Journal of *Personalized Medicine*

MDPI

Article

From Data to Insights: Machine Learning Empowers Prognostic Biomarker Prediction in Autism

Ecmel Mehmetbeyoglu [1,2,*], Abdulkerim Duman [3], Serpil Taheri [2,4], Yusuf Ozkul [2,5] and Minoo Rassoulzadegan [2,6]

[1] Department of Cancer and Genetics, Cardiff University, Cardiff CF14 4XN, UK
[2] Betul-Ziya Eren Genome and Stem Cell Center, Erciyes University, Kayseri 38280, Turkey; serpiltaheri@hotmail.com (S.T.); ozkul@erciyes.edu.tr (Y.O.); minoo.rassoulzadegan@unice.fr (M.R.)
[3] School of Engineering, Cardiff University, Cardiff CF24 3AA, UK; dumana@cardiff.ac.uk
[4] Department of Medical Biology, Erciyes University, Kayseri 38280, Turkey
[5] Department of Medical Genetics, Erciyes University, Kayseri 38280, Turkey
[6] Inserm-CNRS, Université Côte d'Azur, 06107 Nice, France
* Correspondence: mehmetbeyogluse@cardiff.ac.uk

Abstract: Autism Spectrum Disorder (ASD) poses significant challenges to society and science due to its impact on communication, social interaction, and repetitive behavior patterns in affected children. The Autism and Developmental Disabilities Monitoring (ADDM) Network continuously monitors ASD prevalence and characteristics. In 2020, ASD prevalence was estimated at 1 in 36 children, with higher rates than previous estimates. This study focuses on ongoing ASD research conducted by Erciyes University. Serum samples from 45 ASD patients and 21 unrelated control participants were analyzed to assess the expression of 372 microRNAs (miRNAs). Six miRNAs (miR-19a-3p, miR-361-5p, miR-3613-3p, miR-150-5p, miR-126-3p, and miR-499a-5p) exhibited significant downregulation in all ASD patients compared to healthy controls. The current study endeavors to identify dependable diagnostic biomarkers for ASD, addressing the pressing need for non-invasive, accurate, and cost-effective diagnostic tools, as current methods are subjective and time-intensive. A pivotal discovery in this study is the potential diagnostic value of miR-126-3p, offering the promise of earlier and more accurate ASD diagnoses, potentially leading to improved intervention outcomes. Leveraging machine learning, such as the K-nearest neighbors (KNN) model, presents a promising avenue for precise ASD diagnosis using miRNA biomarkers.

Keywords: autism; miRNAs; machine learning

check for updates

Citation: Mehmetbeyoglu, E.; Duman, A.; Taheri, S.; Ozkul, Y.; Rassoulzadegan, M. From Data to Insights: Machine Learning Empowers Prognostic Biomarker Prediction in Autism. *J. Pers. Med.* **2023**, *13*, 1713. https://doi.org/10.3390/jpm13121713

Academic Editors: Richard E. Frye, Richard G. Boles, Daniel Rossignol and Shannon Rose

Received: 10 November 2023
Revised: 10 December 2023
Accepted: 12 December 2023
Published: 15 December 2023

1. Introduction

Autism Spectrum Disorder (ASD) presents a challenge both for society and science, as it affects a large number of children with alterations in communication, social interaction, and behavior with repetitive patterns [1–4]. The Autism and Developmental Disabilities Monitoring (ADDM) Network performs continuous monitoring of ASD. It tracks the occurrence and traits of ASD in children who are 8 years old and whose parents or guardians reside in 11 ADDM Network locations throughout the United States. In 2020, the estimated prevalence of ASD in 8-year-old children was 1 in 36, which amounted to approximately 4% of boys and 1% of girls. These estimates are higher than the previous ADDM Network estimates from 2018 to 2020 [5,6].

ASD impacts the most intricate brain functions, prompting important inquiries from neurobiologists, social scientists, and geneticists. The high correlation between monozygotic twins suggests a genetic cause, but the wide range of phenotypes, even among twins, suggests that epigenetic mechanisms play a significant role [7–9]. In the future, precision in the diagnosis of characters and the comprehension of ASD will need to be included in the molecular study of brain functions, which is the highest level of biological complexity.

However, in the short term, identifying the cause of ASD could result in beneficial medical advancements, expert advice for both medical professionals and parents, and, most importantly, early detection and support for affected children to promote their optimal development. This is crucial for physicians responsible for their care.

Scientific and medical institutions worldwide are currently studying large to extremely large groups of patients, with one recent study including 35,584 individuals, including 11,986 with ASD, their families, and controls [10,11]. Various genetic research avenues are currently being explored using cutting-edge genome analysis techniques. More than 100 mutated loci with protein-coding alterations were identified in the genomes of patients with autism. Moreover, the detection of tandem DNA repeats through genome-wide analysis has indicated their expansion in individuals with autism, with many expressed in early developmental stages in neurons and neuronal precursors [12]. Recently, a study reported the role of RNA in controls of CGG repeats in FXS cells with the possibility of FMR1 reactivation in patients' cells [13]. Uncovering the mechanisms, extent, variability, and connection of the disorder with other neurodevelopmental abnormalities is a high priority. However, none of the genetic variations found were present in all or even most patients, thus disqualifying them as primary determinants of the disorder.

The recent publication from Erciyes University pertains to different areas of ongoing ASD research. The initial phase involved creating a group of 45 patients from 37 separate families, with each family having genetically related individuals (parents, siblings) and 21 unrelated control participants, resulting in a total of 187 serum samples [14]. The study then concentrated on the group of genes that encode miRNAs, which are 22nt-long non-coding regulatory RNAs recognized as significant factors in cellular differentiation processes [15]. The Kayseri cohort study analyzed the expression of 372 miRNAs using miRNA PCR Array profiles and found that 6 miRNAs (miR-19a-3p, miR-361-5p, miR-3613-3p, miR-150-5p, miR-126-3p, and miR-499a-5p) were significantly down-regulated in the serum of all 45 ASD patients, compared to healthy controls. Furthermore, our group extended this investigation by creating autism mice models. In these models, we observed a similar downregulation of the identified miRNA profiles, reinforcing the translational relevance of our findings [14]. The fact that all patients and autism mice models showed the same profile of reduced expression of these six miRNAs indicates that this feature is a significant intrinsic characteristic of ASD. The downregulation of these six miRNAs raises a question: can miRNAs be a reliable biomarker in the clinic as a computer-aided diagnostic model?

Artificial intelligence (AI) can be useful for big data with different parameter interpretation processes [16]. The theories and methods used to create automated machines that can mimic human intelligence are collectively referred to as AI. Recent advances in technology have ushered in a new era of precision medicine that combines machine learning (ML) and biological science to analyze diseases using big data. AI includes deep learning, neural networks, and machine learning. Owing to these strategies, computers are now able to come up with wise selections. ML is commonly employed in the study of cancer and is gaining popularity in the identification and treatment of cancer [17].

The overall goal of ML is to build automated tools that can quickly perform classifications from previously observed examples by designing or learning functional dependencies between selected inputs (features) and outputs (classes) [18]. Therefore, the diagnosis of an individual with multi-factorial diseases such as ASD, which aims to translate the knowledge of the characteristics extracted from the miRNA features into meaningful groups (groups of healthy individuals or with ASD), is fundamentally an ML problem.

