International Journal of
Molecular Sciences

MDPI

Article

Machine Learning-Based Blood RNA Signature for Diagnosis of Autism Spectrum Disorder

Irena Voinsky [1], Oleg Y. Fridland [2], Adi Aran [3,4], Richard E. Frye [5,6,*] and David Gurwitz [1,7,*]

[1] Department of Human Molecular Genetics and Biochemistry, Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel
[2] Independent Researcher, Tel Aviv 69978, Israel
[3] Shaare Zedek Medical Center, Jerusalem 91031, Israel
[4] Obesity and Metabolism Laboratory, Institute for Drug Research, School of Pharmacy, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 91240, Israel
[5] Autism Discovery and Treatment Foundation, Phoenix, AZ 85050, USA
[6] Rossignol Medical Center, Phoenix, AZ 85050, USA
[7] Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 69978, Israel
[*] Correspondence: drfrye@rossignolmedicalcenter.com or drfrye@autismdiscovery.org (R.E.F.); gurwitz@tauex.tau.ac.il (D.G.)

Abstract: Early diagnosis of autism spectrum disorder (ASD) is crucial for providing appropriate treatments and parental guidance from an early age. Yet, ASD diagnosis is a lengthy process, in part due to the lack of reliable biomarkers. We recently applied RNA-sequencing of peripheral blood samples from 73 American and Israeli children with ASD and 26 neurotypically developing (NT) children to identify 10 genes with dysregulated blood expression levels in children with ASD. Machine learning (ML) analyzes data by computerized analytical model building and may be applied to building diagnostic tools based on the optimization of large datasets. Here, we present several ML-generated models, based on RNA expression datasets collected during our recently published RNA-seq study, as tentative tools for ASD diagnosis. Using the random forest classifier, two of our proposed models yield an accuracy of 82% in distinguishing children with ASD and NT children. Our proof-of-concept study requires refinement and independent validation by studies with far larger cohorts of children with ASD and NT children and should thus be perceived as starting point for building more accurate ML-based tools. Eventually, such tools may potentially provide an unbiased means to support the early diagnosis of ASD.

Keywords: machine learning; RNA biomarkers; blood RNA-sequencing; autism spectrum disorder (ASD)

check for updates

Citation: Voinsky, I.; Fridland, O.Y.; Aran, A.; Frye, R.E.; Gurwitz, D. Machine Learning-Based Blood RNA Signature for Diagnosis of Autism Spectrum Disorder. *Int. J. Mol. Sci.* **2023**, *24*, 2082. https://doi.org/10.3390/ijms24032082

Academic Editor: Changjong Moon

Received: 15 December 2022
Revised: 15 January 2023
Accepted: 17 January 2023
Published: 20 January 2023

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder exhibiting a wide phenotypic scope and characterized by impairment in communication skills, social interaction, and behavior (restricted or repetitive) [1]. ASD is usually diagnosed during childhood, and mild autism is sometimes diagnosed only during adulthood [2,3]. It is an extremely heterogeneous disorder and could develop due to inheritable or de novo gene variations. Although hundreds of genes have been associated as contributors, in most cases the etiology remains unknown [4]. Thus, ASD is now assumed to be a disorder of complex interaction involving genetics, epigenetics, and the environment [5]. Common contributors to the development of the disorder include point mutations [6], copy number variants (CNVs) [7], translocations [8], DNA methylation [9], histone modifications [10], miRNAs expression [11], mitochondrial deficiencies [12], viral infections [13,14], aberrant gut microbiome composition [15], parental age, and environmental influences [16,17].

While understanding of the neurobiology and genetics of ASD has greatly improved in recent years, the diagnosis of ASD remains mainly based on defined behavioral and clinical symptoms reported by ASD children's primary caregivers and clinicians' assessment. Early diagnosis, ideally by age of 3–4 years, is crucial for starting behavioral therapy at an early age, which is critical for reducing ASD symptoms, strengthening communication skills, guiding parents, and improving ASD patients' quality of life. Despite increasing awareness and monitoring of ASD rates for two decades, the average age of diagnosis, at least in the USA, has not improved [18]. This may be, in part, due to the fact that no unbiased systematic approach or medical test for the detection of ASD has been adopted into clinical practice. Indeed, although there are many biomarkers under development, most require replication and validation [19].

Machine learning (ML), sometimes referred to as "deep learning", is a subfield of artificial intelligence (AI) research that analyzes data by computerized analytical model building. ML models are built based on statistical algorithms and are fitting for complex problem-solving involving multiple possibilities and combinations where conventional computational models might fail. Consequently, ML may provide tools to considerably increase the function of computational methods in neuroscience as well as improve clinical diagnosis and assist in the selection of treatment options. In recent years, considerable research has been applied in developing ML models to classify neuronal pathways and improve the understanding of mental disorders [20,21], Parkinson's disease [22,23], Alzheimer's disease [24,25], epilepsy [26], gestational diabetes [27], blood infections [28], COVID-19 [29], and more. Studies applying ML tools in ASD research include mainly models based on brain imaging data [30–33], but also behavioral evaluations [34–37], kinematic data [38,39], parental ages [40], eye movement data [41], and audio communication samples [42]. With ASD being a complex heterogeneous disorder, ML models based on genetic and/or genomic information are more limited. Recently published studies in this field focused on data retrieved from rare copy number variations (CNV) [43], long non-coding RNA (lncRNA) gene expression [44], and genome-wide association study (GWAS) meta-analysis [45]. ML models for ASD diagnosis were proposed mostly using DNA variant analysis [46,47]. Some ASD ML-based models examined RNA levels using in vitro cellular systems [48] or in silico data mining [49]. One study proposed a ML tool for ASD diagnosis based on salivary RNA [50]. However, blood samples are more readily available than saliva in toddlers. To our knowledge, no ML tools based on blood RNA expression levels have been reported for pediatric ASD diagnosis.

