A Hybrid mRMR-RFE and AI Framework for Advancing Alzheimer's Biomarkers Discovery

Md. Maniruzzaman
Department of EE
School of Engineering, San Francisco Bay University
Fremont, CA 94539, USA
manir8421@gmail.com

Md. Shahadat Jaman
Department of EEE
Bangladesh University of Business and Technology
Dhaka, Bangladesh
shahadat30eee@gmail.com

Md Amzad Sadik Abid
Lamar College of Business
Lamar University
Texas, USA
mabid@lamar.edu

Zakaria Mahmud
Lab Systems and Automation
GlaxoSmithKline (GSK)
Cambridge, MA, 02142, USA
zakaria.mahmud.bappy@gmail.com

Muhammad Enayetur Rahman
Department of ECE
Old Dominion University
Norfolk, VA, 23508, USA
mrahm011@odu.edu

Md Nurul Absar Siddiky
Department of ECE
UNC at Charlotte
Charlotte, NC, 28223, USA
msiddiky@uncc.edu

***Abstract*—Alzheimer's disease (AD) biomarkers are highly variable, which complicates early prediction. In this study, we provide a comparative study of machine learning (ML) and deep learning (DL) techniques integrated with strong feature selection methods in the context of AD classification. This paper mainly focuses on one of the most important issues in this field: how to construct accurate models based on high-dimensional datasets where the size of the sample is very small. To tackle this problem, we evaluated various feature selection methods: minimum redundancy maximum relevance (mRMR), recursive feature elimination (RFE), and their hybrid variant based on gene expression data. Models like Multilayer Perceptron (MLP) and Random Forest Classifier (RFC) were utilized on a dataset containing 445 samples and 923 features (genes) for which the small sample size made classification challenging. Experiments show that the hybrid feature selection method found the SLC25A46 gene from the entire dataset and, under fine-tuning, substantially boosts Alzheimer's disease classification performance. The best validation average accuracy rate of 95% is achieved for the proposed method. This is simple and can be used in the classification of AD and drug discovery tasks.**

***Index Terms*—Alzheimer's disease, biomarkers, gene expression, mRMR, RFE, random forest.**

I. Introduction

Alzheimer's disease is one such common neurological disorder that manifests as a gradual decrease in cognitive function, leading to memory loss. This impacts millions across the globe, infecting brain neurons responsible for language and memory functions. The disease usually manifests itself after the age of 65, and the risk of developing it grows exponentially with age. Alzheimer's disease (AD) is one of the most frequent types of dementia [1]. It is a slowly progressive chronic neurodegenerative disease that presents with subtle changes. Alzheimer's is a complex neurodegenerative disease that has a strong genetic component. Understanding its onset, development, and pathophysiological origins is critical to future studies and therapeutic approaches. In 2015, approximately 47 million people globally suffered from AD, and the overall costs exceeded $818 billion. These numbers are expected to increase in the future [2].

Microarray technology can identify genes that cause Alzheimer's disease (AD) in order to predict gene expression profiles, create efficient AD treatments, and deliver personalized healthcare. Processing microarray data comes with a number of challenges, including redundancy and overfitting that need to be avoided and managed with caution, given that such techniques work on a large number of genes and samples. Unlike most other methods that extract all association genes (in Alzheimer's disease context), a certain method, called the "gene selection approach," reduces computational cost, improves efficiency, and ensures that researchers can still identify only the key genes for disease classification. For example, gene selection is performed here using unsupervised methods like PCA and SVD for the gene expression microarray data. These approaches produce lower-dimensional representations for classification tasks and reveal the structure of datasets [3]. Microarray gene expression data is a more promising approach for early AD detection than neuroimaging and EEG, each of which has pros and limitations of its own.

In order to predict Alzheimer's disease using gene expression data, Alia et al. [4] employed a hybrid feature selection strategy. They selected genes using LASSO and ANOVA methods. The Support Vector Machine (SVM) classifier attained the maximum performance values in just one dataset, while the Multilayer Perceptron (MLP) classifier achieved the best performance metrics in four datasets. A deep learning algorithm was presented by Chihyun et al. [5] to forecast AD using a dataset that combined DNA methylation and gene expression. PCA-gene expression, PCA-DNA methylation, t-SNE-gene expression, and t-SNE-DNA methylation were the

four training examples. The deep neural network's average accuracy was 82.3%.

