Comparative Evaluation of Feature Extraction Methods for GWAS-Based Alzheimer's Disease Classification

Bruna B. F. de Carvalho[1], **Arysson S. de Oliveira**[2],
Rodrigo F. A. P. de Oliveira[1], **Diego de F. Bezerra**[1]

[1]CESAR School

[2]Centro de Estudos e Sistemas Avançados do Recife (CESAR)

{bbfc,rfapo,dfb2}@cesar.school, aso@cesar.org.br

Abstract. *Genome-Wide Association Studies (GWAS) generate high-dimensional genetic data that challenge machine learning models and increase susceptibility to overfitting. This study compares linear, nonlinear, and deep learning-based feature extraction methods for Alzheimer's disease classification using ADNI data. Latent representations combined with demographic risk factors were evaluated via a Random Forest classifier under cross-validation and independent test subsets. Results show that the non-linear Diffusion Maps technique achieved the highest dimensionality reduction while maintaining predictive performance comparable to the baseline model, preserving the essential predictive variance of the original SNP matrix.*

1. Introduction

Alzheimer's Disease (AD) is the leading cause of dementia, accounting for 60-80% of all cases and posing a critical socioeconomic burden for modern societies [Al-Shoukry et al. 2020]. As a progressive brain disorder that limits daily autonomy [Smid et al. 2022], AD has become the sixth leading cause of death in the United States, affecting approximately 6 million people in 2021 with projections reaching 14 million by 2060 [Malik et al. 2024]. This growing global epidemic demands advanced computational approaches to improve early stratification and understand its complex pathophysiology.

Due to AD's strong hereditary contribution [Saykin et al. 2015] and polygenic nature, where biological signals emerge from combined genetic interactions rather than isolated genes [Fridman et al. 2004], genomic datasets typically feature millions of genetic variants per participant. This severe imbalance relative to the sample size triggers the curse of dimensionality, reducing machine learning performance and increasing overfitting risks [Alatrany et al. 2022]. Consequently, dimensionality reduction is vital for computational feasibility and mitigating overfitting, though it requires rigorous architectural choices to avoid discarding low-effect variants essential to the complex biological signature of AD [Khanal et al. 2021].

To address these high-dimensional challenges and mitigate computational redundancy, dimensionality reduction stands as an indispensable step for efficient machine learning deployment [Borah et al. 2024]. This process typically relies on two distinct paradigms: Feature Selection (FS) and Feature Extraction (FE) [Borah et al. 2024, Berahmand et al. 2024]. While FS filters the dataset to retain only the most informative

original variables-preserving interpretability [Labory et al. 2024]-, FE maps the data onto a reduced mathematical space, deriving new latent variables through linear or non-linear combinations of the initial characteristics [Pölsterl et al. 2016]. Within the FE domain, approaches are broadly categorized into classical statistical methods- divided into linear and non-linear - and deep learning frameworks, which iteratively adjust latent boundaries through sequential layered processing [Berahmand et al. 2024]. In genomic studies, these latent representations function as condensed mathematical summaries capable of capturing hidden biological variation within a significantly lower dimension [Borah et al. 2024].

Given the complexity and high dimensionality of AD genetic data, this study compares linear, nonlinear, and deep learning-based FE methods for AD classification.The goal is to investigate how different latent representations affect predictive performance, assessing whether dimensionality reduction can preserve the discriminative information contained in the original GWAS data relative to a baseline model trained on the full feature space.

This work is organized as follows: Section 2 presents related works. Section 3 discusses the dataset, preprocessing procedures, and the experimental methodology used. Section 4 summarizes and discusses the obtained results. Finally, Section 5 presents the work's conclusions, limitations, and potential directions for future research.

2. Related Work

Although dimensionality reduction is widely used in bioinformatics, the literature on AD classification from GWAS data remains heavily focused on FS strategies. A relevant example is the work of [Khanal et al. 2021], who presented a two-step FS pipeline applied to genetic data, producing a curated list of important genetic markers for AD classification. However, the authors relied exclusively on FS techniques, without investigating whether FE approaches could recover additional discriminative signal from the same data. Therefore, their study was adopted as the main experimental benchmark for evaluating the improvements achieved through FE approaches.

A systematic review by [Dehbozorgi et al. 2025] compared statistical, radiomics, and DL feature extraction techniques across multiple medical imaging modalities, showing that the application of FE techniques in AD research is predominantly concentrated on unstructured data, particularly medical imaging, while remaining largely unexplored in structured genetic matrices such as GWAS data. Another significant limitation is the fragmentation of experimental results. [Labory et al. 2024] evaluated linear and nonlinear dimensionality reduction methods on metabolomics datasets and found that combining FS and FE improved patient classification performance. However, DL-based architectures were not considered in the same study.

