AIMS Mathematics

AIMS Mathematics, 9(7): 17827–17846.
DOI: 10.3934/math.2024867
Received: 29 February 2024
Revised: 10 May 2024
Accepted: 20 May 2024
Published: 24 May 2024

https://www.aimspress.com/journal/Math

Research article

Integrating gene selection and deep learning for enhanced Autisms' disease prediction: a comparative study using microarray data

Mahmoud M. Abdelwahab[1,2,*], Khamis A. Al-Karawi[3,4] and H. E. Semary[1,5]

[1] Department of Mathematics and Statistics, College of Science, Imam Mohammad Ibn Saud Islamic University, Saudi Arabia
[2] Department of Basic Sciences Higher Institute of Administrative Sciences, Osim, Egypt
[3] School of Science, Engineering, and Environment, Salford University, Great Manchester, UK
[4] Diyala University, Baqubah, Diyala, Iraq
[5] Department of Statistics and Insurance, Faculty of Commerce, Zagazig University, Egypt

* **Correspondence:** Email: mmabdelwahab@imamu.edu.sa.

Abstract: In this article, Autism Spectrum Disorder (ASD) is discussed, with an emphasis placed on the multidimensional nature of the disorder, which is anchored in genetic and neurological components. Identifying genes related to ASD is essential to comprehend the mechanisms that underlie the illness, yet the condition's complexity has impeded precise information in this field. In ASD research, the analysis of gene expression data helps choose and categorize significant genes. The study used microarray data to provide a novel approach that integrated gene selection techniques with deep learning models to improve the accuracy of ASD prediction. It offered a detailed comparative examination of gene selection approaches and deep learning architectures, including singular value decompositions (SVD), principal component analyses (PCA), and convolutional neural networks (CNNs). This paper combines gene selection methods (PCA and SVD) with deep learning models (CNN) to improve ASD prediction. Compared to more traditional approaches, the study revealed that its integrated methodology was more effective in improving the accuracy of ASD prediction results through experimentation. There was a difference in the accuracy between the PCA-CNN model, which achieved 94.33% with a loss of 0.4312, and the SVD-CNN model, which achieved 92.21% with a loss less than or equal to 0.3354. These discoveries help in the development of more accurate diagnostic and prognostic tools for ASD, which is a complicated neurodevelopmental disorder. Additionally, they provide insights into the molecular pathways that underlie ASD.

Keywords: autism spectrum disorder; gene selection; microarray technology; convolutional neural networks
Mathematics Subject Classification: 00A05, 00A71, 97K80, 65C60, 92D10

1. Introduction

Autism spectrum disorder (ASD) has been discussed, with an emphasis placed on the multidimensional nature of the disorder, which is anchored in genetic and neurological components. Identifying genes related to ASD is essential to comprehend the mechanisms that underlie the illness; however, the condition's complexity has impeded precise information in this field. In ASD research, the analysis of gene expression data helps choose and categorize significant genes. This research aims to predict ASD using combined gene selection methods, namely principal component analyses (PCA) and singular value decompositions (SVD) with deep learning models, namely convolutional neural networks (CNN). According to statistics from the World Health Organization (WHO), over 0.63% of children are diagnosed with ASD. ASD appears in childhood and continues into adolescence and adulthood, with symptoms usually appearing within the first five years of life [1,2]. ASD imposes substantial healthcare burdens because of its neurodevelopmental characteristics.

ASD is characterized by persistent deficiencies in social interaction and repetitive behaviors, frequently accompanied by notable limitations in communication abilities [3,4]. The illness originates from genetic and neurological elements, leading to difficulties in social interactions, cognitive processing, repetitive behaviors, and communication challenges [5,6]. An early diagnosis of ASD has significant treatment benefits [7,8]. Early detection is crucial in clinical settings, leading to customized therapies designed to improve the welfare of children with ASD and their families [9,10]. The process of diagnosing ASD can be lengthy and expensive, requiring thorough testing. The increase in ASD prevalence worldwide has motivated healthcare providers and researchers to investigate more effective screening methods. ASD is becoming more common in people of all ages, highlighting the importance of early detection to protect an individuals' mental and physical health. Machine learning methods are becoming more popular for predicting diseases, making early identification of ASD possible using many health and physiological factors. This has motivated us to improve the detection and analysis of ASD to enhance treatment methods. Diagnosing ASD is difficult since its symptoms can be similar to those of other mental diseases, making the diagnostic procedure more complex [11,12]. ASD fundamentally relates to human brain development and significantly impacts an individuals' ability to engage in social interactions and communication throughout their lives. Both environmental and genetic factors contribute to ASD onset, with symptoms often emerging by age three and persisting indefinitely. While ASD cannot be fully cured, early detection can temporarily alleviate its effects [13,14]. Despite assuming a genetic basis, scientists have yet to pinpoint the exact causes of ASD, as human genes interact with the environment to influence development. Several risk factors, including low birth weight, having a sibling with ASD, and advanced parental age, contribute to ASD susceptibility [15,16]. As a result, it impacts an individual's entire cognitive, social, emotional, and physical health [17,18]. Both the extent and the intensity of its symptoms are quite variable. Some typical symptoms include difficulty communicating, particularly in social situations, obsessional hobbies, and repeated mannerisms. A comprehensive examination is needed to detect ASD. This also includes a thorough evaluation and a range of assessments performed by child psychologists and other qualified professionals [19–21]. This paper combines gene selection methods (PCA and SVD) with deep learning models (CNN) to improve ASD prediction. Compared to more traditional approaches, the study reveals that its integrated methodology is more effective in improving the accuracy of ASD prediction results through experimentation.

The procedure applied in this work is normalization, which is performed after importing the data using the Min-Max approach. This normalization method guarantees that gene expression values are suitably scaled, thus enabling significant comparisons and minimizing the influence of fluctuations in expression data. Subsequently, gene selection methods pinpoint a subset of pertinent genes. This stage aims to decrease the data's dimensionality and concentrate on genes that powerfully connect with the current classification challenge. Different gene selection techniques might be used based on the specific needs of the analysis. A CNN classifier performs the classification task. The CNN architecture is ideal for analyzing intricate patterns in gene expression data. It utilizes the hierarchical organization of the data and uses convolutional layers to extract crucial information. The remaining research project components are Section 2, the background, and a literature review. Section 3 provides a walkthrough of the microarray technology. Section 4 outlines the experimental setup. Section 5 outlines the methodology, and Section 6 discusses the results. Finally, the conclusion and recommendations for further studies are presented in Section 7.

