Apriori-backed Fuzzy Unification and Statistical Inference in Feature Reduction: An Application in Prognosis of Autism in Toddlers

Shithi Maitra[1], Nasrin Akter[2], Afrina Zahan Mithila[3], Tonmoy Hossain[4], and Mohammad Shafiul Alam[5]

Sheba.xyz[1], East West University[2], and
Ahsanullah University of Science and Technology[3,4,5]
shithi@sheba.xyz[1], nipa.ete@gmail.com[2], azmithila2014@gmail.com[3],
tonmoyhossain.cse@ieee.org[4], shafiul.cse@aust.edu[5]

Abstract. Weak Artificial Intelligence (AI) allows the application of machine intelligence in modern health information technology to support medical professionals in bridging physical/psychological observations with clinical knowledge, thus generating diagnostic decisions. Autism, a highly variable neurodevelopmental condition marked by social impairments, reveals symptoms during infancy with no abatement with time due to comorbidities. There exist genetic, behavioral, neurological actors playing roles in the making of the disease and this constructs an ideal pattern recognition task. In this research, the *Autism Screening Data (ASD)* for toddlers was initially exploratorily analyzed to hypothesize impactful features which were further condensed and inferentially pruned. An interesting application of the business intelligence algorithm: *Apriori* has been made on transactions consisting of ten features and this has constituted a novel preprocessing step derived from *market basket analysis*. The huddling features were fuzzily modeled to a single feature, the membership function of which evaluated to the degree to which a toddler could be called autistic, thus paving the way to the first optimized Neural Network (NN). Features were further eliminated based on statistical t-tests and *Chi*-squared tests, administering features only with *p-values* < 0.05—giving rise to the second and final optimized model. The research showed that the unremitted 16-feature and the optimized 5-feature models showed equivalence in terms of maximum test accuracy: 99.68%, certainly with lower computation in the optimized scheme. The paper follows a 'hard (EDA, inferential statistics) + soft (fuzzy logic) + hard (forward propagation) + soft (backpropagation)' pipeline and similar systems can be used for similar prognostic problems.

Keywords: autism spectrum disorder, exploratory data analysis, *Apriori* algorithm, fuzzy modeling, membership functions/values, inferential statistics, t-test, *Chi*-squared (χ^2) independence test, *ANOVA*-test

1 Introduction

Autism is a neurodevelopmental condition marked by stereotypical, compulsive, repetitive, ritualistic behavior and sometimes by limited interests and tendency towards self-harm, in extreme cases—noticed in early childhood with no diminutive effect along with the increase of age. The said disorder shows a severity-gradient [1] and hence the term Autism Spectrum Disorder (ASD) is coined. The developed world is inflicted more with its scourge and almost 1.5% of children were diagnosed autistic in 2017 [2]. While early intervention is necessary to groom the affected individuals for better self-care, the UK National Autism Plan for Children presently endorses an assessment as lengthy as 30 weeks [3].

Artificially intelligent *Clinical Decision Support Systems (CDSS)* can draw demarcations between neurologically sound and unhealthy subjects given a proper knowledge-base. A CDSS can be developed and deployed using clinical data mining which is the process of extracting medical insights from diagnostic data. This paper implements the steps of building such a system—which can significantly spare the time for prognosis, the working hours and efforts of a physician and can make caregivers mentally, financially prepared—by:

- mining a soundly, consensually collected dataset that captures both historical, behavioral aspects
- performing an intensive preprocessing and producing three variants of the screening dataset
- modeling the data to an appropriate algorithm for generating the most practical prognosis in light of known cases

It is expected that a bulky set of toddlers' diagnostic data may be uncertain, imprecise, partially true and an exact solution might be infeasible—hence the appeal of soft computing methods. The beauty of this recent development in computing is that it has a humanoid, heuristic process of giving a useful, optimal solution. The research at hand complementarily applies two components of soft computing for the prognosis of a probable ASD trait, namely: fuzzy logic and neural networks (NNs). The layered, hierarchical structure of a neural network can recognize complex patterns spread through multiple dimensions by iteratively refining some initial set of parameters using back-propagation.

Personality traits in a particular cohort (in our case, the autistic toddlers) often show up in bundles, due to which association rule mining can serve the purpose of identifying the frequent traits occurring together in toddlers' behavior. *Apriori* algorithm [*Agrawal and Srikant, 1994*] can be used for defining such rules, which may club several features together by aggregation. This research attempts to structure the screening data in a transactional form for the application of *Apriori*. The clubbed outcome of such basket analysis is then fuzzily modeled, to capture the vague and imprecise effect.

This era of big data demands the usage of clean, dimensionally reduced data since an increased dimensionality demands greater numbers of examples for sound training of a soft, predictive model—otherwise known as the curse of dimensionality. Hence this research first produces a conglomeration of as many

as ten features and further prunes the remaining features based on inferential statistics. Concretely, this paper poses a supervised binary classification problem upon extensive, novel preprocessing of toddlers' data, to label them as neurologically typical or autistic instances following a hard-soft-hard-soft pipeline. A concise review of existing literature, rendering of the proposed methodology, followed by tabulation and explanation of results concludes and constructs the paper.

2 Literature Review

The recent academic literature invested in prognosticating autistic behavior encompasses multi-sourced data collection of random-control groups, its analysis (both statistical and predictive), elimination of redundant features and finally generating the output using both structured data, sequence models and images (Fig. 1). A gradual unraveling of such methods to our purpose is demonstrated using four paradigms as stated below.

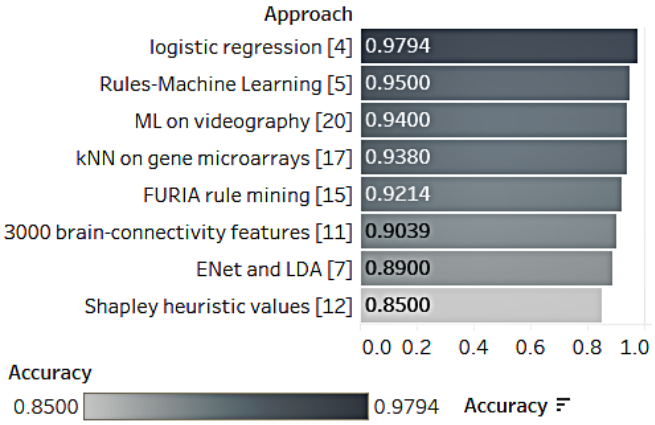


Fig. 1. a synoptic review of related literature

2.1 Data Garnering and Predictive Analyses using Basic ML

Thabtah [4] firstly developed a mobile application through which he could construct the ASD dataset which contained information of toddlers, children, adolescents and adults. The effort created a motley dataset which contained data from different countries, languages, ethnicities and introduced the Q_{10}-chat. He deployed Naive Bayes and logistic regression which earned him an accuracy of 97.94% and a recall of 98%. Next, Thabtah and Peebles [5] developed a medical decision support system introducing Rules Machine Learning, which gave both diagnosis and insights. This is an advancement that utilized their legacy Q_{10}-chat questionnaire [4] which has also been used in this research. This earned them the accuracy of a highest 95% and recall of a highest 97%.

