Wrapper-based Modified Binary Particle Swarm Optimization for Dimensionality Reduction in Big Gene Expression Data Analytics

Hend S. Salem, Mohamed A. Mead, Ghada S. El-Taweel
Department of Computer Sciences, Suez Canal University, Ismailia, Egypt

***Abstract*—Gene expression data has emerged as a crucial aspect of big data in genomics. The advent of high-throughput technologies such as microarrays and next-generation sequencing has enabled the generation of extensive gene expression data. These datasets are characterized by their complexity, fast data generation, diversity, and high dimensionality. Analyzing high dimensional gene expression data offers both challenges and opportunities. Computational intelligence and deep learning techniques have been employed to extract meaningful information from these enormous datasets. However, the challenges related to preprocessing, reducing dimensionality, and normalization continue to exist. This study explored the effectiveness of the Wrapper-based Modified Particle Swarm Optimization (WMBPSO) algorithm in reducing dimensionality of big gene expression data for Alzheimer's disease (AD) prediction, using the GSE33000 dataset. The reduced dataset was then used as input to a CNN-LSTM model for prediction. The WMBPSO method identified 4303 genes out of a total of 39280 genes as being relevant for AD. These genes were selected based on their discriminatory power and potential contribution to the classification task, achieving an accuracy score of 0.98. The performance of the CNN-LSTM model is evaluated using these selected genes, and the results were highly promising. The results of our analysis are 0.968 for mean cross-validation accuracy, 0.995 for AUC, and 0.967 for recall, precision, and F1 score. Importantly, our approach outperforms conventional feature selection methods and alternative machine and deep learning algorithms. By addressing the critical challenge of dimensionality reduction in gene expression data, our study contributes to advancing the field of AD prediction and underscores the potential for improved diagnosis and patient care.**

***Keywords*—Alzheimer disease; big gene expression; binary particle swarm optimization; deep learning; dimensionality reduction**

I. Introduction

In the rapidly evolving landscape of genomics and bioinformatics, the emergence of gene expression data as a prime example of big data presents both opportunities and formidable challenges. Big data, characterized by its immense size and complexity, demands innovative approaches for efficient processing and analysis [1]. Gene expression data, in particular, involves measuring the activity levels of thousands of genes across various biological samples or conditions. This surge in data generation has been fueled by advances in high-throughput technologies, such as microarrays and next-generation sequencing, ushering in an era of information abundance [2]. The following are some key aspects of gene expression data as big data:

- Volume: gene expression data is typically characterized by its sheer volume. Experiments can yield thousands to millions of data points, each representing the expression level of a specific gene in each sample. These large datasets require substantial storage and computational power to manage and analyze effectively [3].
- Variety: gene expression data comes in various formats, such as raw intensity values from microarrays or read counts from RNA sequencing experiments. Additionally, it often includes associated metadata, such as sample annotations, experimental conditions, and clinical information. The integration and analysis of these diverse data types add complexity to the big data challenge [4].
- Velocity: the generation of gene expression data can be incredibly fast due to high-throughput technologies. With the ability to generate a massive amount of data in a short time, there is a need to find new ways to process and analyze the data rapidly [5].
- Complexity: analyzing gene expression data involves complex statistical and computational techniques to identify differentially expressed genes, perform clustering and classification, and infer gene regulatory networks. The complexity of these analyses increases with the size of the dataset [6].
- High dimensionality: each gene expression dataset typically consists of multiple samples (e.g., individuals, cells, or tissues) and thousands of genes. As a result, the data becomes high-dimensional, making it challenging to analyze and interpret effectively [7].
- Diversity of data sources: gene expression data is collected from diverse sources, including different tissues, organs, cell types, and experimental conditions. Integrating data from multiple sources adds complexity to the analysis and requires sophisticated data processing techniques [8].

While these characteristics offer tremendous insights into biological processes, they also present formidable analytical challenges. In light of these challenges, our study sets out to address two pivotal research questions that drive the core of our investigation.

Research Question 1: How can we effectively tackle the inherent complexity and high dimensionality of big gene expression data, specifically in the context of Alzheimer's Disease (AD) prediction?

As gene expression data exhibit substantial volume, diverse formats, rapid generation, inherent complexity, high dimensionality, and diverse data sources, it becomes paramount to devise innovative approaches to streamline the analysis process.

Research Question 1: How can we harness the power of Particle Swarm Optimization (PSO) to enhance deep learning models for gene selection, thereby improving AD prediction?

To overcome the challenges posed by high-dimensional gene expression data, we seek to integrate PSO into the feature selection process of deep learning models. This integration aims to harness PSO's optimization capabilities to select the most relevant genes, ultimately enhancing the performance of AD prediction models.

Alzheimer's Disease (AD) prediction stands out as a compelling application of gene expression data analysis. AD is a complex neurodegenerative disorder marked by progressive cognitive decline and memory loss [9]. Early and precise AD prediction is pivotal for timely interventions and personalized treatment strategies, with gene expression datasets serving as invaluable resources for identifying potential biomarkers and elucidating the underlying molecular mechanisms. Leveraging deep learning models and Particle Swarm Optimization (PSO) holds great promise in enhancing AD prediction accuracy and selecting the most relevant genes associated with the disease [9].

Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have revolutionized diverse fields by excelling in complex tasks like image recognition, natural language processing, and speech synthesis [10] [11]. Their capacity to automatically extract intricate hierarchical representations from data makes them particularly well-suited for gene expression data analysis in AD prediction tasks [12]. However, the effectiveness of deep learning models hinges on the availability of high-quality features that encapsulate pertinent information from the input data.

Feature selection plays a pivotal role in identifying and extracting the most informative features, thereby mitigating computational complexity and enhancing model interpretability [13].

Nonetheless, gene expression datasets often grapple with high dimensionality, encompassing a multitude of genes that may not all be pertinent for AD prediction [14]. Feature selection methods aim to tackle this challenge by combining the merits of filter and wrapper approaches. Filter methods gauge genes based on their statistical relevance to AD, utilizing metrics such as correlation or mutual information. In contrast, wrapper methods assess gene subsets using specific prediction algorithms.

Among these feature selection techniques, PSO has gained traction as an optimization algorithm for identifying the optimal gene subset [15]. Inspired by social behavior, PSO simulates the collective movement of particles within a search space, with each particle representing a potential solution guided by its own best position and the swarm's best position [16].

PSO has demonstrated effectiveness in solving optimization problems, including gene selection for AD prediction, through efficient exploration of the search space and convergence toward promising solutions.

This study's primary objective is to combine the strengths of deep learning models and PSO-based feature selection to enhance feature selection efficiency and deep learning model performance. By integrating PSO into the feature selection process of deep learning models, these hybrid approaches aim to overcome the limitations of traditional feature selection techniques and unlock the full potential of deep learning architectures. The utilization of PSO for feature selection involves two main stages: initialization and iterative optimization [17]. In the initialization stage, the PSO algorithm initializes a population of particles, each representing a potential feature subset. These particles traverse the search space, evaluating their fitness based on a fitness function that quantifies feature subset quality. The iterative optimization phase entails updating particle positions and velocities based on their own best positions and the best position discovered by the swarm, continuing until a termination criterion is met [18].

