AMRITA SCHOOL OF ARTIFICIAL ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

April - 2025

B. TECH ARTIFICIAL INTELLIGENCE IN DATA SCIENCE AND MEDICAL ENGINEERING

AUTISM PREDICTION USING MACHINE LEARNING

24AIM112 Molecular biology & basic cellular physiology

24AIM115 Ethics, innovative research, businesses & IPR

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

BONAFIDE CERTIFICATE

This is to certify that the report entitled “**AUTISM PREDICTION USING MACHINE LEARNING”** submitted by:

Dheeraj Chowdary	**CB.AI.U4AIM24109**
Sai Charan	**CB.AI.U4AIM24124**
Prem Siva	**CB.AI.U4AIM24125**
Chiru deep	**CB.AI.U4AIM24137**

for the final project of 2nd semester in **B. TECH ARTIFICIAL INTELLIGENCE IN DATA SCIENCE AND MEDICAL ENGINEERING** is a Bonafide record of the work carried out at Amrita School of Artificial intelligence, Coimbatore.

Submitted for the final evaluation on 23-04-2025

FACULTY **FACULTY**

INTRODUCTION

Understanding Autism Spectrum Disorder (ASD)

autism spectrum disorder (ASD) is a chronic neurodevelopmental disorder that impacts the way individuals perceive, communicate, and relate to the world around them. Individuals with autism can experience difficulty with social interaction, verbal and non-verbal communication, and repetitive behaviour or limited interests. The condition is highly variable from individual to individual in terms of severity and the character of symptoms, and that is why it is referred to as a "spectrum." Some people will have high support needs, while others will be independent and have special talents or abilities in areas like art, music, mathematics, or memory.

ASD usually becomes apparent in early childhood, usually before the age of three. It is diagnosed based on behavioural observations and developmental milestones. Some people, however, because of the varied presentation of autism, may remain undiagnosed until adolescence or adulthood. The etiology of autism is multifactorial and not yet clearly understood. It is thought to be caused by a combination of genetic, neurological, and environmental factors. Genetic factors are strong, with research suggesting that more than one gene is involved in the risk for autism. Differences in brain development, prenatal issues, and some environmental exposures could also contribute to the risk of developing ASD.

Importance of Early Detection and Current Challenges

Early detection and intervention are essential in enhancing developmental outcomes for individuals with autism. Interventions administered early in the course of brain development can greatly improve communication, learning capacity, and social interaction skills. Although early detection is important, conventional diagnostic procedure s are time-consuming and depend on subjective judgments by clinicians, teachers, or caregivers. These procedures can result in delayed diagnosis or missed cases, particularly in areas with poor access to trained personnel.

Against such challenges, the demand for more objective, effective, and efficient methods to contribute to the identification and analysis of autism is growing. Breakthroughs in technology, specifically artificial intelligence and machine learning, present promising methods that can deal with large and complicated data, find patterns, and assist healthcare decision-making.

Role of Machine Learning in Autism Prediction

Machine learning (ML) is a strong area of artificial intelligence that allows computers to learn from information and make decisions or predictions with little human interaction. In the field of medicine, ML is being used more and more to create predictive models, identify illnesses, and customize treatment plans. With autism, ML can assist in finding patterns in behavioural data, genetic data, or other biological markers that are not so apparent through direct observation.

Through training algorithms on labelled data, ML models can learn to differentiate between autistic and non-autistic subjects or detect genetic markers linked with the condition. The models can potentially aid clinicians by offering supportive diagnostic tools, minimizing human bias, and facilitating early screening in a larger population. Notably, ML models can improve incrementally as additional data becomes available, thus making them extremely adaptable and scalable for practical application.

Project Scope and Model Objectives

The project seeks to use machine learning methods to forecast and examine autism based on both behavioral and genetic data sets. Three individual models were created to tackle the issue from various sides:

Behaviour-Based Autism Prediction Model:

This model employs a dataset of behavioural traits gathered from individuals. We compared three machine learning models [insert models used, Logistic Regression, decision tree, SVM] to identify the best-performing model in predicting autism. The aim was to compare their performance based on accuracy, precision, recall, and F1-score, and determine the most appropriate algorithm for early behavioural screening.

Gene Score Classification Model:

The second model is gene-related data-focused. Its objective is to predict gene classes according to their gene scores so that the genes that play a greater role in the risk of autism can be determined. Gene score is a numerical value defining a gene's potential effect in terms of mutation frequency or any other biological significance. This model aids in gene prioritization for future study and potential clinical assessment.

Syndromic vs Non-Syndromic Gene Prediction Model

The third model applies gene-related information but aims at a different objective. It is programmed to forecast whether a gene is syndromic or non-syndromic for the case of autism. Syndromic autism comprises other medical or neurological conditions that are associated with particular genetic syndromes, while non-syndromic autism does not. Genes classified this way help us comprehend the genetic heterogeneity of autism and potentially facilitate personalized treatment strategies.

