IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.17, No.3, July 2023, pp. 259~270
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.83585

Autism Spectrum Disorder (ASD) Identification Using Feature-Based Machine Learning Classification Model

Anton Novianto[1], Mila Desi Anasanti*[2]
[1]Master Program of Computer Science, Universitas Nusa Mandiri, Jakarta, Indonesia
[2]Departemen of Information Studies, University College London, London, United Kingdom
e-mail: [1]14210243@nusamandiri.ac.id, *[2]**mila.mld@nusamandiri.ac.id**

Abstrak

Autism Spectrum Disorder (ASD) adalah gangguan perkembangan yang mengganggu perkembangan perilaku, komunikasi, dan kemampuan belajar. Deteksi dini ASD membantu pasien mendapatkan pelatihan yang lebih baik untuk berkomunikasi dan berinteraksi dengan orang lain. Dalam studi ini, kami mengidentifikasi individu ASD dan non-ASD menggunakan pendekatan machine learning (ML). Kami menggunakan K-Nearest Neighbor (KNN), Random Forest (RF), Regresi Logistik (LR), Naive Bayes (NB), Support Vector Machine (SVM) dengan fungsi basis linier dan Decision Tree (DT). Kami preprocessing data menggunakan metode imputasi, yaitu regresi linier, Mice forest, dan Missforest. Kami memilih fitur-fitur penting menggunakan teknik pemilihan dan peringkat fitur perturbasi Simultan (SpFSR) dari semua 21 fitur ASD dari tiga kumpulan data yang digabungkan (N = 1.100 individu) dari repositori University California Irvine (UCI). Kami mengevaluasi kinerja diskriminasi metode, kalibrasi, dan utilitas klinis menggunakan metode validasi silang 10 kali lipat bertingkat. Kami mencapai akurasi setinggi mungkin dengan menggunakan SVM dengan memilih 10 fitur terpenting. Kami mengamati integrasi imputasi data dengan linear model, SpFSR dan SVM sebagai model yang paling efektif, dengan tingkat akurasi 100% mengungguli studi prediksi ASD sebelumnya.

Kata kunci*— Gangguan Spektrum Autisme, machine learning, pemilihan fitur, imputasi*

Abstract

A developmental disease known as autism spectrum disorder (ASD) affects how people behave, communicate, and learn. Early detection of ASD helps patients to get better training to communicate and interact with others. This study identified ASD and non-ASD individuals using machine learning (ML) approaches. We used K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM) with linear basis function, and Decision Tree (DT). We preprocessed the data using the imputation methods: linear regression, Mice forest, and Missforest. We selected the essential features using the Simultaneous perturbation feature selection and ranking (SpFSR) technique from all 21 ASD features of three datasets combined (N=1,100 individuals) from the UCI repository at the University of California Irvine. We evaluated the performance of the method's discrimination, calibration, and clinical utility making use of the stratified 10-fold cross-validation technique. We achieved the highest accuracy possible using SVM, selecting the most essential ten features. We observed the integration of imputation using linear regression, SpFSR, and SVM as the most effective models, with an accuracy rate of 100%, outperformed the previous studies in ASD prediction.

Keywords— *Autism spectrum disorder, machine learning, feature selection, imputation*

Received April 2nd,2023; Revised July 27th, 2023; Accepted July 29th, 2023

1. INTRODUCTION

A collection of mental diseases known as autism spectrum disorders (ASD) are distinguished by some difficulty in social interaction and communication [1]. These difficulties include unusual patterns of activity and behavior, such as difficulty switching between activities, difficulty concentrating, and strange responses to sensations. However, a diagnosis of autism is frequently established much later in life, although symptoms may first appear in early childhood. Epilepsy, depression, anxiety, hyperactivity, attention deficit disorders, and other co-occurring disorders of the central nervous system, as well as risky behaviors, including difficulty falling asleep and self-harm, are frequently present in children with autism [2]. Children with autism spectrum disorders have various intellectual abilities, from severe conditions to higher levels [2].

Around one in 100 youngsters globally has autism [3]. Before 2000, there were 2-5 to 15-20 cases of autism per 1,000 live births or 1-2 cases per 1,000 people worldwide [4]. According to ASA (Autism Society of America) statistics from 2000, 1 in 250 people were autistic [4]. However, according to data from the CDC (Centers for Disease Control and Prevention, USA), there were 1 in 150 residents with autism in 2001, and it was between 100 people in various parts of the USA and the UK [4]. CDC also recorded that 1 in 88 children had autism in 2012, a rise of 30% to 1.50%, or 1 in 68, in 2014 [4]. The prevalence of ASD from 2000 to 2018 is shown in Table 1 [5]. ASD has been detected in around 1 in 44 children, according to estimates from the CDC's Autism and Developmental Disabilities Monitoring (ADDM) Network [6]. ASD has been reported in people of all races, ethnic, and economic and social groups and is more significant than four times as typical in boys than in girls [6].