2. Background and literature review

Autism is a condition currently experiencing a worldwide explosion rate that is both numerous and rising at a very high rate. ASD affects around one child out of every 160, as reported by the World Health Organization (WHO) [22,23]. While some persons with this illness can live independently, others will need 24-hour care and assistance for the rest of their lives. ASD is a neurodevelopmental condition that impacts social interactions and communication abilities [11,24]. Individuals with ASD often experience lifelong challenges in these areas. The causes of ASD are believed to involve a combination of genetic and environmental factors. Symptoms typically manifest around three and persist throughout the person's lifetime. Although there is no known cure for ASD, early detection of symptoms can help manage its effects for a certain period. While scientists have not fully understood the precise causes of ASD, genetic factors are believed to play a role in its development. These genes interact with environmental influences to affect a person's development. Certain risk factors, such as a low birth weight, having a sibling with ASD, or having older parents, can also contribute to the likelihood of developing ASD. An early diagnosis of autism can be quite beneficial because it allows doctors to provide patients with the appropriate treatment at an earlier stage. It can potentially halt any further deterioration of the patient's condition. It would help to cut down on the expenditures associated with a delayed diagnosis over the long term. Therefore, there is a significant need for a screening test instrument that is time-efficient, accurate, and simple. Given the significance of ASD and the absence of a definitive treatment, a pioneering approach called microarray has been utilized to identify the genes responsible for the disease. Biologists utilize microarray technology to assess gene expression levels in specific organisms. A microarray data analysis primarily involves identifying optimal treatments for various diseases and achieving precise medical diagnoses through practical applications involving diverse sets of genes [21,25]. However, microarray technologies yield complex gene expression data that pose issues.

Extraneous or superfluous genes can be eliminated without causing substantial data loss. Analyzing microarray data is challenging due to the large number of genes and samples. This may result in a reduced prediction accuracy and problems with overfitting [26]. Researchers use the "gene selection approach" to tackle this difficulty by identifying the most pertinent collection of genes to create classification models. Gene selection (GS) is choosing a smaller group of relevant genes from a more significant number. By concentrating on this specific group of genes, researchers can acquire useful knowledge about the genetic components of disorders. This approach can decrease the computing expenses and improve the classification effectiveness, particularly for ASD [27,28].

Various methods can be used for gene selection, including a PCA and an SVD. These algorithms are frequently employed as unsupervised methods to examine gene expression microarray data and offer insight into the dataset's fundamental structure. They have been used to create concise representations of gene expression data for classification, especially on extensive datasets [29,30]. The bioinformatics community is actively researching different machine-learning algorithms to diagnose and categorize microarray data, which presents substantial hurdles to classification [18,31]. This study utilizes a deep learning (DL) algorithm to identify ASD using gene expression data. Machine learning encompasses a subset known as deep learning, where algorithms such as CNN utilize abundant data to learn and identify unknown class labels based on gene behaviour patterns in the training set. Additionally, the potential of using a CNN architecture to enhance the predictive accuracy is explored. This study focuses on the methods of gene selection and achieving accurate categorization following the gene selection process.

3. Microarray technology

Microarray technology is a potent and extensively utilized instrument within molecular biology and genetics, facilitating the comprehensive study of gene expression on a large scale. It allows scholars to evaluate the expression levels of several genes at once in a single experiment [32]. Microarray technology attaches brief DNA or RNA sequences (probes) to a stable surface such as a glass slide or a microprocessor. These probes are carefully designed to correspond to particular target genes of interest [33]. This technology has multiple uses, such as gene expression profiling, biomarker discovery, the identification of disease-related genes, and therapeutic target identification. It has dramatically enhanced our understanding of biological processes and has revolutionized genomics, transcriptomics, and personalized medicine. However, it is essential to mention that newer sequencing methods, such as RNA sequencing (RNA-seq), have primarily supplanted microarray technology. RNA-seq provides extensive and quantitative gene expression information and can identify new transcripts and splice variants. Microarray technology is helpful for particular applications and is used in many research contexts [34]. Microarray technology is a standard method used in laboratories to analyze nucleic acids. It entails attaching multiple identified nucleic acid fragments to a solid surface, sometimes called a "chip". The chip is subsequently exposed to DNA or RNA from the studied sample, such as cells or tissue. Fluorescence can be observed by specialized equipment through the complementary base pairing between the sample and the immobilized fragments on the chip. Microarray technology is used in research and clinical contexts to measure gene expression levels and detect specific DNA sequences such as single-nucleotide polymorphisms (SNPs) [35]. Microarray technology is a groundbreaking advancement in genetic analysis that allows for thorough investigations in various disciplines of biology and biomedicine without the need for sequencing. This progress has significantly reduced the high costs of in-depth research [36]. Microarrays provide two essential functions: they enable gene expression analyses by measuring the RNA levels of specific genes in cells and streamline the investigation of SNPs. SNPs have proven instrumental in genome-wide association studies (GWAS), which investigate genetic variations across the entire genome. These methods have been widely applied to study prevalent and less common human diseases and research involving model and diverse organisms globally [37]. Figure 1 depicts the surface of a DNA microarray.

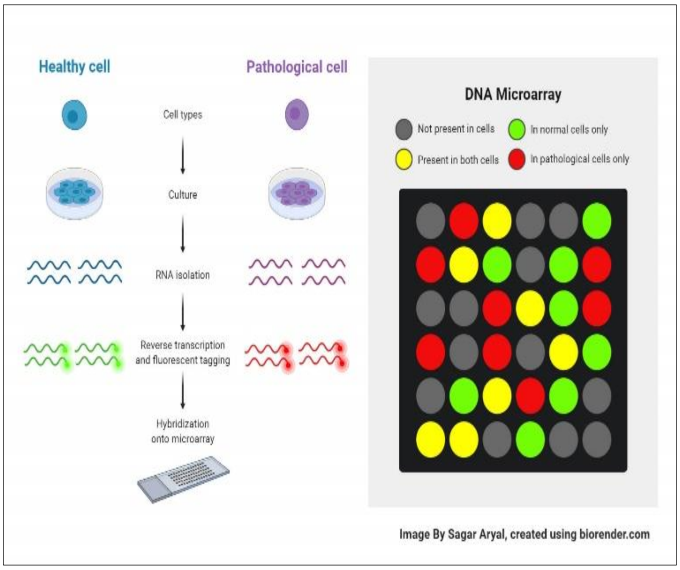

Image By Sagar Aryal, created using biorender.com

Figure 1. DNA microarray's surface [38].