The purpose of this study is to assess whether ML algorithms, trained on a comprehensive set of miRNA findings collected in advance, can differentiate between healthy individuals and those with ASD and predict the likelihood of ASD in individual patients with accuracy.

The miRNA data we have generated (q-RT-PCR) have the potential to be a valuable tool for diagnosing ASD and other medical conditions. The hypothesis is that ML algorithms can successfully identify patterns in miRNA data that distinguish healthy individuals from those with ASD, resulting in a high accuracy in predicting the disease in testing data. Therefore, this study aims to investigate the performance of ML models in analyzing miRNA data and evaluate their potential for clinical application. Our results show K-nearest neighbor (KNN) models, in particular, achieved outstanding accuracy and overall performance, indicating their potential utility in practical applications within the domain of interest.

2. Materials and Methods

2.1. Ethics Statement

The methods involving human participants in this study were conducted in accordance with the applicable guidelines and regulations set forth by the Ethics Committee of Erciyes University School of Medicine. The study received validation from the committee on 20 September 2011, with the assigned committee number: 2011/10. Furthermore, the study was approved by the hospital. Prior to participating in the study, all parents provided written informed consent, which was obtained and validated by the Ethics Committee of Erciyes University School of Medicine (committee number: 09-20-2011, ethics committee number: 2011/10). Patient selection criteria were also adhered to, and a comprehensive explanation of the study was provided to both the participants and their parents before their enrollment.

2.2. Data Set

The study included a total of 217 participants from both multiplex and simplex families. Among them, there were 45 subjects with ASD aged between 2 and 13 years (31 males and 14 females), along with 21 age- and sex-matched typical control subjects aged between 3 and 16 years (10 males and 11 females). Additionally, 33 healthy siblings aged between 1 and 20 years (17 males and 16 females) were included in the study. All participants were of Turkish origin.

For miRNA analysis, plasma samples were collected from patients and healthy controls, and miRNA PCR Array profiles were extracted, containing measurements for 372 miRNA expression levels.

The processing and analysis of the raw data were accomplished using R software (version 4.2.0, Limma package) to identify differentially expressed miRNAs; adjusted $p < 0.05$ was set as the threshold. The differential miRNA expression analysis was performed as explained previously [14].

2.3. Data Pre-Processing

The ML technique includes a crucial step called data pre-processing. Data selection, noise filtering, imputation of missing values, resampling (SMOTE) to handle imbalanced data, feature selection, and normalization are all components of large-scale data preparation.

2.3.1. Feature Selection

Normalized microarray expression values extracted from serum samples served as the primary input for rigorous feature selection. The data set was carefully partitioned into three distinct parts, with two out of the three parts allocated for the training data set. Subsequently, a comprehensive analysis of the correlation matrix between miRNAs was executed using the Spearman's correlation method exclusively within the training data set.

Following the identification of pertinent miRNAs, a robust repeated stratified k-fold validation framework was employed to rigorously apply the selection criteria. This entailed a partitioning scheme of three splits and repeating the process five times within the confines of the training set. This methodological approach was designed to circumvent the pitfalls of data overfitting and ensure the reliability of the results.

2.3.2. Lasso

The LASSO (least absolute shrinkage and selection operator) linear regression model was employed in this study to enhance prediction accuracy. The standard five-fold cross-validation technique was used to assess and validate the model's performance. In recent years, LASSO regression analysis has gained prominence as a valuable tool for identifying diagnostic or prognostic features, and it has been widely utilized in various research studies [19].

Repeated 5-fold cross-validation was used to split the data into training and test sets using Python (v3.8.8) and a Python libraries such as "Sklearn". The Sklearn library was used for feature selection, LASSO implementation, and data normalization/scaling.

Overfitting was prevented by using K-fold cross-validation. We also measured the model accuracy using 5-fold cross-validation. The data were equally divided into 5 parts (repeated 5 times), of which 4 were used for training and 1 for evaluation. The cross-validation was carried out once more after rearranging the data set. Following training, only the test data were used to evaluate the model's prediction.

2.4. Machine Learning Algorithms

Recent research on the prediction and classification of disease using gene expression data has made extensive use of many algorithms created to solve classification problems in machine learning [20,21]. In machine learning, the general classification process entails training a classifier to detect patterns effectively from provided training samples and then classifying test data using the trained classifier. The classification process employs illustrative classification methods such as the k-nearest neighbor (KNN), support vector machine (SVM), random forest (RF) classifier, and decision tree.

KNN, logistic regression, SVM, random forest, and decision tree were used to create diagnostic models for autism utilizing five ML algorithms using Python (v3.8.8) and Python libraries "xgboost", "TensorFlow", "sklearn", and "imblearn".

One of the most popular techniques for memory-based induction is the KNN. Based on similarity metrics, KNN retrieves the k closest vectors from the reference set from an input vector and uses the labels of the k closest neighbors to decide on the input vector's label [22]. Euclidean distance and Pearson's correlation coefficient were utilized as the similarity metric.

Logistic regression is a vital statistical technique employed in various fields to model the relationship between binary outcomes and one or more predictor variables [23]. This versatile method plays a crucial role in binary classification tasks. Logistic regression provides a way to estimate the probability of an event occurring based on the values of independent variables, making it a fundamental tool for decision-making and risk assessment [24].

SVM is an effective technique for creating classifiers. To enable the prediction of labels from one or more feature vectors, it seeks to establish a decision boundary between two classes [25]. The hyperplane, a decision boundary, is oriented to be as far away from each class's nearest data point as is technically possible. These closest points are called support vectors.

An ensemble of learning techniques for classification, regression, and other tasks known as random forest is a machine learning algorithm that works by building a large number of decision trees during the training phase and then displaying the classes (for classification problems) or mean predictions (for regression problems) of the individual trees [26]. The RF algorithm creates every decision tree it contains, training each one with a portion of the problem's data. It randomly selects N records from the data and then uses multiple trees that were built and trained using the bagging principle to learn from them. This makes sure that every decision tree has a unique perspective on the issue. After all the decision trees have been trained, RF votes on all of them and then make decisions according to the classification or regression problem that needs to be solved [27].

Decision trees are recognized as a highly effective data mining methodology, widely embraced in diverse academic disciplines [28]. Their appeal is attributed to their user-friendly characteristics, interpretability, and robustness, even when confronted with missing data. It is worth noting that decision trees demonstrate versatility by accommodating both discrete and continuous variables, which may function as either target or independent variables [29].

Each ML model shares the same training and test sets. As shown in Figure 1, we trained the machine learning models and measured the training time at the training stage;

then, we tested their accuracy and measured the prediction accuracy at the testing stage. To compare various ML methods and parameter settings, the model performance measures such as prediction accuracy and standard deviation were calculated, as well as the area under the receiver operating characteristic curve.

