International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Handling Scarcity of Data in Autism Identification using Binary Imputation Method

Sushama Rani Dutta, Sujoy Datta, Monideepa Roy

Abstract*: Autism is a neuro-developmental disorder. Identifying the type of autism is a very crucial job for a doctor, since each type of autism has a different type of therapy. In rural areas, the identification and prediction of suspected autistic children become difficult because of various factors. This is because, mostly the parents are uneducated and are not able to notice and express the symptoms of their children properly. This in turn leads to the doctors often being left to deal with incomplete datasets, thus making the diagnosis process erroneous or difficult. In our previous work, we had proposed a framework to assist the doctors as well as the parents of the anticipated patients in rural areas to better recall the maximum number of symptoms, by prompting them for associated symptoms, once a first symptom is mentioned by the parent. Our method prompted the parents with possible associated symptoms based on previous autistic children data stored in EHR (Electronic Health Records).However, in case of surveys where the above procedure has not been implemented, the complete set of symptoms for a patient may not be available, thus leading to incomplete datasets. The incomplete datasets are the data sets which are having missing symptoms. Diagnosis of autism with missing symptoms is very difficult. In this paper, we have proposed a Binary Imputation method (BIM) algorithm, to handle such missing symptoms in the collected datasets, which uses the weight factors (influence of parameter on the disease diagnosis) of the symptoms. This method inserts a binary "1" for imputing values in place of some missing attributes, which is decided by the proposed BIM. We use Levenshtein distance (LD) formula for finding the suspected child by imputing '1' in place of only one high weight missing symptom in a dataset. This method has been tested with the collected Asperger syndrome (autism type) datasets for identification of autism. We get better accuracy in diagnosis of autism and finding of the suspected child, as compared to other missing values handling methods like K nearest neighbour imputation method, mean imputation and case deletion methods. This method will help the doctor for easy diagnosis with the datasets having missing symptoms because all the missing symptoms can be handled by BIM algorithm.

Index Terms*: Autism identification; mRMR rule; Machine learning; Weight factor; Missing symptom; Asperger Syndrom; Binary imputation method.

I. INTRODUCTION

Autism spectrum disorder (ASD) has a range of challenging conditions like social skill, verbal and nonverbal communication, repetitive behavior etc. It has up to 12 different types of variants.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Sushama Rani Dutta*, SRF in ITRA project in the School of Computer Engineering, KIIT Deemed to be University
Sujoy Datta, Assistant Professor in the School of Computer Engineering, KIIT Deemed University
Monideepa Roy, Associate Professor at KIIT Deemed University, Bhubaneswar.

The differences in the symptoms of different types of autism are mostly very subtle and hard to identify. However, identifying the type of autism is a very crucial job for a doctor, since each type of autism has a different type of therapy. In addition to this, the symptoms need to be identified during the early years of the child, for an effective and complete cure. But in rural areas, the identification and prediction of suspected children become difficult because of various factors mostly because the parents are uneducated and not able to notice and express properly the symptom of their children. The parents may feel lack of confidence, lack of communication skill, emotions or privacy issue to express the symptom of their child. This causes lack of information leading to difficulty in tracking in the identification of type of autism. In our previous work, we have proposed a method i.e. mutual association selection (MAS) method for the Identification of type of autism from an initial symptom [1]. Once a first symptom is identified, this method assists the doctors as well as parents of the children in rural areas by prompting the possible associated symptoms. This is because a parent may not always be able to express or remember all the symptoms properly. So our method pulls the associated symptoms of a typical autistic patient from the electronic health record (EHR), which stores symptoms of the previous autistic patients. This method pulls the next and appropriate symptoms from the EHR by using Association rule (AR), minimum Redundancy and Maximum Relevance (mRMR rule) and mutual information dependency rule. We used the database of symptom sets of the different types of autism, which is based on the domain expert knowledge to verify the diagnosed set. Our method starts with a single symptom expressed by the parents or by observing the child behaviour. This method was executed by applying machine learning methods. We used Association rule (AR), minimum Redundancy and Maximum Relevance (mRMR rule) and mutual information differences rule to pull the appropriate symptoms from the EHR which were then confirmed by the parents. The final diagnosis needs to be verified with the data sets of the symptoms available in the database of domain expert's knowledge base. In rural areas the autistic children are misguided for the treatment because of poor communication with the doctors. This system will help the parents to express or recall the symptoms that are experienced by their child. This also helps the doctors in the easy identification of the type of autism. A survey carried out in a rural area for estimating the prevalence of autism spectrum disorder, found out that more than 80% of data sets have missing data [2]. Due to the absence of some important data it is very difficult for a doctor to diagnose or analyze these datasets for autisms. The missing information in the collected data sets are due to many reasons, e.g.

less interaction of children with parents, communication problem of the parents, privacy issues, language problems or less attention towards a child's behavior. In order to solve this problem, many conventional methods have already been applied like the case-deletion method, which deletes all the datasets having missing data, Mean imputation method, which imputes the mean value of all other available values of the same attribute and KNN imputation method which imputes the average value of the K nearest neighbors of the missing attribute. The incomplete datasets and the noisy datasets are the main causes of improper diagnosis or improper identification of disease [3].

