Early Detection of Autism Spectrum Disorder (ASD) using Traditional Machine Learning Models

Prasenjit Mukherjee[1], Sourav Sadhukhan[2], Manish Godse[3], Baisakhi Chakraborty[4]
Dept. of Technology, Vodafone Intelligent Solutions, Pune, India[1]
Dept. of Computer Science, Manipur International University, Manipur, India[1]
Dept. of Finance, Pune Institute of Business Management, Pune, India[2]
Dept. of IT, BizAmica Software, Pune, India[3]
Dept. of Computer Science and Engg, National Institute of Technology, Durgapur, India[4]

***Abstract*—Autism Spectrum Disorder (ASD) is a mental disorder among children that is difficult to diagnose at an early age of a child. People with ASD have difficulty functioning in areas such as communication, social interaction, motor skills, and emotional regulation. They may also have difficulty processing sensory information and have difficulty understanding language, which can lead to further difficulty in socializing. Early detection can help with learning coping skills, communication strategies, and other interventions that can make it easier for them to interact with the world. This kind of disorder is not curable but it is possible to reduce the symptoms of ASD. The early age detection of ASD helps to start several therapies corresponding to ASD symptoms. The detection of ASD symptoms at an early age of a child is our main problem where traditional machine learning algorithms like Support Vector Machine, Logistic Regression, K-nearest neighbour, and Random Forest classifiers have been applied to parents' dialog to understand the sentiment of each statement about their child. After completion of the prediction of these models, each positive ASD symptoms-related sentence has been used in the cosine similarity model for the detection of ASD problems. Samples of parents' dialogs have been collected from social networks and special child training institutes. Data has been prepared according to the model for sentiment analysis. The accuracies of these proposed classifiers are 71%, 71%, 62%, and 69% percent according to the prepared data. Another dataset has been prepared where each sentence refers to a particular categorical ASD problem and that has been used in cosine similarity calculation for ASD problem detection.**

***Keywords*—Support vector; logistic regression; cosine similarity; K-nearest neighbor; random forest**

I. Introduction

People with ASD [1] often have difficulty in understanding the social cues and expectations that are necessary for meaningful conversations and relationships with others. This can lead to isolation, difficulty in forming relationships, and, in some cases, difficulty in gaining recognition in society as in [2]. Early detection can help identify the illness sooner, allowing for personalized treatments or preventive measures to be put in place that can help reduce the severity of the illness and improve the chances of recovery as in [3]. It is caused by a combination of genetic and environmental factors that affect the development of the brain. It is characterized by difficulty in social interaction, communication, and repetitive behaviors. Research has been done to identify the causes of this syndrome, which include genetic predisposition, environmental factors, and lifestyle choices. Although the exact cause is still unknown, the available evidence shows that it is a multi-faceted condition. In addition, the lack of trained professionals and resources to diagnose and treat ASD [1] has created a huge gap in access to care. Furthermore, due to the complexity of the disorder, it can be difficult to diagnose and properly classify it, leading to misdiagnosis or delayed diagnosis. This is because autism is a complex disorder, and it can manifest itself differently in each affected individual [4]. As such, it is difficult to create a single biomarker that can accurately detect the disorder. Additionally, research into developing tools and applications, data analysis, and pattern recognition [5][6] to help identify children with autism is challenging, as it requires creating a comprehensive program that can detect subtle signs of autism across a range of contexts as in [7]. People with autism may struggle with understanding social cues, interpreting and responding to others' emotions, and forming relationships. They may also have difficulty with processing sensory information or have strong interests in certain topics or activities. Diagnosis is based on observed behavior, and the process can involve interviews and questionnaires, cognitive assessments, physical examinations, and genetic and neurological tests. All of these evaluations can take time and money, and the cost can be prohibitive for some families. These tests are designed to identify patterns of behavior and symptoms associated with autism, by asking parents and professionals to observe the individual. They then analyze the responses and compare them to a set of criteria established to identify autism or other developmental disorders. For example, if a person is using a metal detector, they must have an understanding of the type of metal they are looking for and the size of the object they are searching for. The quality of the metal detector will also have an impact on the accuracy and efficiency of the screening method. Such systems can use algorithms to analyze large amounts of data and detect patterns with high accuracy, potentially leading to earlier and more accurate diagnoses. Additionally, such systems can help to automate certain labor-intensive tasks and reduce the amount of time needed to complete diagnostic tests. This is because machine learning algorithms can analyze large amounts of data and identify patterns and correlations that would be difficult or impossible for humans to find. The algorithms can then be used to develop predictive models that can accurately identify potential diagnoses and suggest

therapies as in [8]. Some research scholar has done some work on ASD diagnosis using machine learning. The aim of this research is to reduce the classification time of ASD diagnosis process after the detection of the most influential ASD diagnosis items as in [9][10][11][12]. Machine learning (ML) is a powerful tool that can be used to analyze vast amounts of data and identify patterns that can be used to detect mental health issues. ML can also be used to develop personalized treatments based on individual patient characteristics. This could potentially lead to more targeted and effective treatments for mental health issues. Through the use of data-driven techniques, ML enables the analysis of large amounts of data to uncover previously unknown patterns, trends, and correlations. ML can be used to develop predictive models or to recommend interventions that may be tailored to individual needs. These challenges include the need to ensure responsible data collection and storage, to develop equitable access to ML-enabled solutions, to ensure ethical and responsible use of ML and AI, and to ensure that privacy and confidentiality are maintained as in [13].

The proposed work is based on the detection of ASD symptoms from the parents' dialogue. Parents of autistic children have the best experience with their autistic children's symptoms. The data has been collected from many social sites and organizations for special children. The data is related to the parents' dialogue in text mode and a dataset has been prepared using these parents' text inputs. Traditional machine learning models like SVM, Logistic Regression, K-nearest neighbor (KNN), and Random Forest have been used to detect the symptoms from the parents' text. The sentiment analysis process has been used to detect sentences from the parents' text. After completion of the prediction using the proposed machine learning models, the positive sentences have been used as input in the cosine similarity model. This model will calculate the cosine similarity of input sentences and ASD symptoms sentences to detect ASD problems. Many machine learning-based applications related to mental disorders have been discussed in Section II. The proposed dataset, detailed architecture of the proposed system, and machine learning models have been discussed in Section III. The results of this proposed system have been discussed in Section IV. The limitation has been given in Section V whereas conclusion has been discussed in Section VI and ends with the future work in Section VII.

II. Related Works

Today, Autism Spectrum Disorder (ASD) is a highly prevalent disorder problem among children. Now it is one of the main components in the healthcare domain and much research has been done using Artificial Intelligence (AI). A few important AI-based research works on Mental Health related issues have been included in this related work section.

These NLP software tools use a combination of natural language processing (NLP) algorithms and domain-specific ontologies to identify and extract biomedical concepts from unstructured texts. The ontologies provide an organized representation of biomedical concepts and the NLP algorithms enable the software to accurately identify the concepts in the text. This is due to the fact that the existing literature on these disorders is often written in a complex, highly technical language that is difficult to parse and interpret with natural language processing tools. Additionally, many of the diseases are multi-faceted and involve a variety of clinical terms that need to be identified by the NLP tools in order to accurately extract relevant information. The authors evaluated the predictive performance using precision, recall, and F1 score. We also ran a manual evaluation to compare the manual annotation of ASD-related terms with the tools' extracted terms, and found that CLAMP outperformed the other two tools in terms of precision, recall, and F1 score on both the abstracts and full-text articles. The F1 score combines the precision and recall of a system, so it takes into account both the accuracy and completeness of the system. In this case, CLAMP had the highest F1 score, meaning it had both a higher precision and a higher recall than the other two systems. This type of analysis protocol allows researchers to better identify, classify, and quantify the symptoms of a disorder, even when there is not a well-defined terminology set to describe it. This makes it easier to compare the presentation of the disorder across different populations and can help to identify potential biomarkers for the disorder as in [14]. People with ASD had more difficulty in expressing emotions and abstract concepts than typically developing individuals, as well as difficulty in using language to describe events and convey information. This suggests that impairments in the use of pragmatic language are an important aspect of ASD and should be addressed in interventions. This suggests that the differences in narrative production between ASD and control groups are related to difficulties in understanding and expressing emotions, as well as producing more abstract language. The individuals with typical development had a more varied range of vocabulary, which included more words with both positive and negative sentiments, while the participants with ASD displayed a limited vocabulary, resulting in a greater tendency to use negative words. The lower level of language abstraction in the ASD narratives could be due to the limitation of their vocabulary and the difficulty of expressing abstract concepts. This suggests that language abstraction and emotional polarity can be used to measure the narrative abilities of individuals with ASD without relying on age or IQ scores. The strong positive correlation between linguistic abstraction and emotional polarity indicates that the more abstract language used, the more likely it is to contain emotional content. The difference in emotional polarity between the two groups could be due to the fact that individuals with ASD may have difficulty recognizing and expressing emotions. In addition, they may have difficulty understanding abstract language concepts, which could explain why they used fewer abstract words in their narratives as in [15]. One of the most promising areas for developing assistive tools is the use of artificial intelligence (AI) and machine learning (ML) algorithms. These algorithms can be used to analyze data from various sources and can provide insights that may help diagnose ASD earlier and more accurately. The proposed approach is expected to find the underlying patterns in the eye-tracking records which can be used to accurately diagnose the disorders. The results of this study could provide clinicians with a powerful tool that could potentially improve the