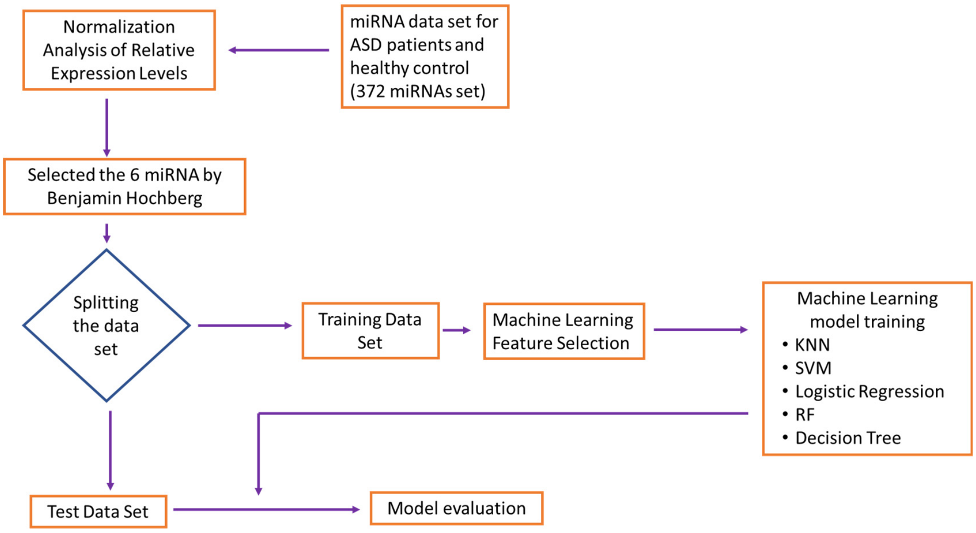
Normalization Analysis of Relative Expression Levels
miRNA data set for ASD patients and healthy control (372 miRNAs set)
Selected the 6 miRNA by Benjamin Hochberg
Splitting the data set
Training Data Set
Machine Learning Feature Selection
Machine Learning model training
• KNN
• SVM
• Logistic Regression
• RF
• Decision Tree
Test Data Set
Model evaluation

Figure 1. The flowchart of the machine learning method for identification of miRNAs. K-nearest neighbors (KNN); support vector machine (SVM); random forest (RF).

2.5. Validation Approaches

The performance of each machine learning algorithm underwent evaluation through two distinct forms of cross-validation. Initially, a random 5-fold cross-validation was executed by randomly assigning each patient to one of three groups. In each iteration of the cross-validation process, one group was set aside as the test set, while the classifiers were trained on the remaining data.

Subsequently, considering that the data originated from an independent cohort, the accuracy of all machine learning models was thoroughly assessed. Performance metrics, including sensitivity and specificity, were diligently evaluated to provide a comprehensive understanding of the algorithms' effectiveness.

2.6. Pathway Analysis and Gene Ontology (GO)

To gain a deeper insight into the functional significance of a chosen miRNA, a pathway prediction analysis was conducted. To accomplish this, the bioinformatics tool DIANA-mirPath (version 3) was employed [30]. This computational approach enabled the identification of all genes and cancer-related molecular pathways influenced by the selected miRNAs, providing valuable insights into their potential roles in biological processes.

3. Results

3.1. Subject Characteristics

Peripheral blood samples were collected from 45 patients with ASD and 37 normal healthy volunteers, the clinical characteristics of which are published in Ozkul et al., 2013 and Rassoulzadegan et al., 2022 [14,31]. In the context of a cohort study, the study encompassed the quantification of 372 miRNAs' expression levels. After the implementation of the Benjamin Hochberg method, it became evident that 180 miRNAs displayed statistically significant disparities among the study groups. To enhance the precision of the selection process, a predefined threshold was introduced to identify miRNAs that exhibited downregulation by at least 90% in comparison to the control group's expression levels. Employing this stringent selection criterion, the study identified merely six miRNAs that exhibited a statistically significant downregulation.

Figure 2 shows the expression levels of six miRNAs in the plasma that were prospectively measured. These six miRNAs including miR-3613-3p, miR-150-5p, miR-19a-3p, miR-361-5p, miR-499a-5p, and miR-126-3p were significantly decreased in patients with ASD and their family.

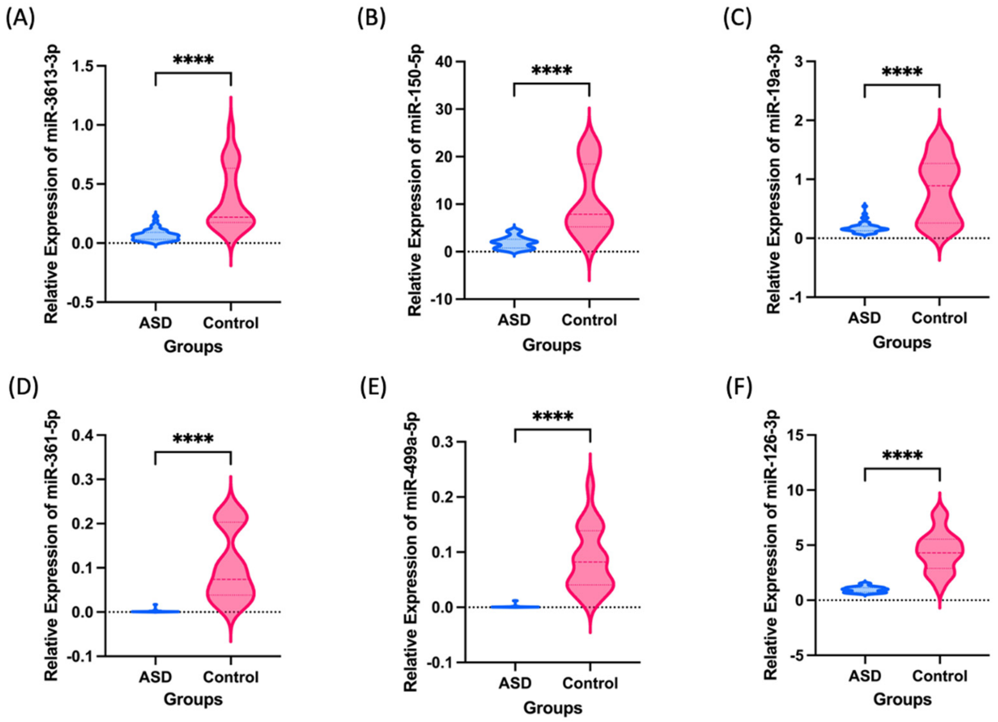

Figure 2. Box violin plots of relative miR-3613-3p (**A**), miR-150-5p (**B**), miR-19a-3p (**C**), miR-361-5p (**D**), miR-499a-5p (**E**), miR-126-3p (**F**); expression level in ASD (n = 45) and healthy control (n = 21). Asterisks denote significant differences by unpaired t-test, **** < 0.0001.

3.2. ROC Analysis

We used data produced by Ozkul et al., 2020 [14] to assess the reliability and performance of the ML program. The study evaluated the diagnostic potential of differentially expressed miRNAs in plasma as biomarkers for the diagnosis of ASD using a receiver operating characteristics (ROC) ML algorithm. As depicted in Figure 1 we implemented the flowchart as our approach to the identification of miRNAs using a machine learning method. The ROC curves of individual miRNAs showed area under curve (AUC) scores ranging from 0.894 to 0.998, indicating good discriminatory power (Figure 3 and Table 1). The miRNA with the highest AUC score, miR-361-5p, was found to have greater potential in discriminating between ASD patients and the control group compared to other miRNAs. A clear separation of six individual miRNAs between ASD and control groups appears with AUC values with a 95% confidence interval. miR-361-5p was the most reproducible diagnostic biomarker for ASD in serum based on the ROC curves.

