG signals from 19 electrodes divided into six distinct frequency bands. The Power Spectral Density (PSD) vector representations are derived and used to generate hypothetical samples reflecting spectral differences between AD and CN. An example of mapping of electrodes (F3, F4, F7, F8, Fz) to Prefrontal Cortex region by taking the mean is also shown at the end of the flowchart.

Hz. We denote the amplitude of the EEG signal at frequency F_k for the mth segment with $X_m(F_k)$. More specifically:

$$X_m(F_k) = \mathrm{FFT}\{w(t) \circ e_m(t)\} \tag{1}$$

Each sample in our dataset has EEG measurements collected over a set of electrodes. For sample a and electrode b, we represent the FFT of the EEG sequence corresponding to the kth frequency component as $X_{a,b}(F_k)$. We compute the power of each discrete frequency component of the FFT as $|X_{a,b}(F_k)|^2$.

Let us denote the total duration of the given signal with T, and the total number of discrete frequencies in band ψ, as determined by the FFT, with N_ψ. The sum of these power values, averaged over the total duration, returns the PSD vector $\vec{P}$. Mathematically, we derive $\vec{P}$ for the bth electrode of ath sample for band ψ as:

$$\vec{P}_{a,b,\psi} = \frac{1}{T} \sum_{k \in B_\psi}^{N_\psi} |X_{a,b}(F_k)|^2 \tag{2}$$

The resulting PSD vectors provide a detailed and powerful representation of the EEG as it shows the power distribution within each frequency band for every electrode location and each sample.

For each sample, we apply the L2 normalization on vector $\vec{P}_{a,b,\psi}$ individually for each frequency band and electrode. This ensures that the sum of squares of the normalized PSD components across a band spectrum equals one. We perform this L2 normalization for $\vec{P}_{a,b,\psi}$, over the frequency index k. Thus, for the ath sample and bth electrode, we compute the L2 norm over N_ψ frequency points of band ψ as:

Table 1
Definition of variables used in the study

Variable	Definition
ψ	Specific frequency band from Ω
T	Total duration of the given signal
E	Number of electrodes
L	Number of frequency bands
d	Number of selected genes in each solution
D	Number of genes in the transcription dataset
N_{pop}	Number of solutions in the population
ϕ	A solution mask in a given population
$\rho_{\psi,s}$	Spearman correlation between EEG and transcriptomics profiles for band ψ and selected genes s
f'_{ϕ}	Normalized fitness of solution ϕ
Ω	Set containing names of all frequency bands
B	Set containing all frequency band ranges
S, S'	Set of solutions in current and next generation respectively
$\vec{F}$	Vector containing discrete frequency components
$\vec{P}_{a,b,\psi}$	PSD vector for the bth electrode of ath sample for band ψ
$\hat{P}_{a,b,\psi}$	Normalized PSD vector for the bth electrode of ath sample for band ψ
$\Delta(\text{EEG})_{b,\psi}$	Distance value for Euclidean distance between AD and CN hypothetical groups for band ψ of electrode b
$\Delta(\overline{\text{EEG}})_{\psi}$	Spectral difference vector for Euclidean distances between AD and CN groups for band ψ across all 11 brain regions
$\Delta(\text{T})_s$	Transcriptional profile differences across all regions for selected genes s
$R_{\Delta(\tau)}$	Rank vector for a list of distances $\Delta(\tau)$
$X_m(F_k)$	FFT of the mth segment of the EEG signal for the kth frequency component
$X_{a,b}(F_k)$	FFT of the EEG sequence for sample a, electrode b, corresponding to kth frequency component
M_A, M_C	Matrices representing transcriptomics data for AD and CN groups respectively
H_{A_i}, H_{C_i}	Hypothetical samples for AD and CN groups for region i
$\Delta(T)_i$	Euclidean distance representing transcriptional profile differences between AD and CN groups for region i

$$\|\vec{P}_{a,b,\psi}\|_2 = \sqrt{\sum_{k=1}^{N_\psi} P^2_{a,b,\psi,k}} \tag{3}$$

We then compute the normalized PSD vector $\hat{P}_{a,b,\psi}$ as:

$$\hat{P}_{a,b,\psi} = \frac{\vec{P}_{a,b,\psi}}{\|\vec{P}_{a,b,\psi}\|_2} \tag{4}$$

Next, for each electrode b and frequency band ψ, we derive a hypothetical sample for each CN and AD groups, represented as $H_{C_{b,\psi}}$ and $H_{A_{b,\psi}}$ respectively. We compute these samples based on the mean of the normalized PSD values over their respective groups. Let us denote the number of CN and AD samples with N_C and N_A, respectively. We compute $H_{C_{b,\psi}}$ and $H_{A_{b,\psi}}$, as:

$$H_{C_{b,\psi}} = \frac{1}{N_C}\sum_{a=1}^{N_C} \hat{P}_{a,b,\psi} \quad \text{and} \quad H_{A_{b,\psi}} = \frac{1}{N_A}\sum_{a=1}^{N_A} \hat{P}_{a,b,\psi} \tag{5}$$

Following this transformation, we measure the Euclidean distance between the hypothetical samples of the AD and CN groups across predefined frequency bands for each electrode. Let E be the total number of electrodes and L be the total number of frequency bands. To capture the spectral differential, we compute the Euclidean distance between the AD and CN groups denoted with $\Delta(\text{EEG})_{b,\psi}$ as:

$$\Delta(\text{EEG})_{b,\psi} = \sqrt{\sum_{b=1}^{E}\sum_{\psi=1}^{L}\left(H_{C_{b,\psi}} - H_{A_{b,\psi}}\right)^2} \tag{6}$$

2.2. Transcriptomics preprocessing

We use a transcriptomics dataset containing 328 AD and 214 CN samples over 11 brain regions. We represent the AD and CN data with matrices M_A and M_C, respectively. Here, each row of these matrices corresponds to a unique brain region sample, and each column a gene. Let us also denote indices for sample and gene axes with i and j, respectively. We first apply L2 normalization for each gene in each matrix. This means that for each gene indexed j, we normalize its expression values across different samples y in the matrices M_A and M_C to have a unit L2 norm at each column. Mathematically, let us denote the value at the ith row and jth column of the M_C with $M_C[i, j]$. We compute the inverse of the L2 norm of the jth column (i.e., gene) of M_C as:

$$\xi_j = 1 \,/\, \sqrt{\sum_{i=1}^{N_C} M_C[i,j]^2}$$

Let us denote the number of genes with D. We denote inverse L2 norms of all genes with vector $\xi = [\xi_1, \xi_2, \ldots, \xi_D]^T$. Let us denote the $D\times D$ identity matrix with I_D. We compute the normalized matrix for CN samples as:

$$\hat{M}_C = M_C I_D \xi \tag{7}$$

We normalize the matrix M_A for the AD group, similarly.

In order to align the transcriptomics data with the EEG-derived hypothetical samples, we construct region-specific hypothetical samples for each brain region. We isolate the normalized transcriptomic data of region i with matrices

$\hat{M}_A[i]$ and $\hat{M}_C[i]$. For a given region i, let us denote the number of CN and AD samples of that region with N_{A_i} and N_{C_i}, respectively. We derive the mean vectors of H_{A_i} and H_{C_i} across all samples as:

$$H_{A_i} = \frac{1}{N_{A_i}} \sum_{h=1}^{N_{A_i}} \hat{M}_A[i,h] \quad \text{and} \quad H_{C_i} = \frac{1}{N_{C_i}} \sum_{h=1}^{N_{C_i}} \hat{M}_C[i,h] \tag{8}$$

Given these hypothetical samples, we then compute the Euclidean distances between the AD and CN groups for each brain region i. The Euclidean distance, denoted as $\Delta(T)_i$, provides means to differentiate transcriptional profiles between the two groups for a specific region. More specifically:

$$\Delta(T)_i = \sqrt{\sum_{j=1}^{D} \left(H_{A_{i,j}} - H_{C_{i,j}} \right)^2} \tag{9}$$

Later, the distances calculated for each gene across all brain regions serve as values that form many different combinations of genes that represent solutions to be utilized in our algorithm in 2.4.

2.3. Brain region mapping

Here, we leverage literature-driven information and retain only those transcriptomics dataset regions where EEG data regions are also a part of. To establish correlation between EEG and transcriptomic profiles we utilize the relative positions of the electrodes, their functional relationships defined within the concept of Broadmann Areas (BA), and the variability analyses [34].

In order to eliminate noise, we exclude transcriptomics samples acquired from deep brain regions in the transcriptomics dataset. Specifically, from these regions we exclude Amygdala, Nucleus Accumbens, Parahippocampal Gyrus, Hippocampus, Caudate Nucleus, and Putamen. Hence, we retain samples from the cerebellar cortex to ensure a higher degree of overlap between the EEG and transcriptomics datasets, as an EEG recording is more likely to capture activity from the outer surface of the brain. This leads to selection of 11 brain regions. In total, 19 electrodes monitor the EEG activities of these 11 brain regions. Note that the mapping between electrodes and brain regions is a many-to-many relationship. That is, an electrode may monitor multiple brain regions, and a brain region may be monitored by multiple electrodes. Table 2 lists the 11 brain regions, and the electrodes monitoring each of these regions.