Significance and Expected Outcomes

By combining behavioural and genetic information using machine learning, this project seeks to investigate both superficial and underlying causes of autism. Co mbining the two methods offers a more complete picture of the disorder. The behavioural model is supportive of early detection, whereas the gene-based models help to find biological markers and enhance the accuracy of genetic studies.

The results of this project should illustrate the potential and utility of machine learning in the study of autism. These models, if they are further improved and validated, would be usable as tools to aid healthcare practitioners in diagnosis, aid genetic counsellors in risk evaluation, and play a role in the expanding discipline of precision medicine in neurodevelopmental disorders.

LITERATURE REVIEW

Si.	Title	Year	Results
1.	Fusion of Features: A Technique to Improve Autism Spectrum Disorder Detection Using Brain MRI Images [1]	2023	Accuracy is high for the classification tree, around 85.96 percent.
2.	Autism Detection for Toddlers using Facial Features with Deep Learning [2]	2024	Used VGG16 and ViT Accuracy is around 89.06%
3.	Innovative Autism Spectrum Disorder Prediction Using Machine Learning [3]	2024	They used the KNN model. The accuracy they achieved is 66.10%
4.	Discovering the Gene-Brain-Behaviour Link in Autism via Generative [4]	2024	89% accuracy in predicting genetic variations, strong correlation between structural brain changes and ASD-related traits.
5.	The Role of Intelligent Technologies in Early Detection of Autism Spectrum Disorder (ASD): A Scoping Review [5]	2022	ML/DL techniques outperform traditional methods, need for robust datasets and multicultural validation.

ETHICAL RIGHTS:

1. Right to Communication and Expression

This person should not be punished or excluded for different communication styles.

2. Right to Accessible Healthcare

They should not be subjected to forced treatments without informed consent.

3. Right to Dignity and Respect

Their differences should be respected as part of natural human beings rather than as defects.

4. Right to Representation

The autistic individuals should have a say in decisions that affect their lives, whether in education, healthcare, etc.

ETHICAL CONSIDERATIONS:

- Patients with mental capacity have the right to accept or decline any medical procedure.
- For Early detection, parents should receive balanced information not only about challenges but also about strengths.
- Instead of eliminating autism, one should focus on helping autistic individuals**.**
- Autistic individuals, particularly minors and those with cognitive or communication challenges, may require additional safeguards to ensure ethical participation.

CASE STUDY 1:

Title: Multiple Classification of Brain MRI Autism Spectrum Disorder by Age and Gender Using Deep Learning

Objectives:

- This study aims to improve the accuracy of autism spectrum disorder using deep learning techniques.

Related Works

- Focused on quadrupled classification on gender, quadrupled classification on age, and octal classification.
- Investigated which group can be predicted with more accuracy for ASD using EMcRBFN, and the accuracy for women is 81%, and for females is 60%.
- They used SVM and got 69% accuracy for females and 66% for males.
- They used MC-CNN for the 2-year-old group and got accuracy of 76.24%.
- They used a DBN (deep belief network) model the accuracy is 65.69%.
- Using DL MRI images, they got an accuracy of 90.39%.

Materials

- Collected the data from the ABIDE database.
- The data collected from ABDIE consists of data from 29 sites.
- After scanning images collected from 29 different sites finally got 1831 images (938 ASD, 893 TD).

Data Pre-Processing:

- In the first stage, eliminate all the unclear images.
- In the second stage, CED (canny edge detection) minimized all the images.
- In the third stage, after pre-processing, the images were rotated at certain angles.

CNN MODELS:

Optimal hyperparameter selection

- Three DL models are used based on the GSO algorithm.
- Used activation function as SoftMax function.
- The first model predicts the ASD, taking only age as the main parameter.
- The second model predicts the ASD, taking only gender as the main parameter.
- The third model predicts the ASD taking both age and gender as parameters.
- The developed model is compared with four pretrained networks using TL.

RESULTS

- For the first model using age as the major parameter, we got an accuracy of 80.94%, which is higher than all pre-trained models designed.
- The second model using gender as the major parameter got an accuracy of 85.42%.
- The third model using both age and gender as the parameters got an accuracy of 67.94%.

CONCLUSIONS

- This model has better accuracy than the pre-trained models such as AlexNet, GoogleNet, ResNet-18, and SqueezeNet.
- The dataset we obtained from the 29 different sources is enlarged by 5 times using DA techniques.
- Planning to do more future applications using Enhanced Probabilistic Neural Network (EPNN) and Neural Dynamic Classification (NDC) algorithm.

