# Behaviour Analysis using Statistical Learning Assisted Fuzzy Rough Set Theory (SL-FRST): A Comprehensive Investigation on Autism Spectrum Disorder (ASD)

Abinaya S[1]
Ph.D. Research Scholar
Department of Information Technology
Bharathiar University
Coimbatore – 641046
contactdrabinaya@gmail.com

Dr. W. Rose Varuna[2]
Assistant Professor
Department of Information Technology
Bharathiar University
Coimbatore – 641046
rosevaruna@buc.edu.in

**Abstract:** *Autism Spectrum Disorder (ASD)* indicates a developmental disorder where they face issues in communication, societal interaction, repetitive behaviour, and tampering. Every individual faces diverse weaknesses and strengths, and wide variations of severity or symptoms can be spotted. Behaviour analysis plays a prominent role in understanding the nature and treating the children. Examining behaviour can give a systematic approach for assessing, intervening, and understanding the behaviour issue related to the disorder. Incorporating statistical analysis and fuzzy-based behaviour analysis can assist *ASD* children. Overall, well-being and quality of life can be enhanced by analyzing behaviour. Initially, the data is scaled to normalize the features, and statistical analysis is applied. *Multivariate Data Analysis (MDA)* is utilized to identify the complex relationships and patterns from within the dataset. *MDA* permits exploring interrelationships among multiple variables, which interact with each other and can collectively impact the outcome. This research uses Principal Component Analysis (PCA) as an *MDA* technique. Analysis and predictions are more accurate and robust. *Fuzzy Rough Set Theory (FRST)* handles complex data, and the fuzzy sets are determined using the feature score. The fuzzy rules are generated to define the feature relation, which assists in behaviour analysis. The experimental results of the proposed *Statistical Learning Assisted Fuzzy Rough Set Theory (SL-FRST)* outperform the existing state-of-the-art technique.

**Keywords:** Behavior analysis, fuzzy rules, intervals, multivariate data, relationship, and pattern.

## Introduction

*Autism Spectrum Disorder (ASD)* is a neurological disorder which disturbs the gaining of linguistic, communicative, cognitive, and social capabilities [1]. Earlier identification of mild symptoms of *ASD* is a tedious task since the different types of neurological diseases pose almost similar symptoms. The traditional *ASD* diagnosis process occurs in the clinical environment with experienced professionals, which is lengthy and costly [2]. Recently, the rise of *Machine Learning (ML)* models acquired promising results in diagnosing neuropsychiatric illness. The *ML* models considered the *ASD* diagnosis process a classification process where the prediction approaches are designed depending on the past data. The *ASD* classification technique ML improves the diagnostic accuracy of *ASD [3]*.

Earlier disease prediction is more essential, which helps to extend the survival rate [4]. Similarly, *ASD* should also be predicted earlier as it enables the development of numerous medical-based models [5]. It is depicted that previous intervention is more efficient in reducing cognitive inabilities and tends to increase the positive outcomes for a toddler. Extensive work with massive children has been addressed with *ASD* affected by genetic as well as environmental aspects [6]. Here, the ecological aspects are a range of influences like parental age, obstetric state, vaccinations consumed, maternal food consumption, exposure to drugs at the prenatal stage, prenatal stress, and so on. An environmental factor relevant to the risk of *ASD* can communicate with one another by making the etiologic of *ASD* highly complicated [7].

In case of complication, the detection is completed at the neonatal phase, which guides in isolating the impacts of postnatal environment risk issues of *ASD* and enhances the knowledge of *ASD* [8]. Additionally, *ASD* is nearby normal development at a young age, before the environmental factors have affected the postnatal deployment [10]. Suppose a child is cheerful about ASD at an earlier stage [9]. In that case, few ecological factors are involved in *ASD* development after birth, which means that primary aetiology and the manifestation have massive homogenous objectives that make simple detection and learn the aetiology contributed to *ASD* by using biomarkers at an early age [11]. Several research works have utilized *ML* models in *ASD* diagnosis, and datasets applied in experiments have different class labels, such as *ASD* and Non-ASD. It considers it as a binary classification issue and avoids the circumstance where *ASD* is degraded as "Light Autism", "Severe Autism", and so on [12].

Furthermore, the autism classification issues are classified into *ASD* and non-*ASD* in the dataset by eliminating the instances with general features and *Pervasive Development Disorder (PDD)* classes like *ADHD* as well as Asperger Syndrome [13]. Specific algorithms have eliminated the preprocessing stage, and it is unclear whether ASD or non-ASD has been used [14]. Such instances are highly complicated to examine as they come under the numerous classes that collapse the ML method in the detection phase and enhance the prediction [15]. In-depth data analysis has been identified before using ML, where a dataset is collected under hierarchical clustering that enables data from *ASD* to be reduced as multiple hierarchies with the application of various agglomerative models [16]. It improves the data presentation before the learning phase and generates beneficial outcomes for multiple stakeholders [17]. The application of predictive methods has produced multi-label classifiers. It also enables additional knowledge and reduces the volume of tossed knowledge in the training phase [18]. Then, cases and controls of *ASD* share general features in overlapping various PDD classes that have recommended multi-label methods with excess knowledge showing overlapped data instances in diagnosis types since the rules have multiple labels. Hence, the application of a multi-label mechanism requires effective data transformation.

*Motivation*

*ASD* is a brain development disorder that mitigates some of the activities and social behaviours resulting from natural cell development. Genetic and neurological factors cause autism. Apart from genetic causes, *ASD* is examined by applying behavioural predictors like social communication, thinking capability, repeated behaviours, and interaction function. A child with *ASD* suffers from earlier developmental complexities than alternate infants. These behavioural actions differ from each other, such as responding to sensory details like hearing, smelling, tasting, and so forth, lags in language acquisition, complicated interaction, impacting early learning, and tedious communication ability.

To maximize the analysis procedure of ASD, the developers have recently concentrated on applying ML-based intelligent models. The central premise of the ML approach is to enhance the *ASD* diagnosing time for accessing health care services, maximize the diagnosing accuracy, and limit the dimension of the input dataset to find the maximum graded features of ASD. Then, the ML model is unified with mathematics, AI, search models, and sciences to retrieve exact prediction approaches from datasets. Some ML technologies are NNs, SVM, DT, and rule-based classifiers, which are considered automatic devices that need a minimum workforce in data computation. The software package samples incorporated with ML models are R, Scikitlearn, Statistics and ML, MATLAB toolbox, and WEKA.

### *Research Contribution*

The research contribution of using Statistical Learning Assisted Fuzzy Rough Set Theory (SL-FRST) for behaviour analysis in children with autism spectrum disorder (ASD) lies in its ability to integrate advanced statistical learning techniques with fuzzy rough set theory to improve the accuracy and interpretability of behavioural analysis results. The research contributions:

- SL-FRST improves the reliability of behaviour analysis models. These algorithms can find out sophisticated relations and tendencies in the data which conventional statistical techniques are fail to find, which enhances the accuracy of the models.
- The fuzzy rough set theory offers ways of dealing with the uncertainty and the vagueness in the behavioural data set to enhance the interpretability of the analysis results to clinicians and researchers. On this basis, SL-FRST uses the fuzzy rough set theory with learning capabilities in conjunction with statistical learning methods to provide a deeper understanding of other behavioural characteristics of children with ASD.
- SL-FRST allow for the individual specific behavior and learning styles in children with ASD by permitting the modeling of different behaviors. Due to the focus on features and relationships with a child, the approach can develop interventions and supportive measures that can be individual for every child.