For each identified region, we compute the mean spectral distances between AD and CN hypothetical samples corresponding to the relevant electrodes, denoted with $\Delta(\overline{\text{EEG}})_{b,\psi}$. More specifically, let us take the 'Frontal Pole' region as an example from Table 2. Assume that this is the first brain region, (i.e., i=1). We compute its mapped region distance $\Delta(\overline{\text{EEG}})_{1,\psi}$ by averaging the hypothetical sample distances from electrodes 'Fp1' and 'Fp2'. More specifically:

Table 2
Determined brain regions and their corresponding mapped electrodes.

Index	Regions	Mapped Electrodes
1	Frontal Pole	Fp1, Fp2
2	Occipital Visual Cortex	O1, O2
3	Inferior Temporal Gyrus	F7, F8
4	Middle Temporal Gyrus	F7, F8
5	Superior Temporal Gyrus	T3, T4, F7, F8, T5, T6
6	Posterior Cingulate Cortex	Pz
7	Anterior Cingulate	Fz
8	Temporal Pole	T3, T4, T5, T6
9	Precentral Gyrus	C3, Cz, C4
10	Superior Parietal Lobule	P3, P4, Pz
11	Prefrontal Cortex	F3, F4, F7, F8, Fz

$$\Delta(\overline{\text{EEG}})_{1,\psi} = \frac{\Delta(\text{EEG})_{(\text{Fp1}),\psi} + \Delta(\text{EEG})_{(\text{Fp2}),\psi}}{2} \tag{10}$$

This representation allows us to capture the differences between AD and CN groups across the distinct frequency bands and brain regions.

2.4. Gene selection algorithm

We develop a genetic algorithm to identify the gene set with a prespecified number of genes, whose variation of transcriptional profile correlates with that of their EEG profiles in 11 brain regions among CN and AD samples. This algorithm designs novel selection, crossover and mutation operators, unique to handle both transcription and EEG profiles, thus allowing efficient navigating of the search space.

Initialization. Our algorithm starts by creating a population of randomly generated solutions. Let us denote the number of selected genes in each solution with d. Also, let us denote the total number of genes in our transcription dataset with D. Let us denote a given solution with a binary vector s of length D. The jth value in s, denoted with s_j corresponds to the jth gene. For $j \in \{1, 2, \dots, D\}$, if the jth gene is a part of that solution, we set $s_j = 1$. Otherwise we set $s_j = 0$. Note that since each solution has d selected genes, exactly d values in s are equal to 1.

Fitness evaluation. Briefly, given a solution s of size D with d selected genes, the fitness function measures how well the differences between the EEG values of AD and CN samples at 11 brain regions mirror the transcription differences of the AD and CN samples of those selected genes in the same 11 brain regions.

More specifically, we compute this function as follows. For each frequency band $\psi \in \Omega$, we construct a vector of spectral distances for the 11 brain regions, denoted with $\Delta(\overline{\text{EEG}})_{\psi}$ as:

$$\Delta(\overline{\text{EEG}})_{\psi} = [\Delta(\overline{\text{EEG}})_{1,\psi}, \Delta(\overline{\text{EEG}})_{2,\psi}, \dots, \Delta(\overline{\text{EEG}})_{11,\psi}]$$

Recall that $\Delta(\text{EEG})_{b,\psi}$ is the spectral difference between AD and CN samples for the bth brain region and frequency

band ψ (see Equation 6). Also recall that $\Delta(T)_i$ consists of transcriptional distances across all genes for region i (see Equation 9).

Firstly, for a given binary solution mask s of size D, we apply element-wise multiplication to our transcriptional distance matrix $\Delta(T)_i$. The resulting distance matrix for each region i, containing the distances of the genes selected by the solution mask is given by $\Delta(T)_{i,s} = \Delta(T)_i \circ s$. Then, we construct the transcriptomics distances $\Delta(\mathrm{T})_s$ for the 11 regions as:

$$\Delta(\mathrm{T})_s = [\Delta(\mathrm{T})_{1,s}, \Delta(\mathrm{T})_{2,s}, \dots, \Delta(\mathrm{T})_{11,s}]$$

Each value in $\Delta(\mathrm{T})_s$ indicates the differences in the transcriptomic activity for the selected genes in one of the 11 brain regions.

After obtaining the EEG and transcriptomics distances, denoted as $\Delta(\overline{\mathrm{EEG}})_\psi$ and $\Delta(\mathrm{T})_s$ respectively, we calculate rank vectors for each brain region. We determine the rank vectors, $R_{\Delta(\overline{\mathrm{EEG}})_\psi}$ and $R_{\Delta(\mathrm{T})_s}$, by sorting each list in ascending order and assigning ranks based on their position. We explain this on a small hypothetical example for three brain regions. Consider a list of distances given by $\Delta(\tau) = [0.5, 0.9, 0.3]$. If we sort this list in ascending order, we obtain $[0.3, 0.5, 0.9]$. Thus, the rank for the value 0.3 is 1, for 0.5 is 2, and for 0.9 is 3 (i.e., 0.3 is the smallest value, 0.5 is the second smallest value and 0.9 is the third). Hence, the rank vector for $\Delta(\tau)$ is $R_{\Delta(\tau)} = [2, 3, 1]$.

Let us denote covariance and standard deviation functions with $cov()$ and σ, respectively. We compute the Spearman correlation $\rho_{\psi,s}$ between the EEG and transcriptomics profiles for band ψ and selected genes s as:

$$\rho_{\psi,s} = \frac{\mathrm{cov}\left(R_{\Delta(\overline{\mathrm{EEG}})_\psi}, R_{\Delta(\mathrm{T})_s}\right)}{\sigma_{R_{\Delta(\overline{\mathrm{EEG}})_\psi}} \times \sigma_{R_{\Delta(\mathrm{T})_s}}} \tag{11}$$

A higher value of $\rho_{\psi,s}$ indicates a stronger relevance between the EEG and transcriptomics profiles. This evaluation metric drives the algorithm searching for genes that effectively capture the underlying relationship between EEG profiles and transcriptional activity across the predetermined 11 brain regions.

2.4.1. Crossover

Let us denote the set of solutions in the current population with S_c. Crossover operator combines genetic information from two existing solutions (called parent solutions) to create new solutions which partially carry characteristics of the two selected solutions. Our custom crossover strategy ensures that the offsprings inherit the genes for which both parents agree. To make sure that the offspring solution has the same number of selected genes as the parents, it then randomly picks more genes among those the parent solutions disagree.

We explain this on a small hypothetical example. Assume that in each solution mask, the total number of genes is D=7, and the number of selected genes in each solution is d=4. Also assume that we pick parent solutions $\phi_1 = [1, 0, 1, 1, 0, 0, 1]$ and $\phi_2 = [1, 1, 0, 1, 0, 1, 0]$ for crossover. Let us denote the logical OR, AND and XOR operators with symbols $\vee$, $\wedge$, and $\oplus$ respectively. Genes common to both parents are $\phi_1 \wedge \phi_2 = [1, 0, 0, 1, 0, 0, 0]$ and genes unique to one parent are $\phi_1 \oplus \phi_2 = [0, 1, 1, 0, 0, 1, 1]$. There are 2 genes common to both parents. Thus, we need $d - 2 = 2$ more genes to create a new valid solution. To do that, we randomly pick two genes from $\phi_1 \oplus \phi_2$ and include them in $\phi_1 \wedge \phi_2$.

Assume that our random selection picks genes 2 and 7 (i.e., $r_1 = [0, 1, 0, 0, 0, 0, 1]$) and genes 6 and 7 (i.e., $r_2 = [0, 0, 0, 0, 0, 1, 1]$). Then, the two new solutions we generate are:

$$\phi'_1 = (\phi_1 \wedge \phi_2) \vee r_1 = [1, 1, 0, 1, 0, 1, 1]$$

and

$$\phi'_2 = (\phi_1 \wedge \phi_2) \vee r_2 = [1, 0, 0, 1, 0, 1, 1]$$

We repeat the procedure of picking two parents and generating new solutions until we generate N_{pop} new solutions. We denote the set of new solutions with S'.

2.4.2. Selection

Previous step doubles the number of solutions by generating N_{pop} new solutions. At this step, we reduce it down to N_{pop} by selecting a subset of the solutions in $S_i \cup S'$. During the selection process, we apply elitism for the best fit member of the population. That is, we guarantee that the solution with the highest fitness is preserved for the next generation.

For the remaining $2N_{pop} - 1$ solutions, we employ the roulette-wheel selection as it provides a balanced level of exploration and exploitation [22]. The main objective of this scheme is to introduce a fitness proportionate selection, thus allowing solutions with lower fitness values to have a non-zero chance to be selected. This helps maintain the gene diversity within the population, preventing premature convergence to a suboptimal gene set.