Machine learning-based binary and multiclass classification for early Alzheimer's disease diagnosis was proposed by M. Sudharsan et al. [6]. In a multiclass grouping of sMRI data from the ADNI dataset, they examined a PCA-based search strategy. In conjunction with feature selection methods, RELM significantly improved the accuracy of classifying AD from MCI and HC individuals. A trustworthy deep-learning model was created by Mahmoud M. Abdelwahab et al. [3] to forecast Alzheimer's disease early. They used GSE63060 and GSE63061 microarray gene expression data, which included 569 samples and 16,383 genes. Outstanding results on the AD dataset demonstrated the PCA–CNN model's efficacy, with an accuracy of 96.60% and a loss of 0.3503. In contrast, the SVD–CNN model demonstrated exceptional accuracy, attaining 97.08% with a loss of 0.2466.

Hala Alshamlan et al. [7] used several feature selection techniques on the large-scale gene expression profiles, GSE33000 and GSE44770, which contained 19,488 genes in total, 257 of which were normal and 439 of which were AD samples, to determine the best machine learning model for identifying risk genes linked to AD. With many genes between 20 and 40, the results show that the mRMR and F-score feature selection approaches with an SVM classifier produced a high accuracy of about 84%.

Aliaa SaadEl-Gawady et al. [8] proposed a machine-learning approach for the prediction of Alzheimer's disease. They identified 1058 significant genes using a dataset of 1157 cases and 39,280 genes. The SVM model (sensitivity/recall: 0.97; specificity: 0.97; precision: 0.98; kappa index: 0.945; AUC: 0.972; accuracy: 0.975) demonstrated the maximum accuracy. Yi Zhang et al. [9] built multiclass eXtreme Gradient Boosting models (XGBoost) to characterize large-scale transcriptomic-based blood biomarkers and evaluated their performances in distinguishing AD from cognitive normal (CN) and mild cognitive impairment (MCI). For multiclass classification, this study's area under the receiver operating characteristic curve (AUC) was 0.81 (sensitivity = 0.81; specificity = 0.63).

Using longitudinal MRI images from the ADNI dataset, Zhentao Hu et al. [10] proposed a novel VGG-TSwinformer model for early Alzheimer's disease prediction. Low-level spatial features of longitudinal sMRI images are extracted using a CNN based on VGG-16, and these low-level features are mapped to high-level feature representations. Accuracy, sensitivity, specificity, and AUC for the sMCI vs. pMCI classification task were 77.2%, 79.97%, 71.59%, and 0.8153, respectively. Wujia Yu et al. [11] examined the potential for genetic and EEG data to be combined for subclassification of AD. Results from this study indicated that the support vector machine (SVM) model had noteworthy success in terms of classification performance with an accuracy of 0.920 and an AUC value of 0.916.

There is evidence that early diagnosis/treatment in patients with Alzheimer's disease (AD) can improve their prognosis. Machine learning (ML) and deep learning (DL) techniques have emerged as powerful tools in all sectors [12]–[15] including medical areas [13], [16], particularly in the early detection of Alzheimer's Disease (AD). Previous studies were performed using brain imaging data, but some others have also used gene or cellular point data. Yet, a proper method for feature selection, not depending on the model, is still needed. Existing approaches often face issues of underfitting and overfitting. To address these challenges, we have analyzed ML and DL models alongside various robust feature selection techniques to enhance the accuracy of AD predictions, providing support for medical professionals in clinical decision-making.

II. Methodology

A. Dataset Description

To define our machine learning (ML) and deep learning (DL) models for early prediction of AD, we increased the sample size in this study by combining datasets with the following identifiers: GSE5281 (161 samples), GSE48350 (253 samples), and GSE1287 (31 samples). This resulted in a total of 445 samples from various brain regions, including the hippocampus, entorhinal cortex, medial temporal gyrus, posterior cingulate, superior frontal gyrus, primary visual cortex, and post-central gyrus [17]. The combined dataset contains 923 features (genes), with 256 samples representing healthy controls and 189 samples associated with Alzheimer's Disease (AD).