Sarkar, Wade et al. [6] drilled down to the toddlers' group and built a tablet-run application that assessed the severity of autistic behavior. The pilot evaluation showed potential in detecting the syndrome with an F1-score of 0.94. Duda et al. [7] classified between ASD, ADHD using 15 feature-like responses that were collected through a rigorous, holistic data collection procedure. They dealt with imbalanced data and obtained an accuracy of 0.89 ± 0.01 using ENet and LDA classifiers.

2.2 Approaches to Eliminate Redundant Features

Achenie, Scarpa et al. [8] applied a feed-forward neural network on the M-CHAT-R dataset for the screening of autism in toddlers. The reduced the scope of features from 20 to 18 to a final 14 and measured its effectiveness gender/ethnicity-wise. They introduced maternal educational period as a feature and found 99.72% accuracy. Thabtah et al. [9], as a sequel, proposed Variable Analysis (Va) that helped shed off features based on correlations.

Abbas et al. [10] accumulated both structured and graphical content and trained two different classifiers, to combine them finally to produce both conclusive and inconclusive outcomes, after much feature-scrambling. Kong, Gao et al. [11] extracted brain connectivity data from images and capped the number of features to 3000 in a descending sequence of F1-score, achieving a 90.39% accuracy. Tariq, Fleming et al. [12] applied Shapley heuristic values to determine the importance of features which led them to 85% accurate outcomes.

2.3 Applications using Fuzzy Logic

Farsi et al. [13] argued that Fuzzy Cognitive Maps (FCMs) have limitations in handling a great level of uncertainty due to the causal inferences they tend to make and use Interval Agreement Approach to assign weights to the links of FCMs in order to model the uncertainties better in predicting ASD. Al-diabat [14] applied fuzzy rule mining and extracted 29 (11 'yes'+18 'no') fuzzy rules, contending FURIA to be the most effective with an accuracy of 90%. He identified the most frequently occuring features to be the most influential. Khan, Alshara [15] similarly endorsed FURIA with 92.14% accurate predictions.

2.4 Usage of Data from Variegated Sources

Xiao, Fang et al. [16] intended to avoid reporting bias for ASD screening and found random forest classifier as the best-performing on regional cortical thickness data obtained by neuroimaging. This performed better than volume and area-based data, using 20 highest importance regions. Kim et al. [17] examined blood gene expression profiles and used microarray data to first verify distinguishability using clustering and finally found 93.8% accuracy using kNN. Karan et al. [18] examined electroencephalogram (EEG) signals' data to classify among autistic and neuro-typical subjects with 71% precision.

Abbas, Garberson [19] et al. applied machine learning to structured data obtained from children's parents' responses to questionnaires and also to short snippets of videos obtained from their homes. They proposed an extension of their work for diagnosing other neurodevelopmental conditions as well. Tariq, Daniels [20] et al. observed 30 behaviors from 3-minute home videos of children with and without ASD traits and found $> 94\%$ accuracy by modeling them to ML algorithms.

This paper is a different endeavor in that it attempts association rule mining to find a grouping tendency among features and hence justifiably clubs them up to a fuzzy feature to find comparable accuracy using fewer features, hence combining ML, feature reduction and fuzzy logic.

3 Proposed Model

The screening data for autism in toddlers [4] was researched in a modular fashion: with the first module performing an intense preprocessing on the raw data, the second module finding and fitting an appropriate predictive model and the final one evaluating the efficacy of the former two in generating the desired outcome (Fig. 2).

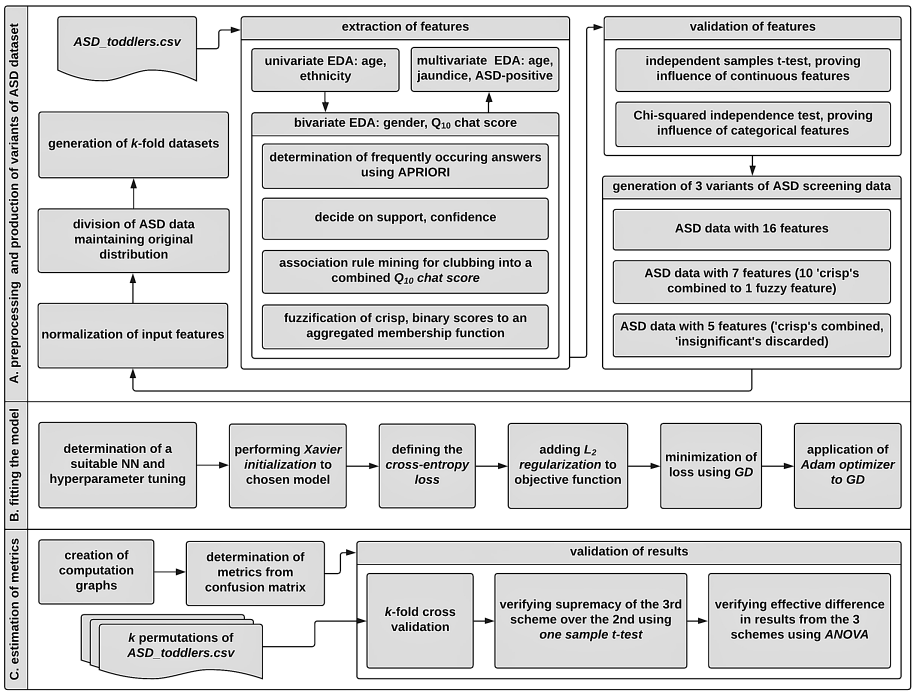


Fig. 2. workflow for implementing the proposed autism-decision support system in toddlers

3.1 Preprocessing ASD Data and Producing of Variants of the Dataset

Exploratory data analysis (EDA) is a preliminary visual summarization of a dataset for formulating hypotheses to be further tested and for developing insights into what mathematical model should fit the data best. In this piece, we employ EDA in three modalities: univariate, bivariate and multivariate—depending on how many variables are under study. We use EDA to initially hypothesize an attribute as a feature and further bolster this hypothesis by inferential statistics. The features thus found are transformed to fit a predictive model, introducing *Initial Data Analysis (IDA)* to the process, which is encompassed by EDA. We firstly explore singularly (hence, univariate) the impact

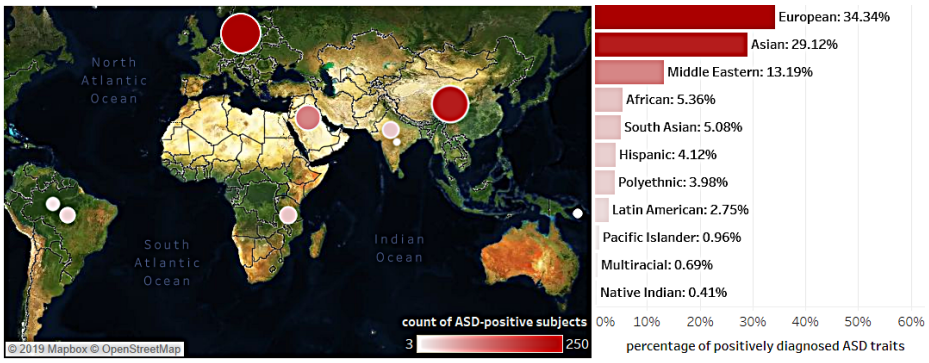


Fig. 3. geographic distribution of ASD-positive toddlers among different ethnic groups

of the continuous variable: age and secondly analyze the casual correlation of the discrete variable: ethnicity upon a toddler being autistic. The analysis (Fig. 3) reveals that among the culturally variegated data, toddlers from a European descent (34.34%) are the most affected with the condition followed by Asians (29.12%), whereas children from South Asian, Hispanic, Latin American ancestry are the least inflicted (less than 5%). The age-range of 12 to 36 months being definitive of toddlers, the exploratory analysis reveals 24 months to 36 months to be appropriate for a positive diagnosis (Fig. 4).