The Spearman correlation values range in the [-1, 1] interval. To avoid potential biases in the roulette wheel selection process, we normalize the fitness of each solution in the population to fall within the [0, 1] range, and revert back to its original range after the selection. If not addressed, negative correlations could result in certain solutions being unfairly penalized, leading to their unwarranted exclusion during the selection step.

We denote the fitness of solution ϕ with f_ϕ and compute the normalized fitness score f'_ϕ relative to a given frequency band ψ as:

$$f'_\phi = \frac{\rho_{\psi,\phi} + 1}{2} \tag{12}$$

Consider a population with a total of $2N_{pop}-1$ solutions. We define the vector $F' = [f'_1, f'_2, \dots, f'_{2N_{pop}-1}]$ as the

normalized fitness vector, which comprises the normalized fitness scores of each member in ascending order.

In roulette wheel selection, each member in the generation population occupy a segment of the wheel, specified by its value in F'. The size of the segment is proportional to the member's normalized fitness f'_ϕ. Thus, the probability of a member being selected is directly proportional to its normalized fitness. More specifically it is equal to the ratio of the normalized fitness of the member to the aggregated normalized fitness of the entire population. Let $\mathcal{S}^*$ represent the union of population sets $\mathcal{S}_c$ and $\mathcal{S}'$, (i.e., $\mathcal{S}^* = \mathcal{S}_c \cup \mathcal{S}'$). Mathematically, the probability of selecting a solution $\phi \in \mathcal{S}^*$ is:

$$P(\phi) = \frac{f_\phi}{\sum_{\phi_q \in \mathcal{S}^*} f_{\phi_q}} \tag{13}$$

We repeat selecting solutions using this strategy from among $2N_{pop} - 1$ current solutions, until we have exactly N_{pop} solutions selected. We call the resulting population of solutions $\mathcal{S}_{c+1}$.

2.4.3. Mutation

While crossover derives offsprings from the recombination of solutions in the population, mutation introduces small changes to gene configuration of a solution. The goal is to ensure genetic diversity in the population and thus explore new regions of the solution space. We design our mutation strategy to ensure in a way that preserves the number of selected genes in each mutated solution. We do not mutate the best member of the population to ensure that it is passed to the next generation.

Let us denote the mutation rate of a gene and a solution with μ_g and the solution mutation rate with μ_s, respectively. Here, μ_g is the probability of a single gene undergoing mutation within an individual solution. For each solution in $\mathcal{S}_{c+1}$ (except for the solution with the highest fitness), we flip a biased coin with success probability of μ_s. If it is successful, for each selected gene in that solution, we flip another biased coin with success probability of μ_g. When the coin flip is successful for a selected gene, we remove that gene from that solution and insert another randomly selected gene to the same solution.

From our earlier crossover example, let us use the solution $\phi'_1 = [1, 1, 0, 1, 0, 1, 1]$. Assume that the algorithm arbitrarily selects the second gene for mutation, and the fifth gene as replacement. After the mutations, the solution transforms into $\phi''_1 = [1, \mathbf{0}, 0, 1, \mathbf{1}, 1, 1]$ (we show the mutant genes in boldface).

2.4.4. Termination condition

After creating the initial population of solutions $\mathcal{S}_0$, our algorithm iteratively updates the current population. At each iteration c, we apply crossover, selection and mutation to $\mathcal{S}_c$ to generate next generation $\mathcal{S}_{c+1}$ of solutions. The highest fitness observed at each generation monotonically increases. After each generation, we track the fitness of the solutions, and dynamically adjust mutation probabilities to avoid stagnation. If our mutation metrics exceed certain thresholds, we take corrective actions, including the possibility of early termination. The adaptive mutation strategy ensures that our genetic algorithm can dynamically adjust its evolutionary behavior based on fitness performance across generations. This design choice especially helps with maneuvering through multiple local optima and promoting further exploration of the solution space.

We set the initial mutation probabilities as $\mu_s = 0.1$ and $\mu_g = 0.2$. As generations progress, the algorithm adjusts the values of μ_s and μ_g based on the predefined rules. Let us denote maximum fitness in given generation $\mathcal{S}_c$ with $f^{(c)}_{\max}$ and denote the maximum fitness in the previous generation with $f^{(c-1)}_{\max}$. We adapt the mutation rates as follows:

1. If $f^{(c)}_{\max} = f^{(c-1)}_{\max}$, the algorithm perceives this as stagnation and increases the stagnation counter. If such a condition persists for a predefined number of consecutive generations, it increases the value of μ_s by 50%. Concurrently, the algorithm checks for the following:
   (a) If $\mu_s \geq 0.2$, the algorithm increments the μ_g by 25% to promote mutation rates across all genes. Following that, it resets μ_s to 0.05 to prevent excessive gene mutations.
   (b) If $\mu_g \geq 0.4$, the algorithm terminates, as the mutation rate becomes too aggressive.
2. If $f^{(c)}_{\max} > f^{(c-1)}_{\max}$, the algorithm perceives this as a progression and resets the stagnation counter.

3. Results

In this section we evaluate the performance of the proposed IDEEA algorithm in terms of its convergence to solutions and ability to find the AD related features, such as AD-associated genes and their protein interactions across different EEG bands.

Dataset Description. We evaluate our method using an EEG and a transcriptomics dataset.

EEG DATASET: This dataset contains resting state-closed eyes recordings from 88 samples, among which 36, 23, and 29 belong to Alzheimer's disease (AD group), Frontotemporal Dementia (FTD group), and healthy subjects (CN group) respectively [28]. For our analysis, we omit the FTD patients, focusing on the 65 subjects in the AD and CN groups. The EEG data is collected using 19 scalp electrodes (Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2) according to the 10-20 international system. The mean age of the AD group is 66.4 with a standard deviation of 7.9 and 67.9 for the CN group with a standard deviation of 5.4. Thus, AD and CN groups have similar age distributions. During the EEG recording, subjects were in a sitting position with eyes closed. Skin impedance was below 5kΩ. The sampling rate of the recording was 500 Hz with 10μV/mm resolution. To eliminate noise in the dataset, the authors applied a Butterworth band-pass filter (0.5-45 Hz),

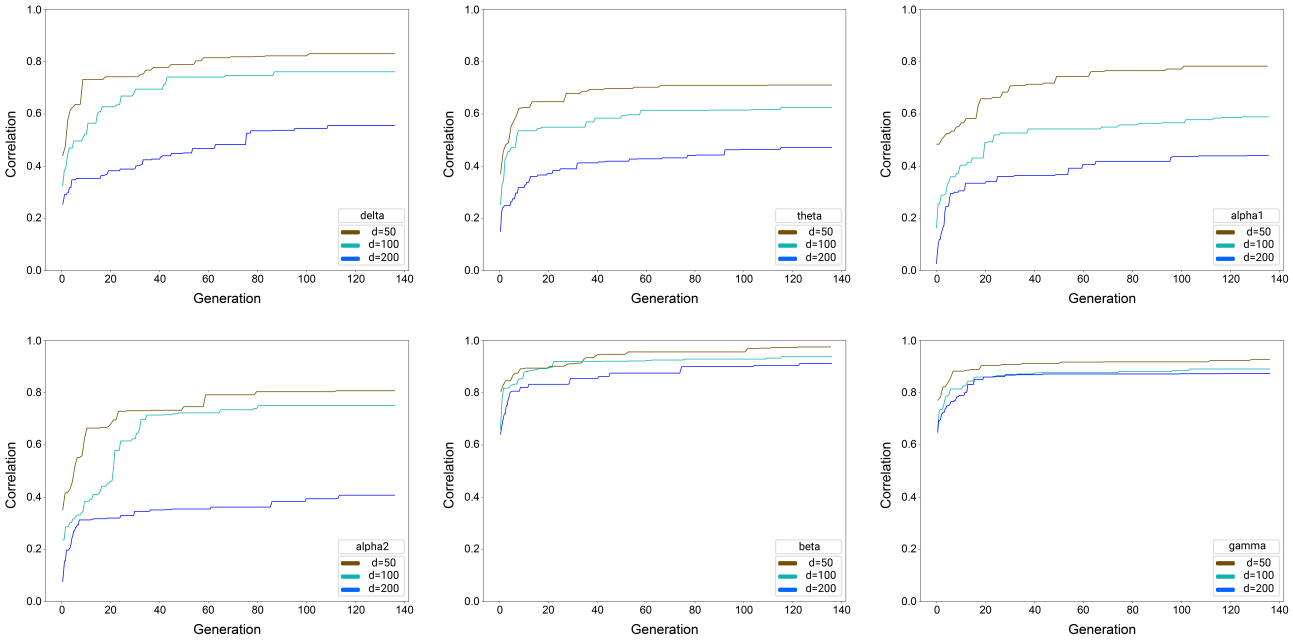

Figure 2: Fitness progress across varying gene set sizes (50, 100, 200) with a constant population size (100) for different EEG frequency bands.