## Related Works

The supervised learning mechanism suggested in this work encloses classification models developed for identifying patterns in applied datasets, resulting in accurate diagnosis. Massive studies have described various supervised ML approaches with numerous supreme approaches deployed from a group. This section explains diverse research approaches in *Autism Spectrum Disorder (ASD)*. Logistic Regression (LR) estimates probabilities using a sigmoid function suitable for binary classification. A nonlinear Support Vector Machine (SVM) finds optimal hyperplanes to separate data points, which is practical for linear and nonlinear classification. Light Gradient Boosting Machine (LGBM) uses gradient boosting with decision trees, excelling in handling large datasets and achieving high accuracy [19].

Researchers have developed filter, wrapper, and embedded FS investigation to find the feature redundancies among SRS queries. To identify the linearity of SRS-based *ASD* infection (Washington et al., 2020) [20]. The 65-question SRS is compressed into low-dimension representations under the application of Principal Component Analysis (PCA), distributed stochastic neighbourhood embedding (t-SNE), and denoising autoencoder (DAE). Hence, the working functions of MLP classification with top-ranking queries as input.

n (Tang et al., 2019) [21], entire-brain functional systems accomplished from resting-state functional Magnetic Resonance Imaging (fMRI) were applied extensively for examining brain infections such as ASD. Automated classification of ASD is a more significant objective that yields better accuracy. Previous classifiers for *ASD* categorization depend upon the features obtained from the whole-brain functional system, which is not different for best function. In (Devika Varshini and Chinnaiyan, 2020) [22], the efficiency of numerous ML and preprocessing methods for classifying clinical datasets was applied to detecting primary autism symptoms in a child and adult. Massive works have applied preprocessing and ML methodologies for computing significant classification tasks. However, simple preprocessing phases are integrated with suitable data encoding in this work.

The primary objective (Abdolzadegan et al., 2020) [23] is to project an effective mechanism for a previous diagnosis of *ASD* from an electroencephalogram (EEG) signal. Here, the study population is composed of children with *ASD* and typical children from the same age group. In this model, linear and nonlinear features like Power Spectrum, Wavelet Transform (WT), Fast Fourier Transform (FFT), etc. Additionally, density-based clustering has been employed to eliminate noise and boost efficiency. Following this, FS has been used based on conditions like Mutual Information (MI), IG, mRmR, and Genetic Algorithm (GA). As a result, K-Nearest-Neighbor (KNN) and SVM classification models are employed for consequential decisions.

In (Xu et al., 2020) [24], a novel method has been developed for estimating the global time-varying nature of the human brain by measuring the change in 1st order of statistical features from the fNIRS time sequence. This is followed by a DL approach integrating LSTM and Convolutional Neural Network (CNN) to make the combinational procedure

with an enhanced bagging scheme to explore effective patterns from temporal differences for *ASD* exploration.

The recent studies aim for an ML model for automated *ASD* diagnosis, while the developers have applied previous ML approaches and used them on autism datasets. The ML works for autism have applied previous software packages for prediction approaches like WEKA, R, and LIBSVM, where input datasets with ASD, Attention Deficit Hyperactivity Disorder (ADHD), non-*ASD* cases and controls are applied. Some of the commonly employed methods are training phase in SVM as depicted in (Kosmicki et al., 2015) [25], Logistic Regression (LR) and Decision Tree (DT) like (Wall et al., 2012) [26] and Self-organizing map (SOM) and Naïve Bayes (NB) like (Pratap et al., 2014) The key objective of these models is to improvise the performance while distinguishing among *ASD* and ADHD [27].

The research proposes a Takagi-Sugeno-Kang (TSK) fuzzy modelling technique for early recognition and severity approximation of Autism Spectrum Disorder (ASD) utilizing EEG signals. Utilizing subtractive clustering and Short Time Frequency Transformation (STFT), the method achieves an accuracy range of 70-97% for ASD classification and 80-100% for crisp decision-making. The model can potentially aid psychiatrists in ASD diagnosis and intervention processes [28]. The research aimed to develop and validate a hierarchical fuzzy autism detection tool called "Fast and Accurate Diagnosis of Autism." This graphical user interface facilitates quick and accurate autism diagnosis, highlighting highly impaired areas in participants. Tested on two groups (autism and normal, N=40 each), the tool achieved 99% accuracy in discriminating between autistic and normal participants. It demonstrated high sensitivity (98.2%) and specificity (99.2%), aiding doctors in diagnosing autism and identifying severity levels efficiently [29].

This study addresses the lack of attention towards developing an autistic triage method for ASD patients, aiming to classify them based on severity using Fuzzy Multi-Criteria Decision Making (fMCDM) methods. Two phases were conducted: data preprocessing and method development. The Processes for Triaging Autism Patients (PTAP) method categorizes patients into minor, moderate, and urgent levels. The triage method achieved high sensitivity (86.67%-90.91%), specificity (88.46%-94.29%), and accuracy (84.78%-93.48%). Utilizing medical and sociodemographic criteria, the developed tool offers promise for early autism diagnosis and clinical treatment support. Correlation analysis highlights the importance of specific criteria, enhancing the triage method's efficacy [30].

A distinct edition of the input dataset undergoes training to enhance the previous metrics and generate an optimal function for the *ASD* prediction method. Therefore, the prediction powers of ASD classification systems in recent studies depend on input features apart from sampling and FS models. The classification task in ML is automated and not a separate issue with static training data, which has undergone training with the help of ML technologies. Additionally, it becomes more complex to a greater extent. It is desired that the task of diagnosing *ASD* be independent and with the help of trained medical employees in making decisions in a clinic.

The gap in literature found in the research is the absence of new methods to improve the efficacy and precision of the diagnosis of autism spectrum disorder (ASD) through ML methods. To the best knowledge of the author, previous work has focused on applying supervised ML models for ASD diagnosis such as logistic regression (LR), support vector machine (SVM), and light gradient boosting machine (LGBM), but the redundancy of the SRS queries still has not been analyzed, as well as FS methods were not examined. Moreover, current ML models have a strong dependence on conventional SW tool sets and approaches and do not consider more sophisticated techniques or dynamic training datasets. Additionally, there is a lack of direct application of ML technologies in clinical work, as diagnosis of ASD should include trained clinicians making rationale decisions based on the data provided by ML algorithms. In conclusion, there is a need to come up with new strategies which can fill these gaps so as to enhance the quality of ASD diagnosis.

Reducing dimensions of the data obtained from the SRS through Principal Component Analysis (PCA) increases interpretability and highlights the most significant factors associated with ASD – the behaviour patterns. On the other hand, the Fuzzy Rough Set Theory deals with uncertainty using fuzzy set for each of the features to be able to accommodate any changes within the data. Combined, they enhance the diagnostic accuracy of an ASD and provide more understandable and robust classification models for clinical practice.

**Proposed Methodology: Statistical Learning Assisted Fuzzy Rough Set Theory (SL-FRST)**

In this research, the dataset is generated from the questionnaire from the *Modified Checklist for Autism in Toddlers (M-CHAT),* where the scoring is accomplished using the *Indian Scale for Assessment of Autism (ISAA). F*eatures related to every question are associated with the behaviour of the child. Initially, the values are normalized in a specified range and transformed using a feature selection technique, Principal Component Analysis (PCA). The disease is classified by generating fuzzy rules based on the fuzzy rough set t*heory (FRST)*. Based on the inter-relationship of diverse behaviour, the level of autism is classified as No, Mild, Moderate, and Severe Autism. The entire process is detailed in this section. The overall architecture is given in Figure 1.

