RESEARCH ARTICLE

Integrating Artificial Intelligence and Machine Learning in Autism Detection via Gut Microbiome Analysis

Shobhita Singh[1], **Shubhani Aggarwal**[2,*], **Aishani Singh**[1] and **Anupriya Sharma**[3]

[1] Amity School of Engineering and Technology, Amity University Punjab, Mohali 140306, India
[2] School of Computer Science, University of Petroleum and Energy Studies, Dehradun 248007, India
[3] School of Computing, Graphic Era Hill University, Dehradun 248002, India

Abstract

The Autism Spectrum Disorder (ASD) diagnosis and detection in its initial stages is a more complex issue in the face of the wide-ranging, diverse nature and causes. Subsequent literature inclined towards a possible correlation of gut microbiome with ASD, and its disclosure presents a more promising attribute for imminent discovery conduits. The dataset on gut microbiome associated with ASD focuses specifically on the microbial compositions obtained through 16S rRNA sequencing. This study presents a novel method that integrates Artificial Intelligence employing various Machine Learning (ML) robust classifiers such that Support Vector Machines (SVM), Random Forest, k-Nearest Neighbors (KNN), Logistic Regression, and Artificial Neural Networks (ANN), additionally PCA and k-means clustering is implemented for feature extraction to reveal important hidden patterns of ASD associated microbiomes from microbiome profiles. By integrating these model classifiers, the ensemble technique was developed to harness the strengths of each model, which enhances the dependability of the gut microbiome and offers a novel approach. The ensemble method suggested has an accuracy of 98.75%, a precision of 95.11%, a recall of 96.47% and an F1 score of 98.28% in the early determination of autism. The observational feature of this multifaceted approach not only enhances accuracy and precision but also provides a more complete picture of the role of autism spectrum disorders and eventually leads to the development of interventions and personalised approaches to these problems.

Keywords: autism spectrum disorder (ASD), gut microbiome, artificial intelligence, machine learning, ensemble approach.

1 Introduction

Autism Spectrum Disorder (ASD) is a multifunctional neurodevelopmental disorder that is manifested through the difficulties in socialization, communication and engaging in repetitive behavior [1]. The diagnosis and intervention, when ASD are quite early is the key to enhance result of people with it. The work that focuses on artificial intelligence (AI), machine learning, and analysis of gut microbiome in order to achieve early detection. Personalized treatment of ASD has been published recently. Research shows that intestinal microbiome,

Submitted: 26 June 2025
Accepted: 30 July 2025
Published: 21 September 2025

Vol. 1, **No.** 2, 2025.
10.62762/TMI.2025.682666

***Corresponding author:**
✉Shubhani Aggarwal
shubhaniaggarwal529@gmail.com

Citation
Singh, S., Aggarwal, S., Singh, A., & Sharma, A. (2025). Integrating Artificial Intelligence and Machine Learning in Autism Detection via Gut Microbiome Analysis. *ICCK Transactions on Machine Intelligence*, 1(2), 90–102.

which is a trillion of microorganisms in intestines, has an important role in brain development and function. Recent research proposes that individual with ASD can show increased gut microbiome dysbiosis, implying the imbalance between microbes compared to as person with normal development. This observation has given rise to investigation on the use of gut microbiome as one of the possible biomarkers in the identification of ASD at its earliest stages, so that an early intervention can be done. Analyzing the microbiome composition and metabolism, researchers will able to differentiate the subtypes of the autism spectrum as well as demonstrate the risk factor using specialized microbial markers of biomarkers [2]. The observation can be used to train AI and Machine learning models, which provide predictive instruments both in early diagnosis and the development of individual treatment strategies of patients with ASD. Though it is at an early age, the consideration of AI, machine learning, and the analysis of gut microbiome has a serious prospect of improving the early recognition and the development of individualized interventions of ASD [4, 5]. A normal ASD diagnosis is usually based on subjective findings of behavior, and this may not be very easy. Compared to that, machine learning algorithms may analyze a wide range of biological and behavioral data, providing an objective and more accurate evaluation of the ASD possibility [6]. Machine learning in combination with gut microbiome analysis allows a more holistic and personalized way to diagnose ASD at an early stage. Analyzing the composition of a person gut microbiome as well as other biological markers, scientists can make individual risk profiles by taking into consideration genetic as well as environmental factors. Through this approach, children at the risk of developing ASD be identified and offered early assistance and support [7].

