A New Machine Learning Model based on Induction of Rules for Autism Detection

[1] Mrs. R. Indra, [2] A. Christy Gilpa

[1]Associate Professor, [2]Research Scholar
[1,2]Department of Computer Science
[1,2]Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India-620002

Abstract : The number of high-dimensional data that endures and is publically accessible on the internet has very developed in the past few years. Therefore, machine learning techniques have the challenge in dealing with the significant number of input features, which is modeling an attractive issue for researchers. To utilize machine learning techniques efficiently, preprocessing of the data is essential. Feature selection is one of the most frequent and prominent methods in data preprocessing, and has become a necessary component of the machine learning process is also known as variable selection, attribute selection, or variable subset selection in machine learning and statistics. It is the method of removing irrelevant and detecting relevant features, noisy data or redundant. This technique speeds up data mining algorithms, enhances comprehensibility and predictive accuracy. Unrelated features are those that give no useful information, and irrelevant features provide no more information than the currently selected features. Regarding supervised inductive learning, feature selection presents a set of candidate features using one of the three approaches. In this research work, a few feature selection techniques like Information Gain, Gain Ratio, Chi-Square and ReliefF are used to reduce the redundant features from the various dataset. These feature selection techniques are implemented by using MATLAB tool. Based on the performance metrics like Accuracy, Kappa Statistics and Error rates, the best feature selection technique to be selected**.**

***IndexTerms* - Autism Spectrum Disorder, MATLAB, Feature Selection, Classification, Machine Learning, Information Gain, Gain Ratio, Chi-Square and ReliefF.**

I. INTRODUCTION

Over the past two decades, there has been a significant rise in the generation, acquisition, and storage of data [1]. The era of big data is marked by both a substantial increase in the rate at which data is generated and the variety of data produced. The growth in volume, velocity, and variety of data is due in large part to the availability and affordability of assorted instruments and infrastructure, used to collect and analyze many different types of information [2]. Furthermore, the availability of various machine learning toolkits, such as Hadoop, TensorFlow, Spark, and R, combined with specialized hardware technologies has led to unique opportunities for researchers to leverage machine learning algorithms. This rise in computing power and processing technologies has opened the door for the application of machine learning theories in different fields of study. One such field is autism spectrum disorder (ASD) research [3]. ASD is a neurodevelopmental disorder characterized by deficits in social communication and social interaction, in addition to restricted, repetitive patterns of behavior (American Psychiatric Association 2013). The Centers for Disease Control and Prevention (CDC) currently estimates that one in 59 children is diagnosed with ASD in the USA (Autism and Developmental Disabilities Monitoring Network 2016) [4]. This may be an underestimate given the results of a recent parent survey finding a prevalence rate of one in 45. ASD is a heterogeneous disorder with diversity observed with respect to symptom presentation and severity, risk factors and etiology, as well as treatment response. The high prevalence rate and heterogeneous nature of ASD have led some researchers to turn to machine learning over traditional statistical methods for data analysis.

Machine learning can be broadly sorted into two categories, unsupervised and supervised learning. This paper focuses on the latter of these approaches. Supervised machine learning involves algorithms that use input variables to predict a target classification (i.e., the dependent variable), which may be categorical or continuous [5][6]. Unlike unsupervised learning (clustering), supervised learning involves datasets where the target prediction (e.g., diagnosis) is known at training time for the data used to learn the model. A supervised learning model is deemed successful when the model can (a) accurately predict the target result for a training dataset to a certain degree of accuracy and (b) be generalized to new datasets beyond those used to train the model. To improve a model's ability to make predictions on future data, a method called cross-validation is often employed. This method allows a subset of the data to be removed before training, so that the model can then be tested on Bnew data. A K-fold cross-validation strategy separates the available data into K-subsets and trains the model on all but one of the subsets and tests on the remainder. The process is repeated until the model has been trained on all the available data. The performance scores across the runs are averaged. In a leave-one-out cross-validation (LOOCV) method [7], all but one data point is used to train the model, which is then evaluated on the held out point. This process is repeated for each of the data points. A supervised machine learning model's success is typically measured according to accuracy (i.e., the ability to correctly classify into separate categories). This may be further broken down to consider sensitivity (i.e., the ability to correctly detect true positives) and specificity (i.e., the ability to correctly detect true negatives) [8][9]. A further measurement of a supervised machine learning model's success is AUC or area under the receiver operating character curve (ROC). The ROC is a plot of sensitivity vs specificity, and the area under the curve depicts how well a method makes positive and negative categorical distinctions [10].