4. Experiment setup

4.1. Dataset

Gene expression data associated with ASD was obtained from the Gene Expression Omnibus (GEO) [39]. The GEO is a public repository managed by the National Center for Biotechnology Information (NCBI), and houses a vast collection of gene expression data. Researchers worldwide contribute to the GEO by depositing their high-throughput functional genomic datasets, including those generated through microarrays and RNA sequencing technologies. The GEO facilitates the sharing, discovery, and analysis of gene expression data, enabling scientists to explore patterns of gene activity across various experimental conditions, tissues, and organisms. By providing free access to diverse datasets, the GEO promotes collaboration, accelerates research, and advances our understanding of gene expression in health and disease. Data was compiled from the GEO database, specifically datasets GSE63060 and GSE63061, which are both publicly accessible sources administered by the NCBI. Subsequently, these two datasets were amalgamated into a singular dataset for analysis. The aggregated dataset encompasses a total of 16,383 genes and 569 samples. Among these samples, 245 individuals were diagnosed with Alzheimer's disease (AD), 142 individuals exhibited symptoms of mild cognitive impairment (MCI), and 182 individuals were classified as healthy controls (CTL). This dataset is valuable for exploring gene expression patterns associated with AD, MCI, and healthy cognition, thus offering insights into the molecular mechanisms underlying these conditions. The GSE6575 dataset is a whole-genome transcriptomics dataset comprised of microarray data from children diagnosed with autism as well as children from the general population. The GSE28521 dataset is an ASD-related dataset comprised of human post-mortem brain tissue samples. The data was preprocessed using GEO2R, and a log2 transformation was performed to achieve normalization. Then, the processed gene expression data were used to calculate the differential gene expression values between the two groups (ASD and control), serving as features in the analysis. Protein-protein interaction data for the genes associated with ASD was obtained from AutDB [40], which is a curated database housing all known direct interactions between proteins, including protein binding, RNA binding, promoter binding, protein modification, auto-regulation, and direct regulation. Interactions involving Homo sapiens were explicitly selected, and redundant interactions were eliminated, resulting in a total of 25,057 unique known interactions involving 12,480 genes. Known ASD genes were collected from the Simons Foundation Autism Research Initiative (SFARI) database [41], and genes that scored from 1 to 3 from the SFARI database were included in the analysis. Additionally, ASD-related genes with high confidence scores (core dataset) were retrieved from the AutismKB 2.0 database [41]. Figure 2. depicts the gene expression data matrix.

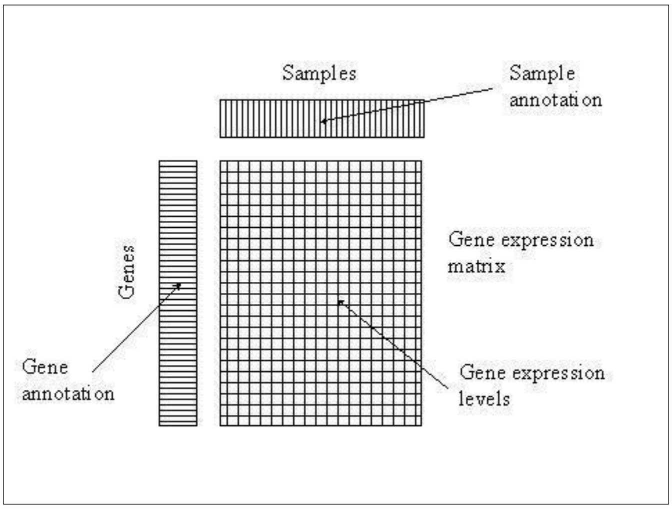

Figure 2. The gene expression data matrix.

4.2. Gene selection

Genetic factors are essential in the development of ASD, which is a complicated neurodevelopmental disease marked by difficulties in social interactions and communication. Advances in machine learning and genetic analyses have enabled the development of more precise and effective models to predict ASD. Gene selection strategies that uncover relevant genes linked to the ASD pathophysiology have emerged as potential tools. By utilizing these methods in conjunction with advanced machine learning algorithms, researchers can improve the precision of ASD prediction models, facilitating early diagnoses and interventions. This work combines gene selection techniques with machine learning methods to enhance ASD disease prediction, potentially leading to significant advancements in ASD research and clinical applications. Microarray experiments facilitate the detection of gene expression variations in various situations, thus producing substantial data. The high dimensionality of microarray data presents issues because most genes are not relevant to the categorization process. Gene selection strategies are used to find a subset of essential genes and decrease data dimensionality [42]. These gene selection strategies aim to find a small number of genes that produce the best results, resulting in an enhanced accuracy of ASD classifiers and decreased computational expenses. The study employed gene selection techniques such as PCAs and SVDs to pinpoint genes that directly affect the disease diagnosis. The methods aim to identify genes that contain significant data linked to the classification job, thus enhancing the efficiency and precision of AD classification.

4.2.1. Principal Component Analysis (PCA)

A PCA is a commonly employed method to reduce dimensions in data analyses and machine learning. Its goal is to convert high-dimensional data into a lower-dimensional space while retaining the key information. A PCA identifies the main components as orthogonal vectors that indicate the directions of maximum variance in the data. A PCA simplifies the dataset's structure and eliminates unnecessary information by only keeping the primary elements that represent the most variability. This approach aids in data visualization and improves the computational efficiency in further analyses. A PCA is beneficial in different fields, such as pattern recognition, image processing, and bioinformatics, as it helps researchers reveal hidden patterns and connections in intricate datasets [43]. By using A PCA as a gene-selection method, valuable gene information can be extracted from large datasets. A PCA allows researchers to identify the most influential genes that contribute significantly to the overall variation in the data, providing a more focused subset for further analysis and classification purposes. In essence, utilizing PCA as an unsupervised gene selection method enables the identification of crucial original genes associated with the principal components. Assuming that a dataset $(X_1, X_2, \ldots\ldots X_m)$ has *m*-dimensional data, a PCA projects m-dimensional data into a k-dimensional sub-space (*k*<*m*). The steps for PCA are described below [44]:

A PCA involves several mathematical steps. Here are the equations used in a PCA:

1) Data standardization

$$Z = \frac{(X - \mu)}{\sigma}, \tag{1}$$

where Z is the standardized data matrix, *X* is the original data matrix, μ is the mean vector of *X*, and σ is the standard deviation vector of *X*.

2) Covariance matrix

$$Z = \frac{1}{n} z^T Z, \tag{2}$$

where *C* is the covariance matrix of the standardized data, z^T is the transposition of the standardized data matrix, and σ is the number of observations.

3) Eigenvalue decomposition

$$C.V = \lambda.V \,, \tag{3}$$

where *V* is the matrix of eigenvectors and *λ* is a vector of eigenvalues.

4) Sorting eigenvalues and eigenvectors

The eigenvalues λ and the corresponding eigenvectors V are sorted in descending order based on the magnitude of the eigenvalues.

5) Selecting principal components

The *k* eigenvectors corresponding to the k largest eigenvalues are selected to form the matrix *P* of principal components.

6) Transforming Data into Principal Components

$$P = Z.V \,, \tag{4}$$

where *P* is the matrix of the principal components, *Z* is the standardized data matrix, and *V* is the matrix of the selected eigenvectors. These equations are the fundamental mathematical operations involved in performing a PCA. They help reduce the data's dimensionality while preserving the maximum variance in the dataset.