B. Block Diagram

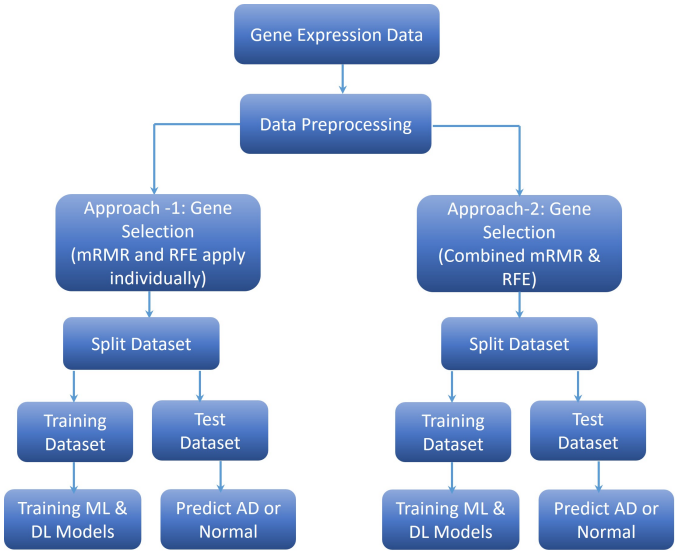

Fig. 1: Flow diagram of the proposed model

C. Feature Selection Methods

Finding genes with the most information regarding the sample class labels is the aim of any pertinent gene selection process about microarray gene expression data. In this study, we have demonstrated two effective features selection methods that can identify most risk genes in AD.

1) Minimum Redundancy and Maximum Relevance (mRMR): mRMR select the features with high correlation to the target vaiable and reduce features when high correlation themselves In this study, the goal is to minimize the redundancy of a gene while maximizing its significance with output. Mutual information difference (MID) has been used as an objective function to select and reduce genes from the score metrics. This approach intended two sight such as the mRMR feature set will be more representative of the target phenotypes, we expect it to have better generalization qualities with an equivalent amount of features. Similarly, a smaller set of mRMR features can effectively cover the same area as a larger conventional feature set [18]. The MID function has been shown in Equation 1.

$$mRMR(X_i) = Relevance(X_i, Y) - \frac{1}{|S|} \sum_{X_j \in S} Redundancy(X_i, X_j) \quad (1)$$

Where X_i represents the genes in the dataset and Y represents the target variable.

Here, Relevance (X_i, Y) = mutual information, $I(X_i, Y)$ and Redundancy $(X_i, Y) = I(X_i, Y)$

The first feature is selected by the using of equation 2.

$$\text{Max}V_I, \quad V_I = \frac{1}{|S|} \sum_{i \epsilon S} I(X_i, Y) \quad (2)$$

The rest of the features are selected in an incremental way and earlier selected feature remain in the feature set.

Algorithm 1 Procedure of mRMR.

1: At beginning, feature set S = {}.
2: Calculate mutual information, $I(X_i; Y)$ to measure relevance.
3: Select first feature from relevance metrics.
4: Iteratively add features to S by selecting the feature X_i that maximizes the $mRMR$ objective:

i. $\arg\max_{X_i \notin S} mRMR(X_i)$

5: Stop when the desired number of features is selected.

2) Recursive Feature Elimination (RFE): In this algorithm, genes are ranked according to their impact on the value of estimator means classification algorithm using an iterative procedure called Recursive Feature Elimination (RFE). First, the entire genes set is used to train the classifier, and the classification accuracy is tracked [19]. The desire value of the classifier is then monitored following each iteration in which several genes are eliminated from the initial set. The whole procedure are given below.

3) Steps to Combine mRMR with RFE:

1) To choose a subset of the most pertinent and non-redundant features, use mRMR. As a result, the dataset's dimensionality is decreased but its most informative properties are preserved.

Algorithm 2 Procedure of RFE.

1: Set ranking criteria w_i, which determine the contribution of classification accuracy, given by:

$$w_i, = (\mu_i, (+) - \mu_i(-))/(\sigma_i(+) + \sigma_i(-)) \quad (3)$$

Where μ_i = mean and σ_i = standard deviation of the ith gene.
2: Train the classifier (Random Forest) with weights optimized for cost function using all genes.
3: Compute ranking criteria for all genes.
4: Remove the r number of genes that have the lowest ranking criterion.
5: Repeat steps 3 to 5.

