International Journal of Computer Sciences and Engineering
Vol.**12**, Issue.**2**, pp.**18-29**, February **2024**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Autism Spectrum Disorder Detection from Parents Dialogues Using Multinomial Naïve Bayes and XGBoost Models

Prasenjit Mukherjee[1*], Sourav Sadhukhan[2], Manish Godse[3]

[1]Dept. of Technology,Vodafone Intelligent Solutions, Pune, India & Dept. of Computer Science, Manipur International University, Manipur, India
[2]Dept. of Business Management, Pune Institute of Business Management, Pune, India
[3]Dept. of IT, Bizamica Software, Pune, India

**Corresponding Author: prasen.mscit09@gmail.com*

Received: 29/Dec/2023; **Accepted**: 31/Jan/2024; **Published**: 29/Feb/2024. **DOI**: https://doi.org/10.26438/ijcse/v12i2.1829

Abstract: To indicate the proper development of a child, there are certain baseline milestones. If a child is not reaching the milestones at the expected rate, it can indicate that there is an issue that needs to be addressed. By early intervention, the development of the child can be improved and the long-term impact of the developmental delays may be reduced. One such constraint of child development is Autism spectrum disorder. The ASD-affected children exhibit difficulties in communication, socialization and challenges in physical, social, and emotional development. This neurodevelopmental disorganization may exhibit an extensive range of effects and symptoms including challenges in communication, social interactions, and physical, social, and emotional behaviours. To identify ASD symptoms in a child, the range of ASD symptoms must be available as datasets to the researchers. The difficult phenomenon is that parents are not able to identify or detect early-age indications of ASD in their children. This proposed research work aims to detect the symptoms of ASD from parents' dialogues. The dataset has collected data from many autism groups from social media and organizations for special children. To understand the sentiment of parents' dialog there are two important and popular machine learning models, the Multinomial Naïve Bayes and the XGBoost. Naïve Bayes is based on a probabilistic machine learning model and XGBoost is an ensemble-oriented model. If new data comes from a new parent, the sentiment of that data is also predicted by these models. By using these two models, sentiment analysis can help to identify ASD symptoms. Based on the prepared data, the accuracy of these two models is 70% and 70% respectively.

Keywords: Autism Spectrum Disorder, Machine Learning, ASD Detection, ML-based Framework, Traditional Machine Learning, Multinomial Naïve Bayes, XGBoost

1. Introduction

ASD, which is a complex neurodevelopmental disorder, affects brain functioning. It also affects the development of communication, social interaction, and behaviour. Since it is often accompanied by sensory processing problems and can cause complications in interpreting and understanding social cues, difficulty with social interaction and communication is mostly like to take place. Moreover, ASD can lead to problems with physical and emotional growth, such as problems with motor skills and emotional management. Detection of ASD in children usually goes unnoticed as most parents are unaware of the SD symptoms. Once, parents come to know that their child is suffering from ASD, they want to make sure that their children are catered to with the best opportunities and support as in [1]. World Health Organization (WHO) indicates that there is one autistic child in 160 children [2]. ASD-affected people often face problems with social interactions and communication skills ability to interpret gestures, facial expressions, and body language, and exhibit repetitive behaviours. This is a cause for concern because it is still unclear what causes autism spectrum disorder and there is no known cure. Early detection and intervention can help children with ASDs to lead more normal lives [3], but if they are not identified until later in life, they may not get the help they need. Early diagnosis and treatment are important because they can help reduce the severity of symptoms and help the child develop more cognitive and social skills. It can ~~also~~ help the family to provide necessary support as caregivers of their autistic child. Additionally, this can help reduce medical costs associated with diagnosing and treating ASD. By using machine learning algorithms, it is possible to detect patterns in the data that can be used to accurately classify children as being autistic or non-autistic. This can help medical professionals with quick and accurate ASD diagnosis in children, leading to earlier and more

effective interventions, which can ultimately result in lower medical costs associated with treating ASD as in [4]. To overcome this problem, new technologies like artificial intelligence, natural language processing, and machine learning have been developed that can accurately detect ASD with significantly less time. AI and ML technologies can be used to analyze vast amounts of data, but they are limited by the quality of the data they are given. If the data is incomplete or incorrect, the results will also be incomplete or incorrect. Therefore, it is important to have high-quality data in order to get accurate results as in [5]. Additionally, machine learning can be used to develop and refine diagnostic criteria that are tailored to individual patient characteristics. Additionally, machine learning may examine data from a range of different sources, such as behavioral, physiological, and biological data, and can provide comprehensive data on autism as in [6]. A new idea to detect Autism Spectrum Disorder has been approached by the authors of [7] using machine learning. The authors [7] have developed Worm Optimized Extreme Learning Machines (WOELM) that help to diagnose autism. The proposed algorithm is the hybrid type that has been developed by the glowworm optimization and single feed-forward extreme learning algorithms as in [7]. Another machine-learning approach has been developed by the authors [8] to detect autism. This machine learning algorithm is able to detect autism without any professional intervention as in [8]. Authors of [9] have developed a system for the detection of autism where convolutional neural network (CNN) and particle swarm optimization algorithm (PSO-CNN) have been used and that has been analyzed with the support vector machine (SVM), Logistic Regression (LR), Naïve Bayes (NB) for better performance and scalability as in [9]. According to ASD detection from the question-answering dataset [10], supervised algorithms have been used where KNN and Random Forest models have given the best accuracy based on question question-answering dataset. This work is related to the development of a web-based tool for ASD detection as in [10].

The early detection of autism is a difficult task because there are no early accurate signs or fixed symptoms that define autism. Brain MRI or EEG report analysis using a machine learning algorithm is time and cost-effective. MRI scans, EEG, or any other medical test always are not available in rural areas. Suggestions from medical practitioners are also difficult for rural people. Many health centers do not have any infrastructure for autism detection. Autism ASD is a problem among children which is not curable and it can affect the parent's life. The overall development path is critical for an autistic child with poor social behavior, Poor communication, poor eye contact, poor education, and poor social awareness. An autistic child contains many symptoms that are directly related to ASD. Early detection of autism among children is a good option for the reduction of ASD symptoms which is the main motive of this research. Most parents have failed to detect ASD symptoms at an early age in their autistic child. ASD symptoms can be detected within the range of 1.5 years to 2 years of age. The best developmental phase of a child is very good up to 6 years. Early detection of the ASD system is required where parents can enter their knowledge about the behaviour of their child as text inputs to the Early Age ASD Detection Systems proposed in this research. The given texts were processed by the proposed systems to understand the sentiment of each sentence and detect the symptoms of ASD from the sentences. The objective of the research is to introduce early-age ASD detection among children using machine learning techniques.