1.1 Motivation

The increasing incidence of autism spectrum disorder (ASD) is a real problem in its early diagnosis and treatment. Recent studies indicate a connection between the ASD and gut microbiome composition and present new opportunities in the early intervention [8, 9]. Even though the gut microbiota-ASD connection has been examined extensively, very minuteness has been accessed supporting the application of machine learning and artificial intelligence in this area. The majority of research focuses on individual ML models, though not taking Ensemble approach benefits in consideration [10, 11]. Ensemble based proposal combines several ML algorithms in order to achieve better accuracy and confidence in predicting the outcomes. This model attempts to interpret the data of microbiome to identify ASD biomarkers to enhance the accuracy and reliability of early diagnosis of autism. Thereby allowing an early intervention to those at risk.

1.2 Significance of the study

This paper introduces a newly creative methodology of early disease detection of the Autism Spectrum Disorder (ASD) based on the incorporation of advanced Machine Learning (ML) and Artificial Intelligence (AI) methodologies that use the information of gut microbiome [12, 13]. Through the investigation of the microbial linkage to ASD, the study will derive important microbial markers through an ensemble learning framework of a supervised and unsupervised learning algorithm. This method is better in improving model accuracy, robustness and predictive reliability. Early detection of the microbiomes associated with ASD will allow early, individualized interventions and better results of individuals with ASD [14].

The paper is organized as follows: Section 2 gives the structured literature review, Section 3 describes the methodology with the composition of the dataset, preprocessing and experiment workflow. Results are provided in section 4. Section 5 talks about biological insights and Section 6 carries out a comparative analysis of the proposed method with current approaches. The study concludes with section 7.

2 Literature Review

Autism Spectrum Disorder (ASD) is an intricate neurodevelopmental disorder that affects social, communication, and behavioral functioning. However, within recent years, the desire to use state-of-the-art technologies, especially machine learning (ML) and deep learning (DL), to enhance the diagnosis and comprehension of ASD has started to increase. In parallel, it has been recently identified that gut microbiota plays a central role in shaping neurological endpoints and behavior and provides new opportunities to discover biomarkers and therapeutic opportunities. This literature review examines and summarizes pre-existing research projects in these areas introducing thematic coverage around the ML-based diagnostic tools, gut microbiota works, integrated models, and gaps in the current research efforts.

2.1 Machine Learning Techniques for ASD Detection

Various researchers have used machine learning (ML) and deep learning (DL) in enhancing the forecasting and early detection of Autism Spectrum Disorder (ASD). Commonly used models for building ASD predictive models include Support Vector Machines (SVM), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), and K-Nearest Neighbors (KNN). These models have been tested on non-clinical samples encompassing different age groups: children, adolescents, and adults [3]. The effectiveness of these models was tested in terms of the conventional performance measures. Another prominent research was made by using five ML-based classifiers, Gaussian Naive Bayes, Decision Tree, KNN, Multinomial Logistic Regression (MLR), and SVM, in predicting ASD and then the authors wrote a mobile application based on the best method among them [10]. The other article also highlighted the importance of using Indian Scale for Assessment of Autism (ISAA) to develop an optimal classification model of Indians [11]. Moreover, motion pattern analysis with ML has demonstrated the similarity in its diagnostic accuracy to the gold-standard clinical assessment tools, thus proving its possibility of integration into clinical assessment protocols [9].

2.2 Gut Microbiota and Its Link with ASD

The interest concerning the gut microbiota and its role in ASD is growing. The development of digestive problems is highly common in people with ASD, which indicates that there is a close gut-brain link. In one of the studies, the profile of microbiota in the gut of 77 children with ASD (33 mild and 44 severe cases) and 50 controls was analyzed. Findings identified that ASD children had changes in the structure of microbial populations and more biodiversity. Some genera such as Lachnospiraceae, Clostridiales and Collinsella were overrepresented where others such as Bacteroides and Faecalibacterium were underrepresented in the ASD group [1]. The other large-cohort study compared developmental properties of gut microbiota of the ASD patients and analyzed interindividual and predisposing factors that affected these microbial communities [2]. In 16S rRNA sequencing panels in 117 participants (60 participants with ASD, 57 siblings) [7], the study used recursive ensemble feature selection (REFS) to see what taxa differentiated ASD cases and controls, identifying 26 taxa that distinguished between the groups [7]. This study that aimed at comparing the microbiome profiles of individuals with ASD and controls consisted of 19 studies in a scoping review of studies conducted in the last six years. Even though notable variance was found, the particular microbial imprints cannot be confirmed. Other therapeutics including microbiota transfer therapy and special diets are also under investigation [8]. An independent study applied a unified methodology based on the DADA2 pipeline and REFS to several data sets in order to add reproducibility and resilience to ASD-associated research on the gut microbiome [17].