The worldwide increase in the prevalence of ASD cases has prompted the need to compile behavioral trait-related data. It is difficult to conduct a thorough investigation to improve the efficacy, sensitivity, specificity, and predictive accuracy of ASD screening. There are currently few clinical or screening datasets about autism, most related to genes [7].

This study aimed to create an efficient ASD prediction approach based on crucial selected features by combining machine learning (ML) classification, imputation, and feature selection (FS) approaches. Recent ASD-related research has been conducted using various classification methods, but only some studies focused on the study of critical features. In a 2016 study, M. Duda et al. conducted research to develop six ML algorithms with an average prediction accuracy of 95.6% [8]. In 2018, Heinsfeld et al. conducted research using deep learning methods and achieved an accuracy of 70% [9]. In 2018, a study by Vaishali et al. using the binary firefly algorithm method achieved an accuracy of 92.12% [10]. In 2019, a Support Vector Machine (SVM) used by In-On Wiratsin et al. achieved a mean prediction accuracy of 90.8% [11]. SVM RFE (Support Vector Machine Recursive Feature Elimination) was used in a study by C. Wang et al. in 2019 that had a prediction accuracy of 90.6% [12].

Numerous factors can result in missing values, such as respondents who did not wish to be questioned or could not be located, data not collected owing to officer errors, equipment malfunctions, and application malfunctions. In addition, missing values can appear as outliers, discordant with the initial value [13], or anomalous data entries. Missing values can be associated with several issues, including inefficiency, difficulty handling and interpreting data, and anomalies or distortions between data containing missing values and complete data [13]. Therefore, additional processing is required to address the issue of missing values using the imputation technique.

Different algorithms and approaches to addressing missing values can result in different estimation outcomes. As a result, this study intends to enhance the ASD's predictive performance by incorporating imputation ML techniques and emphasising the relevant aspects with FS techniques.

2. METHODS

Conda version 22.9.0 and Python version 3.9.12 were used for all research analyses. Multiple modules, among them scikit-learn, a python machine learning module based on

"imbalanced-learn", "miceforest packages", "scipy package", and "missingpy", were utilized to generate and select the most critical features from the data. Figure 1 shows the complete steps taken to get the results for an ASD prediction:

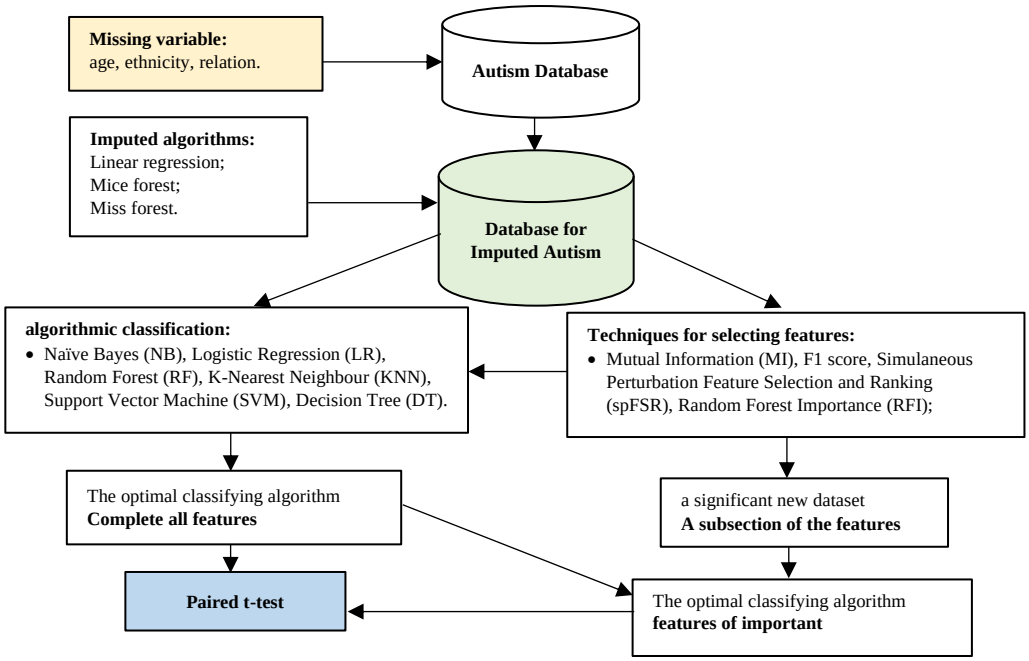

Figure 1. Research Methodology

2.1 Data Collection

This study used three publicly accessible datasets from the machine learning repository at the University of California Irvine [14]. The dataset includes data on routine health examinations from 1,100 participants, ages 4 to 64, gathered in 2017. Due to the similarity in data type and structure among the three datasets used in this investigation, they were merged into a single dataset to increase the power of prediction by taking advantage of increasing the sample size.

Since the dataset combines string and numeric data types, conversions must be made to several data types for the algorithmic methods used in this study to function correctly. The final completed dataset is shown in Table 1. It offers 21 qualities, consisting of 12 numeric variables, one categorical variable, and one category response variable (class).