4.2.2. Singular Value Decomposition (SVD)

An SVD is a robust linear algebra and its matrix is decomposed into three distinct matrices: the left singular vectors, singular values, and appropriate singular vectors. An SVD creates a matrix by combining singular vectors and values, concisely representing the original data. The left singular vectors represent the connections between rows in the matrix, whereas the appropriate singular vectors represent connections between the columns. The singular values show the significance of each singular vector when depicting the original matrix. An SVD is utilized in signal processing, picture compression, recommendation systems, and dimensionality reduction across several domains. Its capacity to identify significant patterns and simplify the complexity of intricate datasets establishes it as a crucial instrument in data analyses and machine learning [32]. The SVD of a matrix *A* is represented as follows:

$$A = U\Sigma V^{T} \,, \tag{5}$$

where:

- U is an $m \times m$ orthogonal matrix containing the left singular vectors of *A*;
- Σ is an $m \times n$ diagonal matrix containing the singular values of *A*; and
- V^{T} is an $n \times n$ orthogonal matrix containing the suitable singular vectors of *A*.

The singular values *σi* (where *i=1,2,...,r*) are arranged in descending order along the diagonal of *Σ*, where *r* is the rank of matrix *A*. The remaining singular values are zero.

The equations to compute *U*, *Σ*, and V^T are as follows:

1) Compute the eigenvalues and eigenvectors of A^TA. Let the eigenvalues be *λ1,λ2,...,λr,* where *r* is the rank of *A* and the corresponding eigenvectors be *v1,v2,...,vr.*

2) Compute the singular values σ_i as $\sigma_i = \sqrt{\lambda_i}$ for *i=1,2,...,r*.

3) Normalize the eigenvectors to obtain the suitable singular vectors υ_i as

$$\upsilon_i = \frac{1}{\sigma_i} A^T \mu_i$$

for *i=1,2,...,r, where* μ_i are the corresponding eigenvectors of AA^T.

4) Compute the left singular vectors μ_i as $\mu_i = \frac{1}{\sigma_i} A\upsilon_i$ for *i=1,2,...,r.*

These equations yield the matrices $A = U\sum V^T$, constituting the SVD of matrix *A*.

4.2.3. Features modalities

Features and modalities are the various qualities or properties that can be derived from data for analysis, interpretation, and modeling. These modalities cover a broad spectrum of information that mirrors the complexity and richness of real-world phenomena. Feature modalities in data science and machine learning can encompass numerical features for quantitative measurements, categorical features for discrete categories, textual features from text data, image features from visual content, and audio features from sound signals. Temporal features represent trends over time, spatial features define geographical traits, and biological features relate to genetic or physiological aspects. Every modality provides distinct perspectives on the data and necessitates specific extractions, processing, and analysis methods. Comprehending and efficiently using different feature types is crucial to identify patterns, predict outcomes, and obtain practical insights from a wide range of information in various fields [32].

4.3. The deep learning model for ASD

The deep neural network model for autism is an innovative technique in neurodevelopmental disorders. This model aims to analyze complex patterns in genetic data, neuroimaging scans, and behavioral assessments related to ASD using artificial neural networks, specifically deep learning architectures such as CNNs and recurrent neural networks (RNNs). Deep-learning algorithms can identify subtle biomarkers and predictive traits related to ASD by analyzing extensive datasets containing genetic markers, brain scans, and clinical records. The models were created to distinguish between usual brain regions and those showing traits associated with ASD, aiding in early identification, precise diagnosis, and tailored treatment plans. The deep learning model for autism has the potential to enhance our comprehension of the genetic and neurological factors that play a role in ASD. This could lead to personalized interventions and treatments for patients. The deep learning model for autism is a cutting-edge innovation in the field of neurodevelopmental diseases, thus providing new opportunities for early intervention, better results, and improved quality of life for those with ASD [45]. CNNs are ubiquitous among the various architectures used in deep learning.

CNNs have proven highly effective in classifying ASD based on gene expression information.

CNNs are deep learning models that process visual input such as images and movies. CNNs excel at learning hierarchical representations of features from raw pixel data, allowing them to extract spatial patterns and structures within images efficiently. This is accomplished by utilizing convolutional layers, which use adaptable filters to conduct convolutions over the input image, capture nearby spatial relationships, and identify important characteristics. CNNs commonly use pooling layers to decrease the size of feature maps, which helps to reduce computing demands while preserving crucial information. By combining convolutional and pooling layers, CNNs can learn more intricate and sophisticated characteristics as data progresses through the network [46]. CNNs have shown exceptional achievements in many computer vision assignments, such as picture categorization, object recognition, and semantic segmentation. CNNs are precious tools in various industries because they can automatically learn distinctive traits from unprocessed data, such as in healthcare, autonomous driving, security, and entertainment. CNNs continue to lead innovation in deep learning research, advancing visual perception and pattern recognition. Deep CNNs offer benefits beyond gene expression research due to their versatility across several application domains. The advantages of these systems include their capacity to derive significant features from intricate data, process inputs with many dimensions, autonomously acquire hierarchical representations, and attain a top-notch performance in tasks like image recognition, natural language processing, and speech recognition. CNNs combine the selection and classification processes into a single learning entity. During training, these networks can extract and optimize significant features from the raw input data. 2D CNNs have demonstrated a potential in diverse areas, such as early disease detection, structural integrity monitoring, data classification, and personalized healthcare. One significant benefit is that 2D CNNs may utilize real-time data and affordable hardware, thus enabling a straightforward, concise design that mainly emphasizes 2D convolution [46]. Figure 3 depicts the CNN model, displaying its structure and components. 2D CNNs can efficiently analyze and extract features from 2D data, such as images, by integrating these layers, thus allowing for complex analyses and classification tasks. The 2D CNN design's simplicity and efficiency make it a versatile and robust tool for numerous applications across diverse areas. The CNN model in our research used gene expression as the input, considering it as a vector. The model utilizes a 2D kernel to analyze the input vector. The model has two convolutional layers, two thick layers, and a flattening layer. This model is called the 2D CNN for convenience [47]. The CNN architecture is specifically developed to process gene expression data by detecting intricate patterns and correlations. Convolution layers employ filters to extract meaningful information, while dense layers handle more complex learning and classification tasks. The flattening layer transforms the convolution layers' output into a format suitable for the following dense layers.