In this study, we aimed to generate predictive ML models for pediatric ASD diagnosis by utilizing our datasets of RNA expression levels in whole blood samples of children with ASD and neurotypical (NT) control children using quantitative real-time quantitative polymerase chain reaction (RT-qPCR) data. The RT-qPCR database applied for our current study is composed of the RNA expression levels of 10 studied genes found dysregulated in ASD and reported in our recently published research article, which was based on genome-wide RNA sequencing (RNA-seq) of peripheral blood samples from 73 American and Israeli children with ASD and 26 NT children [51]. Here, we present our ML-generated tool as a tentative proof-of-concept study that, once validated and improved using far larger cohorts of children with ASD and NT children, may potentially serve as an unbiased adjacent tool for early diagnosis of ASD.

2. Results

2.1. Choosing the Optimal Gene Combinations for ML Models

We first evaluated the utility of a dataset of differently expressed genes in blood samples from ASD and NT individuals to serve as a potential ASD diagnostic tool. We performed ROC analysis with four genes that we recently reported as dysregulated in the blood of 73 children with ASD compared with 26 NT children [51]: *BATF2, LY6E, MT2A,* and *ISG15.* ROC analyses by mRNA expression with AUC > 0.5 for each of the tested genes alone (Figure 1): *BATF2* (AUC = 0.6774, p = 0. 0072), *MT2A* (AUC = 0.6553, p = 0.02), *ISG15* (AUC = 0.7518, p = 0.0001), and *LY6E* (AUC = 0.6538, p = 0.0198). These AUC and p values indicate a statistically significant distribution between the ASD and NT control groups. Pearson r correlation analysis was applied to the above four genes with diagnostic significance as detected by the ROC analysis for determining the gene combination with the highest predictive capacity based on our RT-qPCR data. The chosen gene combinations (predictors) for ML testing were those with correlations of r > 0.3 and $*p \leq 0.05$: (#1) *BATF2, LY6E, MT2A,* and *ISG15;* (#2) *BATF2, SERPING1, MT2A,* and *FBXO6*; (#3) *MT2A, ISG15, FBXO6, SERPING1,* and *BATF2*; (#4) *MT2A, ISG15,* and *FBXO6* (shown in bold fonts in Supplementary Table S1). An additional fifth predictor was chosen based on the results of the random forest classifier feature importance [52]. The fifth predictor is a combination of the following five genes with the highest importance as shown by the classifier: *BATF2, ISG15, SERPING1, LY6E,* and *EFHC2* (Supplementary Figure S1).

