International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I4P131

Volume 72 Issue 4, 306-314, April 2024

Original Article

Identifying Optimal Feature Set for Improved Autism Classification Using Machine Learning Techniques

Karpagam C[1], Deepa C[2]

[1]*Department of Computer Science, Sri Ramakrishna College of Arts and Science, Tamil Nadu, India.*
[1]*Department of Comp. Sci. with Data Analytics, Dr. N.G.P. Arts and Science College, Tamil Nadu, India.*
[2]*Department of Computer Science (AI & DS), Sri Ramakrishna College of Arts and Science, Tamil Nadu, India.*

[1]*Corresponding Author : karpagam@drngpasc.ac.in*

Received: 15 December 2023 Revised: 31 March 2024 Accepted: 09 April 2024 Published: 24 April 2024

Abstract *- Administering standard medical prognosis tools for autism disorder is a time-consuming process. Furthermore, only a trained and experienced professional can supervise the assessment. The attempts that failed to evaluate ASD (Autism Spectrum Disorder) at the right time lead to critical medical care costs and a high impact on an individual's performance in regular activities. More flexible and evident accessible methods would assist parents and caretakers in mitigating the hurdles faced during conventional clinical diagnosis. The previous work represents the exploratory data analysis made on the autism dataset. Here, an expansion to build a model by combining Machine Learning classification algorithms on selected feature sets for improved accuracy is administered. The autism dataset used in the experiment is collected from a public repository that includes 1054 instances. RFE (Recursive Feature Elimination) and the Boruta method are preferred to determine the relevant feature set with the highest rank. A significant improvement in the accuracy of results is noted when a Random Forest (RF) is ensembled with a Support Vector Machine (SVM) with 98.97% accuracy in the toddler dataset. The resultant model maximizes accuracy and minimizes the efforts taken by practitioners with the extensive diagnosis process.*

Keywords *- Autism Detection, Boruta, Random Forest, Recursive Feature Elimination, Support Vector Machine.*

1. Introduction

Autism Spectrum Disorder is a neurodevelopmental disability commonly identified in children in the age group of 4-5 years. This complex disorder in the brain, results in impairment of functioning and affects the cognitive abilities of individuals. A vivid analysis of this spectrum disorder is presented below:

1.1. Background Study

Autism is a developmental disability that causes deficits in communication, attention, social skills, and behavior. ASD is a spectrum disorder where the severity of the syndrome ranges from mild to moderate and severe. The impact of developmental conditions can manifest differently from one person to another. [1,2]

1.2. Causes

The vital cause of the syndrome appears to have different reasons given by researchers; a few among them are genetic influences, environmental causes, malnutrition during pregnancy, childhood vaccines and older age of parents. [3]

1.3. Notable facts about ASD

1 in 59 children get affected with autism. Siblings are at significant risk of developing autism condition, which is considered to be hereditary. Autism is likely to affect more boys than girls over a gender ratio of 4:1, and now it is closer to 3:1. [4-6]

1.4. Diagnosis Methods

The right age to perform a reliable diagnosis for the disorder is 18 - 24 months of age. Below this age limit might result in an erroneous prediction that leads to a false conclusion. The current autism diagnosis practices are evaluated only by trained professionals who last for hours of assessment. The assessment tools evaluate the constructive results by administering various rating scales, which is expected to be a long-awaited moment. [7,8]

Researchers design numerous approaches to detect autism at a young age. Different modalities of clinical data are obtained, and complex algorithms are implemented. It includes neuroimages, genetics, video analysis and facial expressions of ASD patients. Here, the proposed model attempts to identify the traits at an early stage using an optimized feature selection process followed by an ensemble technique in Machine Learning. The proposed model uses machine learning techniques such as Random Forest and

Support Vector Machines to enhance the accuracy of autism assessment to overcome the growing time constraint challenges in conventional diagnosis approaches. The toddler dataset used in this study primarily focuses on children in the age group between 12 – 36 months.

The main objective of this paper is to determine the disorder at an early stage that helps in a fast curative method. When the individual reaches the age of 8 and above, it becomes a tedious task for the therapists to change the behavior patterns of children. Children with ASD have a high risk of developing hypertension, complications of obesity and a rise in depression as they grow. The best choice of recommendation for this problem is to identify the autistic traits of an individual at the early stage of development. In this study, an extension to improvise the assessment evaluation by extracting prominent features of the dataset that have a greater impact in diagnosing ASD is experimented with. [9,10]

The arrangement of the entire content of this paper is as follows. Section 1 states the pilot study of autism disorder, its causes, diagnosis methods and notable facts about ASD. Section 2 elaborates on the related work carried out in research in recent years. Section 3 gives a clear view of the datasets used as input. Section 4 elucidates the feature selection techniques applied in the study. Section 5 compares various classification algorithms in measuring the outcome of results. Section 6 and 7 discuss with results and conclusion and further enhancement of this study respectively.