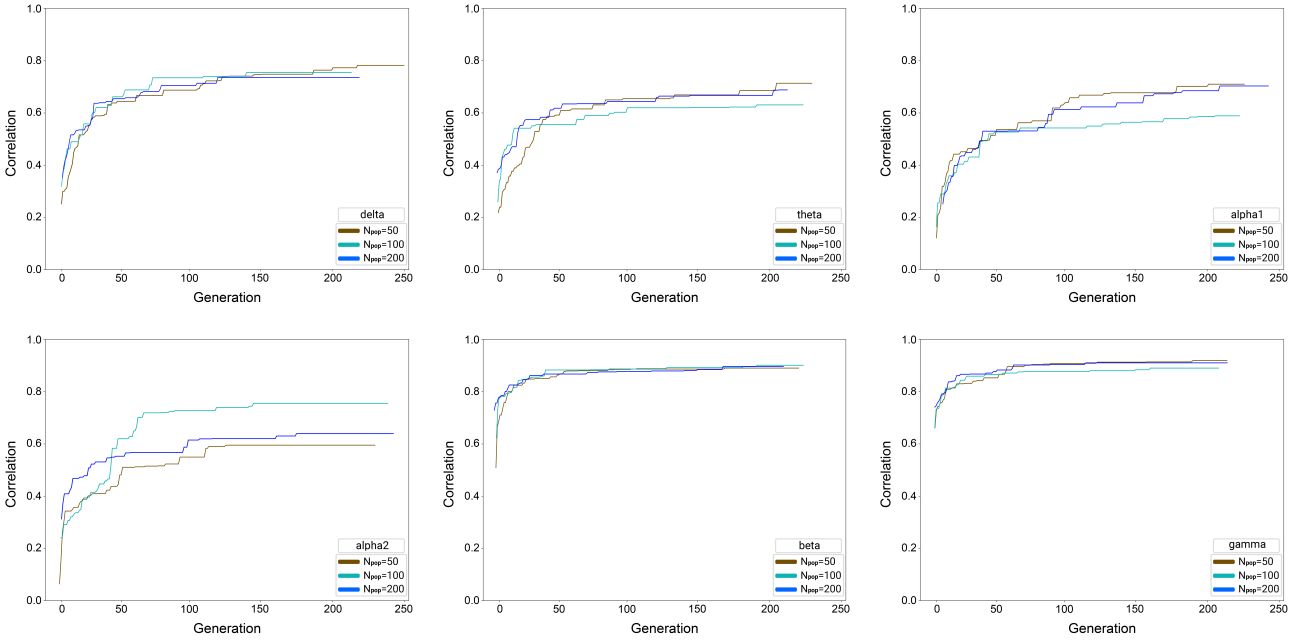

Figure 3: Fitness progress across varying population sizes (50, 100, 200) with a constant gene set size (100) for different EEG frequency bands.

re-referencing to A1-A2, Artifact Subspace Reconstruction (ASR), and Independent Component Analysis (ICA). They also removed the artifacts stemming from eye movement and muscle activity.

TRANSCRIPTION DATASET: For the transcriptomics dataset, we use 1053 post-mortem brain samples, collected from 19 cortical regions of 125 individuals [38] (GSE84422). These individuals represent a diverse range of dementia severity and neuropathology characteristic of AD. The dataset is originally collected using three platforms (GPL96, GPL97, GPL570). It contains gene expression data from 2004 samples classified as CN, definite AD, probable AD, and possible AD. We prioritized our focus on definite AD cases. In our study we use the GPL96 platform as it contains the largest number of samples among the three platforms with 328 AD and 214 CN samples. It features a total of 12,937

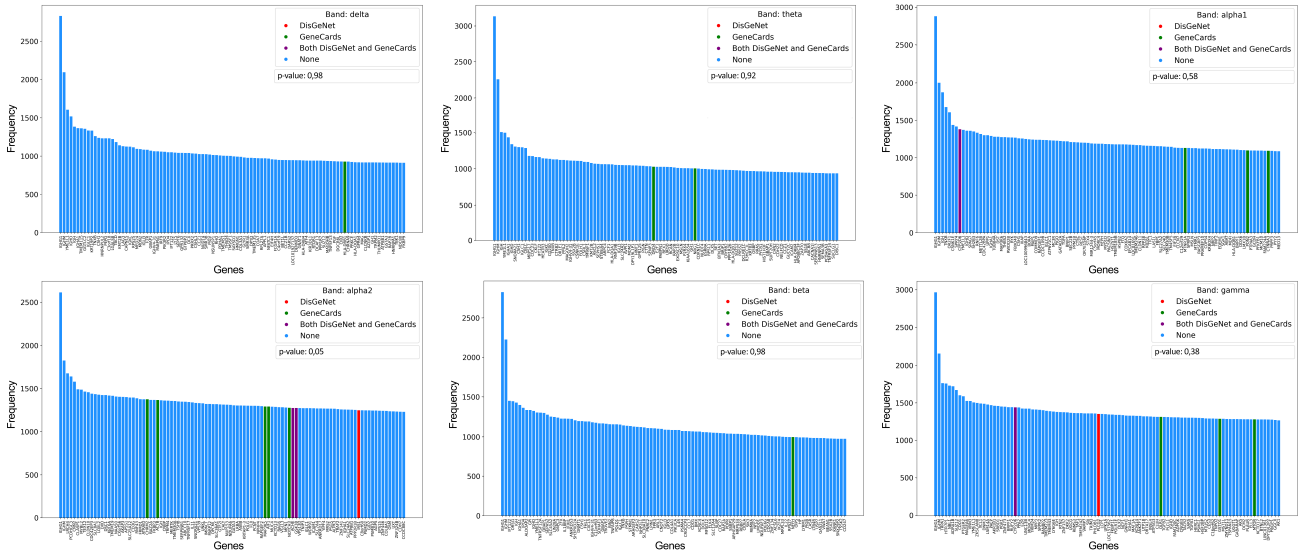

Figure 4: Distribution of the 100 most frequently selected genes with $N_{pop} = 100$, $d = 200$ for different EEG bands.

differentially expressed genes. From the 19 brain regions available, which cover both deep brain and cerebellar cortex areas, we omit the deep brain regions and concentrate on the remaining 11 regions. EEG recording is predominantly sensitive to the activity from the outer surface of the brain. Table 2 lists these regions. This dataset provide Robust Multi-array Average (RMA)-normalized transcription data.

SYSTEM DETAILS: We conduct all computational experiments on a system equipped with 32 GB DDR4 RAM, an AMD Ryzen 5 3600 CPU with 6 cores, and an NVIDIA RTX 2070 GPU with 8 GB of VRAM.

3.1. Evaluation of genetic algorithm performance

We start by evaluating the convergence behavior of our algorithm over iterations. Specifically, we evaluate how its convergence speed is affected by the gene set size d and the population size N_{pop}. We measure the influence of these parameters across different bands.

3.1.1. IDEEA parameters: Role of gene set size

In order to understand the impact of the gene set size, in Figure 2, we vary the gene set size d to values 50, 100 and 200 while keeping the population size N_{pop} constant at 100. For each frequency band, we run our algorithm until convergence of the frequency band with the most number of generations till convergence (which is less than 140 generations). We repeat this process five times with different initializations, leading to 180 experiments (i.e., $5 \times 6 \times 3$ with 5 repeats, 6 frequency bands, and 3 gene set sizes). We compute the fitness value of our top result with the highest fitness value for each generation and report the average of the five runs for each experimental set up.

3.1.2. IDEEA parameters: Role of population size

To study the impact of population size, we fix the number of genes to be emulated by EEG measurements to $d = 100$, and vary the population size as 50, 100, and 200. Similar to the previous experiment, for each parameter setting, we repeat each experiment five times with different random and report the average fitness of the best result of each generation. Thus, we run 180 experiments (i.e., 5 repeats $\times$ 6 frequency bands $\times$ 3 population sizes). We run our algorithm until convergence, which happens in less than 250 generations. We present our results in Figure 3

3.2. Evaluation of gene selection performance

In our next set of experiments we aim to understand the range of biological implications of our IDEEA method for studying the relationship between the EEG signals and transcription patterns for AD disease.