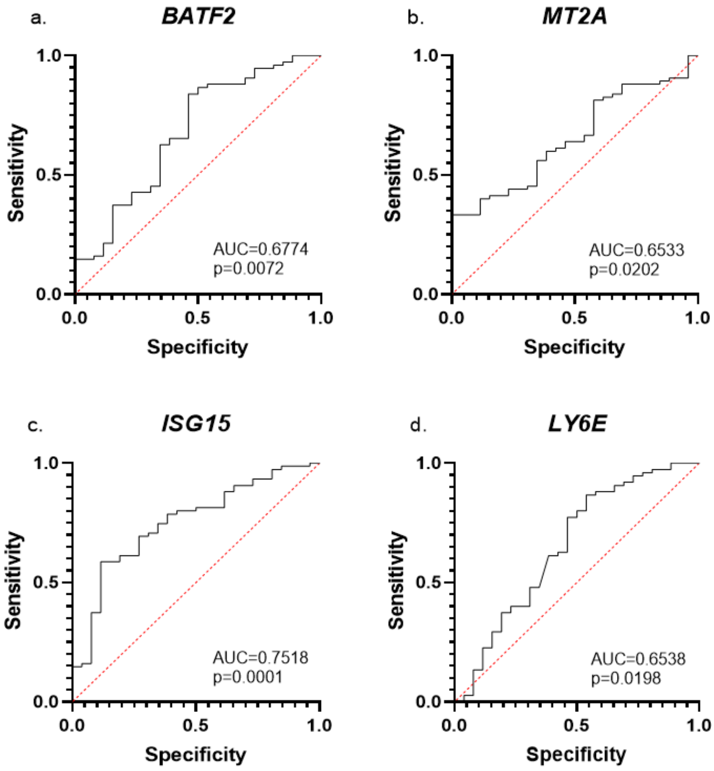


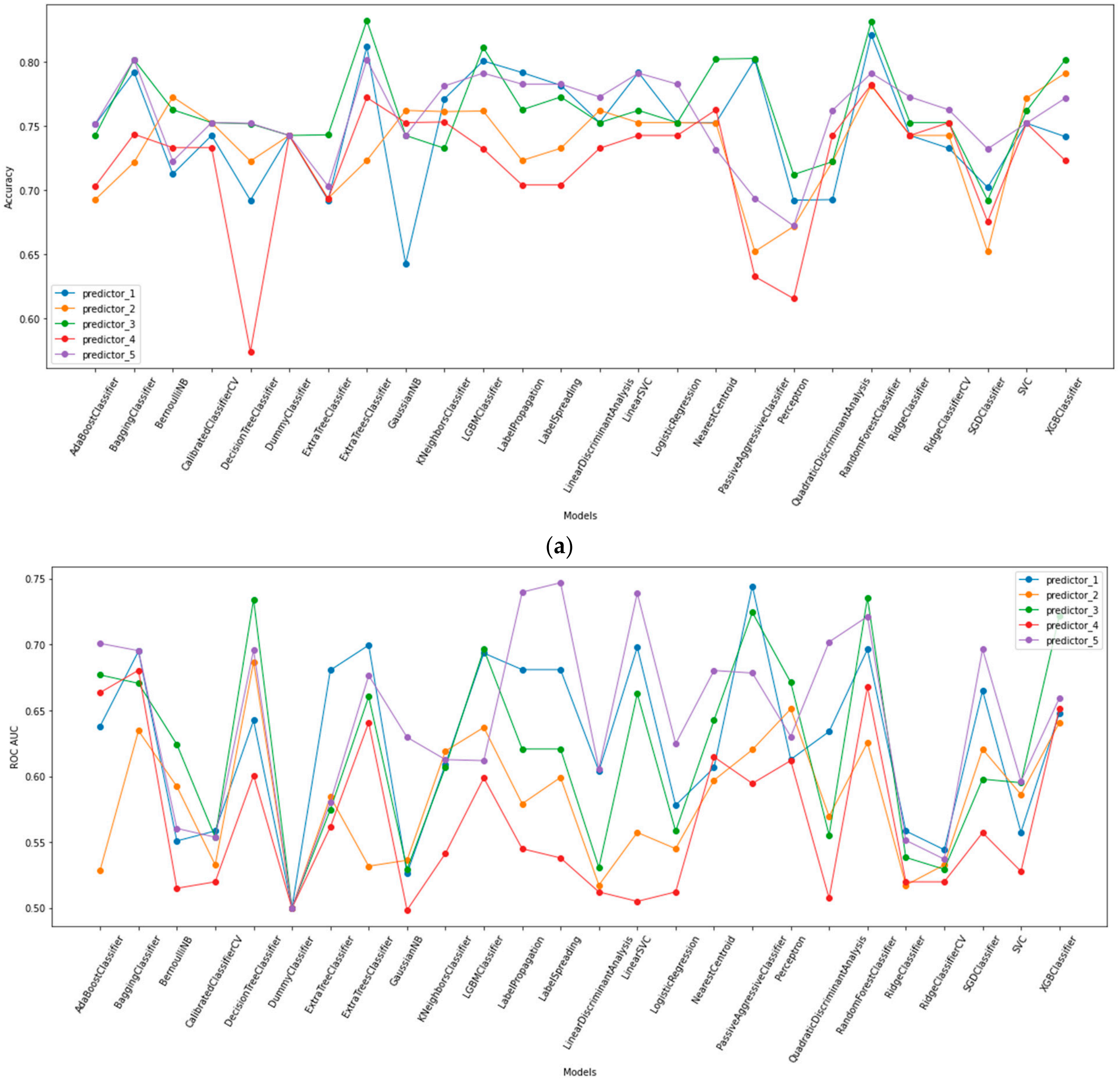
Figure 1. ROC analysis of ASD-specific mRNA expression biomarkers. AUC and *p*-values are displayed for (**a**) *BATF2*, (**b**) *MT2A*, (**c**) *ISG15,* and (**d**) *LY6E.*

2.2. Detection of an Optimized Diagnostic Model

Based on our ROC analysis, Pearson r correlations values (Supplementary Table S1), and MDI feature importance evaluations, we chose the five mRNA expression combinations described above (Section 2.1). Next, we applied the Lazy Predict tool to the chosen predictors to determine which of the 36 ML models is most suitable for our randomized data sets. Extra trees and random forest classifiers presented the highest accuracy and ROC AUC values. Both ML models worked most accurately with predictors #1 (*BATF2*, *LY6E*, *MT2A* and *ISG15*), #3 (*MT2A*, *ISG15*, *FBXO6*, *SERPING1* and *BATF2*), and #5 (*BATF2*, *ISG15*, *SERPING1*, *LY6E* and *EFHC2*; Figure 2). To review the efficiency of extra trees and random forest classifiers, we measured the accuracy score using the leave-one-out cross-validator and performed ROC AUC measurements using stratified K-folds cross-validator. All calculations were made for the three chosen predictors (Table 1 and Figure 3). Results presented in Table 1 show the highest accuracy for predictors #3 and #5 when using the random forest classifier (accuracy = 82.178%; AUC = 0.82, 0.77, respectively). The combination consisting of the four significantly dysregulated and RT-qPCR validated genes in our recently published article (*BATF2*, *LY6E*, *MT2A*, and *ISG15*; predictor #1) also produced a highly accurate result when applying the extra trees classifier (accuracy = 81.188%, ROC AUC = 0.79; Table 1 and Figure 3).

Table 1. Summary of model accuracy results for different gene combinations (predictors), based on RT-qPCR results ($2^{-\Delta Ct}$): (#1) *BATF2*, *LY6E*, *MT2A* and *ISG15*, (#3) *MT2A*, *ISG15*, *FBXO6*, *SERPING1*, and *BATF2* (#5) *BATF2*, *ISG15*, *SERPING1*, *LY6E*, and *EFHC2*.

Predictor Model Accuracy	#1	#3	#5
Extra Trees Classifier	**81.188%**	80.198%	80.198%
Random Forest Classifier	79.208%	**82.178%**	**82.178%**



(**a**)

(**b**)

Figure 2. Summary of Lazy Predict library results of 36 ML models for five established predictors. (**a**) Accuracy score, (**b**) ROC AUC score. Two models were removed due to poor fit. Predictor_1: *BATF2*, *LY6E*, *MT2A* and *ISG15*. Predictor_2: *BATF2*, *SERPING1*, *MT2A*, and *FBXO6*. Predictor_3: *MT2A*, *ISG15*, *FBXO6*, *SERPING1*, and *BATF2*. Predictor_4: *MT2A*, *ISG15*, and *FBXO6*. Predictor_5: *BATF2*, *ISG15*, *SERPING1*, *LY6E*, and *EFHC2*.