TABLE I: Hyperparameter search space for RF and MLP models.

Model	Hyperparameters	Search Space
RF	Max_depth	[10, 20, 40, 100, **None**]
	N_estimators	[100, 200, **300**, 400]
	Min_sample_leaf	[**1**, 2, 3, 4, 5]
	Min_sample_split	[1, **2**, 4, 6, 8]
MLP	Node size in each layer	[64, **32, 16, 8**, 128, 256]
	Activation function	[relu, softmax, sigmoid, tanh]
	Optimizer	[SGD, adam, **RMSprop**, Nadam]
	Batch size	[8, 16, 32, 64, **128**]
	Epoch	[50, 100, **200**, 300]

2) The Random Forest machine learning model is trained, and the least significant features are recursively removed based on the model's performance (i.e., feature coefficients or significance scores).
3) The final feature set balances statistical relevance (mRMR) and predictive performance (RFE).

D. ML and DL Model Developing

This study used RF and MLP architecture to detect between healthy controls and AD. To find the optimal parameter. we used grid search technique in search space for both models. In table 1, [bold] value represent the optimal parameter in search space.

The following random forest algorithm has been done for our study:

Input: Dataset D, the number of trees NT, and the threshold T of Gini impurity.

Output: A random forest for $i = 1$ to NT :

1) Draw a bootstrap sample D_i of size n from the training set D.
2) Construct a decision tree of the bootstrapped data recursively from the root node. Repeatedly perform the following steps until the Gini impurity is less than T:
   a) Randomly select a subset of $\sqrt{m}$ features.
   b) For $j = 1$ to $\sqrt{m}$:
      - Compute Gini impurity for feature x_j.
   c) As the split attribute and split value, select the feature and value with the lowest Gini impurity.

d) Split the internal node into two child nodes according to the split feature and value.

We implemented MLP model with three hidden layers, batch size, and epoch with [32, 6, 8], 128, and 200 respectively followed by dense layers. In this model, we set learning rate 0.01 and also used dropout layer to prevent over fitting problem. The MLP model is implemented with the API of Google TensorFlow (version 1.4.1).

To evaluate of proposed model, five fold cross validation technique is used where the process is repeated five time by one subset for testing and rest of the sets for training for both models.

III. Result and Discussion

To choose genes utilizing hybrid feature selection techniques, such as merging Minimum Redundancy Maximum Relevance (mRMR) and Recursive Feature Elimination (RFE), and comparing the model performance of Multilayer Perceptron (MLP) and Random Forest Classifier (RFC) for Alzheimer's disease (AD) prediction, the suggested method was applied to gene expression datasets. The benchmark dataset consists of 445 samples and 923 genes, where 256 belong to healthy controls and 189 to AD, as shown in Fig. 2. A 70:30 split was made between the input dataset's training and testing sets. The average values of classification accuracy, precision, and recall were measured after the experiment was conducted ten times. To obtain the study's final results, fivefold cross-validation was employed.

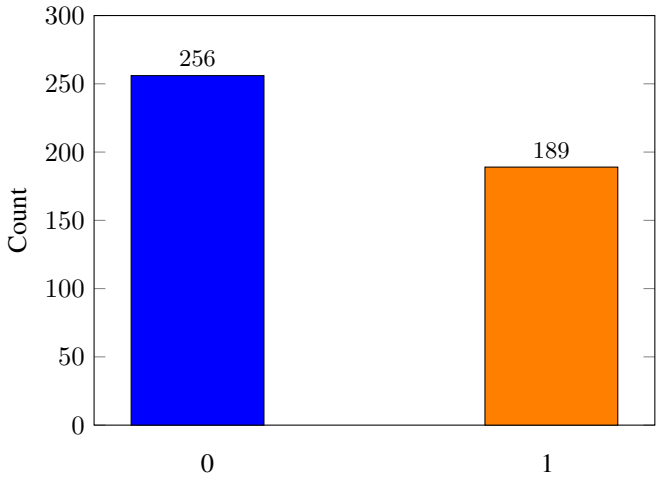

Fig. 2: Class Distribution: (Class 0 = Healthy controls, Class 1 = AD).