In this paper we have proposed a new technique to handle these missing symptoms in the autism dataset. To handle the datasets we categorized them into two categories. The first category has binary complete datasets i.e the total dataset having value 1 in all attributes and the second category has binary incomplete datasets i.e attributes containing both 1 and 0 values. All the confirmed symptoms in the dataset are represented as '1' and the symptoms absent or missing are represented as '0'. This method uses the weight factor of each symptom, which is decided by the domain expert. These binary incomplete datasets are further divided into 3 types:

A. Datasets having only low weight symptoms missing.
B. Datasets having both low weight symptoms and only one high weight symptom missing.
C. Datasets having many number of high weight symptoms missing.

Then applying our new algorithm i.e. Binary Imputation method (BIM), we have imputed the binary '1' to the missing symptoms of 'A' type datasets. We can also find the suspected autistic child by applying Levenshtein distance formula to the 'B' type of datasets. We have rejected the 'C' type of datasets which are having many numbers of high weight symptoms missing. We got very good accuracy in identification of type of autism and found the suspected autistic child than applying other conventional methods like K nearest neighbour imputation method, mean imputation and case deletion methods for handling missing symptoms**.** The experiment has been done with the 500 dataset collected from a rural area of Orissa to estimate the number of children suffering from the Asperger syndrome (Autism). Asperger syndrome is a major and common type of autism in children in rural area. Children suffering from the Asperger syndrome generally have the problem in communications and social interaction. The rest of this paper is organised as follows: Section II represents the related works, Section III explains the mutual association selection (mas) method, Section IV discusses binary imputation method (BIM), Section V discusses the experimental analysis and result, Section VI discusses the conclusion and future work.

II RELATED WORKS

Prevalence of the chronic disease increasing gradually in the world and controlling the disease is a great challenge. New technology using machine learning improving the healthcare system. In order to improve the quality of life for individual, the wearable sensors are used, which can detect the abnormal condition of a person. The physical parameters like temperature, pressure, heart rate, pulse rate, oxygen in blood etc can be measured by the sensors and the abnormal case of the person can be intimated to the concerned doctors and family members [4]. Continuous monitoring of the mental illness patient with the daily activity can increases the effectiveness of therapy of the mental disorder. The multiple sensors embedded in smartphones are used to monitor the multiple parameters of the body of a person such as social interaction dimension, physical and mental stresses [5]. Machine learning technique improves the diagnosis method as well as quality of life. The supervised machine learning technique used in identifying the minimal sets of behavior in a survey of ASD[6] .Machine learning is used to build the quantitative model to distinguish the autism features as mild, medium, sever autism[7].

Different machine learning classifiers are used to classify the autistic children in different survey as mild group, medium group and severe autism patient group. Twelve feature support vector machine is used in [8] to classify autism groups. Five feature LR classifier (LR5) used in [9] for the selection of categories in autism. EHR are used for the early assessment of diseases by the help of previous patient data and time series information [10]. This paper used classification handling technique and sequential mining rule for data abstraction. The different autism parameters are measured through body sensors like skin conductance sensor is used to measure galvanic skin conductance [11].

Autism identification in a rural area is a difficult task due to missing information in dataset. In health domain many doctors handled the missing data differently. In [12] the authors rejected the data sets having missing information. This method is not effective for identification of disease as total number of datasets will be reduced. In [13] the authors described the multiple imputation method i,e trying with multiple values one by one which is a time consuming process and not much effective method for maintaining accuracy. In [14] the mean imputation method is applied to replace the missing value. In which the average value calculated by all other present value, this causes improper diagnosis of disease. Machine learning used in K-Nearest Neighbor (KNN) imputation method [15] in which the complete samples will be selected and calculate the Euclidean distance between the missing and the available attributes. This method is a time consuming and also expensive method of searching complete sample. Proper classification of disease is a challenge with the missing datasets. In [16] the authors used neural network for the classification of missing data. In this paper maximum missing value replaced by the available information and without disturbing the characteristics of datasets. It did not allow any assumed value for replacement.

We have classified the autism datasets into two categories. The first one is complete binary dataset i,e all attributes are '1' and the other is incomplete binary datasets having '1'&'0' values. We have applied BIM algorithm for replacement of missing value depending upon the weight factor of the missing symptoms. We have used Levenshtein distance formula to identify the suspected autistic child [17]. This formula makes a dataset as a complete dataset by adding one attribute or by deleting one attribute. This helps us to make our dataset complete by handling missing attribute and it helps to identify the suspected one.