2. Related Work

Researchers can overcome difficulties with high-dimensional data sets by using feature selection techniques. The primary job is to provide more projection on relevant attributes, surpassing the irrelevant and redundant data values. Due to this factor, the computation time of the model gets reduced, and prediction accuracy gets improved. Numerous studies have shown that feature extraction has sped up the machine learning models' training times. A few prominent findings of feature selection, machine learning and deep learning techniques are below. [11-13]

Vaishali et al. provided a powerful swarm intelligence model for the categorization of autism that chooses features using a firefly wrapper algorithm. Selecting 10 out of 21 attributes in prediction results in better classification accuracy was proven. The dataset obtained from the public repository exhibits an average accuracy range between 92% and 97%. The author concludes that selected feature sets have the opportunity to retain the structure of the complete dataset. [14] Girish et al. demonstrate an outline review of filter, wrapper and embedded methods in feature selection. Collectively, all these methods help in the variable elimination of a high-dimensionality dataset. The author clearly distinguishes the purpose and importance of feature selection methods such as correlation criteria, mutual information, sequential selection, heuristic search and others. The stability of these feature selection techniques was proven by applying them to a standard dataset. [15]

Rahman et. al. reviews different approaches in balancing the dataset, feature selection and machine learning classification. The author categorizes the feature selection methods as flat features (filter methods), streaming features (for dynamically streaming applications) and structured features (tree and graph structure). The author imparts a unified approach to indicate different techniques in feature selection and the significance of choosing the right one. [16]

Wiratsin et al. focused on the Apriori algorithm to determine the strong relationship between the data values. Furthermore, the chi-square test and Mutual information are applied to determine the dominant attributes of categorical features in a dataset. The evaluation of results was proven using SVM by considering the top 3 dominant features of the dataset with 78 and 83% accuracy. [17]

Washington et al. used a filter, wrapper and embedded methods to perform feature selection on the Social Responsiveness Scale (SRS) questionnaire to classify ASD and ADHD traits. Further, methods for dimensionality reduction such as PCA (Principal Component Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding) and denoising autoencoder are employed. The classification performance is measured using Multilayer Perceptron (MLP) with 92% accuracy. [18]

Alzubi et al. implemented a hybrid method based on the filter and wrapper technique, a conditional Mutual Information Maximization (CMIM) method and the SVM Recursive Feature Elimination (SVM-RFE).[19] The delivered model is compared with four classification algorithms: Naive Bayes (NB), Linear Discriminant Analysis (LDA), k Nearest Neighbors (kNN) and Support Vector Machine (SVM), which resulted in 89% classification accuracy [6].

Guruvammal et al. introduce a new technique, Levi Flight Cub Update-based lion algorithm (LFCU-LA), for optimal selection of features. The proposed method is a modification of the lion algorithm. The classification experiment is also a hybrid of a Deep Belief Network and a Neural Network. Finally, the author states that the proposed model exhibits a higher performance rate when compared with conventional methods. [20]

Hossain et al. encompass Information Gain, Correlation, One R, Chi-squared and Relief F, a total of 5 feature selection techniques and implement using classification algorithms for evaluation. Among the above listed, Relief F outperforms all other selection techniques and yields 100% accuracy while

assessing using Multilayer Perceptron (MLP). [21] The above literary works present the merits of the unique feature selection process employed in recent advancements in research.

3. Dataset Description

The datasets used in this experiment are from the public repository Kaggle.com. [22] A mobile application obtains the accumulated values that assess autistic traits in toddlers. Figure 1 presents an introductory exploration by studying and analyzing the data values thoroughly. Data is compiled depending on the age groups of individuals, including toddlers, children, adolescents, and adults. It evaluates an individual's behavioural traits to determine whether or not they exhibit autistic characteristics. The purpose of a set of 10 autism questionnaires (AQ) is to measure and validate the behavioural traits of individuals. The questionnaire is commonly known as (The Quantitative Checklist for Autism in Toddlers) Q-CHAT-10, an autistic assessment tool. The principal objective of Q-CHAT-10 is to concentrate on the early detection of ASD in children between 12 and 36 months of age. The 10-attributes of toddler dataset assesses the behavioural characteristics of a child as illustrated in Table 1.

The main focus of Q-CHAT-10 is to assess the symptoms of autistic conditions, specifically speech, attention, and social interaction. The entire list consists of questions with two alternative answers that are dichotomous. Yes/No. Without clinical support from practitioners/experts, the assessor can effortlessly answer the questionnaire in Q-CHAT-10. [23,24]

Other characteristics such as gender, ethnic background, age, jaundice at the time of birth and relatives with autistic symptoms are thought to be crucial data offered by a person for further investigation. The remaining characteristics are to spot autistic children's behavioural characteristics. There are 1054 instances and 19 feature sets in the toddler dataset.