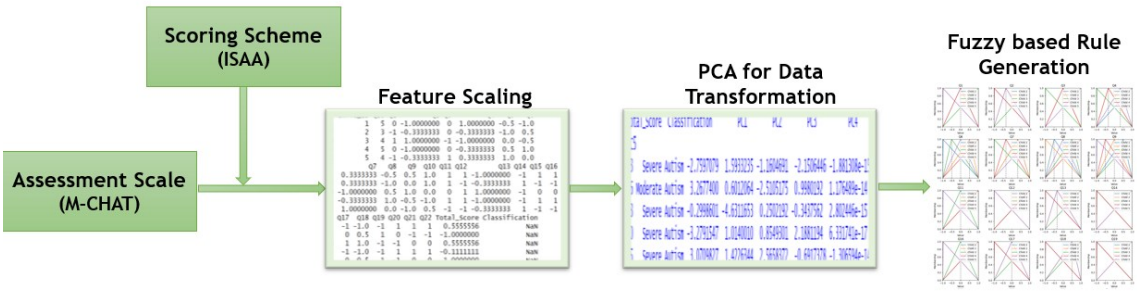


Figure 1. The overall methodology of Statistical Learning Assisted Fuzzy Rough Set Theory (SL-FRST)

Statistical Learning Assisted Fuzzy Rough Set Theory (SL-FRST) is a new approach to data analysis that analyses the data in the complicated fields, such as behavioral analysis in Autism Spectrum Disorder (ASD). This extensive research examines the use of SL-FRST in elucidating behavioral features of ASD, its theoretical background, and significance. SL-FRST combines statistical learning with fuzzy rough set theory that can be used as a reliable approach to analyse behavioural data and its uncertainty and imprecision. It enriches the traditional rough set models with statistical components that increase classification accuracy and reduce complexity. In the case of ASD, where behavior depends on various factors and differs significantly, SL-FRST is a useful approach for selecting features, determining their classification, and making decisions.

SL-FRST behavior analysis starts with data pre-processing where noisy and missing values are dealt with before proceeding to feature extraction to determine the behavioral characteristics. The fuzzy rough set approach then defines approximation spaces that retain the inherent uncertainty in behavioral data and provides a precise classification of the ASD-related behaviors. This research emphasises on how SL-FRST can identify weak features that are associated with ASD, thus enabling early detection and intervention. In this way, employing a more flexible hybrid design, researchers can reveal new associations between behavioral variables and ASD diagnostic criteria.

SL-FRST is not only used for diagnosis of a disease but also for therapeutic monitoring and predicting outcome. By performing real-time analysis of the behavioral data collected from patients through SL-FRST, it is possible to monitor the changes in the patients' reactions to the interventions applied and make adjustments to the interventions in real time to enhance their effectiveness and achieve better long-term outcomes for patients with ASD. Through this research, it is evident that Statistical Learning Assisted Fuzzy Rough Set Theory has the capacity to revolutionize the possibility of behavioral analysis of Autism Spectrum Disorder. With the help of statistical learning techniques and fuzzy rough set theory, SL-FRST does not only improve the understanding of the behaviors related to ASD but also opens the new opportunities for the development of the diagnostic and therapeutic approaches for the patients with the ASD based on the individual needs.

**Feature Scaling**

Normalization is a widely used feature scaling; every feature is scaled to the standard deviation of *1* and mean *0*. The features are linearly scaled based on minimum and maximum values, which lie in the range of [-1, 1]. The feature scaling is accomplished using Equation 1.

$$X_{scaled} = \frac{2(X - X_{min})}{(X_{max} - X_{min})} - 1 \text{---------} (1)$$

where the minimum and maximum value in the feature is indicated as $X_{min}$ and $X_{max}$.

**Transformation using Principal Component Analysis (PCA)**

Let X be the $n \times p$ standardized dataset, where $n$ is the count of samples and $p$ is the count of the variable. The process of standardization assures every variable has unit variance and zero mean. The value of covariance matrix $X$ is estimated in Equation 2.

$$C = \frac{1}{n} X^T X \text{---------(2)}$$

The eigendecomposition is computed on the covariance matrix using Equation 3.

$$C = V \Lambda V^T \text{---------(3)}$$

where the eigenvector matrix is indicated as $V$, and the eigenvalue diagonal matrix is shown as $\Lambda$.

The eigenvalue calculation and normalization are accomplished using Equations 4 and 5.

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^T (x_i - \bar{x}) \text{---------(4)}$$

$$u_i = \frac{v_i}{\|v_i\|} \text{---------(5)}$$

where the eigenvalue is indicated by $\lambda$ and the normalized vector is indicated by $u_i$.

The projection rate at the principal component is calculated using Equation 6.

$$y_i = U^T x_i \text{---------(6)}$$

where the projected data is $y_i$, original data is $x_i$, and the matrix of eigenvectors are $U^T$.

The eigenvalues are sorted in descending order, and the eigenvectors are rearranged to their relevant position. The top k eigenvectors are related to the highest eigenvector to form the $p \times k$ matrix $V_k$, where the *PCA* count is indicated using $k$.

The standardized dataset is projected to the subspace spanned by the designated eigenvectors. The principal component is estimated using Equation 7.

$$Z = XV_k \text{---------(7)}$$

where $Z$ is the $n \times k$ matrix of principal component scores.

After acquiring the principal component scores $Z$, the dataset is efficiently transmitted into a minimal dimensional space with minimized complexity. The prominent features are retrieved to enhance the performance of the subsequent algorithm.

**Fuzzy Rough Set Theory-based Rule Generation**

For every feature, fuzzy sets are determined based on score ranges [-1 to 1]. Let $X_i$ indicate the feature $i^{th}$ and $x_{ij}$ indicate the feature score for $j^{th}$ in $X_i$. The fuzzy sets $A_{ij}$ are determined for every score $x_{ij}$ within the feature set $X_i$. The degree value indicates the fuzzy sets membership function relies on the fuzzy sets. The linguistic labels, namely low, medium and high, are determined by the fuzzy sets. The degree of membership value is assigned accordingly, and the fuzzy sets are partitioned based on the membership score range.

Every feature $X_i$ with a sequence of values $x_{ij}$ determined by fuzzy sets $A_{ij}^k$ where the value of k is 1, 2,……m. The process is given in Equation 8.

$$A_{ij} = \{\mu A_{ij}(x), where\ x \in X_i\ and\ \mu A_{ij}(x) \text{---------} (8)$$

The fuzzy partitioning is converted into continuous intervals, and the membership degree is determined for every interval. The fuzzy discretization process handles the complexity and uncertainty in the dataset. The membership function for the autism feature is given in Equation 9.

$$\mu_{A_i}(x) = \begin{cases} 1 & if\ x\ fully\ belong\ to\ A_i \\ 0 & if\ x\ doesnt\ belong\ to\ A\_i \\ in-between & for\ partial\ membership \end{cases} \text{---------} (9)$$

The logical operations on autism features are given in Equations 10 to 12.

$$A_i \cup B_i = \{(x, \max\left(\mu_{A_i}(x), \mu_{B_i}(x)\right))|x \in X_i\} \text{---------} (10)$$

$$A_i \cap B_i = \{(x, \max\left(\mu_{A_i}(x), \mu_{B_i}(x)\right))|x \in X_i\} \text{---------} (11)$$

$$\overline{A_i} = \{(x, 1 - \mu_{A_i}(x)|x \in X_i\} \text{---------} (12)$$

The feature $X_i$ range is partitioned into fuzzy intervals $m$. The values are partitioned in the range of $R_i=[a_i, b_i]$ of every feature $X_i$ into m range of intervals $I_{ij}^k$ that is given in Equation 13.

$$I_{ij}^k = C_{ij}^k, d_{ij}^k \text{---------} (13)$$

The membership degree is estimated using Equation 14, and it is for $x_{ij}$ feature $X_i$ for every interval of fuzzy $I_{ij}^k$. The appropriate function determines the degree.

$$\mu_{A_{ij}^k}(x_{ij}) \text{---------} (14)$$

The uncertainty of the data is estimated by fuzzy upper and lower approximation. The approximation is given in Equations 15 and 16.

$$U_X(A) = \{x \in X|\mu_A(x) \geq \beta\} \text{---------} (15)$$

$$L_X(A) = \{x \in X|\mu_A(x) \geq \alpha\} \text{---------} (16)$$

where the thresholds are indicated by $\beta$ and $\alpha$.