III MUTUAL ASSOCIATION SELECTION (MAS) METHOD

Our previous was applied when parents are unable to express the symptoms of their child. Our diagnosis method takes only one symptom initially and pulls next most possible symptom from Electronic Health Record (EHR).The Procedure of mutual association selection (MAS) method is shown in Fig. 1. We applied Association Rule (AR) and minimum redundancy and maximum relevance (mRMR) rule for pulling symptoms. The pulled symptom then confirmed by the parents and also matched with the Domain Expert's knowledgebase (DEK), which contains the symptom sets of all type of autism. If any dataset matched with the calculated data set then that dataset considered as a predicted disease dataset. If the symptom set not confirmed by the parent or not matched with DEK datasets, then repeating same method by applying mutual information difference (MID) rule to pull the next most appropriate symptom up to finding matching dataset from DEK. The detail method is explained below.

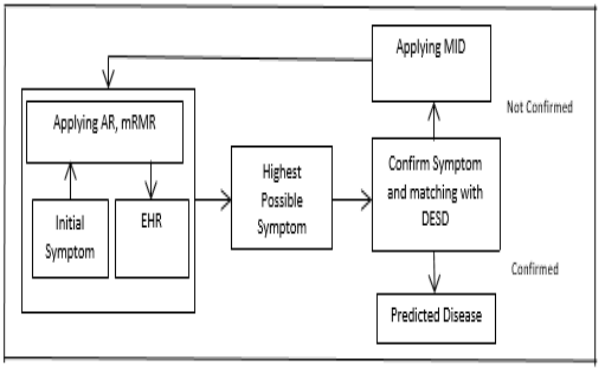

Fig. 1 Procedure of mutual association selection (MAS) method

This method starts with a single symptom applying AR rules as follows

According to the association rule (AR)

X->Y,where

X- Initial symptom of an autistic child

Y- Targeted symptom which can be pulled from EHR

To get Y from X we have applied the interestingness's rule. It has two factors 'support' and 'confidence'

Support(X) means number of time X experienced or available throughout the database or with all datasets available in the database

Confidence (X->Y)

Probability of Conf (X->Y) =

number of time X experience with y / Support of X

The minimum redundancy and maximum relevance rule of machine learning is applied to select the maximum occurrence symptom (Y) with (X) which has minimum occurrences with other symptoms.

The mutual information dependency (MID) rule is used to consider X= X+Y.

X is the initial symptom and Y is the next targeted symptom. For the next phase of consideration, X+Y will be initial symptom X, i.e. mutually X and Y form again X to pull Y from EHR.

X and Y mutually depending upon another target Y. We have taken one example to prove the predicted autism from an initial symptom by applying AR, MRMR and MID rule.

Let the different type of autism are d1, d2, d3--- dn and symptom sets of the autism are:

d1={ S_{13} , S_5 , S_{15} , S_7 }, d2 ={ S_7,S_5,S_{11},S_{13},S_{11} }, d3={S_4,S_2,S_5,S_2}----

d7 = {S_5, S_7, S_{10}, S_2} ---- dn= {S_1,S_3,S_{10},S_7}

Symptoms are used as abbreviated form. Let the parents is expressing only one symptom of his /her child is 'S_5'. In order to prove the child having d7 type autism having {S_5, S_7, S_{10}, S_2} symptoms. To get all other symptom we follow the given methods as follows

Let ' S_5 ' symptom present with S_7,S_{12},S_{10},S_9,S_2,S_{15} symptoms in different autism cases.

Considering support of symptom 'S_5' is 13.

By applying association rule (AR) we get,

'S_5' experiencing 11 times with S_7 Confidence of S_5-> S_7 = 11/13= 0.84 i,e in EHR like wise

'S_5' experiencing 1 time with S_{12} Confidence of S_5-> S_{12} = 1/13 = 0.07

'S_5' experiencing 8 times with S_{10} Confidence of S_5-> S_{10} = 8/13 = 0.61

' S_5' experiencing 2 times with S_9 Confidence of S_5-> S_9= 2/13 = 0.15

' S_5' experiencing 2 times with S_2 Confidence of S_5-> S_2 = 5/13 = 0.38

' S_5' experiencing 2 times with S_{15} Confidence of S_5-> S_{15} =2/13 =0.07

By applying mRMR rule, we get S_5-> S_7 = 0.84 which has highest accuracy. After confirming 'S_7' as next pulled symptom, we take S_5 and S_7 as initial symptom and by applying MID rule, 'S_5' and 'S_7' mutually dependent on the following symptoms.

confidence of (S_5S_7) -> S_{10} = 0.6

Conf (S_5S_7) -> S_{15} = 0.23

Conf (S_5S_7) -> S_2 = 0.5

Conf (S_5S_7) -> S_3= 0.1

Conf (S_5S_7) -> S_{13}= 0.05

By applying mRMR we got S_5S_7 -> S_{10} as the next appropriate symptom set. Each time after getting target symptom, need to confirm from the parent and if the parent confirmed about the symptom then take previous symptom and the confirmed derived symptom as the symptom set for discussion.