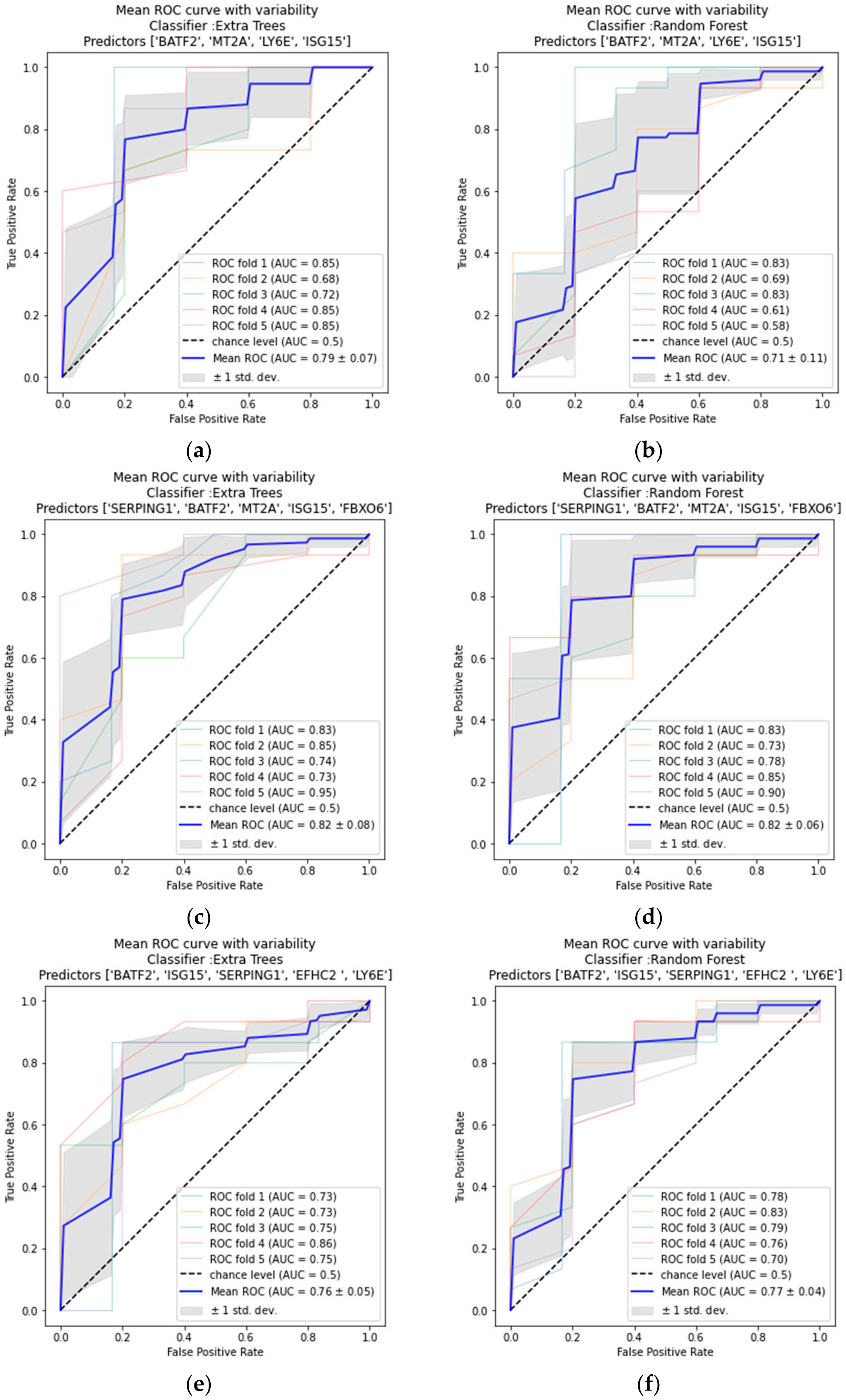


Figure 3. Results of mean ROC curve with variability for different predictors: #1, #3 and #5, based on our ML models. Predictor #1 (**a**) Extra Trees classifier, (**b**) Random Forest classifier. Predictor #3 (**c**) Extra Trees classifier, (**d**) Random Forest classifier. Predictor #5 (**e**) Extra Trees classifier, (**f**) Random Forest classifier. Each ROC AUC consists of a 5-StratifiedKFold average.

3. Discussion

In this proof-of-concept study, we aimed to demonstrate the utility of blood transcriptomic data from small cohorts for building a ML-based tentative tool for distinguishing between children with ASD and NT children. Gene combination predictors were identified based on a combination of ML methods with RT-qPCR data generated blood gene expression values ($2^{\sim\text{-}\Delta Ct}$) that yielded an accuracy of 82% in correctly identifying children with ASD and NT children. Two of our five ML models presented the highest suitability for our dataset: (1) *MT2A, ISG15, FBXO6, SERPING1,* and *BATF2*; (2) *BATF2*, *ISG15*, *SERPING1*, *LY6E,* and *EFHC2*. All the genes included in these two predictors have significance in ASD etiology, as discussed in our recent publication [51].

Our ML-generated tools described here should be considered as a proof-of-concept study and a preliminary guide for further studies on transcriptomics-based ASD diagnostics. Key limitations of the study include its small sample size and the possibility of model overfitting due to the absence of independent validation cohorts. In addition, the connection between the human peripheral blood and the brain transcriptomic profiles in individuals is poorly understood. Studies suggest that between 35% and 80% of known human transcripts are expressed in both the brain and blood, indicating a thoughtful use is needed when purposing peripheral gene expression as a proxy for gene expression in the CNS [53,54]. Therefore, the findings presented in this study should be interpreted carefully. Yet, keeping in mind that RT-qPCR studies (or custom-built gene expression microarrays) of blood samples are more accessible and affordable compared with brain MRI or fMRI scans, the diagnostic potential of ML-based tools for the detection of individuals with ASD following analysis of blood gene expression levels deserves further exploration. Analysis in larger cohorts should be carried out for improving and refining the ML tools proposed here. Future studies should also consider the influence of additional factors, including sex, age, ethnicity, and other confounders affecting associations between ASD phenotypes and blood genomic markers in their ML algorithms. This approach may eventually assist in the identification of a panel of biomarkers, leading to the earlier diagnosis of ASD among children with atypical neurodevelopment and to the stratification of the ASD population to different pathophysiologically relevant subgroups. Hence, RNA-based ML tools may provide better-personalized treatment alternatives for individuals with ASD.