Table 1. ROC analysis of miRNAs. AUC: Area under the curve, CI: Confidence interval, SEM: Standard error of the mean.

miRNAs	AUC	SEM	95% CI	*p*
miR-3613-3p	0.956	0.023	0.9109–1	<0.0001
miR-150-5p	0.946	0.033	0.8803–1	<0.0001
miR-19a-3p	0.894	0.043	0.8094–0.979	<0.0001
miR-361-5p	0.998	0.001	0.9954–1	<0.0001
miR-499a-5p	0.956	0.001	0.9954–1	<0.0001
miR-126-3p	0.988	0.0123	0.9646–1	<0.0001

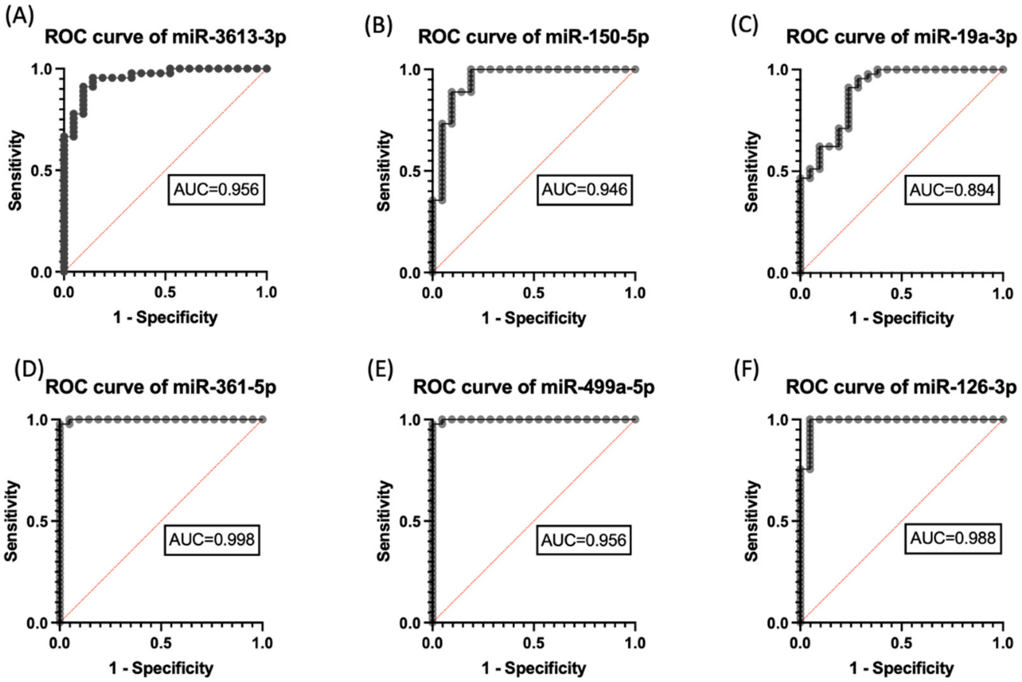

Figure 3. Evaluation of the diagnostic effectiveness of autism-specific miRNAs' biomarkers. ROC curve of (**A**) miR-150-5p, (**B**) miR-19a-3p, (**C**) miR-361-5p, (**D**) miR-499a-5p, (**E**) miR499a-5p, and (**F**) miR-126-3p with area under the curve (AUC) values. ROC, receiver operating characteristics.

3.3. Machine Learning

After obtaining data, feature selection was applied to data in which relevant features or variables were selected from the pre-processed data. This is an important step that can greatly affect the accuracy of an ML model. In this study, the six significantly expressed miRNAs were selected as features for the ML models based on a previous study [14].

Pearson's correlation analysis was conducted on our data set, leading to the simplification of the model and the reduction of overfitting through the removal of highly correlated features. This enhancement not only improves a model's interpretability but also refines its overall performance. Specifically, we focused on the six miRNAs that were previously identified as potential biomarkers for ASD diagnosis. Enrichment scores for four of six miRNAs had a significant Pearson's correlation, and there were significant differences between patients and controls (Welch's *t*-test, $p < 0.05$) (Figure 4). Pearson's r score was between +0.80 and 1, indicating that the module has a high level of predictive power. These results support the potential utility of these miRNAs as biomarkers for ASD diagnosis and disease activity assessment.

We found that miR-499a-5p and miR-361-5p exhibited the greatest positive effect value and were strongly correlated (Figure 4). The strong correlation between miR-499a-5p and miR-361-5p implies that they may be functionally related or co-regulated in the context of ASD. Additionally, we observed a strong positive correlation between miR-150-5p and miR-3613-3p. It suggests that these two miRNAs may also have a cooperative or synergistic effect in influencing the development and manifestation of ASD. Based on these correlations, we selected one of the miRNA pairs that were strongly correlated with each other to be used for diagnostic purposes. These results may allow for the selection of a smaller subset of miRNAs that are most informative for ASD diagnosis. After highly correlated feature elimination, thereby reducing the risk of overfitting, miR-3613-3p, miR-19a-3p, and miR-126-3p were selected. As a second step of feature selection, LASSO regression analysis was performed to determine which miRNAs are the major ASD-associated miRNAs having a significant impact on the detection of ASD. Which miRNA/miRNAs have a high impact on ASD detection?

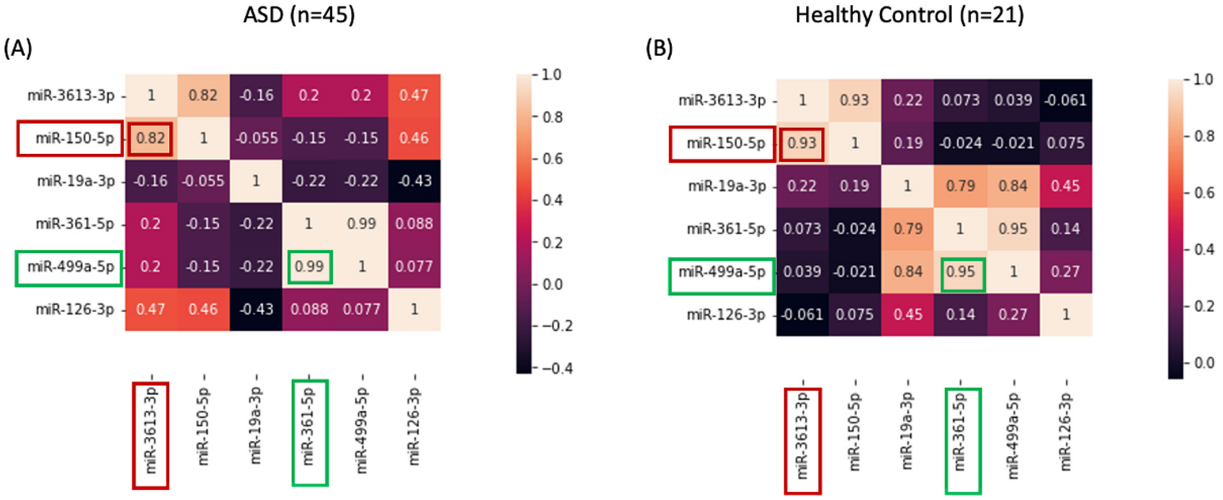

Figure 4. Pearson's correlation analysis of six miRNAs. (**A**) Correlations between miRNAs in the ASD patients. (**B**) Correlations between six miRNAs in the healthy control.