Additionally, the review by [Malik et al. 2024], which surveyed DL architectures applied to AD prediction across a wide range of imaging modalities, highlights the widespread adoption of end-to-end DL architectures, in which the FE and classifier are trained jointly. Although effective, such approaches make it difficult to independently evaluate the quality of the learned representations and may increase the risk of overfitting in high-dimensional genomic settings. Finally, manifold learning methods such as Diffusion Maps remain underexplored in GWAS applications when compared to more established approaches such as Kernel PCA, Sparse PCA, and Autoencoders.

To address these gaps, this work performs a comparison of linear, nonlinear, and DL-based FE methods under the same experimental protocol and using the same genetic input data, including approaches that remain relatively unexplored in GWAS studies, such Diffusion Maps and Variational Autoencoders.

3. Data and Method

This study utilized the Alzheimer's Disease Neuroimaging Initiative (ADNI) database to extract all required data files. As illustrated in the methodology diagram Figure 1, the workflow of this study is structured into two main stages: (i) pre-processing, which encompasses cohort selection, phenotype extraction, demographic covariate definition, and genomic data quality control; and (ii) experimental execution and evaluation, utilizing FE algorithms combined with a Random Forest classifier. Each stage is detailed in the following subsections.

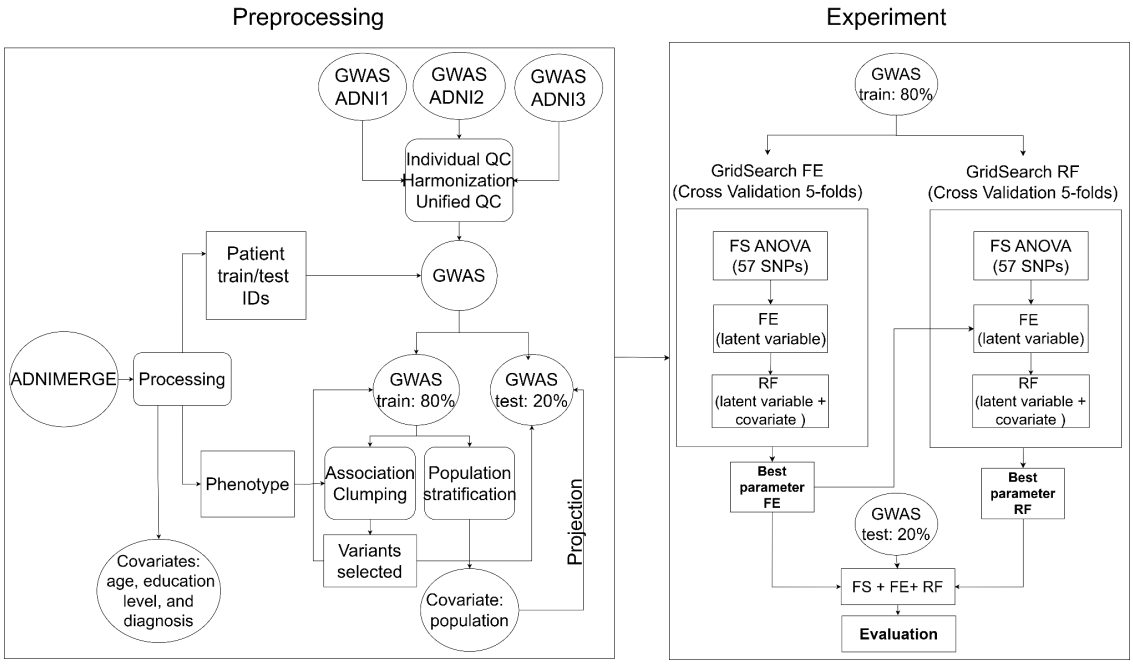


Figure 1. Methodological workflow of the study, detailing the data preprocessing pipeline (QC, GWAS association, clumping, and population stratification check) and the downstream experimental design using nested grid search optimization.

3.1. Data and preprocessing

The study used the *ADNIMERGE* dataset, which integrates longitudinal clinical data from ADNI. To fit the analysis into a cross-sectional framework, only the most recent record of each participant was retained. A cascading filtering strategy was applied: first, the last active phase for each participant was selected, and then the most recent visit within that phase was retained. This approach captures the most advanced and up-to-date clinical status of a progressive neurodegenerative disease [Scheltens et al. 2021]. To ensure a robust binary classification task and reduce noise during model training, intermediate or ambiguous diagnostic categories were excluded, retaining only AD and Cognitively

Normal (CN) participants. Due to this study's cross-sectional design, transitional phenotypes were excluded, as their validation requires longitudinal tracking. Additionally, clinical diagnoses often introduce label noise that can impair machine learning training [Alatrany et al. 2022]. After defining the final cohort, the dataset was split into 80% training and 20% testing subsets. This partition generated participant identification lists for each subset, the phenotype file required for GWAS analyses, and the corresponding feature matrices.