The equivalence relation for the fuzzy autism feature is given in Equation 17.

$$R_i = \{((x, y), \mu_{R_i}(x, y))|x, y \in X_i\} \text{---------} (17)$$

where the equivalence of degree among $x$ and $y$ of the feature $i^{th}$ is indicated as $\mu_{R_i}$.

The rules are generated based on the approximations from FRST, and every rule defines the correlation between classification and feature combination. The rules are articulated

in *IF-THEN,* and their consequences are considered for classification. The form of the rule is given in Equation 18.

$$IF\ X_1\ is\ A_1\ AND\ X_2\ is\ A_2\ AND\ \dots\dots AND\ X_n\ is\ A_n\ THEN\ Classification \text{---------}(18)$$

where the fuzzy sets are represented by $A_i$, feature sets are represented by $X_i$, and classification is autism severity.

The sample fuzzy rules and classification generated from the *SL-FRST* are given as

Rule 1: *IF* Child enjoys being swung, bounced, etc. (Q1) is high *AND* Child takes an interest in other children (Q2) is low *AND* Child likes climbing on things (Q3) is high, *THEN* Classification is Severe Autism.

Rule 2: *IF* Child enjoys playing peek-a-boo/hide-and-seek (Q4) is low *AND* Child ever pretends (Q5) is high *THEN* Classification is Moderate Autism

Rule 3: *IF* Child uses index finger to point to ask for something (Q6) is high, *OR* Child uses index finger to indicate interest in something (Q7) is high, *THEN* Classification is Severe Autism

Rule 4: *IF* Child can play properly with small toys (Q8) is low, *AND* Child brings objects over to show (Q9) is high, *THEN* Classification is Moderate Autism

Rule 5: *IF* the Child looks you in the eye for more than a second or two (Q10) is high *AND* the Child seems oversensitive to noise (Q11) is low, *THEN* Classification is Severe Autism

The redundant and irrelevant rules are pruned to enhance efficiency and interpretability. The quality of the generated rules is identified using the performance metrics, where the rules are refined iteratively to acquire balance and accuracy. The procedure of *SL-FRST* is given in Algorithm 1.

| Algorithm 1. *SL-FRST* for autism classification |
|---|
| Data Preparation:<br>Organize the dataset with the given Child ID and questionnaire responses.<br>Ensure the dataset is clean, with no missing or erroneous values.<br>Feature Encoding:<br>Convert categorical responses into numerical values based on scores.<br>For example, scale scores from -1 to 1.<br>Feature Scaling:<br>Normalize feature values to bring them within a similar range.<br>Apply scaling: $X_{scaled} = \frac{2(X-X_{min})}{(X_{max}-X_{min})} - 1$<br>PCA Statistical Learning:<br>Apply statistical learning techniques such as logistic regression or decision trees to model relationships between principal components and target variables.<br>$Z = XV_k$ |

Fuzzy Rough Set Theory:
Handle uncertainty with fuzzy sets based on score ranges.
Define fuzzy membership functions for each feature.
Perform fuzzy discretization to convert continuous scores into fuzzy intervals.
Rule Generation:
Generate rules using fuzzy rough set theory to describe feature-class relationships.
Rule format: IF feature_1 IS high AND feature_2 IS low, THEN class IS severe autism.

## Result and Discussion

The dataset encompasses 550 responses to the *Modified Checklist for Autism in Toddlers (M-CHAT)* of a child nature. These responses reflect diverse aspects of the child's development and behaviour. The entire data provides needful insights into the traits of behaviour that are associated with *Autism Spectrum Disorder (ASD)* in infants and toddlers. The data has to be split into training, validation, and testing sets, which are very important in model generalization. This helps in preventing the model from being tested on the data used for training and hence gives a real-world evaluation. The training set, which is usually 60% of the entire data set, should be large enough to accommodate the variability of the data while at the same time not being too large so as to cause overfitting. The remaining 40 percent can be divided by two, meaning that the validation set will be 20 percent of the total dataset and the test set, also 20 percent of the total dataset. The data was synthetically created based on statistical characteristics obtained from the 550 records for subsequent analysis. This research is implemented in the latest version of R. The *Statistical Learning Assisted Fuzzy Rough Set Theory (SL-FRST)* is evaluated using the performance metrics: accuracy, precision, recall, F1-score, Specificity, and Sensitivity. The performance of the proposed SL-FRST is compared with existing approaches, namely *Short Time Frequency Transformation (STFT) [28], Hierarchical Fuzzy Autism Detection (HFAD) [29], and Fuzzy Multi-Criteria Decision Making (fMCDM) [30]*

*Principal Component Analysis (PCA)* is a robust data analysis tool for dimensionality reduction and pattern recognition. In the autism dataset, *PC1* captures the most significant variation, likely representing the most dominant behavioural patterns. A negative score on *PC1* suggests behaviours associated with severe autism traits, while a positive score indicates less severe characteristics. For instance, a child with a negative score (-2.7597079) exhibits behaviours aligned with severe autism, whereas a positive score (3.2677400) indicates less severe traits. The *PC2* captures orthogonal variation, possibly representing communication difficulties. For example, a child with a score of 0.6012064 on *PC2* exhibits moderate alignment with communication difficulties.

Interpretation becomes more nuanced with *PC3*, *PC4*, and *PC5*, which capture minor variances. The interpretation may be less straightforward due to the less explained variance. Understanding these components can aid in identifying behavioural patterns and clustering individuals based on their scores. Further analysis can involve examining the loadings of original variables on each principal component to discern which behaviours contribute most significantly to the observed patterns. Comparing scores

across diverse *PC*s can identify behavioural profiles while investigating loadings on original variables can reveal which behaviours contribute most to each component.

The accuracy shows the capability to appropriately classify the instances, which indicates the effectiveness of the autism classification. The accuracy of classification is shown in Table 1 and Figure 2. The accuracy of autism classification is done by Equation 19.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{---------(19)}$$

Table 1. Comparison of Accuracy

| Sample Count | STFT | HFAD | fMCDM | SL-FRST |
|---|---|---|---|---|
| 100 | 79.23 | 81.23 | 83.02 | 86 |
| 200 | 80.03 | 82.7 | 83.67 | 87.5 |
| 300 | 81.23 | 84.3 | 84.45 | 89.33 |
| 400 | 82.43 | 85.66 | 85.82 | 89.76 |
| 500 | 83.23 | 86.2 | 86.88 | 90.4 |

