Journal of Information Systems Engineering and Management
2025, 10 (32s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

An Integrated Model of Natural Language Processing and Machine Learning (INM) for Autism Detection and classification from Random Symptom

S Hima Bindu Sri[1], Sushama Rani Dutta[2]

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India

[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India

Corresponding author: Sushama Rani Dutta (sushamarani.dutta@klh.edu.in).

ARTICLE INFO

Received: 27 Dec 2024

Revised: 12 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

Autism Spectrum Disorder (ASD) is a neurologic disability which affects daily life of autistic children. In general, parent of the autistic child provides symptoms in vernacular language and may express only one or two symptoms out of many. So, it is challenging to perceive this disorder in the early stage of the child. Even though it is challenging to detect at early stage, our proposed Integrated Natural Language Processing and Machine Learning (INM) model can minimize the severity of the condition. We collected Autistic children's data from "Total Solution Rehabilitation Society, Hyderabad, India". The dataset has many similar types of symptoms which may confound the machine learning model for Autism detection and classification. We experimented different NLP techniques i.e. Bag of Words (BoW), Bag of N-grams, TF-IDF, and One-Hot Encoding to normalize the dataset with highest cosine similarity index. We found (BoW) outperformed among all and we replaced similar-meaning symptoms with a unique symptom. The proposed model accepts any random symptom from parent as a preliminary symptom. Further, the frame work used Association rules to generate frequent symptom sets with minimum-support and maximum-confidence to identify the most appropriate symptoms with a preliminary symptom. Then frequent symptom set is classified through various ML classification techniques i.e. Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), ADA Boost (AB) and Linear Discriminant Analysis (LDA) to identify the proper autism type. Finally, the results are evaluated with several statistical evaluation metrics (Accuracy, Precision, Recall, F1-Score and Mathews Correlation Coefficient: MCC). After examining the experimental results, it is identified that Random Forest classifier detected Autism type with maximum accuracy of 90.48%, precision of 96.67%, recall of 91.53%, F1-score of 93.29% and MCC of 87.94%. The proposed INM model guides the decision-making of Autism health care medicos while examining the ASD cases. The proposed INM model is tested in the "Total Solution Rehabilitation Society, Hyderabad, India." organization with 1896 autistic children and achieved 99% accuracy on identification of ASD types.

Keywords: Autism Spectrum Disorder, Natural Language Processing, Bag of Words, Association rules, Machine Learning, Decision Tree.

I. INTRODUCTION

ASD is a neurodevelopmental disability, related to the development of brain which starts early in the age, affecting a person's interactions with society and communication with other people. [1], [2]. ASD can be defined with a very limited and repetitive behavioral conditions. The word "Spectrum" encircles a wide variety of symptoms such as speech impairment issues, behavioral issues, Improper development of social, cognitive, visual, academic, and sensory skills [3], [4],

[5]. Though we do not have any authentic solution to cure ASD, just early identification and a proper therapeutical care will make a considerable difference in the child's capability in communication and behavioral skills [6], [7], [8]. But still identification and ASD diagnosis is tough by using conventional behavioral techniques. Generally, Autism can be diagnosed at early age of two years and even after that depends on its severeness [9], [10], [11]. Nowadays, many kinds of treatments, therapies are available to detect Autism when parents provide a wide range of symptoms. Our proposed INM model works effectively even if the parent could express any preliminary symptom of a child.

Generally, ASD children are identified at the age of one or 2 years. At this early age, a child is unable to speak out. The parents of child may or may not express the all the behavior of the child properly in formal language which are required to diagnose the ASD properly. To achieve this, we have collected a real time data from "Total Solution Rehabilitation Society, Hyderabad, India" therapy center. As we are aware that any biomedical data consists of redundancy as well as it is unstructured and irregular information with casual language in wide variety of medical documents and electronic health records. The medical reports from clinics or therapy center are need to be analyzed to extract hidden meaningful information. [12]. So, to extract meaningful insights and unique information from casual language must be pre-processed by using text pre-processing techniques such as lower casing, tokenization, stop word removal and stemming etc.

To eliminate redundancy or to normalize the data, we need to create word embeddings by using distinct NLP techniques such as- Bag of Words [13], Bag of N-grams [14], TF-IDF [15], One-Hot Encoding [16]. Word embedding is a type of word representation used in natural language processing (NLP) and machine learning. It encode the semantic meaning of words in a continuous vector space.

Bag of Words (BoW) represents text as word frequencies, useful for sentiment analysis and document classification. Bag of N-grams Considers word sequences for tasks like machine translation and text generation. TF-IDF Measures word importance in a document relative to a collection, commonly used in information retrieval and search algorithms. One-Hot Encoding Converts categorical variables into binary format for machine learning models. These algorithms are used to create vector representations to find similarity between symptoms by using cosine similarity. We found Bag of Words outperformed in finding similar symptoms and created word embeddings from the data set. We normalized the data using these word embeddings.

Apriori algorithm of Data mining is used to generate frequent symptom sets by setting threshold as min-support and max-confidence. Data mining is a cross-functional research area, can be the basis for many applications. Decision makers can obtain valuable information from their data sets with its help. It is based on databases, mathematics, and other computer science issues.[12]. Associative classification (AC) [17]is an innovative research area that emerged in the last ten years from the integration of classification and association principles. [18]. To discover rules, AC utilizes primarily rules of association finding methods to train on an input data set. It then adds methods to construct a classifier and forecast the test data. AC has been used recently in several domains and security applications, including as text mining [19], defect prediction [20], systems for recommendation [21], and the web phishing detection [22]. Each rule has two main parameters: confidence and support. In this research, the rules generated from the input dataset based on the minimum support and maximum confidence. The generated rules were sorted from highest rank. Each rule is examined with parent's opinion then the generated frequent sets will be considered for the classification.

The frequent symptoms list generated from association rules are classified by five simple but efficient classification algorithms such as Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), ADA Boost (AB) [23] and Latent discriminant Analysis (LDA). Random Forest (RF) Ensemble learning method using multiple decision trees for classification and regression tasks. Decision Tree (DT) Hierarchical tree structure to make decisions based on features for classification and regression. Support Vector Machine (SVM) classify the data with the help of hyperplanes. ADA Boost (AB) Adaptive boosting technique that combines weak classifiers to improve overall prediction performance. Latent Discriminant Analysis (LDA) Statistical method for pattern recognition and dimensionality reduction in feature space. We applied evaluation metrics such as Matthews correlation Coefficient (MCC), Accuracy, F1 Score, Precision and Recall for each algorithm. We found Decision Tree (DT) outperformed. Hence, considered these into our proposed model to detect Autism disorder type.

Generally, ASD found with the children of age 1- 2 years. The diagnosis of ASD with children and therapy based on children behavior and parent's opinion. In rural area it is found that many parents are not able to express the proper symptom of their children because of illiteracy and lack of time to spend with the children to observe their activities.[24][25][26] To detect Autism type from the casual language of the parent with any random one or two symptoms is a big challenge.

- We resolve the issue by collecting an Autism dataset from "Total Solution Rehabilitation Society, Hyderabad, India." with 182 instances and preprocessed the redundant and unstructured data with NLP techniques. Word embeddings are created by NLP techniques such as Bag of words, Bag of N-grams, TF-IDF and One-Hot encoding to normalize the data and we observed Bag of Words performed well to provide best cosine similarity index for similar symptoms.

- We applied Apriori algorithm to create association rules based on min-support and max-confidence to generate frequent symptom sets by considering the preliminary input from the parent.

- These results are classified with the help of effective classifiers i.e. Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), ADA Boost (AB) and Latent discriminant Analysis (LDA) and identified Decision Tree as the best classifier by evaluating the predictions generated by all classifiers with the help of evaluation metrics i.e. Accuracy, Precision, Recall, F1 Score and Matthews correlation Coefficient (MCC).

The Major contributions of this paper are as follows:

- Overcoming the challenges of casual language of parent and similar kind of symptoms by converting them to formal symptoms through Bag of Words, NLP technique with similarity Index.

- Detecting Autism from any one random symptom is too challenging which we overcome by applying association rules of ML techniques.

- Detecting Autism type with a symptom which is available in various Autism categories is also a challenge and we overcome this by using Decision Tree, Machine Learning classification algorithm and Association rules. We achieved 99% accuracy on predicting to a particular type of Autism.

- We developed an Integrated Model of Natural Language Processing and Machine Learning (INM) for the detection of Autism disorder type in children.