The integration of deep learning models with PSO-based feature selection offers several advantages. Firstly, it reduces input data dimensionality, crucial for managing large-scale datasets and mitigating overfitting risk. Secondly, it enhances model interpretability by selecting a subset of features most relevant to the target task. Lastly, it augments deep learning models' generalization capability by focusing on discriminative features, potentially leading to superior overall performance.

Having outlined these research questions, our study provides comprehensive answers and innovative solutions. We introduce a gene selection method based on a wrapper-based binary PSO (WBPSO) for dimensionality reduction. This method identifies the optimal subset of genes relevant to AD. Additionally, we propose a hybrid convolutional neural network (CNN) and long short-term memory (LSTM) deep learning model for precise AD prediction. Our study investigates the effectiveness of this approach in improving gene selection efficiency and deep learning model performance across various tasks. Comprehensive experiments conducted on benchmark gene expression datasets allow us to compare our method with other gene selection techniques and validate its superiority.

The remainder of this paper is organized as follows: Section II offers an examination of previous research concerning feature selection techniques and their integration with machine and deep learning models for Alzheimer's Disease (AD) prediction. Section III outlines the materials and methods employed in our proposed approach. Section IV delves into the experimental setup and presents an analysis of the results obtained from our experiments. Section V provides the limitations of the study and some future directions. Lastly, in Section VI, we wrap up the paper by summarizing our discoveries, highlighting the most important findings, and specifying our contribution.

II. Literature Review

This section provides an overview of various investigations concerning feature selection techniques and the prediction of Alzheimer's disease (AD) using machine and deep learning approaches with gene expression data. Each study is summarized and evaluated for its contributions and limitations.

Martinez et al. [19] introduced a methodology to identify AD-associated genes using decision trees, quantitative association rules, and hierarchical clustering. While this approach effectively detected genes with significant expression changes, its scalability was limited for large-scale datasets.

In their work, Park et al. [20] proposed a deep learning model for AD prediction by integrating gene expression and DNA methylation data. Although their model showed improved accuracy compared to traditional machine learning methods, it faced limitations, including a small sample size, potential overfitting, and a lack of benchmarking with logistic regression or other deep learning algorithms.

Sharma et al. [21] employed random forest and regularized regression models (specifically LASSO) to analyze microarray datasets across four brain regions. This approach aimed to identify genetic biomarkers for AD prediction, achieving high accuracy. However, it faced challenges in handling high-dimensional data and potential overfitting.

Chen et al. [22] highlighted the significance of differential network analysis to uncover AD-related genes using the JDINAC machine learning method. This method successfully identified differential networks associated with AD pathology, contributing to a better understanding of the disease.

Patel et al. [23] focused on differentiating individuals with AD from others using gene expression biomarkers in blood samples. While their XGBoost classification models achieved success, there was a need to improve sensitivity and establish a more specific blood signature for AD.

Bogdanovic et al. [24] emphasized proper experimental design and preprocessing techniques to analyze a large dataset. Their approach, based on XGBoost, achieved competitive performance and offered interpretability, highlighting the importance of explainable machine learning in AD diagnosis.

In [25], an autoencoder (AE) was employed to integrate DNA methylation and gene expression data for AD prediction. The approach demonstrated improved accuracy, addressing the challenges of high-dimensional, low-sample size datasets.

Mahendran et al. [26] developed a gene selection pipeline for AD, combining mRmR, WPSO, and Autoencoder methods. They used Bayesian Optimization to tune hyperparameters and achieved promising results.

Lee et al. [27] utilized three publicly available datasets to investigate AD-related genes and develop classifiers. Their approach demonstrated predictive performance, even across different datasets.

Kamal et al. [28] employed machine learning techniques to classify AD using both image and gene expression data. The CNN achieved high accuracy for image data, while SVC demonstrated accuracy for gene expression data, with the aid of LIME for interpretability.

Maj et al. [29] combined deep learning and machine learning techniques to analyze gene expression data in AD. Their study highlighted the potential of recurrent neural networks (RNNs) in modeling gene expression data, although limitations included sample size and sex-specific considerations.

Kim et al. [30] used the SpliceAI framework, based on a variant of convolutional neural networks (CNNs) called the residual CNN model, to predict Alzheimer's disease (AD)-specific nucleotide alteration sites in pre-messenger RNA (mRNA) sequences. They identified 14 splicing sites in the PLCG1 gene with single-nucleotide variants (SNVs) occurring at the same position in both humans and the AD mouse model cortex. The study's limitation lies in investigating only one gene and lacking comparison with existing models. Future studies should consider analyzing more genes and incorporating high-quality gene expression data for a comprehensive evaluation of model performance.

The work in [31], a deep learning model based on Wasserstein Generative Adversarial Networks (GANs) with a gradient penalty term was utilized to predict the virtual disease/molecular progression of Alzheimer's disease (AD) using gene expression data from a mouse AD model. The latent space interpolation of GANs was leveraged to describe pathological pathway cascades in AD progression. However, the study had limitations, including a small number of differentially expressed genes (DEGs) used for training data and a small sample size of gene expression profiles, which hindered drawing conclusive results. Additionally, the proposed model was not compared to existing models to demonstrate its performance, and future studies should consider incorporating more genes and high-quality augmentation data.

Xie et al. [32] introduced MLP-SAE, a deep learning regression model for predicting gene expression based on genetic variation. The model outperformed other methods, highlighting the potential of deep learning in genomics data analysis.

Alhenawi et al. [33] conducted a systematic review of feature selection methods for microarray data analysis, highlighting the prevalence of hybrid feature selection methods as a promising research direction.

The existing methods, while contributing significantly to AD prediction using gene expression data, face limitations ranging from scalability to interpretability. These limitations have created a notable research gap, particularly concerning high dimensionality and feature selection accuracy

In this paper, we introduce a novel gene selection method based on a wrapper-based binary Particle Swarm Optimization (PSO) algorithm (WBPSO). Our approach is designed to overcome the limitations of existing feature selection techniques by efficiently selecting informative genes for AD prediction. Furthermore, we extend our approach by integrating the selected genes into a hybrid convolutional neural network (CNN) and long short-term memory (LSTM) deep learning model. This integration aims to enhance model interpretability and significantly reduce dimensionality, potentially improving overall AD prediction performance.

Table I provides a summary of some recent research investigating the prediction of Alzheimer's disease (AD) through

the analysis of gene expression data, utilizing various gene selection (GS) techniques and machine and deep learning (ML) models.