Table 1. The early-stage ASD risk prediction dataset

No	Type of dataset	Attribute of dataset
1	Number	Age
2	String	Gender
3	String	Ethnicity
4	Boolean = yes/no	Jaundiced at birth
5	Boolean = yes/no	Family member with PDD (Pervasive Developmental Disorder) [18][19].
6	String	Who is completing the test
7	String	Country of residence
8	Boolean = yes/no	Used the screening app before
9	Integer = 0,1,2,3	Screening Method Type
10	Binary = 0, 1	(A1_Score), (A2_Score), (A3_Score), (A4_Score), (A5_Score), (A6_Score), (A7_Score), (A8_Score),
20	Integer	Screening Rating
21	Class/ASD	Class (ASD)

To use the imputation approach in this study, the dataset with the sign ("?") is converted to ("NA"), then imputed three using ML-based imputation methods proposed in this study [20]. We pre-processed the input by encoding the class target with a number between 0, 1, and so on. Using a set of independent data points, it ascertains the probabilities of a particular occurrence, such as participating in the vote or not participating in the vote, and then reports that probability. We also used normalization techniques, such as Min-Max scaling for the age variable, to adjust the range of the dataset so that it falls between 0 and 3, utilizing the highest and lowest possible values for each feature and changing the wrong data values to "NA" values with the total number of missing values. We compared the best imputation approach among those based on Linear Regression (LR) [21][22], Mice Forest (MC) [23], and Missforest (MF) [24].

2.2 Multiple Imputation Techniques

Several imputation techniques were used before classifying the data. We compared the performance of each imputation technique to obtain the best-imputed dataset to be integrated with the classifier.

2. 2.1 Linear Regression

Regression is one of the most frequently employed statistical methods. Regression is a form of model that is used to describe the actions of an intriguing random variable. This variable could be the stock market value in the financial sector, the development of a species, or the probability of detecting gravitational waves. It is the dependent variable and is denoted by "y". [21]. Consider the following multiple linear regression model: $Y = 1\theta_0 + X\theta + \epsilon$. Where: $Y = [y_1, y_2,\ldots, y_n]^1$ is a "n x 1" vector of responses, 1 is a "n x 1" vector of one, $X = [x_1, x_2,\ldots, x_q]$ is a "n x q" is a "n x q" non-stochastic design matrix, $\theta = [\theta 1, \theta 2,\ldots \theta q]^1$ is a q x 1 vector of unknown coefficients and $\epsilon = [\epsilon_1, \epsilon_2,\ldots \epsilon_n]^1$ is a vector of independent and identically distributed error terms [21].

2. 2.2 Mice Forest

The MICE (Multivariate Imputation by Chained Equations) algorithm is likely one of the most widely utilized imputation algorithms and a standard interview topic. MICE first calculates each column mean with a missing value and then uses the mean as a substitute [25]. It then executes a sequence of regression models (chained equations) to impute each missing value sequentially [25]. To apply the MICE algorithm: *Iterative Imputer (missing_values = np.nan, max_iter = 10, tol = 0.001, n_nearest_features = None, initial_strategy = 'mean', imputation_order = 'ascending', estimator = None, sample_posterior = False,)*

2. 2.3 Missforest

The Missforest method uses random forests to impute phenomics data. Missforest trained a random forest (RF) on the observed values for each variable, employing an iterative strategy for imputation, anticipating the missing values, and continuing until the stopping requirement was met. In addition, it can be executed in parallel to save computation time and to evaluate the OOB (out of bag) imputation error for the continuous and categorical portions of the imputed datasets. OOB is a method for calculating errors in random forest prediction. Observe the effectiveness of this evaluation by comparing the absolute difference between actual imputation error ($error_{true}$) and OOB imputation error ($error_{OOB}$) across all simulated iterations [26].

2. 3 Feature Selection

The methods of feature extraction and feature selection are two frequently used methods for decreasing data dimensionality. Feature extraction makes new features by mapping the

original features onto a new (lower-dimensional) space [27]. We used several feature selection techniques that represented each approach and were widely used in ML studies [28].

2. 3.1 Feature Selection F-Score

Feature selection is required for a classifier that may use many observational variables to choose a relatively limited subset of variables, decrease computation requirements, and enhance algorithm performance [29]. The F-score formula is as follows in equation (1):

$$F(i) = \frac{\left(\underline{x_i}^{(+)} - \underline{x_i}\right)^2 + \left(\underline{x_i}^{(-)} - \underline{x_i}\right)^2}{\frac{1}{n+-1}\sum_{k=1}^{n+}\left(\underline{x_{k,i}}^{(+)} - \underline{x_i}^{(+)}\right)^2 + \frac{1}{n--1}\sum_{k=1}^{n-}\left(\underline{x_{k,i}}^{(-)} - \underline{x_i}^{(-)}\right)^2} \quad (1)$$

Where: $\underline{x_i}^{(+)}, \underline{x_i}^{(-)}$ (The average of each i[th] feature across positive and negative datasets), $\underline{x_{k,i}}^{(+)}$ (The i[th] feature of the k[th] positive instance), $\underline{x_{k,i}}^{(-)}$ (The i[th] feature of the k[th] negative instance).