II. LITERATURE SURVEY

A scoping analysis using PRISMA guidelines discovered an increase in the usage of technology-based ASD detection methods that analyze multimodal data using ML and DL approaches. This demonstrates the potential of technology to improve current ASD screening and diagnosis procedures. [27]. AI-based innovations have accelerated ASD diagnosis [28], [29], boosted clinical capacity, and improved access to early intervention programs [30]. Thermal imaging and non-invasive tools can help identify emotional states, particularly negative ones. The research produced a thermal-based classifier for the affective state of autistic youngsters.[31]. Semantic relationships between biomedical documents are found through a variety of NLP techniques. Performing statistical computation on a huge number of documents, the Latent Semantic Analysis (LSA) methods select concepts and provides the semantic meaning of words [32]. LSA lowers the size of the vector space model while identifying the hidden classes [33]. Furthermore, LSA suffers from challenges with mathematical complexity and requires a strong statistical approach. [34]. Moreover, redundancy complicates text mining [35]. The high dimensionality issue is brought by the hundreds to thousands of medical features identified in biomedical texts [36]. It is very difficult to handle high-dimensional data. Because handling this type of data gives a lot more parameters to the model and increases the level of complexity, handling this type of data can be troublesome. Moreover, problems regarding noise and sparsity are generally associated with higher-order dimensional data. By using Latent Semantic Analysis (LSA), clinical case summaries can be automatically evaluated in real time [37]. The duplicate data can be removed from the biomedical data through Topic modelling technique.[38]. One of the most popular method Term Frequency (TF) [39] is used in this paper. M. ALHAWARA et.al [40] used word embeddings. word embedding is used to represent text concepts as real-valued vectors in vector space that retain the syntactic and semantic characteristics of the text. Few studies are classifying Arabic text documents use word embedding; instead, traditional text representation methods like bag-of-words and TF-IDF weighting are used. In our proposed method also bag of words works well on our real time Autism dataset. Arabic text categorization has already been done using traditional machine learning techniques, with better results. Many researchers studied text classification problem and traditional ML techniques are applied to get effective results [41]-[45].

Classification and association rule mining are similar activities as the former uses the values of the attributes in a data set to construct a model which can be utilized to predict the response variable of a test case, while the latter uses the values of the attributes to search for hidden relationships among the values in a data set. As an outcome, integrating the two tasks yields Associative Classification (AC) [46], which utilizes association rule techniques to extract classifiers that have "IF-THEN" rules.

In Common there are three steps involved in Associative classification model:

- Understanding the rules (Training): To produce the rules, the algorithm analyses the data.
- Design a rules ranking strategy: During the process of sorting based on parameters like rule confidence and support, the extracted candidate rules are assessed using the training data set in order to choose which rules are to be included in the classifier.
- Classification of test results: This stage uses the classifier's rules to predict the test data's class values. Prediction accuracy is used in this step for evaluating the classifier's prediction capability in terms of Accuracy, Precision, recall, F1-Score and Matthews Correlation Coefficient (MCC).

Over the past decade, numerous research studies, such as those reported in [47], [48], have documented the applicability of AC for a range of applications. These include credit card scoring, fraud detection, online security, bioinformatics, medical evaluation, text classification, and more. In addition to its classifiers' high predicted accuracy, the main reasons for the widespread use of this classification approach are its rules' readability and simplicity. Our proposed algorithm utilizes the Apriori candidate generation method to discover class association rules from data set. These rules used for the prediction of Autism with the help of classifiers. To quickly identify and diagnose ASD as well as other health conditions like diabetes, stroke, and heart failure, multiple studies utilizing Machine Learning (ML) techniques in recent years [49], [50], [51]. The authors of [52] created predictive models for kids, teens, and adults by combining the Random Forest (RF) and Iterative Dichotomiser 3 (ID3) algorithms. Used Decision Trees (DT), Support Vector Machines (SVM), and Logistic Regression (LR) as prognostic and diagnostic classifiers for ASD [53]. In our propose model, we have used Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), Ada Boost (AB) and Linear Discriminant Analysis(LDA) to classify the approved symptoms from the parent to a proper Autism type. We will discuss the accuracy of each algorithm in the results section. In [54], the authors studied the accuracy of the classifiers for adult ASD prediction using Linear Regression (LR), K-Nearest Neighbors (KNN), LDA, Classification and Regression Trees (CART), Naive Bayes (NB) and SVM. They even studied the impact of normalization on classification findings. In [55], a machine learning model for Autism identification through rule induction was presented. This approach involves limited comparison and testing on a single dataset. An ML strategy for classifying Autism was developed by the authors in [56] using LR analysis; however, this approach also requires comprehensive validation and comparison. After carefully examining Autism data, the authors in [57] observed that attention deficit hyperactivity disorder (ADHD) can detect ASD using five of the totals of 65 specifications. Incorporating behavioral data, the authors of [53] constructed an RF-based model for the prediction of ASD in 2019. Further, the authors in [58] identify children with ASD between the ages of 4 and 11 by using the LDA and KNN methods. To classify ASD and Parkinson's disease, the authors in [59] used ensemble machine learning methods such as fuzzy K-nearest neighbor (FKNN) kernel support vector machines (KSVM) and random forest (RF). S. M. MAHEDY HASAN et.al used eight classification algorithms such as, DT, ADA BOOST, RF, K-Nearest Neighbor, LR, Gaussian Naive Bayes (GNB), SVM, and LDA to classify feature scaled databases [60]. In the literature, we found most of the papers used children's behaviors, brain image and activities. We proposed a model which will not depend on complete behavior or activities of the children. Our INM model can diagnose autism with any one random symptom.

The remaining part of the paper is arranged as follows: Section III presents the proposed INM model, Section IV discusses the experimental result analysis, Section V summarizes the paper and concludes.

III. PROPOSED INTEGRATED NLP AND ML (INM)MODEL

In this Section, we have discussed our proposed INM model. As shown in the Fig. 1, the proposed model detects the Autism disorder type from preliminary symptom provided by the parent. We have collected 182 Autistic children's data from "Total Solution Rehabilitation Society, Hyderabad, India.". The collected data is unstructured and redundant with casual language as shown in Table I. The dataset is not reliable for building an effective Machine Learning model. Many symptoms in the dataset have similar significance and need to be normalized. We used NLP techniques with cosine similarity measure to replace with the single symptom. Generated association rules from the normalized data. A random symptom is collected from the parent to find frequent_symptoms_sets with the help of min_support and max_confidance of Association rules. Then identify autism type with a ML classification algorithm.

A. COLLECTED DATASET AS CORPUS

We considered our collected dataset as a corpus, which consists of 182 Autistic children's symptoms. The dataset having the attributes gender, age, symptoms, and disorder type as shown in Table I. The dataset contains noise with casual languages, words having similar meanings but distinct words which needs to be normalized.

B. TEXT PRE -PROCESSING

Preprocessing dataset is an important phase in text mining. Stop words, special characters, word variations, punctuation, and other noise can be found in biomedical text dataset. Therefore, the following techniques are applied to avoid the above noise. Algorithm 1 explains the preprocessing steps.

1) CONVERSION OF DATASET TO LOWER CASE

The noise prone biomedical datasets are changed to lowercase to eliminate word differences. Let us consider an example symptom e.g. "Abnormal sensitivity to touch" and "abnormal Sensitivity to Touch" should be considered same as "abnormal sensitivity to touch" by converting their case.

2) TOKENIZATION

Tokens, or words, are created from text data. From the text data, the meaningful keywords are extracted by tokenization. The tokens serve as the input for the next process. e.g. "abnormal sensitivity to touch" as "abnormal"," sensitivity"," to" and "touch".

3) STOP WORDS

The text datasets have been cleared from the stop words. A document may have a few words that are frequently used, but they are essentially unimportant because they are just used to bind words together in a sentence. The general opinion is that stop words do not work well in textual documents, as stop words appear in documents so often, text mining becomes an obstacle to understanding the content of the document. Stop words include frequently used terms like "and", "are", "this",

" to" and so on. They are not useful in grouping and classifying documents. As a result, all stop words were eliminated from the biomedical text dataset. e.g. "abnormal sensitivity to touch", here the stop word is "to" which does not have any meaning but only a connector between words. They must thus discard from the dataset.

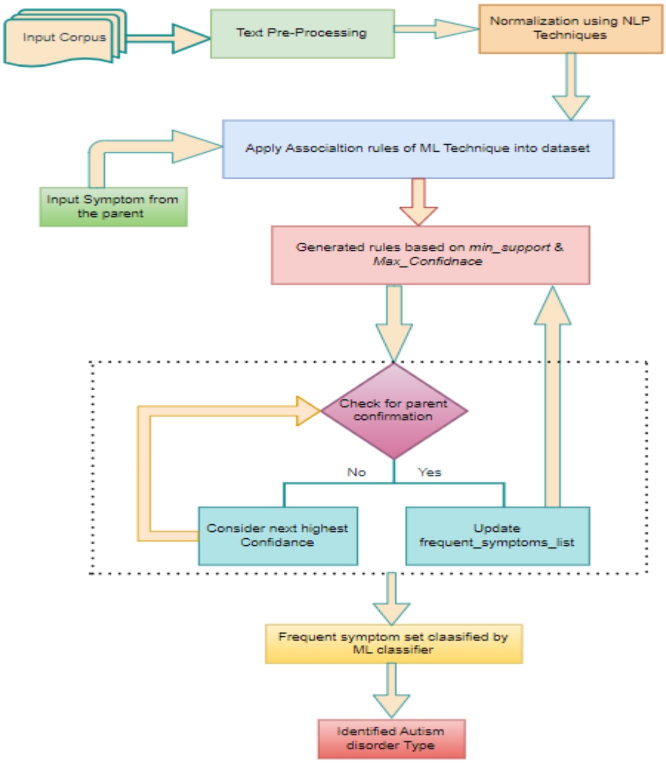
Input Corpus
Text Pre-Processing
Normalization using NLP Techniques
Apply Associaltion rules of ML Technique into dataset
Input Symptom from the parent
Generated rules based on *min_support* & *Max_Confidnace*
Check for parent confirmation
No
Yes
Consider next highest Confidance
Update frequent_symptoms_list
Frequent symptom set claasified by ML classifier
Identified Autism disorder Type

FIGURE 1. Integrated NLP and ML (INM) model for Autism detection.