III. Materials and Methods

This section presents the proposed method, a comprehensive account of the dataset employed, the preprocessing steps applied to the microarray dataset, and the techniques employed for gene selection and AD prediction. As illustrated in Fig. 1, the overall approach consists of four main components: Preprocessing, Gene Selection, AD Prediction, Performance Evaluation. Utilizing this framework enables the opportunity to create a powerful AD prediction system that merges deep learning models with wrapper-based feature selection method that is inspired by nature. This system leverages big gene expression dataset to produce precise and dependable predictions, ensuring accuracy and reliability. The following sections provide a comprehensive explanation of each individual component in the proposed approach.

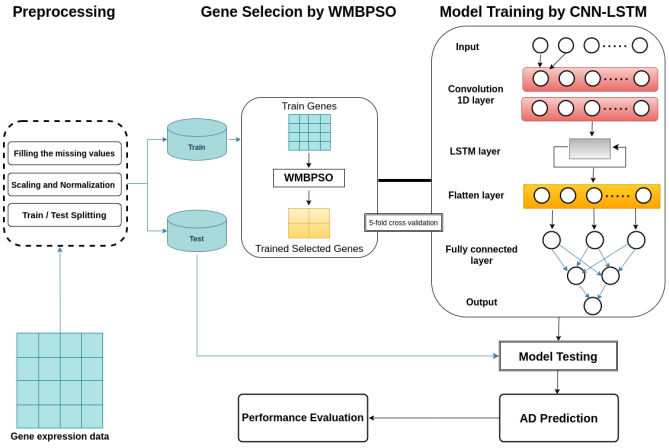

Fig. 1. The overall dimensionality reduction and AD prediction approach.

A. Dataset

The dataset utilized in this study was obtained from the National Center for Biotechnology Information-Gene Expression Omnibus (NCBI-GEO) database [42]. Specifically, the dataset corresponds to the access number GSE33000. It comprises four DNA microarray data profiles, representing multiple tissues in the human brain. These profiles were collected from three distinct brain regions of AD patients: prefrontal cortex (PFC), visual cortex (VC), and cerebellum (CR). However, the focus of the GSE33000 dataset is exclusively on the PFC. In total, the dataset consists of 624 demented and non-demented control cases, each characterized by 39,280 genes.

B. Preprocessing

To prevent training bias, it is crucial to normalize the input data within a specific range when training models using large channel values. The GSE33000 dataset is preprocessed by the following:

1) Filling the missing values (NaN) by using the mean Let G is the set of genes and C is the set of cases that have value, and C' is the complementary set of cases that are missing (denoted as NaN or NULL) For each case ci : If the value of case c_i in gene g_j is missing, then its value is filled using the available values by the following formula:

$$m_{c_i,g_j} = \frac{\sum_{c_i \in C_{g_j}} v_{c_i,g_j}}{|C_{g_j}|} \quad (1)$$

where m_{c_i,g_j} represents the missing value of case i in gene j.

2) Scaling and Normalization: In this step we used the StandardScaler. In this scaler, the mean is subtracted from each sample and then scaled to have a unit variance. The data is re-scaled in a way that ensures it has a mean of 0 and a standard deviation of 1. The standard score z of a sample x is calculated using the formula:

$$z = (x - u)/s \quad (2)$$

Where u represents the mean and s denotes the standard deviation.

C. Wrapper-based Modified Binary Particle Swarm Optimization(WMBPSO)

To address research question 1, which focuses on the utilization of Particle Swarm Optimization (PSO) for gene selection, we employ the Wrapper-based Modified Binary Particle Swarm Optimization (WMBPSO) algorithm. The Particle Swarm Optimization (PSO) algorithm is a metaheuristic optimization technique inspired by the social behavior of bird flocking or fish schooling in a search space. For gene selection, the PSO algorithm seeks to find the optimal set of genes that will maximize the performance of the deep learning model. In order to do this, the PSO algorithm assigns each gene a weight, and then iteratively updates these weights based on the fitness of the current solution. Binary Particle Swarm Optimization (BPSO) is a variant of PSO that is specifically designed for binary optimization problems. The Wrapper-based Modified Binary Particle Swarm Optimization (WMBPSO) algorithm includes some modifications compared to the base BPSO algorithm. Fig. 2 depicts the flowchart of the WMBPSO, and the following are the key modifications:

TABLE I. SUMMARY OF SOME RECENT RESEARCH INVESTIGATING THE PREDICTION OF AD THROUGH THE ANALYSIS OF GENE EXPRESSION DATA, UTILIZING VARIOUS GS TECHNIQUES AND MACHINE AND DEEP LEARNING MODELS

Study	Dataset	GS Method	Model	Performance
[20]	GSE33000 GSE44770 GSE80970	DMP DEG	DNN	Acc = 82.3%
[34]	Proteomic	Belief Network	DBN	Acc > 90%
[35]	GSE33000 GSE5281 GSE122063 GSE97760	NONE	SVM	AUC=0.879
[36]	GSE33000 ADNI	Importance Scores	PINNet	AUC=0.97 F1=0.96
[37]	GSE33000 GSE44770 GSE44771 GSE44768	Chi squared ANOVA MI	SVM	ACC=0.975 AUC=0.972
[38]	GSE33000	DEG LASSO	SVM-RFE RF	AUC=0.954
[39]	GSE63060 GSE63061	LASSO	SVM	Acc= 0.781 AUC=0.859
[40]	GSE63061 DCR	RFE	RF	Acc=0.657 AUC= 0.724
[26]	GSE5281	mRmR WPSO Autoencoder	IDBN	Sensitivity=94.54 Specificity=96.17 Accuracy=96.78 FMeasure=95.09
[27]	GSE63060 GSE63061 ADNI	CFG CFG CFG	DNN SVM DNN	AUC=0.874 AUC=0.804 AUC=0.657
[41]	GSE63060 GSE63061	LASSO	SVM	AUC=0.859 Acc= 0.781
[40]	GSE5281	t-test	SVM	AUC=0.894

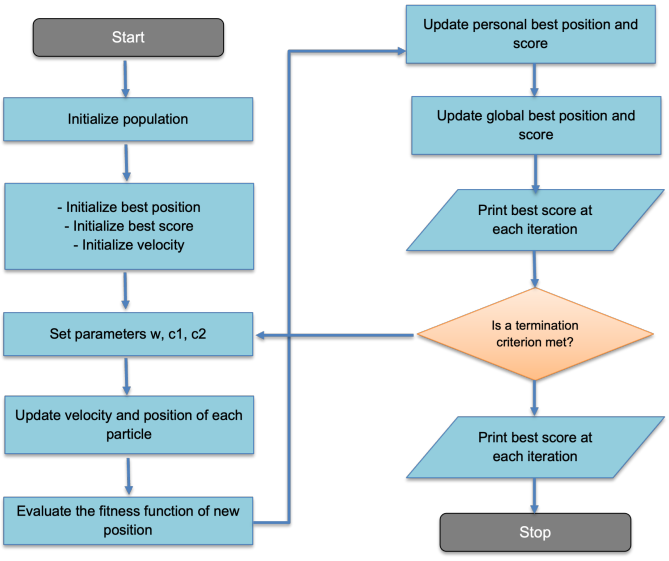

Fig. 2. WMBPSO flowchart.