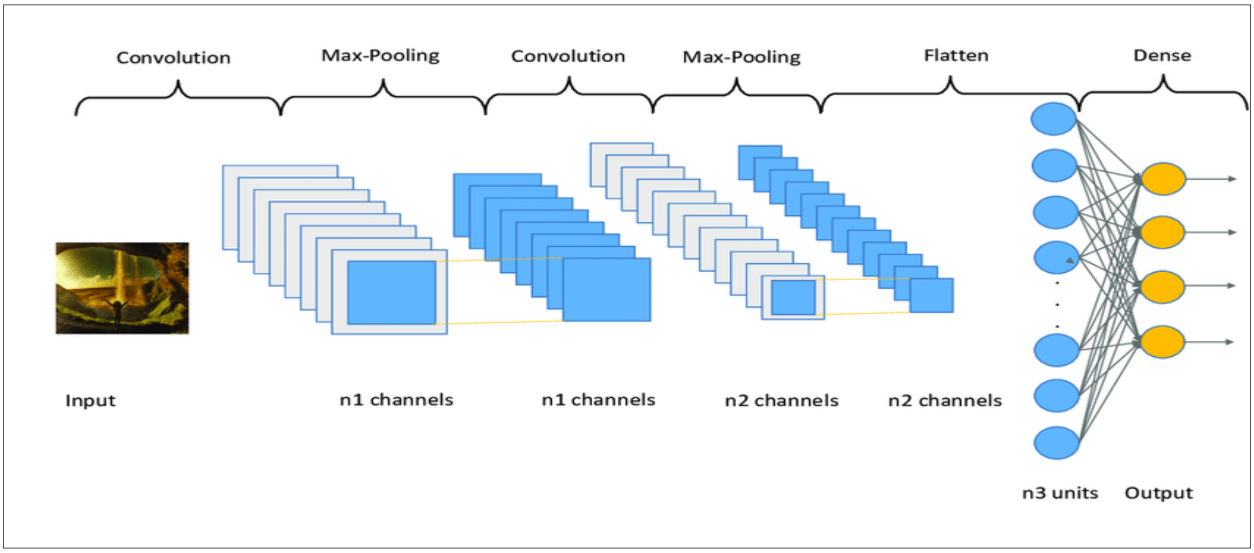

Figure 3. Convolutional neural networks [46].

5. Methodology

This approach includes numerous crucial operations. First, the raw microarray data for ASD is loaded into the system. This data is fundamental to conduct additional analyses. Figure 4 depicts the main steps in the procedure. For example, after importing the data, normalization is performed using the Min-Max approach. This normalization method guarantees that gene expression values are suitably scaled, enabling significant comparisons and minimizing the influence of fluctuations in expression data. Subsequently, gene selection methods pinpoint a subset of pertinent genes. This stage aims to decrease the data's dimensionality and concentrate on genes strongly connected with the current classification challenge. Different gene selection techniques might be used based on the specific needs of the analysis. A CNN classifier performs the classification task. The CNN architecture is ideal to analyze intricate patterns in gene expression data. It utilizes the hierarchical organization of the data and uses convolutional layers to extract crucial information.

5.1. Pre-processing

Microarray data preprocessing is essential to minimize inherent noise and decrease variability in the expression values. [48]. Standardizing datasets is crucial to mitigate substantial discrepancies in the value ranges, particularly following the encoding of nominal values. Without normalization, attributes with more extensive ranges may dominate, potentially biasing the analyses. Moreover, normalization aids the algorithm efficiency by using a narrower range of numbers [49]. Scaling data within the 0 to 1 interval is a standard normalization method, achieved through Eq (7), where *x* represents the original attribute value, *x_Normalized* is the scaled value, *min_a* is the attribute's minimum value, and *max_a* is the maximum value.

$$X_{Normalized} = \left(\frac{x - \min_a}{\max - \min_a}\right) \quad . \tag{6}$$

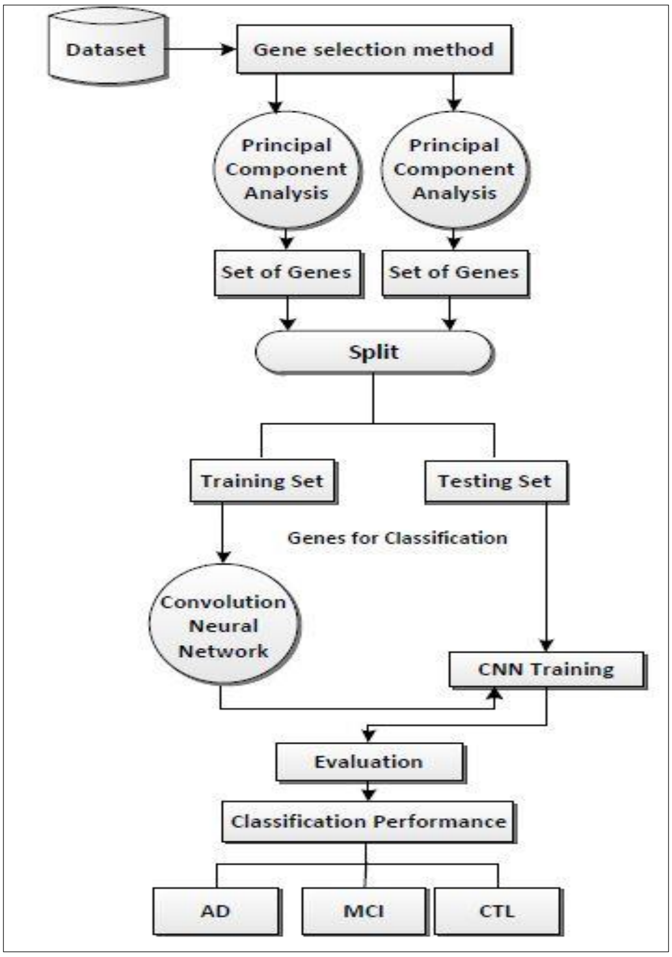

Figure 4. The proposed method.

5.2. Gene selection

Gene selection procedures in the data analysis aim to reduce the computational space dimensionality by selecting a subset of genes from the dataset. Not all genes are informative for the analysis, so selection strategies are applied before machine learning algorithms. By using methods such as PCA and SVD, genes impacting the classification tasks are identified and retained. The study optimized gene sets by applying PCA and SVD for an improved categorization, thus enhancing the result accuracy and interpretability.