II. RELATED WORKS

Sudha, V. Pream, and M. S. Vijaya [11] Autism spectrum disorder (ASD) is characterized by a set of developmental disorders with a strong genetic origin. The genetic cause of ASD is difficult to track, as it includes a wide range of developmental disorders, a spectrum of symptoms and varied levels of disability. Mutations are key molecular players in the cause of ASD, and it is essential to develop effective therapeutic strategies that target these mutations. The development of computational tools to identify ASD originated by genetic mutations is vital to aid the development of disease-specific targeted therapies. This chapter employs supervised machine learning techniques to construct a model to identify syndromic ASD by classifying mutations that underlie these phenotypes, and supervised learning algorithms, namely support vector machines, decision trees and multilayer perceptron, are used to explore the results. It has been observed that the decision tree classifier performs better compared to other learning algorithms, with an accuracy of 94%. This model will provide accurate predictions in new cases with similar genetic background and enable the pathogenesis of ASD.

Bishop- Fitzpatrick, Lauren, et al [12] We analyzed all ICD-9 codes, V-codes, and E-codes available in the electronic health record and Elixhauser comorbidity categories associated with those codes. Diagnostic patterns distinguished decedents with ASD from decedent community controls with 75% sensitivity and 94% specificity solely based on their lifetime ICD-9 codes, V-codes, and E-codes. Decedents with ASD had higher rates of most conditions, including cardiovascular disease, motor problems, ear problems, urinary problems, digestive problems, side effects from long-term medication use, and nonspecific lab tests and encounters. In contrast, decedents with ASD had lower rates of cancer. Findings suggest distinctive lifetime diagnostic patterns among decedents with ASD and highlight the need for more research on health outcomes across the lifespan as the population of individuals with ASD ages.

Payabvash, Seyedmehdi, et al. [13] The Edge Density (ED) maps were computed from probabilistic streamline tractography applied to high angular resolution diffusion imaging (HARDI). Tract-Based Spatial Statistics (TBSS) was used for voxel-wise comparison and coregistration of ED maps in addition to conventional DTI metrics of Fractional Anisotropy (FA), Mean Diffusivity (MD), and Radial Diffusivity (RD). Tract-based average DTI/connectome metrics were calculated and used as input for different machine learning models: naïve Bayes, random forest, support vector machines (SVM), neural networks. For these models, crossvalidation was performed with stratified random sampling (×1000 permutations). The average accuracy among validation samples was calculated. In voxel-wise analysis, the body and splenium of corpus callosum, bilateral superior and posterior corona radiata, and left superior longitudinal fasciculus showed significantly lower ED in children with ASD; whereas, we could not find significant difference in FA, MD, and RD maps between the two study groups. Overall, machine-learning models using tract-based ED metrics had better performance in identification of children with ASD compared to those using FA, MD, and RD. The EDI-based random forest models had greater average accuracy (75.3%), specificity (97.0%), and positive predictive value (81.5%), whereas EDI-based polynomial SVM had greater sensitivity (51.4%), and negative predictive values (77.7%). In conclusion, we found reduced density of connectome edges in the posterior white matter tracts of children with ASD; and demonstrated the feasibility of connectome-based machine-learning algorithms in identification of children with ASD.

Heinsfeld, Anibal Sólon, et al. [14] We investigated ASD patients brain imaging data from a world-wide multi-site database known as ABIDE (Autism Brain Imaging Data Exchange). ASD is a brain-based disorder characterized by social deficits and repetitive behaviors. According to recent Centers for Disease Control data, ASD affects one in 68 children in the United States. We investigated patterns of functional connectivity that objectively identify ASD participants from functional brain imaging data, and attempted to unveil the neural patterns that emerged from the classification. The results improved the state-of-the-art by achieving 70% accuracy in identification of ASD versus control patients in the dataset. The patterns that emerged from the classification show an anticorrelation of brain function between anterior and posterior areas of the brain; the anticorrelation corroborates current empirical evidence of anterior-posterior disruption in brain connectivity in ASD. We present the results and identify the areas of the brain that contributed most to differentiating ASD from typically developing controls as per our deep learning model.