1) Initialize the population: Each particle represents a subset of genes to be selected.
Let $X = [x_1, x_2, ..., x_n]$ be the binary feature vector for a particle, where n is the total number of features. The value of each feature x_i is either 0 (not selected) or 1 (selected).

2) Evaluate the fitness: Train a CNN-LSTM model using the selected subset of features (genes). Evaluate the fitness of each particle based on the performance metrics (accuracy, F1, AUC, recall, precision) achieved by the model on the AD prediction task.

3) Update the velocity of each particle using Eq. 3:

$$v(t+1) = w * v(t) + c_1 * r_1 * (pbest - x(t)) + c_2 * r_2 * (gbest - x(t)) \quad (3)$$

Here, $v(t)$ represents the current velocity, w is the inertia weight, c_1 and c_2 are acceleration coefficients, r_1 and r_1 are random numbers, $pbest$ represents the personal best position (best subset of genes) for the particle, and $gbest$ represents the global best position (best subset of genes) among all particles.

4) Update the position of each particle by rounding the sigmoid output of the velocity using the equation:

$$x(t+1) = \text{round}\left(\frac{1}{1+\exp(-v(t+1)}\right) \quad (4)$$

5) Apply boundary conditions to ensure that the positions of particles (gene subsets) stay within the valid

range of feature selections.

6) Evaluate the fitness of each particle based on the performance of the CNN-LSTM model using the updated feature subset. Update the personal best position ($pbest$) and fitness for each particle if its fitness improves. Update the global best position ($gbest$) and fitness if any particle achieves a better fitness than the current global best.

D. Hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM)

To answer research question 2, which revolves around enhance deep learning models for gene selection, thereby improving AD prediction, we incorporate the architectures of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) along with WMBPSO for AD prediction into our model. The CNN-LSTM model combines the CNN and LSTM architectures to process both spatial and temporal information in the gene dataset.

1) Convolutional Neural Network (CNN): A Convolutional Neural Network (CNN) is a powerful deep learning technique that has found extensive use in various applications, including image classification, object detection, speech recognition, computer vision, video analysis, and bioinformatics [43]. Unlike traditional neural networks, CNNs are characterized by their deep architecture, incorporating multiple layers [44]. These networks utilize weights, biases, and nonlinear activation functions to process input data effectively. At its core, a CNN consists of convolutional layers, pooling layers, and fully connected layers, forming a comprehensive architecture for feature extraction and classification tasks [45]. Fig. 3 portrays the basic structure of CNN network.

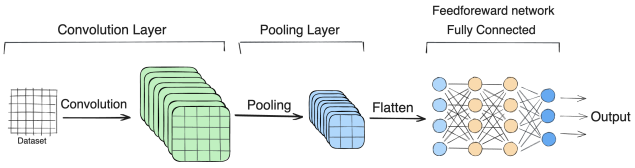

Fig. 3. Basic structure of CNN model.

The CNN's core operation is convolution, which employs convolution kernels to perform convolutions on the inputs. It differs from fully connected structures by leveraging information from adjacent areas of the data matrix. Sparse connections and weight sharing significantly reduce the parameter matrix size. The pooling layer creates feature maps through averaging or taking the maximum value, compressing features and mitigating overfitting. CNN allows for constructing multilayer convolution and pooling operations [46]. Deeper layers extract more abstract features. These abstract features are then merged using a fully connected layer. Finally, classification and regression problems can be addressed using softmax or sigmoid activation functions. In our case, we utilize one-dimensional convolution in CNN to effectively extract spatial features from gene expression data. The convolution layer operates as a filter and subsequently undergoes activation through a non-linear activation function, as described in Eq. 5:

$$a_{i,j} = f\left(\sum_{m=1}^{M}\sum_{n=1}^{N} w_{m,n} \cdot x_{i+m,j+n} + b\right) \tag{5}$$

where a_i, j is the activation, f denotes a non-linear function, w_m, n represents the m×n matrix of convolution kernel weight, $x_{i+m,j+n}$ refers to the activation of the upper neurons and connected to the neuron (i, j), and b represents the bias value. In this study, the convolutional layers utilize rectified linear units (ReLU) for computing the feature maps. The non-linear function associated with ReLU is defined in Eq. 6:

$$\sigma(x) = \max(0, x) \tag{6}$$

Where x is the input value and 0 is a threshold. The ReLU activation function takes an input x and computes the output as follows: If x is greater than or equal to 0, the function returns x. If x is negative, the function returns 0. In essence, the ReLU activation function linearly activates the positive part of the input, while any negative input is turned off (outputting 0).

2) Long Short-Term Memory (LSTM): An LSTM network belongs to the class of recurrent neural networks (RNNs) and offers significant advantages over traditional RNNs, allowing for faster learning and addressing issues such as vanishing and exploding gradients [46]. By incorporating memory blocks and employing a cell state, an LSTM network can effectively store and retrieve long-term information. This is achieved through the utilization of input, forget, and output gates, which enable the network to retain relevant past data and connect it with the present inputs. As a result, LSTM networks are capable of solving complex tasks that were challenging for earlier RNN architectures, making them a valuable tool in various applications [43]. The cell state is the main component of LSTM, which involves three essential processes. The initial step entails deciding the type and quantity of information to be eliminated from the cell state, accomplished through the forget gate. Subsequently, the input gate determines the new information to be incorporated into the cell state. Lastly, the output gate determines the specific information to be outputted.

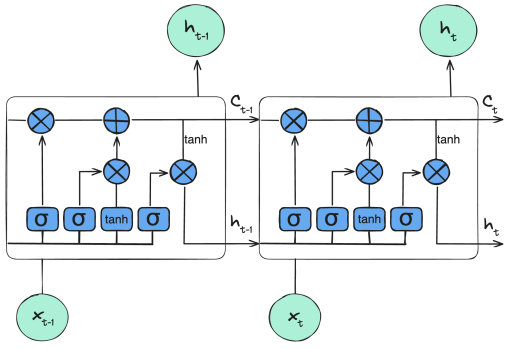

Fig. 4. Typical design of LSTM model.

The LSTM model enhances the original short-term memory unit, represented by h_t, by introducing a memory unit C_t

to preserve long-term memory or the cell state. In Fig. 4, we observe that an LSTM unit receives three inputs at each time step: the current input x_t, the previous state C_{t-1}, and the previous output h_{t-1}. Notably, both x_t and h_{t-1} are simultaneously utilized as inputs for three gates. The LSTM network follows a specific update process, which can be summarized in Eq. 7:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{7}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{9}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{10}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{11}$$

$$h_t = o_t \times \tanh(C_t) \tag{12}$$

where, W_f, W_i, W_c, W_o represent the coefficient matrices, b_f, b_i, b_c, b_o are the matrices of bias, σ is a sigmoid activation function, f_t denotes the forget gate, which regulates the amount of previous memory to be discarded. In contrast, the input gate denoted as i_t determines the amount of new memory $\tilde{C}_t$ to be stored in long-term memory.