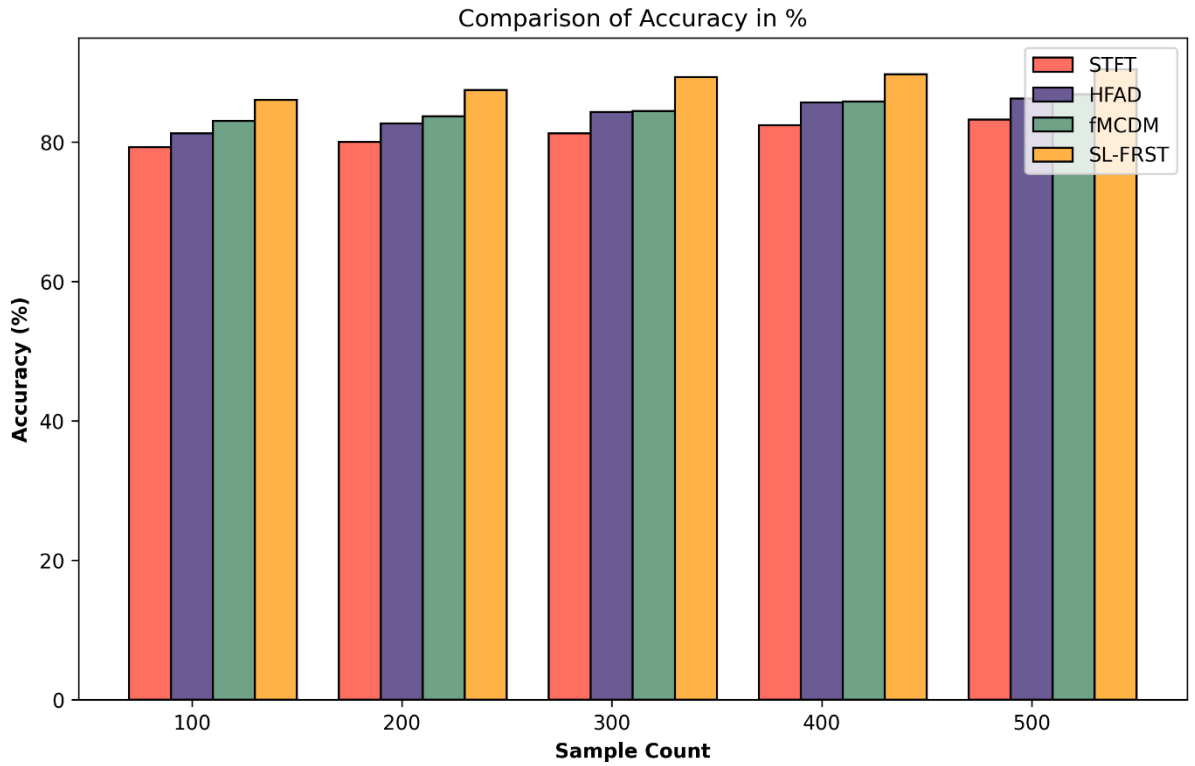


Figure 2. Comparison of Accuracy

The precision identifies the accuracy of the model in predicting the individuals with autism across others predicted to have autism. The precision of classification is given in Table 2 and Figure 3. The accuracy of autism classification is done by Equation 20.

$$Precision = \frac{TP}{TP+FP} \text{---------(20)}$$

Table 2. Comparison of Precision

| Sample Count | STFT | HFAD | fMCDM | SL-FRST |
|---|---|---|---|---|
| 100 | 80.02 | 82.56 | 83.89 | 87.63 |
| 200 | 80.9 | 83.12 | 84.23 | 88.29 |
| 300 | 82 | 85.09 | 85 | 88.74 |
| 400 | 83.67 | 85.99 | 86.11 | 89.91 |
| 500 | 84 | 86.8 | 87.09 | 90.36 |

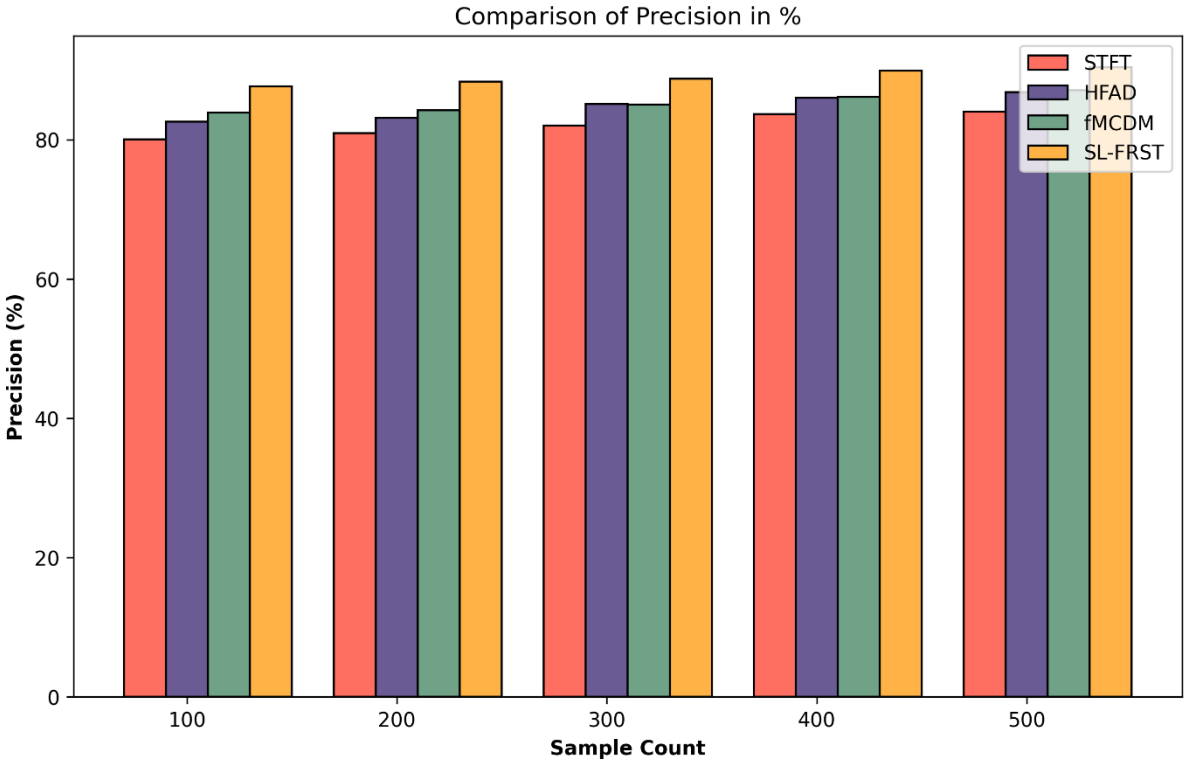


Figure 3. Comparison of Precision

The F1-score is a combination of recall and precision that gives a balanced evaluation distinctly efficient for imbalanced data classes. The F1-score of Classification is shown in Table 3 and Figure 4. The F1-score of autism classification is done by Equation 21.