2. 3.2 Mutual Information (MI)

MI uses the amount of information when the variables exchange can be scaled, and the uncertainty of the random variables can be measured using information entropy [30]. Entropy can be represented as follows [31], as shown in equation (2):

$$H(X) = -\sum p(x) log\ \rho(x) \quad (2)$$

Where: $p(x)$ = marginal probability density.

Mutual reliance, which is defined as: can be measured by mutual information (MI), which is show in equation (3):

$$MI(Y;X) = \sum\sum \rho(y,x)\ log\ \frac{p(x,y)}{p(x)p(y)} = H(Y) - H(Y|X) \quad (3)$$

Where: $\rho(y,x)$ = joint probability density, and $H(Y|X)$ = conditional entropy at X is known, which is computed as show in equation (4):

$$H(Y|X) = \sum\sum p(y|x)\ log(p(y|x)) \quad (4)$$

As show in equation (5), continuous random variables, however,

$$H(X) = -\int_X\ p(x) log\ p(x)dx,\ H(Y|X) = -\int\int_{x,y} p(y,x) log(p(y|x))dxdy \quad (5)$$

$$MI(Y;X) = \int\int_{X,Y} p(y,x) log\ \frac{p(x,y)}{p(x)p(y)}\ dxdy \quad (5)$$

2. 3.3 Random Forest Importance (RFI)

Random Forests (RF) produce many distinct decision trees during the training phase. The mode of the classes for classification is the final prediction or the average forecast for regression, which is a combination of the projections from all trees. They are called ensemble methods because they rely on a set of outcomes to conclude. The probability of a node is computed by dividing the total number of pieces by the number of samples that arrive at the node. The greater the value, the greater the importance of the trait [32], as shown in equation (6).

$$n_i = \frac{N_t}{N}\left[impurit - \left(\frac{N_{t(right)}}{N_t}\ x\ right\ impurity\right) - \left(\frac{N_{t(left)}}{N_t}\ x\ left\ impurity\right)\right] \quad (6)$$

Where: N (total number of rows present in the data), N_t (number of rows in that specific note), $N_{t(right)}$ (number of nodes in the right node), $N_{t(left)}$ (number of notes in the left node), Impurity (a Gini index value).

2. 3.4 Simultaneous perturbation feature selection and ranking (spFSR)

SpFSR is a unique FS and rating technique that extends the stochastic optimization algorithm for general applications. SpFSR begins with the initial solution $\omega 0$ and utilizes recursion to determine the local minimum $\omega^{\wedge}$, as show equation (7):

$$\omega^{\wedge}_k + 1 := \omega^{\wedge}_k - a_k G^{\wedge}(\omega^{\wedge}_k) \quad (7)$$

Where: a_k (order of iteration gain); $a_k \geq 0$ and $G\hat{}(\omega\hat{}_k)$ are gradient estimations at *k*.

2.4 Techniques for machine learning classification

2.4.1 Logistic regression

Regression methods are now a required step in each data research describing the link between a response variable and one or more explanatory factors. Usually discrete, the outcome variable has two or more possible values. The most popular regression model for analyzing these data is the logistic regression model [33]. The conditional probability can characterize the link between the outcome variable y and the independent variable x if x = (x1, x2,..., xp) is a 1 p independent variable [34], as shown in equation (8).

$$p(y = 1) = \pi(x) = \frac{1}{1 + exp(-g(x))} \quad (8)$$

2.4.2 Random Forest

Based on regression trees, the supervised ensemble learners are random forests, a nonparametric model that learns variable interactions through recursive partitioning [35]. High-dimensional nonlinear issues are particularly well-suited for relationship detection by random forests [36]. However, their primary concern is classification or regression. They have just recently been used as time-series predictors [37]. CART is the primary approach for creating regression trees [35], which applies the subsequent formulation. To minimise the prediction error on the output space, z is used to partition the input space X into K regions M_k and assign an output value Y_{Mk} to each region. If a sum of squared errors is used to minimise the prediction error, the optimal output predictor f̀ for a new input observation x(t) is, at shown in equation (9).

$$\hat{Y}(t) = \hat{f}(x(t)) := \frac{1}{k} \sum_{i=1}^{k} Y_{MK} I(x(t) \in M_i) \quad (9)$$

Where: *I* (functional indication) [35].

2.4.3 Naïve Bayes (NB)

Knowing that the marginal density ratio is the best univariate classifier, we enhance this ratio by fusing the prior probability and the computed boundary, as shown in equation (10).

$$\{x : \pi f(x)/ (1-\pi)g(x)) = 1\} = \{x : \log \pi f(x) - \log (1-\pi)g(x)) = 0\} \quad (10)$$

Where: Class conditional densities for classes 0 and 1 are represented by the letters g and f, respectively; π shows the previous likelihood of class 1; $(1 - \pi)$ is the previous probability of class 0 [38].

