Improvised Data Preprocessing using Missing value handling and Deduplication using artificial bee colony optimization for Autism Dataset

Suresh Kumar R[1], Dr.Renugadevi M[2]
[1]*Assistant Professor,Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi.*
[2]*Professor & Head Department of BCA, Sri Krishna Arts and Science College, Coimbatore*
(E-mail: sureshdhanya2002@yahoo.com,renuga.srk@gmail.com)

***Abstract*—** Autistic Spectrum Disorder (ASD) is a neuro developmental condition associated with significant healthcare costs, and early diagnosis can significantly reduce these issues. Unfortunately, waiting periods for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, to help health professionals and inform individuals whether they should pursue formal clinical diagnosis, a time-efficient and accessible ASD screening is imminent. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behavior traits. The quality of autism dataset is an essential characteristic of prediction process. Presence of missing value and duplication in dataset will leads to inaccurate results. Thus, this paper aims to improvise the quality of the autism dataset by handling missing value using machine learning approach which is known as enhanced boosted K-NN and eliminating duplicate record using behavioral approach of artificial bee colony optimization. The simulation results proved that the proposed work outperforms the existing traditional approaches.

***Keywords*—** *Autism, machine learning, missing value, boosted k-NN artificial bee colony, Prediction*

I. INTRODUCTION

Among various development disorders in children, Autism is the most common brain disorder. According to the report of Indian scale assessment of Autism, nearly 2 million children with Autism in India [1]. It is generally defined as a complex development disability that designs showing sign in the first three years of life. autism this a neurodevelopment disorders the samples normal brain function. with adversely Impact the communication skills of child as well as his or her ability to interact with people. The autistic children urgently characterized by lack of communication behavior and social interaction. All the symptoms of Autism spectrum disorders typically image early in life career level diagnosis is usually not achieved before the age of 3. Evidence [1] reported that the best prognosis for ASD presently lies in early diagnosis to improve the outcome by modifying emergent atypical development trajectories.

In every children and adults, the signs and symptoms of the autism spectrum disorders embrace issues with social interaction skills, speech and communication. The autism spectrum disorders are measured based on the presence of multiple symptoms that disrupt the child's ability to talk, make the relationships, explore, play, and to study. The method, an individual communicates and relates to people. The symptoms of autism spectrum disorders: Social skills. Basic social interaction may be troublesome for children with autism spectrum disorders.

Symptoms might include:

1. Unusual or inappropriate visual communication, gestures, and facial expressions (e.g. avoiding eye contact or facial expressions that don't match what he or she is saying).
2. Lack of interest in people or in sharing interests or achievements (e.g. showing you a drawing, pointing to a bird).
3. Unlikely to approach others or to pursue social interaction; comes across as aloof and reserved; prefers to be alone.
4. Problem and difficulty in understanding individual person's feelings, reactions, and nonverbal cues.
5. Resistance to being touched.
6. Difficulty or failure to create friends with children the same age.

II. RELATED WORK

Stahl et al. [2] analyzed event-related potential data of eye gaze performance to discriminate infant groups at high- or low-risk for later diagnosis of autism. Computational methods using regularized discriminant functions analysis (DFA), linear discrimination analysis (LDA) and SVM were explored. The accuracy was around 64% for SVM, 61% for regularized DFA, and 56% for LDA. They highlighted the need for finding the most discriminative features and eliminating irrelevant variables prior to classification in order to increase predictive accuracy.

In Wilson et al. 2014 [3], twelve cognitive scores obtained from two subcomponents of an IQ test and ten

neuropsychological experimental tasks were used to study how cognitive abilities could be useful in characterizing individuals with ASC. Significant differences between 89 participants of Autism Spectrum Condition (ASC) and 89 participants of male controls were found on five cognitive tasks using statistical t-tests. Interestingly, one of these tasks was motor performance. Although it was recognized that not every variable showed a remarkable impact on each group, a SVM model was still trained using all variables so as to achieve the overall identification accuracy of 81%. It was recognized that quantitative methods to determine the most significant variables could improve the accuracy by removing outlier factors.

Applied a machine learning classifier using data from the Autism Diagnostic Interview-Revised and The Social Responsiveness Scale, Bone et al. [4], to a large group of verbal individuals with ASC and those with non-ASC disorders. A novel multiple-level SVM model was proposed, and promising classification results were reported. This work proved that for creating robust algorithms to improve ASD screening and diagnostic methods the ML was useful.