This symptom set then matched with the symptom set of domain expert knowledgebase. If any matching found in DEK, then the symptom set consider as the predicted disease otherwise again by applying MID rule to all symptom present in the symptom set and search for target symptom from EHR.

The $\{S_5, S_7, S_{10}\}$ data set is not sufficient for any disease identification. By applying MID we found

$S_5S_7S_{10}$->S_{15} = 0.1

$S_5S_7S_{10}$-> S_2= 0.4

$S_5S_7S_{10}$-> S_{13} = 0.23

By applying mRMR, we have considered $S_5S_7S_{10}$-> S_2 as the highest relevance symptom given in Fig.2. G also confirmed by the parent and $S_5S_7S_{10}S_2$also matched with DEK symptom set, the matching found , then the data set $\{S_5, S_7, S_{10}, S_2\}$ dataset predicted as d7 type autism, which is already mentioned above as the symptom set of d7 type autism. The details of the explanation of the pulling system of symptom is given in the Fig.3.

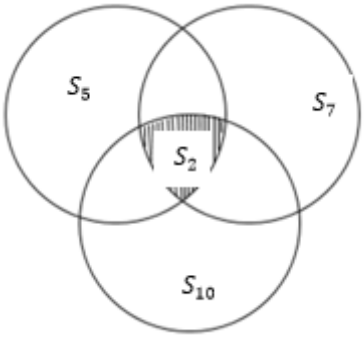

Fig.2 $S_5S_7S_{10}$ mutually selecting symptom S_2

Symptom	possibility of target symptoms	occurrences	selected symptom
S_5	--S_7--	0.84	S_5S_7
	--S_{12}--	0.07	
	--S_{10}--	0.61	
	--S_9--	0.15	
	--S_2--	0.38	
	--S_{15}--	0.07	
S_5S_7	--S_{10}--	0.6	$S_5S_7S_{10}$
	--S_2--	0.23	
	--S_{15}--	0.4	
	--S_9--	0.1	
	--S_{12}--	0.05	
$S_5S_7S_{10}$	--S_{15}--	0.1	$S_5S_7S_{10}S_2$
	--S_2--	0.2	
	--S_9--	0.01	

Fig.3 maximum occurrence of symptom selection procedure

IV BINARY IMPUTATION METHOD (BIM)

The collected datasets for identification of disease are having many number of missing information. While collecting data for identification of autism a set of questionnaire asked to the parent. The answers are gathered as yes or No or blank.

We have converted all datasets into binary datasets. We represent all yes as '1' and No as '0' and blank as '0'. A reference dataset is used while collecting data, the reference datasets are different for different type of autism. Hence all the completed datasets i,e all the field filled with 'yes' are represented as complete binary dataset, like {1,1,1,….1}. The datasets which are having no value or blank information are treated as incomplete datasets, like {1, 0, 1, 10, 0, 1}. It includes few '1' and few '0' value. '0' indicates many indications. The symptoms missing due to many reasons like parents unable to notice or express the symptom or parent hiding information due to privacy issue or health workers missed to enter the attribute. But by handling this '0' values in the dataset is very difficult for proper identification of autism.

Our model shown in Fig.4 based on the weight factor of the symptom. The weight factor is decided by the domain expert. The high weight factor symptom has a high impact or high importance on the disease detection. The different autism symptoms are ranked from 0 to 1.