Abbas, Halim, et al [15] In this work, we apply Machine Learning (ML) to gold standard clinical data obtained across thousands of children at risk for autism spectrum disorders to create a low-cost, quick, and easy to apply autism screening tool that performs as well or better than most widely used standardized instruments. This new tool combines two screening methods into a single assessment, one based on short, structured parent-report questionnaires and the other on tagging key behaviors from short, semi-structured home videos of children. To overcome the scarcity, sparsity, and imbalance of training data, we apply creative feature selection, feature engineering, and novel feature encoding techniques. We allow for inconclusive determination where appropriate in order to boost screening accuracy when conclusive. We demonstrate a significant accuracy improvement over standard screening tools in a clinical study sample of 162 children.

III. PROBLEM STATEMENT

Diagnostic Instruments and Criteria Seven studies examined how the three subtypes differed on diagnostic measures such as the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule. The ADI-R is a semi-structured parent interview that operationalizes DSM-IV and ICD-10 criteria, assessing domains of social interaction, communication, and restricted, repetitive behaviors/interests. The ADOS is a standardized observation of social behavior and communication in a play-based setting.

- Decreased classification accuracy.
- Increased error rates.
- The value of true positive, false positive and false negative are not more accurate.

IV. PROPOSED METHODOLOGY

In this research work, a few feature selection techniques like Information Gain, Gain Ratio, Chi-Square and ReliefF are used to reduce the redundant features from the various dataset. These feature selection techniques are implemented by using MATLAB tool. Based on the performance metrics like Accuracy, Kappa Statistics and Error rates, the best feature selection technique to be selected. To utilize machine learning techniques efficiently, preprocessing of the data is essential. Feature selection is one of the most frequent and prominent methods in data preprocessing, and has become a necessary component of the machine learning process is also known as variable selection, attribute selection, or variable subset selection in machine learning and statistics. It is the method of removing irrelevant and detecting relevant features, noisy data or redundant. This technique speeds up data mining algorithms, enhances comprehensibility and predictive accuracy. Unrelated features are those that give no useful information, and irrelevant features provide no more information than the currently selected features. Regarding supervised inductive learning, feature selection presents a set of candidate features using one of the three approaches.

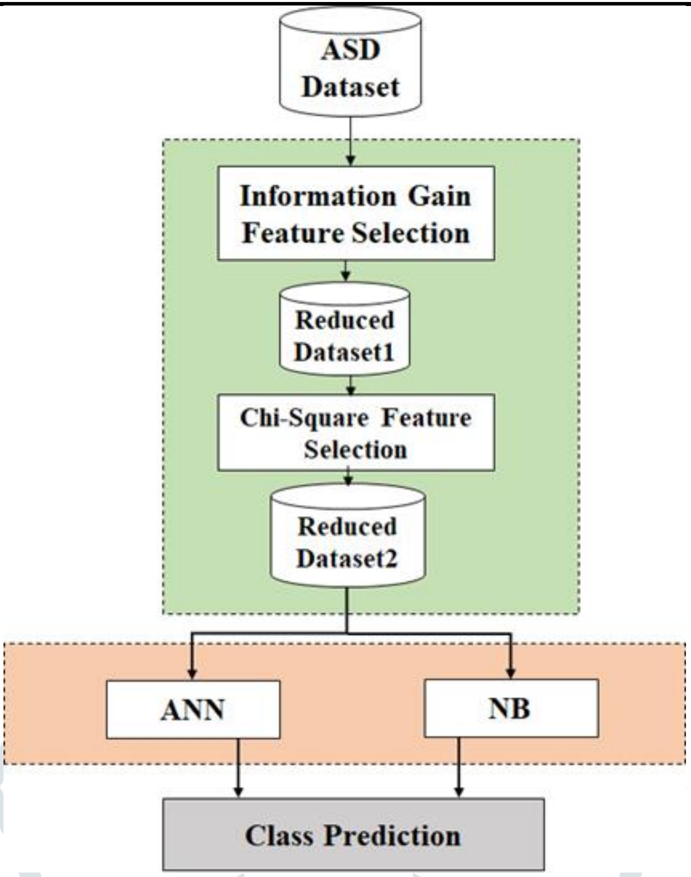

Figure 1: Architecture Diagram of ASD Detection using Data Mining Techniques

4.1Information Gain

Entropy is generally worked in the information theory measure, which symbolizes the clarity of the subjective collection of examples. It is in the establishment of Gain Ratio, Information Gain and Similarity Uncertainty (SU). The entropy criterion is considered through the measure of the system's randomness. The entropy of Y is