3.1. Data Imbalance

Data preprocessing and cleaning are performed to minimize noise and missing values in a data set. The data's essence has been improved and made available for additional analysis to produce an optimum result. Similarly, data imbalance impacts the performance of the predicted model, especially in high-dimensional datasets. [26,27]

The toddler dataset used in this experiment consists of 728 autistic and 326 non-autistic instances. When the class proportion is not balanced correctly, machine learning algorithms will provide erroneous results. Predictions will be less accurate and lower as a result of this variation. The primary issue of class imbalance in datasets is largely addressed by SMOTE (Synthetic Minority Oversampling Technique). It generates synthetic samples by drawing random samples from minority classes.

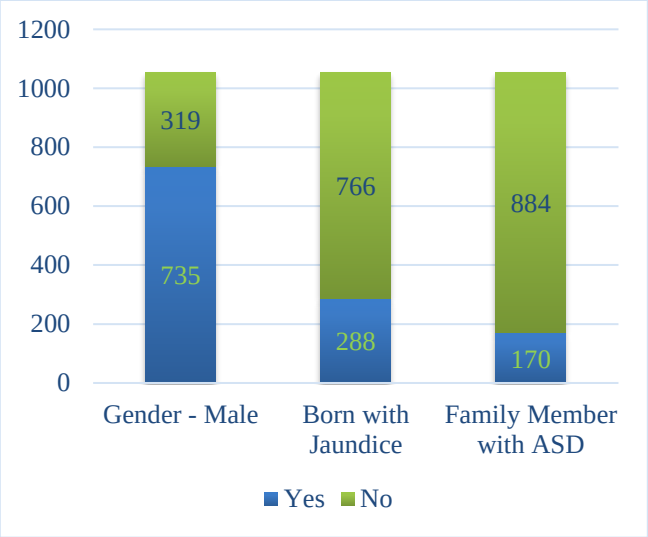

Fig. 1 Dataset Description

This technique is considered to be one pioneering method in preprocessing a dataset. Figure 2 presents the effect of SMOTE in the toddler dataset, where the autism class variable is with an equal count of instances. [28-30]

4. Feature Selection

By reducing the complexity of the model, feature selection techniques help the machine learning algorithm train more quickly [11]. Raw Data includes many irrelevant and redundant attributes. Feature selection is an appropriate method for removing unnecessary or duplicated data. A classification model's performance will be less accurate and inconsistent in case of improper feature selection. [31]

Careful selection of feature sets results in the improvisation of classification accuracy and processing speed of the algorithm. To increase the classification performance, the model has to concentrate on a subset of attributes from the provided dataset. A diverse collection of feature selection approaches is available to evaluate the relevance of each feature with the class/target variable. It extracts the relevant column set to determine the optimal solution [32]. In this study, the implementation of two prime approaches is depicted. They are Recursive Feature Elimination (RFE) and Boruta method.

4.1. Recursive Feature Elimination

The Recursive Feature Elimination (RFE) method is a powerful and simple technique to implement and use. RFE is equipped for distinguishing the highlights in a dataset that are more relevant in anticipating the target variable. RFE is a feature selection algorithm of the wrapper variety. It operates by starting with all feature sets in the training set and determines a subset of reduced features. The feature selection comprises two design decisions: the count of features to obtain and an algorithm model to aid in feature selection. [33, 34]

Table 1. QCHAT-10 Toddler Questionnaire[25]

S.No.	Attribute Description	Type of Response
1.	Does your child look at you when you call his/her name?	Yes/No
2.	How easy is it for you to make eye contact with your child?	Yes/No
3.	Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)	Yes/No
4.	Does your child point to share interests with you?	Yes/No
5.	Does your child pretend? (e.g. care for dolls, talk on a toy phone)	Yes/No
6.	Does your child follow where you are relooking?	Yes/No
7.	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g. stroking hair, hugging them)	Yes/No
8.	Would you describe your child's first words as:	Yes/No
9.	Does your child use simple gestures? (e.g. wave goodbye)	Yes/No
10.	Does your child stare at nothing with no apparent purpose?	Yes/No