CASE STUDY 2:

Title:

Video-based continuous affect recognition of children with autism spectrum disorder using deep learning**.**

OVERVIEW:

Data obtained before October 2023 have been used for training. This research employs an integration of controlled and uncontrolled datasets toward model training on the detection of ASD and affect recognition. The main datasets include SSBD, Affect Net, IEMOCAP, and CK+. SSBD covers 73 YouTube videos of self-stimulatory behaviors that have been manually labeled with arousal and valence. For every video, preprocessing involved face detection and pose detection, during which frames were sampled to capture the temporal context. Image feature extraction was performed by EfficientNet-B0 and ResNet-18, whereas fusion took place at the transformer stage. For the final classification, MLPs were applied with evaluation via five-fold cross-validation. The results emphasize the need for affect predictions to be done on ASD-specific data, while InceptionV3 achieved an F1 score of 97.7% and a Recall of 97.9%, marking it as the most successful one within this research.

Case Study: - 3

Methods for treating autism spectrum disorder and associated symptoms

(AU 2022203294 B2)

Objective:

This invention presents a way to restore gut microbiota to treat autism spectrum disorder (ASD) through microbiota transfer therapy (MTT). Beneficial gut bacteria from healthy donors are given to people with ASD to improve GI symptoms and behaviors associated with the disorder.

Inputs:

The study concentrated on gastrointestinal (GI) symptoms, including bloating, diarrhea, and constipation, in autistic people.

Preparation of Fecal Microbiota

Stool samples from healthy volunteers were treated to eliminate dangerous microorganisms and unnecessary trash while obtaining beneficial gut bacteria.

Methodology:

The patent presents four key steps for treatment:

1. **Direct Oral FMT Treatment**
Patients received fecal microbiota from healthy donors in capsule or rectal form.

2. **Cultured Bacteria Approach**
Instead of using the whole microbiota, selected beneficial bacterial strains (e.g., **Lactobacillus**) were used.

3. **Three-Step Treatment Process:**
   - **Step 1: Antibiotic Pre-Treatment** – Patients first took antibiotics to reduce harmful bacteria.
   - **Step 2: Bowel Cleanse** – A **bowel cleansing procedure** was performed to prepare the gut for microbiota transfer.
   - **Step 3: Fecal Microbiota Transplantation (FMT)** – Beneficial bacteria were introduced into the gut **orally** or rectally over several weeks.
4. **Pharmaceutical Drug Development**
   - The method can be used to develop a commercial medication for ASD treatment.

Data Collection & Analysis

Clinical Assessments – Data was collected using:

- ->GSRS (Gastrointestinal Symptom Rating Scale) – Assessed GI symptoms.

- ->CARS, (CARS2), ABC, SRS – Evaluated ASD symptoms like behavioral changes and social communication improvements.

Outputs

- ->Reduced GI Symptoms – Significant decrease in diarrhea, pain, bleeding

- ->Reduced ASD Symptoms – Improvements in social communication and behavior.

Conclusion:

This patent presents a novel, microbiome-based treatment for ASD, offering a gut-focused alternative to existing therapies. By restoring gut bacteria, ASD symptoms and GI symptoms can be significantly reduced, and the benefits are long-lasting. This approach opens the door for new medical treatments in neurodevelopmental disorders

PATENT:

Overview

The patent US20210256249A1, "Detecting Visual Attention of Children with Autism Spectrum Disorder", suggests an automated approach to measuring visual attention through facial landmark analysis. It is aimed particularly at helping children with autism spectrum disorder (ASD), whose abnormal attention behaviors affect learning.

Methodology involved

The system takes face photos with a webcam while performing an attention task. The system identifies 34 facial landmarks along the jaw, eyes, eyebrows, nose, and lips. It then computes Euclidean distances between pairs of landmarks and chooses the 20 most salient features that distinguish attentive and inattentive states. These 20 features are used as input to a machine learning predictor.

Out of six classifiers tried (SVM, Decision Tree, Logistic Regression, Random Forest, Gradient Boosting, KNN), Support Vector Machine (SVM) was the best. It was tuned with cross-validation and hyperparameter tuning.

Key Results

- Participant-dependent model had high performance (ROC-AUC = 0.959).
- Participant-independent model was above chance (ROC-AUC = 0.561).

The method worked best when processing data from Mild ASD vs. Moderate ASD, typically developing (TD) children, and low distraction tasks.

Conclusions

This technique offers a non-invasive, real-time measure for detecting attention in children with ASD based on facial characteristics. It generalizes fairly well and is optimally suited to customized, participant-based applications. The technique is beneficial in educational and therapeutic practices, where knowledge about attention is imperative.

METHODOLOGY:

1. Autism prediction model (comparison between different ML algorithms)
2. Gene data analysis
    a. Gene score classification model
    b. Syndromic prediction

OVERVIEW:

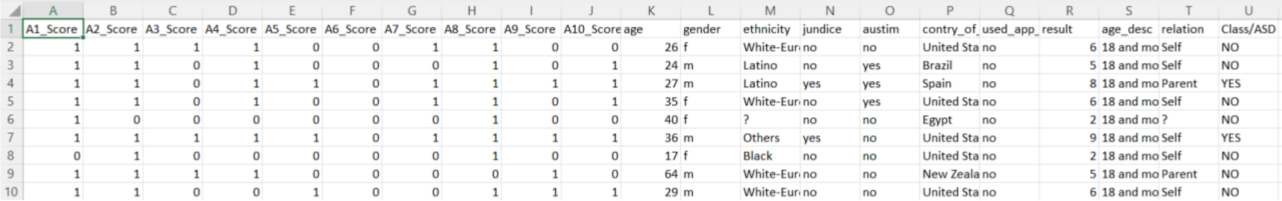
Autism prediction model: The model prepared can predict whether the person has autism or not. We compared the models prepared with different algorithms to check which algorithm gives a good result. we checked on logistic regression, SVM, and decision tree. SVM outperformed the remaining algorithms.