After selecting miR-126-3p as the most promising miRNA for ASD detection, we trained the ML models on the training data. Five different algorithms were used: KNN, logistic regression, SVM, RF, and decision trees, and their results are presented in Table 2. The ML models were trained and subsequently assessed using k-fold cross-validation. This approach was employed to gauge the accuracy and generalizability of the models. Multiple metrics are available to assess a models' performance, including accuracy, precision, recall, F1 score, and others. In our study, we primarily focus on the accuracy score as the key metric of interest.

Table 2. Model performance of machine learning classifiers on the validation set. The table summarizes the results obtained for each classifier, including accuracy, SD, specificity, and sensitivity.

Models	Accuracy	SD	Specificity	Sensitivity
KNN	0.94	0.06	0.81	1
Logistic regression	0.94	0.06	0.81	1
Support vector machine	0.94	0.05	0.82	1
Random forest classifier	0.94	0.05	0.84	0.98
Decision tree	0.96	0.05	0.9	0.99

The analysis of various machine learning models yielded accuracy scores exceeding 90% across the board. Notably, the KNN model exhibited an impressive accuracy of 0.94, indicative of its proficiency in generating accurate predictions. Furthermore, the logistic regression model demonstrated a perfect sensitivity score of 1, signifying its exceptional ability to correctly identify positive instances, while achieving a specificity of 0.81, underscoring its precision in recognizing negative instances.

The SVM model showcased excellent performance, boasting an accuracy of 0.94, coupled with a high sensitivity of 1 and a perfect specificity of 0.82. The RF classifier also achieved a commendable accuracy at 0.94, complemented by a sensitivity of 0.98 and a perfect specificity of 0.84.

The top-performing model, however, emerged as the decision tree, with the highest accuracy score of 0.96. This exceptional accuracy was matched by an excellent sensitivity of 0.99, further bolstered by a commendable specificity of 0.9. These findings underscore the robust predictive capabilities of these machine learning models within the context of the study, with the RF model notably demonstrating superior overall performance.

The confusion matrix can also be seen in Figure 5.

To validate our models, we applied the trained models on unseen test data set. KNN using miR-126-3p exhibited the best predictive performance among the five ML models

(Table 3). The selected biomarker discriminated healthy and ASD groups with over 90% accuracy for all ML models. These findings suggest that reduced levels of circulating miR-126-3p on the selected six miRNAs can serve as an effective biomarker for ASD detection.

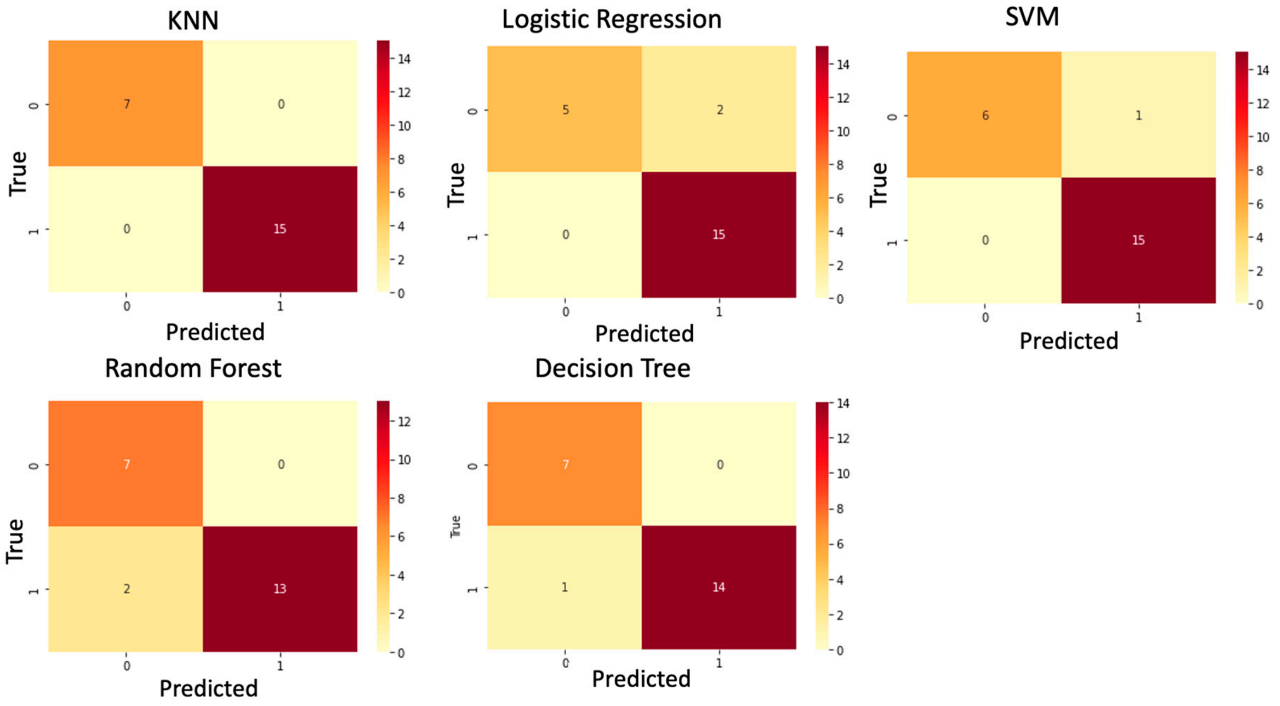

Figure 5. Confusion matrix for all ML models. The confusion matrix was used to compare the different machine learning algorithms.

Table 3. Model performance of machine learning classifiers on the independent cohort. The table summarizes the results obtained for each classifier, including accuracy, SD, specificity, and sensitivity.

Models	Accuracy	Specificity	Sensitivity
KNN	1	1	1
Logistic regression	0.90	0.71	1
Support vector machine	0.95	0.85	1
Random forest classifier	0.90	1	0.86
Decision tree	0.95	1	0.93

3.4. Functional Analysis

We conducted an extensive functional enrichment analysis of miR-126-3p, focusing on its involvement in specific molecular activities based on Gene Ontology (GO). The results, presented in Figure 6A, unveil noteworthy enrichments that provide insights into the regulatory capacity of miR-126-3p. Our comprehensive analysis demonstrates that miR-126-3p is intricately involved in 25 distinct pathways, highlighting its significant role in cellular regulation. Notably, one of the most prominent pathways in which miR-126-3p displayed a high enrichment is the mTOR (mammalian target of rapamycin) signaling pathway, which governs critical cellular processes. Within the mTOR signaling pathway, miR-126-3p exhibited a substantial enrichment and was notably associated with three target genes: Insulin Receptor Substrate 1 (IRS1), Phosphoinositide-3-Kinase Regulatory Subunit 2 (PIK3R2), and Vascular Endothelial Growth Factor A (VEGF-A) (Figure 6B). This observation points toward a compelling regulatory role for miR-126-3p within the mTOR pathway, which, in turn, influences key cellular processes. IRS1 is a central mediator in insulin signaling, influencing cell growth and metabolism [32]. Its interaction with miR-126-3p in the context of the mTOR pathway suggests a regulatory link between miR-126-3p and metabolic processes. PIK3R2 is a regulatory subunit of the PI3K complex [33], a key player in the mTOR pathway. The interaction between miR-126-3p and PIK3R2 suggests a potential role in modulating PI3K-mediated signaling. VEGF-A is a critical factor in angiogenesis, which is intricately connected with mTOR activity [33]. The association between miR-126-3p and VEGF-A hints at miR-126-3p's potential role in regulating angiogenic processes.