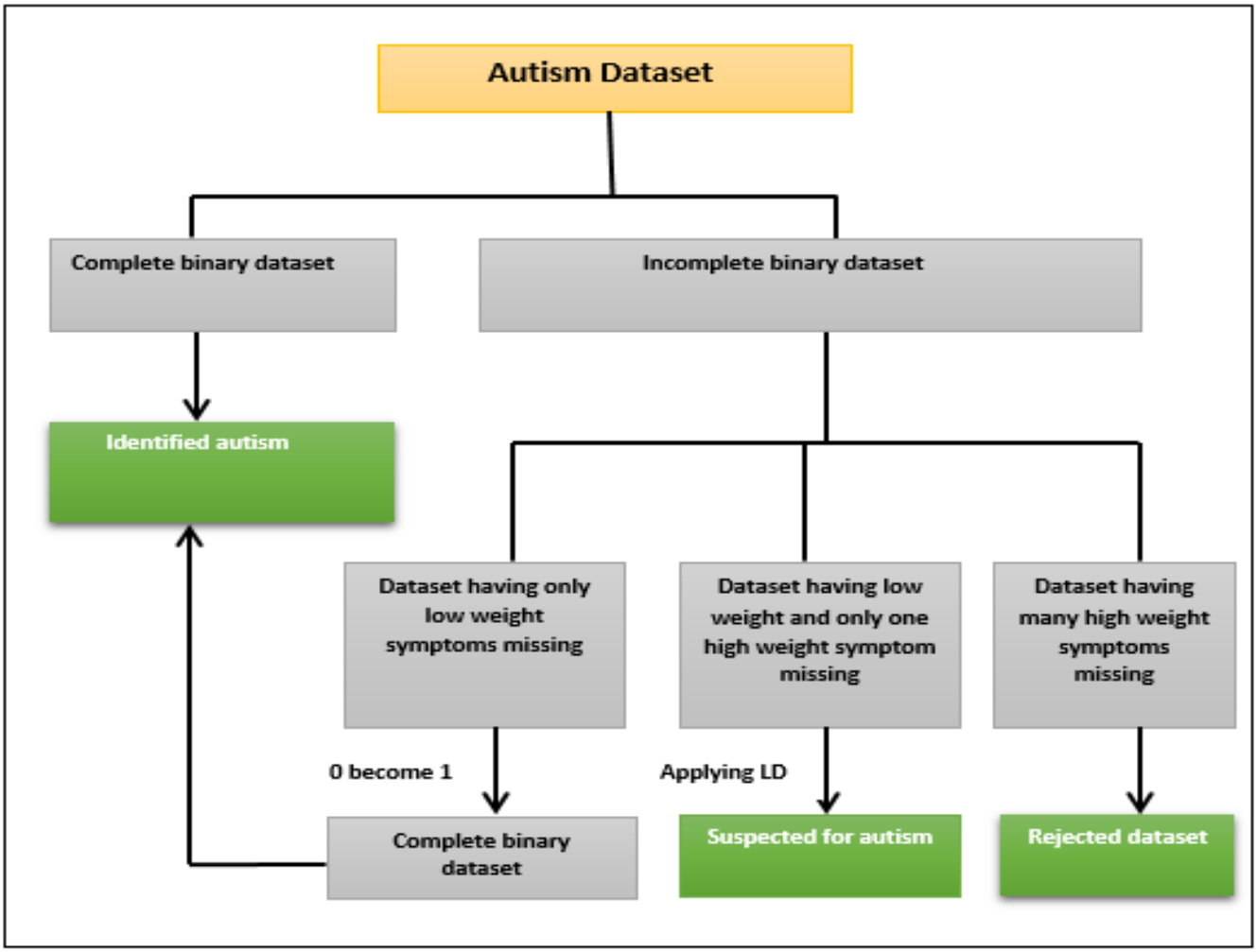

Fig.4 Management of missing data in autism dataset

We have taken the weight factor above 0.5 has a high impact on disease diagnosis and weight factor below 0.5 has a low impact on disease diagnosis. We designed BIM algorithm (given below) for the replacement of missing information in the autism datasets. We use Levenshtein distance (LD) formula for finding the suspected child from the datasets which are having both low weight and only one high weight symptom missing. According to the levenshtein distance (LD) formula.

$$leb_{X,R}(n,n) = \{\min\{ leb_{X,R}(n-1,n)+1\}\}$$

Where X is the patient symptom set and R is the reference symptom set and both the lengths are same and equal to 'n'. If x having some missing data then according to this formula, addition of one symptom to n-1 number of symptoms, it makes the set as a reference set. Hence we can suspect that set as autism set. This type of sets are considered as suspected autism set.

According to the Fig. 4 we have categorised the incomplete datasets in to 3 types

1. Datasets which is having only low weight symptoms missing.
2. Datasets having low weight and only one high weight symptom missing.
3. Datasets having many number of high weight symptoms missing.

Based upon these 3 types of incomplete datasets and the weight factor of the symptom, we have designed the BIM algorithm to replace all type of missing symptom explained in algorithm 1.According to this algorithm,

1. If the symptom dataset 'S' having only low weight symptom missing then all the missing symptoms will be replaced by '1'. Hence we get the complete binary set i.e. the datasets having '1' in all field will indicate the identification of autism.
2. If the symptom set 'S' having many numbers of high weight symptoms missing then we are rejecting that datasets as multiple numbers of high weight symptoms missing, which are very risky to replace by 1.
3. If the symptom sets 'S' having low weight symptoms and only one high weight symptom missing then we replaced all low weight symptom as '1' and replace '1' with the only high weight missing symptom. This is done by applying Levenshtein distance formula.