Gene data analysis: The model helps to know the role of a particular gene in autism and whether the gene causes other conditions than autism.

1. AUTISM PREDICTION MODEL

DATASET OVERVIEW:

The dataset says whether a person has autism, using some behavioural questions and features.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	age	gender	ethnicity	jundice	austim	contry_of_	used_app_	result	age_desc	relation	Class/ASD
2	1	1	1	1	0	0	1	1	0	0	26	f	White-Eur	no	no	United Sta	no	6	18 and mo	Self	NO
3	1	1	0	1	0	0	0	1	0	1	24	m	Latino	no	yes	Brazil	no	5	18 and mo	Self	NO
4	1	1	0	1	1	0	1	1	1	1	27	m	Latino	yes	yes	Spain	no	8	18 and mo	Parent	YES
5	1	1	0	1	0	0	1	1	0	1	35	f	White-Eur	no	yes	United Sta	no	6	18 and mo	Self	NO
6	1	0	0	0	0	0	0	1	0	0	40	f	?	no	no	Egypt	no	2	18 and mo	?	NO
7	1	1	1	1	1	0	1	1	1	1	36	m	Others	yes	no	United Sta	no	9	18 and mo	Self	YES
8	0	1	0	0	0	0	0	1	0	0	17	f	Black	no	no	United Sta	no	2	18 and mo	Self	NO
9	1	1	1	1	0	0	0	0	1	0	64	m	White-Eur	no	no	New Zeala	no	5	18 and mo	Parent	NO
10	1	1	0	0	1	0	0	1	1	1	29	m	White-Eur	no	no	United Sta	no	6	18 and mo	Self	NO

Fig: sample dataset [6]

- Total 704 samples.
- 21 Features:
  >>**A1_Score to A10_Score** – Responses to 10 screening questions related to behaviour and preferences. Each question is binary (0 or 1).
  >>Age, gender
  >>ethnicity – ethnic background
  >> jaundice -whether the person is suffering from frequent jaundice or no
  >> Autism – Family history of autism (yes, no).
  >>contry_of_res – Country of residence.
  >>result – sum of the score of the screening test.
  >>age_desc- age category
  >>relation

DATA PREPROCESSING:

To make the comparison subtle between different algorithms same data preprocessing techniques are used for all three 3 models.

```
import pandas as pd
from google.colab import files
import io

# Step 1: Upload the dataset
uploaded = files.upload()

# Get the uploaded file name
file_name = list(uploaded.keys())[0]

# Step 2: Load the dataset
df = pd.read_excel(io.BytesIO(uploaded[file_name]), sheet_name='autism')

# Step 3: Display missing values before processing
print("Missing values before processing:\n", df.isnull().sum())

# Step 4: Count "?" values before removal
print("\nCount of '?' in ethnicity:", (df['ethnicity'] == '?').sum())
print("Count of '?' in relation:", (df['relation'] == '?').sum())

# Step 5: Remove rows where 'ethnicity' or 'relation' contains "?"
df = df[(df['ethnicity'] != '?') & (df['relation'] != '?')]

# Step 6: Remove rows where 'age' is NaN
df = df.dropna(subset=['age'])

# Step 7: Remove rows where 'age' is greater than 80
df = df[df['age'] <= 80]

# Step 8: Display missing values after processing
print("\nMissing values after processing:\n", df.isnull().sum())

# Step 9: Save the cleaned dataset
cleaned_file_name = "cleaned_autism.xlsx"
df.to_excel(cleaned_file_name, index=False)

# Step 10: Download the cleaned dataset
files.download(cleaned_file_name)
print("\nDownload started: 'cleaned_autism.xlsx'")
```

Code used for preprocessing

- Initially, the dataset is uploaded.
- And the model reads the file uploaded and gives the count of missing values, unusual values, and outliers in each column.
- Removes the rows where ethnicity or relation contains ‘?’.
- The cleaned file is downloaded.