TABLE I

SAMPLE REALTIME AUTISM DATASET

Child S.No	Gender	Age	Symptoms	Disorder Name
C9	M	6	improper development sensory skill, Excitement, laughing, drooling, underdeveloped speech, underdeveloped language skills, Unusual social behaviour	Pervassive developmental disorder
C1	M	3	Poor sitting tolerance, Involuntary hand movements	Rett Syndrome
C2	M	5	can follows 2 step instructions, Stops talking gradually, Lose the ability of playing, Oppose physical contact, Unusual social behaviour	Pervassive developmental disorder

C3	M	5	underdeveloped language skills, follow instructions, pitch low in the center, Unusual social behaviour, Lose the abilities of motor	Rett Syndrome
C4	F	5.5	Abnormal in sensory experiences, stubbornness, poor eye contact, underdeveloped speech	Pervassive developmental disorder
C5	M	4	follows 2 step instructions, repetitive words when instructed, Oppose physical contact	Asperger's Syndrome
C6	F	3.5	speech delay, hyperactivity, stubborn, self-talk, poor sitting tolerance, may have abnormal communication, Abnormal sensitivity to taste, Abnormal sensitivity to sight, Abnormal sensitivity to touch	Childhood disintegrative disorder
C7	M	3	follows instructions, does not accept change, Abnormal sensitivity to taste, Abnormal sensitivity to sight, Abnormal sensitivity to touch	Childhood disintegrative disorder
C8	M	7	screaming, repetition of words, hallucinations in sensory experiences	Rett Syndrome
C10	M	6.7	self-stimming, self-talk, Persistent, repetitive actions such as opening and closing doors, Abnormal sensitivity to taste, Abnormality with sight, Abnormality to touch	Rett Syndrome
C11	M	5	Poor sitting, Involuntary hand movements, breathing problems	Pervassive developmental disorder
C12	F	4	self-talk, irritation, crying behaviour, spends too much of time on 1 topic, Struggele to set body language, Hobbies interfere with daily living	Pervassive developmental disorder
C13	M	8	poor eye contact, poor attention, Unusual social behaviour, Communication problems	Pervassive developmental disorder
C14	M	6	no sitting tolerance, smaller head size, Difficulty in breathing	Pervassive developmental disorder
C15	M	3.7	Poor Sitting, Poor Waiting, Poor Eye Contact, Speech problems, Loss of mobility	Pervassive developmental disorder

4) STEMMING

Stemming improves text analysis in such a way that to identify patterns and relationships within textual data and helps to reduce the dimensionality so that information can be retrieved efficiently. For e.g. from symptom “abnormal sensitivity to touch” "abnormal" is to be stemmed to "abnorm", "sensitivity" is to be stemmed to "sensit", and "touch" remain unchanged. As Algorithm 1 illustrates from NLTK module of python programming used stopwords, SnowballStemmer and RegexpTokenizer to tokenize, remove stop words and create stems for the observations of each autistic child from the dataset. As a result, pre-processed observation retrieved from this process. e.g. “abnorm sensit touch “.

Algorithm 1. Procedure for pre-processing Dataset

```
Input: ASD dataset as corpus
Output: Set of pre-processed Observations
1. stopwords ← list of stop words
2. data ← ASD_data
3. symptoms_set ← {}
4. For each Observation_list in data do
   a. For each observation in observation_list do
      i. tokens ← tokenize_observation(observation)
      ii. preprocessed_observation ← “”
      iii. For each word in tokens do
         1. word ← lowercase(word)
         2. If word not in stopwords then
            a. word ← stem_word(word)
         3. End If
         4. preprocessed_observation ← preprocessed_observation + word
      iv. End For
      v. Add preprocessed_observation to symptoms_set
   b. End For
5. End For
```

C. NORMALIZATION USING NLP TECHNIQUES

The collected data set consists of many symptoms with similar significances, which can be replaced with a single symptom as per doctor’s opinion. So, the normalized dataset can help us to build an efficient Machine Learning model. For e.g. “delayed social speech development”, “delayed speech”,” speech delay” having similar meaning, and can be replaced with a unique word like “communication disorder”. So, text normalization plays a vital role to remove redundancy and to maintain consistency in the biomedical dataset. To get the cosine similarity index between two similar words in our dataset we experimented with various NLP techniques such as Bag of Words, Bag of N-grams, TF-IDF, One-Hot Encoding on pre-processed observations as shown in Fig. 2. The range of cosine similarity index is from -1 to 1. 1 indicates that the two words are most similar, 0 indicates no similarity and -1 indicates complete dissimilar. The experimental result shows that Bag of Words technique gives best performance and our proposed INM model used this Bag of Words technique for finding best similarity index. The detailed description is here:

1) BAG OF WORDS

The Bag of Words is a popular technique in NLP adopted for text representation [13]. It involves data representation for text by means of a bag holding words or as a multiset, which discards word order and grammar but keeps multiplicity. In mathematical terms, Bag of Words model is represented by the following: For each term in the document, calculate the term frequency as:

$$Tf(t,d) = count\ of\ t\ in\ d \tag{1}$$

Where $Tf(t,d)$ is term frequency of term *‘t’* in document *‘d’*. A vector representing the document in a high-dimensional space is created using this term frequency count, with each dimension denoting a distinct word from the corpus lexicon.

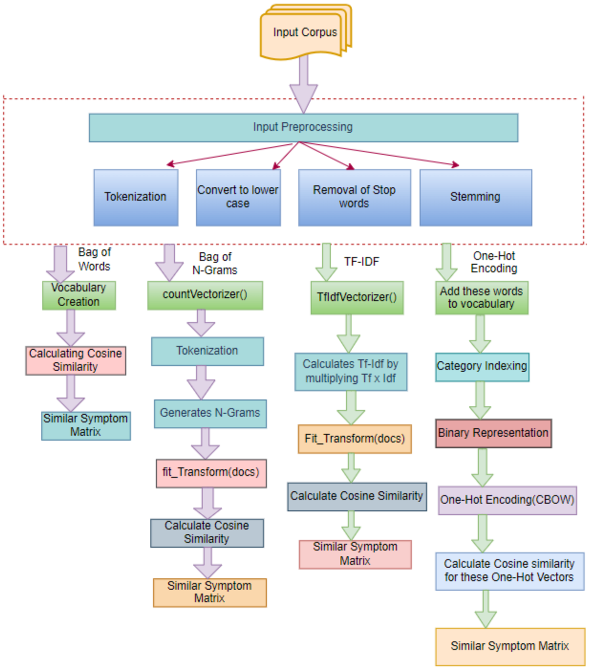

FIGURE 2. An Experimental model for finding best similarity index.

2) BAG OF N-GRAMS

The Bag of N-Grams model is an extension model of Bag of Words. Text data is represented as an unordered collection of its successive components (n-grams) [14]. Here, ‘n’ is number of words. For e.g. Bi-grams for the term "Communication problems” is represented as "Communication" and “problems" treated as a single unit. The Bag of N-Grams model's mathematical representation is comparable to that of the Bag of Words model. However,

n-grams are used in place of single words. In the document *(d)*, the term frequency *(tf (n-gram, d))* is computed as the number of times an n-gram appears. The following equation representing term frequency for the Bag of N-Grams model:

$$tf\,(n\text{-}gram,\,d) = n\text{-}gram\ count\ in\ d \qquad (2)$$

A vector which represents the document in a high-dimensional space is created using this word frequency count. Where each dimension denoting a distinct n-gram from the corpus vocabulary.

3) TF-IDF

TF-IDF stands for Term Frequency-Inverse Document Frequency which indicates the importance of a word for a document relative to a corpus or a collection words [15]. The value of TF-IDF increases with the number of words exists in a document, which allows removing common words that generally appear frequently.

Here is the mathematical equation for TF-IDF:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t,D) \qquad (3)$$

Here,

- t- stands for word or term
- d- stands for a document
- D- stands for group of documents

- $TF(t,d)$ is the term frequency, the number of times the term (t) presents in document (d)
- $IDF(t,D)$ is the inverse document frequency, calculated as:

$$IDF(t,D) = log(|\ \{d \in D: t \in d\}\ ||\ D\ |) \tag{4}$$

In the above equation, $\{d \in D: t \in d\}$ is the number of documents where the term *(t)* appears in document *(d)* and $(|D|)$ is the total number of documents in the input corpus.

4) ONE-HOT ENCODING

One-hot encoding method used to obtain binary vector representation from categorical data [16]. A vector is used to represent each category, with one element like "Come" (set to 1) and all other elements like "Go" (set to 0).

The one-hot encoding for a category (c) in a document (d) can be mathematically expressed as a vector (V) of length (N) for a categorical variable with (N) categories, where:

$$V_i = \begin{cases} 1 & if\ c\ is\ the\ i^{th}\ category \\ 0 & otherwise \end{cases} \tag{5}$$

There is a unique vector (V) for each category (c) such that all other elements of (V) are 0 and only one member is 1, which corresponds to the category (c).

An illustration using the categories (underdeveloped, language, skills) is provided here:

The one-hot encoded vector for the category (underdeveloped) may be ([1, 0, 0]). The one-hot encoded vector for the category (language) may be ([0, 1, 0]). The one-hot encoded vector for the category (skills) may be ([0, 0, 1]).