2.4.4 Support Vector Machine (SVM)

Using the margins approach as its foundation, the SVM is a classification tool, where an ideal hyperplane can most effectively distinguish classes by lowering structural risk. This provides SVM with a robust ability to generalize and resistance to the issue of overfitting. In addition, SVM may handle nonlinear classification problems by selecting kernel functions to transfer a few high-dimensional feature spaces from the original feature space, which cases are linearly separable. Furthermore, SVM can perform novelty identification [39][38].

2.4.5 DT: Decision Tree

DT is a product of the community of Machine Learning (ML). Because multivariate statistics is a broad area of study within machine learning, computer science, bioinformatics, artificial intelligence, and some chemometrics, the notation and machine learning terminology is frequently different, commonly observed in the chemometrics literature. To aid the reader, the following glossary defines a few terms. This syntax is consistent with what is typically found in the machine learning literature [40].

2.4.6 K-Nearest Neighboar (KNN)

Allocating unlabeled observations to the class with the most comparable labelled samples is the goal of the KNN classifier. Both the training and test datasets gain observational properties [41]. By contrasting test dataset observations with training dataset observations, the KNN algorithm categorises test dataset observations. To assess the effectiveness of the KNN model, we are aware of the basic types of observations in the test dataset. The average accuracy, as given by the following equation, is one of the most often employed parameters. Average accuracy, as shown in equation (11):

$$the \sum_{i=1}^{1} \quad \frac{TPi + TNi}{TPi + FNi + FPi + TNi} / 1 \tag{11}$$

Where: TP, TN, FP, and FN stand for the true positive, false positive, and false negative, respectively. Category is indicated by the subscript *i*, while the word "l" stands for "total category," [41].

2.5 Evaluation

By comparing the results of each method, we can figure out which gives the best performance in accuracy by following this formula, as shown in equation (12):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{12}$$

FP (False Positive), FN (False Negative), TP (True Positive), TN (True Negative)

We utilized a stratified 10-fold cross-validation technique (max inter = 10) with three rounds to evaluate performance to reduce variability while maintaining computation speed. To enable future replication and independent confirmation of our results, the random state has been set to 999 (random state = 999). All feature selection techniques were installed and evaluated on the same data partition, and the random state was maintained throughout all processes for cross-validation. This method indicates that our experiments were conducted in pairs with far less variability than when performed individually. To determine whether there is a statistically significant performance difference between the two FS or ML techniques or whether the difference is the result of sampling variation, statistical tests are necessary because the cross-validation technique employs a random procedure. We did a combined t-test on the data before and after imputation to see if there were statistically significant differences between ML-FS methodologies and feature-based ML approaches. Using the "stats.ttest" function from the "Scipy" Python library, we run a combined t-test and then analyze the p-values. A p-value of less than 0.05 indicates that the difference is statistically significant at a 95% confidence level.

3. RESULTS AND DISCUSSION

3.1 Results and Analysis

Using all the features (21 features), Table 2 compares the performance of the ML techniques utilized in this work.

Table 2. The accuracies of various machine learning classification methods on all feature data imputed using different imputation methods

Classifiers		DT	RF	SVM	NB	LR	KNN	Average
Before Imputed		100%	100%	100%	96.60%	97.70%	93.40%	97.95%
Imputation Methods	LR	100%	100%	100%	96.90%	98.50%	93.70%	98.18%
	MiceFo	100%	100%	100%	96.90%	98.50%	93.20%	98.10%
	MissFo	100%	100%	100%	96.90%	98.50%	93.60%	98.16%

LR (Linear Regression), MiceFo (Mice Forrest) and MissFo (Miss Forest)

It is evident in Table 2 that all the imputation methods have slightly higher accuracies in predicting ASD compared to those using the unimputed data. All the imputation methods

performed almost similarly, with Linear Regression (LR) being the best.

3.1.1 Machine learning classification technique performance using all features

In conducting our test using a sample dataset from the LR imputation method, some unrelevant columns were removed, such as result, age_desc, ethnicity, country_of_res, used_app_before, relation, jaundice, and autism. We resulted in 13 features in total. Table 3 evaluates how well the ML algorithms used in this study worked with all available features.

Table 3. Values of various feature selection algorithms using all features

Algorithms	DT	RF	SMV	NB	LR	KNN
Accuracy	91.30%	96.30%	100%	95.40%	99.50%	94.30%
Precision	87.10%	96.40%	100%	94.90%	98.80%	88.70%
Recall	89.10%	93.20%	100%	92.30%	99.90%	96.50%
F1-Score	87.90%	94.70%	100%	93.50%	99.30%	92.40%

It was found that SMV performed the best than other ML algorithms, with LR coming in second. To establish whether the differing accuracies were statistically significant and did not occur by chance, we performed several paired t-tests. We found that all significant P-values were less than 0.05, with the most considerable P-value 0.001, which was the lowest possible value to compare the performance of SMV and LR. Thus, SMV is the finest method for 100% accuracy in predicting ASD.