Additionally, a rigorous quality control (QC) procedure was applied to the genetic data from GWAS containing Single Nucleotide Polymorphisms (SNPs) - which represent single-base variations in the DNA sequence commonly utilized as genetic markers - extracted from the three study phases: ADNI1, ADNI2/GO, and ADNI3. This step was performed using PLINK - a specialized bioinformatics software for high-throughput genomic data processing and GWAS analytics - following the protocol described by [Marees et al. 2018], which documents the standard GWAS quality control procedures and corresponding commands. The procedure was applied independently to each ADNI phase to avoid batch effects, since different genotyping platforms were used across phases (*Illumina 610-Quad*, *OmniExpress*, and *GSA2*, respectively) [Squillario et al. 2020, Nasr et al. 2025]. To ensure statistical consistency across datasets generated by different genotyping arrays, genomic harmonization was conducted in two stages: (1) coordinate conversion of ADNI1 data from *NCBI36/hg18* to *GRCh37/hg19* using ADNI2 as positional reference; and (2) strand and allele alignment, using ADNI1 as the reference dataset for standardizing major and minor allele assignments. Variants with incompatibilities or missing information across phases were excluded. After merging the datasets, the analysis was restricted to autosomal chromosomes due to quality issues in sex chromosomes, resulting in an intersection of 72,867 common SNPs. After dataset integration, additional quality control procedures requiring the complete cohort were applied, including minor allele frequency filtering and relatedness removal to prevent confounding effects during association analyses [Marees et al. 2018], resulting in 62,129 variants.

The genomic dataset was then divided into training and testing subsets using the participant lists previously extracted from ADNIMERGE, preventing data leakage during the final quality control stages. Association analysis and extraction of ten principal components for population stratification correction were performed exclusively on the training set. The resulting ancestry components were projected onto the test set using singular value decomposition with dot-product alignment to resolve sign ambiguities [Bro et al. 2008]. Finally, to reduce statistical redundancy caused by linkage disequilibrium, the clumping procedure was applied to the training set [Marees et al. 2018], selecting the most significant SNP from each correlated block and reducing the feature space to 603 variants.

3.2. Experiments and Evaluation

After preprocessing, the final genomic matrix contained 603 genetic variants from 1,383 unique individuals. The training set consisted of 1,106 participants (561 AD cases and 545 CN controls), including 573 males and 533 females. This subset achieved a call rate of 0.9984, indicating that more than 99.84% of allele dosage values were successfully retained without missing data, demonstrating the effectiveness of the quality control

pipeline. The test set comprised the remaining 277 participants (141 AD and 136 CN), including 148 males and 129 females, with a similar call rate of 0.9980. Notably, this class parity was achieved naturally as a direct result of the cohort definition and quality control pipeline, without the need for any algorithmic balancing or resampling techniques. In addition to genetic data, age, years of education, and the previously computed population stratification components were incorporated as predictors. Age and education are well-established demographic risk factors that improve classification boundaries in AD modeling [Khanal et al. 2021], while the population components ensure that baseline geographical variations are controlled during the downstream representation learning.