Algorithm. 1 Binary Imputation method (BIM)

```
Require: Autism Symptom set S
1:      If set S having only missing data of low weight value
2:          Then Replace missing value by 1  // low weight symptom does not have much effect
            on autism diagnosis
3:      else if S having missing data of both low weight values and only one high weight value.
4.          Replace all low weight missing value by 1
5:          then apply Lavenshetine distance formula
6:          Replace one high weight missing value by 1 // suspect symptom
7:          Find suspected set for autism
8:      else reject dataset // having many number of high weight symptom
9:      end if
10:     end if
```

As we are dealing with two type of scarcity of data handling system, the implementation of our previous work is given in section A by collecting data from Autism Therapy counselling and Help (CATCH),Bhubaneswar,India. Section B implements the binary imputation method (BIM), for this we have collected 500 datasets from a rural area for the identification of Asperger Syndrom (Autism type) in children.

A. Automatic pulling of symptoms for Autism identification

In this phase, we have taken an autism database, which contain 500 symptom sets of autistic children. We have taken a database of symptom sets of 10 type of autism prepared by the domain expert knowledge. This data sets containing 50 different type of symptoms of autism. We have tested 100 children for different type of autism in that center. We have only taken one symptom from the parent and started the diagnosis by using our system which will automatically pulled the symptoms using AR, mRMR, MID rule of machine learning. Fig.5 shows the testing of intellectual disability which was started the diagnosis with only one symptom 'Stress'. Communication disorder pulled in first chance and confirmed by the parent but these two symptoms could not decide the type of autism. So by using MID rule we get more options. First option hearing screening is not confirmed by the parent so next maximum relevant symptom pulled for testing i, e Aphasia and Articulation but still symptom set not matched with the dataset of DEK. Hence next tested for Dyslexia, dyscalculia and Bhatia battery and after getting confirmation from the parent it concluded with the predicted autism is Intellectual disability. The experiment get similar accuracy like real diagnosis with maximum available data. We got 95% accuracy in Intellectual Disability (ID) where the actual diagnosis got 87%. We got 78% accuracy in Pervasive Development Disorder but actual diagnosis got 80%. This variation was due to lack of information at that diagnosis period.

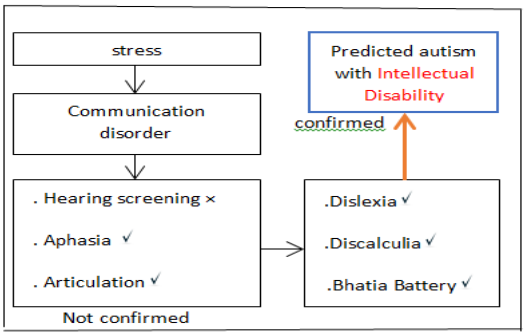

Fig.5 Diagnosis of Intellectual disability from stress only

B. Handling Missing information by binary imputing method (BIM)

The Imputation of binary '1' based on our algorithm BIM, which depends on the weight factor of each missing symptom. For this experiment we have collected data from a rural area of Orissa to test how many children are affected by Asperger Syndrome. Asperger syndrome is a type of Autism which is mostly seen in rural children. It is observed that normal children who are

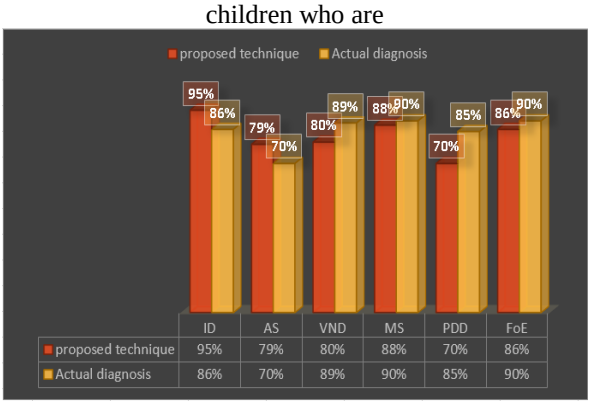

Fig.6 Comparison of proposed and actual diagnosis

staying near (neighbor child or cousins staying in a house) the autism affected children are behaving like Asperger Syndrome affected children. It seems affected children influencing normal children. Our survey collected 500 children data for the identification of the Asperger syndrome symptom. We have collected the data using a reference set of Asperger syndrome symptom. This reference dataset contains 11 attributes or 11 symptoms for the actual identification of Asperger syndrome in that particular area. These symptom are (1) Abnormal eye contact (2) Aloofness (3) Failure to respond when called by name(4) Communication difficulties (5) Lack of social interaction (6) Socially immature (7) Repetitive behavior (8) Anxiety (9) Difficulty in understanding nonverbal communication (Body Language) (10) Sleep problem, frequent awakening (11) Unaware of emotion . Among 500 children we have identified 3 categories of children (1) Asperger Syndrome (AS) affected children (2) Suspected by AS and (3) Not affected by AS. But After collecting the data, we found many numbers of missing attributes in the data sets. The list of missing symptom and percentage is given in Table 1.