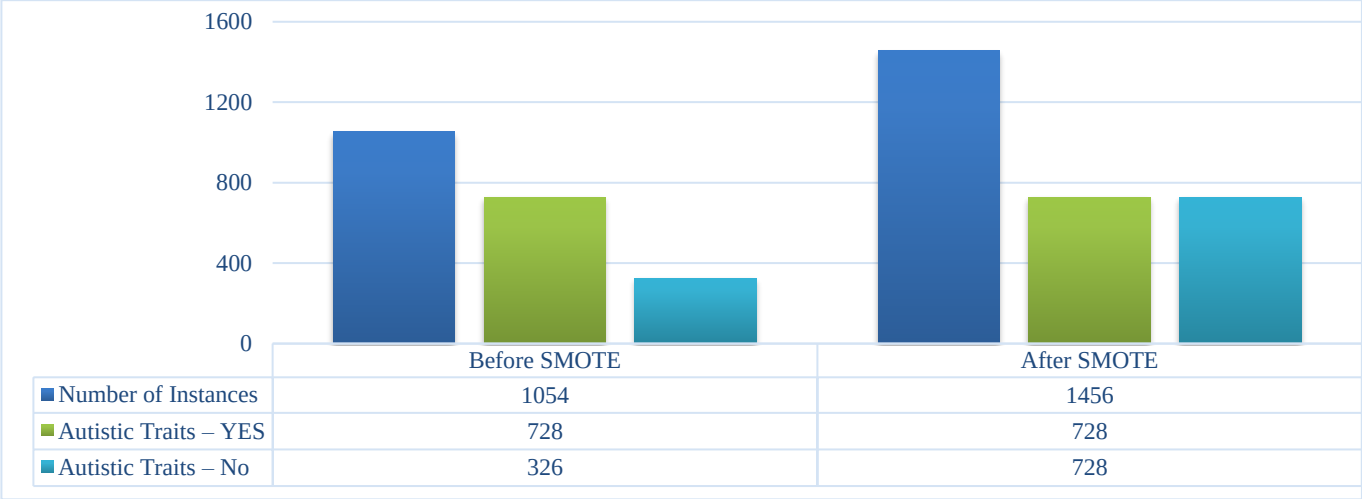

Fig. 2 Data Imbalance - SMOTE

A Machine Learning algorithm or a statistical method can be employed to score the features. Apply a filter to assign a ranking to each attribute and select the one with the highest ranking. To achieve this, the machine learning algorithms in the core model are first tuned, then the feature sets are ranked according to their relevance, the least influential features are removed, and the model is tuned again.

A predetermined set of attributes is left after repeating this method [34]. The following pseudocode describes the entire RFE process in a sequence of steps. The two hyperparameters used in this experiment are the statistical approach to score the features and a total count of six features to be determined.

Pseudocode: RFE Feature Elimination Process
Input: Set of attributes in the dataset X= $\{x_1,x_2…..x_n\}$

Parameters:
no_of_features: count of attributes
estimator: Support Vector Machine
cross_validation: 10-fold
Output: List of attributes ranked R = $\{r_1,r_2,….r_n\}$

Methodology:
Step 1: Read the dataset, X
Step 2: Perform preprocessing
Step 3: Set the hyper-parameters values as input P=$\{p_1,p_2…p_n\}$
Step 4: Apply 10-fold cross-validation to partition the dataset randomly
Step 5: Train and fit each subset of the model using SVM
Step 6: Determine the weight factor by computing the measure of variable importance
Step 7: Repeat steps 5 and 6 until all the attributes are ranked.

4.2. Boruta Method

A reliable feature selection algorithm for low-dimensional datasets is Boruta. It chooses the most favourable subset of features by implementing a random forest model using metrics. Boruta processes the dataset by generating randomized duplicates of each feature in the input data set to introduce randomization. A random forest classifier is then trained using the enlarged data set, and the relevance of each feature is evaluated using a feature importance measure. Higher values denote higher importance.

The process is continued further by eliminating the characteristics that are regarded to be of extremely low relevance. The method terminates either when all features are accepted or rejected or when the number of random forest runs reaches a predetermined threshold. [35-37]

Pseudocode: Boruta Feature Selection Process
Input: Set of attributes in the dataset X= {x1,x2…..xn}

Parameters:
no_of_features: count of attributes
classifier: Random Forest
Output: List of attributes ranked R = {r1,r2,….rn}

Methodology:
Step 1: Read the dataset, X
Step 2: Create a copy of the dataset values.
Step 3: Shuffle the column values in random to create a shadow of the dataset
Step 4: Merge the shuffled column values with the original data set.
Step 5: Compute the statistical value z-score by building an RF classifier
Step 6: Compare the resultant value with the original attribute to maximize the importance.
Step 7: If the attribute is significant, then retain it or else drop it.
Step 8: Repeat steps 5 - 8 until all the attributes are ranked.

The selected feature set from the toddler set is common in both RFE and Boruta methods. The lists of features are A1, A4, A5, A6, A7 and A9. The following are considered to be prime features that measure the autistic traits of an individual, which is the same using RFE and Boruta.

1) A1 & A6: measures the attention skill and level of concentration.
2) A4 & A5: measures the communication traits to give and receive information from one another.
3) A7 & A9: measures the social skills that facilitate interaction both verbally and non-verbally.

5. Classification Techniques

To identify an optimal model that produces the maximum accuracy, a few benchmark classification strategies are used in this paper. No one strategy applies with equal efficacy to all classification problems or all datasets. By considering this as input, the identified classification methods Random Forest, Support Vector Machine, and the proposed hybrid of RF and SVM are chosen for implementation.