3.2.1. Distribution of AD-related genes

Recall that the goal of our IDEEA algorithm is to identify the genes which have differential transcriptional behaviors for Alzheimer's patients without knowing the actual transcription values, by studying the EEG patterns. Here, we evaluate how well we achieve this goal. To do that we focus on the top 100 most frequently selected genes by our algorithm in a total of 5 runs per frequency band. We also obtain the list of Alzheimer's Disease associates genes from two databases; Gene Disease Association (GDA) scores from DisGeNet [29] and GeneCards [36]. These scores provide a quantifiable measure of the relationship between individual genes and Alzheimer's disease, with higher scores indicating a stronger association. We define thresholds for both database GDA scores based on the 95th percentile, aiming to focus on the top 5% of genes most strongly associated with AD. We then sort the top 100 genes our algorithm finds in descending order of their frequency (i.e., the number of times each gene appears in a top solution found by our algorithm across different iterations). We plot the frequencies of these genes with the colors green, red,

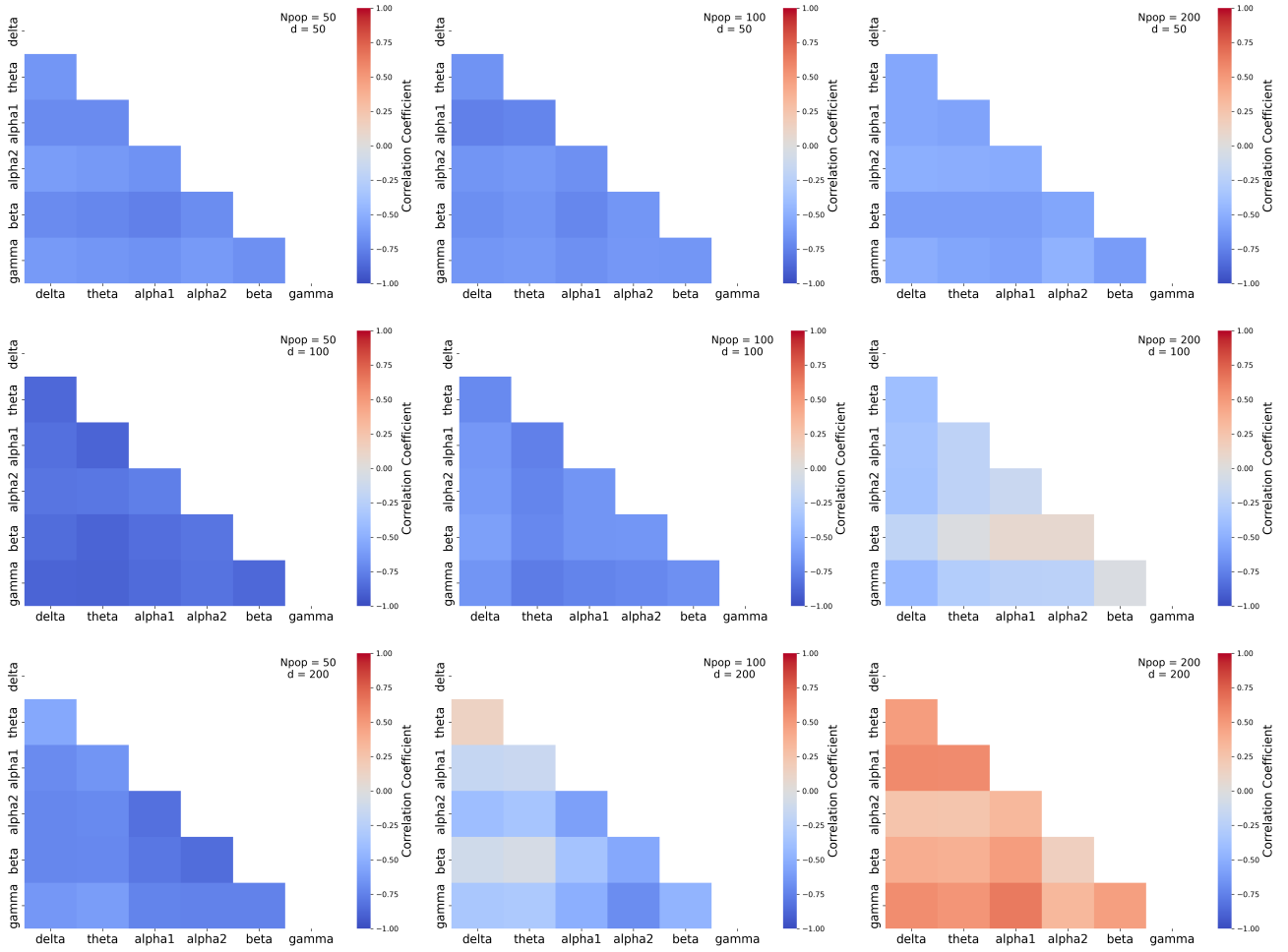

Figure 5: Heatmaps for correlation of 100 most frequently selected genes between bands with varying N_{pop} (50, 100, 200) and d (50, 100, 200) sizes.

and purple if they appear in GeneCards, DisGeNet, and both databases, respectively. In summary, we perform this evaluation by conducting 30 runs of experiments using our method (5 random initial populations × 6 frequency bands) and compare the results against four combinations of gene classifications with respect to two disease/gene association databases. Figure 4 presents our results.

3.2.2. Evaluation of band-pair association

An important question which follows from our findings in Figure 4 is how much variation does our method exhibit in selecting top genes across different EEG frequency bands. To answer this question, we compare the top 100 genes identified by our method for every pair of frequency bands. More specifically, consider two frequency bands $\psi_1, \psi_2 \in \Omega$ with $\psi_1 \neq \psi_2$. Let us denote the top 100 genes our IDEEA algorithm identifies as G_{ψ_1} and G_{ψ_2} respectively. For each parameter combination $N_{pop} \in \{50, 100, 200\}$ and $d \in \{50, 100, 200\}$ we compute the Pearson's correlation between the frequencies of the genes in $G_{\psi_1} \cup G_{\psi_2}$ using the band ψ_1, and those using the band ψ_2 for every band pair (ψ_1, ψ_2).

Thus, we carry out 135 experiments (i.e., 15 band pairs × 3 population sizes × 3 gene set sizes) to perform this analysis. We plot the resulting correlations as a heatmap in Figure 5.

While Figure 5 assesses the correlation in the top 100 gene histograms across bands, we also perform a quantitative evaluation of the similarity across bands. In order to do that, we identify the intersecting gene sets for each band pair (ψ_1, ψ_2) by applying $G_{\psi_1} \cap G_{\psi_2}$. The heatmap in Figure 6 shows the number of genes that are shared between bands in their top 100 genes, regardless of their occurrence counts. The purpose of this experiment is to understand whether the correlation values is affected by common gene selection.

3.2.3. Evaluation of interactions among selected genes

The final question we seek to answer is whether the genes selected by our IDEEA method are acting alone or they are interacting with each other. By answering this question, we aim to understand the functional connection among the identified genes. To do this, for each frequency band, we pick the top 100 genes identified by our IDEEA algorithm, and extract the protein-protein interaction (PPI) networks

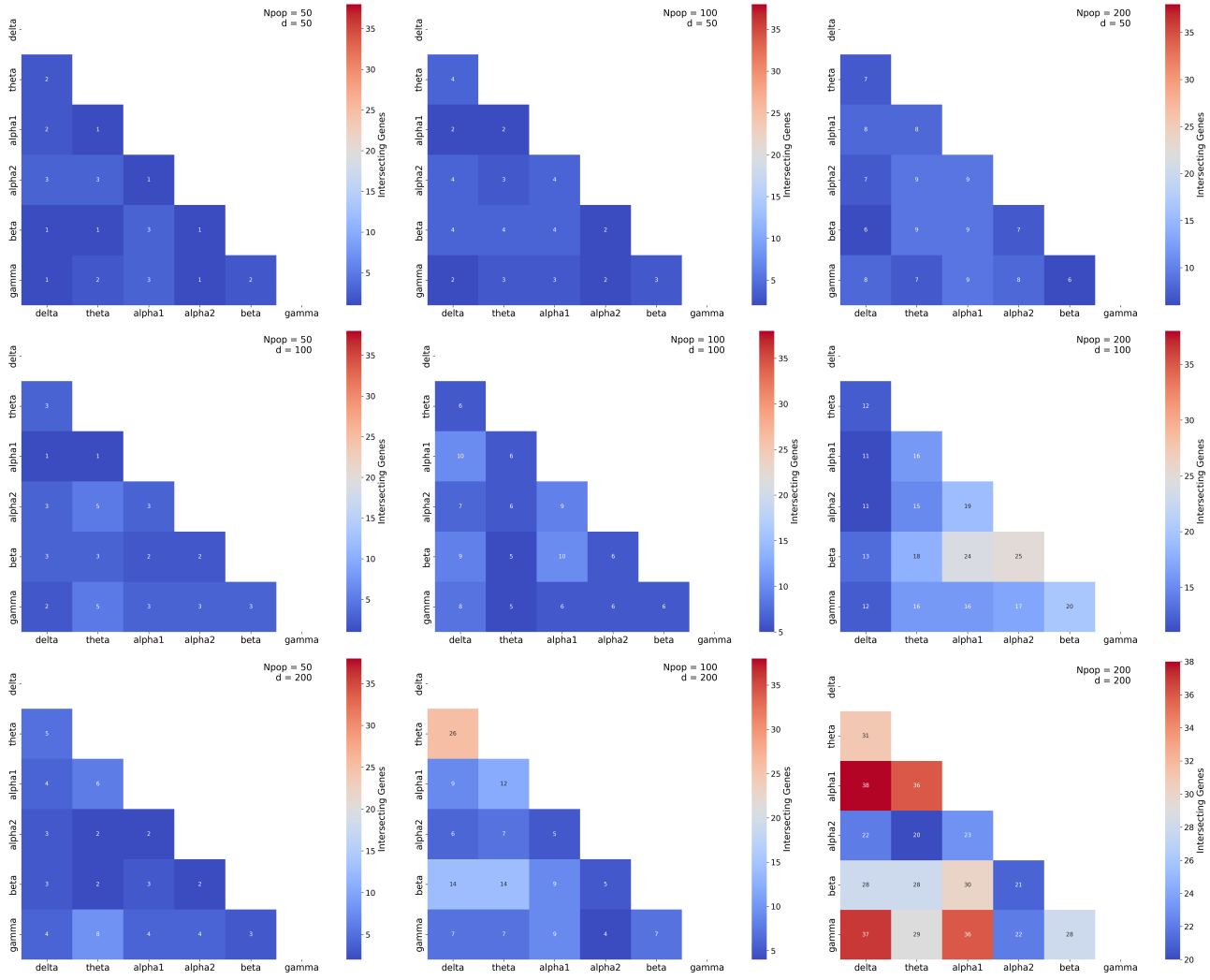

Figure 6: Heatmaps for common genes of 100 most frequently selected genes between bands with varying N_{pop} (50, 100, 200) and d (50, 100, 200) sizes.