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \text{---------(21)}$$

Table 3. Comparison of F1-Score

| Sample Count | STFT | HFAD | fMCDM | SL-FRST |
|---|---|---|---|---|
| 100 | 85.2 | 85.34 | 88.23 | 92.39 |
| 200 | 85.93 | 86.34 | 89.43 | 92.96 |
| 300 | 86.76 | 87.39 | 90.3 | 93.41 |
| 400 | 86.92 | 88.23 | 90.9 | 93.88 |
| 500 | 87.3 | 88.99 | 92.23 | 93.9 |

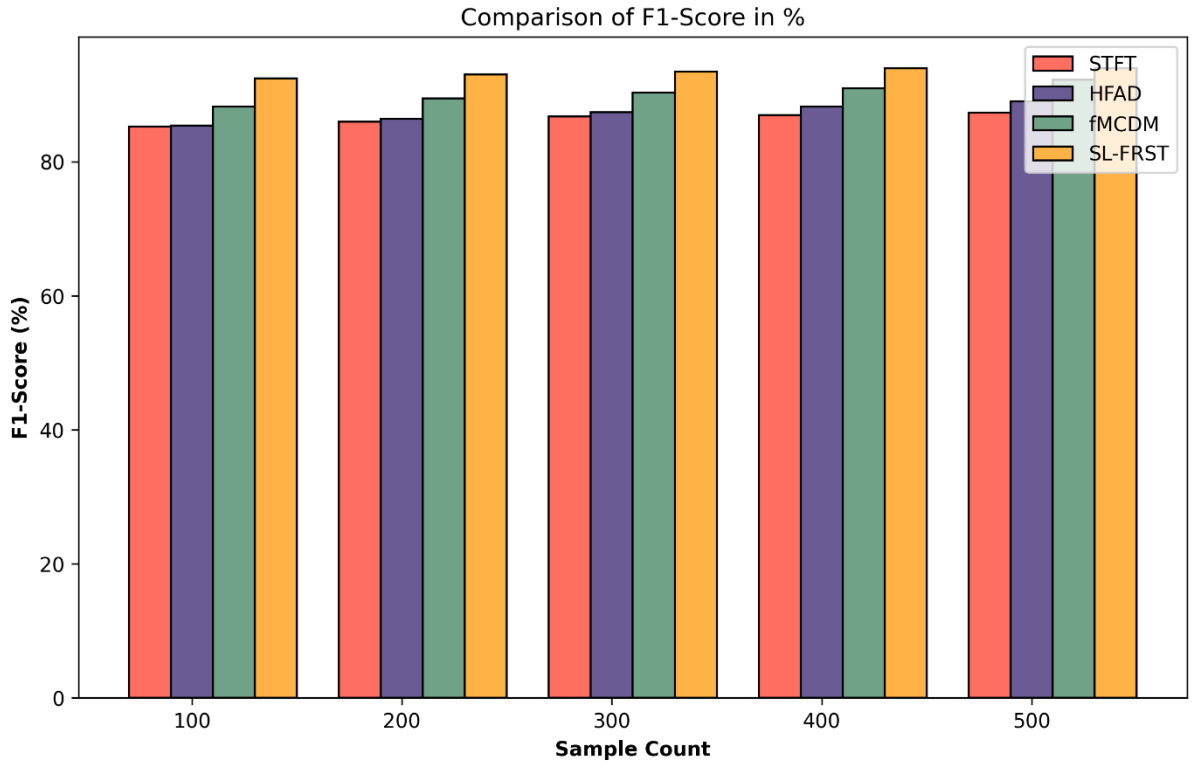


Figure 4. Comparison of F1-Score

The autism classification model can precisely predict the individuals from the autism data where the children have autism. The sensitivity of classification is given in Table 4 and Figure 5. The sensitivity of autism classification is done by Equation 22.

$$Sensitivity = \frac{TP}{TP+FN} \text{---------(22)}$$

Table 4. Comparison of Sensitivity

| Sample Count | STFT | HFAD | fMCDM | SL-FRST |
|---|---|---|---|---|
| 100 | 86.23 | 86.8 | 88.91 | 91.7 |
| 200 | 86.93 | 87.09 | 89.12 | 91.67 |
| 300 | 87.9 | 87.67 | 89.56 | 92.19 |
| 400 | 88 | 88.67 | 89.99 | 92.36 |
| 500 | 89.3 | 89.63 | 90.02 | 92.46 |

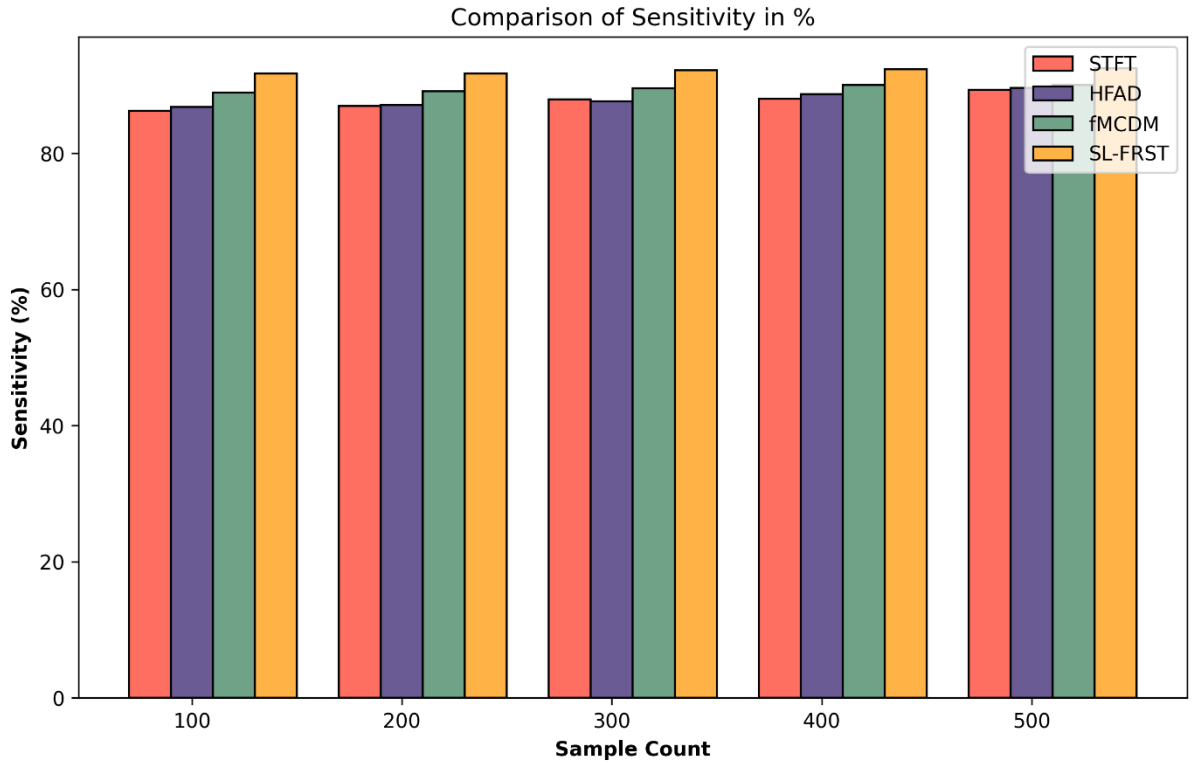


Figure 5. Comparison of Sensitivity

The ability to classify children without autism exactly across the autism dataset is essential. The specificity of classification is given in Table 5 and Figure 6. The specificity of autism classification is done by Equation 23.

$$Specificity = \frac{TN}{TN+FP} \text{---------(23)}$$

Table 5. Comparison of Specificity

| Sample Count | LR | HFAD | fMCDM | SL-FRST |
|---|---|---|---|---|
| 100 | 2.32 | 2.65 | 3.23 | 7.69 |
| 200 | 3.34 | 3.76 | 4.33 | 18.15 |
| 300 | 13.76 | 13.91 | 14.45 | 21.05 |
| 400 | 13.97 | 14.29 | 15.43 | 20 |
| 500 | 14.12 | 14.56 | 16.45 | 20 |