Next, we delve into a bivariate analysis in that we examine the interaction between ASD traits and Q_{10}-chat score, gender. We discover that the 10 questions, answered by parents or surrogate caregivers, elicit an affirmative response for neurologically sound subjects (except for Q_7). We examine the inverse answers and find that there exists a considerably large gap when it comes to the unsound subjects (Fig. 5). We perform an *Apriori*-based aggregation to the data to generate the Q_{10}-chat score and find its positive correlation with ASD-positive subjects. We further find that despite a 69.73% share of male toddlers in the dataset, a greater 73.35% of the ASD-positive cases are among the males (Fig. 6).

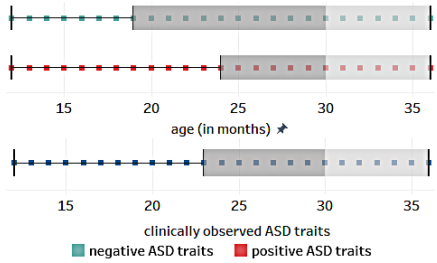


Fig. 4. box-and-whisker plot demonstrating effective age-ranges for prognosis

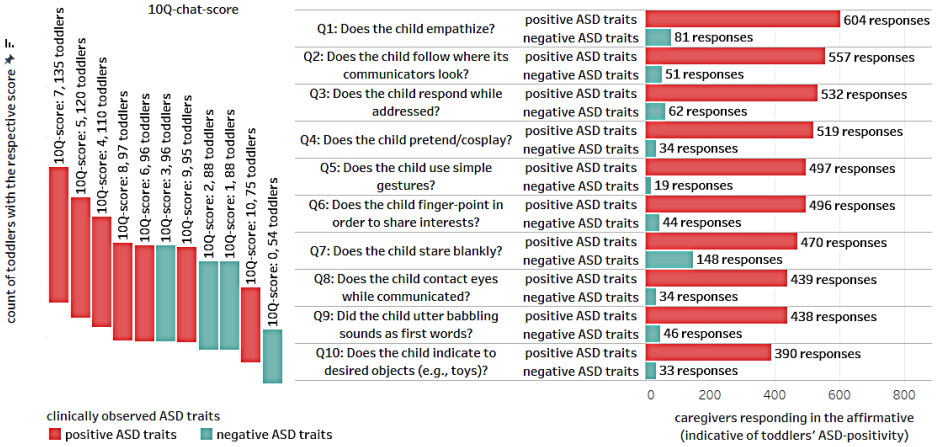


Fig. 5. toddlers scoring more (more than 3) in the Q_{10} chat seem more likely to have autism

It is understandable that a representative, combined feature instead of 10 different features would be convenient and more computationally efficient for running expensive soft computing algorithms. According to the jargon of database management, we treat the dataset as a relation and apply *Apriori* association rule mining and find frequently occurring positive answer-sets. There exists its application in *market basket analysis*, but the application here is relevant because a tendency of appearing together had been detected in EDA (Fig. 5). The hyperparameters (Fig. 7): support ≥ 0.02 and confidence ≥ 0.8, have led to 6 association rules (Fig. 8) with full confidence that justify the amalgamation of the answers to a single feature.

Fuzzy models are capable of representing and utilizing partly true, vague and imprecise information mathematically. It is particularly useful in medical decision-making frameworks because it can appropriately define the degree to which a symptom is being visible. After discovering that Q_{10} answers appear frequently together using *Apriori*, we define a fuzzy process in the following way (Fig. 9) which transformed 10 attributes into an enveloped feature:

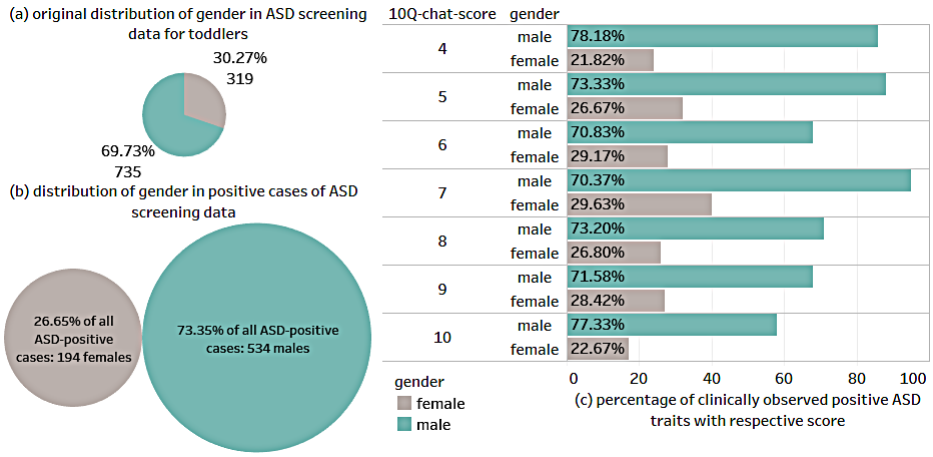


Fig. 6. males showing a greater tendency towards a positive diagnosis

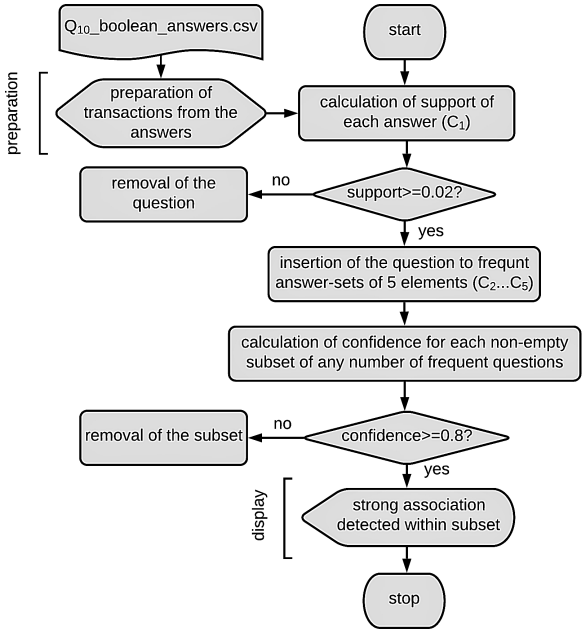


Fig. 7. ASD-specific *Apriori* algorithm devised and deployed for effective clubbing of Q_{10} answers