For comparative analysis, a Random Forest (RF) model trained directly on the raw genotype data without dimensionality reduction was established as the baseline. In addition to being a widely adopted traditional machine learning algorithm in GWAS classification studies, RF offers a favorable balance between predictive performance, robustness, interpretability, and computational cost when handling high-dimensional genetic data. Although literature-such as the study by [Khanal et al. 2021]-highlights both XGBoost and RF as leading algorithms for similar tasks, preliminary experiments conducted on our specific dataset demonstrated that RF achieved more stable predictive behavior and superior performance metrics than boosting-based alternatives. Therefore, RF was selected as the common downstream classifier to ensure a consistent and computationally efficient comparison across all feature extraction workflows. The raw genotype data were encoded using an additive inheritance model (0, 1, or 2 copies of the alternative allele) and subsequently standardized using *StandardScaler*. To evaluate the impact of latent representations against the baseline, a standardized experimental pipeline was established for the alternative workflows using a stratified 5-fold cross-validation framework combined with Grid Search for hyperparameter optimization. To guarantee a rigorous evaluation, prevent data leakage, and mitigate overfitting [Alatrany et al. 2022], FS was embedded directly within this cross-validation loop. Specifically, for each split, FS was performed strictly on the active training folds using *SelectKBest* with the ANOVA scoring function-a widely adopted and validated approach for genetic data screening in bioinformatics [Pölsterl et al. 2016]. This step filtered the high-dimensional space down to the top 57 genetic predictors. This strategy aligns with the findings of [Najar et al. 2021]. The choice of retaining 57 SNPs was guided by the findings of [Najar et al. 2021], who demonstrated that polygenic profiles constructed from this specific number of variants were associated with a clinical hazard ratio for dementia onset. Furthermore, given the available cohort size, limiting the feature space to 57 SNPs helped maintain an appropriate sample-to-feature ratio, reducing dimensionality-related issues and minimizing the risk of overfitting when integrating genetic, demographic, and population stratification predictors. Following this feature reduction, the optimization and classification workflow followed a nested, sequential design across all workflows. In the first stage, a Grid Search was applied to each of the six FE techniques within the active training folds. This step optimized each method's intrinsic parameters to compress the 57 selected markers into the most representative latent spaces. In the second stage, a subsequent Grid Search was executed to train and tune the final RF classifier using these optimized latent features concatenated with the standardized demographic and population covariates. The complete hyperparameter search spaces for both stages are detailed in Table 1. Finally, the optimal configurations selected from this nested pipeline were frozen and evaluated on the

completely isolated testing subset along with the demographic predictors to validate the model's generalizability and ultimate predictive performance.

The evaluated FE techniques were selected to explore diverse mathematical assumptions and structural data geometries within the GWAS dataset, with their formal mathematical foundations and theoretical formulations detailed in the respective original or foundational references cited below:

Linear Methods: Sparse PCA (SPCA) and Non-Negative Matrix Factorization (NMF) were implemented to model linear genomic variance. SPCA was evaluated for its potential to filter out background genomic noise by concentrating variance onto fewer components, isolating targeted blocks of correlated SNPs [Zou et al. 2006, Berahmand et al. 2024]. Conversely, NMF operates under non-negativity constraints to decompose the genomic matrix into positive components that combine additively. This structural design is hypothesized to mirror the multi-hit, cumulative genetic risk profile of AD by isolating additive mutational signatures that reflect cumulative genetic burden in an interpretable manner [Lee and Seung 1999, Berahmand et al. 2024].

Nonlinear Methods: Kernel PCA (KPCA) and Diffusion Maps were employed to map non-linear topological configurations. KPCA was evaluated for its potential to capture epistasis-where the phenotypic effect of a SNP depends on alternative modifiers-potentially uncovering complex genetic interactions that linear models overlook [Labory et al. 2024, Schölkopf et al. 1998, Borah et al. 2024].Diffusion Maps construct a graph-based probabilistic framework to measure geometric connectivity between genetic profiles, exploring whether clustering individuals based on localized graph alignment can reveal structural patterns associated with the target cohorts [Coifman and Lafon 2006].

Deep Learning Methods: Deterministic Autoencoders (AE) and Variational Autoencoders (VAE) were designed as neural network architectures, restricted to a single hidden layer to mitigate overfitting under sample constraints. The AE compresses features by minimizing reconstruction error, acting as a structural filter against statistical noise and variant redundancies [Berahmand et al. 2024]. The VAE replaces this rigid bottleneck with a probabilistic framework, encoding genomic data into a distribution optimized via Kullback-Leibler divergence. This technique evaluates whether constructing a continuous, regularized latent space can successfully map the subtle biological variations separating the cohorts [Kingma and Welling 2014].

The dimensionality reduction details, including input dimensions, latent space sizes, and corresponding compression ratios for each workflow, are summarized in Table 3. Predictive performance across all workflows was evaluated using Area Under the ROC Curve (AUC-ROC) as the primary discriminative metric, given its reliability in AD classification [Alatrany et al. 2022]. Additionally, Accuracy, Precision, Recall, and F1-Score were computed to provide a comprehensive analysis of classifier behavior. To assess stability and generalizability, 95% confidence intervals were calculated over the cross-validation folds, with the calculation based on Student's t-distribution utilizing the standard error of the mean and the appropriate degrees of freedom for the sample size $(n - 1)$, followed by final evaluation on the completely isolated testing subset.

4. Results and Discussion

The statistical validity of the initial association screening was evaluated via Manhattan and Quantile-Quantile (Q-Q) plots (Figures 2 and 3). The distributed landscape in the Manhattan plot confirms the polygenic nature of AD, characterized by multiple sub-threshold genetic variations rather than isolated major peaks. Concurrently, the tight adherence of data points to the diagonal identity line in the Q-Q plot demonstrates that population stratification and genomic inflation were successfully controlled, while the sharp upward deviation at the upper-right tail confirms a true biological association signal.