The proposed research work focused on ASD detection at an early age among children. The early age detection of ASD is a good time to start many therapies according to the need for reducing the symptoms of ASD. The data has been collected from the organizations of special children and many autism-related groups from social sites. These data are parents' dialogue in text mode. Thoughts and experiences as dialogue from parents of ASD children are very important because only parents spend most of their time with their ASD children. They have good knowledge of ASD symptoms. The dataset has been prepared using these parents' dialogues where a sentence is related to the positive symptoms or not. Multinomial Naïve Bayes and XGBoost algorithms have been appliedfor symptom detection from the sentence. Sentiment has been predicted by these two machine learning algorithms. The cosine similarity model accepts these positive sentences as input and processes it for finding the ASD symptoms that have been discussed in section 3. ASD-related similar types systems have been elaborated in section 2. The proposed system architecture has been elaborated in section 3 and the result of each machine-learning model has been discussed in section 4 and ends with Conclusion and Future Work in section 5.

2. Related Work

Autism Spectrum Disorder (ASD) is a healthcare issue and to understand and treat this disorder, much research is going on. Artificial intelligence (AI) has come into this field as an important tool not only for advancing our knowledge but also for improving developmental outcomes for a person with ASD. In this section, the discussion is about AI-based research in the healthcare domain including ASD. A given transcript may be analysed manually to identify impairment patterns in autistic patients, but NLP-based approaches perform the same task with increased efficiency ad accuracy. This is particularly helpful to ASD-affected populations with difficulty in communicating thoughts or necessities in verbal mode. By using an unsupervised NLP technique, we were able to develop a measure that could be applied to the transcripts without relying on any predetermined reference data. This allowed us to generate an unbiased evaluation of the lexico-semantic similarity between two conversations. The research was conducted to determine if this method of measuring semantic coherence could be used to accurately identify children with ASD. It was predicted that children with ASD would have more variability in their semantic coherence scores than children without ASD and that the scores would be a useful way to distinguish between the two groups. Authors [11] have used these tests to assess the participants' language abilities, social communication skills, and repetitive behaviours. This allowed us to determine if

there were any differences between the two groups in terms of language development and social communication skills. Authors [11] have used data from 70 numbers of males between 4 to 8 years with autism (N = 38) who are enrolled in a language study. After the transcripts were converted to vectors, the pairwise similarity between all of the subjects could be compared to the grand mean. This allowed ~~us~~ to measure how close each subject's language output was to the grand mean. From this, a pseudo-value was generated which represented the contribution of each subject to the grand mean. The data suggest that TD subjects were more likely to provide responses that were similar to the overall group mean than the ASD subjects. This suggests that TD subjects had a better understanding of the task and were able to provide more accurate responses than their ASD counterparts. The results showed that the variance of pseudo-values was significantly greater in the ASD group than TD group, indicating that the children with ASD had greater difficulty understanding the meaning of words. Furthermore, the results showed that this difficulty was not related to nonverbal IQ, mean length of utterance, or NDR (Neuro-developmental Regression), indicating that the difficulty was specific to understanding the meaning of words as in [11]. By studying the behaviour of individuals with ASD, researchers have found that they often struggle with the ability to predict how a situation will play out, and how their choices will affect the outcome. This suggests that deficits in prediction are an underlying factor in the development of ASD. The results indicated that the autistic children made more anticipatory fixations when compared to the NT (Neuro-typical) children, suggesting that they may be better at processing language in a predictive manner. Additionally, these results suggest that predictive language processing may be differentially preserved in autistic children, even in the early stages of language development as in [12]. This suggests that autistic children can process language in the same way as their neurotypical peers. Moreover, when controlling for age differences, it was found that the diagnostic group had a significant effect on looking behaviours, indicating that the behaviour of autistic children and their NT peers was different. This suggests that the ability to process language to inform visual search in the NT group was more successful than the ASD group, and that language skills were associated with improved visual search in both groups. The findings from the studies suggest that autistic participants when given the same set of stimuli, are able to anticipate or predict what will happen next. Further research is needed to determine if this predictive processing follows the same developmental trajectory as non-autistic (NT) children as in [12]. FIRST created an AI-based tool to help people with ASD process written documents. The tool is capable of summarizing complex sentences, detecting non-literal text, and recognizing uncommon and technical terms. By using the tool, people with ASD can understand the gist or meaning of the document without having to go through the entire document. This is because the editor uses natural language processing (NLP) to convert the original texts into language that is easier for people with ASD to comprehend. It also optimizes the layout of the text to make it easier to read and allows for customization of the text to suit the user's needs. The tool has been tested on individuals with ASD and the results show that it can improve communication skills, reduce anxiety, and teach social skills. It also has the potential to improve their quality of life by helping them to better navigate the world as in [13]. The causes of mental illness are diverse and include biological, genetic, environmental, and psychological factors. These can lead to an increased risk of developing mental health problems, including depression, anxiety, and other mental disorders. Additionally, certain lifestyle factors, such as poor diet and lack of physical activity, can also contribute to mental health problems. Through NLP methods, such as sentiment analysis, machine learning, and deep learning algorithms, it is possible to detect the presence of certain emotions, topics, and other complex associations that may be indicative of mental health issues. This can help to provide a more holistic view of mental health and enable more proactive healthcare services. The studies included in this review used a variety of NLP techniques, such as text classification, natural language processing, and machine learning, to detect mental illness. The review also looked at the challenges and limitations of current methods, as well as potential future directions for the field. This is due to the fact that deep learning models are better suited for dealing with unstructured data, such as text, and therefore are better able to detect patterns in the data and make more accurate predictions. Furthermore, deep learning models are relatively easy to implement and can be quickly adapted to new datasets, making them attractive for researchers as in [14]. ASD is characterized by deficits in communication, social interaction, restricted interests, and repetitive behaviours. Additionally, traditional autism detection methods rely heavily on subjective assessments and do not take into account the complexity of the disorder. By utilizing omics data, researchers may understand the genetic and molecular mechanisms of autism, which can then be used to develop more accurate and reliable detection methods. Multi-omics data analysis combines data from different sources, such as genetics, proteins, and metabolites, to get a more complete picture of the disorder. It can provide more insight into the underlying causes of the disorder and how it is distinct from other disorders. Supervised machine learning algorithms can then be used to make better predictions about the development of the disorder and to develop more targeted treatments. The integration of genomics and proteomics data allows for a more comprehensive understanding of the data, which can be used to accurately identify genes that are associated with autism. The machine learning algorithms help to identify patterns in the data that would not be possible with manual analysis. Additionally, the performance of the system is evaluated using standard datasets to ensure the accuracy of the results as in [15].