accuracy and speed of diagnosis. By applying NLP methods to the raw eye-tracking data, the study was able to extract meaningful features from the data that could be used to train classification models. The experiment showed that using these features could yield better results than using the raw data alone. The authors [16] used a customized loss function to adjust the weights of the model, which allowed them to achieve a high level of accuracy. Additionally, authors [16] utilized transfer learning to fine-tune the model, allowing us to further improve the accuracy of the model. The author's [16] approach could realize a promising accuracy of classification (ROC-AUC up to 0.8) as in [16]. Social behavior issues are often the most noticeable in children with autism, and they may include difficulty forming relationships, lack of eye contact, and difficulty understanding nonverbal communication. Clinical tests can also be used to look for developmental delays, such as difficulty with speech and language, as well as repetitive behaviours like hand flapping or rocking. The assessment process is designed to identify key characteristics of autism in individuals, such as difficulty in communication and social interaction, and to determine the severity of the condition. By using semi-structured data posted in Twitter, the team of doctors can gain insight into the individual's behavior, which can then be used to develop a more accurate and effective assessment. Analyzing the tweets, it allows researchers to detect the sentiment of people's opinions on autism, the topics that are most commonly discussed, and the language used to discuss autism. This helps researchers gain a better understanding of how people think and talk about autism, and can help inform policy decisions NLP and topic modeling allow for more efficient processing of data by automatically recognizing patterns and keywords, saving time and effort. Furthermore, the results of the analysis are highly accurate, making them an ideal choice for studying topics such as genetic analysis, the effect of vaccination, and behavior analysis. The 10k tweets dataset is enough to provide in-depth analysis and insight into these topics. The analytical results are used to learn the genetic impact on ASD, the vaccination effect on ASD and also used to learn the behavior changes and population of autistic children as in [17]. It is characterized by a persistent pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development. It is often accompanied by other mental health disorders, such as anxiety and depression, which can further impair functioning and quality of life. We applied the CNN model to the EEG data in order to distinguish between ADHD patients and healthy controls. The CNN was able to accurately classify the EEG data with an accuracy of 90.3%, significantly outperforming other methods, particularly of event-related potentials (ERP) from ADHD patients (n = 20) and healthy controls (n = 20) collected during the Flanker Task, with 2800 samples for each group. By exploiting invariances, deep networks are able to classify data even when there are variations in the data, such as changes in lighting or orientation of an image. Compositional features are combinations of basic elements that form a more complex representation of the data, such as edges and shapes in an image. Deep networks are able to identify these features, which enables them to accurately classify data. This was achieved by using a Convolutional Neural Network (CNN) that was trained on EEG data from patients with Parkinson's Disease in order to classify them as either having the disease or not. The CNN was able to extract relevant features from the data without any manual input, resulting in a higher accuracy than other machine learning approaches. This is because CNNs can learn more complex patterns from the data and have the ability to generalize to new data. Event-related spectrograms capture more information about the events of interest, which can be used to extract more accurate features than resting state EEG spectrograms. This suggests that these techniques can be used to identify and visualize the underlying physiological differences between neurological disorders and healthy brains, potentially leading to a better understanding of their underlying pathophysiology. Deep networks are useful because they can extract meaningful patterns from EEG signals and are capable of handling large amounts of data. These results suggest that deep networks can also be used to analyze EEG dynamics from smaller datasets, which could be used to develop biomarkers for clinical use as in [18]. EEG can provide valuable information to help diagnose ADHD in children because it can measure electrical activity in the brain and detect any abnormal electrical activity that may be indicative of ADHD. Additionally, EEG can help to differentiate ADHD from other mental disorders that may be present in the child. Symptoms of ADHD include difficulty paying attention, impulsivity, and hyperactivity. These symptoms can interfere with a child's ability to learn, manage emotions, and interact with peers. Video long-range EEG monitoring can provide more accurate and detailed information about the brain activity of children with ADHD compared to ambulatory EEG monitoring, as it allows for more frequent data collection and better visualization of the EEG data. It also helps to identify abnormal brain electrical activities which may be associated with ADHD, thus aiding in the diagnosis of the condition. By doing this, they were able to accurately identify children with ADHD and study their behavioral patterns in order to better understand and treat the disorder. This allowed for a more precise and detailed analysis than traditional methods of observation. Comparing the results of various models can help to identify which model is best suited for recognizing signs of ADHD in EEG data. By selecting the most accurate and appropriate model, researchers can then use it to build a recognition method that can diagnose children with ADHD more accurately. This is because long-term video EEG can detect the abnormal EEG patterns associated with ADHD, such as slow wave activity, and can also detect the degree of attention fluctuation in children with ADHD as in [19]. With the recent advances in artificial intelligence, computers can now analyze EEG data and provide results much faster than a neurologist. This has enabled the field of neurology to become much more efficient and provide more accurate results in a fraction of the time. This is made possible because AI is able to quickly analyze and process large amounts of data. It can quickly identify patterns and draw conclusions from the data that would take human hours or even days to detect ADHD. Additionally, AI can look for indicators of diseases or abnormalities that would be difficult for humans to find on their own. This is because it can automate the process of analyzing EEG signals, thus allowing neurologists to quickly and accurately identify

patterns associated with different neurological diseases. Furthermore, this technology can also help neurologists to identify subtle changes in EEG signals that could potentially signal the onset of a neurological disorder. The ML model can process the EEG signals quickly and accurately to detect patterns that may indicate ADHD. By making use of the data generated from the EEG signals, the ML model can diagnose ADHD more accurately and quickly than traditional methods. By analyzing the EEG signals, the ML model can identify patterns that are indicative of ADHD. Additionally, the ML model can be trained to recognize these patterns more quickly and accurately than traditional methods. With the right pre-processing techniques and machine learning algorithms, the ML model can provide a more accurate diagnosis of ADHD than traditional methods as in [20]. This allows individuals to stay connected with their friends and family and to keep up with what is going on in the world. Additionally, it makes it easier to stay in touch with people who are not in the same physical location, making it a great way to stay connected during this time. The pandemic has had a negative impact on the mental health of many people, and it has become harder for them to access in-person support. As a result, online tools and resources have become more important than ever for those struggling with mental health issues, allowing them to get the help they need even when they are unable to leave their homes. Mental health conditions can have a significant impact on an individual's overall well-being, affecting their ability to work and their relationships with others. Additionally, research has found that mental illnesses can increase an individual's risk of developing chronic physical health conditions, such as heart disease and diabetes. AI methods can help mental health providers to detect patterns in patient data that might otherwise go unnoticed, as well as to generate insights into the patient's current state. This can lead to more accurate diagnoses and better treatment plans, leading to better overall outcomes for the patient. AI can help to analyze patient data quickly and accurately, identify patterns and correlations, and make predictions about the best course of action for a patient's diagnosis and treatment. AI can also help reduce the time and resources required for manual data analysis and provide more efficient and cost-effective care. The models were tested on a labeled dataset of Reddit posts from users with self-reported mental illnesses and compared against a baseline model. The results showed that the machine learning, deep learning, and transfer learning models outperformed the baseline model in correctly classifying the different mental illnesses. This will help to reduce the amount of time it takes to identify and respond to medical emergencies, which will ultimately lead to more lives being saved. Additionally, it will also help to reduce the burden on healthcare workers, which will make the public health system more efficient and cost-effective as in [21]. A variety of factors can contribute to depression, such as genetics, brain chemistry, environmental influences, traumatic experiences, and other medical conditions. Additionally, depression can be caused by a combination of these factors, making it difficult to pinpoint a single cause. Genetics and brain chemistry can predispose someone to depression, while environmental factors and traumatic experiences can trigger its onset. Other medical conditions such as chronic illnesses can also be associated with depression. Recognizing the early signs of depression can help to identify and address the issue before it becomes a more serious problem. The CNN is used to extract high-level features from speech signals, while the SVM classifier is used to classify the extracted features. The hybrid model is trained on a dataset of Arabic speech from people with depression and those without, to produce a model that is capable of distinguishing between the two. The hybrid model uses a combination of convolutional neural networks (CNNs) and support vector machines (SVMs) to analyze while 30% of data were used to test the proposed model. A hybrid model (CNN + SVM) attained a 90.0% and 91.60% accuracy rate to predicting the depression from the data and make predictions. This combination of techniques allows for the model to process the data quickly and accurately, resulting in the high accuracy rates it achieved. This is likely because the hybrid model combines the strengths of both models. The RNN can accurately make predictions based on the context of the data, while the CNN can detect the most important features in the data. By combining both models, the predictive power of the hybrid model is enhanced ,e RNN achieved an 80.70% and 81.60% accuracy rate. This indicates that the combined model was more effective in classifying depression than either of the individual models alone. The results suggest that incorporating multiple models into one prediction system can increase the accuracy of the diagnosis. This is because the achieved findings can be used to identify key indicators of depression in spoken Arabic, such as speech patterns, intonation, and pauses. These indicators can then be used to identify individuals who may be suffering from depression and help physicians, psychiatrists, and psychologists provide more effective treatment as in [22]. The mental health issues, such as depression and anxiety, are becoming more common, and people are recognizing the need to prioritize their mental health as well as their physical health. Additionally, with the development of telehealth services, it's become easier for people to access mental health services regardless of their location. This means that most people who suffer from mental health issues are unable to get access to the right diagnosis and treatment, resulting in an overall decrease in the mental health of the population. The model will be trained on a dataset of speech samples from people with and without depression. Exploring the acoustic features and patterns in the speech samples of people with depression will help to identify the differences between those with and without depression. By doing so, it will be possible to detect signs of depression in an individual and provide an initial diagnosis of mental health problems. This model uses Natural Language Processing (NLP) techniques to analyze the text and determine the sentiment of the posts. The sentiment of the posts is then used to assess an individual's mental health status as in [23].