E. Performance Evaluation

In this research, we assessed the performance of our approach using the test dataset. We employed five metrics to evaluate the predictive capability: Accuracy, Recall, Precision, F1 score, and AUC. These metrics quantify the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and the following are the details of each metric.

- Accuracy: is a metric that quantifies the ratio of correct predictions $(TP + TN + FP + FN)$ made by the predictor or classifier to the total number of data points $(TP + TN)$ in a dataset. The accuracy metric is calculated using Eq. 13 as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{13}$$

- Recall (also known as sensitivity): measures the ability of a model to correctly identify positive instances out of all the actual positive instances. It quantifies the proportion of true positives that are correctly predicted. The recall metric is calculated by using Eq. 14 as follows:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

- Precision (also known as positive predictive value): measures the proportion of true positives out of all the instances that the model predicted as positive. It focuses on the accuracy of the positive predictions.The precision metric is calculated by using Eq. 15:

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

- F1 score: is a metric that combines precision and recall into a single value. It provides a balanced measure of a model's performance by taking into account both false positives and false negatives. The F1 score is calculated by using Eq. 16:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{16}$$

- AUC: stands for Area Under the Curve provides a single scalar value that summarizes the overall performance of a binary classification model in terms of its ability to rank and discriminate between positive and negative instances. Once the Receiver Operating Characteristic (ROC) curve is created by plotting the true positive rate (TPR), which is synonymous with sensitivity or recall, on the y-axis, and the false positive rate (FPR), calculated as (1 - specificity), on the x-axis , the AUC is computed as the area under this curve.

IV. Results and Discussion

For the experimental work, the code was executed using Python version 3.8.10. The libraries employed were Keras, Tensorflow, and Scikit-learn. The experimental setup included an Intel® CoreTM i5-8250U CPU @ 1.60 GHz, 8 GB of main memory, and a 64-bit OS running Ubuntu 20.04.1 LTS. In this study, the performance of WMBPSO algorithm was investigated for dimensionality reduction of big gene expression data in the context of AD prediction. The reduced dataset was then used as input for a CNN-LSTM model for prediction.

A. Dimensionality Reduction using WMBPSO

The WMBPSO gene selection technique identified a total of 4303 genes as being relevant for AD prediction, achieving an accuracy score of 0.98. These genes were selected based on their discriminatory power and potential contribution to the classification task. This dimensionality reduction significantly improved model performance. The dimensionality reduction achieved through WMBPSO has profound implications. It not only improved AD prediction accuracy but also streamlined the feature set, making it more interpretable.

B. Comparative Analysis of Gene Selection Methods

We conducted a comparative analysis of the WMBPSO-based approach with three commonly used methods for gene selection: the lasso-based approach, the ANOVA method, and a hybrid ANOVA-lasso-PSO method. The performance of the lasso approach compared to WMBPSO-CNN-LSTM is depicted in Fig. 5. The lasso approach achieved an accuracy of 0.920 and an AUC of 0.915. The F1 score, recall, and precision for the lasso approach were 0.929, 0.961, and 0.90, respectively. Fig. 6 reports the scores of ANOVA method, it achieved an accuracy of 0.89 and an AUC of 0.88, which are lower than those obtained by the WMBPSO approach.

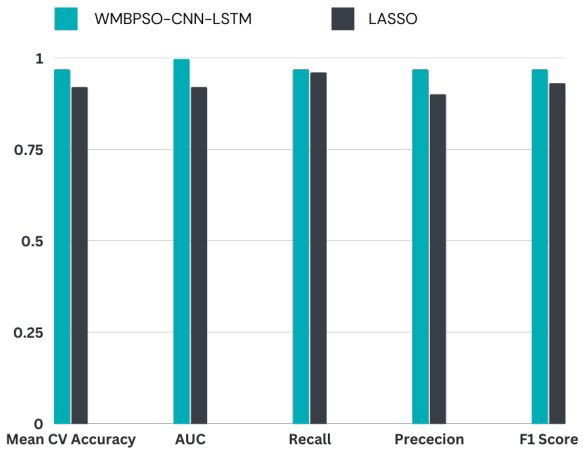

Fig. 5. Comparison of the proposed WMBPSO with lasso method.

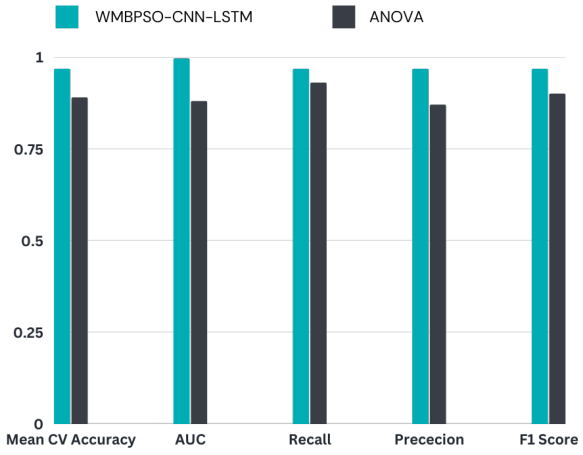

Fig. 6. Comparison of the proposed WMBPSO with anova method.

F1, recall, and precision scores were 0.90, 0.93, and 0.87, respectively.

Fig. 7 describes the performance of the hybrid ANOVA-LASSO-PSO method. This method achieved an accuracy of 0.84 and an AUC of 0.84. Also, the method obtained the value 0.83 for F1 score, recall, and precision. The comparison between the WMBPSO-based approach and the other gene selection methods highlights the effectiveness of the WMBPSO approach and showed competitive performance with a strong AUC score.

C. AD Prediction with WMBPSO-CNN-LSTM

The CNN-LSTM model, trained with the selected genes, yielded highly promising results. The performance of the CNN-LSTM model is evaluated using these selected genes, and the results were highly promising. The results of our analysis are presented in Fig. 8. To train the CNN-LSTM model for AD prediction, we utilized a cross-validation approach with k = 5 folds to further assess the robustness of

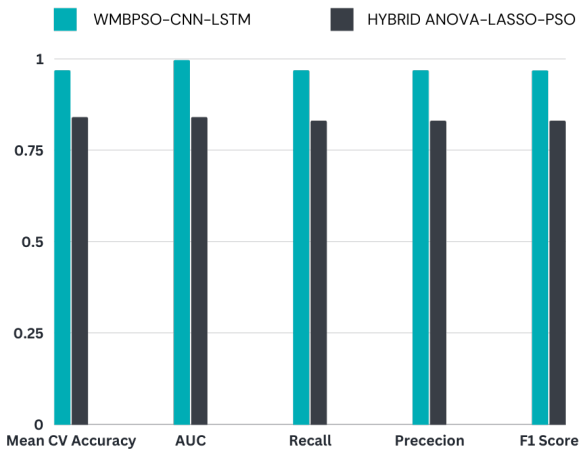

Fig. 7. Comparison of the proposed WMBPSO with hybrid anova-lasso-WMBPSO method.

the model. The model was trained over 10 epochs, with a batch size of 32. The mean cross-validation (CV) accuracy, calculated over multiple iterations, was found to be 0.968. This value indicates a consistently high level of accuracy across different folds of the dataset, reinforcing the reliability of the proposed model. Moreover, the area under the curve (AUC) was used to evaluate the model's performance in terms of its ability to discriminate between AD and non-AD cases. The AUC value obtained was 0.9958, suggesting a high level of discrimination power. Additional performance metrics were computed, the recall value was 0.9677; this indicates that the model effectively identified a high percentage of AD cases. Similarly, the precision value was also 0.9677. This indicates that the model made a high percentage of correct positive predictions. Also, the F1 score was found to be 0.9677. This value indicates a balanced trade-off between precision and recall, demonstrating the model's ability to achieve both high precision and high recall simultaneously.