$$H(Y) = \sum_{y\in Y} p(y) \log_2(p(y))$$

where *p(y)* is the marginal probability density function for the random variable *Y*. If the experiential values of *Y* in the training data set *S* are apportioned according to the values of a second feature *X*, and the entropy of *Y* with reference to the partitions prompted by *X* is less than the entropy of *Y* prior to partitioning, then there is an association between the features *Y* and *X*. The entropy of *Y* after observing *X* is then:

$$H(Y|X) = \sum_{x\in X} p(x) \sum_{y\in Y} p(y|x)\ \log_2\left(p(y|x)\right)$$

where *p(y |x)* is the conditional probability of *y* given *x*. The given entropy is a criterion of impurity in a training set *S*, we can state the measure which reflects further information about *Y* provided by *X* that epitomizes the quantity by which, the entropy of *Y* decreases. This measure is acknowledged as IG. It is given by

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

IG is a symmetrical measure and it is given by equation (3.3). The information gained about *Y* after detecting *X* is equal to the information gained about *X* after observing *Y*. A weakness of the IG criterion is that it is predisposed in approval of features with further values even when they are not more informative [16][17][19][20].

4.2 Chi-Square Analysis

Feature Selection via chi square χ^2test is another, regularly used technique. Chi-squared attribute evaluation assesses the value of a feature by calculating the value of the chi-squared statistic with reverence to the class [16][17][18]. The preliminary hypothesis H_0 is the assumption that the two features are distinct, and it is tested by chi-squared formula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{O_{ij} - E_{ij}}{E_{ij}}\right)^2$$

where O_{ij} is the observed frequency and E_{ij} is the expected (theoretical) frequency, proclaimed by the null hypothesis. The greater the value of χ^2, the greater the evidence against the hypothesis H_0

4.3 Artificial Neural Network Classification Method

Artificial Neural Network (ANN) is an efficient computing system whose central theme is borrowed from the analogy of biological neural networks. ANNs are also named as "artificial neural systems," or "parallel distributed processing systems," or "connectionist systems." ANN acquires a large collection of units that are interconnected in some pattern to allow communication between the units. In order to form a feed-forward multi-layer in MLP, the collection of non-linear neurons is connected to one another. This technique is known to be very useful for prediction and classification issues. Cross-validation is used to determine the 'optimal' number of hidden layers and neurons which were relied on the experimental design of the ASD classification framework. These units, also referred to as nodes or neurons, are simple processors which operate in parallel [18][21][22].

4.4 Naïve Bayes Classification Method

Bayesian network classifiers are a popular supervised classification paradigm. A well-known Bayesian network classifier is the Naïve Bayes' classifier is a probabilistic classifier based on the Bayes' theorem, considering Naïve (Strong) independence assumption. It was introduced under a different name into the text retrieval community and remains a popular(baseline) method for text categorizing, the problem of judging documents as belonging to one category or the other with word frequencies as the feature. An advantage of Naïve Bayes' is that it only requires a small amount of training data to estimate the parameters necessary for classification. Abstractly, Naïve Bayes' is a conditional probability model. Despite its simplicity and strong assumptions, the naïve Bayes' classifier has been proven to work satisfactorily in many domains. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. In Naïve Bayes' technique, the basic idea to find the probabilities of categories given a dataset by using the joint probabilities of words and categories. It is based on the assumption of word independence [23].

V. RESULTS AND DISCUSSION

5.1 Description of the Dataset

Table 1 depicts the description of the dataset.