A comparative analysis has been done on proposed systems that are equipped with machine learning models and similar types of systems that are also based on machine learning models. Table I contains 'Models' as the first attribute where each model name is defined. The 'Description' attribute contains details about the models. The third attribute is 'Dataset' which refers to the dataset details and the fourth attribute is 'Accuracy' where each model's accuracy has been given. The last attribute is 'Remarks' about each model. Fig. 1

shows the accuracy graph of similar machine learning models and proposed machine learning models.

TABLE I. COMPARATIVE STUDY OF PROPOSED MODELS WITH SIMILAR TYPE MODELS IN MENTAL DISORDERS

Sl.No.	Models	Description	Dataset	Accuracy	Remarks
	Similar Type Machine Learning Models in Mental Disorders				
1	CNN, RNN, SNN [18]	Deep learning CNN, Recurrent Neural Network, and Recurrent Neural Network are used for classification and comparison to detect Attention deficit hyperactivity disorder (ADHD).	EEG data has been used.	88%, 86%, and 78%	EEG is a medical test that measures electrical activity in the brain. This data is a very high volume and time and cost-effective.
2	Fully connected neural network model [19]	Neural Network-based Deep Learning Model to detect disorders like ADHD.	Deep learning long-range EEG big data.	97.7%	The data is long-range EEG big data which is a very high volume data for analysis.
3	KNN, SVM, and RF [20]	KNN, SVM, and RF Models are used trained with the EEG signals data to detect ADHD.	EEG signals data of ADHD	69%, 72%, and 74%	Much time has to be given for preprocessing to improve the quality of EEG signals.
4	Linear Support Vector Classifier, LR, NB, and RF [21]	Depression, anxiety, bipolar disorder, ADHD, and PTSD detection from unstructured data.	Unstructured user data on the Reddit platform has been used.	79%, 79%, 74%, and 75%	Reddit's post-dataset cleaning process is related to removing personal information, punctuation marks, and URLs.
5	CNN +SVM[22]	Intelligent system to detect depressive symptoms using speech analysis	Basic Arabic Vocal Emotions Dataset (BAVED)	90 and 91.60	The dataset has been prepared from the audio format for sentiment analysis.
6	RNN+CNN [22]	Intelligent system to detect depressive symptoms using speech analysis	Basic Arabic Vocal Emotions Dataset (BAVED)	88.50 and 86.60	The dataset has been prepared from the audio format for sentiment analysis.
	Proposed Models in Mental Disorder (Autism Spectrum Disorder)				
7	Proposed SVM	SVM model to predict positive ASD symptoms from parents' dialogue.	Parents' Dialogues of Autistic Children in text format from SAHAS- Durgapur, India, and Social Sites.	71%	The data has been collected in text form. The parents' dialogues about their autistic children are very useful because they shared their experiences and thoughts about their autistic children. A parent of an autistic child is the best source to understand the ASD symptoms patterns.
8	Proposed Logistic Regression	SVM model to predict positive ASD symptoms from parents' dialogue.	Parents' Dialogues of Autistic Children in text format from SAHAS- Durgapur, India, and Social Sites.	71%	The data has been collected in text form. The parents' dialogues about their autistic children are very useful because they shared their experiences and thoughts about their autistic children. A parent of an autistic child is the best source to understand the ASD symptoms patterns.
9	Proposed K Nearest Neighbor (KNN)	SVM model to predict positive ASD symptoms from parents' dialogue.	Parents' Dialogues of Autistic Children in text format from SAHAS- Durgapur, India, and Social Sites.	62%	The data has been collected in text form. The parents' dialogues about their autistic children are very useful because they shared their experiences and thoughts about their autistic children. A parent of an autistic child is the best source to understand the ASD symptoms patterns.
10	Proposed Random Forest	SVM model to predict positive ASD symptoms from parents' dialogue.	Parents' Dialogues of Autistic Children in text format from SAHAS- Durgapur, India, and Social Sites.	69%	The data has been collected in text form. The parents' dialogues about their autistic children are very useful because they shared their experiences and thoughts about their autistic children. A parent of an autistic child is the best source to understand the ASD symptoms patterns.

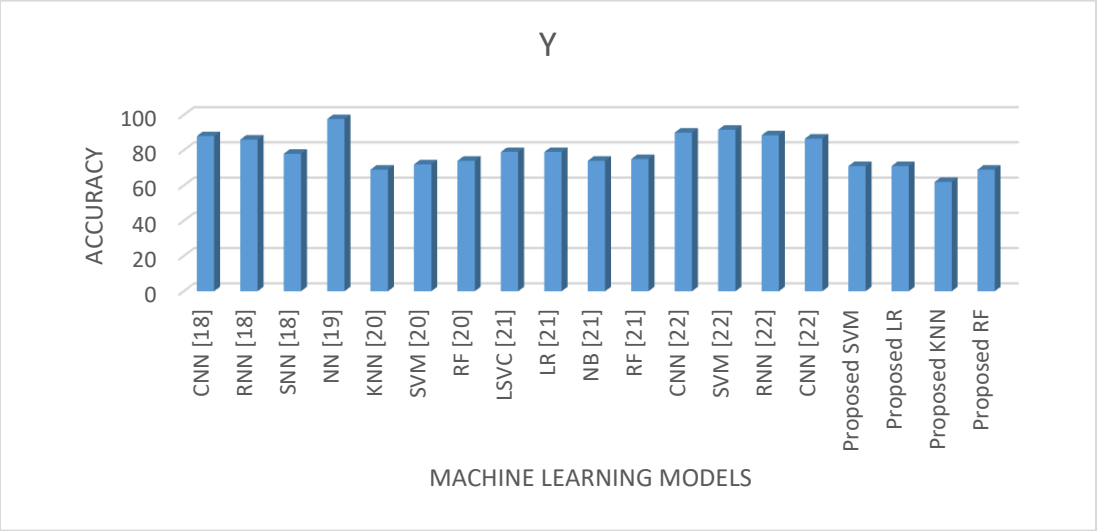

Fig. 1. Accuracy graph of similar ML models and proposed ML models.

III. Architecture of Proposed Models

A few traditional machine learning classifiers have been used to identify ASD symptoms from parents' dialogues. SVM has been used as the first classifier to identify the symptoms from the parents' dialogue. Logistic regression is a second classifier that is also identifying the ASD symptoms from the given dataset. KNN and Random forest are the last two classifiers that are also used to identify ASD symptoms from a given dataset.

A. Dataset of Proposed System

The Dataset has been prepared using the parents' dialogue where parents are describing their thoughts and experiences about their own autistic child. These data have been collected from several different social networks and organizations where special children are taking their therapies on communication, speech, and behavior. A few parent dialogue example has been given in Table II. Parents' dialogues are very important data from where all possible symptoms of ASD can be identified. The given dialogues are used to make the dataset for proposed machine learning models training and testing.

TABLE II. Example of Parents' Dialogues

Sl. No.	Parents' Dialogues
1.	My second son is 4 and also autistic; he's on the move always and always into something and he's also a big momma's boy, loves hugging and cuddling me. I'm nervous about bringing baby home. Idk how he'll handle it. Any advice?
2.	Hi. Please I need some advice. My son is 10 and from a few years is very hard to make him do some activities (writing and staff like that) At school he refuse. They are not able to make him do anything. At school just play and if say no to him he just scream. He doesn't want to do anything; (in terms of studying or activities). I really don't know what to do.
3.	I'm currently having problems with washing my (almost 2 year old) daughter's hair. Whenever i try, she basically goes ballistic and throws a fit. She's scared and I'm trying to figure out how to support her and make her feel safe because she does have to get hair washed. Any suggestions and things that have worked for you?
4.	My youngest with autism, learning disabilities and is non-verbal, will be 4. She has to be in a pushchair whilst out and about for safety as has zero sense of danger. I'm struggling to find a double pushchair suitable for a newborn and my will be 4 year old. If anyone can send any links or pictures that would be great.
5.	From few days my son eye movements strangely like keeping head down n seeing up and moving eye balls to the corners of the eyes. Can anyone suggest why he is doing so? Please... thanks!

The Dataset has been prepared from the text in Table II. Each sentence has been taken into consideration to identify whether it is a symptom of ASD or not. There are no fixed symptoms in ASD for identification. Increment of those parents' dialogues who are actually parents of autistic children can be a good idea to identify more symptoms as well as a good advantage to train the machine learning models for better accuracy. A few examples of data from the proposed dataset have been given in Table III.