– First, we take as input crisp binary answers and fuzzify them using a simple, discrete membership function $\mu_{autistic}(toddler)$ of aggregation.
– Then we execute applicable rules from the rule-base we define for different severity-levels of ASD [21] and calculate the answers' average, which serves

as a fuzzy output value (discrete membership value, $support(autistic) = \{toddler \mid 0 \leq \mu_{autistic}(toddler) \leq 1\}$).
– Finally, the said average serves as a crisp output once we convert it into percentage and this defines the extent to which a toddler can be called autistic according to its Q_{10}-score.

```
> inspect(head(sort(itemsets, by="support"), 20))
     items              support   count
[1]  {a1,a5,a6,a7,a9}   0.2125237 224
[2]  {a4,a5,a6,a7,a9}   0.2115750 223
[3]  {a1,a4,a6,a7,a9}   0.2115750 223
[4]  {a1,a2,a4,a6,a7}   0.2096774 221
[5]  {a1,a4,a5,a6,a7}   0.2077799 219
[6]  {a1,a2,a5,a6,a7}   0.2068311 218
[7]  {a1,a2,a6,a7,a9}   0.2011385 212
[8]  {a3,a4,a6,a7,a9}   0.1992410 210
[9]  {a1,a4,a5,a7,a9}   0.1982922 209
[10] {a3,a4,a5,a6,a7}   0.1963947 207
[11] {a1,a4,a5,a6,a9}   0.1963947 207
[12] {a3,a4,a5,a7,a9}   0.1925996 203
[13] {a1,a3,a4,a6,a7}   0.1925996 203
[14] {a3,a4,a5,a6,a9}   0.1916509 202
[15] {a1,a2,a6,a7,a10}  0.1888046 199
[16] {a1,a2,a4,a7,a9}   0.1888046 199
[17] {a1,a2,a4,a5,a7}   0.1878558 198
[18] {a1,a2,a4,a5,a6}   0.1850095 195
[19] {a4,a6,a7,a8,a9}   0.1840607 194
[20] {a5,a6,a7,a8,a9}   0.1840607 194

> inspect(strong_rules)
     lhs                          rhs  support    confidence lift
[1]  {a2,a3,a5,a7,a8,a9,a10}      => {a4} 0.07685009 1.0000000  1.951852
[2]  {a2,a3,a5,a6,a7,a9,a10}      => {a4} 0.08918406 1.0000000  1.951852
[3]  {a1,a2,a3,a5,a7,a8,a9,a10}   => {a4} 0.07495256 1.0000000  1.951852
[4]  {a2,a3,a5,a6,a7,a8,a9,a10}   => {a4} 0.07210626 1.0000000  1.951852
[5]  {a1,a2,a3,a5,a6,a7,a9,a10}   => {a4} 0.08633776 1.0000000  1.951852
[6]  {a1,a2,a3,a5,a6,a7,a8,a9,a10} => {a4} 0.07115750 1.0000000  1.951852
[7]  {a2,a3,a5,a7,a9,a10}         => {a4} 0.09582543 0.9901961  1.932716
[8]  {a1,a2,a3,a5,a7,a9,a10}      => {a4} 0.09108159 0.9896907  1.931730
[9]  {a1,a3,a5,a7,a8,a9,a10}      => {a4} 0.08918406 0.9894737  1.931306
[10] {a3,a5,a6,a7,a8,a9,a10}      => {a4} 0.08918406 0.9894737  1.931306
```

Fig. 8. output generated from applying hyperparametrically tuned *Apriori*

Having concluded the bivariate analysis, we explore the relationships individually between age, family history, neonatal jaundice history, caregivers' association and positive ASD traits (Fig. 10). The multivariate analysis (Fig. 10(a)) exposes that among the subjects with positive ASD traits, 29.53% have had a history of neonatal jaundice while among normal subjects jaundice was less pervasive, 22.39%. If family-history (Fig. 10(b)) or the test-administering caregiver (Fig. 10(c)) plays a role in ASD is unclear from EDA, and is being left for statistical inference to clarify. EDA thus helped to investigate and hypothesize features, which were further validated/invalidated by statistical inference.

Table 1. Welch Two Sample t-test results

continuous features	t-score	degrees of freedom	p-*value*	$H_0 : \mu_1 = \mu_2$	$H_a : \mu_1 \neq \mu_2$
10 questions' quiz-score	-55.072	1006.4	< 2.2E-16	reject	retain
age (months)	-2.0323	537.95	4.26E-02	reject	retain

Hypothesized features need validations that the tendencies shown by EDA will indeed hold if further data-points are added and that their impact on the classes (autistic/typical) is not due to chance. Inferential statistics does just that by applying the *Chi*-squared independence test on discrete and independent samples t-test on continuous variables. The t-test verified that the sample

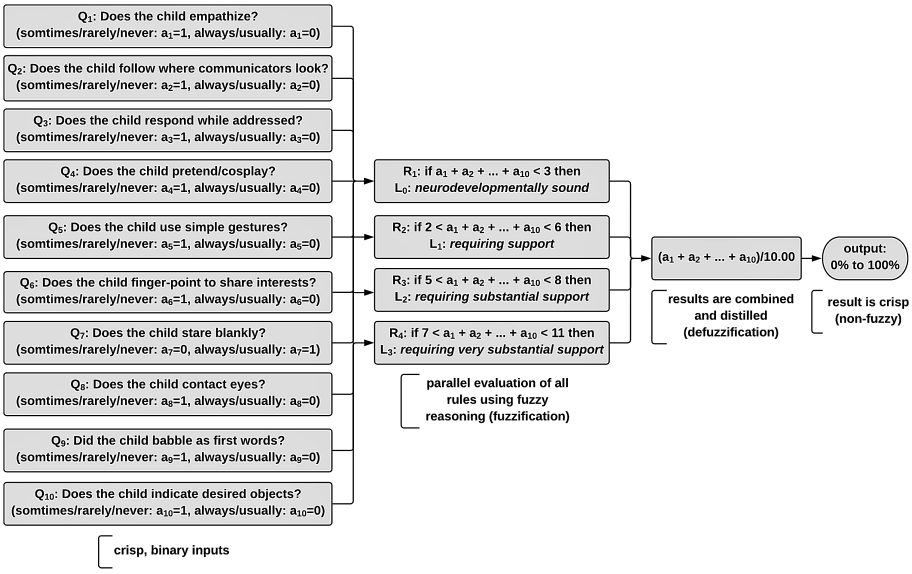


Fig. 9. application of fuzzy logic for the conversion of crisp inputs into a fuzzy input