In table II and III, the details of the MLP model are presented. The model has 3,466 total parameters, all of which are trainable, meaning there are no frozen layers. The small number of parameters indicates that this is a compact model suitable for tasks requiring low computational resources. Dropout layers follow each dense layer, which helps prevent overfitting by regularizing the network. This improves the model's generalization on unseen data.

TABLE II: MLP Model architecture.

Layer (type)	Output Shape	No. of Param
Dense	(None, 64)	704
Dropout	(None, 64)	0
Dense	(None, 32)	2080
Dropout	(None, 32)	0
Dense	(None, 16)	528
Dropout	(None, 16)	0
Dense	(None, 8)	136
Dropout	(None, 8)	0
Dense	(None, 2)	18

TABLE III: MLP Model Parameters.

Total Parameters	3,466
Trainable Params	3,466
Non-trainable Params	0

Table IV presents the optimal hyperparameters used for the Random Forest model in this study. The model was configured with 300 estimators (`n_estimators`) to enhance the robustness and stability of predictions. The `max_depth` parameter was set to `None`, allowing each tree to grow fully until all leaves are pure or contain fewer than the minimum required samples. The minimum number of samples required to split a node (`min_samples_split`) is set to 2, while the minimum samples required at a leaf node (`min_samples_leaf`) is set to 1. For selecting features during the splits, the `max_features` parameter was set to `'sqrt'`, which optimizes the trade-off between accuracy and computational cost. The model employs bootstrapping (`bootstrap=True`) to sample data with replacement and balances class weights (`class_weight=balanced`) to handle class imbalances effectively. These parameter settings ensure that the Random Forest model achieves high performance while maintaining fairness in predictions.

The significance of features ranked using hybrid (mRMR & RFE) feature selection techniques is depicted in the bar plot in Fig. 3. This investigation identifies the most important characteristics influencing the model's predictive performance.

TABLE IV: Optimal value of parameters for random forest.

Parameter	Value
n estimators	300
max depth	None
min samples split	2
min samples leaf	1
max features	'sqrt'
bootstrap	True
class weight	balanced

The hybrid gene selection approach highlighted *SLC25A46* as the most crucial trait, followed by several other significant attributes. These findings can guide focused research or simplify models, thereby increasing efficiency and interpretability.

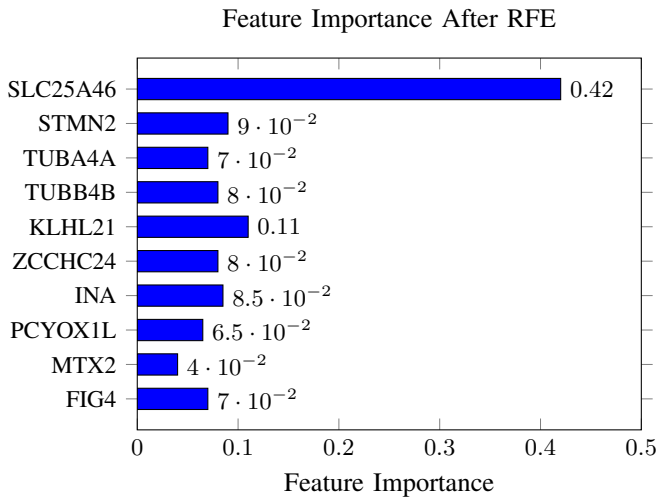

Fig. 3: Desired Selected Gene by Hybrid Method

The hybrid mRMR-RFE approach was used initially separately and later in tandem for gene selection in the classification of Alzheimer's disease (AD). When compared to separate approaches, the combined mRMR and RFE technique performed better. The combined strategy with the RF classifier produced the highest accuracy, precision, and recall values, which are 0.95, 0.94, and 0.94, respectively, as shown in Fig. 4. This approach outperformed the MLP model and the individual gene selection techniques. Notably, compared to RFE alone, the mRMR filter approach continuously increased classification accuracy when used in conjunction with RFE. These results demonstrate that the mRMR-RFE technique is promising for identifying pertinent genes and effectively removing redundant ones.