Machine learning models are applied to the medical domain, for disease detection, diagnosis, and prediction. Using UCI repository, various algorithms of machine learning algorithms such as Naive Bayes (NB), Support Vector Machines (SVM), and k-Nearest Neighbors (KNN) have been used successfully with k-fold validation. Clustered cross-validation strategy enabled us to identify parameters that influence the dataset the most, which can estimate the results more accurately. This

empowers us to get more reliable outcomes as well as validates the estimated accuracy. Hyper parameter tuning process can find the best values for parameters of a model that can produce the most accurate results. To validate a model, the clustered cross-validation method partitions the dataset into clusters, and after that, it applies cross-validation within each cluster. This can help to minimize the bias in the evaluation and produce ~~more~~ valid results. Since the SVM-based model had 99.6% accuracy, it exhibits the productiveness of the hyperparameter tuning and clustered cross-validation strategy as in [16]. Researchers can have a better understanding of the process of visual information by following the eye movements of individuals with ASD. This enables them to recognize behavioral patterns which are then effective in diagnosing ASD and providing more successful treatment. Researchers are able to identify patterns of visual activity that indicate autism by analyzing the projection points of the eye. This technique enables the doctors an early and proper diagnosis, and therefore they can start treatment in right away. By using neural networks, Deep learning algorithms process the data and make decisions. For a more proper diagnosis, the two approaches are combined by the hybrid technique. FFNNs and ANNs are designed so that they can parody the biological structure of the brain and replicate the way neurons interact with one another. The FFNN and ANN produced an accuracy of 99.8 % whereas the second technique (Google Net and ResNet 18) marked the accuracy as 93.6% and 97.6%. The third technique (Google Net + SVM and ResNet-18 + SVM) makes an accuracy of 95.5% and 94.5% as in [17].

Machine learning algorithms are enabled to analyze huge datasets to recognize patterns that are useful to diagnose properly and trace the development of ADHD. It is observed that between the ADHD and TD groups, there are differences in the prefrontal cortex oxygenated hemoglobin concentration. These findings also advise that NIRS has the capability of providing brain activity information for those children having ADHD. SVM performs ~~about the~~ differentiation between the ADHD and TD groups according to a specificity of 83.78%, a sensitivity of 88.71%, and also an overall discrimination rate of 86.25% as in [18]. The aim of this research is to examine whether EEG data are useful to properly diagnose ADHD based on the activity of an individual's brain. The deep learning methods applied in this study enabled the researchers not only to compare the EEG data between ADHD patients and healthy controls but also help to determine differences between them in terms of brain activity. A combination of electroencephalogram (EEG) data and machine learning models are used by the model, which has the usefulness to identify patterns from data that are connected with the two subtypes of ADHD. Patterns that were not visible to the bare~~d~~ eye were possible to be recognized by doing so and thus the accuracy of diagnosis improved. The deep learning method was unable to recognize the~~se~~ precise differences between subtypes of different cognitive disorders, resulting in misclassification of disorders with the application of deep learning methods, it is possible to identify patterns in EEG data that may help to clinically diagnose ADHD. It is required to perform furthermore research that can refine the approach and maximize the accuracy of the results as in [19]. To allow for better differentiation between the various types of ADHD based on the power spectra of the EEG measurements, the technique combines some classifiers. When classifiers are combined, it is easy for the algorithm to recognize patterns in the data that would otherwise be overlooked. This enabled researchers to compare the performance of the two groups based on various conditions and thus the effectiveness of the treatments can be measured. Furthermore, it can provide an understanding of the varying effects of ADHD based on different cognitive functions, like emotion regulation, attention, and executive functioning. With 117 adults (67 ADHD, 50 controls), the sample has been analyzed. The authors [20] have divided the sample into four data sets and created four distinct models that are able to identify the differences between the ADHD and control groups with their subtypes. The outputs of these classifiers are combined with the expressions derived from the Karnaugh map as in [20]. In Section 5 of this paper, the proposed models in the paper have been compared with state of art models, and a comparative study has been done and presented in Table 7.

3. Architecture of Proposed System

From the dialogues of parents, the process of identifying the symptoms of ASD has been made possible by applying two powerful machine-learning classifiers, the Naïve Bayes and the XGBoost, in tandem with the dataset. A comprehensive and robust analysis of the dialogue data is ensured by this tandem approach, which promises valuable outputs regarding the presence of ASD symptoms.

3.1 Dataset of Proposed System

Parents of probable autistic children discuss their personal experiences and thoughts and these parental dialogs help to constitute the proposed dataset which consists of dialogs of parents for the experiment. The dataset has been collected from several social networks and organizations which provide treatment for children with speech, communication, and behavioural challenges. Table 1 provides examples of the parent dialogues collected for this work. These dialogues contain valuable information that allows us to identify possible symptoms of ASD. We used this data to create a dataset that we will use to train and test our proposed machine-learning models.

Table 1: Text Data from Parents' Dialogues

Sl. No.	Text from Parents' Dialogues
1	Hello everyone, my son has gradually stated a one word conversation, like where are you going? Eating what? How are you? Etc..still he needs prompt for his answer. How can I improve further? Most of the time he will first ans huh bussh , like this, but after hearing the answer he will say right. But his articulation is poor. Please guide me further .....there is a long long way ahead. He is 5.1 years. Almost nil behaviour issue, stimming is there.
2	She is being bulliedather workplace and bullying has been affecting her from childhood. She feels uncomfortable and scared at her workplace.
3	My daughter is 16 month old. She is not responding to her name. Playing with toys. Has eye contact. When she is busy with toys she doesn't respond. She babble Manama, papapappa, bababba randomly. She always use to be with me and not with her

	relatives..plz suggest what should I do
4	Hi everyone ma daughter 3.8 years non verbal totally 0 i found out before 2 months she's Austistic now her behaviour is going worse if she wants anything she screams like a whistling sound n also hitting me n ma husband... behaviour is going worse in public also ppls watch her... anyone having such problems
5	I'm new on here. We recently just found out my 4 year old son is Autistic. I just need some advice/support. He has really bad tantrums where he will screams at the top of his lungs and throws things. He has also started hitting. We are struggling.

According to Table 2, the sentiment data has been prepared from the text data in Table 1. Each sentence was evaluated to determine whether it represented ASD symptoms or not. There are no fixed symptoms of ASD, so we included input from parents of autistic children to help identify additional symptoms and improve the proposed machine-learning model accuracy. Table 3 provides examples of the data we collected and included in the proposed dataset.

Table 2: Sentiment Analysis data in the proposed dataset

Sl. No.	Parents Comments	Sentiment
1.	So my son turned 18 in January & at the end of the school yr	0
2.	I actually stated out loud in front of him that I wondered why they didn't call me or send me a paper to let me know	1
3.	If I call her she does not look at me.	1
4.	I knew right then how his last yr was gonna go	0
5.	when I call him not much eye contact and also he's not talking	1

Table 2 presents the structure of the sentiment analysis dataset for the proposed research. The proposed dataset consists of three columns that are Serial Number, Parents Comments, and Sentiment. The text has been extracted from parents' dialogues and analyzed each sentence to determine if it represented a symptom of ASD. Sentences that were deemed to be symptoms of ASD were labeled with a 1 (true), while non-symptomatic sentences were labeled with a 0 (false). In Table 2, serial numbers 2, 3, and 5 in the Parents Comments column represent ASD symptoms that are positive, while serial numbers 1 and 4 are false symptoms. The proposed system has utilized this dataset to train MNB and XGBoost models.