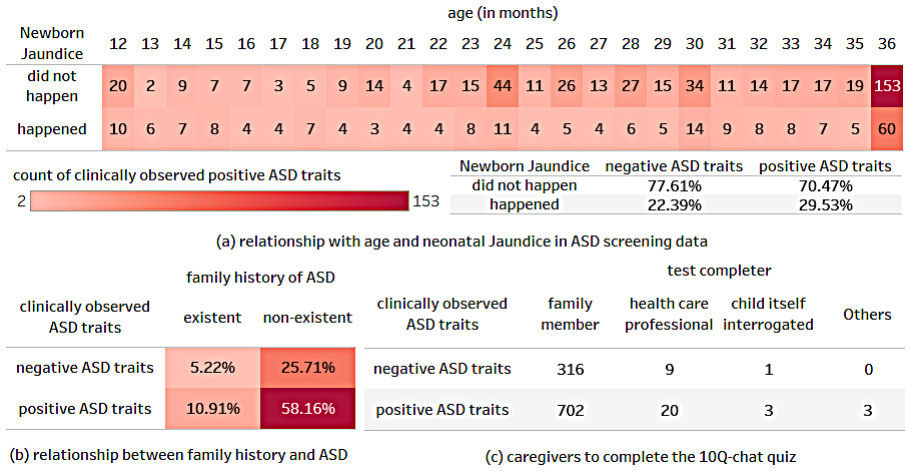


Fig. 10. neonatal jaundice may play a role in ASD, while the roles of family, test-administrator remain unknown

means showed statistical evidence of being considerably segregated while the *Chi*-squared test compared existing and expected frequencies for statistical independence. We recognize an attribute as a feature only upon getting statistically significant ($p\text{-}value < 0.05$) results (Table 1, Table 2).

Preprocessing thus far leaves us with three variants of the ASD toddlers' data: one with all features intact, another with the 10 Q_{10}-chat responses converged to

Table 2. Pearson's χ^2-test results

discrete features	χ^2	degrees of freedom	*p-value*	H_0	H_a
ethnicity, ASD traits	43.571	10	3.93E-06	reject	retain
gender, ASD traits	14.044	1	1.79E-04	reject	retain
Jaundice history, ASD traits	5.427	1	0.01983	reject	retain
test responder, ASD traits	1.4153	3	7.02E-01	retain	reject
family history, ASD traits	0.12094	1	7.28E-01	retain	reject

```r
# setting working directory
setwd("F:/Autism Research")

# reading data into a dummy dataframe
dummy <- read.csv("autism_toddler.csv")

# shuffling data within the dummy frame
dummy <- dummy[sample(1:nrow(dummy)), ]

# separating autistic/mentally sound subjects
dummy_ones <- dummy[dummy$group==1, ]
dummy_zeros <- dummy[dummy$group==0, ]

# preparing test set with 30% data
# 31% of test data
test_zeros <- dummy_zeros[1:98, ]
# 69% of test data
test_ones <- dummy_ones[1:218, ]
test <- rbind(test_zeros, test_ones)

# preparing training set with 70% data
# 31% of training data
train_zeros <- dummy_zeros[99:326, ]
# 69% of training data
train_ones <- dummy_ones[219:728, ]
train <- rbind(train_zeros, train_ones)

# assembling prepareed dataset
dummy2 <- rbind(train, test)
write.csv(dummy2, "autism_toddler_1.csv", row.names=FALSE)
```

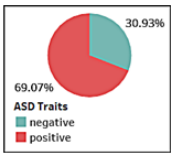


Fig. 11. R-script to partition data, maintaining the original class-distribution (shown in inset) within ASD screening data

a single feature and the other finally excluding statistically insignificant (Table 1, Table 2) features (test administrator, family history). We generate k-fold ($k = 5$) permutations of each variant for getting an unbiased estimate of the metrics. For gradient descent to converge following an unfaltering trajectory, we normalize the input numerals within [0, 1] for a fair comparison. Finally, we split the data fairly into training (70%) and test (30%) sets with each set maintaining the representative distribution (69% autistic + 31% typical) of the two classes (Fig. 11).

3.2 Finding, Fitting, Tuning Predictive Model

We choose to fit a neural network (Fig. 12) that takes as input the numeric representation of both categorical and continuous inputs, propagates them through *ReLU*-activated hidden layers and finally maps them to a *SoftMax* classifica-

tion layer. Hyperparameters, which impact performance greatly alongside the architecture, were tuned in the following way:

- **the number of layers, neurons:** The best-fitting NNs shrunk in architectural complexity as we gradually narrowed down on features as shown in (Fig. 12(a, b, c)).
- **the number of epochs:** For the highest refinement of parameters, the models were trained for 250 epochs.
- **learning rate, α:** A small learning rate of 0.001 was maintained not to overshoot minima.
- **regularization parameter, λ:** To prevent overfitting by penalizing the parameters, this was set to 0.08.
- **size of minibatch:** Minibatch gradient descent was run using 256 training examples at a time so as to not run out of primary memory.

The weights, mapping the neurons hierarchically from one layer to the other, were initialized using *Xavier* initialization assuming the inputs to hail from a *Gaussian* or uniform distribution. The cross-entropy loss, appended by an L_2 regularizer (to prevent overfitting) has been optimized for the classification problem. In the equation below, n, $y^{(n)}$, $\hat{y}^{(n)}$, i, λ, L and w are representative of count of training examples, gold labels for separate examples, model's predicted labels, sequence of a layer's activation, regularization parameter and weights being refined, respectively; with F denoting *Frobenius* norm.

$$-logL(\{y^{(n)}\}, \{\hat{y}^{(n)}\}) = \sum_{n} H(\{y^{(n)}\}, \{\hat{y}^{(n)}\}) + \frac{\lambda}{2n} \sum_{L} ||w^{L}||_{F}^{2} \tag{1}$$

An initial set of parameters θ is refined by optimizing loss $J(\theta)$ through running gradient descent [21] repeatedly for a specified number of epochs or until convergence—parallelly for all features i.e., for $j = 0, 1, ..., n$ where α is the learning rate. For m training examples $(x^{(i)}, y^{(i)})$, where $h_\theta(x^{(i)})$ is the machine-prediction, gradient descent is run like the following:

Repeat until convergence {

$$\theta_j = \theta_j + \alpha \sum_{i=1}^{m} [y^{(i)} - h_\theta(x^{(i)})] x_j^{(i)} \quad (for\ every\ j) \tag{2}$$

}

We apply the gradient-based *Adam* optimizer to optimize gradient descent—harmonizing present parameters with lower-order moments. We select the exponential decays, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a very small number $\epsilon =$ 10E-8 for preventing division by 0.