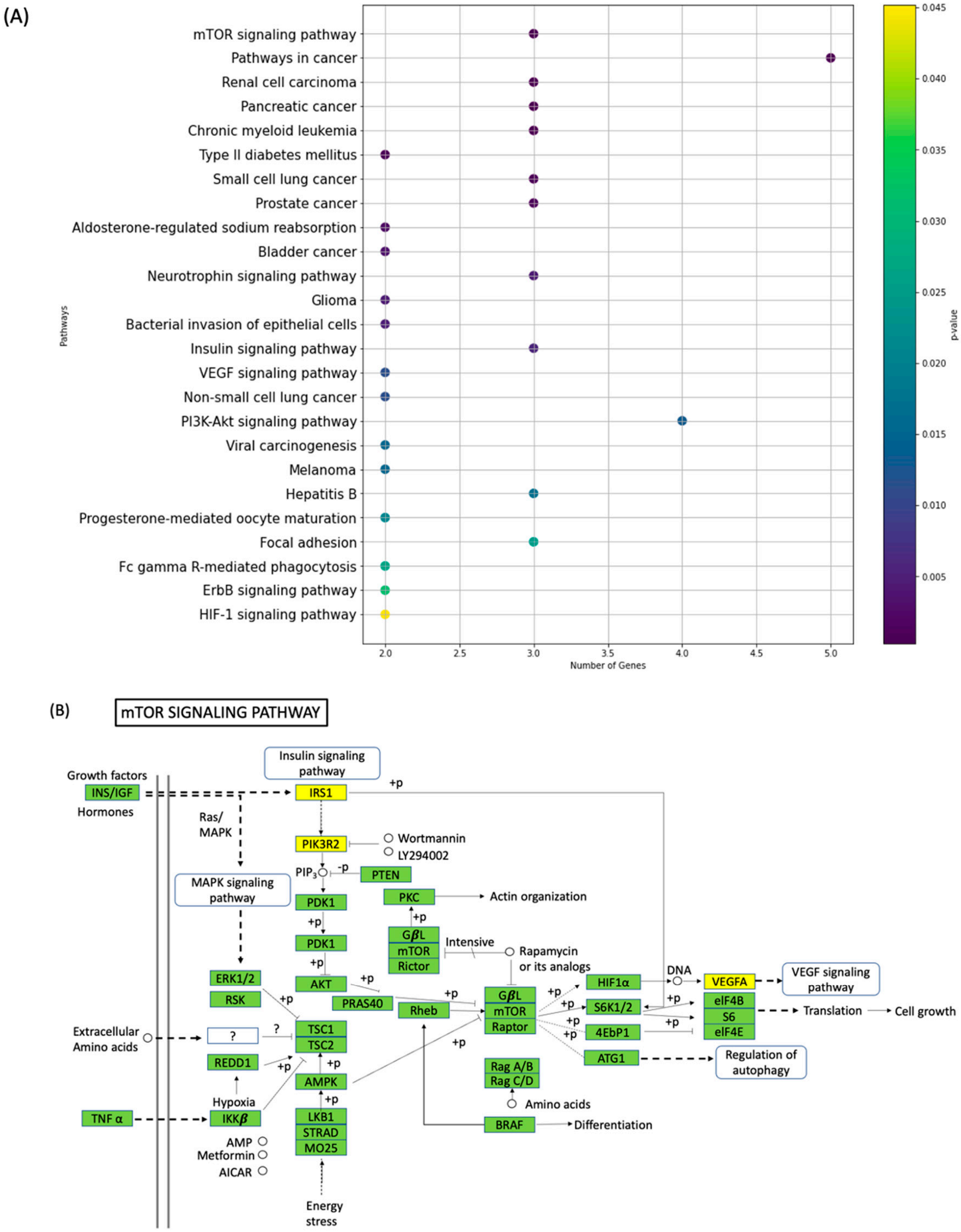

Figure 6. Functional enrichment analysis of target genes based on gene ontology regarding molecular function and (**A**) KEGG pathway analysis. (**B**) mTOR pathway analysis IRS1, PIK3R2, and VEGFA.

4. Discussion

miRNAs have attracted increasing attention in recent years as a crucial tool for gaining a deeper understanding of a wide range of biological processes, including the emergence of many human diseases such as cancer, cardiovascular problems, and neurological disorders [34–36]. The quantitative measurement of small miRNAs can be established routinely in the clinical laboratory, but the interpretation of the results requires reliable and automatic methods. In our previous study, we investigated the miRNA expression profiles of the parents of autism patients, which provided important insights into the potential genetic and epigenetic factors underlying the development of ASD [14]. The rationale behind conducting this study is to identify a reliable diagnostic biomarker panel for ASD using miRNAs. There is currently a need for a non-invasive, reliable, and cost-effective biomarker for ASD diagnosis, as current diagnostic methods are subjective and time-consuming.

These findings suggest that alterations in miRNA expression patterns may contribute to the risk of developing ASD, and a further investigation of these miRNAs may be valuable in understanding the underlying mechanisms of ASD pathogenesis. Building on these previous findings, the current study aimed to identify specific miRNAs that could serve as diagnostic biomarkers for ASD, which is a challenging condition to diagnose accurately and early. The significance of the findings is that this study identified the potential for miR-126-3p to serve as a diagnostic biomarker for ASD with a high accuracy. This can potentially lead to earlier and more accurate diagnoses of ASD, allowing for earlier interventions and improved outcomes. Additionally, the use of ML methods, such as KNN, provides a promising approach to accurately predicting ASD diagnosis using miRNA biomarkers.

We have demonstrated that our KNN model can accurately estimate the risk of ASD by using a data set that contains miRNA levels. This study used a large data set to develop the computer-aided diagnostic model. Our results show that KNN may have the potential to aid physicians in discriminating between healthy patients and patients with ASD.

The potential functional implications of miR-126-3p in molecular pathways offer valuable information for understanding the molecular interactions. We showed that miR-126-3p highly enriched in mTOR, a serine/threonine kinase, plays a pivotal role in regulating cellular responses to nutrient availability, influencing critical processes such as protein synthesis and cell growth and proliferation [37–39]. Dysregulation in the mTOR pathway has been linked to the development of autism, contributing to aberrant synaptic protein synthesis and associated symptoms, including macrocephaly, seizures, and learning deficits [40–42]. Notably, approximately 8–10% of autism cases have been associated with abnormalities in the mTOR signaling pathway [42]. Moreover, a substantial proportion of autism predisposition genes directly or indirectly intersect with mTOR signaling activity [43]. A study showed that in autism mice models, synaptic alterations dependent on mTOR lead to a distinctive functional hyperconnectivity signature, and this effect can be reversed by inhibiting mTOR [44]. These findings collectively underscore the central role of mTOR signaling in autism. Particularly, it has already been shown that miR-126 directly targets the PIK3R2 gene [33]. The enrichment of miR-126-3p within the mTOR signaling pathway opens new avenues for research. Another study has demonstrated the role of miR-126-3p in influencing the stress response to mild traumatic brain injury (mTBI). The findings suggest that changes in the expression levels of miR-126-3p during embryonic development might have enduring impacts on the adult brain and metabolism [45]. This suggests that miR-126-3p's regulatory role can significantly impact key cellular processes, such as cell growth, metabolism, and angiogenesis. Further in-depth investigations are warranted to unravel the precise molecular interactions between miR-126-3p and its target genes within the mTOR pathway.