Table 3: ASD Related Problems with Label

Sl. No.	ASD Related Problems	Label
1.	Speech Problem	1
2.	Sensory Problem	2
3.	Behaviour Problem	3
4.	Special Education	4
5.	Social Interaction	5
6.	Eye Contact	6
7.	Cognitive Behaviour	7
8.	Hyper Active Problem	8
9.	Child Psychological Problem	9
10.	Attention Problem	10

Table 3 presents the association between each ASD-related problem and its corresponding label. In this table, Label 1 represents the "Speech Problem," while 2 and 3 represent the "Sensory" and "Behavior" problems, respectively. Other ASD problems are also mentioned in the labels listed in Table 3. The proposed system predicts the sentiment of a sentence that is related to ASD symptoms. In the next step,The proposed system tries to identify the label using Table 4 that indicates an ASD-related problem. If a sentence is detected as a positive sentence (1) then it needs to be sent as an input in Spacy cosine similarity model. The cosine similarity model will check the similarity between the input sentence and each sentence from Table 4. The label will be selected according to a sentence that has scored the highest similarity with the input sentence and that label will be searched in Table 3 for correct ASD-related problems.

Table 4: Dataset for Cosine Similarity check

Sl. No.	Positive Sentences	Label
1.	he starting running shortly after up down the stairs by him no problem.	8
2.	Failed to show her how to do potty during the day.	7
3.	I'm like no wonder why he doesn't understand us as his parents.	9
4.	So I need to know how to handle him before time to meet his specialist.	3
5.	but because being non verbal I'm afraid someone else won't	1

The analysis of the results of each model is in the Results and Discussion section.

3.2 Multinomial Naïve Bayes Model

The Multinomial Naïve Bayes algorithm has been used for binary classification in this proposed research work to identify the symptoms from parents' texts. The main advantage of the Naïve Bayes algorithm is feature independent means each feature classification is not dependent on other features. The Naïve Bayes algorithm [21] is a purely probability-based algorithm where the probability of class A when predictor B is already provided.

P(B) = probability of B

P(A) = probability of class A

P(B|A) = occurrence of predictor B given class A probability

The equation will be

$$P(A|B) = P(A) * P(B|A)/P(B)$$ as in [21]

This equation work inside the Naïve Bayes algorithm to calculate the probability of the event.

Proposed Multinomial Naïve Bayes Algorithm:

Pseudo Code:

Step 1: Initialize path of CSV file for data reading.

Step 2: D=data in row-column format from csv

$p_1=[a_1,a_2,a_3,a_4,a_5,\ldots\ldots a_n]$ it refers to the parents' comments data inside the dataset.

$q_1=[r_1,r_2,r_3,r_4,r_5,\ldots\ldots r_n]$ it refers to label data inside the dataset

Step 3: Stop words removal Vector representation

tokens = RegexpTokenizer(r'[a-zA-Z0-9]+')

cv = CountVectorizer(stop_words='english',ngram_range = (1,1),tokenizer = token.tokenize)

textcounts = cv.fit_transform(p_1)

Step 4: Splitting dataset in train and test combination.

Xtrain, Xtest, Ytrain, Ytest = train_test_split(textcounts, q_1, test_size=0.25, random_state=5)

Step 5: Naïve Bayes model creation and prediction

MNB = MultinomialNB()

MNB.fit(Xtrain, Ytrain)

Predicted_value = MNB.predict(Xtest)

3.3 XGBoost Model

XGBoost is a powerful machine learning algorithm and it comes with an optimized distributed gradient boosting algorithm. This algorithm is efficient and scalable for machine learning model training. XGBoost is equipped with multiple weak models that are able to produce a strong prediction. XGBoost is widely used in the machine learning field because this model is able to produce good predictions when the dataset is large. The best feature of XGBoost is missing value handling without any pre-processing activity. Customization is allowed for fine-tuning the parameters of various model in XGBoost which enhance the accuracy. Many decision trees are created in sequential form at the first stage of the XGBoost model. Weights are included in each independent variable and fed into the decision trees for good prediction.

Proposed XGBoost algorithm:

Pseudo code:

Step 1: Initialize path of CSV file for data reading.

Step 2: D= data in row-column format from csv

$p_1=[a_1,a_2,a_3,a_4,a_5,\ldots\ldots a_n]$ it refers to the parents' comments data inside the dataset.

$q_1=[r_1,r_2,r_3,r_4,r_5,\ldots\ldots r_n]$ it refers to label data inside the dataset.

Step 3: Stop words removal Vector representation

// Split the data in train and test format

xtrain,xtest,ytrain,ytest=train_test_split(p_1, q_1,stratify= q_1,test_size=0.33)

// Stop words remove from input text

en_stopwords = set(stopwords.words("english"))

no_stopwords = ["not","don't","aren','don','ain","aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't","won't", 'wouldn', "wouldn't"]

forno_stopword in no_stopwords:

en_stopwords.remove(no_stopword)

Step 4: Input text to Vector representation.

vc = CountVectorizer(binary = True)

vc.fit(x_train)

trainx = vc.transform(xtrain)

testx = vc.transform(xtest)

Step 5: XGBoost model creation and prediction

xgb1_train = xgb.DMatrix(trainx, xgb_train_labels)

xgb1_test = xgb.DMatrix(testx, xgb_test_labels)

param = {'eta': 0.75, 'max_depth': 50, 'objective': 'binary:logitraw'}

// Training the Modelecture

xgb1_model = xgb.train(param, xgb1_train, num_boost_round = 30)

// Prediction using the Model

prediction = xgb1_model.predict(xgb1_test)

The overall architecture has been given in Fig. 3.

3.4 Proposed System Architecture

Figure 1 has given a clear overview of the proposed system that is equipped with two popular machine-learning models. The first model is multinomial naïve bayes and second model is XGBoost. The data and the dataset preparation has been discussed above of this research paper. The proposed system will read the dataset and start some preprocessing tasks which is related to NLP. At first, the proposed system will read each sentence from the 'Parents Comment' column of the dataset and tokenize it. Each token is a word, again this system will remove all unwanted words like 'am','is',a','an', etc. using the stop words removal method of NLP. In the next step, the remaining tokens or words are transformed into vectors as input for each machine-learning model. These all steps are the preprocessing task of NLP. Now, the data has been split into two parts, the first data part is for the training and the second data part is for testing the model performance. After model training and testing, the proposed system is ready to accept new text from the user in order to detect positive sentences that denote symptoms of Autism Spectrum Disorder (ASD). The system will predict whether each sentence is positive (1) or negative (0).