Table 1: Description of the ASD Dataset

Attribute	Type	Description
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician ,etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2= adolescent, 3= adult)
Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

5.2 Performance Analysis of the Feature Selection Method

Table 2 gives the result obtained by using Information Gain, and Chi-Square Feature selection method. Table 3 gives the performance analysis of the IG and CS by using Naïve Bayes classification method. Table 4 presents the performance analysis of IG and CS by using Artificial Neural Network classification method.

Table 2: Number of features obtained by IG and CS feature selection methods

Sl.No	Information Gain	Chi-Squared analysis
1	(A4_Score)	country_of_res
2	country_of_res	(A9_Score)
3	(A9_Score)	(A10_Score)
4	(A10_Score)	(A8_Score)
5	(A8_Score)	(A6_Score)
6	(A6_Score)	(A3_Score)
7	(A1_Score)	(A1_Score)
8	(A5_Score)	(A5_Score)
9	ethnicity	(A7_Score)
10	(A2_Score)	(A2_Score)
11	autism	ethnicity
12	gender	autism
13	used_app_before	gender
14	jaundice	used_app_before
15	age	jaundice
16	age_dec	age
17		age_dec

Table 3: Performance Analysis of Information Gain, Chi-Square analysis by using Naïve Bayes classification method

Performance Metrics	Feature Selection Methods	
	Information Gain	Chi-Square
Accuracy	69.4779 %	69.4779 %
Kappa Statistic	0.1869	-0.0028
Mean Absolute Error	0.3721	0.3627
Root mean squared error	0.4543	0.4575
Relative absolute error	88.9114 %	97.7234 %
Root relative squared error	99.376 %	106.3606 %
TP Rate	0.695	0.695
FP Rate	0.527	0.697
Precision	0.667	0.629
Recall	0.695	0.695
F-Measure	0.673	0.652
ROC Area	0.654	0.559

Table 4: Performance Analysis of Symmetrical Uncertainty, Information Gain, Chi-Square analysis by using Artificial Neural Network classification method

Performance Metrics	Feature Selection Methods	
	Information Gain	Chi-Square
Accuracy	62.249 %	65.0602 %
Kappa Statistic	0.243	0.3008
Mean Absolute Error	0.4582	0.4544
Root mean squared error	0.497	0.4923
Relative absolute error	91.6985 %	90.954 %
Root relative squared error	99.4351 %	98.4887 %
TP Rate	0.622	0.651
FP Rate	0.38	0.35
Precision	0.622	0.651
Recall	0.622	0.651
F-Measure	0.622	0.651
ROC Area	0.634	0.645

VI. CONCLUSION

Autism spectrum disorder being an incurable neuro-development disorder, needs to be detected at an early stage. Its detection at an early stage will not only prepare and parents but will also help in devising the education tools and tricks to help them educate and mature their children. Classification is relatively the most suitable technique for detecting the disease. Although, there have been other techniques like clustering which has been applied in four fields like environment, biotic, irrigated area, corp yields. There has also been further improvement in early detection and finding causes of this disease by using algorithms like feature selection methods and classification techniques like Naïve Bayes, Multi-Layer perceptron neural network are utilized in this project to detect the accuracy of the ASD children using feature selection.

In future work, a more thorough comparative analysis of all these approaches. It has been believing that such reflection may help us better address the classification problems experienced by the dataset. The size of the dataset can also be increased to improve the accuracy of the classification of ASD. Using this project, automation of the finding of ASD in early stage can be done.

REFERENCES

[1] P. Szatmari, S. Georgiades, E. Duku, T.A. Bennett, S. Bryson, E. Fombonne, et. al., ‘Developmental trajectories of symptom severity and adaptive functioning in an inception cohort of preschool children with autism spectrum disorder’, JAMA psychiatry, vol. 72(3), 2015, pp. 276-283.

[2] C. Lord, and R.M. Jones, ‘Annual Research Review: Re- thinking the classification of autism spectrum disorders’, Journal of Child Psychology and Psychiatry, vol. 53(5), 2012, pp. 490-509.

[3] C. Lord, S. Risi, L. Lambrecht, E.H. Cook, B.L. Leventhal, P.C. DiLavore, et. al., ‘The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism’, Journal of autism and developmental disorders, vol. 30(3), 2000, pp. 205-223 International Journal of Pure and Applied Mathematics Special Issue 434.