4. Materials and Methods

4.1. Data Collection

The data used for this study are RNA expression levels of genes quantified by RT-qPCR ($2^{-\Delta Ct}$). Data were obtained as described by Voinsky et al., 2022. Briefly, whole blood samples were collected from 73 ASD children and 26 NT controls in two cohorts (Israel and USA). RNA sequencing was performed on a subset of the samples. Next, the top 10 genes which were differentially expressed between the ASD and NT groups ($p_{adj} < 0.05$) were validated by RT-qPCR experiments, containing all samples in the Israeli and American cohorts. RNA expression levels of the following 10 genes were studied: *SERPING1*, *EFHC2*, *BATF2*, *CDC20*, *FCGR1A*, *MT2A*, *ISG15*, *FBXO6*, LINC00869, and *LY6E; GAPDH* was used as the qPCR control gene. The description of these genes, including their Gene ID codes, is provide in the work of Voinsky et al. [51]. All procedures and protocols were previously explained. Notably, of these 10 genes, two (*BATF2* and *LY6E*) were found upregulated and two (*ISG15* and *MT2A*) were found downregulated in blood samples from our combined American and Israeli cohorts of 73 children with ASD and 26 NT children [51].

4.2. Data Pre-Processing

Pre-processing of data was required for handling null values, missing in some samples due to removal in cases of a low quantity of tested samples. As such data were lacking at random, there is no specific structure to explain this absence, and missing values were replaced with median imputation [55,56]. Original and processed data were compared and found to present no significant statistical difference ($p > 0.05$).

4.3. Selecting Feature Importance

Random forest classifier was applied, working with the scikit-learn library [57] and using the RandomForestClassifier method (Python software v. 3.9), to determine the contribution of each of the 10 genes to the model prediction. Feature importance in the random forest classifier is based on a mean decrease in impurity (MDI). Thus, a score is computed based on the mean and standard deviation of accumulation of the impurity decline within each tree. In the scikit-learn library, for each decision tree, the library calculates an importance node using an MDI, with only two child nodes assumed (a binary tree):

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (1)$$

where ni_j = node j importance, w_j = weighted number of samples reaching node j, C_j = node j impurity value, left (j)= child node from left split on node j, and right (j) = child node from right split on node j.

Next, the importance of each feature on a decision tree is determined as:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \quad (2)$$

where *fi* $_i$ = the importance of feature I, n_{ij} = the importance of node j.

Later, this value can then be normalized by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \quad (3)$$

The random forest's final feature importance is its average over all the trees. The total value of the feature's importance on each tree is determined and then divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \quad (4)$$

where $RFfi_i$ = the importance of feature i calculated from all trees in the random forest model, $normfi_{ij}$ = the normalized feature importance for i in tree j, and T= sum of trees (total). For the predictor combination, we used genes with an MDI score > 1. Additionally, we focused on the four dysregulated genes from our previously published study that were validated by RT-qPCR in the combined American and Israeli cohorts [51]: *BATF2*, *LY6E*, *MT2A*, and *ISG15*. To evaluate the dysregulated genes' diagnostic value, ROC (receiver operating characteristic) analysis was utilized [58]. Next, Spearman r correlation analysis was used for each dysregulated gene with other significantly differentially expressed genes (as defined in "Section 4.1. Data Collection"). For the correlation test, $p \leq 0.05$ was considered significant. ROC and Spearman analyses were performed using GraphPad Prism v. 9 software (San Diego, CA, USA).

4.4. Machine Learning Algorithms

The Lazy Predict python library (https://lazypredict.readthedocs.io/en/latest/, accessed on 1 September 2022) was used to evaluate the most applicable ML algorithms for the prediction of ASD transcriptomic signature. Lazy Predict is a library that builds 36 basic ML models, suggesting the most suitable model for prediction variables prior to testing against hyperparameters. Consequently, two ML models were selected for further inspection, random forest classifier and the extra trees (extremely randomized trees) classifier. Random forest is a controlled ensemble learning algorithm that consists of many small decision trees (estimators), each generating its own prediction [59]. The random forest algorithm creates and combines multiple decision trees into one "forest" to deliver a more accurate prediction. Extra trees classifier is also an ensemble learning algorithm, like random forest, except for the random selection of split values in the data [60]. That is, while random forest selects cut points to split connections at an optimal split, extra trees chooses them randomly. Next, we applied grid search, an optimization tool used to select the best combination of parameters, for tuning the hyperparameters in our models. The chosen models and tools were applied using the scikit-learn methods RandomForestClassifier, ExtraTreesClassifier, and GridSearchCV. All the methods were computed using Python software v. 3.9.

4.5. Accuracy and ROC AUC Validation

For validating the accuracy of our ML algorithms, we applied the leave-one-out cross-validator, using the scikit-learn method LeaveOneOut. This method is favored when analyzing small data sets, such as the one we used for this study. In this form of validation, the number of folds equals the number of cases in the data set. Hence, it uses a selected case as a single-item test set, where the learning algorithm is applied once for each case, and all other cases are used as a training set. To assess our models, we visualized the variance of the ROC metrics using cross-validation. Scikit-learn library methods were utilized. The RocCurveDisplay method was used to draw the curves, StratifiedKFold method computed the fold groups, and the auc method was used to calculate the area under the curve (AUC) using the trapezoidal rule. All the methods were computed using Python software v. 3.9.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms24032082/s1.

Author Contributions: I.V. conceived the study, performed the statistical and correlation analyses, prepared all figures and tables, and wrote the first article draft. O.Y.F. wrote the code for running the machine learning analysis of the study dataset. A.A. and R.E.F. provided the ASD and NT blood samples of the Israeli and American cohorts, respectively. A.A., R.E.F. and D.G. took part in writing the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The study was supported by a grant from the Israel-United States Science Foundation (BSF grant 2019049) to D.G. and R.E.F. In addition, I.V. was supported by a Ph.D. scholarship from the Yoran Institute for Human Genome Research at Tel Aviv University.

Institutional Review Board Statement: The protocol for the Israeli cohort was approved by Shaare Zedek Medical Center Institutional Review Board (Jerusalem, Israel; IRB# 0501-20-SZMC, original approval 2 December 2020). The protocol for the American cohort was approved by the Institutional Review Board at Phoenix Children's Hospital (Phoenix, AZ, USA; IRB# IRB-19-606, original approval 13 January 2020).