5.1. Random Forest

The supervised learning approach includes the popular machine learning algorithm, Random Forest. [38] The ability of the random forest algorithm to process datasets with continuous variables and categorical variables used in regression and classification is one of its key features. For issues involving classification, it yields better outcomes.

Random Forest classification implementation is carried out by applying the following steps:

1) N random records are taken in random order from a data set of k records.
2) Every sample receives a different decision tree, and each decision tree's output is generated. [39]
3) Classification is performed using the outcome and assessed using the majority vote.

5.2. Support Vector Machine

SVM is a popular method in Machine Learning technique that can improve the expected result. SVM's remarkable generalization capability and ideal solution have recently increased its appeal to the data mining, pattern recognition, and machine learning sectors. SVM has proven to be an effective method for handling real-world challenges with binary classification. [40]

Support Vector Machine supports classification and regression but best results in classification problems. SVM applies the statistical approach to determine the hyperplane that separates the two classes perfectly. The linear function of the support vector machine adopts Equation 1, and the non-linear function adopts Equation 2. The main goal is to create a decision boundary that separates the given autism dataset into two class values. As the given dataset is binomial classification, linear SVM is adopted to create the hyper-plane. [41,42]

$$y = wx + \gamma \quad (1)$$

$$y = w\phi(x) + \gamma \quad (2)$$

5.3. Proposed Model

The proposed model employs an integrated approach to address the classification problem and parallelly combines Random Forest and SVM capabilities. Figure 3 depicts the architecture of the proposed framework. Random forest is a machine learning technique that builds several decision trees using randomly chosen attributes and predicts the class of a test instance through voting. SVM determines a hyperplane that divides two classes by applying a linear function.

Random Forest can be highly accurate and interpretable for the dataset with high dimensions. Similarly, the Support Vector Machine proves to be a better classification algorithm in many applications [43]. These two models are ensembled together to improve the prediction accuracy. The resultant prediction values of RF and SVM are created and appended to the validation samples. The new dataset values created are trained and validated to obtain maximum accuracy.

Pseudocode: Proposed RF-SVM
Input: Set of attributes in the dataset X= {x1,x2…..xn}
Output: Accuracy of the model

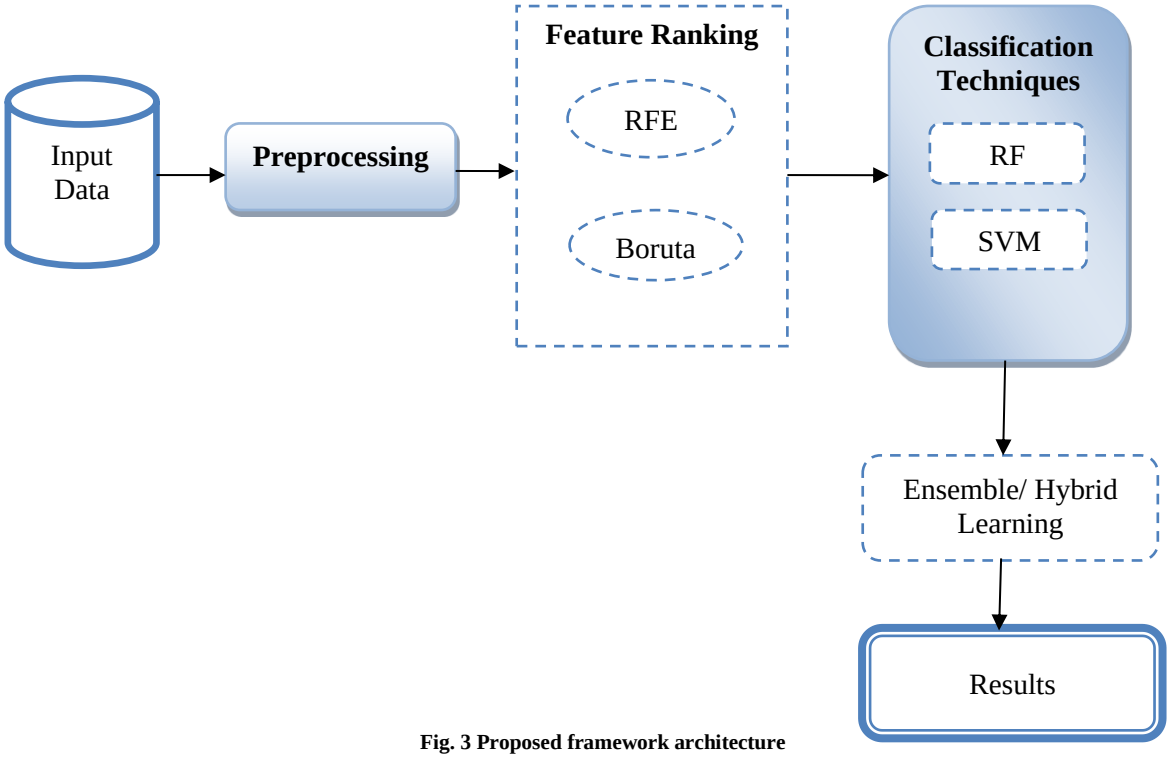

Fig. 3 Proposed framework architecture

Methodology:

Step 1: Consider a dataset $X = \{x_1, x_2 \ldots . x_n\}$ with n instances and the target variable $Y = \{y_1, y_2 \ldots . y_n\}$

Step 2: Compute values on the training set X_t

Step 3: Select K features randomly from the given X_t

Step 4: Compute the best split among the K

Step 5: Split the node into sub-nodes

Step 6: Repeat steps 2 to 4 till N number of trees generated

Step 7: Apply trained classifiers to each testing sample to determine the majority of votes X_p.

Step 8: Concatenate the predicted results to the dataset Equation 3

$$f(x) = \sum_{i=1}^{n} (X + X_p) \quad (3)$$

Step 9: Load the generated training dataset as X and class labels as Y

Step 10: Assign weight vector and offset value

Step 11: Construct Gaussian RBF kernel to optimize values

Step 12: Generate a hyperplane

Step 13: Based on the highest value of the above function, the prediction accuracy is determined.

6. Results and Discussion

This paper aims to present a hybrid classification model to detect autism disorders in toddlers. In the first phase, a comprehensive description of data is provided with the collection of 1054 instances. Using QCHAT-10, the three key assessment criteria are communication, attentiveness, and social interaction. To solve the overfitting issue, the dataset was balanced using the SMOTE technique in the second phase. The third phase includes feature selection methods such as RFE and Boruta. The last phase includes the implementation of Random Forest, SVM, a hybrid approach of RF-SVM methods.

Equations 4, 5 and 6 are the performance evaluation metrics considered for the assessment of the proposed model.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

Here, the hybrid model results in a higher accuracy of 98.97%. Table 2 shows the classification accuracy, precision and recall of all three algorithms. The notable work of the researchers [44] specifies an accuracy of 97.95 using CNN. Comparatively, this approach yields a better accuracy of 98.97%. Figure 4 depicts the same using visual representation.

Figure 5 presents the training and validation accuracy of the proposed model. After 15 epochs, the performance of validation accuracy rises from 76% to 98.9%. The Y-axis depicts the accuracy measure, and the x-axis represents the number of epochs. During the training phase, the accuracy ranges from 65% to 92%.

Table 2. Performance measure

Algorithms	Accuracy	Precision	Recall
Random Forest	94.86	93.12	92.66
SVM	95.19	97.56	96.35
Hybrid RFSVM	98.97	97.89	96.15