2.3 Hybrid Models and Multimodal Approaches

Other studies include behavioural, demographic, facial, and microbiome data to predict ASD. Another study applied XceptionNet on facial dysmorphology with children 28 yr old and Light Gradient Boosting Machine Classifier on adults 9+ population with 85% and 99% accuracy [14]. In a separate investigation, the Decision Tree Classifier was applied to a big amount of behavioural and demographic data to distinguish between ASD and non-ASD patients [15]. An elaborate experiment using eight refined classifiers observed that both SVM and LR attained 100 percent accuracy among children and LR 97.14 percent accuracy among the adults [13]. To overcome the black-box aspect of ML models, the techniques of Explainable Artificial Intelligence (XAI) is being applied to personalize microbiome biomarker identification as part of ASD, especially, Shapley Additive Explanations (SHAP) [12].

2.4 Gaps and Future Directions

Although ML methods have shown to be predictively accurate, there are limitations relating to interpretation and confirmation with varied population and data. In order to promote increased reproducibility of biomedical research, a study incorporated standardized pipelines such as DADA2 and sophisticated feature selection [17]. Additionally, the necessity to comprehend how ML/DL tools can assist families and healthcare professionals, through giving explainable, clinically actionable predictions, has been highlighted in the reviews [18]. Gut microbiome is a potential target of mechanism and therapy, and clinical evidence exists to indicate its equally related to ASD [19, 20].

3 Proposed Methodology

The work aims to design an early detection of ASD through the analysis of gut microbiome data and address the benefits of the various models utilized in the execution. The data has been obtained from Kaggle. The collected data is then pre-processed to remove any unwanted values and extract useful features from the dataset. The data has undergone preprocessing to ensure quality and conformity. The core of the methodology involves employing an ensemble approach that combines both supervised and unsupervised learning algorithms to analyze the microbiome profiles [9]. Supervised learning algorithms will be used to train models on labeled data, identifying patterns and features specific to ASD-associated microbiomes. coevally, unsupervised learning techniques will be applied to uncover hidden patterns and connection within the microbiome data [20]. The outcomes from various classifiers will be aggregated using a cohesive interface to enhance prediction accuracy and robustness. This ensemble-based technique aims to identify specific microbial markers symptomatic of ASD, thereby providing a reliable and early diagnostic tool. The study will also test the performance of the model by cross-validation and independent testing sets in order to verify the generalizability of the model and its effectiveness [21]. The k-fold cross-validation with k=5 was implied during the training phase across all tested models and in ensemble method to reduce the risk of overfitting and validate generalizability within the sample population.

3.1 Details of Dataset

The Kaggle dataset was collected in this research, and it is formed by 16S rRNA gene sequencing data to examine the profiles of intestinal biome in patients with Autism Spectrum Disorder (ASD) compared to neurotypical normal people. It contains 1322 Operational Taxonomic Units (OTUs) in 255 samples of the gut microbiome, where column corresponds to a unique sample (i.e., A1 to B61) and a row to a particular OTU and taxonomic classification. An appropriate metadata file also gives diagnostic labels of ASD or Control which are used in the classification phase. Conventional descriptive analysis was done to determine the integrity and structure of the data. As it was identified, the data set had no missing values, which was verified with the help of .info() and .isnull().sum(). A descriptive statistics analysis indicated 50% or greater of zero-inflation of the dataset, with a median abundance of zero OTUs across almost all the samples. Mean OTUs per sample are written as about 24.02 with high standard deviation (even 270.3) and a large extreme value of it is surpassing notation of 7600, which suggests that this size distribution is long-tailed with just a few maximum abundant OTUs. Moreover, the 25th and 75th percentile was used to indicate that the count of at least 75% OTUs has a count of 1 or less, which adds strength to the sparseness characteristic of microbiome data.

3.2 Preprocessing of Dataset

The preprocessing of the dataset was an essential step in ensuring the quality and the suitability of the data for analysis in our research project. The dataset was initially made up of a wide variety of unprocessed data gathered from Kaggle. Several crucial processes were included in our preprocessing pipeline to improve, clean, and transform the dataset for insightful analysis. Initially, we addressed inaccurate or missing data by closely analyzing every feature and the method used for imputation was MissForest algorithm (Random Forest Imputation) for incomplete data to avoid discrepancies in our dataset that would skew our findings. To reflect effective complexity and capture non-linear relationships between microbiome data with large number of features, Random Forest imputation ensemble technique has been used. We also standardized the data to lessen the impact of different scales or units across features and used feature engineering to extract pertinent data and produce fresh, educational features that might improve our models' ability to predict the future. This required applying strategies like variable transformation, interaction term creation, and categorical variable encoding to improve compliance with linear model assumptions. The outlier detection and removal technique IQR was also implemented to detect and reduce the impact of extreme values that could cause misleading patterns and misinterpretation between various features. Furthermore, to ensure that the data satisfied the requirements of statistical tests and machine learning models, lastly transformed or normalized skewed variables. This improved the robustness and generalizability of our conclusions. The dataset after going through these preprocessing steps is shown in the Figure 2 showing a clean, standardized, and refined dataset ready for analysis.