MODEL DEVELOPMENT:

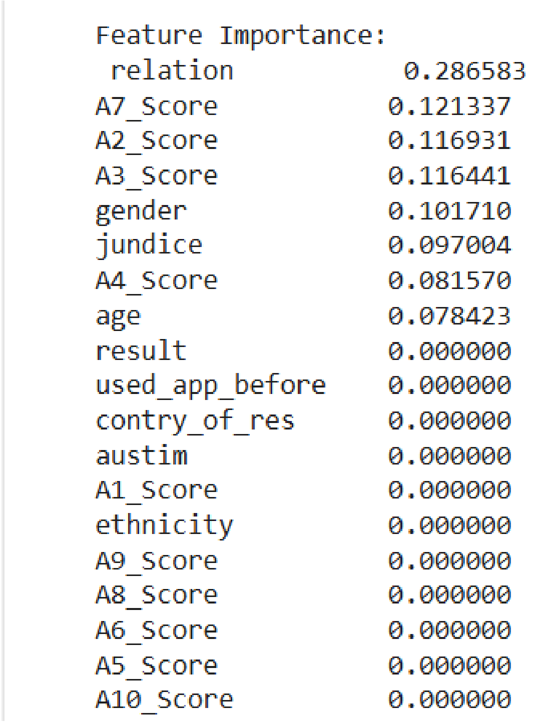
Correlation analysis:

This is a way to measure how strongly a feature (or column) in a dataset is related to the target value.

The result is a correlation coefficient, which ranges from -1 to +1:

- +1 → strong positive relationship (as one increases, the other also increases)
- 0 → no relationship
- -1 → strong negative relationship (as one increases, the other decreases)

Results of correlation analysis:



```
Feature Importance:
 relation          0.286583
A7_Score          0.121337
A2_Score          0.116931
A3_Score          0.116441
gender            0.101710
jundice           0.097004
A4_Score          0.081570
age               0.078423
result            0.000000
used_app_before   0.000000
contry_of_res     0.000000
austim            0.000000
A1_Score          0.000000
ethnicity         0.000000
A9_Score          0.000000
A8_Score          0.000000
A6_Score          0.000000
A5_Score          0.000000
A10_Score         0.000000
```

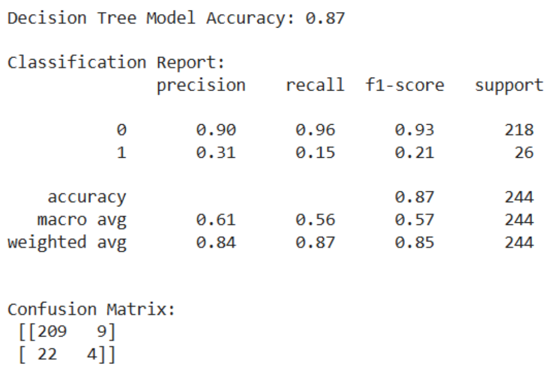
This shows that the relationship between the person and the person in the family who has autism plays a key role in determining whether a person has autism or not.

- The data preprocessing code will be the same for all three algorithms, and the core code has built-in algorithm functions.

Results:

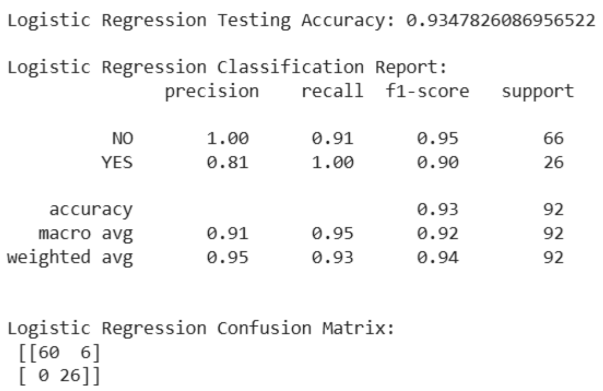
Final accuracies and confusion matrices for each algorithm.

Decision tree result:



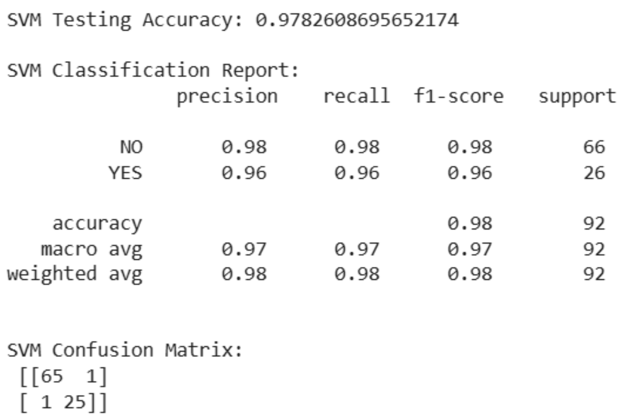
```
Decision Tree Model Accuracy: 0.87

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.96      0.93       218
           1       0.31      0.15      0.21        26

    accuracy                           0.87       244
   macro avg       0.61      0.56      0.57       244
weighted avg       0.84      0.87      0.85       244


Confusion Matrix:
 [[209   9]
 [ 22   4]]
```