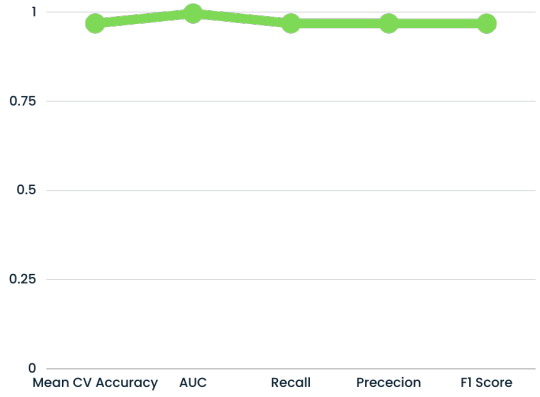

Fig. 8. Performance metrics of the proposed WMBPSO-CNN-LSTM approach.

To evaluate the superiority of the CNN-LSTM model

in combination with the WMBPSO-based gene selection method, we conducted a performance comparison between the WMBPSO-CNN-LSTM model and two other deep learning models, WMBPSO-RNN and WMBPSO-CNN, which also utilized the WMBPSO algorithm for gene selection. For the CNN model, as shown in Fig. 10, the results obtained were 0.94, 0.93, 0.95, 0.93, and 0.94 for mean cross-validation accuracy, AUC, recall, precision, and F1 score, respectively. Fig. 9 portrays the performance of WMBPSO-RNN. The results obtained were 0.84, 0.89, 0.84, 0.85, and 0.85 for mean cross-validation accuracy, AUC, recall, precision, and F1 score, respectively. The comparison among the three models highlights the superior performance of the WMBPSO-CNN-LSTM model in AD prediction using gene expression data. It achieved higher accuracy, AUC, F1 score, and recall compared to both the WMBPSO-CNN and WMBPSO-RNN models. The precision values were comparable between the WMBPSO-CNN-LSTM and WMBPSO-CNN models, indicating similar abilities to identify positive cases accurately. However, the WMBPSO-RNN model exhibited lower performance across all metrics, suggesting it may be less effective in capturing the complex relationships present in the gene expression data for AD prediction.

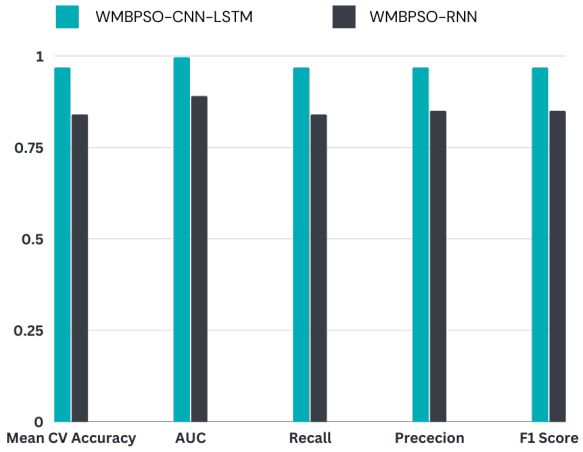

Fig. 9. Comparison of the proposed WMBPSO-CNN-LSTM model with RNN model.

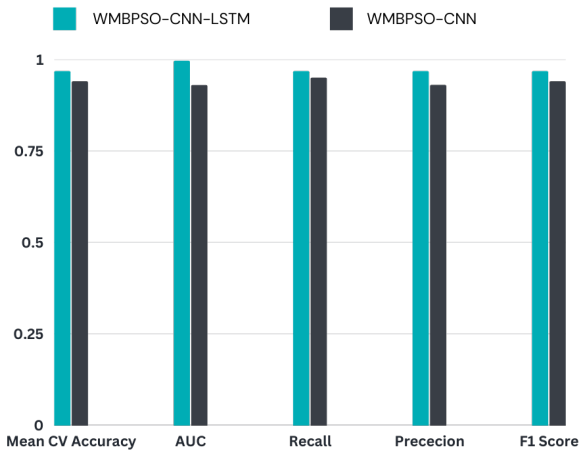

Fig. 10. Comparison of the proposed WMBPSO-CNN-LSTM model with CNN model.

TABLE II. PERFORMANCE EVALUATION VALUES OF DIFFERENT MACHINE LEARNING MODELS COMPARED TO THE PROPOSED WMBPSO-CNN-LSTM

Model	Recall	Precision	F1 Score	Accuracy
Logistic Regression	0.97	0.91	0.94	0.93
Boosted Random Forest	0.83	0.86	0.84	0.83
Decision Tree	0.82	0.79	0.80	0.78
SVM	0.95	0.93	0.94	0.93
KNN	0.80	0.87	0.83	0.82
MLP	0.66	0.97	0.79	0.80
Guassian NB	0.72	0.80	0.76	0.74
ANN	0.97	0.86	0.91	0.90
WMBPSO-CNN-LSTM	**0.97**	**0.97**	**0.97**	**0.97**

Further comparisons with various machine learning methods are presented in Table II.

The remarkable performance of the CNN-LSTM model highlights its effectiveness in handling the reduced dataset. This suggests that by effectively managing dimensionality, we can harness the full potential of deep learning models. In conclusion, according to the findings of this study, the WMBPSO-CNN-LSTM model demonstrated superior performance compared to the WMBPSO-CNN, WMBPSO-RNN, as well as other machine learning models and prevalent feature selection techniques in reducing dimensionality and predicting Alzheimer's Disease using big gene expression data. The WMBPSO-CNN-LSTM model exhibited outstanding AUC, higher accuracy, F1 score, recall, and precision. These results underscore its superior ability to capture relevant features and patterns related to big gene expression data.

D. Implications and Suggestions

Our findings carry significant implications for AD prediction and gene expression analysis:

- The dimensionality reduction techniques employed in this study have the potential to revolutionize AD prediction, making it more interpretable and precise.