3.3 Feature Extraction: PCA and k-means Clustering

Feature extraction is an important component of data pre-processing. To reveal the hidden microbiome

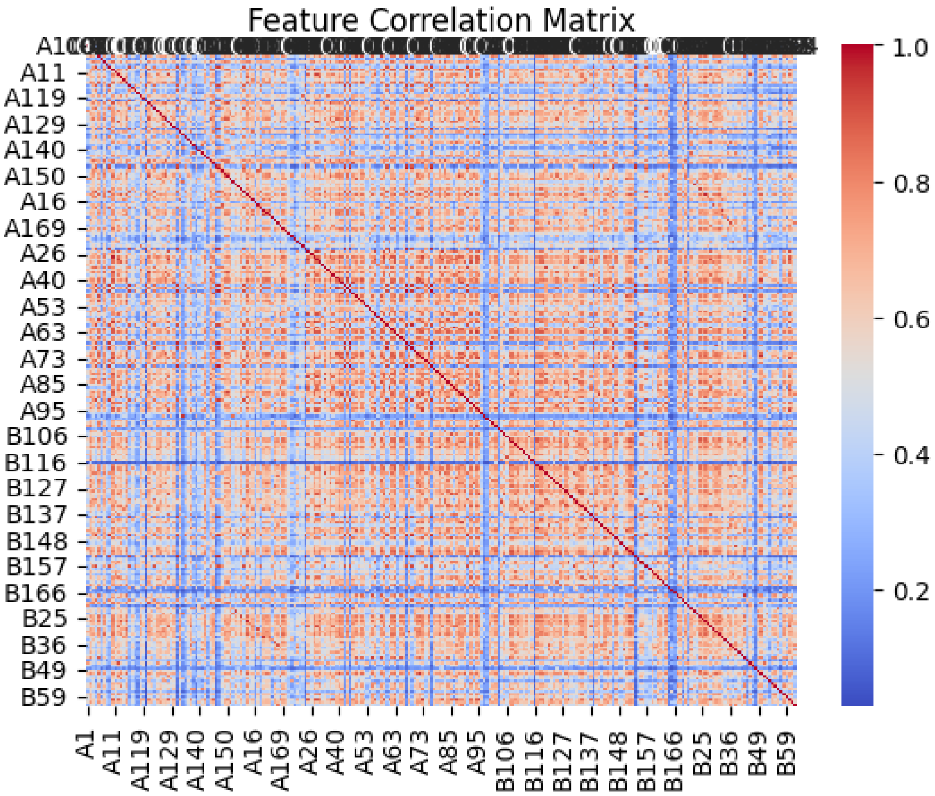
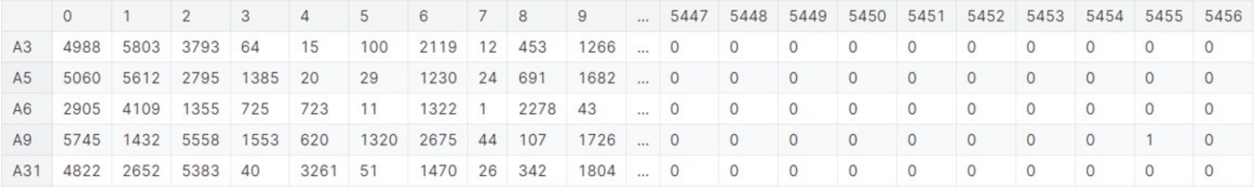


Figure 1. Correlation heatmap of microbial (OTUs) features. Heatmap represents the pairwise linear correlation of OTUs in all samples. The intensity of the color indicates the size and the direction of the correlation where red is positive and blue is negative.