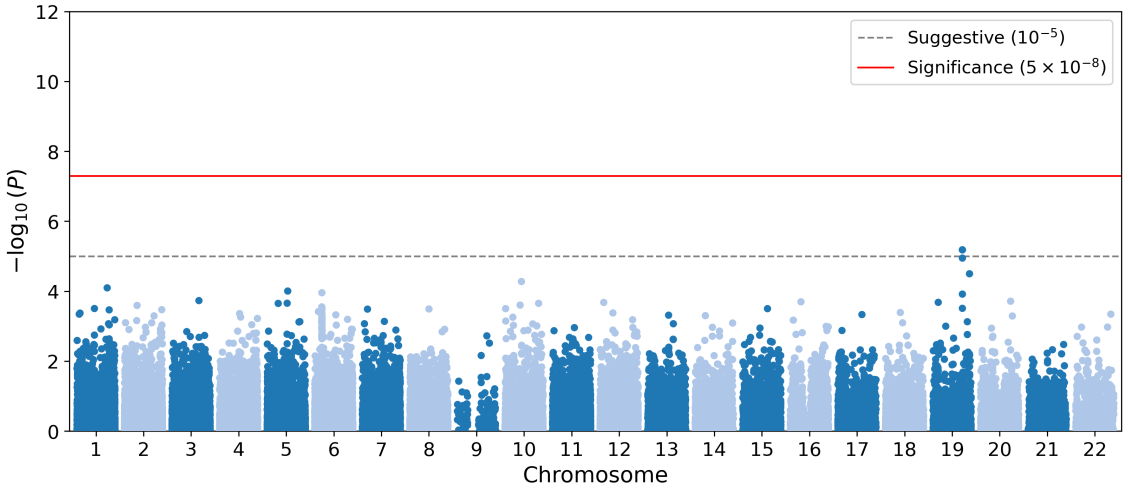


Figure 2. Manhattan plot showing the genetic association results across chromosomes. Individual points represent variants, where higher points indicate stronger association with Alzheimer's disease.

Grid Search optimization Table 1 revealed distinct geometric behaviors. While SPCA, KPCA, NMF, and the neural architectures required moderately sized latent spaces ($d = 25$) and specific inductive biases (e.g., sparsity penalties, non-negative constraints, or SELU activations) to stabilize boundaries, Diffusion Maps efficiently mapped the underlying topological connectivity onto an compact manifold ($d = 5, k = 10$).

The complete predictive performance metrics obtained across both the cross-validation folds and the independent testing subset are detailed in Table 2. As shown by the results, with the exception of the VAE, the adopted experimental pipeline achieved higher predictive performance than the benchmark reported by [Khanal et al. 2021], which achieved a maximum AUC of 0.64 using ADNI GWAS data and demographic covariates without FE.

Table 3 summarizes the dimensionality reduction details, including input dimensions, latent space sizes, and corresponding compression ratios for each workflow. Although the baseline model outperformed the other approaches in absolute AUC-ROC terms, Diffusion Maps utilized only 8.7% of the original data (reducing from 57 to 5 latent features) and achieved a test AUC-ROC of 0.70 (cross-validation AUC-ROC of 0.71 ± 0.03). This small difference of only 0.01 compared to the baseline indicates that the method preserved nearly all essential predictive variance after dimensionality reduction, demonstrating efficiency in separating patient classes while operating under a high compression ratio of 11.4.

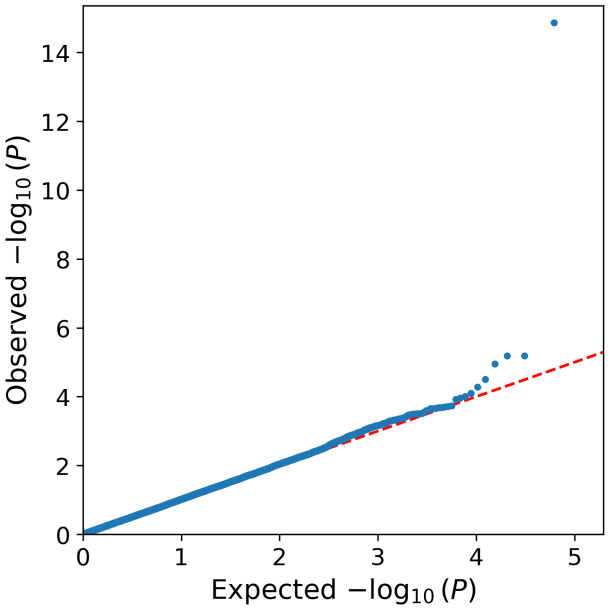


Figure 3. Quantile-Quantile (Q-Q) plot evaluating data quality. Points aligned with the diagonal line confirm that background noise was controlled, while the deviation at the upper tail shows true genetic signals.