induced on these genes using the STRING database [37] (see Figure 4). To simplify the figure, we show the connected components with more than 3 genes in Figure 7. After this filtering, the genes selected by our IDEEA method using the EEG signals form two connected components for the delta and alpha1 bands, and a single connected component for the remaining four bands theta, alpha2, beta, and gamma.

4. Discussion

AD is a complex disease that is extensively studied with no effective cure present. It is difficult to elucidate the genetic disease mechanism since AD is a neurodegenerative brain disease that may require invasive approaches to explore the genetic factors that induce neuropathological changes in the brain system. In this study, we build an alternative machine learning approach to study AD, without risky invasive brain tissue harvesting process. We introduced IDEEA algorithm to associate non-invasively acquired EEG signals with the genetic basis of AD using transcriptional data. In our algorithm we employed machine learning techniques to analyze the PSD differences between AD and healthy samples across six different frequency bands — delta, theta, alpha1, alpha2, beta, and gamma and identified the genes whose differential transcription patterns are reflected in differential EEG signals for each frequency band.

Impact of the gene set size. From Figure 2, we observe that as the number of selected genes d decreases, convergence speed and the best fitness score improves. We conjecture that this happens for three possible reasons. First, our algorithm is more effective in exploring smaller gene sets, and thus finds those gene sets whose transcription patterns have high correlation with the EEG patterns. Second, EEG patterns may decipher the transcription patterns of a limited set of genes. Third, only a limited number of genes are associated with the differential behavior of the transcription patterns of AD and CN groups, and thus imposing more genes into the subset introduces noise.

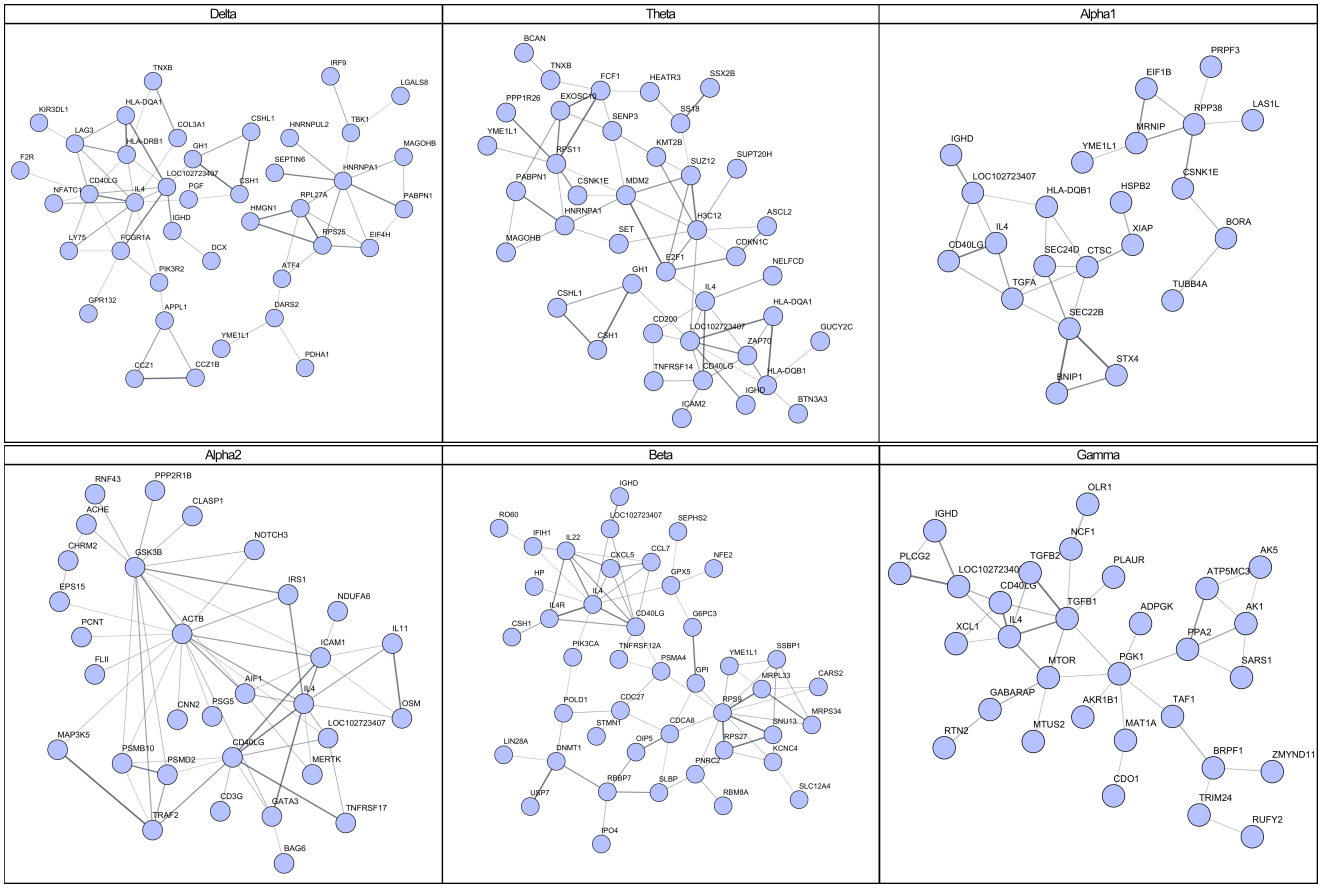

Figure 7: Protein interactions between 100 most frequently selected genes with $N_{pop} = 100$, $d = 200$ for different EEG bands.

Our results demonstrate that the IDEEA algorithm can always achieve high correlation values (i.e., over 0.75), particularly for small gene subsets. The correlation values are larger than 0.8 for 4/6 frequency bands, and larger than 0.9 for 2/6 bands (beta and gamma bands). In this context, while delta and theta bands show relatively slower convergence and lower fitness scores, especially when $d > 100$, this is not the case for beta and gamma bands. The faster convergence and higher fitness scores in beta and gamma bands, even with larger gene sets, essentially have to do with algorithm solving a simpler optimization problem due to less differentiation in these bands for Alzheimer's disease. In contrast, a higher fluctuation in spectral differences in delta, theta and alpha sub-bands suggest a more complex relationship between spectral patterns and gene expressions in these bands, potentially due to more significant impact of Alzheimer's disease on the cognitive states they represent. Overall, this implies the functional difference between these bands in finding a fitting transcriptomics profile to the spectral distances provided.

Impact of population size. Firstly, our results in Figure 3 suggest that our method can achieve high fitness values across all bands, which confirm our results above. We observe that higher population size provide a relative improvement in fitness convergence along with higher fitness scores. This is because, larger population sizes allows us to widen the exploration of key genes especially for the bands that are differentially impacted by Alzheimer's, as there would be more opportunities for the algorithm to evaluate diverse genetic combinations including potential biomarkers for Alzheimer's disease that are not as prominent by themselves. The gap between the performance of our method across different population sizes are however marginal, except for the alpha2 band. Thus if our algorithm is run long enough, it will converge to a similar local optimal value without necessitating very large populations. All in all, these findings further inform us regarding the optimization strategy for different spectral characteristics in our algorithm when applied to EEG data for the detection of Alzheimer's disease.

Evaluation of AD-related genes. From our results in Figure 4, we observe that alpha1, alpha2, and gamma yield a significantly high number of AD-associated genes among all bands. We show the statistical significance of these findings with hypergeometric test for each frequency band and provide p-values. We hypothesize that the frequency range of alpha2 (10-13 Hz) has a potential relation to neural activities or processes that are significantly impacted in Alzheimer's disease. The p-value for alpha2 band is 0.05, showing the

significance of this finding. Along with alpha2, alpha1 and gamma bands also result in relatively more AD-related genes compared to the other frequency bands.