After analyzing the NLP algorithms, we found Bag of Words and One-Hot Encoding techniques were more efficient in creation of word embeddings for our real time dataset. In our proposed INM model, we preferred Bag of Words technique over One-Hot Encoding technique, as extensive sparse matrices were created to form the vectors for each symptom in One-Hot Encoding technique. Algorithm 2 illustrates the procedure to create word embeddings with outperformed Bag of Words technique. From *sklearn* module called *CountVectorizer()* to vectorize the observations and identified the similarity between these vectors using *cosine_similarity* . Fig. 3 shows the sample vocabulary, 'poor' is a word added to the bag with unique index '83'. Our dataset contains 'poor' word many times but all of them are having same index '83'.

Algorithm 2. Procedure to create similar symptoms with cosine similarity

```
Input: ASD dataset, pre-processed Symptoms list
Output: Word Embeddings
1. symptom_embeddings ←{}
2. vectorizer←CountVectorizer()
3. Set threshold value
4. X ← vectorizer.fit_transform(pre-processed_symptoms_list)
5. cosine_sim <- cosine_similarity(X)
6. For each row in the len(cosine_sim) do
   a. key_symptom ← pre-processed_symptoms_list[row]
   b. If key_symptom not in symptom_embeddings then
      i. s_list ←[]
      ii. For each col in the len(cosine_sim) do
         1. If cosine_sim [row][col] > threshold then
```

```
  a.   s_list ← s_list + pre-processed_symptoms_list [col]
  2.   End If
iii.   End For
 iv.   symptom_embeddings[key_symptom] ← s_list
  c.   End If
  7.   End For
```

```
Vocabulary: {'poor': 83, 'sit': 103, 'toler': 120, 'involuntari': 56, 'hand': 44, 'movement': 68, 'follow':
38, 'step': 113, 'instruct': 53, 'stop': 114, 'talk': 117, 'gradual': 42, 'lose': 63, 'abil': 0, 'play': 82, 'oppos':
77, 'physic': 80, 'contact': 19, 'unusu': 126, 'social': 110, 'behaviour': 7, 'motor': 67, 'skill': 106,
'stubborn': 116, 'eye': 35, 'underdevelop': 124, 'speech': 111, 'languag': 58, 'abnorm': 1, 'sensit': 98,
'tast': 118, 'sight': 102, 'touch': 123, 'repeat': 89, 'word': 130, 'reduc': 88}
```

FIGURE 3. Sample vocabulary of Bag of Words

D. ASSOCIATION RULES GENERATION

The Apriori algorithm identifies frequent-item-sets to discover hidden relationships between items in any kind of dataset. The sequence of steps involved to generate rules are:

1. Generate Candidate Symptom sets: In the initial stage, consider individual symptoms are considered as 1-symptom sets. Then, from frequents symptoms of length k-1 generate candidate symptom sets of length k.

2. Prune Candidate Symptom sets: Candidate Symptom sets are pruned by removing those which cannot reach the min_support threshold.

3. Calculate Support: Each candidate support is calculated by counting the number of transactions containing the symptom set.

4. Generate Frequent Symptom sets: Only the candidate Symptom sets that reach to the minimum support threshold are retained as frequent Symptom sets.

5. Repeat the Process: Steps 1-4 are repeated iteratively to find frequent Symptom sets of higher lengths until no new frequent Symptom sets can be found.

The algorithm uses support and confidence metrics to analyze the robustness of association between items. Support is the proportion of transactions containing a Symptom set, while confidence measures the likelihood of a rule given the presence of another Symptom set. By leveraging these metrics, the Apriori algorithm efficiently uncovers associations and patterns in large datasets, assisting organizations in making profitable choices based on customer behaviour and market trends. Our proposed INM model utilized this technique to generate frequent_symptom_sets to identify the hidden relationships between symptoms(observations) in our dataset. A frequent_symptom_set is a collection of symptoms that occur together frequently in the dataset above a certain support threshold. From *mlxtend* library of python used *TransactionEncoder*, *apriori*, *association_rules* modules to generate rules. Here, *TransactionEncoder* converts symptoms into transaction data. *Apriori* module analyzes this transaction data to identify frequent_symptom_sets and association rules.

Algorithm 3 illustrate the procedure to generate frequent_symptom_set from the preliminary symptom *p*. Initially *p* is added to frequent_symptom_set. Association rules are filtered when *p* matches with antecedents and then sorted in descending order based on confidence. Then extracted consequents from sorted Association rules. Afterwards with the confirmation of parent the highest confidence consequent is added to frequent_symptom_set. With the frequent_symptom_set generated, we repeated the process as above and updated frequent_symptom_set until there are no more matching Association rules. Finally, the entire frequent_symptom_set generated by the Association rules is classified through Machine Learning techniques.

E. CLASSIFICATION ALGORITHMS IN MACHINE LEARNING

To detect the Autism disorder type, we classified the frequent_symptom_set using several ML classifiers such as ADA Boost (AB), Random Forest (RF), Linear Discriminant Analysis (LDA), Decision Tree (DT) and Support Vector Machine (SVM). We experimented with the following algorithms:

Algorithm 3. Procedure to generate symptoms_set from the preliminary symptom.

```
Input: preliminary_symptom, association_rules (antecedents → consequents), ASD_dataset
Output: Generated symptoms with Autism type
1. Read preliminary_symptom and ASD_dataset
2. generated_symptoms_set ← preliminary_symptom
3. Filter association_rules where preliminary_symptom matches with antecedents
4. Sort filtered association_rules by confidence in descending order
5. Extract consequents from sorted association_rules
6. While consequents available do
    a. For each consequent in consequents do
        i. If consequent approved by parent then
            1. generated_symptoms_set ← generated_symptoms_set + current consequent
        ii. End if
    b. End For
7. End While
8. Repeat from 3 to 7 with generated_symptoms_set until there is no matching consequents.
```

ADA BOOST

AdaBoost is a hybrid ensemble model which is used many weak classification algorithm to enhance the performance of classification [61]. It assigns the weights based on the precision of the previous training to retrain the algorithm. Any random subsets of the training sets are used to train the weak classification model. Each instance is assigned by a weight in the classification model.

The below equation is a set of weak classifiers:

$$H(x) = Sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{6}$$

$H(x)$ represents the final model's output by combining weak classifiers, $h_t(x)$ defines for input x the output of the classifier t, and α_t specifies classifier's weight. $\boldsymbol{\alpha_t}$ is determined as follows.

$$\alpha_t = \frac{0.5 * ln\,(1 - E)}{E} \tag{7}$$

Where E reflects as error rate. The below mentioned equation is helpful to modify the weights of every training sample label pair $(\boldsymbol{x_i}, \boldsymbol{y_i})$.

$$D_{t+1}(i) = \frac{D_t(i) exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t} \tag{8}$$

where D_{t+1} specifies the modified weight, D_t defines the weight of the previous level, and D_t denotes the sum of all weights.

1) RANDOM FOREST (RF)

Random Forest is the example of ensemble classification approach and it is also a Decision Tree (DT) based algorithm. It uses the DC (Divide and Conquer) approach with corpus to produce many Decision Trees which are known as to be Forest [62]. It is mainly based on two stages. Firstly, it creates a collection of Decision Trees and finds the prediction for each DT which was created in the first stage. The sequence of steps involved in RF are explained below.

1. From the training sample choose the random samples.
2. For each training sample create DT's.
3. *'N'* needs to be selected to specify the number of DT's
4. Repeat the 1 and 2 steps.
5. Based on majority voting assign a class value and find prediction for each DT.

2) DECISION TREE (DT)

DT develops a model that predicts class values from the top-down approach by training data-driven decision-making rules [63]. To identify the best feature, information gain approach is used. Let us assume that, *Pi* as the probability The existence of *xi* is in *D* is predicted by the ratio of |*Ci, D*| to |*D*| for a class *Ci*. To categorize occurrences in the dataset D, the necessary information is required, which is obtained from the below equation:

$$Info(D) = -\sum_{i=1}^{m} P_i log_2 \left(P_i\right) \tag{9}$$

Where *Info(D)* - To identify $C_i$of an instance, $x_i \in D$ the average amount of information required. The goal of Decision Tree is to divide D into sub data $D_1, D_2 \ldots \ldots D_n$. The below equation assesses the $Info_A\,(D)$:

$$Info_A\,(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info\,\left(D_j\right) \tag{10}$$

The gain value is calculated with below equation:

$$Gain\,(A) = Info\,(D) - Info_A\,(D) \tag{11}$$

3) SUPPORT VECTOR MACHINE (SVM)

Both Nonlinear and linear data is classified by using SVM. Even with high dimension data and mapping of nonlinear it performs well. To make difference from class to class it identifies a boundary. We are using a Radial Basis(RB) as a kernel function, By minimizing the higher limit of test error which was predicted SVM defines weights, thresholds and centers [64],[65]. The RB function is denoted with the below equation.

$$K(x, x') = \exp\left(-\frac{(\|x - x'\|)^2}{2\delta^2}\right) \tag{12}$$

Here, δ is a free parameter and $(\|x - x'\|)^2$ is squared Euclidean Distance.