3.1.2 Support vector machine classification technique performance with various number of features

We used the SMV method as a wrapper for classifiers in the ML-FS framework to determine as few features as possible to achieve the same result as the full-feature prediction. We began with four features and increased the number until we reached the same level of performance accuracy with all features, as can be seen in Table 4.

Table 4. Values of various feature selections with various numbers of features

FS techniques	Number of features						
	4	5	6	7	8	9	10
F-Score	88,3%	89,4%	90,9%	92,6%	95,2%	95,4%	100%
MI	88,3%	89,4%	90,9%	92,6%	92,4%	92,4%	95,4%
RFI	88,3%	89,4%	89,4%	90,9%	92,6%	95,2%	94,2%
SpSFR	88,5%	91,4%	90,8%	93,7%	92,2%	96,6%	100%

MI (Mutual Information), RFI (Random Forest Importance), SpSFR (Simulaneous Perturbation Feature Selection and Ranking).

To evaluate all the key indicator performances of accuracy, precision, recall, and F1-score, Table 5 presents the results of FS performances using ten key selection features.

Table 5. Values of various feature selection techniques utilizing ten features

FS techniques	F-Score Feature Importance	Mutual Information	RFI	SpFSR
Accuracy	100%	95.4%	94.2%	100%
Precision	100%	89.8%	93.1%	100%
Recall	100%	98.6%	90.7%	100%
F1-Score	100%	93.9%	9.17%	100%

The outcomes showed that the 10-feature spFSR-SVM and all techniques outperformed the other FS techniques. It also shows that the predictive ML performances using the full and ten key features can attain the same highest accuracy of 100%. With ten features, we could reduce the number of features while still determining the most critical aspects, which was the aim of this research. The top features of the spFSR are shown in Figure 2, together with the accompanying critical scores, indicating that the A9_Score is the most crucial variable in predicting ASD.

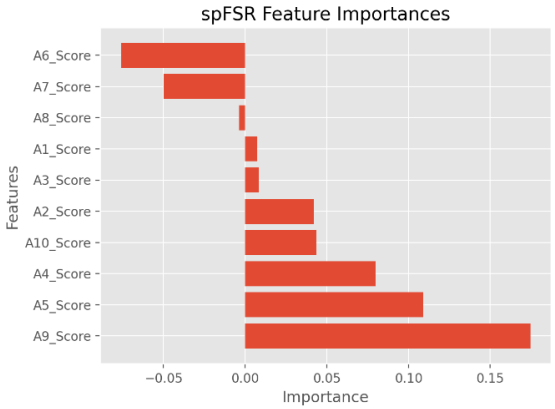

Figure 2. displays the importance of the top ten spFSR features

3.2 Discussion

Our study applied several FS approaches to a previously normalized ASD dataset. In diagnosing ASD, FS reduces the number of features, resulting in an accurate, efficient, and cost-effective prediction. The ASD prediction using SVM with the full features was equivalent to SVM with a subset of ten features, yielding the maximum predicted performance across all attributes (100% F1-score, recall, and accuracy). Importantly, we accomplished a similar outcome utilizing only the ten features chosen by spFSR, which incorporates spFSR-SVM.

Our study yielded the highest accuracy of 100% in predicting ASD using the FS technique; this exceeds previous studies, which only achieved accuracy between 70% and 95.6% [11][12][8][9][10]. However, all reported results cannot be directly compared due to different datasets and validation methodologies. The previous study by M. Duda et al., using the same dataset with complete features, only achieved the highest accuracy result of 95.6% using SVM [8]. By integrating the FS-ML approach, we can still achieve slightly higher accuracy by combining only half of the features that demonstrate the method's efficiency.

Then this research also provides an alternative method of ML classification by using fewer features, which is faster because collecting a complete set of features will require more effort, time, additional costs, and computational complexity [42]. The FS technique aims to reduce variables and adequately represent relevant and needed data. Integrating FS techniques into ML methods can aid in the efficient and low-cost prediction of ASD.

From the results of our study, the ten recommended features for predicting ASD attributes are A6_Score, A7_Score, A8_Score, A1_Score, A3_Score, A2_Score, A10_Score, A4_Score, A5_Score and A9_Score. These features become essential for the ten-feature method, especially the outstanding A9_Score feature where the "Usually, I can tell what someone is feeling or thinking by looking at their face" classification is high enough to affect the patient's condition, which can result in ASD status. The findings also indicate that emotional challenges including the inability to discern another person's feelings are common in people with autism. There isn't much scientific evidence to support the idea that this trait is a component of autism, despite the fact that this trait is nearly universally acknowledged as such.