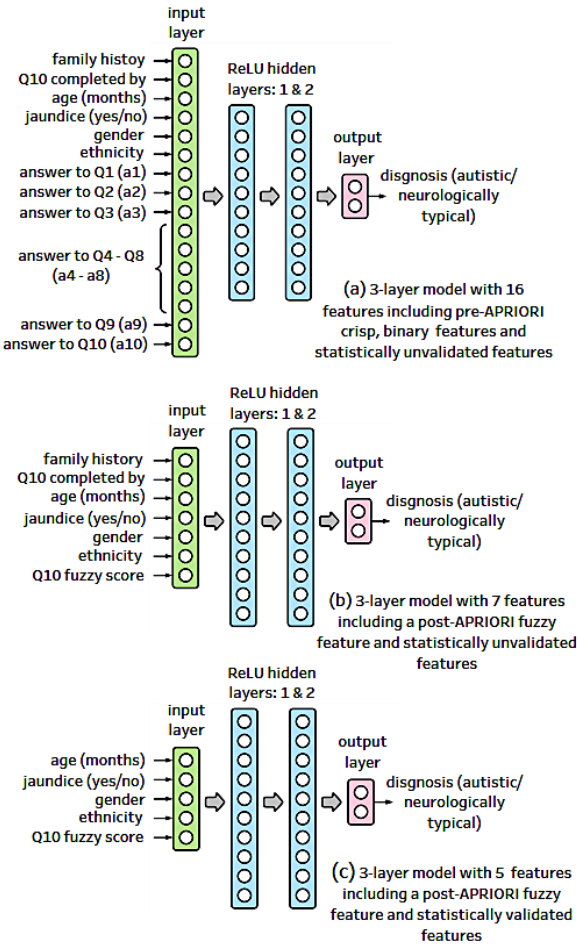


Fig. 12. evolutions in the neural networks employed for the research

3.3 Estimating Performance Metrics

The ML framework of *TensorFlow* uses graph theory's computation graphs (Fig. 13) for defining its sessions with the circular nodes rendering operations and the rectangular ones denoting operators on one-hot representations. The research places the predictions thus computed in a contingency table having both classes along both its dimensions, to evaluate medically important metrics: accuracy, precision (proportion of truly correct autistic identifications), recall (proportion of the actually autistic, classified correctly) and F1-score (harmonic mean of precision and recall). This cross-tabular layout is called a confusion matrix (Fig. 15).

One sample t-test and *ANOVA*-test have been used to strengthen that the performance of the three neural networks was systematically different. We use

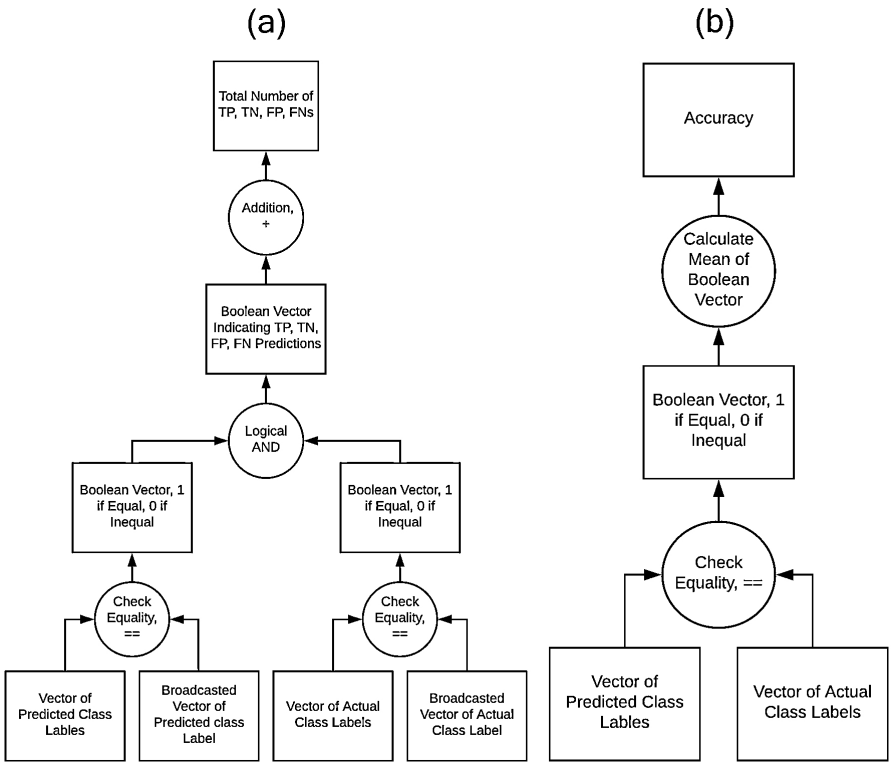


Fig. 13. (a) generalized computation graph for determining entries associated with confusion matrix (b) computation graph for the computation of accuracy

the one-sample t-test in order to compare the mean of the second scheme with k = 5-fold accuracies of the third scheme. We finally applied $ANOVA$ which null hypothesized equality of the average F1-scores of the three schemes, H_0: $\mu_1 = \mu_2 = \mu_3$ (no association exists). To simplify, the $ANOVA/t$-statistic calculated the ratio of the variance between and the variance within the random k = 5-sample groups. The greater the ratio, the more the probability of the alternative hypothesis, H_a (association exists), being justified.

4 Experimental Results and Discussion

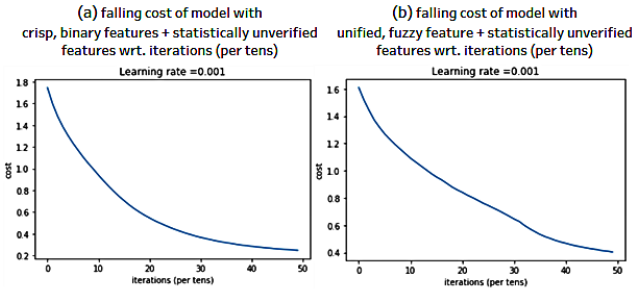
The experimental results produced in this study draw much from the methods followed since the results influenced the methods to devolve to the final, most successful, highly optimized model (Fig. 12(c)). The results initiate a funnel-like pavement which gradually accumulates the fragmented modules to a single, meaningful piece. The methodology, intertwined with the results, show a gradual reduction in the complexity of the neural networks and the number of

features—lowering them from 16 to 7 to a final 5, whilst not compromising the accuracy (Table 4), 99.77% (maximum).

Table 3. the finally extracted features with *p-values* in increasing order, defining their statistical priority

statistical priority	discrete/continuous feature	inferential *p-value*
1	10 questions' quiz-score	$<$ 2.2E-16
2	ethnicity	3.93E-06
3	gender	1.79E-04
4	Jaundice history	0.01983
5	age (months)	4.26E-02

Before delving into the practical results evaluated upon the models, we initially shed light on the results obtained from preprocessing the data. We first applied the *Apriori* algorithm on the transactional forms of the Q_{10}-chat answers and found a high tendency among them of being together (Fig. 8). To recapitulate the methods, observing the ten strong rules reveals this tendency and hence we perform a fuzzy conglomeration of these to a single feature. Next, we shed off features based on t/*Chi*-metrics (Table 1, Table 2) and find the prior features (Table 3).