TABLE 1. Details of collected data with missing information

S .no	Attribute having '0' value	Number of sample having '0' value	Percentage (%) of patient
1	Abnormal eye contact	457	91.4 %
2	Aloofness/ not friendly	430	86 %
3	Failure to respond when called by name	487	97.4 %
4	Communication difficulties	347	69.4 %
5	Social interaction	457	91.4 %
6	Socially immature	235	47 %
7	Repetitive behaviour	278	55.6 %
8	Anxiety and depression	348	69.6 %
9	Difficult to understand nonverbal (body language)	357	71.4 %
10	Sleep problem , frequent awakening	275	55 %
11	Unaware of emotion	387	77.4 %

After collecting data, we have converted data to binary '1' and '0'. The symptoms present is taken as '1' and absent or missing is taken as '0' as discussed above. We observed only 5 children i.e. 1% of the datasets have complete data in each field i,e each field contains 1 in the entire dataset but many datasets are suspected due to presence of missing data .We have evaluated the performance of our algorithm by handling these incomplete datasets and compared with the performance of KNN imputation, Mean Imputation and Case deletion method for handling these incomplete data.

Our algorithm is based on the weight factor of the symptom. Based on the weight factor, the symptoms values are decided from 0 to 1 by the domain expert. The list of the weight factor of the symptom for the Asperger syndrome is given in Table 2.

TABLE 2. Weight factor of the symptoms of Asperger syndrome

s. no	Symptom	Impact value
1	Abnormal eye contact	0.9
2	Aloofness/ not friendly	0.5
3	Failure to respond when called by name	0.6
4	Communication difficulties	0.7
5	Social interaction	0.7
6	Socially immature	0.4
7	Repetitive behaviour	0.3
8	Anxiety and depression	0.2
9	Difficult to understand non verbal (body language)	0.3
10	Sleep problem , frequent awakening	0.2
11	Unaware of emotion	0.4

The weight above 0.5 considered as high weight symptom and below 0.5 considered as low weight symptom. Based on this concept we have implemented our algorithm to get the suspected child and affected child from this 495 incomplete datasets. We got 5 more children affected and 3 children are suspected for Asperger syndrome. The remaining 487 data set are rejected. The 10 fold cross validation method is used to increase the accuracy for the training and testing data. Among the 495 incomplete dataset, the case detection method rejected all 495 datasets.

The Mean Imputation method rejected 493 datasets and suspected 1 datasets and identified 1 dataset as affected. The KNN imputation method rejected 491 datasets identified 3 as affected dataset and 1 dataset as suspected dataset. The comparison table is given in Table-3. This is an experiment through 500 children but when it is tested for a large database our algorithm give more accuracy than other conventional methods. The comparison Graph is shown in Fig. 5.

TABLE 3 Comparison between proposed method and KNN, Mean, Case Deletion method

	proposed	KNN	mean	case deletion
suffered	5	3	1	0
suspected	3	1	1	0
rejection	487	491	493	495

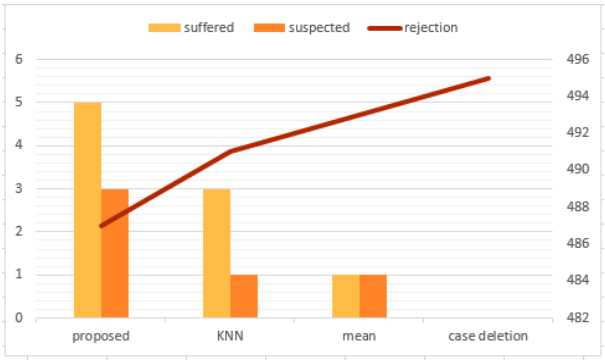

Fig.7 Comparison graph between proposed method and KNN, Mean, Case deletion method

VI CONCLUSION AND FUTURE WORK

Autism diagnosis with the datasets having missing symptoms is very difficult task for a doctor. We used BIM algorithm for handling the missing symptoms. BIM algorithm used the weight factor of symptom to make the symptom high weighted symptom or low weighted symptom to replace its missing value with binary value '1'. We have handled with the low weighted symptoms and with one high weighted missing symptom. Still future research is there to handle the binary incomplete datasets which are having many numbers of high weighted symptoms. Our algorithm got good accuracy over case deletion, mean imputation and KNN imputation methods for autism. Hence rejection of incomplete datasets are reduced. The paper also solved the problem of rural parents who are unable to express the symptom of their children. We used AR, mRMR and MID to pull the most expected symptoms from EHR. We have taken only one prominent symptom which can be noticed and expressed by any parent easily. The system dragging all other symptoms from EHR and confirming from parent. It also matched with the data set of DEK to reduce error and predict the Autism .This system helps the doctors for easy and faster diagnosis of Autism.