Grossi et al. reported that to build up a predictive model based on 27 potential pregnancy risk factors in autism development [5] by the use of specialized artificial neural networks (ANNs). The artificial neural network approach selecting 16 out of 27 variables and achieved 80% global accuracy. Their work demonstrates that ANNs are able to build up a predictive model, which could represent the basis for a diagnostic screening tool.

To date, there have been some studies that have applied machine learning to autistic motor data. Crippa et al. [6] reported a proof-of-concept study to explore that using kinematic analysis of a simple motor task, whether low-functioning children with ASC could be identified. Using a motion tracker, fifteen ASC and 15 non-autistic children were asked to pick up a ball and drop it into a hole while their movements were recorded. Seventeen kinematic parameters were extracted from the upper-limb movement and seven of these were found significant for discrimination by using Fisher discrimination ratio-based technology. A maximum classification accuracy of 96.7% using SVM was reported.

Yang [7] in his work due to the inspiration of behavioral nature of Bats in case of hunting its prey and finding its destination in a problem space like many external barriers he formulated the artificial bat algorithm for optimistic solution space.

Yılmaz et al [8] in their work used proben to determine the duplicate among them by introducing a balanced bat algorithm which satisfies the required criteria.

Kaveh and Zakian [9] in their work done optimization in size of skeletal structures was done using the proposal of enhance bat behavior inspired algorithm. Which produce a potential result on the goal. A trust structure was framed to improve the conventional bat algorithm to enhance its efficiency.

Anibal et al [10] investigated about 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. Achieving 70% accuracy in identification of ASD versus control patients in the dataset, this results improved the state-of-the-art.

Bhaskar Sen et al [11] in their work using structural texture and functional connectivity features obtained from 3-dimensional structural magnetic resonance imaging (MRI) and 4-dimensional resting-state functional magnetic resonance imaging (fMRI) scans of subjects, presented a novel method for learning a model that can diagnose Attention Deficit Hyperactivity Disorder (ADHD), as well as Autism,.

Muhammad et al [12] aims based on ML technique to propose an effective prediction model and to develop a mobile application for predicting ASD for people of any age. As outcomes of this research, an autism prediction model was developed by merging Random Forest-CART (Classification and Regression Trees) and Random Forest-ID3 (Iterative Dichotomiser 3) and also a mobile application was developed based on prediction model.

Koyamada et al. (2015) [13] investigated brain states from measurable brain activities by using Deep Neural Networks (DNN). A task-based fMRI data from 499 subjects classify into seven categories related to the tasks: Emotion, Gambling, Language, Motor, Relational, Social and Working Memory, they trained an artificial neural network with two hidden layers and a softmax output layer.

Plis et al. (2014) [14] using data from four different sites, used deep learning and structural T1-weighted images in order to classify patients with schizophrenia versus matched healthy controls. The authors concluded that deep learning holds great potential for clinical brain imaging applications.

III. PROBLEM STATEMENT

It is difficult to handle huge volume of data due to its growing nature and thus leads to the response time to be slow, accessibility, refuge, and excellence assurance are not under satisfying factor. Now days the status of the organization primarily depends on the quality of data handled by its system and providing reliable service to the users. Under these circumstances, the selection of maintaining data repositories with "ugly" data such as the presence of duplicated or not under proper format it ultimately creates downfall in whole speed and degradation in system performance.

The work of deduplication technique is to detect the presence of duplicates in the data warehouse or repositories due to typing mistake, spelling mistake and style of writing. The difficulty of identifying and eradicating duplicate occurrence in record sets is one of the toughest troubles in the huge volume area of data cleansing and quality of information maintenance in repositories. So, this paper focuses on creating an algorithm that can detect and eliminate maximum duplications and handling missing values.

A. Materials and Methods

The original dataset was cleaned from all kind of inconsistencies. All tuples with same conditional attributes and different classification attribute were removed. This will

clearly improve efficiency as it will remove all suspected cases. The experiments were then designed to check the best model to fill in the missing values that generates the highest coverage of the data set.