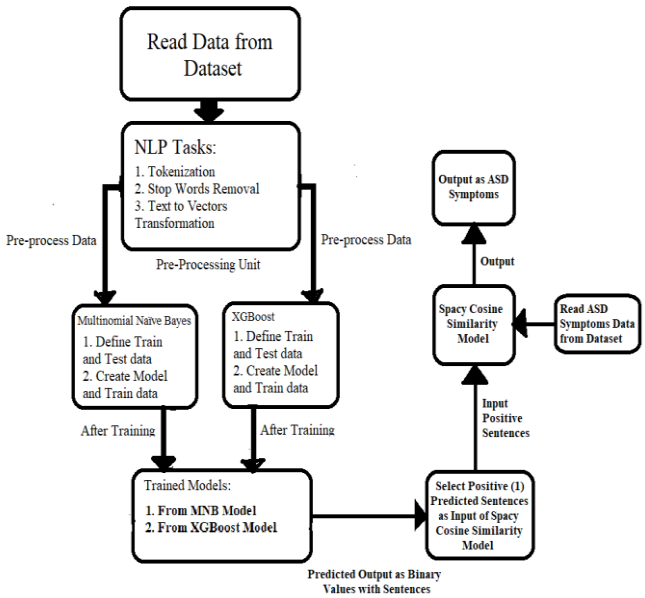

Figure 1: Proposed System Architectural Diagram

Only the positive sentences will be selected for further processing, in the upcoming stage of our process, we will filter out negative sentences to focus solely on positive ones. These selected affirmative sentences will play a pivotal role as input for our Spacy Cosine Similarity Model. This sophisticated model will meticulously analyze each positive sentence drawn from the ASD symptoms dataset (referenced as Table 3) and perform intricate calculations to determine their cosine similarity with the input sentence. Through this rigorous comparison of cosine similarity scores, the Spacy Cosine Similarity Model will discern the sentence that exhibits the highest degree of similarity with the input sentence. Subsequently, the system will make its selection based on the label associated with this most similar sentence. This selection will effectively indicate the specific ASD issue in accordance with the information laid out in Table 4.

Method:

Each input sentence will be processed by the cosine similarity model to identify the corresponding ASD problems. Please find the algorithm provided below for reference:

1. Input a paragraph text.
2. To store positive sentences, do the initialization of an empty list.
3. In the paragraph text, iterate over each sentence.
4. For predicting whether the sentence is positive or negative, use a trained model.
5. If the sentence turns out positive in prediction, include it to the list of positive sentences.
6. After all the sentences are being processed, for maximum cosine similarity score and label, initialize a variable.
7. Do the iteration over each positive sentence.
8. The input sentence and each sentence in the ASD symptoms dataset are calculated for similarity, where cosine similarity scores are used as a metric for our experiment.
9. Watch on the maximum cosine similarity score and the corresponding label.
10. The selection of label will be based on the highest cosine similarity score as the detected ASD problem.
11. Do the repetition from steps 7 to 10 for each positive sentence.
12. According to the cosine similarity scores, the ASD problems will be identified.

The proposed system is allowed by this approach to detect positive sentences associated with ASD signs and then to determine the corresponding ASD problem, matching them with the most alike sentences in the dataset of ASD signs. In pseudo-code, the Cosine similarity algorithm has been provided here.

The pseudocode of Cosine Similarity algorithm:

```
# Step 1: Import Spacy and Python packages
import spacy
import pandas as pd

# Load the English language model
nlp1 = spacy.load('en_core_web_lg')

# Step 2: Data frame to be initialized with positive ASD symptoms data.
df1 = pd.read_csv("ASD_Symptoms.csv")

# Initialize empty lists to store cosine similarity values, comments, and sentiment
input_comments = []
sentiment_labels = []
cosine_similarity_scores = []
# Step 3: Cosine Similarity method has been defined here.
defcalculate_cosine_similarity(input_sentence):
for index in df1.index:
    sentence1 = nlp1(df1['Comments'][index])
    sentence2 = nlp1(input_sentence)
    # Remove stop words
    sentence1_no_stop_words = nlp1(' '.join([str(token) for token in sentence1 if not token.is_stop]))
    sentence2_no_stop_words = nlp1(' '.join([str(token) for token in sentence2 if not token.is_stop]))

input_comments.append(df1['Comments'][index])
sentiment_labels.append(df1['Sentiment'][index])
    # Calculate cosine similarity
similarity_score = sentence2_no_stop_words.similarity(sentence1_no_stop_words)
cosine_similarity_scores.append(similarity_score)

  # Adata frame to be created to store the results
df1_results = pd.DataFrame({
    'Comments': input_comments,
    'Sentiment': sentiment_labels,
    'Cosine_Scores': cosine_similarity_scores
  })

  # Save the results to a CSV file
df_results.to_csv(r'ASD_Cosine_Data.csv')
  df1_results['Cosine_Scores'] = df1_results['Cosine_Scores'].astype('float64')

  # Find the index of the highest cosine similarity score
index_with_max_similarity = df1_results['Cosine_Scores'].idxmax()

return df1_results['Sentiment'][index_with_max_similarity]

# Step 4: predicted positive sentences have to be selected as input
predicted_positive_sentences = [
  "List of predicted positive sentences here",
  "Another positive sentence here",
  # Add more positive sentences as needed
]

for sentence in predicted_positive_sentences:
result = calculate_cosine_similarity(sentence)
print(sentence, "=", result)
```

The result of this system has been discussed in section 4.

4. Results and Discussion

The results of the proposed Multinomial Naïve Bayes and XGBoost model have been discussed in this section. All important metrics for model evaluation have been discussed one by one.

4.1 Result and Discussion of the Naïve Bayes Model

Table 5: Multinomial Naïve Bayes Model Metrics

Sl. No.	Metrics	Values
1	Accuracy	0.70
2	Precision	0.70
3	Recall	0.72
4	F1	0.71
5	AUC	0.70

The performance of the proposed Multinomial Naïve Bayes model can be thoroughly assessed using the metrics which are outlined in Table 5. Among these metrics, the Area bounded by the ROC Curve (AUC) is a vital indicator, that indicates the extent of the area under the curve of Receiver Operating

Characteristic (ROC). According to our demonstration, the AUC value is 0.70, which is equivalent to 70%. This value denotes the ability of the model to discriminate successfully between positive and negative results. The F1 score, an important metric derived from precision and recall scores, further explains the performance of our model. For our Multinomial Naïve Bayes model, the F1 score is 0.71, which reflects a coherence balance between precision (0.70 or 70%) and recall (0.72 or 72%). This balance emphasizes the robustness of the model in rightly pointing out relevant cases when false positives and false negatives are minimized.