TABLE III. Example Data in the Proposed Dataset

Sl. No.	Comments	Sentiment
1.	because all they do there is play with toys with him every time	1
2.	I'm confused guys help my son is 3years old now	0
3.	My little girl is 3 and a half and still non verbal	1
4.	he does is mumbles only no proper words	1
5.	I was really surprised when he came home with iep papers	0

The dataset structure in the proposed research has been described in Table III where the first column is Serial Number, the second column is Comments, and the third column is Sentiment. Paragraph text from parents' dialogues has been taken to prepare the dataset. Each sentence has been taken from the paragraph text and identifies whether it is a symptom of ASD or not. If it is a symptom of ASD then it is labeled as 1 (true) otherwise 0 (false). According to Table III, Sentences in the Comments column with serial numbers 1, 3,

and 4 are true symptoms of ASD whereas serial numbers 2 and 5 are false symptoms. Now this ASD symptom-based dataset has been prepared to train some traditional machine learning models like SVM, Logistic Regression, KNN, and Random Forest.

TABLE IV. LIST OF LABELS WITH ASD PROBLEMS

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

Table IV shows that each ASD problem is associated with the label. Label 1 denotes the "Speech Problem" whereas Label 2 and 3 denotes the "Sensory" and "Behaviour" problems. The other problems also mention in the label in Table IV. This table has been used after the prediction of the sentiment of a sentence according to the ASD symptoms. If the sentence is positive (1) then the proposed system will use this positive sentence as input of the Spacy cosine similarity model. Table V shows a dataset that contains a number of positive sentences with labels. Each label indicates an ASD problem according to Table V. Each sentence will be used for a similarity check with predicted positive sentences in the cosine similarity model and that has been discussed in the Proposed System Flow sections.

TABLE V. DATASET FOR COSINE SIMILARITY CHECK

Sl. No.	Positive Sentences	Label
1.	I can't show him how to potty during the day while his dad is at work	7
2.	he does is mumbles only no proper words	1
3.	when I call him he doesn't come to me	10
4.	He needs to visualize what I'm saying	6
5.	She gets so frustrated it breaks my heart I guess I'm looking for success stories	9

Each model has been described with the proposed algorithm in the next sections where this dataset has been utilized to train these models and the result of each model has been discussed in the Result and Discussion section.

B. Support Vector Machine (SVM)

Support vector machine (SVM) is the first approach to identify the symptoms of ASD. SVM is a supervised machine learning algorithm that can be used for classification or regression problems. It is a good idea to use SVM on a small dataset and it generates good predictive results according to the problem. SVM is based on the finding of the best hyperplane that divides data points either in two classes or multiclass. The proposed approach is binary classification where data points either true (1) or false (0).

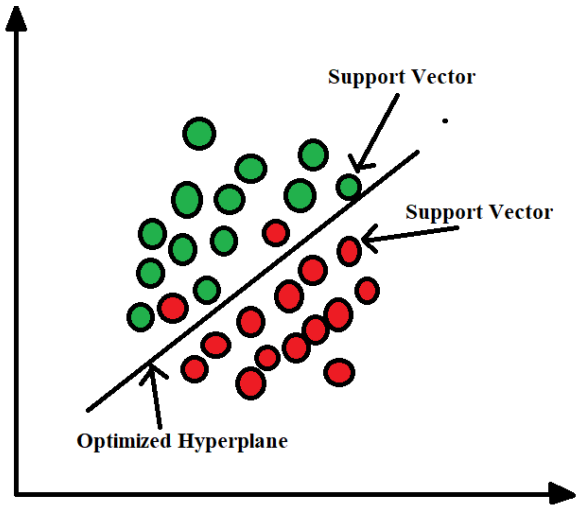

Fig. 2. Support Vector Machine (SVM).

It can be observed according to the above image that it is a two features classification problem. The optimized hyperplane has been drawn to linearly separate support vectors. The support vectors can be seen as red and green circles in Fig. 2. It is a binary classification problem where the SVM algorithm draws many lines to separate vectors according to true and false. After optimization, the SVM algorithm returns the best-fitted line for classifying the support vectors.

According to the equation of hyperplane:

w.X+b=0 Where X is a vector and w is a vector normal to hyperplane and b is an offset value.

The decision rules have been applied to classify the positive and negative value.

$$\vec{X} \cdot \vec{w} - c \geq 0$$
$$\text{putting } -c \text{ as b, we get}$$
$$\vec{X} \cdot \vec{w} + b \geq 0$$

Hence,

$$y = \begin{cases} +1 & \text{if } \vec{X} \cdot \vec{w} + b \geq 0 \\ -1 & \text{if } \vec{X} \cdot \vec{w} + b < 0 \end{cases}$$

According to the above equation as in [24], the value w.X+b>0 then it will be detected as a positive value (1) otherwise it will be a negative value (0). The proposed algorithm is used the ASD symptoms dataset to train the SVM model. The proposed algorithm to train the SVM model for the prediction of ASD symptoms has been given below.

Proposed SVM Algorithm:
Pseudo Code:
Step 1: Read data from csv file.

Step 2. X=data from csv

$x_1=[a_1,a_2,a_3,a_4,a_5,\ldots\ldots a_n]$ is a user text column inside the dataset.

$x_2=[r_1,r_2,r_3,r_4,r_5,\ldots\ldots r_n]$ is a label data column inside the dataset.

Step 3. Split the dataset as train data and test data.

```
train, test = train_test_split(X, test_size=0.2, random_state=1)
X_train = train['text'].values
X_test = test['text'].values
y_train = train['label']
y_test = test['label']
```

Step 4. Define NLP functions to pre-process text from X_train and X_test.

```
// Text tokenization
 tokenize_text=tokenizer(text)
// Stop Words removal from text
fresh_text = stopwords.words(text)
// text to vector conversion using vectorization method
vectorizer = CountVectorizer(
                analyzer = 'word',
                tokenizer = tokenize_text,
                lowercase = True,
                ngram_range=(1, 1),
                stop_words = fresh_text)
```

Step 5. Call method to train SVM model.

```
// kfolds has been used to send data as a bunch into the SVM model.
kfolds = StratifiedKFold(n_splits=5, shuffle=True, random_state=1)
// Make the pipeline to send data inside the SVM model.
pipeline_svm = make_pipeline(vectorizer, SVC(probability=True, kernel="linear", class_weight="balanced"))
// SVM model initialization with parameters
grid_svm = GridSearchCV(pipeline_svm,
            param_grid = {'svc__C': [0.01, 0.1, 1]},
            cv = kfolds,
            scoring="roc_auc",
            verbose=1,
            n_jobs=-1)
// fit data inside the model to train
grid_svm.fit(X_train, y_train)
```

Step 6. Predict the result using SVM model.

```
model= grid_svm.best_estimator_
prediction = model.predict(X_test)
```

The result of this proposed algorithm has been discussed in the Result and Discussion section.

C. Logistic Regression

The next approach is logistic regression which is able to identify ASD symptoms from user text. This is another machine-learning algorithm for binary classification problems. The logistic regression model works on finding the value between 0 and 1 and this algorithm is bounded. The logistic regression does not contain any relationship between input and output variables because of the nonlinear transformation to the odds ratio. Logistic regression can be defined as-

$$\text{Log}(p(M)/1-p(M))=\beta_0+\beta_1 X$$

p()-> refers to the probability function.

M-> refers to the input

Where *p(M)/1-p(M)* is in the left side is termed as odds and the left side is called logit. The odds are the ratio of chance of success according to the chance of failure. In logistic regression, the linear input combination is transformed to log(odds).

The inverse of the above function will be: $p(M)=(e^{\beta 0+\beta 1x} / 1+ e^{\beta 0+\beta 1x})$

This function is a sigmoid function that can be produced an S-shaped curve and it returns a value between 0 and 1. The main work of the sigmoid function is to generate a probability value from the expected value and this value always will be bounded between 0 and 1. The mathematical representation of the sigmoid function can be - $f(m) = 1/(1+e^{-m})$

Fig. 3 shows the S-shape curve according to the function- $f(m) = 1/(1+e^{-m})$

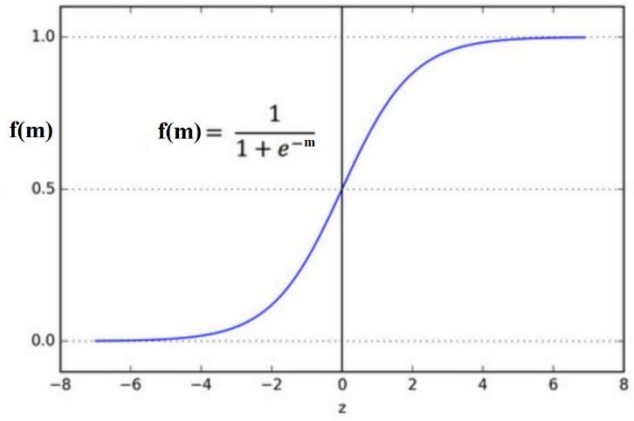

Fig. 3. Sigmoid function according to the equation.

The proposed algorithm which is based on logistic regression has been given below.