B. Mean Imputation

Missing values are replaced with estimates derived from applying statistics methods to available data. Impute Missing Values allows the estimation of missing data for all or selected attributes. The Substitution with a measure of central Tendency which is commonly known as mean substitution is one of the most frequently used method. It is replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs. Let us consider that the value Zij of the t-th class, Ct, is missing then it will be replaced by

$$\overline{z}_{ij} = \sum_{i:z_{ij} \in C_t} \frac{z_{ij}}{n_t}$$

Algorithm:
Input: Data set, DS.
Output: Data set, DS, contains instances with no missing values.
Method:
for each selected Attribute Att in DS
Calculate the Mean Value (MV) of Att
{end for}
for each selected Attribute Att in DS
for each case C of Att
if the value of Att is null
fill the value of Att as MV
{endif }
{endfor}
{endfor}

C. Crisp K-NN

In the crisp nearest neighbor-based imputation an attribute Att with missing value is imputed by finding its k nearest neighbor and assigning its value to the attribute Att [15].
Algorithm
Input :
Split the input Dataset DS into two:
DSm – dataset containing the instances in which at least one of the attribute value is missing
DS c – dataset containing complete attribute information
Output: Dataset DS containing no missing values
Method
For each vector x in D_m:
Divide the instance vector into observed and missing parts as x = [x_o; x_m].
Calculate the distance between the xo and all the instance vectors from the set D_c.
Use only those features in the instance vectors from the complete set D_c,
which are observed in the vector x.
Use the k closest instances vectors (k-nearest neighbors) and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes replace the missing value using the mean value of the attribute in the k nearest neighborhood.

IV. PROPOSED WORK - IMPROVISED AUTISM DATASET USING BOOSTED K-NN AND RECORD DEDUPLICATION USING ARTIFICIAL BEE COLONY OPTIMIZATION

A. Improvised Autism Dataset using Boosted k-NN

The main drawback of the conventional k-nearest neighbor is that it looks for the most similar instances, the algorithm searches through all the data set. This limitation can be very critical for KDD, since this research area has, one of its main objectives, the analysis of large databases. This proposed work adopted the usage of Enhanced Boosted k-NN for handling missing values instead of classification as shown in the figure 1. The proposed method reduces the search space of the k-NN by just finding the similar instances of same class which it belongs. Based on the membership on each classes the instances which belong to the highest membership value of the missing instances is alone considered for the similarity measurement and among them k nearest neighbor is selected and depending the type of the attribute whether they belong to the discrete or continuous the values are assigned.

In traditional k-NN as shown in the figure 2 there are two main issues they are dependent and sample category balance. The dependent value k has to be preset. Sample category balance denotes that when the sample numbers belonging to different categories have large gap, so the classification results inclined to inaccurate.