5.3. Evaluation measures

Accuracy and loss are crucial performance metrics to evaluate ASD categorization methods. Accuracy, which is a widely used metric, indicates the ratio of correctly predicted observations to the total observations. It offers a simple and intuitive assessment criterion, as depicted in Eq (10). On the other hand, loss measures the dissimilarity between the predicted and actual values, quantifying the model's effectiveness in minimizing errors. Equation (10) outlines a generic formula for calculating loss. These metrics enable researchers to accurately assess the performance and effectiveness of the proposed methodology. They offer dependable insights to evaluate and improve the method.

$$Accuracy = \frac{\text{TP+ TN}}{\text{TP+TN+FP+FN} \times 100}. \tag{7}$$

The approach uses the following abbreviations: FP for false positives, TN for true negatives, TP for true positives, and FN for false negatives. A loss function is used to calculate the error score, where N is the number of genes, Xi ′ is the actual class label, and Xi ′ is a projected label. Cross-entropy measures the loss when categorical outcomes are non-binary and more than two.

$$\text{Loss}=-\text{X}i' \, log2 \ \text{X}i \ \text{N} \ i. \tag{8}$$

6. Results and discussion

Here, we will outline the outcomes of the procedures and their precision in identifying ASD. This study was intended to investigate the intricate genetic pathways involved in ASD. Gene selection methods, namely SVD and PCA, were used to analyze extensive genetic data from microarray datasets to find genes associated with AD. We thoroughly analyzed the gene expression patterns using microarray technology, establishing a strong basis for our inquiry. We aimed to identify genetic variables involved in AD and elucidate their impact on the advancement of the disease. We used gene selection approaches and microarray technology to explore the complex genetic makeup of AD and obtain valuable insights into the genes that influence disease progression. This versatile strategy shows potential to create more focused therapies to address this incapacitating neurodegenerative condition. The "curse of dimensionality" in microarray datasets is a substantial impediment caused by low sample sizes and high dimensionality. Gene selection approaches have been developed as possible solutions to this dilemma. This study employed SVD and PCA to tackle this issue. The study effectively decreased the dataset's complexity by using these methodologies, making it easier to explore the genes linked to ASD more efficiently and meaningfully. Integrating CNNs as classifiers and DL techniques marked significant progress in predicting autism disorder.

The effectiveness of the PCA-CNN model was shown through an empirical examination of the ASD dataset. The model achieved an accuracy rate of 94.33% and a low loss of 0.4312 by utilizing a seven-layer CNN with various configurations. The SVD-CNN model demonstrated an outstanding performance, with an accuracy of 92.21% and a loss of 0.3354. The results highlight how the proposed methodology improves the classification accuracy by choosing a subset of essential genes through gene selection techniques. The study highlights the effectiveness of PCA-CNN and SVD-CNN models in decreasing the gene dimensions and improving the classification accuracy. This dimensionality reduction allows for a more targeted and accurate analysis of the genetic variables associated with ASD. Combining gene selection methods with DL models shows potential to improve the ASD prediction. This progress benefits neurodegenerative disease research and advances precision medicine by providing more precise and personalized methods to diagnose and treat ASD. This work utilized a PCA and an SVD as efficient gene selection techniques. The classification algorithm's performance was assessed by accurately determining the ideal number of genes and categorizing the gene expression data. This step entailed reading and normalizing the gene expression data using the Min-Max method. The gene selection strategies decreased the number of genes to match the available samples. The PCA and SVD techniques were utilized for gene selection, as detailed in Table 1. The gene selection methods suggested in this work reduced the number of genes while improving their informativeness, thereby improving the classification accuracy by removing unnecessary genes. It is important since numerous genes in the original dataset have little influence on determining the class labels. Table 1 summarizes the selected data, outlining the genes selected using the proposed gene selection approaches.

Table 1. Summarize the selected data.

Method	Samples	Genes	Selected genes
PCA			540
SVD	549	15452	490

The suggested approach, which combines a PCA with a CNN model, demonstrates an improved classification of the ASD dataset compared to the raw datasets and other gene selection techniques. The PCA-CNN model achieved an impressive accuracy of 96.60%, as shown in Table 2, which displays the average classification accuracy and loss trends. When paired with the CNN model, the SVD-based gene selection strategy achieves an accuracy of 97.08%, demonstrating a strong performance. The results highlight the efficacy of a PCA and an SVD as gene selection methods when used with the CNN model, thus leading to a notable improvement in classification accuracy. The results provide vital insights into the potential of PCAs and SVDs for gene selection, leading to more accurate and efficient classification models in gene expression studies. Figure 5 shows the performance comparison of the PCA and SVD methods. For measuring the potential of these 4 parameters, a confusion matrix was utilized from the model: F_n (false negative), T_n (true negative), F_p (false negative), and T_p (true positive). Table 3 shows the confusion matrix generated by the suggested model for both methods.

Table 2. Shows the accuracy of PCA-CNN and SVD-CNN method.

Method	Performance metrics	
	Accuracy	Loss
Original dataset	87.72%	0.6732
PCA	94.33	0.4312
SVD	92.21	0.3354

Table 3. Confusion matrix values obtained PCA-CNN and SVD-CNN method.

Model name	TP	FP	TN	FN
PCA	94	7	278	7
SVD	86	9	206	9

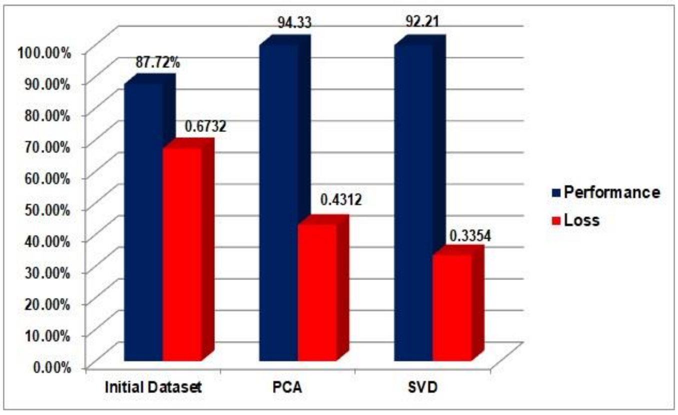

Figure 5. PCA and SVD methods performance comparison.

The study investigated how a PCA and an SVD can improve the performance of CNN models in predicting ASD. The study evaluated the effect of several dataset preparation techniques on the classification accuracy and loss metrics using various cross-validation methods, such as *k*-fold cross-validation. Utilizing a PCA and an SVD resulted in a notable enhancement in accuracy and a decrease in loss compared to the initial dataset. The PCA achieved an accuracy of 96.60%, and the SVD improved it to 97.08%. The results highlight the effectiveness PCAs and SVDs in enhancing the model accuracy to predict AD. Computational efficiency is essential for practical healthcare applications to make timely predictions and to optimize resource usage. The work highlights the significance of assessing the generalizability of AI models on various AD datasets to confirm their dependability and suitability in real-world situations. Studying the biological relevance of genes identified by SVDs and PCAs offers valuable information on AD pathogenesis and possible therapeutic targets. The suggested PCA-CNN and SVD-CNN models outperformed typical machine learning algorithms and state-of-the-art methods in a comparative analysis, indicating their potential to improve the AD prediction accuracy. Ensuring the ethical application of AI in healthcare requires addressing ethical concerns such as model interpretability, bias mitigation, data protection, and transparency. It is essential to balance the potential of AI with ethical norms to foster trust and responsibility in healthcare practices.