(c) falling cost of model with
unified, fuzzy feature + statistically verified features wrt. iterations (per tens)

Epoch	k=1	k=2	k=3	k=4	k=5
25	1.540	1.539	1.538	1.540	1.541
50	1.293	1.294	1.288	1.296	1.296
75	1.126	1.124	1.119	1.126	1.120
100	0.928	0.924	0.916	0.929	0.920
125	0.683	0.687	0.673	0.684	0.676
150	0.557	0.561	0.543	0.560	0.555
175	0.495	0.498	0.481	0.498	0.494
200	0.454	0.459	0.441	0.458	0.453
225	0.426	0.430	0.412	0.430	0.424
250	0.404	0.407	0.390	0.408	0.403

training error (cross-entropy loss)
0.390 1.541

Fig. 14. learning curves learned upon training the different schemes, with the loss plotted once per ten epochs

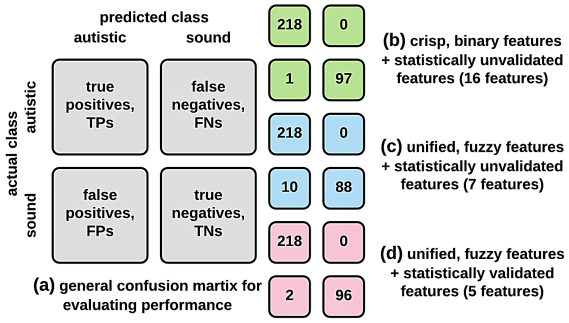


Fig. 15. confusion matrices filled against k = 1-st cross-validation set for each of the schemes

First, we attempt to show that we conducted a sound training of the ML models. With a minuscule learning rate of 0.001, the learning curves obtained showed a gradual reduction in the loss function with each of 250 epochs (Fig. 14(a, b)). For the final model, we choose to show each of $k = 5$ cross-validations converging to a minimal error with the warmer shades showing higher errors that gradually cooled down to errors as low as 0.390 (Fig. 14(c)). The training led us to machine-generated predictions that filled out the confusion matrices (shown for $k = 1$ in Fig. 15).

Table 4. raw results and metrics delivered by the models for $k = 5$ validation-sets each

prediction scheme	k-fold	optimized training loss	test accuracy	precision	recall	F1-score
3-layer NN with 16 features	-	0.259586	0.9968355	0.9954338	1	0.9977117
3-layer NN with 7 features	1	0.426662	0.9683544	0.95614034	1	0.97757847
3-layer NN with 7 features	2	0.417105	0.9778481	0.9688889	1	0.984198651
3-layer NN with 7 features	3	0.432018	0.9746835	0.96460176	1	0.981981977
3-layer NN with 7 features	4	0.404772	0.9746835	0.96460176	1	0.981981977
3-layer NN with 7 features	5	0.529996	0.9778481	0.9688889	1	0.984198651
3-layer NN with 5 features	1	0.403833	0.99050635	0.98642534	1	0.993166287
3-layer NN with 5 features	2	0.40695	0.9936709	0.9909091	1	0.995433795
3-layer NN with 5 features	3	0.389566	0.9968355	0.9954338	1	0.997711676
3-layer NN with 5 features	4	0.40799	0.99050635	0.98642534	1	0.993166287
3-layer NN with 5 features	5	0.402822	0.9936709	0.9909091	1	0.995433795

For each optimized model, we evaluate k-fold cross-validated results with k = 5 (Table 4) to make sure that the models are showing consistent performance. The results were obtained upon using an identical set of hyperparameters and showed brilliance in the test accuracies given that above 95% is generally held as an excellent performance by any ML algorithm. We perform a comparative analysis among the maximum of the metrics evaluated using each model and find that the results of the final 5-feature model showed equal promise as the 16-feature model, albeit with less computation (Fig. 16).

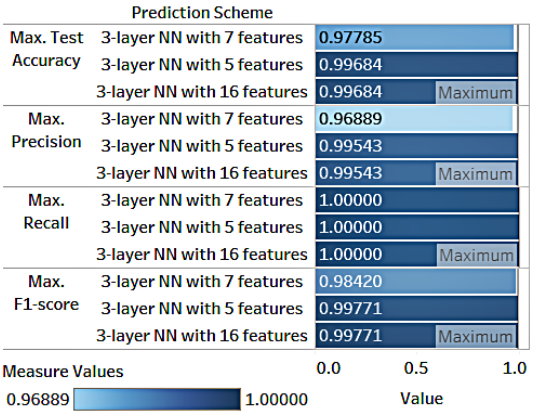


Fig. 16. comparison among the proposed schemes' results

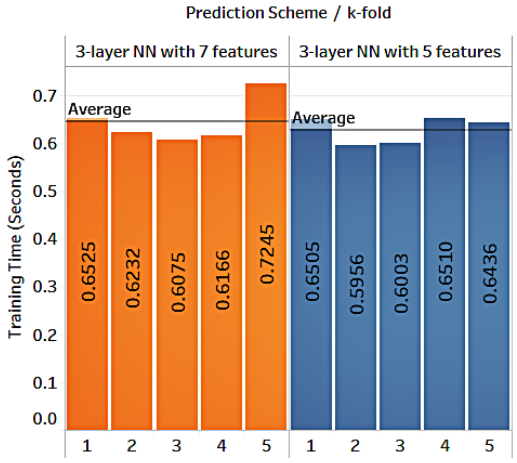


Fig. 17. training time in the final model is reduced

After having tabulated the results we demonstrate that the third model, i.e. the second optimized scheme with 5 features took less time to be trained than the first optimized model, i.e. the model with 7 features (Fig. 17). We show the comparison employing all the $k = 5$ validations and find a decrement in the average training time in the model with 5 features. The intuition behind such results is that a model using fewer features will require less time for backpropagation since there will be fewer features to calculate the contribution in the total error.

We finally apply statistical inference-tests to make sure that the supremacy of one model over the other actually holds water. We apply one-sample t-test on all $k = 5$ cross-validated accuracies of the second optimized scheme (Fig. 12(c)) against the average accuracy of the first optimized scheme (Fig. 12(b)) and find t-statistic $= 15.501$ with a $p\text{-}value = 0.0001011$ (Table 5), repudiating the null

Table 5. t-test results verifying better results yielded by the second optimized model over the first

one sample t-test paramters/metrics	evaluations
comparison of avg. accuracy (model using 5 features) with	k-fold accuracies of model using 7 features
t-statistic	15.501
degrees of freedom	4
p-value	0.0001011
95% confidence interval	0.9897505 to 0.9963255
sample mean	0.993038
alternative hypothesis, H_a: true mean is not equal to 0.9746835	true

Table 6. *ANOVA*-test verifying systematic difference in the models' performances

ANOVA (Analysis of Variance) test metrics	values
degrees of freedom for numerator (ind)	2
degrees of freedom for denominator (residuals)	12
sum of squares of numerators (ind)	0.0007059
sum of squares of denominators (residuals)	0.0000437
mean of squares of numerators (ind)	3.53E-04
mean of squares of denominators (residuals)	3.60E-06
analysed value	96.98
p-value, $Pr(> F)$	3.91E-08

hypothesis and proving the alternative hypothesis that the 5-feature model works significantly and consistently better. We apply a statistical *ANOVA* (Table 6) on the F1-scores of the three models and find: $F(2, 12) = 96.98$, *p-value* $= 3.91E\text{-}08 \ll 0.05$, leading us to accept the safe conclusion that the models are indeed performing with measures separating them from each other. We finally visualize our efforts in comparison with reviewed literature in terms of accuracy and find equivalent or in some cases, superior results (Fig. 18). The point of the discussion is that by initially combining some features into a fuzzy one, we get a tremendous accuracy and by further pruning of insignificant features, we manage to get even better performance, equivalent to what we had previously achieved by retaining all the features—obviously with less computation.