Existing literature supports our results as the relationship between the alpha band with electrophysiological and neurophysiological processes is suggested in several studies as being crucial for understanding spectral responses to AD [12, 31, 35]. Studies have shown changes in the alpha power and coherence in AD, indicating a disruption in neural synchrony and connectivity [12, 35]. Moreover, the alpha2 band has been associated with distinct frequency features in AD, showing evidences of potential pathological changes [12, 31].

Evaluation of band-pair association. Our results in Figure 5 demonstrate that for all population sizes (N_{pop}) and number of selected genes (d), except for the largest values of N_{pop} and d, there is no significant correlation among different EEG bands. This indicates that our algorithm captures the differential transcription behavior of different sets of genes when it is fed with different EEG band values. This is promising as it implies that our method can help us mimic and monitor the transcriptional patterns of alternative gene sets at different frequencies, thus providing a rich set of observations. We observe some high positive correlation when we increase the number of selected genes and the population size of our algorithm to 200. Particularly (alpha1, gamma), (alpha1, delta) and (alpha1, theta) pairs show high correlation while the other pairs remain low.

The heatmaps in Figure 6 provide an alternative view of the association between frequency band pairs, as they show the number of genes that are commonly shared between bands in their top 100 genes, regardless of their occurrence counts. By cross-referencing the two heatmaps from these figures, we can discern whether a high correlation coincides with high intersection counts, which would suggest not only similar gene frequencies but also a substantial overlap in the specific set of genes that are considered important across those bands. In parallel to Figure 5, we observe a mostly similar trend for selected genes and the population size. There are also several nuances where we see that the correlation between certain bands are not necessarily accompanied by the relatively similar level of intersecting gene counts. This is particularly observed when we decrease the number of selected genes and the population size in the algorithm. As we decrease the values of these two parameters the number of genes shared across different band pairs dramatically drops (see the heatmaps in the first row and first column of Figure 6 in the 3×3 heatmap organization). While reduced number of common genes is expected with smaller gene set sizes, for the number of genes we select among 12,937 differentially expressed genes becomes as low as 50, we observe that this drop is affected more prevalent with respect to reduced population size. More specifically when we set the parameters as $d = 50$ and $N_{pop} = 200$, the total number of common genes across all band pairs is 117 (i.e., the sum of the values in the top-right heatmap in Figure 6). On the other hand, the same number when we set the parameters as $d = 200$ and $N_{pop} = 50$ is 55. This inidicates the success of our IDEEA algorithm in identifying gene sets when it uses large enough population to explore the search space. Ultimately, from both figures we can further confirm when we increase the selected genes and set the population size to 200, the highly correlated band pairs also show a similar increase in the number of intersecting genes. For $d = 200$ and $N_{pop} = 200$, we obtain the largest number of common gene sets for the band pairs (alpha1, delta), (gamma, delta), (gamma, alpha1), and (alpha1, theta) with more than 35 common genes among 200. Alpha1 band shares the most genes with any other band across all the frequency bands, with a total of 163 times a gene is shared. Alpha2 shares the least number of genes with other bands with a total of 108 genes.

Interaction topology of selected genes. From Figure 7, we observe that the genes we found are highly connected through known PPI relationships at all frequency bands. Furthermore, our method was able to select hubs (i.e., highly connected genes) at each frequency band as well. Among them, some of the notable ones are MDM2 in theta band (top-middle in Figure 7), whose inhibition reduces the neurogenic defects, such as AD [16]. Another one is IL4 which appears in all six bands delta, theta, alpa1, alpha2, beta, and gamma. The stimulation of IL4 increases the proportion of oligodendrocytes and neurons, thereby having positive effects on cognition [11, 30]. TGF in alpha1 band (top-right figure) and gamma band (bottom-right figure) affects the mediation of microglia, which are resident macrophages in brain thus altering the neurodegenrative disease promotion, such as AD [25].

The networks acquired in delta and theta bands resulted in statistically significant PPI enrichment p-values, 0.01 and 0.03 respectively. This suggests that these genes do not interact randomly but are part of a larger, interconnected network. Functional enrichment analysis within these networks indicates several prominent pathways. This includes the pathways related to immune response (e.g., T cell receptor signaling pathway, FDR=0.0447). This connection is particularly relevant as neuroinflammation has a prominent role in the pathogenesis of AD as recently suggested by the studies [19, 26, 10].

Limitations and opportunities. Main challenges in associating EEG readings with gene transcription value for AD patients arise from the nature of the AD disease that collecting tissue samples is a challenge as it is a risky and invasive procedure. Typically, tissue samples are collected from deceased patients. There can be variety in the brain regions where the tissue samples are collected, which make the dataset less homogeneous. Furthermore, the transcription patterns of more than one brain region may be affected in some AD patients. These factors may negatively influence machine learning strategies as data is diverse. Our results already demonstrate that our method can associate EEG

readings with known AD genes. We conjecture that, more transcription data availability will further improve the accuracy of our IDEEA method.

Another limitation arise from the EEG readings. Some of the electrodes may read EEG values from multiple brain regions (see Table 2). As a result, the EEG values from some electrodes are aggregate readings from multiple brain regions rather than those focused on specific regions. In this study, we use a linear aggregate model to describe such relationships. In our model, we use equal weight to the regions listed in Table 2 for the electrodes which map to multiple brain regions if we do not have established weights in the literature. This brings an opportunity for us the train a machine learning model to compute the contribution of different brain regions for the EEG readings of each electrode. Since this problem deviates from the central hypothesis of this study, in order to maintain the focus of this paper, we defer that as a separate future work. We conjecture that deep learning methods integrated with our IDEEA method will address this important challenge.

In summary, across the spectral plane, our algorithm identified genes whose transcriptional profiles correlate with the EEG patterns observed in AD patients compared to healthy individuals. The genes determined through our algorithm, particularly in the alpha2 frequency range, included a significant number of AD-associated genes. The statistical significance of these findings was supported by hypergeometric testing, suggesting that certain frequency bands are more closely associated with neural activities disrupted by AD. Protein-protein interaction networks further confirmed the biological relevance of the selected genes through the significant PPI enrichment p-values observed in the delta and theta bands. In addition, functional enrichment in immune response pathways supported the hypothesis that the identified genes are biologically connected and play a role in the pathogenesis of AD.

Authorship contribution statement

Enes Ozelbas: Data collection, Literature study, Methodology development, Implementation, Experimentation, Writing. **Tuba Sevimoglu:** Literature study, Data collection, Supervision, Methodology, Formal analysis, Writing. **Tamer Kahveci:** Algorithm development, Experimental set up, Evaluation, Writing, Supervision.

Declaration of competing interest

The authors have no competing financial interests or personal relationships that could influence this study.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] WHO World Health Organization. https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed: 2023-08-24.

[2] 2023 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19, 2023.

[3] Murtala Bello Abubakar, Kamaldeen Olalekan Sanusi, Azizah Ugusman, Wael M. Y. Mohamed, Haziq Kamal, Nurul Husna Ibrahim, Ching Soong Khoo, and Jaya Kumar. Alzheimer's disease: An update and insights into pathophysiology. *Frontiers in Aging Neuroscience*, 14, 2022.

[4] Hind Alamro, Maha A. Thafar, Somayah Albaradei, Takashi Gojobori, Magbubah Essack, and Xin Gao. Exploiting machine learning models to identify novel alzheimer's disease biomarkers and potential targets. *Scientific Reports*, 13, 2023.

[5] Jesús Andrade-Guerrero, Alberto Santiago-Balmaseda, Paola Jeronimo-Aguilar, Isaac Vargas-Rodríguez, Ana Ruth Cadena-Suárez, Carlos Sánchez-Garibay, Glustein Pozo-Molina, Claudia Fabiola Méndez-Catalá, María del Carmen Cárdenas-Aguayo, Sofía Díaz-Cintra, Mar Pacheco-Herrero, José Luna-Muñoz, and Luis O. Soto-Rojas. Alzheimer's disease: An updated overview of its genetics. *International Journal of Molecular Sciences*, 24, 2023.

[6] Shea J. Andrews, Alan E. Renton, Brian Fulton-Howard, Anna Podleśny-Drabiniok, Edoardo Marcora, and Alison M. Goate. The complex genetic architecture of alzheimer's disease: novel insights and future directions. *eBioMedicine*, 90, 2023.

[7] Charles Arber, Christopher Lovejoy, and Selina Wray. Stem cell models of alzheimer's disease: progress and challenges. *Alzheimer's Research & Therapy*, 9, 2017.

[8] Aurina Arnatkeviciute, Ben D. Fulcher, Mark A. Bellgrove, and Alex Fornito. Imaging transcriptomics of brain disorders. *Biological Psychiatry Global Open Science*, 2:319 – 331, 2021.

[9] Eva Bagyinszky, Vo Van Giau, and SeongSoo A. An. Transcriptomics in alzheimer's disease: Aspects and challenges. *International Journal of Molecular Sciences*, 21, 2020.

[10] Brianne M. Bettcher, Malú G. Tansey, Guillaume Dorothée, and Michael T. Heneka. Peripheral and central immune system crosstalk in alzheimer disease — a research prospectus. *Nature Reviews Neurology*, 17(11):689–701, September 2021.