According to the overall accuracy of our proposed model, which is 0.70 or 70%, it experimentally exhibits the ability of the model to provide proper classifications over the dataset. Fig 2 describes the graph related to the training and testing accuracy of our proposed model, which highlights the learning dynamics of the model and its capability to generalize unseen data. The training accuracy of the proposed model stands to 0.92 (92%) whereas the accuracy of the testing of the proposed model ~~is~~ turned out to be 0.70 (70%). These metrics are very crucial for the assessment of the model.

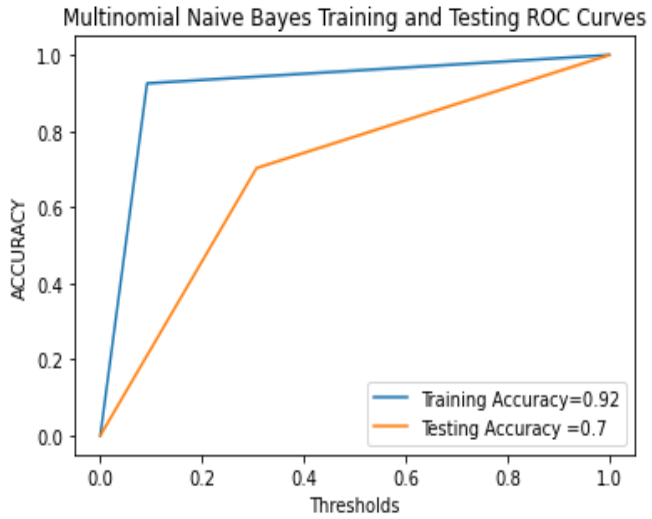

Figure 2: ROC Curve of Multinomial Naïve Bayes Model

4.2 Result and Discussion of the XGBoost Model

In the field of machine learning, the XGBoost is considered as one of the most accepted machine learning models. Based on the dialogues of parents, the proposed XGBoost model is proficient in producing prediction results through the use of ensemble techniques. Table 6 is used to serve as our reference, holding essential metrics with their corresponding values, critical for evaluating the performance of the model. Particularly, in Table 6, the Area under the ROC Curve (AUC), serves as an important parameter that depicts a significant measure of the discriminative capability of the model. Notably, this XGBoost model gains the value of AUC as 0.72, denoting an impressive success rate of 72% in distinguishing between positive and negative outcomes. The F1 score, an extensive metric that can balance Precision and Recall, provides further understanding of the effectiveness of the model. With an F1 score of 0.63 (63%), this model exhibits an admirable equilibrium between Precision (0.69 or 69%) and Recall (0.59 or 59%). This equilibrium shows the capability of the model to make positive predictions properly while sufficiently recognizing actual positive cases. With regard to overall accuracy, the model gains an admirable score of 0.70 or 70%, affirming its capability to arrange proper classifications across the dataset. In summary, the XGBoost model is projected as a strong performer, with remarkable AUC, Precision, Recall, F1 score, and overall accuracy values, which ensures the efficacy of the model in extracting valuable understandings from the dialogues of parents.

Table 6: XGBoost Model Metrics

Sl. No.	Metrics	Values
1	Accuracy	0.70
2	Precision	0.69
3	Recall	0.59
4	AUC	0.72
5	F1	0.63

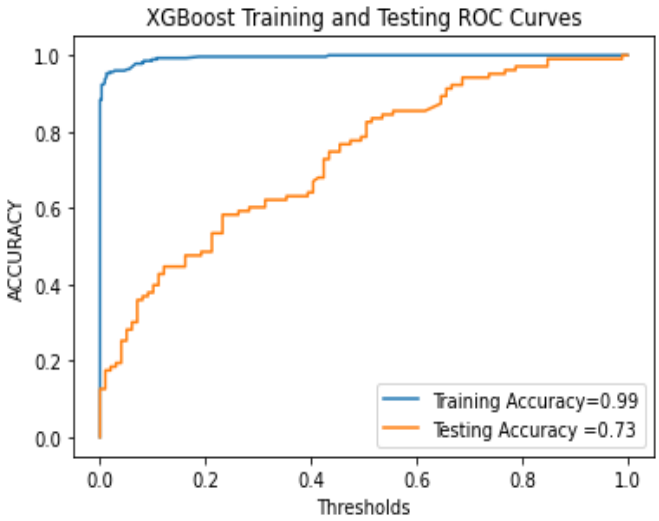

Figure 3: ROC Curves of the XGBoost Model

According to Fig. 3, Two ROC curves have represented the accuracy of the proposed XGBoostmodel. The first blue colour curve is referred to the training accuracy of the proposed model which is 0.99 (99%). The second yellow curve stated the testing accuracy of the proposed XGBoost model which is 0.73 (73%). These ROC curves are generated after training the model on the dataset that has been stated in section 3.

4.3 Result and Discussion of the Cosine Similarity Model

Predicted positive sentences by machine learning algorithms are input in a cosine similarity model that returns the output. The output results are illustrated in Figure 4, where sentences like *"her behavior is too rude with others"* is labeled with 3. Similarly, the sentences *"she doesn't like to see in my eyes"* and *"we are trying to teach her potty and pee training"* are labeled with 6 and 7, respectively. In Table 4, label 3 corresponds to Behavior problems, label 6 indicates Eye Contact problems, and label 7 represents Cognitive Behaviour problems. These problems can be managed by applying appropriate therapies. These targeted therapies can prove highly beneficial in reducing the symptoms associated with

ASD. This personalized approach has the potential to make a significant positive impact on individuals with ASD and their overall well-being.

Figure 4: Example of Output of Cosine Similarity Model

5. Comparative Analysis with Similar Models

A comparative analysis table (Table 7) has been stated here. This table refers to a complete comparative analysis of proposed models and similar-type models. Table 7 describes the models, model description, dataset, accuracy, and remarks about the models.

Fig. 5 shows the accuracies of similar types of models and proposed models. The proposed system uses the machine learning algorithms multinomial naïve Bayes and XGBoost for sentiment understanding of a sentence. The text data has been used which reduces the time and cost. This model will be effective for the detection of ASD in rural area.