4) LINEAR DISCRIMINANT ANALYSIS (LDA)

To estimate the probability Bayes theorem is used in LDA. It is also a dimension reduction method which is used for classification [66]. Let us take k classes and n training samples determined as$\{a_1, a_2 \ldots \ldots \ldots a_n\}$ with classes $z_i \in \{1 \ldots \ldots k\}$. The Gaussian distribution is taken as $\phi(a \mid \mu_k, \Sigma)$ in every class. The approximation is mentioned as below:

$$p_k = \frac{\sum_{i=1}^{n} l * (z_i = k)}{n} \tag{13}$$

$$\mu_k = \frac{\sum_{i=1}^{n} a_i * l * (z_i = k)}{\sum_{i=1}^{n} l * (z_i = k)} \tag{14}$$

$$\Sigma = \frac{\sum_{i=1}^{n} (a_i - \mu_{zi})(a_i - \mu_{zi})^T}{n} \tag{15}$$

Here, p_k is prior probability, μ_k is mean of all the classes, Σ is class means sample covariance.

After analyzing the ML techniques, we found Decision Tree algorithm is more efficient in classification of frequent_symptom_set to appropriate autism type.

IV. EXPERIMENTAL RESULTS ANALYSIS

A. EXPERIMENTAL SETUP

To continue with our experiment, the open-source service Kaggle, supplied by Google LLC, is used. Data pre-processing, normalization, Association rule generation and classification operations are completed using Python's *pandas*, *numpy*, *nltk*, *scikit-learn* and *mlxtend* modules. For visualization Python’s *matplotlib* module is used.

B. DATASET NORMALIZATION WITH NLP TECHNIQUES

The collected dataset normalized with the help of word embeddings which is a crucial task. We analyzed the similarity of the symptoms provided in the dataset using various NLP techniques such as Bag of Words, Bag of N-grams, TF-IDF, and One-Hot Encoding in the process of creating word embedding. Fig. 4 to Fig. 7 shows 3D visualization of cosine similarity between a symptom *S1* i.e. ‘'improper development sensory skill’ with other symptoms *S9, S35, S39, S44, S60, S72* which are 'improper development cognitive skill', 'improper development social skill', 'improper development academic skill', 'improper development visual skill', 'improper development behaviour skill', 'unusual social behaviour' from a set of redundant symptoms. Table II represents the detailed comparison of cosine similarity between symptom *S1* and other symptoms (*S9,S35,S39,S44,S60,S72*) with different NLP techniques. Whereas Fig. 8 shows the 3D visual representation of comparison of cosine similarity of symptoms with different NLP techniques.

Bag of Words, represents the similarity between *S1* and *S9* is 70.53^o shown in Fig. 4. Bag of N-grams shows 78.46^o in Fig. 5. TF-IDF shows 76.86^o in Fig. 6. One-Hot Encoding shows 70.53^o in Fig. 7. The less cosine angle represents the more similarity between the symptoms. In this analysis, we found Bag of Words and One-Hot Encoding techniques were more efficient in creation of word embeddings for our real time dataset.

We preferred Bag of Words technique over One-Hot Encoding technique, as extensive sparse matrices were created to form the vectors for each symptom in One-Hot Encoding technique. Based on the similarity score calculated by cosine similarity using the bag of words techniques, we grouped similar symptoms and assigned a key symptom which is recommended by doctor and created the word embeddings as shown in Table III. Then the redundant dataset is normalized by replacing similar symptoms with key symptom as shown in Table IV.

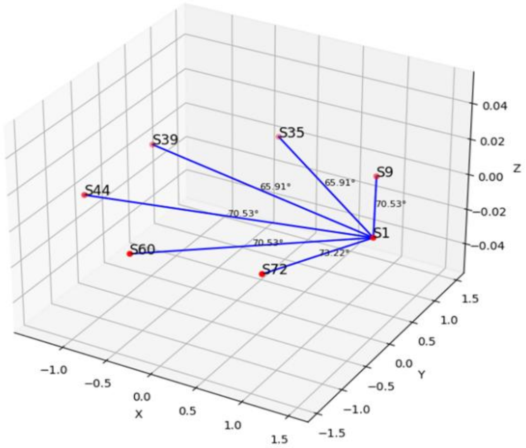

FIGURE 4. **3D visualization of cosine similarity using Bag of words.**

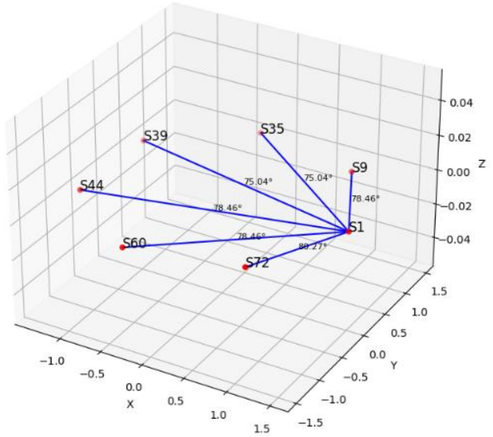

FIGURE 5. **3D visualization of cosine similarity using Bag of N-grams**

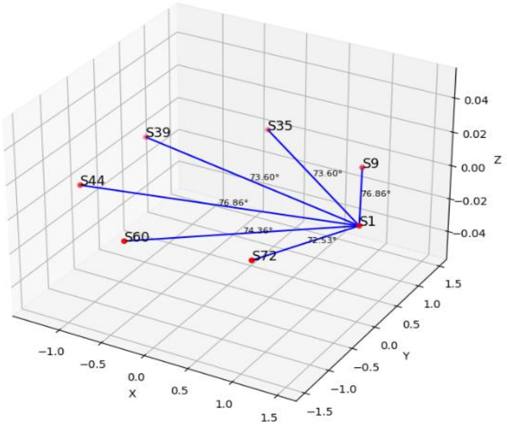

FIGURE 6. **3D visualization of cosine similarity using TF-IDF**

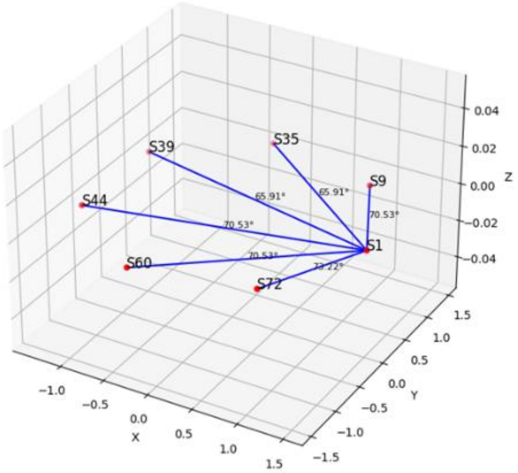

FIGURE 7. 3D visualization of cosine similarity using One-Hot encoding

TABLE II

COMPARISON OF COSINE SIMILARITY OF SYMPTOMS WITH DIFFERENT NLP TECHNIQUES

Symptoms	Bag of words	Bag of N grams	TF-IDF	One hot encoding
S9	0.333333	0.2	0.227323	0.333333
S35	0.408248	0.258199	0.282345	0.408248
S39	0.408248	0.258199	0.282345	0.408248
S44	0.333333	0.2	0.227323	0.333333
S60	0.333333	0.2	0.269592	0.333333
S72	0.288675	0.169031	0.300278	0.288675

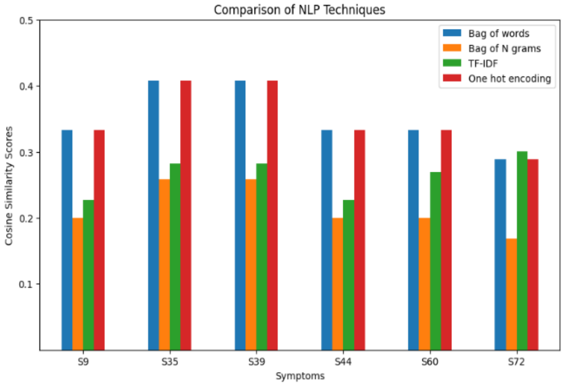

FIGURE 8. Comparison of cosine similarity of symptoms with different NLP techniques

TABLE III

SAMPLE DICTIONARY CREATED FROM THE BAG OF WORDS RESULTS

Key Symptom	Similar Symptoms
communication disorder	['underdeveloped language skills', 'speech delay', 'speech impairment', 'underdeveloped speech', 'speech problems']
hallucination	['hallucination in hearing', 'hallucination to smell', 'hallucination to see', 'hallucinations in sensory experiences']
nonverbal communication difficulty	['struggle to set facial expressions', 'struggle to set gestures', 'struggle to set body language']
developmental delay	['improper development sensory skill', 'improper development cognitive skill', 'improper development social skill', 'improper development academic skill', 'improper development visual skill', 'improper development behavior skill', 'unusual social behavior']
hypersensitivity	['abnormal sensitivity to sight', 'abnormal sensitivity to touch', 'abnormal sensitivity to taste']
self-regulation difficulties	[‘Poor Sitting’, ‘Poor Waiting’, ‘Poor Eye Contact’, ‘Poor sitting tolerance’, and ‘poor attention’]

TABLE IV

BEFORE AND AFTER REPLACEMENT OF SYMPTOM PHRASES

Patient	Before Normalization	After Normalization
C3	underdeveloped language skills, follow two step instructions, pitch low in the center, Unusual social behavior, Lose the abilities of motor	**communication disorder**, follow two step instructions, pitch low in the center, **developmental delay**, lose the abilities of motor

C4	Abnormal in sensory experiences, stubbornness, poor eye contact, underdeveloped speech	Abnormal in sensory experiences, stubbornness, poor eye contact, **communication disorder**
C7	follows instructions, does not accept change, Abnormal sensitivity to taste, Abnormal sensitivity to sight, Abnormal sensitivity to touch	follows instructions, does not accept change, **hypersensitivity**
C8	screaming, repetition of words, hallucinations in sensory experiences	screaming, repetition of words, **hallucinations**
C15	Poor Sitting, Poor Waiting, Poor Eye Contact, Speech problems, Loss of mobility	**self-regulation difficulties**, **communication disorder**, loss of mobility

C. GENERATION OF FREQUENT SYMPTOM SETS THROUGH ASSOCIATION RULES

We applied Apriori algorithm in normalized dataset to generate Association rules as shown in Table V. Support indicates the frequency of occurrence of a symptom in the dataset and confidence represents how often consequent appear in the dataset when antecedent is present. We received preliminary symptom provided by parent and generated frequent symptom sets based on minimum support and maximum confidence of Association rules as shown in Table VI. The table shows the frequent symptom sets generated from a random preliminary symptom of some selected patients 7,13,18,26, and 28.