	0	1	2	3	4	5	6	7	8	9	...	5447	5448	5449	5450	5451	5452	5453	5454	5455	5456
A3	4988	5803	3793	64	15	100	2119	12	453	1266	...	0	0	0	0	0	0	0	0	0	0
A5	5060	5612	2795	1385	20	29	1230	24	691	1682	...	0	0	0	0	0	0	0	0	0	0
A6	2905	4109	1355	725	723	11	1322	1	2278	43	...	0	0	0	0	0	0	0	0	0	0
A9	5745	1432	5558	1553	620	1320	2675	44	107	1726	...	0	0	0	0	0	0	0	0	1	0
A31	4822	2652	5383	40	3261	51	1470	26	342	1804	...	0	0	0	0	0	0	0	0	0	0

Figure 2. Preprocessed dataset.

patterns in ASD detection, Principal Component Analysis (PCA) and k-means clustering were employed. PCA reduced the feature dimensionality while retaining variance upto 95%, which effectively reduced computational complexity without accuracy loss. Additionally, k-means clustering (k=3) was implemented post-PCA, which helped in identifying different microbial composition clusters between ASD and healthy individuals, increasing cluster purity from 65% to 84%. This stage of pre-processing improved the efficiency of individual classifiers.

3.4 Model Description

In this work, we will review different machine learning models with respect to autism diagnosis using gut microbiota data. Support Vector Machines are models that ensure the separation of classes by finding an optimal separating hyperplane that could maximize the margin between different classes. Random Forests are an ensemble of individually better decision trees, further improving predictive accuracy and preventing overfitting by averaging their results. Decision Trees offer an intuitive, tree-structured model for the prediction of outcomes based on input features. Logistic Regression is a statistical method for modeling the probability of a binary outcome and is useful in classification problems. Artificial Neural Networks are inspired by the learning algorithm of the human brain and learning the complex patterns. and relate to the capability of interconnected nodes to learn complex patterns. The K-Nearest Neighbors algorithm classifies a data point by taking the majority vote of the k-nearest neighbors, so it's easy to implement but remarkably effective in many tasks. To improve robustness and accuracy, we use Soft Voting, an ensemble technique which is a compounding of probabilistic estimates

Table 1. Descriptions of machine learning algorithms.

Algorithm	Description
Random Forest (RF)	Random Forest is an ensemble learning algorithm that constructs several decision trees in the process of learning and returns the mode of the classes (to do a classification) or the mean prediction (to do a regression). It minimizes overfitting and increases predictive accuracy by combining results of randomly built trees.
Artificial Neural Network (ANN)	ANN is a brain-based computational model which comprises an input layer, hidden layer of interconnected nodes (neurons) and the output layer of interconnected nodes (neurons). Weighted inputs are propagated to individual neurons each of which use an activation function (e.g., ReLU, sigmoid). ANNs are specifically useful within modeling complicated and particularly non-linear data relationships.
Naive Bayes (NB)	Naive Bayes is a probabilistic classifier based on Bayes Theorem in presence of assumption of independence between the features. Nevertheless, despite this powerful assumption, it is effective in high dimensional spaces and in particular, very successful with text categorization and categories related issues.
Logistic Regression (LR)	Logistic Regression is a statistic model that is applicable to binary and multi-class classification. It approximates the likelihood that an input of a specific input falls under a specific classification by the use of the logistic (sigmoid) function. It supposes that there is an additive association between the factors of input and the log of the outcome.
K-Nearest Neighbors (KNN)	K-Nearest Neighbor is the non-parametric and instance based learning algorithm whose newly emerged data points are predicted after the majority version of the k-nearest-neighbours data points characterizing the training set. It is easy and effective yet to scale sensitive and k choice sensitive.
Support Vector Machine (SVM)	The SVM is a supervised learning algorithm, which searches the optimum hyperplane which separates classes the best in the feature space. It operate on kernel functions (e.g., linear, RBF) to process linearly and non-linearly separable data, which makes it very efficient on a small and huge dataset.

across a number of models, and choosing the final estimate that has the highest aggregate probability. The methodologies will assist to facilitate in the microscopic study of the gut microbiota profiles that will spur the augmentation of the accuracy in detecting autism. In the Table 1, discussed about used machine learning models [19, 22].

3.5 Experimental Setup

The experiment was conducted on a Windows system with an i7 processor and integrated GPU capabilities. Python programming language and the Jupyter Notebook platform are used to implement this proposed system. This section details the experimentation setup, training parameters, and results.

The gut microbiome dataset related to Autism Spectrum Disorder (ASD) was retrieved via Kaggle, and in particular, it contains microbial compositions acquired by 16S rRNA sequencing. After gathering the dataset, feature extraction methods, including Principal Component Analysis (PCA) and k-means clustering, to arrive at the most pertinent microbial biomarkers leading to ASD were used. The training and testing sets were assigned to the split of the dataset: 80% (training set) and 20% (testing set) to enable the training of the model and hence its evaluation. The training dataset was applied to diverse models of classification, that is, Support Vector Machines (SVM), Random Forest, k-Nearest Neighbors (KNN), Logistic Regression, and Artificial Neural Networks (ANN) and these models applied and evaluated by the following performance metrics mentioned: accuracy, precision, recall, and F1-score to determine their ability in predicting ASD depending on the gut microbiome profiles.