ACKNOWLEDGEMENT

This work has been carried out with the support and funding of ITRA Media Lab Asia and DeitY through the project "Remote Health: A Framework for Healthcare Services using Mobile and Sensor-Cloud Technologies".

REFERENCES

1. Dutta,Sushama Rani et al. "A Machine Learning-Based Method for Autism Diagnosis Assistance in Children." 2017 International Conference on Information Technology (ICIT)(2017): 36-41.
2. Barua M, Daley TC. Autistic Spectrum Disorders: A Guide for Paediatricians in India. New Delhi, India: Naveen Printers; 2008. pp. 13–14.
3. Robertson, Colin, et al. "Review of methods for space–time disease surveillance." Spatial and spatio-temporal epidemiology 1.2 (2010): 105-116.
4. Park, Sungmee and Sundaresan Jayaraman (2003). \Enhancing the quality of life through wearable technology". In: IEEE Engineering in medicine and biology magazine 22.3, pp. 41-48
5. Grunerbl, Agnes et al. (2015). Smartphone-based recognition of States and state changes in bipolar disorder patient
6. Duda M, Daniels J, Wall DP. Clinical Evaluation of a Novel and Mobile Autism Risk Assessment. J Autism Dev Disord. 2016; 46(6):1953–61. https://doi.org/10.1007/s10803-016-2718-4 PMID: 26873142. PMCID: PMC4860199.
7. Bone D, Bishop SL, 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; 57(8):927–37. https://doi.org/10.1111/jcpp. 12559 PMID: 27090613. PMCID: PMC4958551.
8. Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Translational Psychiatry. 2015; 5(2):e514. https:// doi.org/10.1038/tp.2015.7 PMID: 25710120. PMCID: PMC4445756.
9. Levy S, Duda M, Haber N, Wall DP. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. Mol Autism. 2017; 8(1):65. https://doi.org/10.1186/ s13229-017-0180-6 PMID: 29270283. PMCID: PMC5735531.
10. Cheng, Yi-Ting, et al. "Mining Sequential Risk Patterns From Large-Scale Clinical Databases for Early Assessment of Chronic Diseases: A Case Study on Chronic Obstructive Pulmonary Disease." IEEE Journal of Biomedical and Health Informatics 21.2 (2017): 303-311.
11. Nehme, B., et al. "Developing a skin conductance device for early Autism Spectrum Disorder diagnosis." Biomedical Engineering (MECBME), 2016 3rd Middle East Conference on. IEEE, 2016.
12. Zhang, Zhaoyang, et al. "Cluster-based epidemic control through smartphone-based body area networks." IEEE Transactions on Parallel and Distributed Systems 26.3 (2015): 681-690
13. Salgado, Cátia M., "Missing Data." Secondary Analysis of Electronic Health Records. Springer International Publishing, 2016. 143-162.
14. Dong, Yiran, and Chao-Ying Joanne Peng. "Principled missing data methods for researchers." SpringerPlus 2.1 (2013): 222.
15. Jonsson, Per, and Claes Wohlin. "An evaluation of k-nearest neighbour imputation using likert data." Software Metrics, 2004. Proceedings. 10th International Symposium on. IEEE, 2004.
16. Tax, David MJ, and Robert PW Duin. "Combining one-class classifiers." International Workshop on Multiple Classifier Systems. Springer Berlin Heidelberg, 2001.
17. Navarro, Gonzalo. "A guided tour to approximate string matching." ACM computing surveys (CSUR) 33.1 (2001): 31-88 nry8

AUTHORS PROFILE

Sushama Rani Dutta has done her M.Tech. from MMU,Haryana. Currently she is working as SRF in ITRA project in the School of Computer Engineering, KIIT Deemed to be University. Her areas of research include Wireless networks, Data analysis. Sensor network. She has several publications in various conferences and journals.

Sujoy Datta has done his M.Tech. from IIT Kharagpur. Currently he is working as an Assistant Professor in the School of Computer Engineering, KIIT Deemed University since the last seven years. His areas of research include Wireless networks, security, Elliptic curve cryptography and neural networks. He has several publications in various conferences and journals. He has co-organised several workshops and international conferences.

Dr. Monideepa Roy did her Masters in Mathematics from IIT Kharagpur, and her PhD in CSE from Jadavpur University. Currently she is working as an Associate Professor at KIIT Deemed University, Bhubaneswar since the last seven years. Her areas of interest include Wireless Networks, Mobile Computing and WSNs, and Cognitive WSNs. At present she has three research scholars working with her in these areas. She has several publications in conferences and journals. She has been the Organizing Chair of the first two editions of the International Conference on Computational Intelligence and Networks CINE 2015 and 2016 and organised several workshops.