Logistic regression result:



```
Logistic Regression Testing Accuracy: 0.9347826086956522

Logistic Regression Classification Report:
              precision    recall  f1-score   support

          NO       1.00      0.91      0.95        66
         YES       0.81      1.00      0.90        26

    accuracy                           0.93        92
   macro avg       0.91      0.95      0.92        92
weighted avg       0.95      0.93      0.94        92

Logistic Regression Confusion Matrix:
 [[60  6]
 [ 0 26]]
```

SVM result:



```
SVM Testing Accuracy: 0.9782608695652174

SVM Classification Report:
              precision    recall  f1-score   support

          NO       0.98      0.98      0.98        66
         YES       0.96      0.96      0.96        26

    accuracy                           0.98        92
   macro avg       0.97      0.97      0.97        92
weighted avg       0.98      0.98      0.98        92


SVM Confusion Matrix:
 [[65  1]
 [ 1 25]]
```

Comparison table:

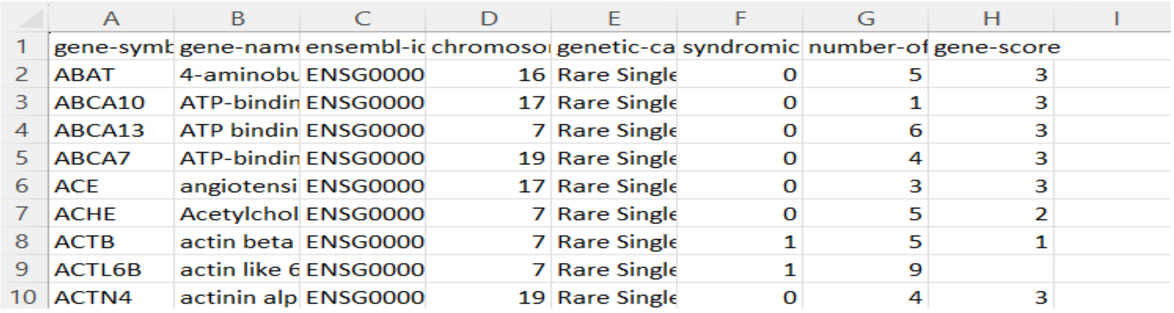
	Algorithm	Accuracy
1.	Decision tree	87%
2.	Logistic regression	93.47%
3.	Support vector machine	97.82%

From the comparison table, we can say that among the three ML algorithms used in the autism prediction model, SVM outperformed the other algorithms. This is due to the nature of the SVM algorithm, which is best suited for well-structured, mostly binary or numerical features. These features allow the algorithm to draw a clear decision boundary between ASD and non-ASD cases. In addition to this, SVM performs well for small and medium-sized datasets (like the one we used).

2. GENE DATA ANALYSIS

DATASET OVERVIEW:

This data shows how much relevance a gene has in causing autism.



	A	B	C	D	E	F	G	H	I
1	gene-symb	gene-nam	ensembl-ic	chromoso	genetic-ca	syndromic	number-of	gene-score	
2	ABAT	4-aminobu	ENSG0000	16	Rare Single	0	5	3	
3	ABCA10	ATP-bindin	ENSG0000	17	Rare Single	0	1	3	
4	ABCA13	ATP bindin	ENSG0000	7	Rare Single	0	6	3	
5	ABCA7	ATP-bindin	ENSG0000	19	Rare Single	0	4	3	
6	ACE	angiotensi	ENSG0000	17	Rare Single	0	3	3	
7	ACHE	Acetylchol	ENSG0000	7	Rare Single	0	5	2	
8	ACTB	actin beta	ENSG0000	7	Rare Single	1	5	1	
9	ACTL6B	actin like 6	ENSG0000	7	Rare Single	1	9		
10	ACTN4	actinin alp	ENSG0000	19	Rare Single	0	4	3	

FIG: sample dataset [7]

- A total of 1023 types of genes.
- 8 features

  >> gene-symbol, gene-name, ensemble-id – these are noted based on how the gene is recognised in the gene database.
  >> chromosome – chromosome location where the gene is present.
  >> syndromic - shows whether the gene is associated with syndromic autism (i.e., autism as part of a broader syndrome). (0-non syndromic)
  >> Number of reports - reports or studies linking the gene to autism.
  >> gene-score – shows the strength of association between the gene and autism.

A) GENE SCORE CLASSIFICATION MODEL:

PURPOSE OF THIS MODEL:

The model is designed to classify genes based on their associated risk scores. By using information like the chromosome number, clinical characteristics, and other gene-related features, the model predicts whether a gene falls into a **low-risk or high-risk category**.