7. Conclusions

This work combined gene selection methods (PCA and SVD) with deep learning models (CNN) to improve ASD prediction compared to traditional approaches. The study utilized a convolutional neural network (CNN) model to classify multiclass microarray samples. It aimed to tackle data dimensionality challenges by employing two gene selection techniques: principal component analysis (PCA) and singular value decomposition (SVD). The suggested approach's success was assessed by performance metrics, including accuracy and loss, where cross-entropy classification was a crucial loss function for non-binary categorization issues. The model was fine-tuned using the ADAM optimization technique. The results showed that the suggested method efficiently lowered the high-dimensional data by generating a subset that included pertinent information, thus enhancing the classification accuracy. The strategy improved the categorization efficiency and decreased the processing time, resulting in a smaller subset for diagnosing ASD. The suggested approach, which combined a PCA with a CNN model, demonstrated an improved classification accuracy on the ASD dataset compared to raw datasets and other gene selection techniques. The PCA-CNN model achieved an impressive accuracy of 96.60%. Ongoing research aims to generalize the approach for broader applications beyond ASD and explore alternative gene selection methods and DL architectures to optimize the predictive performance and applicability across various biomedical contexts.

Author contributions

Mahmoud M. Abdelwahab: Conceptualization, Methodology, Reviewing and Editing; Khamis A. Al-Karawi: Software, Data curation, Writing-Original draft preparation; H. E. Semary: Visualization, Investigation, Validation. All authors have read and approved the final version of the manuscript for publication.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors extend their appreciation to the King Salman Centre for Disability Research for funding this work through Research Group No. KSRG-2023-556.

Conflicts of interest

The authors declared no conflicts of interest.

References

1. W. H. Organization, *Autism spectrum disorders*, Regional Office for the Eastern Mediterranean, 2019. Available from: https://iris.who.int/handle/10665/364128.
2. M. M. Abdelwahab, K. A. Al-Karawi, E. Hasanin, H. Semary, Autism spectrum disorder prediction in children using machine learning, *J. Disability Res.*, **3** (2024), 1–9. https://doi.org/10.57197/JDR-2023-0064
3. P. Hlavatá, T. Kašpárek, P. Linhartová, H. Ošlejšková, M. Bareš, Autism, impulsivity and inhibition a review of the literature, *Basal Ganglia*, **14** (2018), 44–53. https://doi.org/10.1016/j.baga.2018.10.002
4. H. Semary, K. A. Al-Karawi, M. M. Abdelwahab, A. Elshabrawy, A review on internet of things (IoT)-related disabilities and their implications, *J. Disability Res.*, **3** (2024), 1–16. https://doi.org/10.57197/JDR-2024-0012
5. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, F. Meneguzzi, Identification of autism spectrum disorder using deep learning and the ABIDE dataset, *NeuroImage: Clinical*, **17** (2018), 16–23. https://doi.org/10.1016/j.nicl.2017.08.017

6. H. Semary, K. A. Al-Karawi, M. M. Abdelwahab, Using voice technologies to support disabled people, *J. Disability Res.*, **3** (2024), 1–8. https://doi.org/10.57197/jdr-2023-0063
7. L. Franz, K. Adewumi, N. Chambers, M. Viljoen, J. N. Baumgartner, P. J. De Vries, Providing early detection and early intervention for autism spectrum disorder in South Africa: stakeholder perspectives from the Western Cape province, *J. Child Adolesc. Mental Health*, **30** (2018), 149–165.
8. K. A. Al-karawi, Real-time adaptive training for forensic speaker verification in reverberation conditions, *Int. J. Speech Technol.*, **26** (2023), 1079–1089. https://doi.org/10.1007/s10772-023-10074-5
9. M. Pagnozzi, E. Conti, S. Calderoni, J. Fripp, S. E. Rose, A systematic review of structural MRI biomarkers in autism spectrum disorder: a machine learning perspective, *Int. J. Dev. Neurosci.*, **71** (2018), 68–82. https://doi.org/10.1016/j.ijdevneu.2018.08.010
10. S. Alenizi, K. A. Al-karawi, Cloud computing adoption-based digital open government services: challenges and barriers, In: X. S. Yang, S. Sherratt, N. Dey, A. Joshi, *Proceedings of Sixth International Congress on Information and Communication Technology*, Singapore: Springer, **216** (2022), 149–160. https://doi.org/10.1007/978-981-16-1781-2_15
11. F. Thabtah, Machine learning in autistic spectrum disorder behavioral research: a review and ways forward, *Inform. Health Soc. Care*, **44** (2019), 278–297. https://doi.org/10.1080/17538157.2017.1399132
12. K. A. Al-Karawi, D. Y. Mohammed, Early reflection detection using autocorrelation to improve robustness of speaker verification in reverberant conditions, *Int. J. Speech Technol.*, **22** (2019), 1077–1084. https://doi.org/10.1007/s10772-019-09648-z
13. U. Frith, F. Happé, Autism spectrum disorder, *Curr. Biol.*, **15** (2005), R786–R790. https://doi.org/10.1016/j.cub.2005.09.033
14. K. A. Al-Karawi, B. Al-Bayati, The effects of distance and reverberation time on speaker recognition performance, *Int. J. Inform. Technol.*, 2024. https://doi.org/10.1007/s41870-024-01789-y
15. H. K. Tripathy, P. K. Mallick, S. Mishra, Application and evaluation of classification model to detect autistic spectrum disorders in children, *Int. J. Comput. Appl. Technol.*, **65** (2021), 368–377. https://doi.org/10.1504/IJCAT.2021.117286
16. K. A. Al-Karawi, D. Y. Mohammed, Improving short utterance speaker verification by combining MFCC and Entrocy in Noisy conditions, *Multimedia Tools Appl.*, **80** (2021), 22231–22249. https://doi.org/10.1007/s11042-021-10767-6
17. K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, M. N. Islam, A machine learning approach to predict autism spectrum disorder, *2019 International conference on electrical, computer and communication engineering (ECCE)*, 2019. https://doi.org/10.1109/ECACE.2019.8679454
18. S. Alenizi, K. A. Al-Karawi, Effective biometric technology used with big data, In: X. S. Yang, S. Sherratt, N. Dey, A. Joshi, *Proceedings of Seventh International Congress on Information and Communication Technology*, Singapore: Springer, **464** (2023), 239–250. https://doi.org/10.1007/978-981-19-2394-4_22