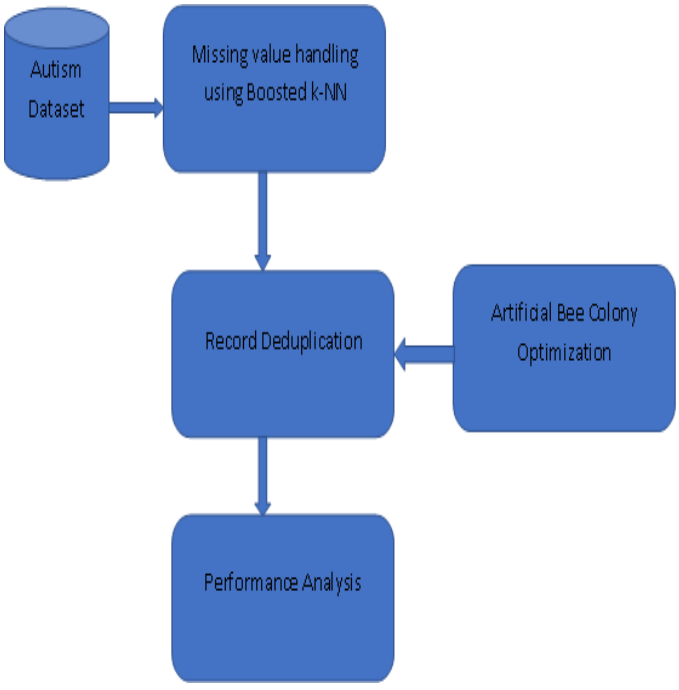


Figure 1: Overall Framework of the proposed model

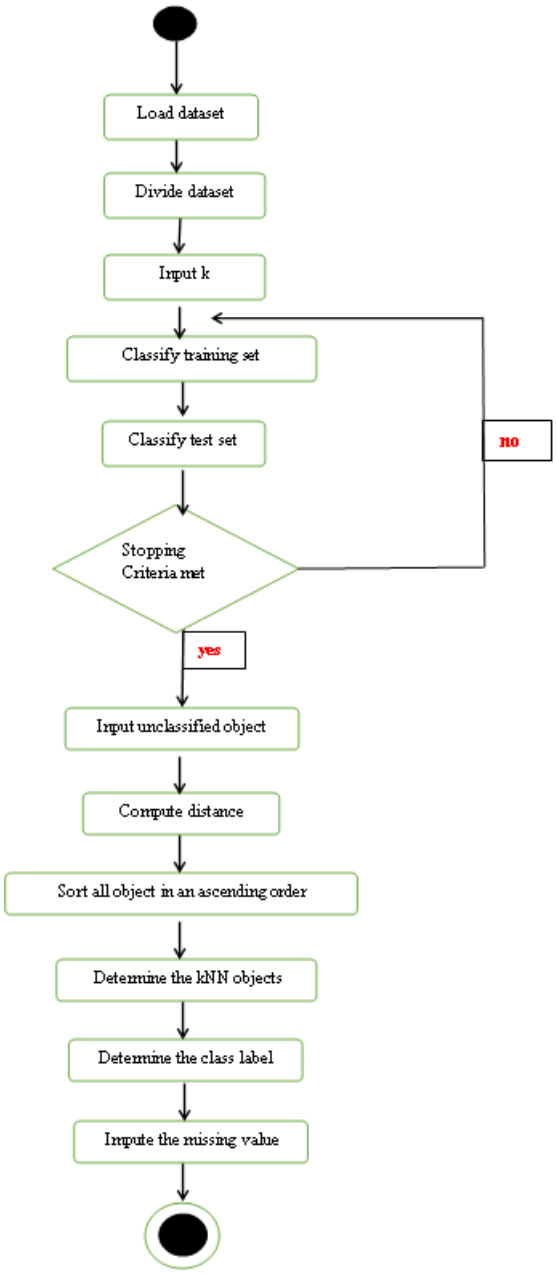


Figure 2: Work flow of kNN in missing value handling

Algorithm 1: Steps for enhanced boosted k-NN

Input: Autism Dataset

$S = \{S_i\} = \{wt_i^0, x_{i,}, y_i)\}, T, \gamma$

$wt_i^0 = 0$, I = 1…n

$S_0 = S$

For t = 1 to T

$S_0 = S_{t-1}$

For $S_q \in S_t$

N_q = kNN of S_q using Dist(S_q, S_i)

Label(S_q) = argmax$_{y \in Y} \Sigma S_q \in N_q$ Dist(S_q, S_i)τ(y,y_i)

If label(S_q) ≠ y_q then

For $S_i \in N_i$

If $y_i \neq y$ then

$wt_i^t = wt_i^t$ - Dist(x_q, x_i)

$wt_i^t = wt_i^t$ + Dist(x_q, x_i)

if label(s_q) = yq, ∀sq then

break

Let S be a training set with n instances, and the i^{th} instance S_i is described by (wt_i^0,x_i,y_i) where wt_i^0 is a weight term initialized to zero, vector x_i is a in the feature space X, and class label y_i is in the label set Y. dist(x_1,x_2) is defined as the Euclidean distance between two instances S_1 and S_2 ($\|x_1 - x_2\|^2$). The distance between the query instance S_q and the i^{th} instance is then defined as the function Dist(S_q, S_i) where

$$Dist(s_q, s_i) = \frac{1}{\left(1 + e^{-wt_i^0}\right) dist(x_q, x_i)}$$

The distance function Dist(s_q,s_i) is designed to be the product of a sigmoid function 1/ $\left(1 + e^{-wt_i^0}\right)$ and the traditional distance function 1/Dist(x_q,x_i). When the weight term is set to the initial value of 0, the value of the distance function Dist(S_q, S_i) is half that of the traditional distance function 1/dist(x_q,x_i). Modifying the weight term then change the value of the sigmoid function between 0 and 1 before it is multiplied with the traditional distance function. Constraining the weight term with a sigmoid function works well to prevent weights from modifying the distance function Dist(Sq, Si) too drastically and/or too quickly.