DATA PREPROCESSING:

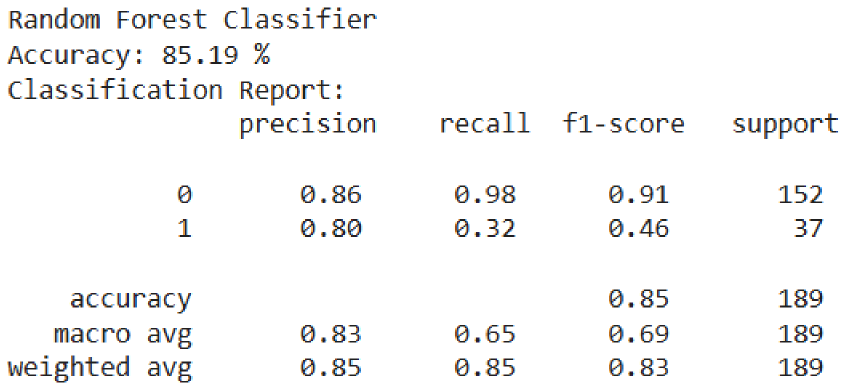
- As the gene score is the target value of this model, if there are any missing gene scores in a row, then the entire row is eliminated.
- The chromosome column is numerical, but at some points it has 'X' as a value. To train the model, all the values in a particular column should be either numerical or categorical. The chromosomes column generally indicates where the gene is located on a chromosome pair. 'X' is not wrong, but the column must be numerical, so we change these 'x' into 23.
- Gene scores generally say how much the gene influences autism. If the gene score is high, it means the gene plays the main role in causing autism. In the dataset, these are from the 1 to 3 range.
  To simplify the prediction task, the scores are converted into binary categories. Gens with a score of 1 or less were labelled as 'low', while those with a gene score greater than 1 were labelled as 'high'. These are further encoded as low=0, high=1. This binarization is done to change the problem into a binary classification model.

MODEL SELECTION AND TRAINING:

A random forest classifier was used for classification because of its robustness and ability to handle numerical and categorical data. The data is divided into 80 per cent for training and 20 per cent for testing. The model is trained using the training data and will be evaluated on test data, and its performance is checked using accuracy.

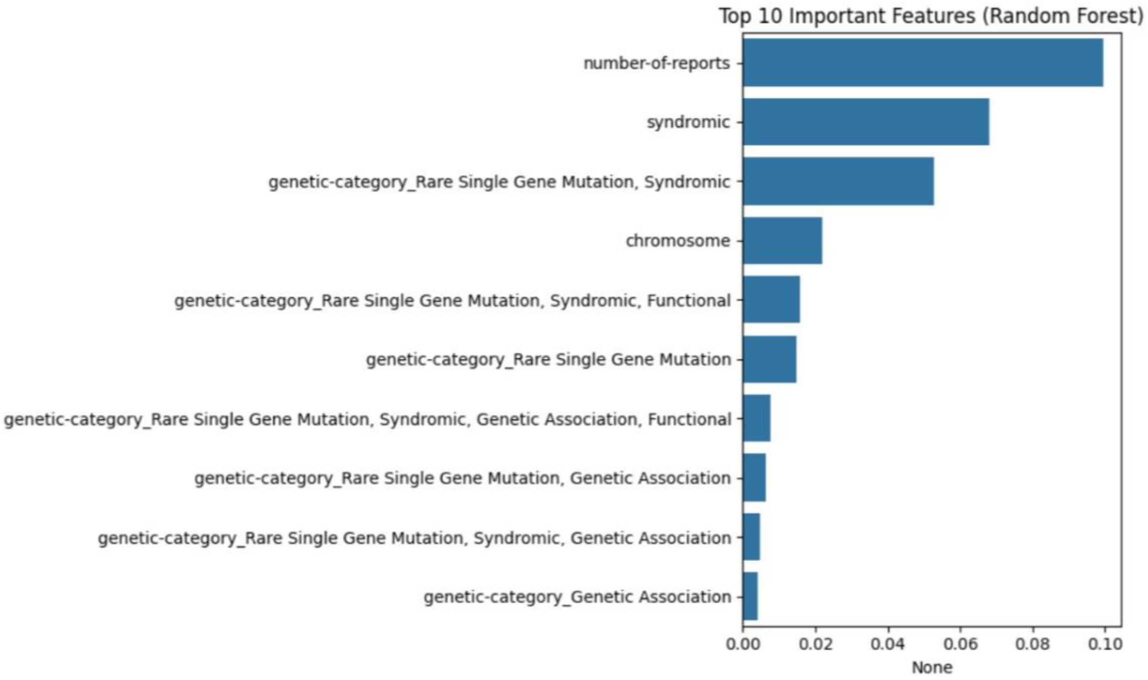
RESULT:

Random forest accuracy:



```
Random Forest Classifier
Accuracy: 85.19 %
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.98      0.91       152
           1       0.80      0.32      0.46        37

    accuracy                           0.85       189
   macro avg       0.83      0.65      0.69       189
weighted avg       0.85      0.85      0.83       189
```

In addition to calculating accuracy, we did a correlation analysis, which says how much an attribute influences the target value.



This is the bar graph generated to show the correlation coefficients. And from the plot number of reports and syndromic attributes are said to have more influence on gene score.