Informed Consent Statement: Informed consent was obtained from the children's parents for both the Israeli and the U.S. cohorts.

Data Availability Statement: The datasets and code generated for the current study, including the real-time qPCR spreadsheets, are available from I.V. or D.G. on reasonable request.

Acknowledgments: This work was carried out as part of the fulfillment requirements for the Ph.D. thesis of I.V.

Conflicts of Interest: The authors declare no competing interests or other interests that might be perceived to influence the results and/or discussion reported in this article.

References

1. Doernberg, E.; Hollander, E. Neurodevelopmental Disorders (ASD and ADHD): DSM-5, ICD-10, and ICD-11. *CNS Spectr.* **2016**, *21*, 295–299. [CrossRef] [PubMed]
2. Pham, H.H.; Sandberg, N.; Trinkl, J.; Thayer, J. Racial and Ethnic Differences in Rates and Age of Diagnosis of Autism Spectrum Disorder. *JAMA Netw. Open* **2022**, *5*, e2239604. [CrossRef] [PubMed]
3. Jadav, N.; Bal, V.H. Associations between Co-Occurring Conditions and Age of Autism Diagnosis: Implications for Mental Health Training and Adult Autism Research. *Autism. Res.* **2022**, *15*, 2112–2125. [CrossRef] [PubMed]
4. Willsey, H.R.; Willsey, A.J.; Wang, B.; State, M.W. Genomics, Convergent Neuroscience and Progress in Understanding Autism Spectrum Disorder. *Nat. Rev. Neurosci.* **2022**, *23*, 323–341. [CrossRef]
5. Yoon, S.H.; Choi, J.; Lee, W.J.; Do, J.T. Genetic and Epigenetic Etiology Underlying Autism Spectrum Disorder. *J. Clin. Med.* **2020**, *9*, 966. [CrossRef]