The pseudo code for the Boosted k-NN algorithm is shown in Algorithm 1. Given the training set S, number of iterations T, and weight update term γ, enhanced boosted k-NN constructs an ensemble of up to T k-NN classifiers with modified weight terms. During each iteration t, a k-NN classifier is constructed by iterating through the weighted training set querying each instance against the rest of the training set. When an instance is misclassified, the weights of its k-nearest neighbors will be modified as follows.

For each neighbor instance that belongs to a different class than the query instance, its weight term for the next iteration will be decreased by γ /dist(x_q,x_i), where dist(x_q,x_i) is the Euclidean distance between the query instance and the nearest neighbor being modified.

On the other hand, a neighbor that belongs to the same class as the query instance will have its weight for the next iteration increased by γ /dist(x_q,x_i). The modified weight term affects the distance function Dist(S_q, S_i) by increasing the distance of neighboring opposite-class instances and decreasing the distance of neighboring same-class instances. The label of the query instance is the class that has the highest weighted sum of votes among its k nearest neighbors based on the distance function Dist(S_q, S_i); therefore, modifying the weight terms in this way improves the chances of misclassified instances being correctly labeled the next iteration.

B. Deduplication of Autism Dataset using Artificial Bee colony Optimization

In the Artificial Bee Colony (ABC)Algorithm the possible solution of optimization is represented as the position of food resource problem [16, 17]. The amount of a food source is correlated to the quality of the solution associated with it. The amount of onlooker bees and the employed bees are equal to the possible number of solutions to the given problem.

In basic ABC there are three important control parameters they are amount of food sources which is equal to the number of onlooker bees or employed bees (SN), the limit value and the number of maximum cycle (MCN). The recruitment rate of honeybees denotes how soon the bee colony determines and reveals a new discovered food source. Depending on the efficiency of the fast discover and utilization of the best resource the progress and the survival rate of bee colony will be reached. Proper case study of the existing problem and the way of deploying the proposed method in an effective way is very important.

In the ABC algorithm, though onlookers and employed bees perform the process of discovery in search space the scout bees perform the process of exploration. Comprehensive pseudo-code of the ABC algorithm is given below [21].

Procedure for Artificial Bee colony Optimization Algorithm
Step 1: Set the round =1
Step 2: Set ABC variables parameters
Step 3: Estimate the fitness of all instances
Step 4: Repeat
Step 5: Build solutions using bees designated as employees
Allocate feature subset patterns which is represented using binary bit string to all employed bee
Create fresh feature subsets FV_i
Input the resultant feature subset to the concern classifier
Calculate the fitness ($fitness_i$) of the feature subset of instances
$fitness_i$= 1/(1+$finess_i$)
Compute the probability $prob_i$ of solution of feature subset
$Prob_i$ = $fitness_i$/($\sum_{i=1}^{m} fitness_i$)
Step 6: Build solution space using the onlookers
Choose a feature related to probability $prob_i$
Create the new solutions Si for the onlooker bees from the existing solutions xi selected depending on $prob_i$ and evaluate them
Perform the process of greedy selection for the onlooker bees
Step 7:Define the unrestricted solution for the scout, if survives, and substitute it with a new arbitrarily created solution xi by
$X_i^j = X_{min}^j$ + rand [0,1] ($X_{max}^j - X_{min}^j$)
Step 8: Remember the finest solution attained so far
Step 9: round = round + 1
Step 10: until round = MCN

Proposed Artificial Bee Colony Optimization on Duplicate Record Detection

Step 1: Similarity Computation for all Records pair

Similarity functions assign a similarity value for each field by compute the similarity of each field with other record field. The similarity metrics used in the proposed work are Levenshtein distance and cosine similarity.

Levenshtein Distance: The chosen name fields of the records are “Record 1” and “Record 2”. The "Levenshtein distance" is computed by calculating the minimum number of operations that has to be made to transform one string to the other, usually these operations are: replace, insert or deletion of a

character. The levenshtein distance between the records is finding out by considering the record as a whole.