TABLE V

GENERATED ASSOCIATION RULES

Antecedents	Consequents	Support	Confidence
{'poor eye contact'}	{'abnormal sensory experiences'}	0.01639	0.0625
{'poor name call response'}	{'repetitive actions opening closing doors', 'developmental delay', 'persistant'}	0.03278	0.2000
{'repetitive actions opening closing doors',	{'stops talking gradually'}		0.1428

'hypersensitivity'}		0.04918	
{'restlessness', 'developmental delay'}	{'poor sitting tolerance', 'visual stimulation'}	0.06557	0.0926
{'hyperactivity', 'stubborn', 'hypersensitivity'}	{'poor sitting tolerance', 'communication disorder'}	0.01439	1

TABLE VI

GENERATED SYMPTOMS FROM A PRELIMINARY SYMPTOM

P.No	Priliminary Symptom	Generated Symptoms
7	lose ability playing	['follows 2 step instructions', 'oppose physical contact', 'communication disorder', 'developmental delay']
13	struggle set body language	['developmental delay', 'hobbies interfere daily living', 'irritated', 'spends much time 1 topic', 'communication disorder']
18	oppose physical contact	['follows 2 step instructions', 'lose ability playing', 'stops talking gradually', 'unusual social behaviour']
26	hobbies interfere daily living	['irritated', 'spends much time 1 topic', 'communication disorder', 'nonverbal communication difficulty', 'developmental delay']
28	screaming	['clapping', 'poor name response', 'communication disorder']

D. ANALYSIS ON CLASSIFICATION ALGORITHMS

The generated frequent symptom set is taken to be classified using various ML classification techniques i.e. Random Forest, Decision Tree, Support Vector Machine, ADA Boost and Linear Discriminant Analysis.

Various statistical evaluation criteria, including accuracy, precision, recall, F1-Score and Mathews Correlation Coefficient (MCC), are used to verify the findings of the experiments with all the above ML algorithms. The evaluation measures are computed using the formulas mentioned in table VII.

Accuracy evaluates a classifier's actual prediction performance. Higher accuracy shows higher prediction and fewer miss-classification. Fig. 9 represents the comparison of accuracy of all ML classifiers. We found Decision Tree algorithm delivers the best accuracy of 90.48% for the normalized real time Autism dataset.

Precision evaluates the high precision value, i.e. low false positive values and high true positive values. Fig. 10 represents the comparison of precision of all ML classifiers. After analyzing the precision values of our dataset Decision Tree provides the best precision of 96.67%.

Recall evaluates how frequently a ML model effectively identifies true positives from all the actual positive values. Fig. 11 represents the comparison of recall of all ML classifiers. After analyzing the recall values of our dataset Decision Tree provides the best precision of 91.53%.

F1-Score evaluates as the harmonic mean of the recall and precision of a ML classification model. Here both precision and recall contribute equally to the F1-score to ensure the ML model's reliability. Fig. 12 represents the comparison of F1-Score of all ML classifiers. After analyzing the F1-Score values of our dataset, Decision Tree provides the best F1-Score of 93.29%.

MCC considers the coefficients of confusion matrix which involves true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) to evaluate the degree of correlation value. Fig. 13 represents the comparison of MCC of all ML classifiers. After analyzing the MCC values of our dataset DT provides the best precision of 87.94%.

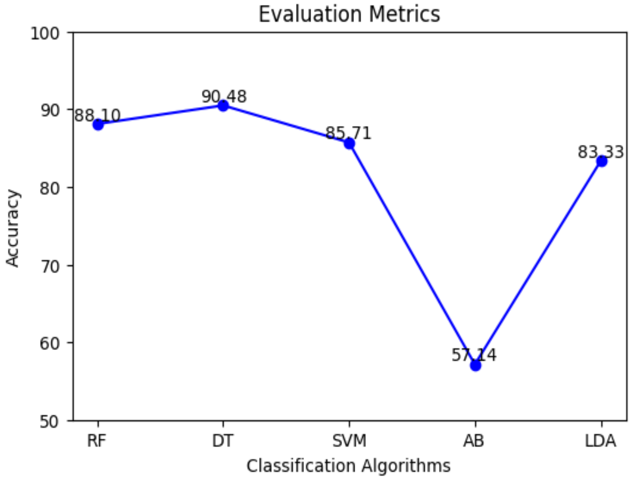

FIGURE 9. Comparison of Autism detection accuracy with ML classifiers

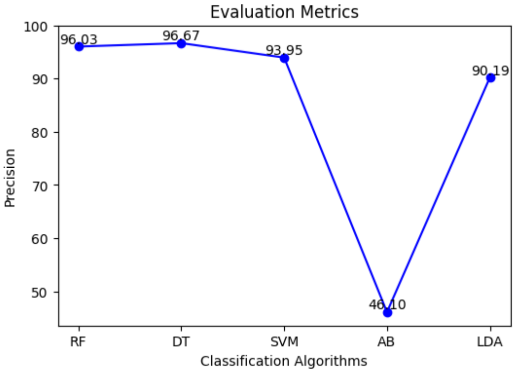

FIGURE 10. Comparison of Autism detection precision with ML classifiers

TABLE VII

DETAILED DESCRIPTION OF EVALUATION MEASURES

Name	DEFINITION	Formula
Accuracy	It evaluates the efficiency of a classification model. It indicates the percentage of correctly classified instances out of all instances analyzed.	$\text{Accuracy} = \frac{TN + T}{TN + TP + F}$ where TN denotes True Negatives, TP denotes True Positives; FN denotes False Negatives and FP denotes False Positives
Precision	It measures the correctly predicted positive instances out of all instances predicted as positive by the model.	$\text{Precision} = \frac{TP}{FP + TP}$ High precision indicates that the model is likely to be correct.
Recall	It measures the proportion of correctly predicted positive instances out of all actual positive instances in the dataset.	$\text{Recall} = \frac{TN}{TP + FN}$ A high recall means that the model can capture a significant portion of the instances that are positive in the dataset. .
F1-Score		$F1 - \text{Score} = \frac{TP}{FN + FP + 2TP}$ A high recall suggests that the model correctly identifies a large fraction of the positive cases in the dataset.

MCC	When dealing with imbalanced datasets, it takes into account all four values in the confusion matrix.	$MCC = \frac{(TP*TN-FP*FN)}{((TP+FP)(TP+FN)(TN+FP)(TN+FN}$ A high MCC value, closer to +1, indicates a highly accurate and reliable model in binary classification, whereas a lower value indicates that the model's predictions are less dependable.

Fig. 14 represents the comparison graph of different evaluation metrics of all the above classification algorithms. The overall comparison of evaluation metrics for above classifiers in tabular form are shown in Table VIII.