[4] C. Gillberg, ‘Autism and related behaviours’, Journal of Intellectual Disability Research, vol. 37(4), 1993, pp. 343-372

[5] C. Lord, M. Rutter, and A. Couteur, ‘Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders’, Journal of autism and developmental disorders, vol. 24(5), 1994, pp. 659-685

[6] K. Gotham, A. Pickles,and C. Lord, ‘Trajectories of autism severity in children using standardized ADOS scores’, Pediatrics, vol. 130(5), 2012,pp. e1278-e1284

[7] M.P. van den Heuvel and O. Sporns, ‘Network hubs in the human brain’, Trends in cognitive sciences, vol. 17(12), 2013, pp. 683-696

[8] M. Rubinov, and O. Sporns, ‘Complex network measures of brain connectivity: uses and interpretations’, Neuroimage, vol. 5 (3), 2010, pp. 1059-1069

[9] L.E. Libero, T.P. DeRamus, A.C. Lahti, G. Deshpande, and R.K. Kana, ‘Multimodal neuroimaging based classification of autism spectrum disorder using anatomical, neurochemical, and white matter correlates’, Cortex, vol. 66, 2015, pp. 46-59

[10] G. Deshpande, L.E. Libero, K.R. Sreenivasan, H.D. Deshpande, and R.K. Kana, ‘Identification of neural connectivity signatures of autism using machine learning’, vol.7, 2013.

[11] Sudha, V. Pream, and M. S. Vijaya. "Machine Learning-Based Model for Identification of Syndromic Autism Spectrum Disorder." *Integrated Intelligent Computing, Communication and Security*. Springer, Singapore, 2019. 141-148.

[12] Bishop- Fitzpatrick, Lauren, et al. "Using machine learning to identify patterns of lifetime health problems in decedents with autism spectrum disorder." *Autism Research* 11.8 (2018): 1120-1128.

[13] Payabvash, Seyedmehdi, et al. "White Matter Connectome Edge Density in Children with Autism Spectrum Disorders: Potential Imaging Biomarkers Using Machine-Learning Models." *Brain connectivity* 9.2 (2019): 209-220.

[14] Heinsfeld, Anibal Sólon, et al. "Identification of autism spectrum disorder using deep learning and the ABIDE dataset." *NeuroImage: Clinical* 17 (2018): 16-23.

[15] Abbas, Halim, et al. "Machine learning approach for early detection of autism by combining questionnaire and home video screening." *Journal of the American Medical Informatics Association* 25.8 (2018): 1000-1007.

[16] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems,* Preprint: 1-18.

[17] M. Durairaj, T S Poornappriya, “Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM.”, *International Journal of Engineering & Technology*, 7 (4) (2018) 5856-5861.

[18] M. Durairaj, T. S. Poornappriya, “Importance of MapReduce for Big Data Applications: A Survey”, *Asian Journal of Computer Science and Technology,* Vol.7 No.1, 2018, pp. 112-118.

[19] M. Lalli, V.Palanisamy,(2016), “Filtering Framework for Intrusion Detection Rule Schema in Mobile Ad Hoc Networks”, International Journal of Control Theory and Applications –(IJCTA),9(27), pp. 195-201, ISSN: 0974-5572

[20] M. Lalli, V.Palanisamy,(2017), “Detection of Intruding Nodes in Manet Using Hybrid Feature Selection and Classification Techniques”, Kasmera Journal, ISSN: 0075-5222, 45(1) (SCIE)(Impact Factor:0.071).

[21] M. Lalli, V.Palanisamy, (Sep 2014), “A Novel Intrusion Detection Model for Mobile Adhoc Networks using CP-KNN”, International Journal of Computer Networks & Communications- (IJCNC), Vol.6, No.5, ISSN:0974-9322.

[22] M. Lalli, “Statistical Analysis on the KDD CUP Dataset for Detecting Intruding Nodes in MANET”, *Journal of Applied Science and Computations,* Volume VI, Issue VI, JUNE/2019, 1795-1813.

[23] M. Lalli, “Intrusion Detection Rule Structure Generation Method for Mobile Ad Hoc Network”, *Journal of Emerging Technologies and Innovative Research*, June 2019, Volume 6, Issue 6, 835-843.