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Cosine Similarity: The cosine similarity between the two records name field "Record 1" and "Record 2" are calculated as follows: First, the dimension of both strings are obtained by taking the union of two string elements in the record 1 and "record 2" as ($word_1$, $word_2$, …….$word_N$) and then the frequency of occurrence vectors of the two elements are calculated i.e. "Record 1" = (, < vector value1>,< vector value2> ……<>) and "Record 2"= (< vector value1>,< vector value2>,……<>). Finally, we obtain the dot Product and magnitude of both strings.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i-1}^{n} A_i^2}\sqrt{\sum_{i-1}^{n} B_i^2}}$$

Step 2: Generate List of Evidence

In this approach each pair of evidence that represents the use of a specific similarity function over the values of a specific attribute found in the data being analyzed [18]. For example, if we want to deduplicate a database table with the attributes (eg Att_1, Att_2 and Att_3) using a specific similarity function (e.g., the Levenshtein function), generate the following list of evidence: a <att_1, Levenshtein >, b<att_2, Levenshtein > , c<att_3, Levenshtein > . Develop a set of expressions by using these evidences with the simple mathematical functions (+,*,-,/,)

Step 3: Optimized Expression of Fitness Evaluation

The fitness value is a value one of the most important components in this process which is generated from the fitness function. If the fitness function is badly chosen then it will surely fail to find the best expression. This approach, used F1 metric as the fitness function and can be calculated as Likewise find the fitness value for each expression in the population based on threshold value. It is necessary to choose an optimized threshold to classify the dataset as duplicates and non-duplicates accurately since F1-value varies with different threshold. Hence, intelligence swarm algorithms named ABC is to discover global optimal solution.

Step 4: Optimal threshold using ABC

In ABC, a population starts with a random set of thresholds (particle) on each record. This set is considered as an employed bee. Find the fitness for all bees. This set passed on to onlooker-bee and scout bee. Thus, ABC selects the best threshold value on each such expression which classify the set of records as duplicate and non-duplicate.

$$\text{Precision} = \frac{\text{No of true duplicates identified}}{\text{Total no of duplicate records identified in autism datas}}$$

$$\text{Recall} = \frac{\text{No of true duplicates identified}}{\text{Total no of duplicate records present in autism data}}$$

$$\text{F-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

V. EXPERIMENTAL RESULT

The autism dataset data preprocessing is done using MATLAB software. The dataset for Autism in toddlers is obtained for Kaggle repository [19]. **Number of Instances (records in your data set):** 1054. **Number of Attributes are**18 including the class variable

TABLE I. DATASET DESCRIPTIONS

Feature	Type	Description
A1: Question 1 Answer	Binary (0, 1)	Does your child look at you when you call his/her name?
A2: Question 2 Answer	Binary (0, 1)	How easy is it for you to get eye contact with your child?
A3: Question 3 Answer	Binary (0, 1)	Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)
A4: Question 4 Answer	Binary (0, 1)	Does your child point to share interest with you? (e.g. poin9ng at an interes9ng sight)

A5: Question 5 Answer	Binary (0, 1)	Does your child pretend? (e.g. care for dolls, talk on a toy phone)
A6: A6: Question 6 Answer	Binary (0, 1)	Does your child follow where you're looking?
A7: Question 7 Answer	Binary (0, 1)	If you or someone else in the family is visibly upset, does your child show signs of wan9ng to comfort them? (e.g. stroking hair, hugging them)
A8: Question 8 Answer	Binary (0, 1)	Would you describe your child's first words as:
A9: Question 9 Answer	Binary (0, 1)	Does your child use simple gestures? (e.g. wave goodbye)
A:10 Question 10 Answer	Binary (0, 1)	Does your child stare at nothing with no apparent purpose?
Age	Number	Toddlers (months)
Score by Q-chat-10	Number	1-10 (Less that or equal 3 no ASD traits; > 3 ASD traits
Sex	Character	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 ASD history	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.
Why_are_you_taken_the_screening	String	Use input textbox
Class variable	String	ASD traits or No ASD traits (automatically assigned by the ASDTests app). (Yes / No)

Precision: It is a fraction of correct instances among those that the algorithm accepts as true belonging to the relevant class.

Precision $(CL_r, CS_i) = N_{ri}/N_i$

Where class CL_r, whose size N_r , classifier CS_i of its size is N_i, N_{ri} data instance in CS_i from the class CL_r

Recall: It is calculated as the fraction of actual instances that were identified.