6. Durand, C.M.; Betancur, C.; Boeckers, T.M.; Bockmann, J.; Chaste, P.; Fauchereau, F.; Nygren, G.; Rastam, M.; Gillberg, I.C.; Anckarsäter, H.; et al. Mutations in the Gene Encoding the Synaptic Scaffolding Protein SHANK3 Are Associated with Autism Spectrum Disorders. *Nat. Genet.* **2007**, *39*, 25–27. [CrossRef]
7. Gregory, S.G.; Connelly, J.J.; Towers, A.J.; Johnson, J.; Biscocho, D.; Markunas, C.A.; Lintas, C.; Abramson, R.K.; Wright, H.H.; Ellis, P.; et al. Genomic and Epigenetic Evidence for Oxytocin Receptor Deficiency in Autism. *BMC Med.* **2009**, *7*, 62. [CrossRef]
8. Vincen, J.B.; Herbrick, J.A.; Gurling, H.M.D.; Bolton, P.F.; Roberts, W.; Scherer, S.W. Identification of a Novel Gene on Chromosome 7q31 That Is Interrupted by a Translocation Breakpoint in an Autistic Individual. *Am. J. Hum. Genet.* **2000**, *67*, 510–514. [CrossRef]
9. Bakulski, K.M.; Dou, J.F.; Feinberg, J.I.; Aung, M.T.; Ladd-Acosta, C.; Volk, H.E.; Newschaffer, C.J.; Croen, L.A.; Hertz-Picciotto, I.; Levy, S.E.; et al. Autism-Associated DNA Methylation at Birth From Multiple Tissues Is Enriched for Autism Genes in the Early Autism Risk Longitudinal Investigation. *Front. Mol. Neurosci.* **2021**, *14*, 775390. [CrossRef]
10. Shulha, H.P.; Cheung, I.; Whittle, C.; Wang, J.; Virgil, D.; Lin, C.L.; Guo, Y.; Lessard, A.; Akbarian, S.; Weng, Z. Epigenetic Signatures of Autism: Trimethylated H3K4 Landscapes in Prefrontal Neurons. *Arch. Gen. Psychiatry.* **2012**, *69*, 314–324. [CrossRef]
11. Wu, X.; Li, W.; Zheng, Y. Recent Progress on Relevant MicroRNAs in Autism Spectrum Disorders. *Int. J. Mol. Sci.* **2020**, *21*, 5904. [CrossRef] [PubMed]
12. Frye, R.E. Mitochondrial Dysfunction in Autism Spectrum Disorder: Unique Abnormalities and Targeted Treatments. *Semin. Pediatr. Neurol.* **2020**, *35*, 100829. [CrossRef]
13. Deykin, E.Y.; Macmahon, B. VIRAL EXPOSURE AND AUTISM. *Am. J. Epidemiol.* **1979**, *109*, 628–638. [CrossRef]
14. Zerbo, O.; Qian, Y.; Yoshida, C.; Grether, J.K.; van de Water, J.; Croen, L.A. Maternal Infection During Pregnancy and Autism Spectrum Disorders. *J. Autism. Dev. Disord.* **2016**, *45*, 4015–4025. [CrossRef]
15. Svoboda, E. Could the Gut Microbiome Be Linked to Autism? *Nature* **2020**, *577*, S14–S15. [CrossRef] [PubMed]
16. Hultman, C.M.; Sparén, P.; Cnattingius, S. Perinatal Risk Factors for Infantile Autism. *Epidemiology* **2002**, *13*, 417–423. [CrossRef] [PubMed]
17. Frye, R.E.; Cakir, J.; Rose, S.; Palmer, R.F.; Austin, C.; Curtin, P. Physiological Mediators of Prenatal Environmental Influences in Autism Spectrum Disorder. *Bioessays* **2021**, *43*, 2000307. [CrossRef]
18. Christensen, D.; Maenner, M.; Bilder, D.; Constantino, J.; Daniels, J.; Durkin, M.; Fitzgerald, R.; Kurzius-Spencer, M.; Pettygrove, S.; Robinson, C.; et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 4 Years—Early Autism and Developmental Disabilities Monitoring Network, Seven Sites, United States, 2010, 2012, and 2014. *MMWR Surveill. Summ.* **2019**, *68*, 1–19. [CrossRef]
19. Jensen, A.R.; Lane, A.L.; Werner, B.A.; McLees, S.E.; Fletcher, T.S.; Frye, R.E. Modern Biomarkers for Autism Spectrum Disorder: Future Directions. *Mol. Diagn. Ther.* **2022**, *26*, 483–495. [CrossRef]
20. Vassileva, J.; Young-Jin Cho, R.; Whelan, R.; Jollans, L. Neuromarkers for Mental Disorders: Harnessing Population Neuroscience. *Front. Psychiatry* **2018**, *1*, 242. [CrossRef]
21. Pintelas, E.G.; Kotsilieris, T.; Livieris, I.E.; Pintelas, P. A Review of Machine Learning Prediction Methods for Anxiety Disorders. In Proceedings of the ACM International Conference Proceeding Series, Thessaloniki Greece, 20 June 2018; Association for Computing Machinery: New York, NY, USA; pp. 8–15.
22. Avuçlu, E.; Elen, A. Evaluation of Train and Test Performance of Machine Learning Algorithms and Parkinson Diagnosis with Statistical Measurements. *Med. Biol. Eng. Comput.* **2020**, *58*, 2775–2788. [CrossRef] [PubMed]
23. Mei, J.; Desrosiers, C.; Frasnelli, J. Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature. *Front. Aging Neurosci.* **2021**, 13. [CrossRef] [PubMed]
24. Mirzaei, G.; Adeli, A.; Adeli, H. Imaging and Machine Learning Techniques for Diagnosis of Alzheimer's Disease. *Rev. Neurosci.* **2016**, *27*, 857–870. [CrossRef] [PubMed]
25. Trambaiolli, L.R.; Lorena, A.C.; Fraga, F.J.; Kanda, P.A.M.; Anghinah, R.; Nitrini, R. Improving Alzheimer's Disease Diagnosis with Machine Learning Techniques. *Clin. EEG Neurosci.* **2011**, *42*, 160–165. [CrossRef]
26. Abbasi, B.; Goldenholz, D.M. Machine Learning Applications in Epilepsy. *Epilepsia* **2019**, *60*, 2037–2047. [CrossRef]
27. Yoffe, L.; Polsky, A.; Gilam, A.; Raff, C.; Mecacci, F.; Ognibene, A.; Crispi, F.; Gratacós, E.; Kanety, H.; Mazaki-Tovi, S.; et al. Early Diagnosis of Gestational Diabetes Mellitus Using Circulating MicroRNAs. *Eur. J. Endocrinol.* **2019**, *181*, 565–577. [CrossRef]
28. Zoabi, Y.; Kehat, O.; Lahav, D.; Weiss-Meilik, A.; Adler, A.; Shomron, N. Predicting Bloodstream Infection Outcome Using Machine Learning. *Sci. Rep.* **2021**, *11*, 20101. [CrossRef]
29. Zoabi, Y.; Deri-Rozov, S.; Shomron, N. Machine Learning-Based Prediction of COVID-19 Diagnosis Based on Symptoms. *NPJ Digit. Med.* **2021**, *4*, 3. [CrossRef]
30. Ghiassian, S.; Greiner, R.; Jin, P.; Brown, M.R.G. Using Functional or Structural Magnetic Resonance Images and Personal Characteristic Data to Identify ADHD and Autism. *PLoS One* **2016**, *11*, e0166934. [CrossRef]
31. Abraham, A.; Milham, M.P.; di Martino, A.; Craddock, R.C.; Samaras, D.; Thirion, B.; Varoquaux, G. Deriving Reproducible Biomarkers from Multi-Site Resting-State Data: An Autism-Based Example. *Neuroimage* **2017**, *147*, 736–745. [CrossRef]
32. Heinsfeld, A.S.; Franco, A.R.; Craddock, R.C.; Buchweitz, A.; Meneguzzi, F. Identification of Autism Spectrum Disorder Using Deep Learning and the ABIDE Dataset. *Neuroimage Clin.* **2018**, *17*, 16–23. [CrossRef] [PubMed]
33. Qiu, A.; He, L.; Li, H.; Parikh, N.A. A Novel Transfer Learning Approach to Enhance Deep Neural Network Classification of Brain Functional Connectomes. *Front. Neurosci.* **2018**, *1*, 491. [CrossRef]
34. Bone, D.; Goodwin, M.S.; Black, M.P.; Lee, C.-C.; Audhkhasi, K.; Narayanan, S. Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises. *J. Autism. Dev. Disord.* **2015**, *45*, 1121–1136. [CrossRef]