Proposed Logistic Regression Algorithm:
Pseudo Code:
Step 1: Read data from CSV file.
Step 2: X=data from csv

$x_1=[a_1,a_2,a_3,a_4,a_5,\ldots\ldots a_n]$ is a user text column inside the dataset.

$x_2=[r_1,r_2,r_3,r_4,r_5,\ldots\ldots r_n]$ is a label data column inside the dataset

Step 3: Features generation using Vectorizer function.

```
// Vectorizer function converts the string value to number values.
vectorizer = CountVectorizer(
                analyzer = 'word',
                lowercase = False,)
// Feature creation using vectorizer.fit_transform function
features = vectorizer.fit_transform(x1)
// Feature array creation
features_nd = features.toarray()
```

Step 4: Model creation and training

```
//Logistic model creation
log_model = LogisticRegression()
// Logistic model train
log_model = log_model.fit(X=X_train, y=y_train)
```

Step 5: Prediction using Logistic Regression model
y_pred = log_model.predict(X_test)

The output as a result of this proposed algorithm has been discussed in the Result and Discussion section.

D. K-Nearest Neighbor (KNN)

The third approach to identifying ASD symptoms from user text. KNN is a supervised algorithm that can be used in classification problems. This algorithm uses feature similarity to predict the value for a new data point that comes as input. KNN uses the similarity between new data points with available categorical data points and identifies this data point in a particular similar data point's category. KNN is very popular in binary classification. Fig. 4 shows before KNN prediction the new data point plotted on a graph where two categories of data points are present. Category A and Category B have been classified according to the nearest data points. According to Fig. 5, after applying the KNN algorithm, the new data point has been assigned as Category B because the nearest neighbor of the new data point is the data point of Category B.

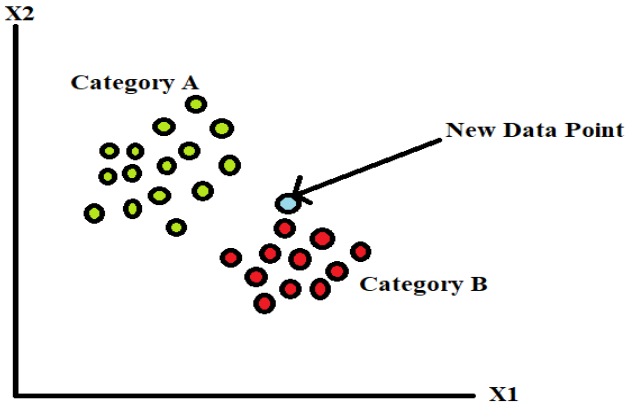

Fig. 4. Before the KNN algorithm is applied on a new data point.

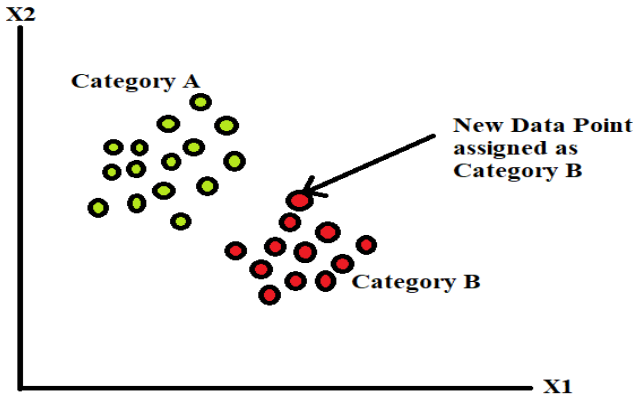

Fig. 5. After the KNN algorithm applied on a new data point.

K is a parameter in KNN that is related to the number of nearest neighbors that are used to count the majority process. The first step of KNN is to transform data points into vectors where the KNN algorithm will calculate the distance of these vectors. KNN computes the distance of each data point of training data then it will calculate the probability of a new data point is similar to the training data. Euclidean, Minkowski or Hamming distance functions can be used here to calculate the distance. The proposed algorithm is based on KNN that uses the proposed dataset.

Proposed KNN Algorithm:
Pseudocode:
Step 1: Read data from CSV file.
Step 2: *X=data from csv*
$x_1=[a_1,a_2,a_3,a_4,a_5,\ldots\ldots a_n]$ is a user text column inside the dataset.
$y_1=[r_1,r_2,r_3,r_4,r_5,\ldots\ldots r_n]$ is a label data column inside the dataset
// Split the data in train and test format
x_train,x_test,y_train,y_test=train_test_split(x_1, y_1,stratify= y_1,test_size=0.33)
Step 3: String value to Vectorizer transformation.
// Vector function declaration
vectorizer=CountVectorizer()
// Vector transformation of x_train
x_train_bow=vectorizer.fit_transform(x_train)
// Vector transformation of y_train
x_test_bow=vectorizer.transform(x_test)
Step 4: KNN Model creation
grid_params = { 'n_neighbors' : [40,50,60,70,80,90],'metric' : ['manhattan']}
knn=KNeighborsClassifier()
Step 5: KNN model training using prepared dataset
clf = RandomizedSearchCV(KNN, grid_params, random_state=0,n_jobs=-1,verbose=1)
clf.fit(x_train_bow,y_train)
Step 6: Prediction using KNN model
Prediction=clf.predict_proba(x_test_bow)

The result of this proposed KNN-based algorithm has been discussed in Result and Discussion section.

E. Random Forest

The last approach is a Random forest machine learning algorithm to identify the ASD Symptoms from user text. This is one of the important machine learning algorithms which is constructed from decision tree algorithms. The Random forest algorithm is used to solve regression and classification problems. This algorithm is trained through bagging which is an ensemble algorithm. The ensemble algorithm is used to improve the accuracy of the machine learning algorithms. The outcomes of the random forest are based on the prediction of the decision tree. The mean of various decision trees is used to calculate the prediction value by the random forest algorithm. Decision trees in random forest algorithms use the tree view to generate prediction value from a series of feature-based splits where it starts from a root node and ends in a leaf node with a decision. Feature selection and the splitting process is depending on the impurity which means either result will be 'yes' or 'no'. To know about the impurity of the dataset, the Gini index [25] is a good option and that can be written mathematically-

$$\text{Gini Index} = 1 - \Sigma (P_i)^2$$
$$= 1 - [(P_+)^2 + (P_-)^2]$$

Where P_+ is denoted as a probability of positive class and P_- is denoted as a probability of negative class. Gini Index will find out all the possibilities of splits and will choose the root node and this root node will be a low impurity means the lowest Gini index.

The proposed random forest-based algorithm has been given below which is utilizing the proposed dataset to train the model for the prediction of ASD symptoms from user text.

Proposed Random Forest algorithm:
Pseudo code:
Step 1: Read data from CSV file.
Step 2: X=data from csv
$x_1=[a_1,a_2, a_3,a_4,a_5,\ldots\ldots\ldots a_n]$ is a user text column inside the dataset.
$x_2=[r_1,r_2, r_3,r_4,r_5,\ldots\ldots\ldots r_n]$ is a label data column inside the dataset
Step 3: Features generation using TFIDF Vectorizer function.
// Split data and assign for training and testing purpose
X_train, X_test, y_train, y_test = train_test_split(x_1, x_2,test_size = 0.90, random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X_train,y_train,test_size = 0.5, random_state=42)
X_val, X_test, y_val, y_test = train_test_split(X_test,y_test,test_size = 0.5, random_state=42)
// TFIDF Vectorizer Function declaration
def vectorize(data,tfidf_vect_fit):
X_tfidf = tfidf_vect_fit.transform(data)
words = tfidf_vect_fit.get_feature_names()
X_tfidf_df = pd.DataFrame(X_tfidf.toarray())
X_tfidf_df.columns = words
return(X_tfidf_df)

tfidf_vect = TfidfVectorizer(analyzer=clean)
tfidf_vect_fit=tfidf_vect.fit(X_train['text'])
X_train=vectorize(X_train['text'],tfidf_vect_fit)
Step 4: Random Forest model initialization
model=RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
criterion='gini', max_depth=20, max_features='auto',
max_leaf_nodes=None, max_samples=None,
min_impurity_decrease=0.0,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=100,
n_jobs=None, oob_score=False, random_state=None,
verbose=0, warm_start=False)
X_val=vectorize(X_val['text'],tfidf_vect_fit)
rf1 = RandomForestClassifier(n_estimators=100,max_depth=20)
rf1.fit(X_train, y_train.values.ravel())
Step 5: Prediction of Random Forest Model.
Prediction = model.predict(X_val)

The result of this algorithm has been discussed in the Result and Discussion section.