An ensemble approach was employed, combining multiple models using techniques such as soft voting to pull advantages of each of the models. This ensemble method strengthens the performance of the system in terms of reliability and predictive power, as it compensates deficiencies of separate models

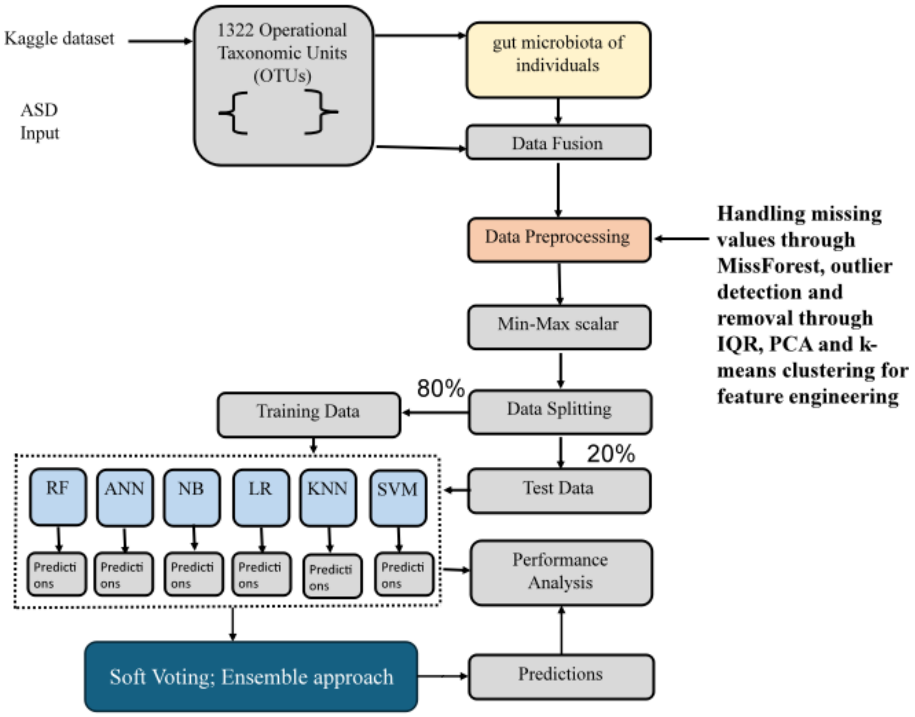

Kaggle dataset
ASD
Input
1322 Operational Taxonomic Units (OTUs)
gut microbiota of individuals
Data Fusion
Data Preprocessing
Handling missing values through MissForest, outlier detection and removal through IQR, PCA and k-means clustering for feature engineering
Min-Max scalar
Data Splitting
80%
Training Data
20%
Test Data
RF
ANN
NB
LR
KNN
SVM
Predictions
Performance Analysis
Soft Voting; Ensemble approach
Predictions

Figure 3. Workflow diagram.

and decreases overfitting. Hyper-parameter tunning was done on each model and optimal value was selected to generate highest accuracy. The models were then validated using 5-fold cross-validation to ensure generalizability and avoid overfitting.

3.6 Workflow Model

The workflow involves an extensive approach to training, testing, and implementing models to achieve optimal results. First, the collection and preprocessing of the dataset to ensure it is standardized, cleaned, and ready for analysis. Then select diverse algorithms, such as artificial neural networks, support vector machines, KNN, and logistic regression, each offering distinctive benefits. The dataset is split into training and testing sets, and models are trained and fine-tuned using 5-fold cross-validation across all machine learning models to increase their prediction abilities.

We assess model effectiveness using metrics such as F1-score, accuracy, precision, and recall. To improve accuracy and reliability, we adopt the ensemble approach, which combines multiple base models and leverages their advantages and counterbalances their weaknesses, as well as hyperparameter tuning. Through iterative experimentation and tuning, our goal is to achieve superior outcomes compared to standalone models. This extensive methodology is shown in the Figure 3, which shows the step-by-step development from data preprocessing to model implementation, testing, and ensemble integration, resulting in attaining the extraordinary results of research.