Although this research has many advantages, it is constrained. Since the quantity of data provided is not high-dimensional, we train all the complete data using the FS method. Then we test it using an iterative cross-validation procedure on the full dataset. It can result in overfitting with a simple technique. The combined split-train-test approach will be suggested for a better strategy. The data set can be split into training and test halves, and the most significant key features in the training data can then be selected using cross-validation procedures. The performance of the features on the test data can be re-evaluated using iterative cross-validation approaches. Another way to ensure the same high accuracy can still be attained is to replicate the procedure on different datasets.

4. CONCLUSIONS

The computational complexity of disease diagnosis will be reduced by incorporating spFSR in the SMV approach. In this study, an accuracy of 100% was achieved by using ten features, representing the highest performance. This study shows that to accurately and reliably predict ASD on the initial dataset, only half of the features can be proposed for efficiency while highlighting the most important ones. Future studies could consider applying this approach to larger or different disease datasets.

REFERENCES

[1] M. M. Rahman, O. L. Usman, R. C. Muniyandi, S. Sahran, S. Mohamed, and R. A. Razak, "A review of machine learning methods of feature selection and classification for autism spectrum disorder," *Brain Sciences*, vol. 10, no. 12. MDPI AG, pp. 1–23, Dec. 01, 2020. doi: 10.3390/brainsci10120949.

[2] "Gangguan Spektrum Autisme (ASD)." https://pantirapih.or.id/rspr/tag/gangguan-spektrum-autisme-asd/.

[3] "WHO Autism." https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders.

[4] "HARI PEDULI AUTISME SEDUNIA." https://kemenpppa.go.id/index.php/page/read/31/1682/hari-peduli-autisme-sedunia-kenali-gejalanya-pahami-keadaannya.

[5] "Data & Statistics on Autism Spectrum Disorder." https://www.cdc.gov/ncbddd/autism/data.html.

[6] M. J. Maenner *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.*, vol. 70, no. 11, pp. 1–16, 2021, doi: 10.15585/MMWR.SS7011A1.

[7] C. for D. C. and P. National Center on Birth Defects and Developmental Disabilities, "Autism Spectrum Disorder. Data & Statistics on Autism Spectrum Disorder.," *Https://Www.Cdc.Gov/Ncbddd/Autism/Data.Html*, pp. 1–2, 2022, [Online]. Available: https://www.cdc.gov/ncbddd/autism/data.html.

[8] M. Duda, R. Ma, N. Haber, and D. P. Wall, "Use of machine learning for behavioral distinction of autism and ADHD," *Transl. Psychiatry*, vol. 6, no. 2, pp. 1–5, 2016, doi: 10.1038/tp.2015.221.

[9] A. R. Franco, R. C. Crad-, A. Buchweitz, and F. Meneguzzi, "AC Identification of Autism Spectrum Disorder using Deep," 2017, doi: 10.1016/j.nicl.2017.08.017.

[10] V. Ravindranath and S. Ra, "A machine learning based approach to classify Autism with optimum behaviour sets," no. August, 2018, doi: 10.14419/ijet.v7i3.18.14907Published.

[11] I. S. Here, I. Wiratsin, and L. Narupiyakul, "Feature Selection Technique for Autism Spectrum Disorder Insert Subtitle Here," pp. 53–56.

[12] C. Wang, Z. Xiao, and J. Wu, "Physica Medica Functional connectivity-based classi fi cation of autism and control using SVM-RFECV on rs-fMRI data," *Phys. Medica*, vol. 65, no. August, pp. 99–105, 2019, doi: 10.1016/j.ejmp.2019.08.010.

[13] I. J. Fadillah and S. Muchlisoh, "PERBANDINGAN METODE HOT-DECK IMPUTATION DAN METODE KNNI DALAM MENGATASI MISSING VALUES Penerapan Pada Data Susenas Maret Tahun 2017," vol. 2017, no. March, pp. 275–285, 2017.

[14] "Center for Machine Learning and Intelligent Systems." https://archive.ics.uci.edu/ml/datasets.php.