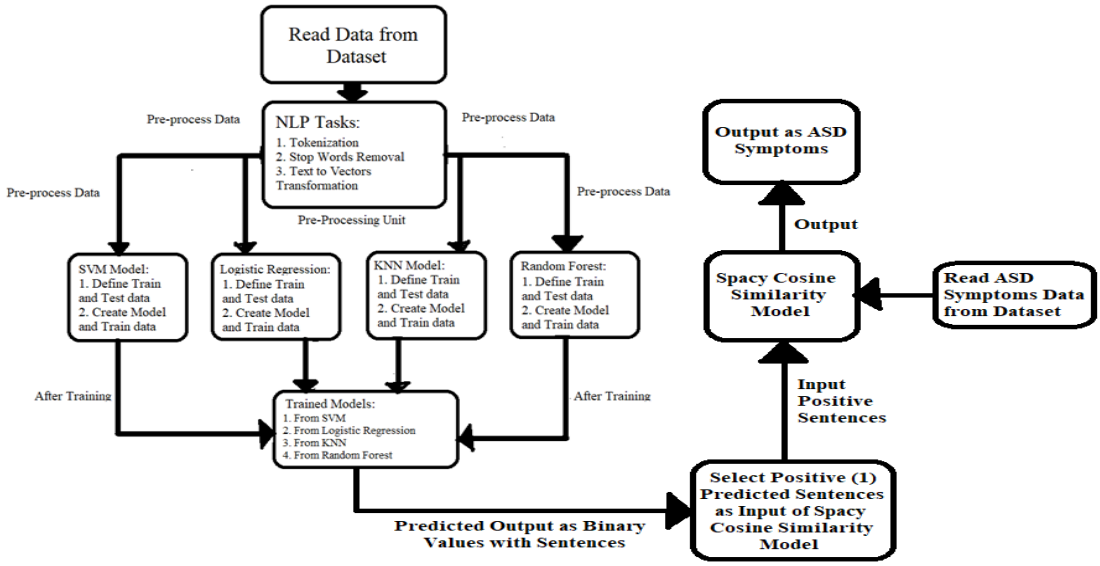

Fig. 6. Flow diagram of Proposed System Architecture

F. Proposed System Flow

Fig. 6 shows the overall architectural diagram of the proposed system to identify ASD symptoms from user text. The proposed system will read data from the ASD symptoms dataset in the first step. Each sentence will be passed through some NLP tasks like tokenization, stop words removal, and text-to-vector transformation. Sentences are tokenized by the tokenization process of NLP where stop words mean unwanted words (tokens) like 'am','is',a','an', etc. are removed from the sentence. The final task is to transform each token into vectors. These vectors are the main input in each machine-learning model with labeled data. After vector transformation, data are separated into two parts which are training and testing data. According to Fig. 6, SVM, Logistic Regression, KNN, and Random Forest models are trained with the training data, and testing the prediction results with test data. After completion of the model training and testing, the proposed system is ready to accept new paragraph text from the user to identify a number of positive sentences from the given text that denotes ASD symptoms. The predicted sentences will be in two modes either it will positive (1) or negative (0). The proposed system will select only the sentences that are positive and the negative sentences will be discarded in the next step. The selected positive sentences will be the input to the Spacy Cosine Similarity Model. This model will read each positive sentence from the ASD symptoms dataset (Table V) and calculate the cosine similarity with the input sentence. The Spacy cosine similarity model will check a sentence that has the highest cosine similarity score with the input sentence and the Label will be selected of this sentence by the system. The Label will indicate the ASD problem according to Table IV. Each input sentence will be handled by

this cosine similarity model to identify ASD problems. The algorithm has been given below.

```
Proposed Cosine Similarity algorithm:
Pseudo code:
Step 1: // Declare Python and Spacy packages
import spacy
import pandas as pd
nlp = spacy.load('en_core_web_lg')
// Initialize positive ASD symptoms data in a Dataframe
Step 2: df = pd.read_csv("ASD_Smptoms.csv")
// Three list variable has been declared to store each cosine
   similarity value with sentence and label
comments=[]
sentiment=[]
cosine_value=[]
Step 3: Define Cosine Similarity Calculation Method
def Spacy_Cosine(strs):
  for ind in df.index:
    sen1 = nlp(df['Comments'][ind])
    sen2 =nlp(strs)

    sen1_no_stop_words = nlp(' '.join([str(t) for t in sen1 if not
     t.is_stop]))
    sen2_no_stop_words = nlp(' '.join([str(t) for t in sen2 if not
    t.is_stop]))

    comments.append(df['Comments'][ind])
    sentiment.append(df['Sentiment'][ind])

    score=sen2_no_stop_words.similarity(sen1_no_stop_words)
    # score=sen2.similarity(sen1)
    cosine_value.append(score)

    dfc=pd.DataFrame(
      {'Comments': comments,
        'Sentiment': sentiment,
        'Cosine_Scores': cosine_value
      })

dfc.to_csv(r'ASD_Cosine_Data.csv')
  dfc['Cosine_Scores']=dfc['Cosine_Scores'].astype('float64')
  i = dfc['Cosine_Scores'].idxmax()

    return dfc['Sentiment'][i]
Step 4: // Select only predicted positive (1) sentences as input
Strs= List of predicted positive sentences
for st in strs['Comments']:
    result=Spacy_Cosine(st)
    print(st,"=",result)
```

The output of this proposed algorithm has been given and discussed in Result and Discussion section.

IV. Result and Discussion

The proposed system uses multiple traditional machine learning models which are SVM, Logistic Regression, KNN, and Random Forest. The proposed dataset has been utilized to train and test these models. The result of each model according to the dataset has been discussed here one by one.

A. Result and Discussion of SVM Model

Table VI has been given here to show the SVM model metrics after training and testing.

TABLE VI. SVM Model Metrics

Sl. No.	Metrics	Value
1.	AUC	0.77
2.	F1	0.74
3.	Accuracy	0.71
4.	Precision	0.71
5.	Recall	0.77

The SVM model has multiple metrics to understand the model's performance and scalability. According to Table VI, the AUC score is 77% which is a good score for any trained SVM model. The AUC refers to the area under the ROC curve that is a popular metric of SVM. If AUC = 1, then the model can distinguish correctly between positive and negative. If the condition is 0.5<AUC<1 then there is a high chance to distinguish between positive and negative. The F1 score of this proposed SVM model is 74% which refers to the combination of precision and recall scores which are 71% and 77% respectively. The overall accuracy of this proposed SVM model is 71% and this score is a good approach. According to the ROC curve, the higher Y-axis value denotes a higher number of true positives than false negatives as well as the higher X-axis value denotes a higher number of false positives than true negatives. According to Fig. 7, the ROC curve of this proposed SVM model shows a higher true positive rate than the false positive rate. This signifies that the proposed is able to generate good prediction results and this ROC curve indication satisfied this.

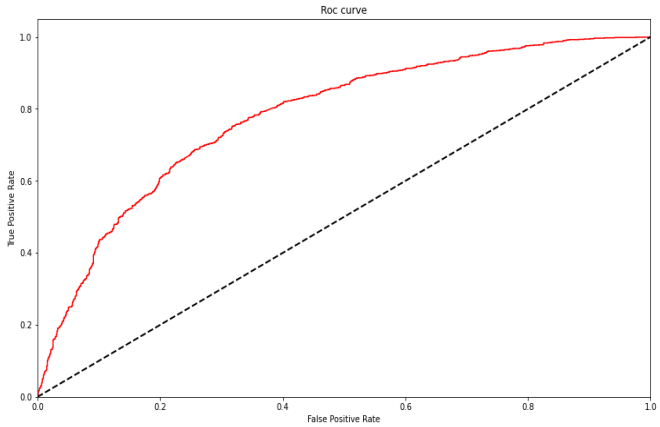

Fig. 7. ROC curve of proposed SVM model.

According to Fig. 8, the training scores line on the graph is between 0.99 and 0.94 (approx.) and the cross-validation scores line is between 0.70 and 0.79 (approx.). The gap between the two score lines is not very high. This proposed model is able to generate good prediction results according to the given Fig. 8.

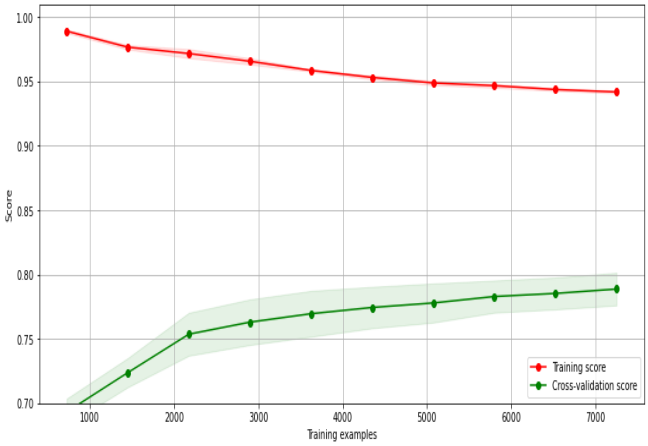

Fig. 8. Training scores and cross-validation scores of SVM.

A few sentences have been sent to the proposed SVM model for the prediction. According to Fig. 9, the proposed model shows the output as 1 or 0, which is attached to each sentence. One (1) refers to a positive sentence regarding ASD detection whereas zero (0) refers to a negative sentence.

```
Result:How does speech therapy help with a nonverbal to speak=[1]
Result:because all they do there is play with toys with him every time=[1]
Result:I'm confused guys help my son is 3years old now=[0]
Result:he does is mumbles only no proper words=[1]
Result:but he goes to speech therapy every month=[1]
Result:Does it help=[0]
Result:My son is 5 and started speech therapy at 3=[1]
Result:My Son can't speak and always spin the wheels of a toy car=[1]
Result:She is 4 years old and reapiting words=[1]
```

Fig. 9. Prediction result of SVM model as output.