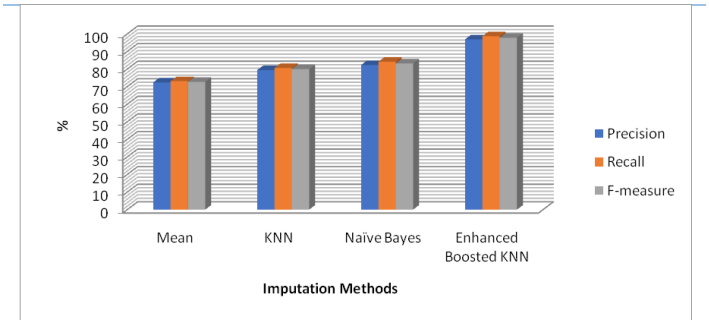
Recall$(CL_r, CS_i) = N_{ri}/N_r$

F-Measure: It is considered as the harmonic mean of both precision and recall and it tries to produce a good combination of these two measures

$$F(CL_r, CS_i) = \frac{2*\text{Recall}(CL_r,CS_i)*\text{Precisi}}{\text{Recall}(CL_r,CS_i)+\text{Precisi}}$$

TABLE II. PERFORMANCE COMPARISON BASED ON PRECISION, RECALL AND F-MEASURE

Imputation Approaches	Precision	Recall	F-measure
Mean	72.4	73.1	72.7
KNN	79.4	80.6	80.0
Naïve Bayes	82.3	84.2	83.2
Enhanced Boosted KNN	96.9	98.7	97.8



From the table 2 and figure 3 it is proved that the performance of enhanced boosted k-NN produces more result because the weakness of k-NN is overwhelmed by boosting it each time by learning the pattern of learning done on the training dataset and a partial of the records within it is used for imputation once it completes its learning process then starts to impute missing values of records which are considered as testing records.

TABLE III. PERFORMANCE COMPARISON BASED ON DEDUPLICATION OF AUTISM DATASET

	Firefly algorithm	Bat Algorithm	Artificial Bee Colony
Accuracy	73.4	76.8	93.6

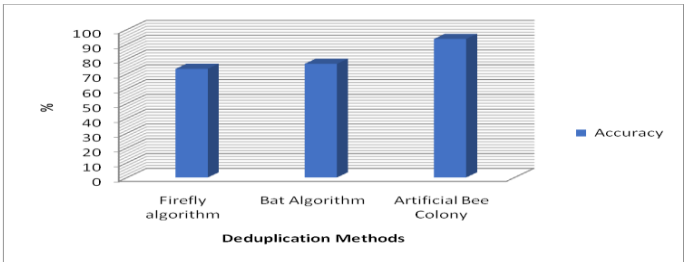


Figure 3 Performance comparison based on Accuracy

The table 4 and the figure 3 illustrate that the proposed work Artificial Bee colony produces highest percentage of accuracy. The other two methods hold less value because the ability to determine instance which belongs to either ambiguity and noise (outlier) are not handled in in enhanced boosted KNN. The boosted KNN is used for improving the quality of autism dataset which consist of missing values. Both the boosted KNN and ABC improves the quality of dataset in a precise manner.

VI. CONCLUSION

The bulk of children with autism even have a learning disability (mental retardation), though many have brain disorder and visual and hearing impairment are over-represented in this group. This paper aims at improving the quality of autism dataset by performing optimized data preprocessing which focused on two important aspects missing value handling using boosted k-NN and record deduplication using Artificial bee colony optimization. The simulation results proved the performance of the proposed approach increases the quality of dataset considerably.

REFERENCES

[1] https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/autism-causes-signs-symptoms-prevention/articleshow/61655180.cms

[2] Stahl D, Pickles A, Elsabbagh M, Johnson MH, Team B, et al. Novel machine learning methods for ERP analysis: a validation from research on infants at risk for autism. Developmental neuropsychology. 2012;37(3):274–298.

[3] Wilson CE, Happé F, Wheelwright SJ, Ecker C, Lombardo MV, Johnston P, et al. The Neuropsychology of Male Adults with High-Functioning Autism or Asperger Syndrome. Autism Research. 2014;7(5):568–581.

[4] Bone D, Bishop S, Black MP, Goodwin MS, Lord C, Narayanan SS. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. Journal of Child Psychology and Psychiatry. 2016.