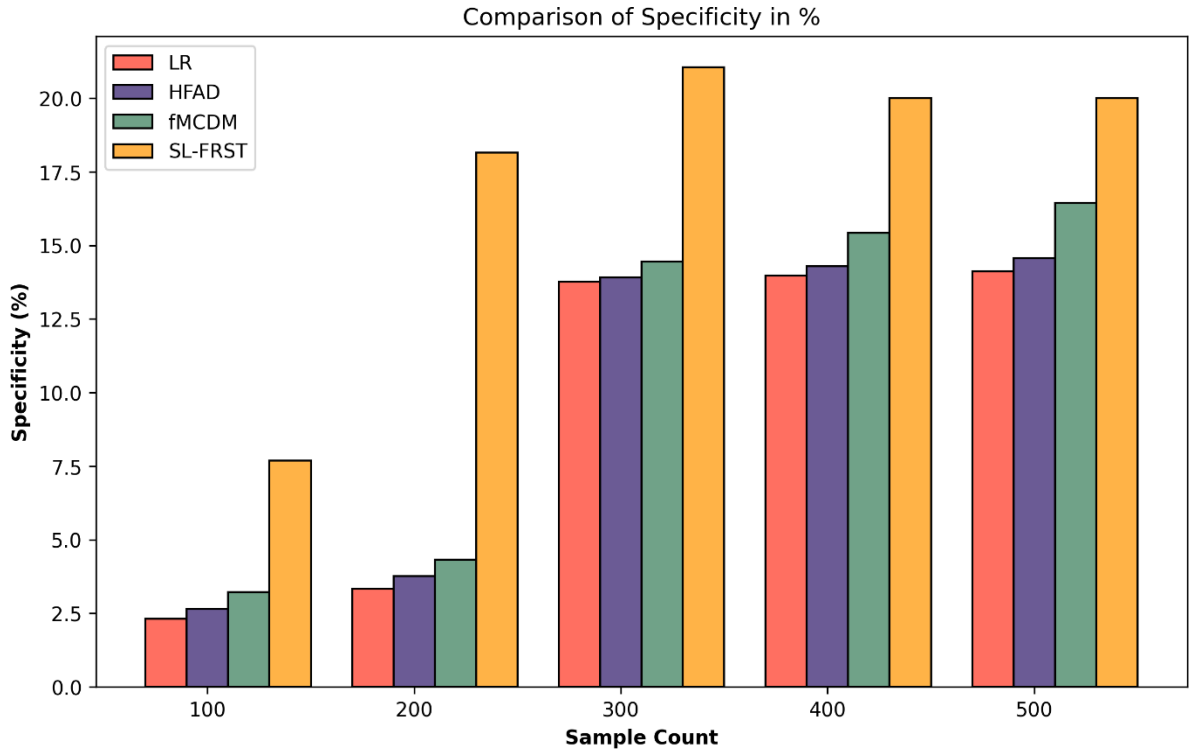


Figure 6. Comparison of Specificity

The comparative analysis of autism classification metrics provides insights into the effectiveness of various models in accurately identifying instances of autism. As depicted in Table 1 and Figure 2, accuracy indicates the models' ability to classify instances, indicating their overall effectiveness appropriately. Across different sample counts, *SL-FRST* consistently outperforms other models, with accuracy increasing from 86% at 100 samples to 90.4% at 500 samples. The precision, illustrated in Table 2 and Figure 3, observes the models' accuracy in predicting individuals with autism among all those predicted to have autism. *SL-FRST* demonstrates superior performance, with precision reaching 90.36% for 500 samples, indicating its reliability in correctly identifying autistic individuals.

The F1-score, illustrated in Table 3 and Figure 4, combines precision and recall, providing a balanced evaluation distinctly functional for imbalanced data classes. *SL-FRST* consistently achieves the highest F1 scores across different sample counts, indicating its effectiveness in autism classification. Sensitivity, illustrated in Table 4 and Figure 5, measures the model's ability to precisely identify individuals with autism across all those who have autism in the dataset. *SL-FRST* shows remarkable sensitivity in achieving 92.46% for 500 samples, highlighting its efficiency in capturing *TP* instances.

The specificity, as shown in Table 5 and Figure 6, evaluates the model's ability to classify individuals without autism accurately. *SL-FRST* maintains higher specificity than other models, indicating its ability to correctly identify individuals without autism. The proposed *SL-FRST* consistently outperforms other models across all evaluated metrics, indicating its effectiveness in autism classification across varying sample counts.

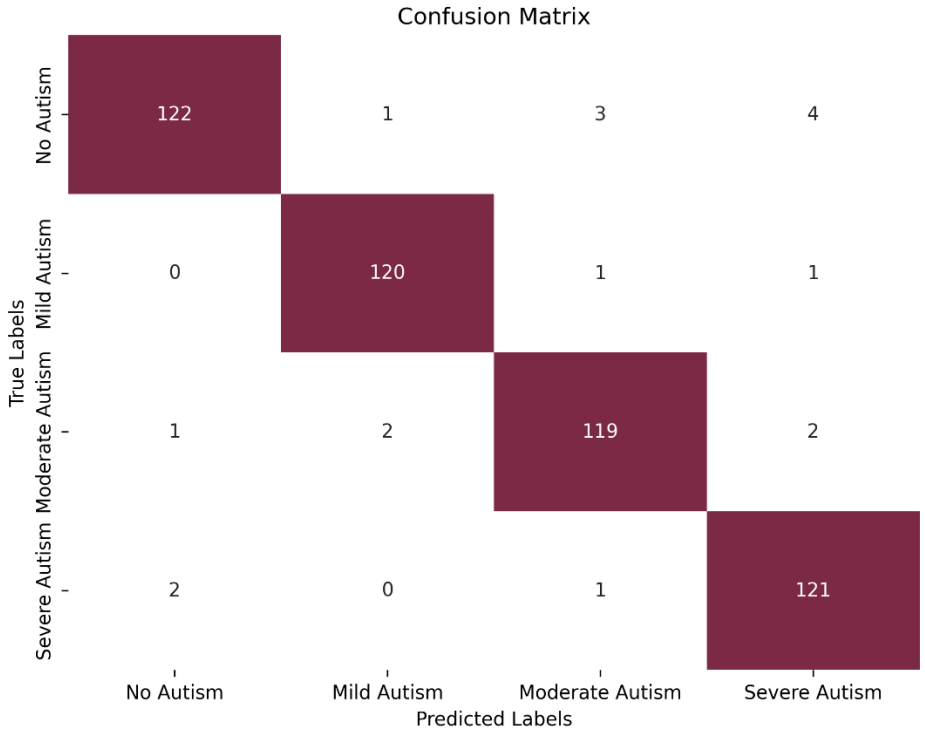


Figure 7. Comparison of Confusion Matrix

In the confusion matrix presented in Figure 7, it is possible to observe the results of the classification of autism according to the severity level. Assessment of the autism supports based on the limitation and possibility of the classification model. The confusion matrix is an important in estimating and improving the accuracy of the classification models for the different levels of autism.

## Conclusion

The use of both the PCA and the Fuzzy Rough Set Theory reduces a lot of drawbacks in the diagnosis of the Autism Spectrum Disorder. PCA helps in the dimensionality reduction of the data from the SRS, which we acquire in high dimensions, but the analysis in lower dimensions is more convenient while not losing critical information. Conversely, Fuzzy Rough Set Theory deals with the vagueness and uncertainty that characterise ASD data and offers a better way of interpreting the data. Combined, these techniques help in the construction of more accurate and reliable models for classification of ASD. Due to the improvements in the accuracy and interpretability of PCA and Fuzzy Rough Set Theory, clinical decision making and intervention in the management of ASD is enhanced hence providing better outcomes for the individuals with ASD and their families.