[15] "Autism Screening Adult Data Set." https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult.
[16] "Autistic Spectrum Disorder Screening Data for Children Data Set." https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++.
[17] "Autistic Spectrum Disorder Screening Data for Adolescent Data Set." https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent+++.
[18] "Pervasive Developmental Disorder (PDD)." https://kidshealth.org/CHOC/en/parents/101543.html.
[19] F. Thabtah and D. Peebles, "A new machine learning model based on induction of rules for autism detection," *Health Informatics J.*, vol. 26, no. 1, pp. 264–286, 2020, doi: 10.1177/1460458218824711.
[20] G. Chhabra, V. Vashisht, and J. Ranjan, "A Comparison of Multiple Imputation Methods for Data with Missing Values," *Indian J. Sci. Technol.*, vol. 10, no. 19, pp. 1–7, 2017, doi: 10.17485/ijst/2017/v10i19/110646.
[21] Q. H. Luu, M. F. Lau, S. P. H. Ng, and T. Y. Chen, "Testing multiple linear regression systems with metamorphic testing," *J. Syst. Softw.*, vol. 182, p. 111062, 2021, doi: 10.1016/j.jss.2021.111062.
[22] H. Tunc and B. Genç, "A column generation based heuristic algorithm for piecewise linear regression," *Expert Syst. Appl.*, vol. 171, no. September 2019, p. 114539, 2021, doi: 10.1016/j.eswa.2020.114539.
[23] S. Van Buuren, "Package 'mice' Title Multivariate Imputation by Chained Equations," 2014.
[24] P. Arriagada, B. Karelovic, and O. Link, "Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm," *J. Hydrol.*, vol. 598, no. May, p. 126454, 2021, doi: 10.1016/j.jhydrol.2021.126454.
[25] I. Ismiguzel, "Imputing Missing Data with Simple and Advanced Techniques," pp. 1–15.
[26] M. K. Saggi and S. Jain, "Reference evapotranspiration estimation and modeling of the Punjab Northern India using deep learning," *Comput. Electron. Agric.*, vol. 156, no. October 2018, pp. 387–398, 2019, doi: 10.1016/j.compag.2018.11.031.
[27] V. Aksakalli, Z. D. Yenice, M. Malekipirbazari, and K. Kargar, "Feature selection using stochastic approximation with Barzilai and Borwein non-monotone gains," *Comput. Oper. Res.*, vol. 132, no. April, p. 105334, 2021, doi: 10.1016/j.cor.2021.105334.
[28] Mila Desi Anasanti, Khairunisa Hilyati, and Annisa Novtariany, "The Exploring feature selection techniques on Classification Algorithms for Predicting Type 2 Diabetes at Early Stage," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 5, pp. 832–839, 2022, doi: 10.29207/resti.v6i5.4419.
[29] W. Huang, H. Yan, R. Liu, L. Zhu, H. Zhang, and H. Chen, "F-score feature selection based Bayesian reconstruction of visual image from human brain activity," *Neurocomputing*, vol. 316, pp. 202–209, 2018, doi: 10.1016/j.neucom.2018.07.068.
[30] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8520–8532, 2015, doi: 10.1016/j.eswa.2015.07.007.
[31] J. Tang, Z. Wu, M. Jia, and Z. Liu, "Mutual Information-Based Modified Randomized Weights Neural Networks," *J. Comput. Commun.*, vol. 03, no. 11, pp. 191–197, 2015, doi: 10.4236/jcc.2015.311030.
[32] "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark." https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

[33] E. R. Ziegel and S. Menard, *Applied Logistic Regression Analysis*, vol. 38, no. 2. 1996. doi: 10.2307/1270433.

[34] X. Song, "Parameterized fragility analysis of steel frame structure subjected to blast loads using Bayesian logistic regression method," *Struct. Saf.*, vol. 87, no. June, p. 102000, 2020, doi: 10.1016/j.strusafe.2020.102000.

[35] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Breiman Classification and regressions Trees. Wadsworth Int," *Group*, vol. 37, no. 15, pp. 237–251, 1984.

[36] M. Benali, J. Jeaidi, B. Mansoornejad, O. Ajao, B. Gilani, and N. Ghavidel Mehr, "Decision support systems for assessment of biorefinery transformation strategies," *Can. J. Chem. Eng.*, vol. 96, no. 10, pp. 2155–2175, 2018, doi: 10.1002/cjce.23301.

[37] A. Lahouar and J. Ben Hadj Slama, "Random forests model for one day ahead load forecasting," *2015 6th Int. Renew. Energy Congr. IREC 2015*, 2015, doi: 10.1109/IREC.2015.7110975.

[38] J. Gu and S. Lu, "An effective intrusion detection approach using SVM with naïve Bayes feature embedding," *Comput. Secur.*, vol. 103, p. 102158, 2021, doi: 10.1016/j.cose.2020.102158.

[39] W. Alam, Q. Khan, R. A. Riaz, R. Akmeliawati, I. Khan, and K. S. Nisar, "Gain-Scheduled Observer-Based Finite-Time Control Algorithm for an Automated Closed-Loop Insulin Delivery System," *IEEE Access*, vol. 8, no. May, pp. 103088–103099, 2020, doi: 10.1109/ACCESS.2020.2997776.

[40] S. D. Brown and A. J. Myles, "Decision Tree Modeling," *Compr. Chemom.*, pp. 625–659, Jan. 2020, doi: 10.1016/B978-0-12-409547-2.00653-3.

[41] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Ann. Transl. Med.*, vol. 4, no. 11, pp. 1–7, 2016, doi: 10.21037/atm.2016.03.37.

[42] F. H. Alfebi, M. D. Anasanti, P. Studi, S. Ilmu, and U. N. Mandiri, "Improving Cardiovascular Disease Prediction by Integrating Imputation , Imbalance Resampling , and Feature Selection Techniques into Machine Learning Model," vol. 17, no. 1, pp. 55–66, 2023.