Table 1. Hyperparameter search spaces and corresponding optimal configurations determined via grid search across all evaluated models.

Algorithm	Hyperparameter	Evaluated Range	Optimal
SPCA	Sparsity penalty (α)	[0.1, 0.5, 1.0]	0.5
	Latent dimensions (d)	[5, 10, 15, 20, 25]	25
NMF	Initialization method	['nndsvd', 'nndsvda', 'random']	nndsvd
	Latent dimensions (d)	[5, 10, 15, 20, 25]	25
KPCA	Kernel function	['cosine', 'rbf', 'poly']	cosine
	Latent dimensions (d)	[5, 10, 15, 20, 25]	25
Diffusion Maps	Number of neighbors (k)	[10, 20, 30]	10
	Latent dimensions (d)	[5, 10, 15, 20, 25]	5
Autoencoder (AE)	Hidden layer size	[32, 64]	64
	Activation function	['SELU', 'ReLU']	SELU
	Learning rate (η)	[0.0001, 0.0005, 0.001]	0.001
	Latent dimensions (d)	[5, 10, 15, 20, 25]	25
VAE	Hidden layer size	[32, 64]	32
	Activation function	['SELU', 'ReLU']	SELU
	KL divergence weight (β)	[0.01, 0.05]	0.01
	Learning rate (η)	[0.0001, 0.0005, 0.001]	0.0001
	Latent dimensions (d)	[5, 10, 15, 20, 25]	25
RF	Number of trees ($n_{\text{estimators}}$)	[1200, 1500]	1200
	Maximum depth (max_{depth})	[6, 8]	6
	Minimum samples per leaf	[20, 25, 30]	30

This performance can be interpreted as a result of a structural alignment between the graph-based framework and the localized splitting mechanics of the Random Forest classifier. By modeling geometric connectivity via graph-probabilistic diffusion paths,

Table 2. Comparative performance of the evaluated feature extraction methods and the Baseline RF. Validation results are presented as mean ± 95% confidence interval from stratified 5-fold cross-validation, while Test results correspond to the independent test set.

	AUC		Accuracy		Precision		Recall		F1-Score	
Method	**CI**	**Test**	**CI**	**Test**	**CI**	**Test**	**CI**	**Test**	**CI**	**Test**
SPCA	0.71 ± 0.03	0.68	0.67 ± 0.02	0.64	0.66 ± 0.03	0.63	0.70 ± 0.04	0.69	0.68 ± 0.02	0.66
NMF	0.69 ± 0.03	0.68	0.64 ± 0.02	0.64	0.64 ± 0.03	0.64	0.67 ± 0.04	0.67	0.65 ± 0.01	0.66
Kernel PCA	0.70 ± 0.03	0.69	0.64 ± 0.04	0.62	0.65 ± 0.05	0.64	0.65 ± 0.02	0.60	0.65 ± 0.03	0.62
Diffusion Maps	0.71 ± 0.03	0.70	0.65 ± 0.02	0.65	0.66 ± 0.02	0.69	0.64 ± 0.12	0.55	0.65 ± 0.06	0.61
Autoencoder	0.68 ± 0.01	0.67	0.62 ± 0.02	0.62	0.62 ± 0.01	0.63	0.67 ± 0.02	0.64	0.65 ± 0.02	0.65
VAE	0.65 ± 0.03	0.62	0.62 ± 0.04	0.57	0.62 ± 0.03	0.58	0.67 ± 0.05	0.57	0.64 ± 0.04	0.58
Baseline RF	0.73 ± 0.03	0.71	0.66 ± 0.05	0.65	0.65 ± 0.04	0.67	0.71 ± 0.07	0.64	0.68 ± 0.05	0.65

Table 3. Dimensionality reduction metrics and architectural compression ratios achieved across the feature extraction frameworks, showcasing the mapping from original selected SNPs to latent features.

Method	Input Features	Latent Features	Compression Ratio
Baseline RF	57	57	1.0
SPCA	57	25	2.28
NMF	57	25	2.28
Diffusion Maps	57	5	11.4
KPCA	57	25	2.28
Autoencoder	57	20	2.85
VAE	57	10	5.7

Diffusion Maps cluster individuals based on non-linear structural similarities. When these 5 unsupervised topological coordinates are fed into the decision trees, the classifier constructs effective decision boundaries while achieving an $11.4\times$ reduction in input dimensionality without significant performance degradation. These results suggest that, for this genomic dataset, AD-related patterns may be more effectively captured through localized graph connectivity than through global variance maximization.