B. Result and Discussion of Logistic Regression

According to Fig. 10, a confusion matrix has been represented that refers to how many are true actual 1s, actual 0s, predicted 0s, and predicted 1s.

According to the test data, the proposed logistic regression model selects 75 sentences as actual 0s and predicted as 0s. Fifteen (15) sentences are actual 0s but predicted as 1s whereas 31 sentences are actual 1s and predicted as 0s. Forty (40) sentences are actual 1s and predicted as 1s. The following metrics for model evaluation have been given in Table VII. The AUC value is 0.69 (69%) which covers the ROC curve. The F1 score is 0.63 (63%) which combines the precision and recall values. The precision value is 0.72(72%) and the recall value is 0.56 (56%). The overall accuracy of the proposed Logistic regression model is 0.71 (71%). According to Fig. 11, the ROC curve, the higher Y-axis value denotes a higher number of true positives than false negatives as well as the higher X-axis value denotes a higher number of false positives than true negatives. The training accuracy is 0.97 (97%) whereas the testing accuracy is 0.70 (70%) on the proposed dataset according to Fig. 11.

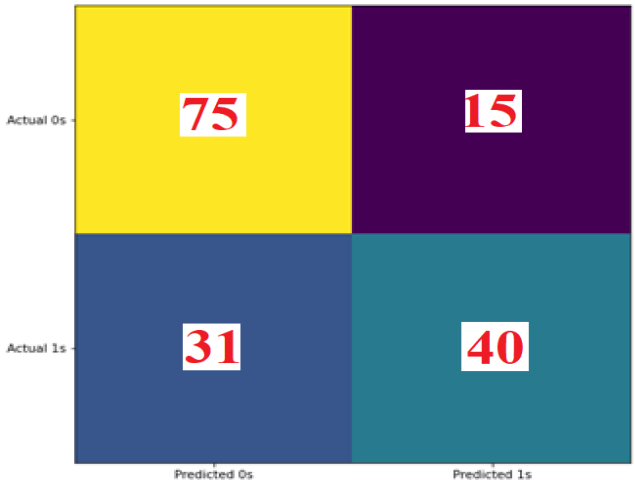

Fig. 10. Confusion matrix of logistic regression model.

TABLE VII. LOGISTIC REGRESSION MODEL METRICS

Sl. No.	Metrics	Value
1.	AUC	0.69
2.	F1	0.63
3.	Accuracy	0.71
4.	Precision	0.72
5.	Recall	0.56

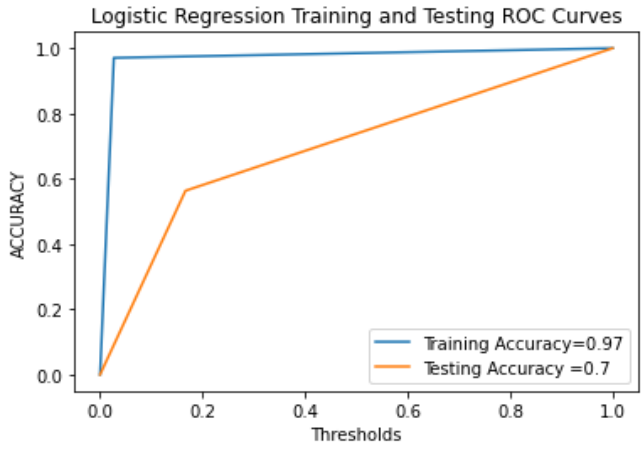

Fig. 11. ROC curves of logistic regression.

C. Result and Discussion of KNN model

According to Fig. 12, two ROC curve has been represented that shows the accuracy of the proposed KNN model.

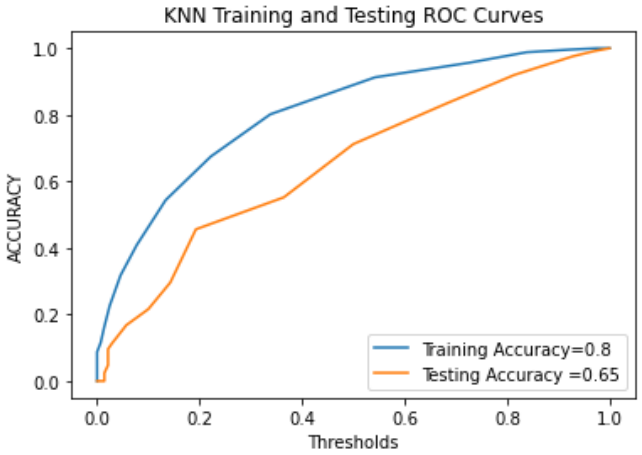

Fig. 12. ROC curves of KNN training and testing.

The AUC value of the proposed KNN model is 0.67 (67%) which refers to the high chance to distinguish positive and negative sentences. The AUC refers to the area of the ROC curves. The given two ROC curves stated that they are moving in almost the same manner from 0 to 1. The training accuracy is 0.80 (80%) as well as testing accuracy is 0.65 (65%) on the proposed dataset. It is another good metric that shows the ability of the prediction of the proposed KNN model. Accuracy is an important metric for machine learning model determination according to the task. The popular metrics have been given in Table VIII are useful to evaluate the machine learning model. The AUC value is already given as 0.67 (67%). The F1 score is 0.65 (65%) which combines the precision and recall values. The precision value is 0.65 (65%) and the recall value is 0.66 (66%). The overall accuracy of the proposed KNN model is 0.62 (62%).

TABLE VIII. KNN MODEL METRICS

Sl. No.	Metrics	Value
1.	AUC	0.67
2.	F1	0.65
3.	Accuracy	0.62
4.	Precision	0.65
5.	Recall	0.66

D. Result and Discussion of Random Forest model

The last proposed model is Random Forest which is a good classifier. The proposed dataset has been applied to this model to predict the sentiment of the sentences from the parents' dialogues. This model trains with the features after extracting these from the sentences.

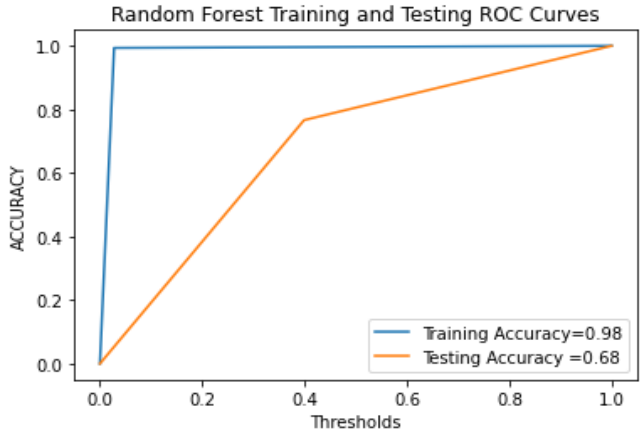

Fig. 13. ROC curves of random forest.

Fig. 13 shows the two lines that are showing the accuracy graph of the proposed Random Forest model. The blue color line shows the accuracy of the training data as well as yellow line shows the accuracy of the testing data. According to Fig. 13, the accuracy score of this model on the testing data is 0.98 (98%), and 0.68 (68%) accuracy score on the testing data. According to Table IX, the F1 score is *0.73 (73*%) which is combined two metrics values that are Precision and Recall. Precision refers to the measurement of the positive prediction of a model whereas recall refers to the positive cases that are correctly predicted by the model. *The* Precision value is 0.70 (70%) and the Recall value is 0.76 (76%). These two values have been defined in Table IX. The overall accuracy value of the proposed model is 0.69 (69%).

TABLE IX. RANDOM FOREST METRICS

Sl. No.	Metrics	Value
1.	AUC	0.68
2.	F1	0.73
3.	Accuracy	0.69
4.	Precision	0.70
5.	Recall	0.76

Fig. 14 shows the top 20 important features from all sentences of the prepared dataset that are also used in the model training. The frequency of each feature can be seen in Fig. 14. "poo", "autism", and "toilet", are three noted words with the highest frequencies. Other words are also given in Fig. 14.

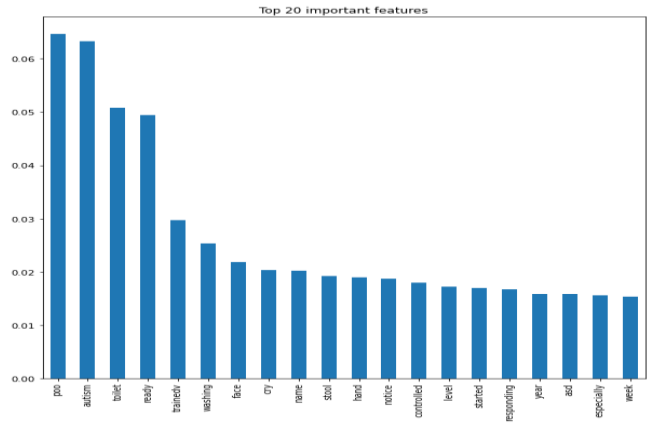

Fig. 14. Important features of proposed dataset.