It is worth noting that our study has some limitations. First, since only one center provided the data, the study should be repeated with more samples. To confirm the efficacy of the models and differentially expressed miRNAs, additional data using the same sample preparation processes and techniques from other nations and ethnic groups are needed, and future studies with larger sample sizes are needed to validate our results. Second, the sample size was relatively small. Additionally, our study only focused on circulating miRNAs, and it remains to be determined if other types of miRNAs, such as tissue-specific miRNAs, can be used as diagnostic biomarkers for ASD.

Additionally, the strength of these findings is further emphasized by the fact that ASD is notoriously difficult to diagnose, with a lack of reliable and objective biomarkers. The identification of a panel of miRNAs that can accurately distinguish ASD from healthy subjects has important implications for early detection and intervention, ultimately improving outcomes for individuals with ASD.

In conclusion, our study provides evidence that miR-126-3p has a high potential as a diagnostic biomarker panel for ASD. These findings have important implications for the early detection and treatment of ASD, which can improve outcomes for individuals with this disorder. Further studies are needed to validate these findings and to explore the underlying mechanisms by which these miRNAs contribute to the pathogenesis of ASD. In conclusion, our results imply that circulating miRNAs have the potential to serve as biomarkers for the identification of ASD. The miR-126-3p, in all ML models, showed the potential to be a useful diagnostic tool to advance the diagnosis of ASD.

Author Contributions: Conceptualization, E.M. and A.D.; methodology, A.D.; software, A.D.; validation, A.D. and E.M.; formal analysis, E.M. and A.D.; resources, A.D.; data curation, A.D.; writing—original draft preparation, E.M.; writing—review and editing, A.D., M.R. and E.M.; visualization, A.D.; supervision, M.R. and S.T.; project administration, M.R.; funding acquisition, Y.O. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Cardiff University.

Institutional Review Board Statement: All methods involving human participants were performed in accordance with the relevant guidelines and regulations of the Ethics Committee of Erciyes University School of Medicine and validated on 20 September 2011, with committee number: 2011/10. This study was approved by the hospital. All parents gave written informed consent before participation (Ethics Committee of Erciyes University School of Medicine 09-20-2011 committee number: 09-20-2011 , ethics committee number: 2011/10).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent was obtained from the patient(s) to publish this paper.

Data Availability Statement: The data used in this study were collected by Ozkul and colleagues [14]. The research article titled "A heritable profile of six miRNAs in autistic patients and mouse models" was published in Nature Scientific Reports in 2020 [14]. The data set utilized in the study contains miRNA (microRNA) levels measured from human samples. Detailed information about the miRNA data set can be found in the Supplementary Materials of the article [14].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lord, C.; Elsabbagh, M.; Baird, G.; Veenstra-vanderweele, J. Seminar Autism spectrum disorder. *Lancet* **2018**, *392*, 508–520. [CrossRef] [PubMed]
2. Iakoucheva, L.M.; Muotri, A.R.; Sebat, J. Getting to the Cores of Autism. *Cell* **2019**, *178*, 1287–1298. [CrossRef] [PubMed]
3. Carroll, L.; Braeutigam, S.; Dawes, J.M.; Krsnik, Z.; Kostovic, I.; Coutinho, E.; Dewing, J.M.; Horton, C.A.; Gomez-nicola, D.; Menassa, D.A. Autism Spectrum Disorders: Multiple Routes to, and Multiple Consequences of, Abnormal Synaptic Function and Connectivity. *Neuroscientist* **2021**, *27*, 10–29. [CrossRef] [PubMed]
4. Miles, J.H. Genetics in Medicine Autism spectrum disorders—A genetics review. *Genet. Med.* **2011**, *13*, 278–294. [CrossRef] [PubMed]
5. Maenner, M.J.; Shaw, K.A.; Bakian, A.V.; Bilder, D.A.; Durkin, M.S.; Esler, A.; Furnier, S.M.; Hallas, L.; Hall-Lande, J.; Hudson, A.; et al. Prevalence and Characteristics of Autism Spectrum Disorder among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018. *MMWR Surveill. Summ.* **2021**, *70*, 1–16. [CrossRef] [PubMed]