[5] Grossi E, Veggo F, Narzisi A, Compare A, Muratori F. Pregnancy risk factors in autism: a pilot study with artificial neural networks. Pediatric research. 2016;79:339–347.

[6] Crippa A, Salvatore C, Perego P, Forti S, Nobile M, Molteni M, et al. Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities. Journal of autism and developmental disorders. 2015;45(7):2146–2156.

[7] Yang, X. S., ―A New Metaheuristic Bat-Inspired Algorithm, in: Nature Inspired Cooperative Strategies for Optimization‖, Studies in Computational Intelligence, Springer Berlin, 284, 2010, pp. 65–74.

[8] S. Yılmaz, E. Ugur Kucuksille, and Y. Cengiz, ―Modified Bat Algorithm‖, ISSN 1392-1215, 20, 2014, pp.71-78.

[9] Kaveh and P. Zakian, ―Enhanced Bat Algorithm for Optimal Design of Skeletal Structures", Asian Journal of Civil Engineering (BHRC), 15, 2014, pp.179-212.

[10] Anibal Sólon Heinsfelda, Alexandre Rosa Franco, R. Cameron Craddock, Augusto Buchweitzb, Felipe Meneguzzia, Identification of autism spectrum disorder using deep learning and the ABIDE dataset, NeuroImage: Clinical 17 (2018) 16–23

[11] Bhaskar Sen, Neil C. Borle, Russell Greiner, Matthew R. G. Brown, A general prediction model for the detection of ADHD and Autism using structural and functional MRI, MRI. PLoS ONE 13(4): 2018, pp 1-28

[12] Muhammad Nazrul Islam, Kazi Shahrukh Omar, Prodipta Mondal, Nabila Shahnaz Khan, Rezaul Karim Rizvi, A Machine Learning Approach to Predict Autism Spectrum Disorder, International Conference on Electrical, Computer and Communication Engineering (ECCE 2019), February 2019

[13] Koyamada, S, Shikauchi, Y, Nakae, K., Koyama, M., Ishii, S, Deep Learning of fMRI Big Data: A Novel Approach to Subject-Transfer Decoding, Neural Networks SI: NN Learning in Big Data Pp 1-21,2015.

[14] SergeyM.Plis, DevonR.Hjelm, RuslanSalakhutdinov, ElenaA.Allen, HenryJ.Bockholt, JeffreyD.Long, HansJ.Johnson, JaneS.Paulsen, JessicaA.Turner and VinceD.Calhoun, Deep learning for neuroimaging: a validation study, Front., Neurosci. 8 (August), 229. Pp 1-11, 2014.

[15] Cover, T., Hart, P.: Nearest neighbour pattern classification. IEEE Trans. Inf. Theor. 13(1), 21–27 (1967)

[16] H. Zhang, Y. Zhu, W. Zou, X. Yan, A hybrid multi-objective artificial bee colony algorithm for burdening optimization of copper strip production, Applied Mathematical Modelling, 36:6(2012) 2578-2591.

[17] X. Liao, J. Zhou, R. Zhang, Y. Zhang, An adaptive artificial bee colony algorithm for long-term economic dispatch in cascaded hydropower systems, International Journal of Electrical Power & Energy Systems, 43:1(2012) 1340-1345.

[18] Moises G. de Carvalho, Alberto H. F. Laender, Marcos Andre Goncalves and Altigran S. da Silva(2011). A Genetic Programming Approach to Record Deduplication. IEEE Transaction on Knowledge and Data Engineering,pp 399-412.

[19] Fadi Fayez Thabtah Department of Digital Technology Manukau Institute of Technology, Auckland, New Zealand, https://www.kaggle.com/fabdelja/autism-screening-for-toddlers

Mrs. Dr.M.Renuka Devi , has nearly 18 years of post graduate teaching experience in Computer Science. She has indulged in training the post graduate students to complete real time projects and also guides research scholars in Computer Science. She has published 40 papers in various international journal. There are 8 scholars doing PhD under her guidance. She has acted as chair in international conferences. Currently she is working as Professor and Head in the Department of BCA at Sri Krishna Arts and Science College.

Mr.R.Suresh Kumar is working as Assistant .Professor, Department of Computer Science, Sree Saraswathi Thyagaraja College. His research interests include classification, clustering of Data Mining for different type of datasets.