35. Kosmicki, J.A.; Sochat, V.; Duda, M.; Wall, D.P. Searching for a Minimal Set of Behaviors for Autism Detection through Feature Selection-Based Machine Learning. *Transl. Psychiatry* **2015**, *5*, e514. [CrossRef] [PubMed]
36. Georgescu, A.L.; Koehler, J.C.; Weiske, J.; Vogeley, K.; Koutsouleris, N.; Falter-Wagner, C. Machine Learning to Study Social Interaction Difficulties in ASD. *Front Robot AI* **2019**, *6*, 132. Available online: https://pubmed.ncbi.nlm.nih.gov/33501147/ (accessed on 14 December 2022). [CrossRef] [PubMed]
37. Thabtah, F. Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. In Proceedings of the 1st International Conference on Medical and Health Informatics 2017, Taichung City, Taiwan, 20–22 May 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 1–6.
38. Crippa, A.; Salvatore, C.; Perego, P.; Forti, S.; Nobile, M.; Molteni, M.; Castiglioni, I. Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities. *J. Autism. Dev. Disord.* **2015**, *45*, 2146–2156. [CrossRef] [PubMed]
39. Li, B.; Sharma, A.; Meng, J.; Purushwalkam, S.; Gowen, E. Applying Machine Learning to Identify Autistic Adults Using Imitation: An Exploratory Study. *PLoS ONE* **2017**, *12*, e0182652. [CrossRef]
40. Grether, J.K.; Anderson, M.C.; Croen, L.A.; Smith, D.; Windham, G.C. Risk of Autism and Increasing Maternal and Paternal Age in a Large North American Population. *Am. J. Epidemiol.* **2009**, *170*, 1118–1126. [CrossRef]
41. Liu, W.; Li, M.; Yi, L. Identifying Children with Autism Spectrum Disorder Based on Their Face Processing Abnormality: A Machine Learning Framework. *Autism. Res.* **2016**, *9*, 888–898. [CrossRef]
42. Nakai, Y.; Takiguchi, T.; Matsui, G.; Yamaoka, N.; Takada, S. Detecting Abnormal Word Utterances in Children With Autism Spectrum Disorders: Machine-Learning-Based Voice Analysis Versus Speech Therapists. *Percept. Mot. Ski.* **2017**, *124*, 961–973. [CrossRef]
43. Engchuan, W.; Dhindsa, K.; Lionel, A.C.; Scherer, S.W.; Chan, J.H.; Merico, D. Performance of Case-Control Rare Copy Number Variation Annotation in Classification of Autism. *BMC Med. Genom.* **2015**, *8*, S7. [CrossRef] [PubMed]
44. Gök, M. A Novel Machine Learning Model to Predict Autism Spectrum Disorders Risk Gene. *Neural. Comput. Appl.* **2019**, *31*, 6711–6717. [CrossRef]
45. Polimanti, R.; Gelernter, J. Widespread Signatures of Positive Selection in Common Risk Alleles Associated to Autism Spectrum Disorder. *PLoS Genet.* **2017**, *13*, e1006618. [CrossRef] [PubMed]
46. Lin, Y.; Afshar, S.; Rajadhyaksha, A.M.; Potash, J.B.; Han, S. A Machine Learning Approach to Predicting Autism Risk Genes: Validation of Known Genes and Discovery of New Candidates. *Front. Genet.* **2020**, *11*, 500064. [CrossRef]
47. Zhou, J.; Park, C.Y.; Theesfeld, C.L.; Wong, A.K.; Yuan, Y.; Scheckel, C.; Fak, J.J.; Funk, J.; Yao, K.; Tajima, Y.; et al. Whole-Genome Deep-Learning Analysis Identifies Contribution of Noncoding Mutations to Autism Risk. *Nat. Genet.* **2019**, *51*, 973–980. [CrossRef]
48. Chiocchetti, A.G.; Haslinger, D.; Stein, J.L.; de La Torre-Ubieta, L.; Cocchi, E.; Rothämel, T.; Lindlar, S.; Waltes, R.; Fulda, S.; Geschwind, D.H.; et al. Transcriptomic Signatures of Neuronal Differentiation and Their Association with Risk Genes for Autism Spectrum and Related Neuropsychiatric Disorders. *Transl. Psychiatry* **2016**, *6*, e864. [CrossRef]
49. Shen, L.; Lin, Y.; Sun, Z.; Yuan, X.; Chen, L.; Shen, B. Knowledge-Guided Bioinformatics Model for Identifying Autism Spectrum Disorder Diagnostic MicroRNA Biomarkers. *Sci. Rep.* **2016**, *6*, 39663. [CrossRef]
50. Hicks, S.D.; Rajan, A.T.; Wagner, K.E.; Barns, S.; Carpenter, R.L.; Middleton, F.A. Validation of a Salivary RNA Test for Childhood Autism Spectrum Disorder. *Front. Genet.* **2018**, *9*, 534. [CrossRef]
51. Voinsky, I.; Zoabi, Y.; Shomron, N.; Harel, M.; Cassuto, H.; Tam, J.; Rose, S.; Scheck, A.C.; Karim, M.A.; Frye, R.E.; et al. Blood RNA Sequencing Indicates Upregulated BATF2 and LY6E and Downregulated ISG15 and MT2A Expression in Children with Autism Spectrum Disorder. *Int. J. Mol. Sci.* **2022**, *23*, 9843. [CrossRef]
52. Archer, K.J.; Kimes, R.V. Empirical Characterization of Random Forest Variable Importance Measures. *Comput. Stat. Data Anal.* **2008**, *52*, 2249–2260. [CrossRef]
53. Tylee, D.S.; Kawaguchi, D.M.; Glatt, S.J. On the Outside, Looking in: A Review and Evaluation of the Comparability of Blood and Brain "-Omes. " *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2013**, *162*, 595–603. [CrossRef]
54. Sullivan, P.F.; Fan, C.; Perou, C.M. Evaluating the Comparability of Gene Expression in Blood and Brain. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2006**, *141B*, 261–268. [CrossRef] [PubMed]
55. Gu, K.M.; Min, S.H.; Cho, J. Sleep Duration and Mortality in Patients with Diabetes: Results from the 2007-2015 Korea National Health and Nutrition Examination Survey. *Diabetes. Metab.* **2022**, *48*, 101312. [CrossRef] [PubMed]
56. Brotherton, A.; Evison, F.; Gallier, S.; Sharif, A. Pre-Operative Waterlow Score and Outcomes after Kidney Transplantation. *BMC Nephrol.* **2022**, *23*, 273. [CrossRef] [PubMed]
57. Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
58. Mandrekar, J.N. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *J. Thorac. Oncol.* **2010**, *5*, 1315–1316. [CrossRef]
59. Fawagreh, K.; Gaber, M.M.; Elyan, E. Random Forests: From Early Developments to Recent Advancements. *Syst. Sci. Control. Eng. Open Access J.* **2014**, *2*, 602–609. [CrossRef]
60. Zhu, R.; Wang, Y.; Liu, J.X.; Dai, L.Y. IPCARF: Improving LncRNA-Disease Association Prediction Using Incremental Principal Component Analysis Feature Selection and a Random Forest Classifier. *BMC Bioinform.* **2021**, *22*, 175. [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.