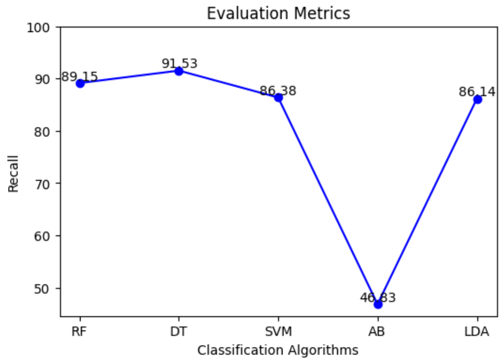

FIGURE 11. Comparison of Autism detection recall with ML classifiers

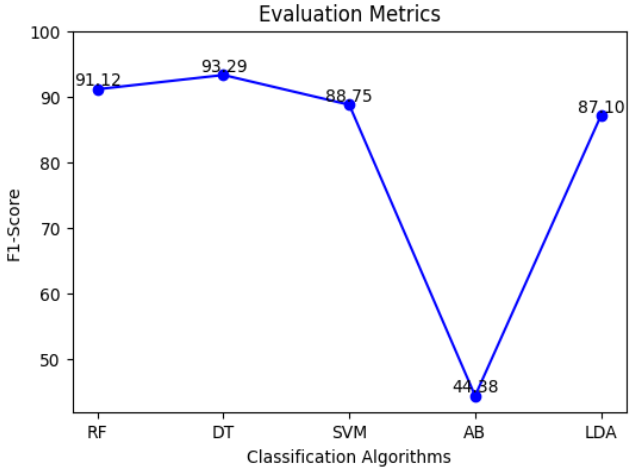

FIGURE 12. Comparison of Autism detection F1-Score with ML classifiers

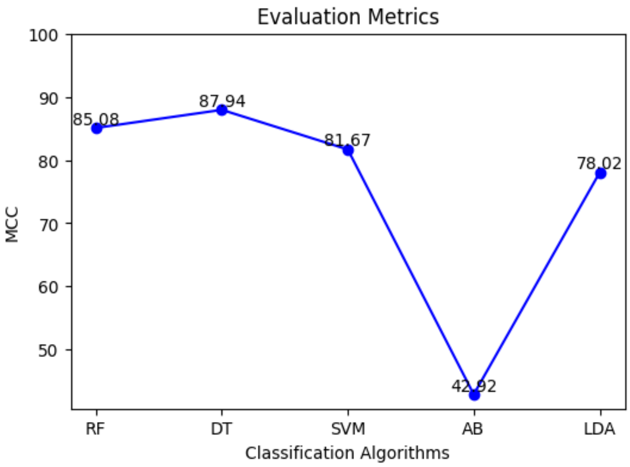

FIGURE 13. **Comparison of Autism detection MCC with ML classifiers**

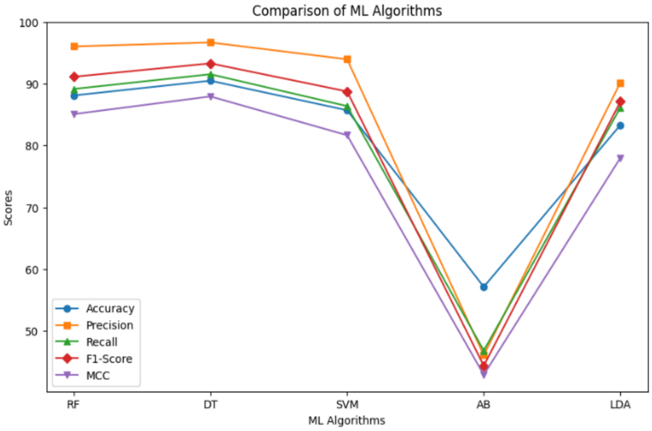

FIGURE 14. **Comparison of classification algorithms based on evaluation metrics**

TABLE VIII

COMPARISON OF CLASSIFICATION ALGORITHMS BASED ON EVALUATION METRICS

Classification Algorithms	Accuracy	Precision	Recall	F1-Score	MCC
DT	88.1	96.03	89.15	91.12	85.08
RF	90.48	96.67	91.53	93.29	87.94
SVM	85.71	93.95	86.38	88.75	81.67
AB	57.14	46.1	46.83	44.38	42.92
LDA	83.33	90.19	86.14	87.1	78.02

V. CONCLUSION

In this research work, we proposed an Integrated model of Natural Language Processing and Machine Learning (INM) for autism detection and classification from a random symptom in children. We analyzed that predictive models based on Natural Language Processing and Machine Learning techniques are advantageous tools for this work. The collected real time autism data from "Total Solution Rehabilitation Society, Hyderabad, India" is preprocessed and created word embeddings using cosine similarity to replace the similar meaning symptoms to achieve consistency in the dataset. We applied Apriori algorithm to create Association rules based on min-support and max-confidence to generate frequent symptom sets by considering the preliminary input from the parent. Afterwards we classified the generated frequent symptom sets with the help of effective classifiers i.e. Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), ADA Boost (AB) and Latent discriminant Analysis (LDA) and identified Random Forest as the best classifier by evaluating the predictions generated by all the classifiers with the help of evaluation metrics i.e. Accuracy, Precision, Recall, F1-Score and Matthews Correlation Coefficient (MCC). Subsequently, our proposed INM model can be utilized as a helpful tool for the decision-making of Autism health care medicos while examining the ASD cases. The proposed INM model is tested in the "Total Solution Rehabilitation Society, Hyderabad, India." organization with 1896 autistic children and achieved 99% accuracy on identification of ASD types. The main novelty of our model is to identify the autistic types from only one random symptom, whereas existing systems needs all symptoms to identify the autism in the children. In future, we would like to extend this work by collecting more ASD dataset, then develop a model using advanced NLP techniques like Word2Vec and Glove model to generalize the detection of Autism for any age of people and helpful for other neurologic disorders.

REFERENCES

[1] M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni, "Efficient machine learning models for early stage detection of autism spectrum disorder," Algorithms, vol. 15, no. 5, p. 166, May 2022.

[2] D. Pietrucci, A. Teofani, M. Milanesi, B. Fosso, L. Putignani, F. Messina, G. Pesole, A. Desideri, and G. Chillemi, "Machine learning data analysis highlights the role of parasutterella and alloprevotella in autism spectrum disorders," Biomedicines, vol. 10, no. 8, p. 2028, Aug. 2022.

[3] R. Sreedasyam, A. Rao, N. Sachidanandan, N. Sampath, and S. K. Vasudevan, "Aarya—A kinesthetic companion for children with autism spectrum disorder," J. Intell. Fuzzy Syst., vol. 32, no. 4, pp. 2971–2976, Mar. 2017.

[4] J. Amudha and H. Nandakumar, "A fuzzy based eye gaze point estimation approach to study the task behavior in autism spectrum disorder," J. Intell. Fuzzy Syst., vol. 35, no. 2, pp. 1459–1469, Aug. 2018.

[5] H. Chahkandi Nejad, O. Khayat, and J. Razjouyan, "Software development of an intelligent spirography test system for neurological disorder detection and quantification," J. Intell. Fuzzy Syst., vol. 28, no. 5, pp. 2149–2157, Jun. 2015.

[6] F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba, "A deep learning approach to predict autism spectrum disorder using multisite resting-state fMRI," Appl. Sci., vol. 11, no. 8, p. 3636, Apr. 2021.

[7] K.-F. Kollias, C. K. Syriopoulou-Delli, P. Sarigiannidis, and G. F. Fragulis, "The contribution of machine learning and eye-tracking technology in autism spectrum disorder research: A systematic review," Electronics, vol. 10, no. 23, p. 2982, Nov. 2021.

[8] I. A. Ahmed, E. M. Senan, T. H. Rassem, M. A. H. Ali, H. S. A. Shatnawi, S. M. Alwazer, and M. Alshahrani, "Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques," Electronics, vol. 11, no. 4, p. 530, Feb. 2022.

[9] P. Sukumaran and K. Govardhanan, "Towards voice based prediction and analysis of emotions in ASD children," J. Intell. Fuzzy Syst., vol. 41, no. 5, pp. 5317–5326, 2021.

[10] S. P. Abirami, G. Kousalya, and R. Karthick, “Identification and exploration of facial expression in children with ASD in a contact less environment,” J. Intell. Fuzzy Syst., vol. 36, no. 3, pp. 2033–2042, Mar. 2019.

[11] M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, “Detecting autism spectrum disorder using machine learning techniques,” Health Inf. Sci. Syst., vol. 9, no. 1, pp. 1–13, Dec. 2021.

[12] Hasan, SM Mahedy, et al. "A machine learning framework for early-stage detection of autism spectrum disorders." IEEE Access 11 (2022): 15038-15057.

[13] Y. Zhang, R. Jin, and Z. Zhou, “Understanding bag-of-words model: A statistical framework,” Int. J. Mach. Learn. Cybern., vol. 1, nos. 1–4, pp. 43–52, Dec. 2010.

[14] Kang, Ruizhi, et al. "Learning chinese word embeddings with words and subcharacter n-grams." IEEE Access 7 (2019): 42987-42992.

[15] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, 1988.

[16] He, Yulin, et al. "A hybrid method to measure distribution consistency of mixed-attribute datasets." IEEE Transactions on Artificial Intelligence 4.1 (2022): 182-196.

[17] Rajab, Khairan D. "New associative classification method based on rule pruning for classification of datasets." IEEE Access 7 (2019): 157783-157795.

[18] B. L. W. H. Y. Ma and B. Liu, “Integrating classification and association rule mining,” in Proc. 4th KDD, 1998, pp. 80–86.

[19] F. Thabtah, O. Gharaibeh, and R. Al-Zubaidy, “Arabic text mining using rule based classification,” J. Inf. Knowl. Manage., vol. 11, no. 1, 2012, Art. no. 1250006.

[20] B. Ma, H. Zhang, G. Chen, Y. Zhao, and B. Baesens, “Investigating associative classification for software fault prediction: An experimental perspective,” Int. J. Softw. Eng. Knowl. Eng., vol. 24, no. 1, pp. 61–90, 2014.

[21] J. P. Lucas, A. Laurent, M. N. Moreno, and M. Teisseire, “A fuzzy associative classification approach for recommender systems,” Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 20, no. 4, pp. 579–617, 2012.

[22] K. D. Rajab, “New hybrid features selection method: A case study on websites phishing,” Secur. Commun. Netw., vol. 2017, Mar. 2017, pp. 9838169:1–9838169:10.

[23] Hasan, SM Mahedy, et al. "A machine learning framework for early-stage detection of autism spectrum disorders." IEEE Access 11 (2022): 15038-15057.

[24] Dutta, Sushama Rani, Sujoy Datta, and Monideepa Roy. "Using cogency and machine learning for Autism detection from a preliminary symptom." 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2019.