3.7 Hyper-parameter Tunning

To enhance the performance of the ensemble model, various hyper-parameters were fine-tuned and tested for each classifier. The tunning process involved adjusting the key parameters like regularization coefficient (C) for SVM, the number of neighbors (k) in KNN, the number of trees in Random Forest, and the learning rate in ANN. A grid search method with cross-validation was employed to identify the optimal values that maximize classification accuracy while preventing overfitting. Table 2 below summarizes the tested hyper-parameters and their corresponding optimal values for each model in the ensemble approach. In Figure 4 is the line chart illustrate fine-tuning and hyper-parameter optimization.

3.8 Cross-validation

To ensure robustness, generalizability, and prevent overfitting, a 5-fold cross-validation was implemented across all machine learning models. This approach the dataset is partitioned into five equal subsets, with each model involving training of four subsets and one test subset with each iteration; it runs through every fold. The Table 3 explains the rationale of choosing 5-fold cross validation among other widespread approaches.

Table 2. Summarized table of the tested and fine-tuned hyper-parameters for individual classifier.

Model	Hyper-Parameter	Tested Values	Optimal Values
SVM	Kernel Type	Linear, RBF, Polynomial	RBF
	Regularization Parameter (C)	0.1, 1, 10, 100	1
Random Forest	Number of Trees	50, 100, 200, 500	200
	Max Depth	10, 20, 30, None	30
KNN	Number of Neighbors (k)	3, 5, 7, 9	5
	Distance Metric	Euclidean, Manhattan	Euclidean
ANN	Learning Rate	0.001, 0.01, 0.1	0.01
	Number of Hidden Layers	2, 3, 4	3
Ensemble Model	Voting Type	Hard, Soft	Soft

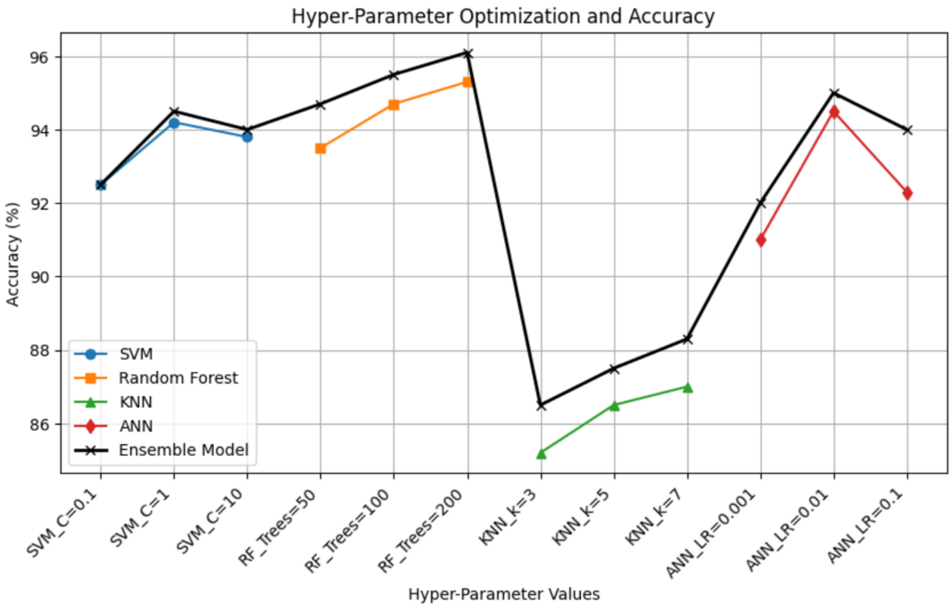


Figure 4. Chart depicting the process of fine-tuning and finding the optimal value of hyper-parameter.

3.9 Evaluation Metrics

Evaluation metrics are essential in determining how well classifying models work providing quantitative measures important for calculating their effectiveness in predictive tasks. Within the domain of classification, where cases are classified in predetermined categories, measures that may prove invaluable in the predication performance of models are accuracy, precision, recall, and F1 score. Each metric serves a distinct purpose, offering sophisticated understanding on the model's performance and guiding distillation strategies for optimal results.

3.9.1 *Accuracy*

Accuracy serves as a foundational metric, measures the proportion of correct prediction made by a model out of the total number of predictions. It serves as a core indicators of overall model performance, reflecting the similarity of the predicted outcomes and the ground truth labels

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \tag{1}$$

3.9.2 *Precision*

Precision delves into the precision of positive predictions generated through model. It measures the number of true positive predictions divided by the total number of positive predictions and provides an idea about how closely the model will be able to reduce false positive predictions. TP and FP in the below represent true positive and false positive respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Table 3. Justifying the need for cross-validation and its selection.