6. Maenner, M.J.; Warren, Z.; Williams, A.R.; Amoakohene, E.; Bakian, A.V.; Bilder, D.A.; Durkin, M.S.; Fitzgerald, R.T.; Furnier, S.M.; Hughes, M.M.; et al. Prevalence and Characteristics of Autism Spectrum Disorder among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *MMWR Surveill. Summ.* **2023**, *72*, 1–14. [CrossRef] [PubMed]
7. Hyman, S. *Autism: The Science of Mental Health*; Routledge: Abingdon, UK, 2013. Available online: https://books.google.com.hk/books/about/Autism.html?id=_WZGAQAAQBAJ&source=kp_book_description&redir_esc=y (accessed on 14 July 2023).
8. Bailey, A.; Couteur, A.L.E.; Gottesman, I.; Bolton, P.; Simonoff, E.; Yuzda, E. Autism as a strongly genetic disorder: Evidence from a British twin study. *Psychol. Med.* **1995**, *25*, 63–77. [CrossRef]
9. Hu, V.W. From Genes to Environment: Using Integrative Genomics to Build a "Systems-Level" Understanding of Autism Spectrum Disorders. *Child Dev.* **2013**, *84*, 89–103. [CrossRef]
10. Mojarad, B.A.; Yin, Y.; Dov, A.; Chandrakumar, I.; Prasolava, T.; Shum, N.; Hamdan, O.; Pellecchia, G.; Howe, J.L.; Whitney, J.; et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **2020**, *586*, 80–86. [CrossRef]
11. Satterstrom, F.K.; Kosmicki, J.A.; Wang, J.; Breen, M.S.; De Rubeis, S.; An, J.-Y.; Peng, M.; Collins, R.; Grove, J.; Klei, L.; et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **2020**, *180*, 568–584.e23. [CrossRef]
12. Schaaf, C.P.; Yuen, R.K.C.; Gallagher, L.; Skuse, D.H.; Buxbaum, J.D.; Bolton, P.F.; Cook, E.H.; Scherer, S.W.; Vorstman, J.A.S. A framework for an evidence-based gene list relevant to autism spectrum disorder. *Nat. Rev. Genet.* **2020**, *21*, 367–376. [CrossRef] [PubMed]
13. Lee, H.G.; Imaichi, S.; Kraeutler, E.; Aguilar, R.; Lee, Y.W.; Sheridan, S.D.; Lee, J.T. Site-specific R-loops induce CGG repeat contraction and fragile X gene reactivation. *Cell* **2023**, *186*, 2593–2609.e18. [CrossRef]
14. Ozkul, Y.; Taheri, S.; Bayram, K.K.; Sener, E.F.; Mehmetbeyoglu, E.; Öztop, D.B.; Aybuga, F.; Tufan, E.; Bayram, A.; Dolu, N.; et al. A heritable profile of six miRNAs in autistic patients and mouse models. *Sci. Rep.* **2020**, *10*, 9011. [CrossRef] [PubMed]
15. Liu, Y.; Liu, X.; Lin, C.; Jia, X.; Zhu, H.; Song, J.; Zhang, Y. Noncoding RNAs regulate alternative splicing in Cancer. *J. Exp. Clin. Cancer Res.* **2021**, *40*, 11. [CrossRef] [PubMed]
16. Rahmani, A.M.; Azhir, E.; Ali, S.; Mohammadi, M.; Ahmed, O.H.; Yassin Ghafour, M.; Hasan Ahmed, S.; Hosseinzadeh, M. Artificial intelligence approaches and mechanisms for big data analytics: A systematic study. *PeerJ. Comput. Sci.* **2021**, *7*, e488. [CrossRef]
17. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef]
18. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef] [PubMed]
19. Wang, H.; Lengerich, B.J.; Aragam, B.; Xing, E.P. Genetics and population analysis Precision Lasso: Accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics* **2019**, *35*, 1181–1187. [CrossRef]
20. Wang, Q. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *17*, 2131–2140.
21. Article, O. Breast Cancer Risk Estimation with Artificial Neural Networks Revisited. *Cancer* **2010**, *116*, 3310–3321. [CrossRef]
22. Cho, S.; Won, H. Machine Learning in DNA Microarray Analysis for Cancer Classification. In Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003, Adelaide, Australia, 1 February 2003.
23. Agresti, A. *Categorical Data Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 792.
24. Sperandei, S. Understanding logistic regression analysis. *Biochem. Medica* **2014**, *24*, 12–18. [CrossRef] [PubMed]
25. Huang, S.; Cai, N.; Pacheco, P.P. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genom. Proteom. January* **2018**, *15*, 41–51. [CrossRef]
26. Id, J.L.; Jew, B.; Zhan, L.; Hwang, S.; Id, G.C.; Freimer, B.; Sul, J.H. ForestQC: Quality control on genetic variants from next-generation sequencing data using random forest. *PLoS Comput. Biol.* **2019**, *15*, e1007556.
27. Pellegrino, E.; Jacques, C.; Beaufils, N.; Nanni, I.; Carlioz, A.; Metellus, P.; Ouafik, L.H. Machine learning random forest for predicting oncosomatic variant NGS analysis. *Sci. Rep.* **2021**, *11*, 21820. [CrossRef]
28. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009; Volume 2.
29. Song, Y.Y.; Lu, Y. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130–135. [CrossRef] [PubMed]
30. Vlachos, I.S.; Zagganas, K.; Paraskevopoulou, M.D.; Georgakilas, G.; Karagkouni, D.; Vergoulis, T.; Dalamagas, T.; Hatzigeorgiou, A.G. DIANA-miRPath v3.0: Deciphering microRNA function with experimental support. *Nucleic Acids Res.* **2015**, *43*, W460–W466. [CrossRef]
31. Rassoulzadegan, M.; Mehmetbeyoglu, E.; Yilmaz, Z.; Taheri, S.; Ozkul, Y. Progressive decline in the levels of six miRNAs from parents to children in autism. *bioRxiv* **2022**. [CrossRef]
32. Shaw, L.M. The insulin receptor substrate (IRS) proteins: At the intersection of metabolism and cancer. *Cell Cycle* **2011**, *10*, 1750–1756. [CrossRef]
33. Tang, X.; Chen, Y.; Luo, H.; Bian, Q.; Weng, B.; Yang, A.; Chu, D.; Ran, M.; Chen, B. miR-126 controls the apoptosis and proliferation of immature porcine sertoli cells by targeting the pik3r2 gene through the PI3K/AKT signaling pathway. *Animals* **2021**, *11*, 2260. [CrossRef]

34. Lin, S.; Gregory, R.I. MicroRNA biogenesis pathways in cancer. *Nat. Rev. Cancer* **2015**, *15*, 321–333. [CrossRef]
35. Wang, F.; Chen, C.; Wang, D. Circulating microRNAs in cardiovascular diseases: From biomarkers to therapeutic targets. *Front. Med.* **2014**, *8*, 404–418. [CrossRef] [PubMed]
36. Absalon, S.; Kochanek, D.M.; Raghavan, V.; Krichevsky, A.M. MiR-26b, upregulated in Alzheimer's disease, activates cell cycle entry, tau-phosphorylation, and apoptosis in postmitotic neurons. *J. Neurosci. Off. J. Soc. Neurosci.* **2013**, *33*, 14645–14659. [CrossRef] [PubMed]
37. Hay, N.; Sonenberg, N. Upstream and downstream of mTOR. *Genes Dev.* **2004**, *18*, 1926–1945. [CrossRef] [PubMed]
38. Huber, K.M.; Klann, E.; Costa-Mattioli, M.; Zukin, R.S. Dysregulation of mammalian target of rapamycin signaling in mouse models of autism. *J. Neurosci.* **2015**, *35*, 13836. [CrossRef] [PubMed]
39. Wang, B.; Qin, Y.; Wu, Q.; Li, X.; Xie, D.; Zhao, Z.; Duan, S. mTOR Signaling Pathway Regulates the Release of Proinflammatory Molecule CCL5 Implicated in the Pathogenesis of Autism Spectrum Disorder. *Front. Immunol.* **2022**, *13*, 818518. [CrossRef] [PubMed]
40. Chen, J.; Alberts, I.; Li, X. Dysregulation of the IGF-I/PI3K/AKT/mTOR signaling pathway in autism spectrum disorders. *Int. J. Dev. Neurosci.* **2014**, *35*, 35–41. [CrossRef] [PubMed]
41. Yeung, K.S.; Tso, W.W.Y.; Ip, J.J.K.; Mak, C.C.Y.; Leung, G.K.C.; Tsang, M.H.Y.; Ying, D.; Pei, S.L.C.; Lee, S.L.; Yang, W.; et al. Identification of mutations in the PI3K-AKT-mTOR signalling pathway in patients with macrocephaly and developmental delay and/or autism. *Mol. Autism* **2017**, *8*, 66. [CrossRef] [PubMed]
42. Bhandari, R.; Paliwal, J.K.; Kuhad, A. Neuropsychopathology of Autism Spectrum Disorder: Complex Interplay of Genetic, Epigenetic, and Environmental Factors. In *Personalized Food Intervention and Therapy for Autism Spectrum Disorder Management*; Springer: Cham, Switzerland, 2020; Volume 24. [CrossRef]
43. Trifonova, E.A.; Klimenko, A.I.; Mustafin, Z.S.; Lashin, S.A.; Kochetov, A.V. The mTOR signaling pathway activity and vitamin d availability control the expression of most autism predisposition genes. *Int. J. Mol. Sci.* **2019**, *20*, 6332. [CrossRef]
44. Pagani, M.; Barsotti, N.; Bertero, A.; Trakoshis, S.; Ulysse, L.; Locarno, A.; Miseviciute, I.; De Felice, A.; Canella, C.; Supekar, K.; et al. mTOR-related synaptic pathology causes autism spectrum disorder-associated functional hyperconnectivity. *Nat. Commun.* **2021**, *12*, 6084. [CrossRef]
45. Tufan, E.; Taheri, S.; Karaca, Z.; Mehmetbeyoglu, E.; Yilmaz Sukranli, Z.; Korkmaz Bayram, K.; Ulutabanca, H.; Tanrıverdi, F.; Unluhizarci, K.; Rassoulzadegan, M.; et al. Alterations in serum miR-126-3p levels over time, a marker of pituitary insufficiency following head trauma. *Neuroendocrinology* **2023**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.