However, despite the high compression efficiency and robust precision achieved by this non-linear manifold mapping, the lower recall metric (0.55) observed for Diffusion Maps compared to the baseline Random Forest (0.64) presents a clear technical limitation. Because the topological graph embedding operates in an unsupervised manner, the geometric compression prioritizes preserving the structural connectivity of the genomic variants without direct knowledge of the clinical outcomes. Consequently, heterogeneous or early-stage AD profiles may remain geometrically close to cognitively normal individuals within the latent space, hindering the classifier's ability to minimize false negatives.

This localized geometry premise is supported by the lower performance of Ker-

nel PCA (Test AUC = 0.69). Despite using a cosine kernel to model complex non-linear boundaries, KPCA's constraint to maximize global variance was outmatched by the localized connectivity of Diffusion Maps, suggesting that relevant discriminative information may be concentrated in localized regions of the genomic manifold rather than being distributed across the entire feature space. Within the linear domain, SPCA and NMF achieved macro-level parity (Test AUC-ROC = 0.68, Accuracy = 0.64) at $d = 25$, but diverged granularly in their classification dynamics. SPCA exhibited higher sensitivity (Test Recall = 0.69 vs. 0.67 for NMF), making it more effective at identifying true positive AD cases. Conversely, NMF demonstrated superior positive predictive value (Test Precision = 0.64 vs. 0.63 for SPCA), indicating a lower rate of false positive assignments.

Finally, the deep learning frameworks underperformed, with the VAE yielding the lowest discriminative metrics (Test AUC-ROC = 0.62). Although both networks used restricted hidden layers to prevent overfitting, the discrete and sparse nature of alternative allele frequencies challenges deep architectures. Crucially, the continuous probabilistic space enforced by the VAE via Kullback-Leibler divergence smoothed out the boundaries required to segregate the cohorts. By mapping discrete multi-locus variants into a continuous Gaussian latent distribution, the VAE reduced the biological contrasts between CN and AD groups, demonstrating that generative smoothing can obscure sparse pathogenic boundaries in high-dimensional GWAS datasets.

5. Conclusion

This study compared linear, non-linear, and Deep Learning feature extraction methods for Alzheimer's disease genomic classification. Diffusion Maps achieved near-parity with the baseline model using only 5 latent components instead of the 57 original SNPs, suggesting that AD genetic patterns are effectively captured by local graph connectivity rather than global variance. Concurrently, linear frameworks (SPCA/NMF) showed solid macro parity by balancing sensitivity and precision, whereas generative deep architectures (VAE) underperformed due to the sparse, discrete nature of allele frequencies. Although lower recall presents a technical limitation—and a clinical risk of delaying early treatment due to false negatives—the precision, AUC-ROC, and compression capacity offer a solid basis for future work. Future studies can explore alternative classifiers beyond Random Forest, use supervised mappings, adjust decision thresholds, or expose the variants underlying latent representations to improve biological interpretation—a capability our framework already supports. Despite computational constraints, this study establishes a clear benchmark for geometric feature extraction and polygenic mapping in neurodegenerative genomics.

References

Al-Shoukry, S., Rassem, T. H., and Makbol, N. M. (2020). Alzheimer's Diseases Detection by Using Deep Learning Algorithms: A Mini-Review. *IEEE Access*, 8:116122–116133. DOI: https://doi.org/10.1109/ACCESS.2020.2989396.

Alatrany, A. S., Hussain, A. J., Mustafina, J., and Al-Jumeily, D. (2022). Machine Learning Approaches and Applications in Genome Wide Association Study for Alzheimer's Disease: A Systematic Review. *IEEE Access*, 10:62402–62423. DOI: https://doi.org/10.1109/ACCESS.2022.3182543.

Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., and Xu, Y. (2024). Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review*, 57(3):1–38.

Borah, K., Das, H. S., Seth, S., Mallick, K., Rahaman, Z., and Mallik, S. (2024). A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. *Functional & Integrative Genomics*, 24(5):155. DOI: https://doi.org/10.1007/s10142-024-01415-x.

Bro, R., Acar, E., and Kolda, T. G. (2008). Resolving the sign ambiguity in the singular value decomposition. *Journal of Chemometrics*, 22(2):135–140. DOI: https://doi.org/10.1002/cem.1122.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30. DOI: https://doi.org/10.1016/j.acha.2006.04.003.