[11] Virginia Boccardi, Eric Westman, Luca Pelini, Olof Lindberg, J.-Sebastian Muehlboeck, Andrew Simmons, Roberto Tarducci, Piero Floridi, Pietro Chiarini, Hilkka Soininen, Iwona Kloszewska, Magda Tsolaki, Bruno Vellas, Christian Spenger, Lars-Olof Wahlund, Simon Lovestone, and Patrizia Mecocci. Differential associations of il-4 with hippocampal subfields in mild cognitive impairment and alzheimer's disease. *Frontiers in Aging Neuroscience*, 10, 2019.

[12] Giordano Cecchetti, Federica Agosta, Silvia Basaia, Camilla Cividini, Marco Cursi, Roberto Santangelo, Francesca Caso, Fabio Minicucci, Giuseppe Magnani, and Massimo Filippi. Resting-state electroencephalographic biomarkers of alzheimer's disease. *NeuroImage: Clinical*, 31:102711, 2021.

[13] Nicholas Chedid, Judie Tabbal, Aya Kabbara, Sahar Allouch, and Mahmoud Hassan. The development of an automated machine learning pipeline for the detection of alzheimer's disease. *Scientific Reports*, 12(1), October 2022.

[14] Eleanor Drummond and Thomas M. Wisniewski. Alzheimer's disease: experimental models and reality. *Acta Neuropathologica*, 133:155–175, 2017.

[15] Mr Amir Ebrahimighahnavieh, Suhuai Luo, and Raymond Chiong. Deep learning to detect alzheimer's disease from neuroimaging: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 187:105242, April 2020.

[16] Vita M. Golubovskaya, Baotran Ho, Mingzhong Zheng, Andrew T. Magis, David A. Ostrov, and William Cance. Mitoxantrone targets the atp-binding site of fak, binds the fak kinase domain and decreases fak, pyk-2, c-src, and igf-1r in vitro kinase activities. *Anti-cancer agents in medicinal chemistry*, 13 4:546–54, 2013.

[17] Boris Guennewig, Julia Lim, Lee L. Marshall, Andrew N. McCorkindale, Patrick Jarmo Paasila, Ellis Patrick, Jillian J. Kril, Glenda M. Halliday, Antony A. Cooper, and Greg Trevor Sutherland. Defining early changes in alzheimer's disease from rna sequencing of brain

regions differentially affected by pathology. *Scientific Reports*, 11, 2021.

[18] Huiwen Gui, Qi Gong, Jun Jiang, Mei Liu, and Huanyin Li. Identification of the hub genes in alzheimer's disease. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

[19] Michael T Heneka, Monica J Carson, Joseph El Khoury, Gary E Landreth, Frederic Brosseron, Douglas L Feinstein, Andreas H Jacobs, Tony Wyss-Coray, Javier Vitorica, Richard M Ransohoff, Karl Herrup, Sally A Frautschy, Bente Finsen, Guy C Brown, Alexei Verkhratsky, Koji Yamanaka, Jari Koistinaho, Eicke Latz, Annett Halle, Gabor C Petzold, Terrence Town, Dave Morgan, Mari L Shinohara, V Hugh Perry, Clive Holmes, Nicolas G Bazan, David J Brooks, Stéphane Hunot, Bertrand Joseph, Nikolaus Deigendesch, Olga Garaschuk, Erik Boddeke, Charles A Dinarello, John C Breitner, Greg M Cole, Douglas T Golenbock, and Markus P Kummer. Neuroinflammation in alzheimer's disease. *The Lancet Neurology*, 14(4):388–405, April 2015.

[20] 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 : JAD*, 2023.

[21] Nesma Houmani, François B. Vialatte, Esteve Gallego-Jutglà, Gérard Dreyfus, Vi-Huong Nguyen-Michel, Jean Mariani, and Kiyoka Kinugawa. Diagnosis of alzheimer's disease with electroencephalography in a differential framework. *PLoS ONE*, 13, 2018.

[22] Abid Hussain and Yousaf Shad Muhammad. Trade-off between exploration and exploitation with genetic algorithm using a novel selection operator. *Complex & Intelligent Systems*, 6(1):1–14, April 2019.

[23] Laura Ibañez, Carlos Cruchaga, and Maria Victoria Fernández. Advances in genetic and molecular understanding of alzheimer's disease. *Genes*, 12, 2021.

[24] Bin Jiao, Rihui Li, Hui Zhou, Kunqiang Qing, Hui Liu, Hefu Pan, Yanqin Lei, Wenjin Fu, Xiaoan Wang, Xue wen Xiao, Xi xi Liu, Qijie Yang, Xinxin Liao, Yafang Zhou, Liangjuan Fang, Yanbin Dong, Yuanhao Yang, Haiyan Jiang, Shan Huang, and Lu Shen. Neural biomarker diagnosis and prediction to mild cognitive impairment and alzheimer's disease using eeg technology. *Alzheimer's Research & Therapy*, 15, 2023.

[25] Mahima Kapoor and Subashchandrabose Chinnathambi. Tgf-ß1 signalling in alzheimer's pathology and cytoskeletal reorganization: a specialized tau perspective. *Journal of Neuroinflammation*, 20(1), March 2023.

[26] Fangda Leng and Paul Edison. Neuroinflammation and microglial activation in alzheimer disease: where do we go from here? *Nature Reviews Neurology*, 17(3):157–172, December 2020.

[27] Marco Magistri, Dmitry Velmeshev, Madina Makhmutova, and Mohammad Ali Faghihi. Transcriptomics profiling of alzheimer's disease reveal neurovascular defects, altered amyloid-β homeostasis, and deregulated expression of long noncoding rnas. *Journal of Alzheimer's Disease*, 48:647 – 665, 2015.

[28] Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Evripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. "a dataset of 88 eeg recordings from: Alzheimer's disease, frontotemporal dementia and healthy subjects", 2023.

[29] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 11 2019.

[30] Hongjian Pu, Yangfan Wang, Tuo Yang, Rehana K. Leak, R. Anne Stetler, Fang Yu, Wenting Zhang, Yejie Shi, Xiaoming Hu, Ke-jie Yin, T. Kevin Hitchens, C. Edward Dixon, Michael V.L. Bennett, and Jun Chen. Interleukin-4 mitigates anxiety-like behavior and loss of neurons and fiber tracts in limbic structures in a microglial ppar-dependent manner after traumatic brain injury. *Neurobiology of Disease*, 180:106078, May 2023.

[31] Elias Mazrooei Rad, Mahdi Azarnoosh, Majid Ghoshuni, and Mohammad Mahdi Khalilzadeh. Diagnosis of mild alzheimer's disease by EEG and ERP signals using linear and nonlinear classifiers. *Biomedical Signal Processing and Control*, 70:103049, sep 2021.

[32] Yelluru Lakshmisha Rao, B. Ganaraja, Bukkambudhi V. Murlimanju, Teresa Joy, Ashwin Krishnamurthy, and Amit Agrawal. Hippocampus and its involvement in alzheimer's disease: a review. *3 Biotech*, 12, 2022.

[33] Seyed-Ali Sadegh-Zadeh, Elham Fakhri, Mahboobe Bahrami, Elnaz Bagheri, Razieh Khamsehashari, Maryam Noroozian, and Amir M. Hajiyavand. An approach toward artificial intelligence alzheimer's disease diagnosis using brain signals. *Diagnostics*, 13(3):477, January 2023.

[34] Catriona L. Scrivener and Arran T. Reader. Variability of EEG electrode positions and their underlying brain regions: visualizing gel artifacts from a simultaneous EEG-fMRI dataset. *Brain and Behavior*, 12(2), jan 2022.

[35] Chanda Simfukwe, Su-Hyun Han, Ho Tae Jeong, and Young Youn. qeeg as biomarker for alzheimer's disease: Investigating relative psd difference and coherence analysis. *Neuropsychiatric Disease and Treatment*, Volume 19:2423–2437, November 2023.

[36] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, Sergey Kaplan, Dvir Dahary, David Warshawsky, Yaron Guan-Golan, Asher Kohn, Noa Rappaport, Marilyn Safran, and Doron Lancet. The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, 54(1), jun 2016.

[37] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, oct 2014.

[38] Minghui Wang, Panos Roussos, Andrew McKenzie, Xianxiao Zhou, Yuji Kajiwara, Kristen J. Brennand, Gabriele C. De Luca, John F. Crary, Patrizia Casaccia, Joseph D. Buxbaum, Michelle Ehrlich, Sam Gandy, Alison Goate, Pavel Katsel, Eric Schadt, Vahram Haroutunian, and Bin Zhang. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to alzheimer's disease. *Genome Medicine*, 8(1), nov 2016.

[39] P. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

[40] Jianfeng Wu, Yanxi Chen, Panwen Wang, Richard J. Caselli, Paul M. Thompson, Junwen Wang, and Yalin Wang. Integrating transcriptomics, genomics, and imaging in alzheimer's disease: A federated model. *Frontiers in radiology*, 1, 2021.