E. Result and Discussion of Spacy Cosine Similarity Model

The proposed model returns the ASD problem according to the positive sentences that contain ASD symptoms.

```
In [51]: runfile('F:/DeepL/Spacy_Cosine.py', wdir='F:/DeepL')
she cant understand where to do pee = 7

In [52]: runfile('F:/DeepL/Spacy_Cosine.py', wdir='F:/DeepL')
She is very aggressive and throwing objects to other = 8

In [53]: runfile('F:/DeepL/Spacy_Cosine.py', wdir='F:/DeepL')
Eyes are scrolling and hand flapping = 6
```

Fig. 15. Output of spacy cosine similarity model.

The output can be seen here in Fig. 15 where the sentence “she can’t understand where to pee” is labeled with 7. The sentences like “She is very aggressive and throwing objects to others” and “Eyes are scrolling and hand flapping” are labeled with 8 and 6. According to Table IV, 7 denotes Cognitive Behaviour problems whereas 8 refers to Hyper Active problems and 6 refers to the Eye contact problem. After the detection of ASD problems, therapies can be started according to the detected problem and that will be very helpful to reduce the ASD symptoms.

V. Limitation of the Proposed System

The proposed system has been equipped with traditional machine-learning models. The Probabilistic model like Naïve Bayes or ensemble model like XGBoost models can be applied to this dataset for better accuracy. More data can be collected for the training score and testing score enhancement. More accurate ASD-related parent dialogs-related to ASD are needed to train the model. If the dataset is large then this traditional machine learning model will not work better and that will downgrade the proposed system. If one part of this system is not responding then the cosine similarity part will not work perfectly.

VI. Conclusion

The proposed system will accept natural language text from the parents’ dialogues. The proposed system will generate positive or negative sentences using sentiment analysis. A sentence that contains ASD symptoms is 1 and a sentence that does not contain any ASD symptoms is 0. The sentiment analysis has been done using SVM, Logistic Regression, KNN, and Random Forest models. These models are trained with the proposed dataset. After prediction, the proposed system will select all positive sentences as input for the cosine similarity model. An ASD symptoms dataset has been proposed where each sentence is labeled with a value that indicates particular ASD symptoms. The proposed system will calculate the cosine similarity value between the input sentence and each ASD sentence of the ASD symptoms dataset. The proposed system will select a label value of an ASD symptoms sentence that has the highest cosine similarity value with the input sentence and this label value will indicate the ASD problem. This system is based on text and does not need to use MRI or Image data for the prediction of ASD at the early age of a child. This system may be used in many health centers in rural areas because people in rural areas are not aware of ASD as well as many of them are financially weak to spend money for MRI or other ASD diagnosis processes.

VII. Future Work

The proposed dataset can be utilized to train the Naïve Bayes and XGBoost models for better output and accuracy. The cosine similarity model of this system depends on the prediction result of the traditional machine learning models. These models are good for small datasets but these models will not work with the best performance when the dataset is large. XGBoost is an ensemble model which is a very powerful model for prediction as well as the Naïve Bayes model is a probabilistic model that works on Bayes theorem. These two models implementation using a proposed dataset for ASD detection is the future development of this proposed system.

Acknowledgment

The authors extend their appreciation to the Manipur International University, Imphal, India for supporting this Post-Doctoral (D.Sc.) research work on Autism.

References

[1] Raj, Suman, Masood, Sarfaraz, “Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques”, Procedia Computer Science, vol. 167, pp. 994-1004, 2020.

[2] A.S. Mohanty , K.C. Patra , P. Parida, "Toddler ASD classification using machine learn- ing techniques", Int. J. Online Biomed. Eng. vol. 17, 2021.

[3] Ashima Sindhu Mohanty, Priyadarsan Parida, Krishna Chandra Patra, "ASD classification for children using deep neural network", Global Transitions Proceedings, pp.461–466, 2021.

[4] K. K. Hyde, M. N. Novack, N. LaHaye, C. Parlett-Pelleriti, R. Anden, D.R. Dixon, and E. Linstead, “Applications of supervised machine learning in autism spectrum disorder research: a review”, Review Journal of Autism and Developmental Disorders, vol. 6(2), pp.128-146, 2019.

[5] L. Xu, X. Geng, X. He, J. Li and J. Yu, “Prediction in Autism by Deep Learning Short-Time Spontaneous Hemodynamic Fluctuations”. Frontiers in Neuroscience, vol. 13, 2019.

[6] A.L. Georgescu, J.C. Koehler, J. Weiske, K. Vogeley, N. Koutsouleris, C. Falter-Wagner, “Machine Learning to Study Social Interaction Difficulties in ASD.” Computational Approaches for Human-Human and Human-Robot SocialInteractions, 2019.

[7] Shomona Gracia Jacob, Majdi Mohammed Bait Ali Sulaiman, Bensujin Bennet, "Algorithmic Approaches to Classify Autism Spectrum Disorders: A Research Perspective", Procedia Computer Science, vol. 201, pp. 470–477, 2022.

[8] Fadi Thabtah, David Peebles, "A new machine learning model based on induction of rules for autism detection",

[9] D. P. Wall, R. Dally, R. Luyster R, et al., "Use of artificial intelligence to shorten the behavioral diagnosis of autism", PLoS ONE, 2012.

[10] M. Duda, R. Ma, N. Haber, et al., "Use of machine learning for behavioral distinction of autism and ADHD", Transl Psychiat, vol. 9(6), 2016.

[11] A.Pratap, C.S. Kanimozhiselvi, R. Vijayakumar, et al., "Predictive assessment of autism using unsupervised machine learning models, Int J Adv Intell Paradig, vol.6(2), pp. 113–121, 2014.

[12] M. Al-Diabat, "Fuzzy data mining for autism classification of children", Int J Adv Comput Sci Appl, vol. 9(7), pp. 11–17, 2018.

[13] ANJA THIEME, DANIELLE BELGRAVE, GAVIN DOHERTY, "Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems", Trans. Comput.-Hum. Interact, vol. 27(5), Article 34, 2020.

[14] Jacqueline Peng, Mengge Zhao, James Havrilla, Cong Liu, Chunhua Weng, Whitney Guthrie, Robert Schultz, Kai Wang, Yunyun Zhou, "Natural language processing (NLP) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder", BMC Med Inform Decis Mak, pp. 1-9, 2020.

[15] Izabela Chojnicka, Aleksander Wawer, "Social language in autism spectrum disorder: A computational analysis of sentiment and linguistic abstraction", PLOS ONE, pp. 1-16, 2020.

[16] Mahmoud Elbattah, Jean-Luc Guérin, Romuald Carette, Federica Cilia, Gilles Dequen, "NLP-Based Approach to Detect Autism Spectrum Disorder in Saccadic Eye Movement", IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1581-1587, 2020.

[17] T. Lakshmi Praveena, N. V. Muthu Lakshmi, "Sentiment Analysis on Autism Spectrum Disorder using Twitter Data", International Journal of Recent Technology and Engineering (IJRTE), vol. 7(4), pp. 204-208, 2018.

[18] Laura Dubreuil-Vall, Giulio Ruffini, Joan A. Camprodon1, "Deep Learning Convolutional Neural Networks Discriminate Adult ADHD From Healthy Individuals on the Basis of Event-Related Spectral EEG", Front. Neurosci, vol. 14, pp. 1-12, 2020.

[19] Dingfu Zhou, Zhihang Liao, Rong Chen, "Deep Learning Enabled Diagnosis of Children’s ADHD Based on the Big Data of Video Screen Long-Range EEG", Journal of Healthcare Engineering, pp. 1-9, 2022.

[20] Shubham Dhuri, Nitin Ahire, Deepak Kamat, Sunil Nayak, Bhavesh Maurya, "ADHD EEG signal analysis using Machine Learning", International Research Journal of Engineering and Technology (IRJET), vol. 8(5), pp. 2572-2575, 2021.

[21] Iqra Ameer, Muhammad Arif,Grigori Sidorov, Helena Gomez-Adorno, Alexander Gelbukh, "Mental Illness Classication on Social Media Texts using Deep Learning and Transfer Learning", arXiv:2207.01012, pp. 1-12, 2022.

[22] Tanzila Saba, Amjad Rehman Khan, Ibrahim Abunadi, Saeed AliBahaj, Haider Ali, Maryam Alruwaythi, "Arabic Speech Analysis for Classification and Prediction of Mental Illness due to Depression Using Deep Learning", Computational Intelligence and Neuroscience, vol. 2022, pp. 1-9, 2022.

[23] Amanda Sun, Zhe Wu, "Early detection of mental disorder via social media posts using deep learning models", Proceedings of Asia Pacific Computer Systems Conference, pp. 149-158, 2021.

[24] Anshul Saini, "Support Vector Machine(SVM): A Complete guide for beginners", https://www.analyticsvidhya.com/blog/2021/10/ support-vector-machinessvm-a-complete-guide-for-beginners/, 2023.

[25] Himanshi Singh, "How to select Best Split in Decision trees using Gini Impurity", https://www.analyticsvidhya.com/blog/2021/03/how-to-select-best-split-in-decision-trees-gini-impurity/, 2021.

AUTHORS’ PROFILE

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.

Dr. Baisakhi Chakraborty received the PhD. degree in 2011 from National Institute of Technology, Durgapur, India in Computer Science and Engineering. Her research interest includes knowledge systems, knowledge engineering and management, database systems, data mining, natural language processing, and software engineering. She has several research scholars under her guidance. She has more than 60 international publications. She has a decade of industrial and 22 years of academic experience.