[25] Dutta, Sushama Rani, et al. "A machine learning-based method for Autism diagnosis assistance in children." 2017 International Conference on Information Technology (ICIT). IEEE, 2017.

[26] Dutta, Sushama Rani, et al. "Iot in Autism detection in its early stages." Internet of Things: Enabling Technologies, Security and Social Implications (2021): 47-58.

[27] Kohli, Manu, Arpan Kumar Kar, and Shuchi Sinha. "The role of intelligent technologies in early detection of autism spectrum disorder (asd): A scoping review." IEEE Access 10 (2022): 104887-104913.

[28] T. Akter, M. Shahriare Satu, M. I. Khan, M. H. Ali, S. Uddin, P. Lió, J. M. W. Quinn, and M. A. Moni, “Machine learning-based models for early stage detection of autism spectrum disorders,” IEEE Access, vol. 7, pp. 166509–166527, 2019.

[29] P. Pandey, P. Ap, M. Kohli, and J. Pritchard, “Guided weak supervision for action recognition with scarce data to assess skills of children with autism,” in Proc. AAAI Conf. Artif. Intell., Apr. 2020, vol. 34, no. 1, pp. 463–470.

[30] A. J. Kumm, M. Viljoen, and P. J. de Vries, “The digital divide in technologies for autism: Feasibility considerations for low- and middle-income countries,” J. Autism Develop. Disorders, vol. 52, no. 5, pp. 2300–2313, May 2022.

[31] Rusli, Nazreen, et al. "Implementation of wavelet analysis on thermal images for affective states recognition of children with autism spectrum disorder." IEEE Access 8 (2020): 120818-120834.

[32] T. K. Landauer, D. Laham, B. Rehder, and M. E. Schreiner, “How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans,” in Proc. 19th Annu. Meeting Cogn. Sci. Soc., 1997, pp. 412–417

[33] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” J. Amer. Soc. Inf. Sci., vol. 41, no. 6, pp. 391–407, 1990.

[34] T. Hofmann, “Probabilistic latent semantic indexing,” in Proc. ACM SIGIR Forum, 2017, pp. 211–218.

[35] R. Cohen, M. Elhadad, and N. Elhadad, “Redundancy in electronic health record corpora: Analysis, impact on text mining performance and mitigation strategies,” BMC Bioinf., vol. 14, p. 10, Jan. 2013.

[36] C. Lee, Z. Luo, K. Y. Ngiam, M. Zhang, K. Zheng, G. Chen, B. C. Ooi, and W. L. J. Yip, “Big healthcare data analytics: Challenges and applications,” in Handbook of Large-Scale Distributed Computing in Smart Healthcare. Cham, Switzerland: Springer, 2017, pp. 11–41.

[37] W. Kintsch, “The potential of latent semantic analysis for machine grading of clinical case summaries,” J. Biomed. Inform., vol. 35, no. 1, pp. 3–7, Feb. 2002.

[38] Rashid, Junaid, et al. "Topic modeling technique for text mining over biomedical text corpora through hybrid inverse documents frequency and fuzzy k-means clustering." IEEE Access 7 (2019): 146070-146080.

[39] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, 1988.

[40] Alhawarat, Mohammad, and Ahmad O. Aseeri. "A superior Arabic text categorization deep model (SATCDM)." IEEE Access 8 (2020): 24653-24661.

[41] I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig, and N. A. Mahyoub, “Automatic Arabic text categorization: A comprehensive comparative study,” J. Inf. Sci., vol. 41, no. 1, pp. 114–124, Feb. 2015.

[42] A. H. Mohammad, T. Alwada’n, and O. Al-Momani, “Arabic text categorization using support vector machine, Naïve Bayes and neural network,” GSTF J. Comput., vol. 5, no. 1, p. 108, 2016.

[43] I. Hmeidi, B. Hawashin, and E. El-Qawasmeh, “Performance of KNN and SVM classifiers on full word Arabic articles,” Adv. Eng. Inform., vol. 22, no. 1, pp. 106–111, Jan. 2008.

[44] M. A. H. Madhfar and M. A. H. Al-Hagery, “Arabic text classification: A comparative approach using a big dataset,” in Proc. Int. Conf. Comput. Inf. Sci. (ICCIS), Apr. 2019, pp. 1–5.

[45] M. S. Khorsheed and A. O. Al-Thubaity, “Comparative evaluation of text classification techniques using a large diverse Arabic dataset,” Lang Resour. Eval., vol. 47, no. 2, pp. 513–538, Jun. 2013.

[46] B. L. W. H. Y. Ma and B. Liu, “Integrating classification and association rule mining,” in Proc. 4th KDD, 1998, pp. 80–86.

[47] J. P. Lucas, A. Laurent, M. N. Moreno, and M. Teisseire, “A fuzzy associative classification approach for recommender systems,” Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 20, no. 4, pp. 579–617, 2012.

[48] F. Padillo, J. M. Luna, and S. Ventura, “Evaluating associative classification algorithms for big data,” Big Data Anal., vol. 4, no. 1, p. 2, 2019

[49] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, “Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison,” Comput. Biol. Med., vol. 136, Sep. 2021, Art. no. 104672.

[50] E. Dritsas and M. Trigka, “Stroke risk prediction with machine learning techniques,” Sensors, vol. 22, no. 13, p. 4670, Jun. 2022.

[51] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, “Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms,” Neural Comput. Appl., early access, pp. 1–17, Mar. 2022.

[52] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, “A machine learning approach to predict Autism spectrum disorder,” in Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE), Feb. 2019, pp. 1–6

[53] F. Thabtah, “Machine learning in autistic spectrum disorder behavioral research: A review and ways forward,” Inform. Health Social Care, vol. 44, no. 3, pp. 278–297, 2018.

[54] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” Appl. Soft Comput., vol. 97, Dec. 2020, Art. no. 105524.

[55] M. Duda, R. Ma, N. Haber, and D. P. Wall, “Use of machine learning for behavioral distinction of Autism and ADHD,” Transl. Psychiatry, vol. 6, no. 2, pp. e732–e732, Feb. 2016.

[56] F. Thabtah and D. Peebles, “A new machine learning model based on induction of rules for Autism detection,” Health Informat. J., vol. 26, no. 1, pp. 264–286, Mar. 2020

[57] M. Duda, R. Ma, N. Haber, and D. P. Wall, “Use of machine learning for behavioral distinction of Autism and ADHD,” Transl. Psychiatry, vol. 6, no. 2, pp. e732–e732, Feb. 2016.

[58] O. Altay and M. Ulas, “Prediction of the Autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children,” in Proc. 6th Int. Symp. Digit. Forensic Secur. (ISDFS), Mar. 2018, pp. 1–4.

[59] A. S. Haroon and T. Padma, “An ensemble classification and binomial cumulative based PCA for diagnosis of Parkinson’s disease and Autism spectrum disorder,” Int. J. Syst. Assurance Eng. Manage., early access, pp. 1–16, Jul. 2022

[60] Hasan, SM Mahedy, et al. "A machine learning framework for early-stage detection of Autism spectrum disorders." IEEE Access 11 (2022): 15038-15057.

[61] D. Mease, A. J. Wyner, and A. Buja, “Boosted classification trees and class probability/quantile estimation,” J. Mach. Learn. Res., vol. 8, no. 3, pp. 409–439, 2007.

[62] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, “DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values,” IEEE Access, vol. 7, pp. 102232–102238, 2019.

[63] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, and M. A. Hossain, “Comparative analysis of classification approaches for heart disease prediction,” in Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (ICME), Feb. 2018, pp. 1–4.

[64] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, “Applications of support vector machine (SVM) learning in cancer genomics,” Cancer Genomics Proteomics, vol. 15, no. 1, pp. 41–51, Jan./Feb. 2018.

[65] D. Ramesh and Y. S. Katheria, “Ensemble method based predictive model for analyzing disease datasets: A predictive analysis approach,” Health Technol., vol. 9, no. 4, pp. 533–545, Aug. 2019.

[66] A. Arabameri and H. R. Pourghasemi, “Spatial modeling of gully erosion using linear and quadratic discriminant analyses in GIS and R,” in Spatial Modeling in GIS and R for Earth and Environmental Sciences. Amsterdam, The Netherlands: Elsevier, pp. 299–321, 2019.

S HIMA BINDU SRI received the B.Tech. degree in computer science and engineering from Nagarjuna University (NU), India, and the M.Tech. degree in computer science and engineering from Jawaharlal Nehru Technological University Hyderabad (JNTUH), India. She is currently pursuing the Ph.D. degree with KL University Hyderabad, India. She is currently serving as an Assistant Professor with the Department of Computer Science and Engineering, Keshav Memorial Institute of Technology (KMIT), Hyderabad, India. Her research interests include Natural Language Processing, Machine Learning, and Artificial Intelligence.

SUSHAMA RANI DUTTA received the MTech. degree in Computer Science and Engineering from MM University, Haryana, India, in 2011, and the Ph.D. degree in Computer Science and Engineering from KIIT University, Bhubaneswar, in 2020. She is an Associate Professor in the Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad. She worked as JRF and SRF in MeitY project, Govt of India for 4 years. She has more than 20 research articles in international journals and conferences. Her research interests include Artificial Intelligence, Missing Data Handling, Machine Learning, NLP, and Data Mining.