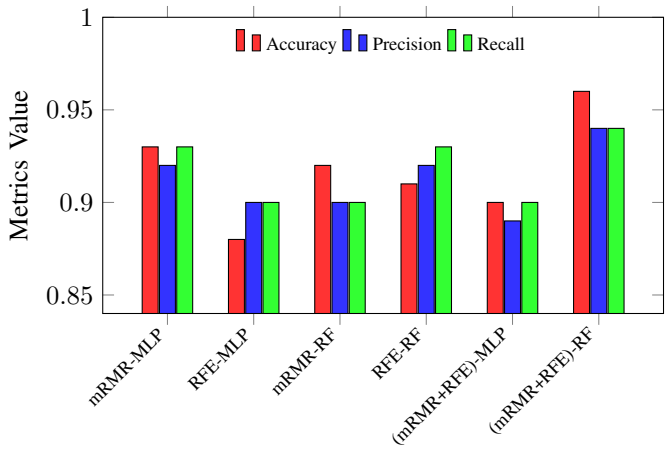

Fig. 4: Model Performance Comparison Using Various Gene Selection Methods.

Using the efficacy of hybrid (mRMR-RFE) and individual gene selection techniques, we have demonstrated the optimal performance of RF by following different numbers of genes in Fig. 5. Our study used the Mutual Information Difference (MID) scheme, which calculated the mRMR score for every gene in the dataset while reducing gene redundancy. The top 10 genes were then chosen using the RFE approach by the random forest classifier model. This approach achieved 95% accuracy from the hybrid method using 10 effective genes.

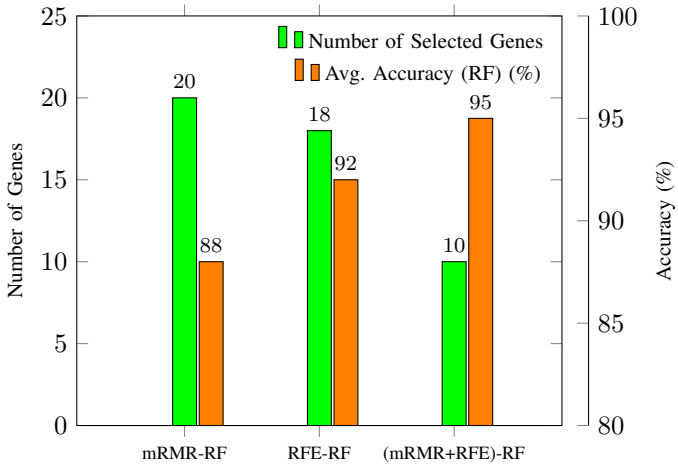

Fig. 5: Accuracy of Random Forest Model for Different Number of Genes.

IV. Conclusion

The useful suggested feature selection method in Alzheimer's Disease (AD) has been examined with machine and deep learning methods in this study. By altering each step of the procedure, we demonstrated a number of comparative experiments. The study's high dimensionality and small sample sizes, which impacted the models' effectiveness, were its drawbacks. This work used combinaton of minimum redundancy maximum relevance (mRMR) and recursive feature elimination (RFE) for gene selection due to the restrictions of a small number of samples and high-dimensional data. These techniques lower the dimensionality of the data, which is crucial for processing gene expression data.

Each model's performance was assessed using a variety of criteria, including recall, accuracy, and precision. In the gene expression dataset for Alzheimer's Disease, the hybrid feature selection with the random forest classifier produced impressive metric scores. In the future, we intend to use the vast amount of samples combined with two or more multi-omics datasets to create a generalizability of our model.

References

[1] S. Ahmed and S. Kadhem, "Early alzheimer's disease detection using different techniques based on microarray data: A review," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, pp. 106–126, 03 2022.

[2] S. Bringas, S. Salomón, R. Duque, C. Lage, and J. L. Montaña, "Alzheimer's disease stage identification using deep learning models," *Journal of Biomedical Informatics*, vol. 109, p. 103514, 2020.

[3] M. M. Abdelwahab, K. A. Al-Karawi, and H. E. Semary, "Deep learning-based prediction of alzheimer's disease using microarray gene expression data," *Biomedicines*, vol. 11, no. 12, p. 3304, 2023.