B) SYNDROMIC OR NON-SYNDROMIC PREDICTION MODEL:

PURPOSE OF THE MODEL:

The main objective of this model is to categorise genes linked to autism as syndromic or non-syndromic. This differentiation is important in genetic studies and clinical genetics. A syndromic gene causes autism in combination with other medical or developmental conditions, like intellectual disability, epilepsy, or facial dysmorphisms, typically as part of a clearly defined genetic syndrome. A non-syndromic gene, on the other hand, is linked to autism without additional conditions.

By developing a machine learning model that can perform this classification, clinicians and researchers can learn about the genetic aetiology of autism and whether certain genes are more likely to produce isolated autism or part of a syndrome. This can be used to inform personalised diagnosis, guide clinical genetic testing, and prioritise genes for research.

DATA PREPROCESSING:

The model was trained on a dataset with gene-related features. Before training the model, the data had to be pre-processed so that it could be used for machine learning. Preprocessing involved the following steps:

- **Column Removal:** Some columns in the data, like "gene-symbol", "gene-name", "ensemble-id", and "status", were removed. These are identifiers or text descriptions that do not directly assist in model training. Having such columns can add noise or unnecessary complexity to the model.

- **Missing Data Handling:** Missing value rows (also referred to as null values) were dropped. Machine learning models only work with complete data for optimal predictions. Having incomplete data might decrease the model's performance or lead to training errors.

- **Feature Scaling:** For all the features (i.e., prediction columns) to be in the same range, we implemented Standard Scaling via StandardScaler. Scaling is necessary for sensitive algorithms such as Support Vector Machines (SVM) that are particular about the range of input values. Standard scaling converts the data so that every feature has a mean of 0 and a standard deviation of 1, which allows the model to work more consistently and train faster.

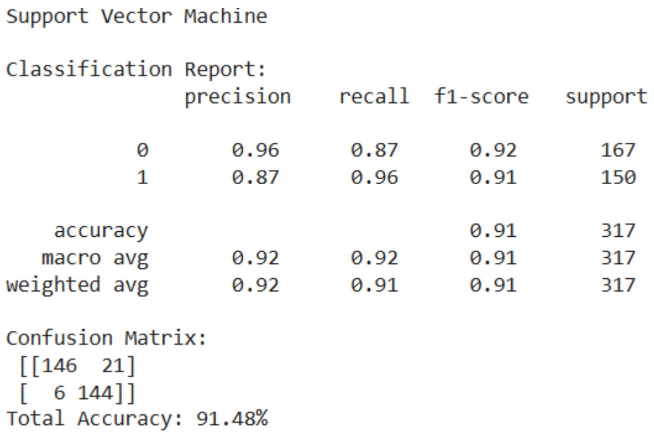
MODEL SELECTION AND TRAINING

We used a commonly used machine learning approach known as a classification algorithm in this model. The algorithm aims at training from examples (data) and then making predictions on whether a gene is syndromic or not.

We employed a technique that is accurate and reliable when dealing with this kind of problem, where the data are complex and difficult to divide. To allow the model to learn best, we also made sure the training data was well prepared and balanced. This includes both classes of genes (syndromic and non-syndromic) equally distributed so that the model would not be biased towards one over the other.

Once we've loaded the data, we divide it into two halves: half was used to train the model, and half was used to test how well the model had been trained. This informs us about how the model would perform on unseen new data.

RESULT:



```
Support Vector Machine

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.87      0.92       167
           1       0.87      0.96      0.91       150

    accuracy                           0.91       317
   macro avg       0.92      0.92      0.91       317
weighted avg       0.92      0.91      0.91       317

Confusion Matrix:
 [[146  21]
 [  6 144]]
Total Accuracy: 91.48%
```

CONCLUSION

This model is highly helpful in supporting autism-related research by helping identify whether a given gene can potentially be involved in syndromic or non-syndromic forms of autism. Using machine learning to help in this process aids in speeding up the process and integrating accuracy into predictions that otherwise would take a long time to be processed manually.

By doing this, based on learning from trends in genetic information, the model can enable researchers, doctors, and geneticists to make more prudent choices while studying genes that pertain to autism. The model also creates a platform for undertaking more sophisticated studies in the future, with such technology aiding early diagnosis in addition to personalised medical treatment in autism.

References:

[1] https://biomedpharmajournal.org/vol16no4/fusion-of-features-a-technique-to-improve-autism-spectrum-disorder-detection-using-brain-mri-images/

[2] https://ieeexplore.ieee.org/document/10575487

[3]https://www.researchgate.net/publication/383947643_Innovative_Autism_Spectrum_Disorder_Prediction_Using_Machine_Learning

[4] https://www.science.org/doi/10.1126/sciadv.adl5307

[5] https://www.turkarchpediatr.org/en/the-role-of-artificial-intelligence-for-early-diagnostic-tools-of-autism-spectrum-disorder-a-systematic-review-131746

[6] autism prediction dataset:
https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults

[7] autism gene-related dataset:

https://gene.sfari.org/database/gene-scoring/