Table 7: Comparative Analysis Between Proposed MNB and XGBoost Models and Similar Types of Machine Learning Models in Healthcare Domain

Sl.No.	ML Models	Model Description	Training and Testing Dataset	Model Accuracy	Remarks about Models
	Similar Type Machine Learning Models in Healthcare				
1	Adaboost, Random Forest, J48 and KNN [11]	These models are classifying the candidate gene from the non-autistic gene.	Genomics and proteomics data	91.6%, 92.4%, 91.4%, and 91.8	The independent omics data has been analyzed for prediction. The NCBI and SFARI standard data have been tested on these models.
2	Naive bayes (NB), Support Vector Machines (SVM) and k-Nearest Neighbours (KNN) [12]	These models are used to diagnose Autism using 3 datasets from the UCI repository.	Autism Toddler Dataset Autism Adult Dataset Autism Adolescent Dataset	a) 93%, 98%, and 97% b) 86%, 96.2%, and 88% c)81%, 88%, and 99%	Each model has been trained categorically according to the datasets. The accuracy has been given on each dataset.
3	GoogleNet + SVM ResNet-18 + SVM [13]	These models have used Eye tracking techniques to detect ASD.	The dataset has been used from Figshare data repository.	95.5% and 94.5%	Two blocks have been used where the first block refers to the CNN to extract deep feature map and the second block will use SVM to classify the extracted features from first block.
4	Support Vector Machine [14]	This model will predict the ADHD severity of the ADHD patient.	Physiological indicators, age, and reverse Stroop task (RST) data of 108 children	sensitivity of 88.71%, specificity of 83.78%	This work related to the ADHD severity prediction using the SVM model that uses data of 108 children.
	Proposed Models in ASD				
5	Proposed Multinomial Naïve Bayes (MNB) Model	MNB model has been to predict positive ASD symptoms from parents' dialogue.	Textual data containing dialogues from parents of autistic children has been sourced from multiple channels, including SAHAS-Durgapur in India and various social media platforms. This corpus of text captures valuable insights and perspectives from	70%	The data has been meticulously gathered in textual format, offering a treasure trove of insights. Within this collection, the dialogues of parents discussing their autistic children emerge as an invaluable resource. These parents have graciously shared their personal experiences and thoughts, providing a unique window into the world of autism from a firsthand perspective. Indeed, a parent of an autistic child serves as an unparalleled source for comprehending the intricate patterns and manifestations of ASD symptoms.

			parents, shedding light on their experiences and interactions related to autism.		
6	Proposed XGBoost Model	XGBoost model has been used to predict positive ASD symptoms from parents' dialogues.	Textual data containing dialogues from parents of autistic children has been sourced from multiple channels, including SAHAS-Durgapur in India and various social media platforms. This corpus of text captures valuable insights and perspectives from parents, shedding light on their experiences and interactions related to autism.	70%	The data has been meticulously gathered in textual format, offering a repository of parents' dialogues about their autistic children. These parental dialogs have served as primary resource to the model's dataset to identify patterns to identify symptoms of ASD.

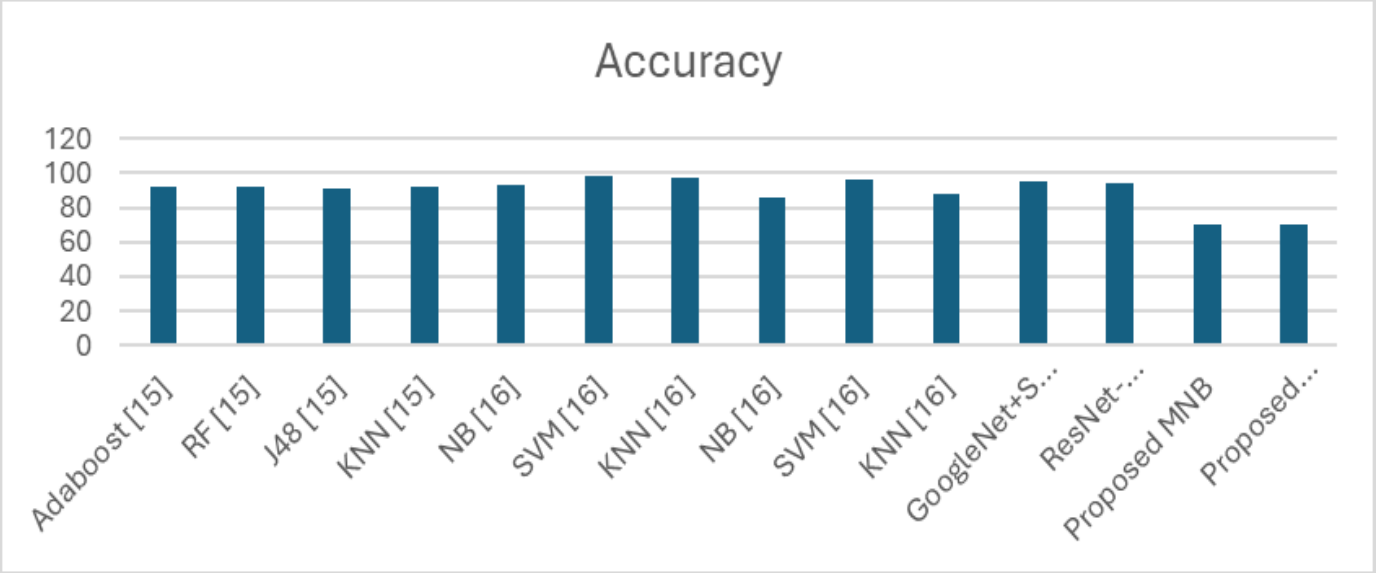

Figure 5: Accuracy of Similar Type Models and Proposed Models

6. Limitation of the Proposed System

The proposed system employs probabilistic and ensemble machine-learning models. To improve ~~the~~ accuracy of the proposed model, deep learning techniques like BERT or Large Language models like ChatGPT may be applied to the dataset. However, the limitation of this model is that it performs well with small to medium-sized databases, for large-sized databases, the performance degrades.

7. Conclusion and Future Scope

The proposed system accepts parents' dialogues as natural language text. It employs sentiment analysis techniques to generate positive or negative sentences based on the presence or absence of ASD symptoms. The sentiment analysis task is handled by the Multinomial Naïve Bayes and XGBoost models using the parent's comments dataset. These two models predict the positive sentences from the user text using sentiment analysis (binary classification). The predicted positive sentences are treated as input for the cosine similarity model. This model accepts all positive sentences from previous models and calculate the scores of similarity to find out the actual ASD problem using the dataset of ASD symptoms. The proposed system has been developed using text data. However, it is possible to include image processing techniques to enhance this system. Image processing techniques will support to processing MRI scan images or any other medical images that are related to ASD. This processing

will help to increase the accuracy of ASD detection. This is the future direction of this research. The proposed application can be utilized by organizations for spreading purposes in rural areas. parents in rural areas are not aware of autism and financially they are weak. MRI scans or other medical tests are very expensive for them to detect ASD. This proposed system can be developed as an end-to-end application for the healthcare sector. This system can be enhanced further using BERT and ChatGPT models. Today, Large Language Models (LLM) are taking positions in many projects inside the industries to solve critical problems. This system can be equipped with LLM models like ChatGPT, Palm, or Gemini. These LLM models are very advanced models according to LLM and these are able to solve critical tasks in the healthcare domain. According to the accuracy, scalability, and stability, these models are performing very well and this enhancement will be the future work of this research.

Data Availability

The data has been arranged from various Autism Groups that are available on various social media. The data has been prepared with the support of the Speech and Hearing Society (SAHAS), Durgapur, India.