19. J. A. Bastiaansen, M. Thioux, L. Nanetti, C. van der Gaag, C. Ketelaars, R. Minderaa, et al., Age-related increase in inferior frontal gyrus activity and social functioning in autism spectrum disorder, *Biol. Psychiatry*, **69** (2011), 832–838. https://doi.org/10.1016/j.biopsych.2010.11.007
20. S. Alenizi, K. A. Al-Karawi, Internet of things (IoT) adoption: challenges and barriers, In: X. S. Yang, S. Sherratt, N. Dey, A. Joshi, *Proceedings of Seventh International Congress on Information and Communication Technology*, Singapore: Springer, **464** (2023), 217–229. https://doi.org/10.1007/978-981-19-2394-4_20
21. S. Alenizi, K. A. Al-karawi, Machine learning approach for diabetes prediction, In: X. S. Yang, S. Sherratt, N. Dey, A. Joshi, *Proceedings of Eighth International Congress on Information and Communication Technology, ICICT 2023*, Lecture Notes in Networks and Systems, Singapore: Springer, **695** (2023), 745–756. https://doi.org/10.30534/ijiscs/2019/13822019
22. G. Suhas, N. Naveen, M. Nagabanu, N. Kumar, Premature identification of autism spectrum disorder using machine learning techniques, *Adv. Innovations Comput. Program. Languages*, **3** (2021), 1–10.
23. K. A. Al-Karawi, Face mask effects on speaker verification performance in the presence of noise, *Multimedia Tools Appl.*, **83** (2023), 4811–4824. https://doi.org/10.1007/s11042-023-15824-w
24. R. Vaishali, R. Sasikala, A machine learning based approach to classify autism with optimum behaviour sets, *Int. J. Eng. Technol.*, **7** (2018), 18.
25. M. S. Othman, S. R. Kumaran, L. M. Yusuf, Gene selection using hybrid multi-objective cuckoo search algorithm with evolutionary operators for cancer microarray data, *IEEE Access*, **8** (2020), 186348–186361. https://doi.org/10.1109/ACCESS.2020.3029890
26. W. Zhongxin, S. Gang, Z. Jing, Z. Jia, Feature selection algorithm based on mutual information and lasso for microarray data, *Open Biotechnol. J.*, **10** (2016), 278–286. https://doi.org/10.2174/1874070701610010278
27. J. Zahoor, K. Zafar, Classification of microarray gene expression data using an infiltration tactics optimization (ITO) algorithm, *Genes*, **11** (2020), 819. https://doi.org/10.3390/genes11070819
28. K. A. Al-Karawi, S. T. Ahmed, Model selection toward robustness speaker verification in reverberant conditions, *Multimedia Tools Appl.*, **80** (2021), 36549–36566. https://doi.org/10.1007/s11042-021-11356-3
29. M. Babu, K. Sarkar, A comparative study of gene selection methods for cancer classification using microarray data, *2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2016, 204–211. https://doi.org/10.1109/ICRCICN.2016.7813657
30. K. A. Al-Karawi, D. Y. Mohammed, Using combined features to improve speaker verification in the face of limited reverberant data, *Int. J. Speech Technol.*, **26** (2023), 789–799. https://doi.org/10.1007/s10772-023-10048-7
31. L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, 856–863.

32. D. H. Lim, Principal component analysis using singular value decomposition of microarray data, *Int. J. Math. Comput. Phys. Quantum Eng.*, **7** (2013), 1390–1392.
33. M. Dufva, Introduction to microarray technology, In: M. Dufva, *DNA microarrays for biomedical research*, Methods and Protocols, **529** (2009), 1–22. https://doi.org/10.1007/978-1-59745-538-1_1
34. U. R. Müller, D. V. Nicolau, *Microarray technology and its applications*, Springer, 2005. https://doi.org/10.1007/b137842
35. R. Govindarajan, J. Duraiyan, K. Kaliyappan, M. Palanisamy, Microarray and its applications, *J. Pharm. Bioallied Sci.*, **4** (2012), S310–S312. https://doi.org/10.4103/0975-7406.100283
36. R. Kothapalli, S. J. Yoder, S. Mane, T. P. Loughran, Microarray results: how accurate are they, *BMC Bioinf.*, **3** (2002), 22. https://doi.org/10.1186/1471-2105-3-22
37. D. H. Blohm, A. Guiseppi-Elie, New developments in microarray technology, *Curr. Opin. Biotech.*, **12** (2001), 41–47. https://doi.org/10.1016/S0958-1669(00)00175-0
38. M. M. Abdelwahab, K. A. Al-Karawi, H. E. Semary, Deep learning-based prediction of Alzheimer’s disease using microarray gene expression data, *Biomedicines*, **11** (2023), 3304. https://doi.org/10.3390/biomedicines11123304
39. S. Abrahams, D. E. Arking, D. B. Campbell, H. C. Mefford, E. M. Morrow, L. A. Weiss, et al., SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs), *Mol. Autism*, **4** (2013), 36. https://doi.org/10.1186/2040-2392-4-36
40. C. Yang, J. Li, Q. Wu, X. Yang, A. Y. Huang, J. Zhang, et al., AutismKB 2.0: a knowledgebase for the genetic evidence of autism spectrum disorder, *Database*, **2018** (2018), bay106. https://doi.org/10.1093/database/bay106
41. L. Kolberg, U. Raudvere, I. Kuzmin, J. Vilo, H. Peterson, gprofiler2--an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler, *F1000Res.*, **9** (2020), ELIXIR-709. https://doi.org/10.12688/f1000research.24956.2
42. H. Ahmed, H. Soliman, M. Elmogy, Early detection of Alzheimer’s disease based on single nucleotide polymorphisms (SNPs) analysis and machine learning techniques, *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, 2020, 1–6. https://doi.org/10.1109/ICDABI51230.2020.9325640
43. M. Lenz, F. J. Müller, M. Zenke, A. Schuppert, Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data, *Sci. Rep.*, **6** (2016), 25696. https://doi.org/10.1038/srep25696
44. N. Parveen, H. H. Inbarani, E. N. S. Kumar, Performance analysis of unsupervised feature selection methods, *2012 International Conference on Computing, Communication and Applications*, 2012, 1–7. https://doi.org/10.1109/ICCCA.2012.6179181
45. Y. Zhang, J. M. Gorriz, Z. Dong, Deep learning in medical image analysis, *J. Imaging*, **7** (2021), 74. https://doi.org/10.3390/jimaging7040074
46. M. Mostavi, Y. Chiu, Y. Huang, Y. Chen, Convolutional neural network models for cancer type prediction based on gene expression, *BMC Med. Genomics*, **13** (2020), 44. https://doi.org/10.1186/s12920-020-0677-2

47. S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, D. Inman, 1D convolutional neural networks and applications: a survey, *Mech. Syst. Signal Process.*, **151** (2021), 107398. https://doi.org/10.1016/j.ymssp.2020.107398
48. T. Ragunthar, S. Selvakumar, Classification of gene expression data with optimized feature selection, *Int. J. Recent Technol. Eng.*, **8** (2019), 4763–4769. https://doi.org/10.35940/ijrte.b1845.078219
49. J. R. Vergara, P. A. Estévez, A review of feature selection methods based on mutual information, *Neural Comput. Appl.*, **24** (2014), 175–186. https://doi.org/10.1007/s00521-013-1368-0