In the future, the performance of the presented works can be improved more by using clustering techniques. From another angle, the proposed research work can be used in other areas apart from the healthcare field. Furthermore, the authors have pointed out that the performance of the SL-FRST model can be enhanced if DL models are used rather than fuzzy sets. Besides, the tuning of hyperparameter of DL models can be incorporated to achieve improved classifier outcomes. Based on the future work, the above-mentioned models can be employed in real-time to help physicians diagnose ASD.

## Reference

1. Yang, X., Zhang, N., & Schrader, P. (2022). A study of brain networks for autism spectrum disorder classification using resting-state functional connectivity. *Machine Learning with Applications*, *8*, 100290.
2. Maye, M. P., Kiss, I. G., & Carter, A. S. (2022). Definitions and classification of autism spectrum disorders. In *Autism Spectrum Disorders* (pp. 3-26). Routledge.
3. Jiang, W., Liu, S., Zhang, H., Sun, X., Wang, S. H., Zhao, J., & Yan, J. (2022). CNNG: a convolutional neural networks with gated recurrent units for autism spectrum disorder classification. *Frontiers in Aging Neuroscience*, *14*, 948704.
4. Jacob, S. G., Sulaiman, M. M. B. A., & Bennet, B. (2022). Algorithmic approaches to classify autism spectrum disorders: a research perspective. *Procedia Computer Science*, *201*, 470-477.
5. Elshoky, B. R. G., Younis, E. M., Ali, A. A., & Ibrahim, O. A. S. (2022). Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images. *ETRI Journal*, *44*(4), 613-623.
6. Kim, J. I., Bang, S., Yang, J. J., Kwon, H., Jang, S., Roh, S., ... & Kim, B. N. (2022). Classification of preschoolers with low-functioning autism spectrum disorder using multimodal MRI data. *Journal of Autism and Developmental Disorders*, 1-13.
7. Gaspar, A., Oliva, D., Hinojosa, S., Aranguren, I., & Zaldivar, D. (2022). An optimized Kernel Extreme Learning Machine for the classification of the autism spectrum disorder by using gaze tracking images. *Applied Soft Computing*, *120*, 108654.
8. Mohanta, A., & Mittal, V. K. (2022). Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features. *Computer Speech & Language*, *72*, 101287.
9. Du, Y., He, X., Kochunov, P., Pearlson, G., Hong, L. E., van Erp, T. G., ... & Calhoun, V. D. (2022). A new multimodality fusion classification approach to explore the uniqueness of schizophrenia and autism spectrum disorder. *Human brain mapping*, *43*(12), 3887-3903.
10. Mishra, M., & Pati, U. C. (2023). A classification framework for Autism Spectrum Disorder detection using sMRI: Optimizer based ensemble of deep convolution neural network with on-the-fly data augmentation. *Biomedical Signal Processing and Control*, *84*, 104686.
11. Kaur, P., & Kaur, A. (2023). Review of progress in diagnostic studies of autism spectrum disorder using neuroimaging. *Interdisciplinary Sciences: Computational Life Sciences*, *15*(1), 111-130.
12. Khudhur, D. D., & Khudhur, S. D. (2023). The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups. *Measurement: Sensors*, *27*, 100774.
13. Zhu, H., Wang, J., Zhao, Y. P., Lu, M., & Shi, J. (2022). Contrastive multi-view composite graph convolutional networks based on contribution learning for autism spectrum disorder classification. *IEEE Transactions on Biomedical Engineering*.
14. Koehler, J. C., Dong, M. S., Bierlich, A. M., Fischer, S., Späth, J., Plank, I. S., ... & Falter-Wagner, C. M. (2024). Machine learning classification of autism spectrum disorder

based on reciprocity in naturalistic social interactions. *Translational Psychiatry*, *14*(1), 76.

15. Wang, Z., Xu, Y., Peng, D., Gao, J., & Lu, F. (2023). Brain functional activity-based classification of autism spectrum disorder using an attention-based graph neural network combined with gene expression. *Cerebral Cortex*, *33*(10), 6407-6419.
16. Sadiq, A., Al-Hiyali, M. I., Yahya, N., Tang, T. B., & Khan, D. M. (2022). Non-oscillatory connectivity approach for classification of autism spectrum disorder subtypes using resting-state fMRI. *IEEE Access*, *10*, 14049-14061.
17. Hasan, S. M., Uddin, M. P., Al Mamun, M., Sharif, M. I., Ulhaq, A., & Krishnamoorthy, G. (2022). A machine learning framework for early-stage detection of autism spectrum disorders. *IEEE Access*, *11*, 15038-15057.
18. Kareem, A. K., AL-Ani, M. M., & Nafea, A. A. (2023). Detection of Autism Spectrum Disorder Using A 1-Dimensional Convolutional Neural Network. *Baghdad Science Journal*, *20*(3 (Suppl.)), 1182-1182.
19. Lu, P., Li, X., Hu, L., & Lu, L. (2022). Integrating genomic and resting State fMRI for efficient autism spectrum disorder classification. *Multimedia Tools and Applications*, 1-12.
20. Washington, P., Paskov, K.M., Kalantarian, H., Stockham, N., Voss, C., Kline, A., Patnaik, R., Chrisman, B., Varma, M., Tariq, Q. and Dunlap, K., 2020, January. Feature selection and dimension reduction of social autism data. In Pac SympBiocomput (Vol. 25, pp. 707-718).
21. Tang, L., Mostafa, S., Liao, B. and Wu, F.X., 2019. A network clustering based feature selection strategy for classifying autism spectrum disorder. BMC Medical Genomics, 12(7), p.153.
22. DevikaVarshini, G. and Chinnaiyan, R., Optimized Machine Learning Classification Approaches for Prediction of Autism Spectrum Disorder. Ann Autism Dev Disord. 2020; 1 (1), 1001.
23. Abdolzadegan, D., Moattar, M.H. and Ghoshuni, M., 2020. A robust method for early diagnosis of autism spectrum disorder from EEG signals based on feature selection and DBSCAN method. Biocybernetics and Biomedical Engineering, 40(1), pp.482-493.
24. Xu, L., Liu, Y., Yu, J., Li, X., Yu, X., Cheng, H. and Li, J., 2020. Characterizing autism spectrum disorder by deep learning spontaneous brain activity from functional nearinfrared spectroscopy. Journal of Neuroscience Methods, 331, p.108538.
25. Kosmicki, J.A., Sochat, V., Duda, M. and Wall, D.P., 2015. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Translational psychiatry, 5(2), pp.e514-e514.
26. Wall, D.P., Dally, R., Luyster, R., Jung, J.Y. and DeLuca, T.F., 2012. Use of artificial intelligence to shorten the behavioral diagnosis of autism. PloS one, 7(8), p.e43855.
27. Pratap, A. and Kanimozhiselvi, C.S., 2014. Predictive assessment of autism using unsupervised machine learning models. International Journal of Advanced Intelligence Paradigms, 6(2), pp.113-121.
28. Shams, W. K., Wahab, A., & Qidwai, U. A. (2012). Fuzzy model for detection and estimation of the degree of autism spectrum disorder. In *Neural Information*

*Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part IV 19* (pp. 372-379). Springer Berlin Heidelberg.

29. Sharma, A., Khosla, A., Khosla, M., & Rao, Y. (2018). Fast and accurate diagnosis of autism (FADA): a novel hierarchical fuzzy system-based autism detection tool. *Australasian physical & engineering sciences in medicine*, *41*, 757-772.
30. Joudar, S. S., Albahri, A. S., & Hamid, R. A. (2023). Intelligent triage method for early diagnosis autism spectrum disorder (ASD) based on integrated fuzzy multi-criteria decision-making methods. *Informatics in Medicine Unlocked*, *36*, 101131.