[4] A. El-Gawady, B. S. Tawfik, and M. A. Makhlouf, "Hybrid feature selection method for predicting alzheimer's disease using gene expression data," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 74, no. 3, pp. 5559–5572, 2023.

[5] C. Park, J. Ha, and S. Park, "Prediction of alzheimer's disease based on deep neural network by integrating gene expression and dna methylation dataset," *Expert Systems with Applications*, vol. 140, p. 112873, 2020.

[6] M. Sudharsan and G. Thailambal, "Alzheimer's disease prediction using machine learning techniques and principal component analysis (pca)," *Materials Today: Proceedings*, vol. 81, pp. 182–190, 2023.

[7] H. Alshamlan, S. Omar, R. Aljurayyad, and R. Alabduljabbar, "Identifying effective feature selection methods for alzheimer's disease biomarker gene detection using machine learning," *Diagnostics*, vol. 13, no. 10, p. 1771, 2023.

[8] A. El-Gawady, M. A. Makhlouf, B. S. Tawfik, and H. Nassar, "Machine learning framework for the prediction of alzheimer's disease using gene expression data based on efficient gene selection," *Symmetry*, vol. 14, no. 3, p. 491, 2022.

[9] Y. Zhang, S. Shen, X. Li, S. Wang, Z. Xiao, J. Cheng, and R. Li, "A multiclass extreme gradient boosting model for evaluation of transcriptomic biomarkers in alzheimer's disease prediction," *Neuroscience Letters*, vol. 821, p. 137609, 2024.

[10] Z. Hu, Z. Wang, Y. Jin, and W. Hou, "Vgg-tswinformer: Transformer-based deep learning model for early alzheimer's disease prediction," *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107291, 2023.

[11] W.-Y. Yu, T.-H. Sun, K.-C. Hsu, C.-C. Wang, S.-Y. Chien, C.-H. Tsai, and Y.-W. Yang, "Comparative analysis of machine learning algorithms for alzheimer's disease classification using eeg signals and genetic information," *Computers in Biology and Medicine*, vol. 176, p. 108621, 2024.

[12] M. N. A. Siddiky, M. E. Rahman, M. S. Uzzal, and H. M. D. Kabir, "A comprehensive exploration of 6g wireless communication technologies," *Computers*, vol. 14, no. 1, 2025. [Online]. Available: https://www.mdpi.com/2073-431X/14/1/15

[13] M. Arifuzzaman, M. J. U. Chowdhury, I. Ahmed, M. N. A. Siddiky, and D. Rashid, "Heart disease prediction through enhanced machine learning and diverse feature selection approaches," in *2024 IEEE 10th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, 2024, pp. 119–124.

[14] M. E. Rahman, M. S. Munir, and S. Shetty, "An attention-based ai model for 3d beam prediction in thz unmanned aerial vehicle communication," in *2024 International Conference on Computing, Networking and Communications (ICNC)*, 2024, pp. 761–766.

[15] S. Sakib, M. N. A. Siddiky, M. Arifuzzaman, and M. J. U. Chowdhury, "Optimizing facial recognition: An analytical comparison of traditional and deep learning approaches," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, 2024, pp. 271–276.

[16] S. M. K. Pathan, S. B. Imran, M. M. S. Iqbal, M. E. Rahman, M. N. A. Siddiky, M. R. Rahman, M. R. Hasan, N. L. Dey, and M. S. Hossain, "Comparative analysis of machine learning models for predicting healthcare traffic: Insights for optimized emergency response," *Magna Scientia Advanced Research and Reviews*, vol. 12, no. 2, pp. 054–061, 2024.

[17] H. Alamro, M. A. Thafar, S. Albaradei, T. Gojobori, M. Essack, and X. Gao, "Exploiting machine learning models to identify novel alzheimer's disease biomarkers and potential targets," *Scientific reports*, vol. 13, no. 1, p. 4979, 2023.

[18] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[19] N. Koul and S. S. Manvi, "A scheme for feature selection from gene expression data using recursive feature elimination with cross validation and unsupervised deep belief network classifier," in *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*. IEEE, 2019, pp. 31–36.