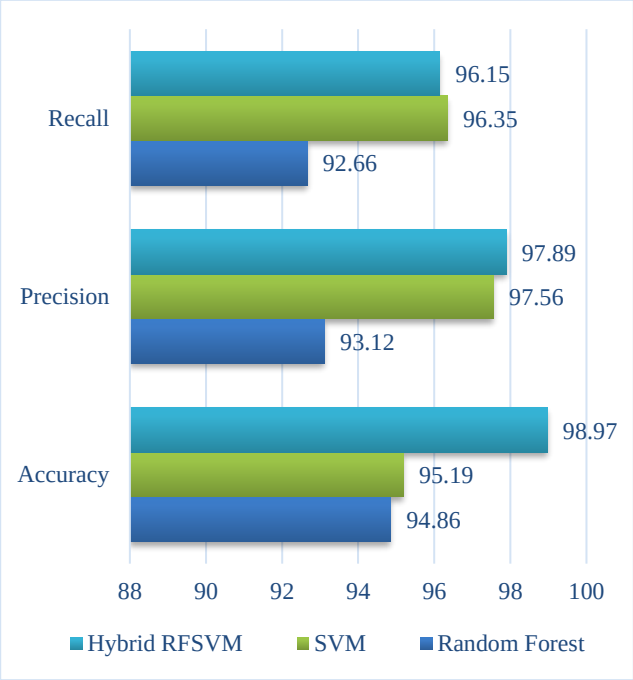

Fig. 4 Performance evaluation metrics

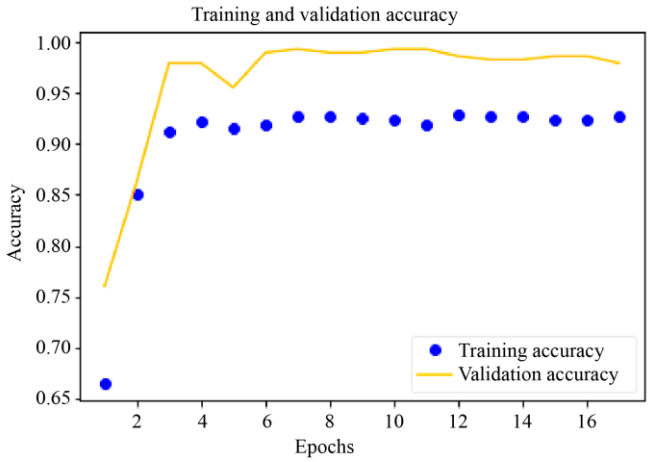

Fig. 5 Training and validation accuracy of proposed model

7. Conclusion

The physical, cognitive, and behavioural development of children with developmental disorders is delayed, which has an adverse effect on daily functioning. The major factors that affect lifelong development have an impact on individual urges to identify the problem and find a cure. In this study, a machine learning technique is used to predict autistic disorder. As a continuation of the earlier pre-processing module, now an optimized feature selection and a hybrid classification approach is demonstrated effectively. In future, a real-time dataset can be collected at a particular geographical region and report the disorder's impact on society.

References

[1] Catherine Lord et al., "Autism Spectrum Disorder," *The Lancet*, vol. 392, no. 10146, pp. 508-520, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Haylie L. Miller, and Nicoleta L. Bugnariu, "Level of Immersion in Virtual Environments Impacts the Ability to Assess and Teach Social Skills in Autism Spectrum Disorder," *Cyberpsychology, Behavior, and Social Networking*, vol. 19, no. 4, pp. 246-256, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[3] Joseph K. Gona et al., "Parents' and Professionals' Perceptions on Causes and Treatment Options for Autism Spectrum Disorders (ASD) in a Multicultural Context on the Kenyan Coast," *Plos One*, vol. 10, no. 8, pp. 1-13, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] Lauren Rylaarsdam, and Alicia Guemez-Gamboa, "Genetic Causes and Modifiers of Autism Spectrum Disorder," *Frontiers in Cellular Neuroscience*, vol. 13, pp. 1-15, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Tony Charman et al., "Non-ASD Outcomes at 36 Months in Siblings at Familial Risk for Autism Spectrum Disorder (ASD): A Baby Siblings Research Consortium (BSRC) Study," *Autism Research*, vol. 10, no. 1, pp. 169-178, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[6] Rachel Loomes, Laura Hull, and William Polmear Locke Mandy, "What is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 56, no. 6, pp. 466-474, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Susan L. Hyman et al., "Identification, Evaluation, and Management of Children with Autism Spectrum Disorder," *Pediatrics*, vol. 145, no. 1. pp. 1-69, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] John Knutsen et al., "A Systematic Review of Telemedicine in Autism Spectrum Disorders," *Review Journal of Autism and Developmental Disorders*, vol. 3, pp. 330-344, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[9] Katherine Shedlock et al., "Autism Spectrum Disorders and Metabolic Complications of Obesity," *The Journal of Pediatrics*, vol. 178, pp. 183-187, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[10] Melissa DeFilippis, "Depression in Children and Adolescents with Autism Spectrum Disorder," *Children*, vol. 5, no. 9, pp. 1-9, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[11] Cai Jie et al., "Feature Selection in Machine Learning: A New Perspective," *Neurocomputing*, vol. 300, pp. 70-79, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] E. Emary, Hossam M. Zawbaa, and Aboul Ella Hassanien, "Binary Gray Wolf Optimization Approaches for Feature Selection," *Neurocomputing*, vol. 172, pp. 371-381, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[13] Ryan J. Urbanowicz et al., "Relief-Based Feature Selection: Introduction and Review," *Journal of Biomedical Informatics*, vol. 85, pp. 189-203, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[14] R. Vaishali, and R. Sasikala, "A Machine Learning based Approach to Classify Autism with Optimum Behaviour Sets," *International Journal of Engineering and Technology*, vol. 7, no. 4, pp. 1-6, 2018. [Google Scholar] [Publisher Link]

[15] Girish Chandrashekar, and Ferat Sahin, "A Survey on Feature Selection Methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mokhlesur Rahman et al., "A Review of Machine Learning Methods of Feature Selection and Classification for Autism Spectrum Disorder," *Brain Sciences*, vol. 10, no. 12, pp. 1-23, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] In-On Wiratsin, and Lalita Narupiyakul, "Feature Selection Technique for Autism Spectrum Disorder," *Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence*, Sanya China, pp. 53-56, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Peter Washington et al., "Feature Selection and Dimension Reduction of Social Autism Data," *Pacific Symposium on Biocomputing 2020*, pp. 707-718, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Raid Alzubi, Naeem Ramzan, and Hadeel Alzoubi, "Hybrid Feature Selection Method for Autism Spectrum Disorder SNPs," *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Manchester, UK, pp. 1-7, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[20] S. Guruvammal, T. Chellatamilan, and L. Jegatha Deborah, "Optimal Feature Selection and Hybrid Classification for Autism Detection in Young Children," *The Computer Journal*, vol. 64, no. 11, pp. 1760-1774, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Delowar Hossain et al., "Detecting Autism Spectrum Disorder using Machine Learning Techniques: An Experimental Analysis on Toddler, Child, Adolescent and Adult Datasets," *Health Information Science and Systems*, vol. 9, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Autism Screening Data for Toddlers, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers

[23] K.K. Mujeeb Rahman, and M. Monica Subashini, "A Deep Neural Network-Based Model for Screening Autism Spectrum Disorder Using the Quantitative Checklist for Autism in Toddlers (QCHAT)," *Journal of Autism and Developmental Disorders*, vol. 52, pp. 2732-2746, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Ashima Sindhu Mohanty, Krishna Chandra Patra, and Priyadarsan Parida, "Toddler ASD Classification Using Machine Learning Techniques," *International Journal of Online and Biomedical Engineering*, vol. 17, no. 7, pp. 156-171, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Arjun Singh et al., "Using Machine Learning Optimization to Predict Autism in Toddlers," *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management*, Singapore, pp. 1-12, 2021. [Google Scholar] [Publisher Link]

[26] Haseeb Ali et al., "Imbalance Class Problems in Data Mining: A Review," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1560-1571, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[27] Mostofa Ahsan, Rahul Gomes, and Anne Denton, "SMOTE Implementation on Phishing Data to Enhance Cybersecurity," *2018 IEEE International Conference on Electro/Information Technology (EIT)*, Rochester, MI, USA, pp. 531-536, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[28] Shahzad Ashraf, and Tauqeer Ahmed, "Machine Learning Shrewd Approach for an Imbalanced Dataset Conversion Samples," *Journal of Engineering and Technology*, vol. 11, no. 1, pp. 1-22, 2020. [Google Scholar] [Publisher Link]

[29] Abid Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707-39716, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[30] Ahmed Jameel Mohammed, Masoud Muhammed Hassan, and Dler Hussein Kadir, "Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3161-3172, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[31] Jundong Li et al., "Feature Selection: A Data Perspective," *ACM Computing Surveys*, vol. 50, no. 6, pp. 1-45, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[32] A. Jović, K. Brkić, and N. Bogunović, "A Review of Feature Selection Methods with Applications," *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, pp. 1200-1205, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[33] Ebrahime Mohammed Senan et al., "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques," *Journal of Healthcare Engineering*, vol. 2021, pp. 1-10, 2021. [CrossRef] [Google Scholar] [Publisher

Link]

[34] Herve Nkiama, Syed Zainudeen Mohd Said, and Muhammad Saidu, "A Subset Feature Elimination Mechanism for Intrusion Detection System," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 4, pp. 148-157, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[35] Frauke Degenhardt, Stephan Seifert, and Silke Szymczak, "Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets," *Briefings in Bioinformatics*, vol. 20, no. 2, pp. 492-503, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[36] Lee Kuok Leong, and Azian Azamimi Abdullah, "Prediction of Alzheimer's Disease (AD) Using Machine Learning Techniques with Boruta Algorithm as Feature Selection Method," *Journal of Physics: Conference Series*, vol. 1372, pp. 1-9, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[37] Rong Tang, and Xiaojun Zhang, "CART Decision Tree Combined with Boruta Feature Selection for Medical Data Classification," *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, Xiamen, China, pp. 80-84, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Dragutin Petkovic et al., "Improving the Explainability of Random Forest Classifier-User Centered Approach," *Pacific Symposium on Biocomputing 2018: Proceedings of the Pacific Symposium*, pp. 204-215, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[39] Amrita Roy Chowdhury, Tamojit Chatterjee, and Sreeparna Banerjee, "A Random Forest Classifier-based Approach in the Detection of Abnormalities in the Retina," *Medical & Biological Engineering & Computing*, vol. 57, pp. 193-203, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[40] Jair Cervantes et al., "A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends," *Neurocomputing*, vol. 408, pp. 189-215, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[41] Diana C. Toledo-Pérez et al., "Support Vector Machine-Based EMG Signal Classification Techniques: A Review," *Applied Sciences*, vol. 9, no. 20, pp. 1-28, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[42] Shan Suthaharan, "Support Vector Machine," *Machine Learning Models and Algorithms for Big Data Classification, Integrated Series in Information Systems*, vol. 36, pp. 207-235, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[43] Zuherman Rustam, Ely Sudarsono, and Devvi Sarwinda, "Random-Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease (CKD)," *IOP Conference Series: Materials Science and Engineering*, vol. 546, no. 5, pp. 1-6, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[44] Seyed Reza Shahamiri, and Fadi Thabtah, "Autism AI: A New Autism Screening System Based on Artificial Intelligence," *Cognitive Computation*, vol. 12, pp. 766-777, 2020. [CrossRef] [Google Scholar] [Publisher Link]