Dehbozorgi, P., Ryabchykov, O., and Bocklitz, T. W. (2025). A comparative study of statistical, radiomics, and deep learning feature extraction techniques for medical image classification in optical and radiological modalities. *Computers in Biology and Medicine*, 186:109768.

Fridman, C., Gregório, S. P., Neto, E. D., and Élida P. Benquique Ojopi (2004). Alterações genéticas na doença de Alzheimer. *Arch. Clin. Psychiatry (São Paulo)*, 31(1):19–25. DOI: https://doi.org/10.1590/S0101-60832004000100004.

Khanal, S., Chen, J., Jacobs, N., and Lin, A.-L. (2021). Alzheimer's Disease Classification Using Genetic Data. In *Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1112–1117. DOI: https://doi.org/10.1109/BIBM52615.2021.9669730.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–14. arXiv preprint arXiv:1312.6114. Disponível em: https://arxiv.org/abs/1312.6114.

Labory, J., Njomgue-Fotso, E., and Bottini, S. (2024). Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data. *Computational and Structural Biotechnology Journal*, 23:1274–1287. DOI: https://doi.org/10.1016/j.csbj.2024.03.016.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791. DOI: https://doi.org/10.1038/44565.

Malik, I., Iqbal, A., Gu, Y. H., and Al-antari, M. A. (2024). Deep learning for Alzheimer's Disease Prediction: A Comprehensive Review. *Diagnostics*, 14(12):1281. DOI: https://doi.org/10.3390/diagnostics14121281.

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2):e1608. DOI: https://doi.org/10.1002/mpr.1608.

Najar, J., van der Lee, S. J., Joas, E., Wetterberg, H., Hardy, J., Guerreiro, R., Bras, J., Waern, M., Kern, S., Zetterberg, H., Blennow, K., Skoog, I., and Zettergren, A. (2021). Polygenic risk score for alzheimer's disease is related to dementia risk in individuals with negative apoe $\varepsilon 4$ allele. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 13(1):e12142. PMCID: PMC7821873, PMID: 33532541.

Nasr, M. K., König, E., Fuchsberger, C., Ghasemi, S., , et al. (2025). Removing array-specific batch effects in GWAS mega-analyses by applying a two-step imputation workflow. *Bioinformatics Advances*, 6(1):vbaf317. DOI: https://doi.org/10.1093/bioadv/vbaf317.

Pölsterl, S., Conjeti, S., Navab, N., and Katouzian, A. (2016). Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection. *Artificial Intelligence in Medicine*, 72:1–11. DOI: https://doi.org/10.1016/j.artmed.2016.07.004.

Saykin, A. J., Shen, L., Yao, X., Kim, S., Nho, K., Risacher, S. L., Ramanan, V. K., Foroud, T. M., Faber, K. M., Sarwar, N., Munsie, L. M., Hu, X., Soares, H. D., Potkin, S. G., Thompson, P. M., Kauwe, J. S. K., Kaddurah-Daouk, R., Green, R. C., Toga, A. W., and Weiner, M. W. (2015). Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, Opportunities, and Plans. *Alzheimers Dement.*, 11(7):792–814.

Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., Cummings, J., and van der Flier, W. M. (2021). Alzheimer's disease. *The Lancet*, 397(10284):1577–1590. DOI: https://doi.org/10.1016/S0140-6736(20)32205-4.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319. DOI: https://doi.org/10.1162/089976698300017467.

Smid, J., Studart-Neto, A., César-Freitas, K. G., Dourado, M. C. N., Kochhann, R., Barbosa, B. J. A. P., Schilling, L. P., Balthazar, M. L. F., Frota, N. A. F., de Souza, L. C., Caramelli, P., Bertolucci, P. H. F., Chaves, M. L. F., Brucki, S. M. D., Nitrini, R., de Paula França Resende, E., and Vale, F. A. C. (2022). Declínio cognitivo subjetivo, comprometimento cognitivo leve e demência - diagnóstico sindrômico: recomendações do Departamento Científico de Neurologia Cognitiva e do Envelhecimento da Academia Brasileira de Neurologia. *Dement. neuropsychol.*, 16(3 suppl 1):1–16. DOI: https://doi.org/10.1590/1980-5764-DN-2022-S101PT.

Squillario, M., Abate, G., Tomasi, F., Tozzo, V., Barla, A., and Uberti, D. (2020). A telescope GWAS analysis strategy, based on SNPs-genes-pathways ensamble and on multivariate algorithms, to characterize late onset Alzheimer's disease. *Scientific Reports*, 10(1):12063. DOI: https://doi.org/10.1038/s41598-020-67699-8.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286. DOI: https://doi.org/10.1198/106186006X113430.