- The combination of WMBPSO and CNN-LSTM demonstrates the power of integrating feature selection with deep learning for complex biological data analysis.
- Future research should explore applications of these techniques in other disease prediction tasks and investigate novel approaches for feature selection and deep learning integration.

V. Limitations and Future Work

While our study has yielded promising results, it is essential to acknowledge its limitations and outline potential avenues for future research.

A. Limitations

- Data Size: The study utilized a single gene expression dataset. Future work could explore the integration of multiple datasets to enhance the robustness and generalizability of the model.
- Generalization: Although our model exhibited exceptional performance on the specific dataset, further validation on diverse datasets and populations is necessary to establish its broader applicability.
- Feature Interpretation: While dimensionality reduction improved model performance, interpreting the biological significance of selected genes remains a challenge. Future research should focus on developing methods for gene function interpretation.

B. Future Work

Several potential avenues of future research can be summed up as following: 1) Integrate biological constraints, such as gene pathway information or known gene-disease associations, to guide the gene selection process and ensure that the selected gene groups are biologically meaningful. 2) Group-based Velocity Update of the WMBPSO; modify the velocity update process to consider interactions between feature groups. The velocity update not only involves individual features but also considers the collective behavior of gene groups in the swarm. 3) Expanding the dataset by including additional samples from diverse populations and incorporating longitudinal data to enhance the generalizability and robustness of the WMBPSO-CNN-LSTM model. 4) Integration of Multi-Omics data such as DNA methylation, microRNA expression, or proteomics data, in combination with gene expression data, could provide a chance to validate the behavior of WMBPSO-CNN-LSSTM algorithm on such large-scale combined datasets. 5) Further validation of the WMBPSO-CNN-LSTM model on independent datasets to assess its performance and generalizability in real-world scenarios. 6) Advanced Deep Learning Architectures: Investigating state-of-the-art deep learning architectures and techniques, such as Transformers and attention mechanisms, may further enhance AD prediction accuracy.

VI. Conclusion

In this study, we aimed to leverage the WMBPSO algorithm for dimensionality reduction in big gene expression data. The accuracy score achieved by the WMBPSO algorithm in selecting genes was 0.98, indicating a significantly high level of accuracy. The objective was to develop an accurate AD prediction model using the WMBPSO algorithm in conjunction with a CNN-LSTM deep learning architecture. Through our investigations, we compared the performance of the WMBPSO-CNN-LSTM model with other deep learning and machine learning methods. Also, the performance of WMBPSO was compared with other common feature selection methods. The results obtained demonstrate the effectiveness of the WMBPSO algorithm for dimensionality reduction in big gene expression data. The WMBPSO-CNN-LSTM model achieved outstanding performance in AD prediction, as indicated by the high mean cross-validation accuracy (0.968), AUC (0.9958), F1 score (0.9677), recall (0.967), and precision (0.967). These metrics validate the potential of the WMBPSO algorithm for effectively selecting informative genes and improving the classification accuracy of the AD prediction model. Comparative analyses were conducted with other deep learning models, including WMBPSO-RNN and WMBPSO-CNN, as well as traditional feature selection methods such as ANOVA, lasso, and hybrid approach. The results indicated that the WMBPSO-CNN-LSTM model outperformed these approaches in terms of accuracy, AUC, F1 score, recall, and precision. In conclusion, our study has made significant strides in addressing the challenges of Alzheimer's Disease (AD) prediction using gene expression data. We have demonstrated that effective dimensionality reduction with the WMBPSO algorithm, coupled with the power of CNN-LSTM, can yield highly accurate predictions. Our research contributes by:

- Introducing an innovative approach to gene selection using WMBPSO, which outperforms traditional methods.
- Highlighting the potential of combining feature selection and deep learning for AD prediction.
- Offering valuable insights into the management of high-dimensional biological data.

While there are limitations to our study, such as dataset size and generalization, the future holds promising prospects for improving AD prediction, advancing gene function interpretation, and ultimately aiding in early diagnosis and intervention. Our work underscores the importance of interdisciplinary research at the intersection of bioinformatics and machine learning, paving the way for more precise and reliable disease prediction models in the era of precision medicine.

References

[1] X. Hou, J. Hou, and G. Huang, "Bi-dimensional principal gene feature selection from big gene expression data," *PLOS ONE*, vol. 17, 2022.

[2] J. Yang, Y. jie Li, Q. Liu, L. Li, A. Feng, T. Wang, S. Zheng, A. Xu, and J. Lyu, "Brief introduction of medical database and data mining technology in big data era," *Journal of Evidence-Based Medicine*, vol. 13, pp. 57 – 69, 2020.

[3] K. Wang, W. Wang, and M. Li, "A brief procedure for big data analysis of gene expression," *Animal Models and Experimental Medicine*, vol. 1, pp. 189 – 193, 2018.

[4] S. Zhao, C. K. Hong, C. Myers, D. Granas, M. A. White, J. C. Corbo, and B. A. Cohen, "A single-cell massively parallel reporter assay detects cell-type-specific gene regulation," *Nature Genetics*, vol. 55, pp. 346–354, 2023.

[5] H. Satam, K. Joshi, U. Mangrolia, S. Waghoo, G. Zaidi, S. Rawool, R. P. Thakare, S. Banday, A. K. Mishra, G. Das, and S. K. Malonia, "Next-generation sequencing technology: Current trends and advancements," *Biology*, vol. 12, 2023.

[6] Z. Sha, L. Zhu, Z. Jiang, Y. Chen, and T. Hu, "How complex is the microarray dataset? a novel data complexity metric for biological high-dimensional microarray data," *ArXiv*, 2023.

[7] H. Pan, S. X. Chen, and H. Xiong, "A high-dimensional feature selection method based on modified gray wolf optimization," *Appl. Soft Comput.*, vol. 135, 2023.

[8] M. Oliva, K. Demanelis, Y. Lu, M. Chernoff, F. Jasmine, H. Ahsan, M. G. Kibriya, L. S. Chen, and B. L. Pierce, "Dna methylation qtl mapping across diverse human tissues provides molecular links between genetic variation and complex traits," *Nature genetics*, vol. 55, no. 1, pp. 112–122, 2023.

[9] E. Lin, C.-H. Lin, and H.-Y. Lane, "Deep learning with neuroimaging and genomics in alzheimer's disease," *International Journal of Molecular Sciences*, vol. 22, 2021.

[10] P. Lavanya and E. Sasikala, "Deep learning techniques on text classification using natural language processing (nlp) in social healthcare network: A comprehensive survey," *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, pp. 603–609, 2021.

[11] W. Zhang, H. Li, Y. Li, H. long Liu, Y. min Chen, and X. chen Ding, "Application of deep learning algorithms in geotechnical engineering: a short critical review," *Artificial Intelligence Review*, vol. 54, pp. 5633 – 5673, 2021.

[12] S. Gao and D. Lima, "A review of the application of deep learning in the detection of alzheimer's disease," *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 1–8, 2022.