5 Conclusion

The problem of prognosticating autism has so far been a lengthy medical process that this study strives to overcome by taking mostly social and interactional elements into account. The novelty of this study has been the application of business intelligence procedures into a neuro-psychological problem, which was obtained through the following milestones:

- The study uses the *Apriori* algorithm not for producing the final output or merely for generating rules, the algorithm has rather obtained endorsement as a preprocessing step for creating better, fewer representative features.
- The paper has shown the effectiveness of the *Apriori*-based clubbing by converting the club into a fuzzy feature, the membership values of which constituted the foremost important feature.

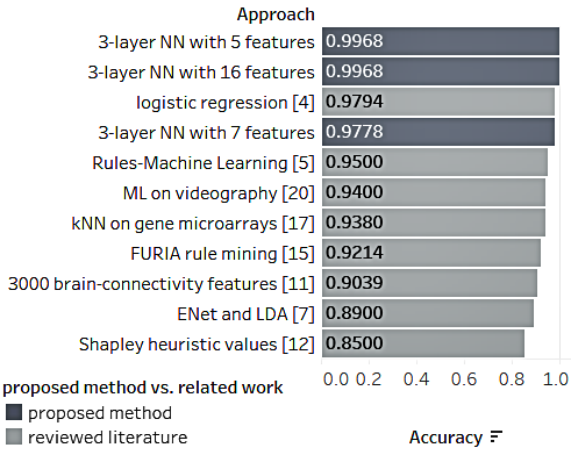


Fig. 18. comparison among proposed and reviewed methods for the detection of autism in toddlers

- The research has reckoned the impact of features by using statistical inference and has assigned priorities to them using no machine learning method, hence delineating the effectiveness of statistical preprocessing.
- The endeavor has shown that hard computing (statistical preprocessing, association rule mining), if applied prior to soft computing (neural networks), can significantly reduce the load on soft computing while maintaining the same scintillating performance.

The research-work presented adds value to not only computer science or data science, but also attempts to help the psychological or neurological community to investigate the aspects of autism at a young age. The work has future prospects of being extended over varieties of age-groups and the limitation of a class-imbalance can be overcome from a data-scientific front. The study lends a message to the research community to explore novel preprocessing techniques instead of making their first leap to machine learning.

References

1. https://www.verywellhealth.com/what-are-the-three-levels-of-autism-260233 (Accessed on 6th March, 2020)
2. Lyall K, Croen L, Daniels J, Fallin MD, Ladd-Acosta C, Lee BK, Park BY, Snyder NW, Schendel D, Volk H, Windham GC. The changing epidemiology of autism spectrum disorders. Annual review of public health. 2017 Mar 20;38:81-102.
3. Dover CJ, Le Couteur A. How to diagnose autism. Archives of disease in childhood. 2007 Jun 1;92(6):540-5.
4. Thabtah F. An accessible and efficient autism screening method for behavioural data and predictive analyses. Health informatics journal. 2019 Dec;25(4):1739-55.

5. Thabtah F, Peebles D. A new machine learning model based on induction of rules for autism detection. Health informatics journal. 2019 Jan 29:1460458218824711.
6. Sarkar A, Wade J, Swanson A, Weitlauf A, Warren Z, Sarkar N. A data-driven mobile application for efficient, engaging, and accurate screening of ASD in toddlers. InInternational Conference on Universal Access in Human-Computer Interaction 2018 Jul 15 (pp. 560-570). Springer, Cham.
7. Duda M, Haber N, Daniels J, Wall DP. Crowdsourced validation of a machine-learning classification system for autism and ADHD. Translational psychiatry. 2017 May;7(5):e1133-.
8. Achenie LE, Scarpa A, Factor RS, Wang T, Robins DL, McCrickard DS. A Machine Learning Strategy for Autism Screening in Toddlers. Journal of Developmental & Behavioral Pediatrics. 2019 Jun 1;40(5):369-76.
9. Thabtah F, Kamalov F, Rajab K. A new computational intelligence approach to detect autistic features for autism screening. International journal of medical informatics. 2018 Sep 1;117:112-24.
10. Abbas H, Garberson F, Glover E, Wall DP. Machine learning approach for early detection of autism by combining questionnaire and home video screening. Journal of the American Medical Informatics Association. 2018 Aug;25(8):1000-7.
11. Kong Y, Gao J, Xu Y, Pan Y, Wang J, Liu J. Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. Neurocomputing. 2019 Jan 9;324:63-8.
12. Tariq Q, Fleming SL, Schwartz JN, Dunlap K, Corbin C, Washington P, Kalantarian H, Khan NZ, Darmstadt GL, Wall DP. Detecting developmental delay and autism through machine learning models using home videos of bangladeshi children: Development and validation study. Journal of medical Internet research. 2019;21(4):e13822.
13. Al Farsi A, Doctor F, Petrovic D, Chandran S, Karyotis C. Interval valued data enhanced fuzzy cognitive maps: Torwards an appraoch for Autism deduction in Toddlers. In2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) 2017 Jul 9 (pp. 1-6). IEEE.
14. Al-Diabat M. Fuzzy data mining for autism classification of children. Int J Adv Comput Sci Appl. 2018 Jul 1;9(7):11-7.
15. Khan S, Alshara M. Fuzzy Data Mining Utilization to Classify Kids with Autism. IJCSNS. 2019 Feb;19(2):147.
16. Xiao X, Fang H, Wu J, Xiao C, Xiao T, Qian L, Liang F, Xiao Z, Chu KK, Ke X. Diagnostic model generated by MRI-derived brain features in toddlers with autism spectrum disorder. Autism Research. 2017 Apr;10(4):620-30.
17. Kim SH, Kim IB, Oh DH, Ahn DH. Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning. European Neuropsychopharmacology. 2017 Oct 1;27:S1090.
18. Karan P, Pirouz M. EEG Analysis for Predicting Early Autism Spectrum Disorder Traits.
19. Abbas H, Garberson F, Glover E, Wall DP. Machine learning for early detection of autism (and other conditions) using a parental questionnaire and home video screening. In2017 IEEE International Conference on Big Data (Big Data) 2017 Dec 11 (pp. 3558-3561). IEEE.
20. Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, Wall DP. Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS medicine. 2018 Nov;15(11).
21. Ng A. CS229 Lecture notes. CS229 Lecture notes. 2000;1(1):1-3.