Conflict of Interest

The authors have no conflicts of interest to declare.

Funding Source

No funding was received for conducting this research.

Authors' Contributions

The authors equally contributed to the present research, at all stages from the formulation of the problem to the final findings and solution.

Acknowledgments

The authors extend their appreciation to the Manipur International University, Imphal, India for supporting this post-doctoral research work on Autism.

References

[1] A. Tewari, S. D. Khanna, R. Bathija, R. Daryani, A. Reejhsinghani, "ASD (Autism Spectrum Disorder): Early Detection Intervention using Machine Learning", *International Journal of Innovative Science and Research Technology*, Vol.**5**, No.**1**, pp.**7-13**, **2020**.

[2] P. Mesa-Gresa, H. Gil-G´omez, L. Quilis, G. Gomez, "Effectiveness of virtual reality for children and adolescents with autism spectrum disorder: an evidence-based systematic review", *Sensors*, Vol.**18**, No.**8**, Art. No.**2486**, **2018**.

[3] S. E. Bryson, L. Zwaigenbaum, W. Roberts, "The early detection of autism in clinical practice", *Pediatrics & Child Health*, Vol.**9**, No.**4**, pp.**219–221**, **2004**.

[4] S. Islam, T. Akter, S. Zakir, S. Sabreen, M. I. Hossain, "Autism Spectrum Disorder Detection in Toddlers for Early Diagnosis Using Machine Learning", *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pp.**1-6**, **2020**.

[5] R. Murugan, L. Senbagamalar, "Prediction of Autism Spectrum Disorder using Machine Learning", *Journal of Emerging Technologies and Innovative Research (JETIR)*, Vol.**9**, No.**2**, pp.**c950-c959**, **2022**.

[6] D. Bone, M. S. Goodwin, M. P. Black, C. C. Lee, K. Audhkhasi, and S. Narayanan, "Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and promises", *Journal of Autism and Developmental Disorders*, pp.**1-48**, **2014**.

[7] A. Saranya, R. Anandan, "Autism Spectrum Prognosis using Worm Optimized Extreme Learning Machine (WOEM) Technique", *International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp.**636–641**, **2021**.

[8] N. Zaman, J. Ferdus, A. Sattar, "Autism Spectrum Disorder Detection Using Machine Learning Approach", *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, pp.**1-6**, **2021**.

[9] V. Kavitha, R. Siva, "Classification of Toddler, Child, Adolescent and Adult for Autism Spectrum Disorder Using Machine Learning Algorithm", *9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, pp.**2444-2449**, **2023**.

[10] S. Islam, T. Akter, S. Zakir, S. Sabreen, M. I. Hossain, "Autism Spectrum Disorder Detection in Toddlers for Early Diagnosis Using Machine Learning", *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Gold Coast, Australia, pp.**1-6**, **2020**.

[11] J.R. Adams, A. C. Salem, H. MacFarlane, R. Ingham, S. D. Bedrick, E. Fombonne, J. K. Dolata, A.P. Hill, and J. V. Santen, "A Pseudo-Value Approach to Analyze the Semantic Similarity of the Speech of Children With and Without Autism Spectrum Disorder", *Front. Psychol*., pp.**1-10**, **2021**.

[12] K. E. Prescott, J. M. Scott, T. Reuter, J. Edwards, J. Saffran, S. E. Weismer, "Predictive language processing in young autistic children", *Autism Research*, pp.**1-12**, **2022**.

[13] C. Orasan, R. Evans, R. Mitkov, "Intelligent text processing to help readers with autism", *Intelligent Natural Language Processing: Trends and Applications*, pp.**713-740**, **2017**.

[14] T. Zhang, A. M. Schoene, S. Ji, S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review", *NPJ Digital Medicine*, pp.**1-13**, **2022**.

[15] G. Anurekha, P. Geetha, "Performance Analysis of Supervised Approaches for Autism Spectrum Disorder Detection", *International Journal of Trend in Research and Development*, pp.**1-3**, **2017**.

[16] N. Ajaypradeep and R. Sasikala, "Child Behavioral Analysis: Machine Learning based Investigation for Autism Screening and Early Diagnosis", *International Journal of Early Childhood Special Education*, Vol.**13**, No.**2**, pp.**1199-1208**, **2021**.

[17] I. A. Ahmed, E. M. Senan, T. H. Rassem, M. 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", *MDPI Electronics*, Vol.**11**, pp.**1-27**, **2022**.

[18] A. Yasumura, M. Omori, A. Fukuda, J. Takahashi, Y. Yasumura, E. Nakagawa, T. Koike, Y. Yamashita, T. Miyajima, T. Koeda, M. Aihara, H. Tachimori, M. Inagaki, "Applied Machine Learning Method to Predict Children With ADHD Using Prefrontal Cortex Activity: A Multicenter Study in Japan", *Journal of Attention Disorders*, pp.**1-9**, **2017**.

[19] A. Vahid, A. Bluschke, V. Roessner, S. Stober, and C. Beste, "Deep Learning Based on Event-Related EEG Differentiates Children with ADHD from Healthy Controls", *Journal of Clinical Medicine*, pp.**1-15**, **2019**.

[20] A. Tenev, S. M. Simoska, L. Kocarev, J. P. Jordanov, A. Müller, G. Candrian, "Machine learning approach for classification of ADHD adults", *International Journal of Psychophysiology*, pp.**1-5**, **2013**.

[21] S. P. Barus, "Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University", *International Conference on Science Education and Technology (ICOSETH)*, pp.**1-7**, **2020**.

AUTHORS PROFILE

Dr. Prasenjit Mukherjee has 14 years of experience in academics and industry. He completed his Ph.D. in Computer Science and Engineering in the area of Natural Language Processing from the National Institute of Technology (NIT), Durgapur, India under the Visvesvaraya PhD Scheme from 2015 to 2020. Presently, He is working as a Data Scientist at Vodafone Intelligent Solutions, Pune, Maharashtra, India, and doing his Post Doctoral (D.Sc.) in Computer Science from Manipur International University, Imphal, Manipur, India.

Sourav Sadhukhan has above 5 years of experience in Law and Management. He completed his Graduation in LLB from Calcutta University, Kolkata, India, and Post Graduate Diploma in Management from Pune Institute of Business Management, Pune, India. Presently he is a student of Executive Post Graduation in Data Science and Analytics from the Indian Institute of Management, Amritsar, India.

Dr. Manish Godse has 27 years of experience in academics and industry. He holds Ph.D. from Indian Institute of Technology, Bombay (IITB). He is currently working as an IT Consultant in the Bizamica Software, Pune in the area of Artificial Intelligence and Analytics. His research areas of interest include automation, machine learning, natural language processing and business analytics. He has multiple research papers indexed at IEEE, ELSEVIER, etc.