[13] T. Thaher, M. A. Awad, M. Aldasht, A. F. Sheta, H. Turabieh, and H. K. H. Chantar, "An enhanced evolutionary based feature selection approach using grey wolf optimizer for the classification of high-dimensional biological data," *J. Univers. Comput. Sci.*, vol. 28, 2022.

[14] F. Han, S. Zhu, Q. Ling, H. Han, H. Li, X. Guo, and J. Cao, "Genecwgan: a data enhancement method for gene expression profile based on improved cwgan-gp," *Neural Computing and Applications*, vol. 34, pp. 16 325 – 16 339, 2022.

[15] D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: an overview," *Soft computing*, vol. 22, pp. 387–408, 2018.

[16] R. Poli, J. Kennedy, and T. M. Blackwell, "Particle swarm optimization," *Swarm Intelligence*, vol. 1, pp. 33–57, 1995.

[17] R. C. Eberhart, Y. Shi, and J. Kennedy, *Swarm Intelligence (Morgan Kaufmann series in evolutionary computation)*. Morgan Kaufmann Publishers, 2001.

[18] A. H. Alsaeedi, A. L. Albukhnefis, D. Al-Shammary, and M. Al-Asfoor, "Extended particle swarm optimization for feature selection of high-dimensional biomedical data," *Concurrency and Computation: Practice and Experience*, vol. 34, 2020.

[19] M. Martínez-Ballesteros, J. M. García-Heredia, I. A. Nepomuceno-Chamorro, and J. C. Riquelme-Santos, "Machine learning techniques to discover genes with potential prognosis role in alzheimer's disease using different biological sources," *Information Fusion*, vol. 36, pp. 114–129, 2017.

[20] 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, 2020.

[21] A. Sharma and P. Dey, "A machine learning approach to unmask novel gene signatures and prediction of alzheimer's disease within different brain regions," *Genomics*, vol. 113, no. 4, pp. 1778–1789, 2021.

[22] H. Chen, Y. He, J. Ji, and Y. Shi, "A machine learning method for identifying critical interactions between gene pairs in alzheimer's disease prediction," *Frontiers in Neurology*, vol. 10, 2019.

[23] H. Patel, R. Iniesta, D. Stahl, R. J. Dobson, and S. J. Newhouse, "Working towards a blood-derived gene expression biomarker specific for alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 74, no. 2, pp. 545–561, 2020.

[24] B. Bogdanovic, T. Eftimov, and M. Simjanoska, "In-depth insights into alzheimer's disease by using explainable machine learning approach," *Scientific Reports*, vol. 12, no. 1, pp. 1–26, 2022.

[25] Z. Abbas, H. Tayara, and K. T. Chong, "Alzheimer's disease prediction based on continuous feature representation using multi-omics data integration," *Chemometrics and Intelligent Laboratory Systems*, vol. 223, 2022.

[26] N. Mahendran, P. M. D. R. Vincent, K. Srinivasan, and C.-Y. Chang, "Improving the classification of alzheimer's disease using hybrid gene selection pipeline and deep learning," *Frontiers in Genetics*, vol. 12, 2021.

[27] T. Lee and H. Lee, "Prediction of alzheimer's disease using blood gene expression data," *Scientific reports*, vol. 10, no. 1, 2020.

[28] M. S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "Alzheimer's patient analysis using image and gene expression data and explainable-ai to present associated genes," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–7, 2021.

[29] C. Maj, T. Azevedo, V. Giansanti, O. Borisov, G. M. Dimitri, S. Spasov, A. D. N. Initiative, P. Lió, and I. Merelli, "Integration of machine learning methods to dissect genetically imputed transcriptomic profiles in alzheimer's disease," *Frontiers in genetics*, vol. 10, 2019.

[30] S.-H. Kim, S. Yang, K.-H. Lim, E. Ko, H.-J. Jang, M. Kang, P.-G. Suh, and J.-Y. Joo, "Prediction of alzheimer's disease-specific phospholipase c gamma-1 snv by deep learning-based approach for high-throughput screening," *Proceedings of the National Academy of Sciences*, vol. 118, no. 3, 2021.

[31] J. Park, H. Kim, J. Kim, and M. Cheon, "A practical application of generative adversarial networks for rna-seq analysis to predict the molecular progress of alzheimer's disease," *PLoS computational biology*, vol. 16, no. 7, 2020.

[32] R. Xie, A. Quitadamo, J. Cheng, and X. Shi, "A predictive model of gene expression using a deep learning framework," *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 676–681, 2016.

[33] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," *Computers in Biology and Medicine*, vol. 140, 2022.

[34] N. An, L. Jin, H. Ding, J. Yang, and J. Yuan, "A deep belief network-based method to identify proteomic risk markers for alzheimer disease," *arXiv preprint arXiv:2003.05776*, 2020.

[35] Y. Lai, X. Lin, C. Lin, X. Lin, Z. Chen, and L. Zhang, "Identification of endoplasmic reticulum stress-associated genes and subtypes for prediction of alzheimer's disease based on interpretable machine learning," *Frontiers in Pharmacology*, vol. 13, 2022.

[36] Y. Kim and H. Lee, "Pinnet: a deep neural network with pathway prior knowledge for alzheimer's disease," *arXiv preprint arXiv:2211.15669*, 2022.

[37] 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, 2022.

[38] B. Jin, X. Cheng, G. Fei, S. Sang, and C. Zhong, "Identification of diagnostic biomarkers in alzheimer's disease by integrated bioinformatic analysis and machine learning strategies," *Frontiers in Aging Neuroscience*, 2023.

[39] N. Voyle, A. Keohane, S. Newhouse, K. Lunnon, C. Johnston, H. Soininen, I. Kloszewska, P. Mecocci, M. Tsolaki, B. Vellas *et al.*, "A pathway based classification method for analyzing gene expression for alzheimer's disease diagnosis," *Journal of Alzheimer's Disease*, vol. 49, no. 3, pp. 659–669, 2016.

[40] L. Wang and Z.-P. Liu, "Detecting diagnostic biomarkers of alzheimer's disease by integrating gene expression data in six brain regions," *Frontiers in genetics*, vol. 10, 2019.

[41] X. Li, H. Wang, J. Long, G. Pan, T. He, O. Anichtchik, R. Belshaw, D. Albani, P. Edison, E. K. Green *et al.*, "Systematic analysis and biomarker study for alzheimer's disease," *Scientific reports*, vol. 8, no. 1, 2018.

[42] Ncbi gene expression omnibus (geo). [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi

[43] L. Ma and S. Tian, "A hybrid cnn-lstm model for aircraft 4d trajectory prediction," *IEEE access*, vol. 8, pp. 134 668–134 680, 2020.

[44] H. Nguyen and N. N. Chu, "An introduction to deep learning research for alzheimer's disease," *IEEE Consumer Electronics Magazine*, vol. 10, no. 3, pp. 72–75, 2020.

[45] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.

[46] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis," *Multimedia Tools and Applications*, vol. 78, pp. 26 597–26 613, 2019.