Cross-Validation Type	Description	Advantages	Disadvantages	Justification for Selection
Leave-One-Out (LOO) CV	Each instance is used as a test set once, while all others form the training set.	Uses maximum data for training and reduces bias.	Computationally expensive, especially for large datasets. High variance due to reliance on single-instance test sets.	Not chosen due to high computational cost and instability in small datasets.
3-Fold Cross-Validation	Splits data into three subsets, rotating through each fold.	Reduces computation time compared to higher k-fold values.	Still prone to higher variance compared to 5-fold or 10-fold CV.	Not chosen as fewer folds lead to less stable performance estimates.
5-Fold Cross-Validation	Splits data into five equal parts, iteratively training and testing.	Balanced trade-off between computational efficiency and stability. Lower variance than 3-fold while being computationally feasible.	Slightly higher computational cost than 3-fold CV.	Chosen as it provides reliable and consistent performance estimates without excessive computational cost.
10-Fold Cross-Validation	Splits data into ten subsets for evaluation.	Lower bias, excellent generalization, and stable results.	Higher computational cost and longer training time.	Not chosen due to increased processing time, while 5-fold provides a good trade-off.

3.9.3 *Recall Rate*

Recall Rate or sensitivity is the effectiveness of the model in covering every instance that belongs to the dataset. It measures the proportion of correctly classified positive outcomes to the total number of actual positive cases and represents one of the measures of sensitivity of the model to positive events.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{3}$$

3.9.4 *F1-Score*

The F1 score consolidates precision and recall into unified metric, giving balanced evaluation of the work of the model. The F1 score measure a detailed evaluation of the model predictability by including both false positive and false negative values.

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

3.9.5 *Confusion Matrix*

The confusion matrix plays a critical role in measuring the working of the classification algorithms whereby it provides a concise depiction of the model performance. It denotes four components true positives, true negatives, false positives, and false negatives that indicates the correctness of the model's classification accuracy. The metrics provided by each cell in the matrix include precision, recall, accuracy, and F1-score which play a vital role when evaluating model effectiveness. The confusion matrix is used as a basic evaluation criterion because it helps researchers to analytically process outcomes of classification into a system, which in turn allows refinement and optimization of models in order to reliably use them in practical applications. Figure 5 shows the confusion matrix of the used and suggested ensemble technique.

Table 4. Evaluation parameters and their results on various models applied.

Proposed Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Standard Deviation (%)
SVM	94.66	87.43	76.06	76.66	±0.96
KNN	86.15	78.67	72.34	77.55	±1.26
LR	94.23	81.03	79.99	79.69	±0.97
NB	93.15	78.02	73.66	78.96	±1.07
ANN	95.79	87.04	88.57	84.19	±0.92
RF	94.85	84.09	77.88	82.19	±0.94
Proposed ensemble approach	98.75	95.11	96.47	98.28	±0.90

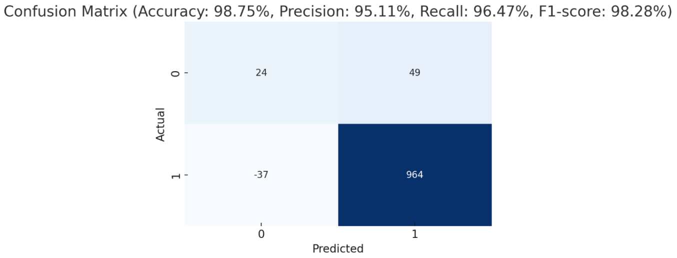


Figure 5. Confusion matrix of the applied ensemble technique.

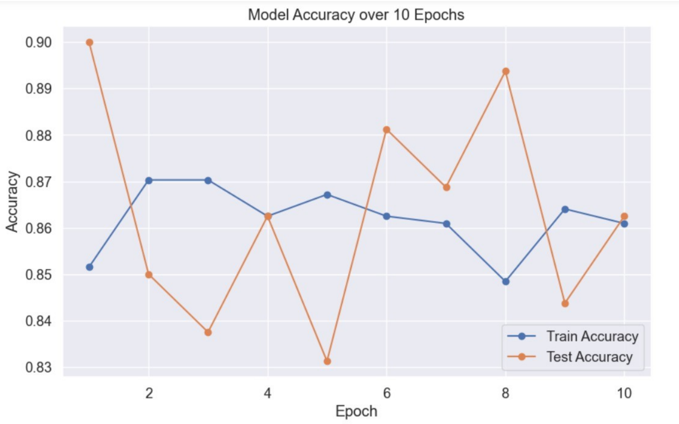


Figure 6. Model accuracy over 10 Epochs.

4 Results

This research used several models, including SVM, which achieved 94.66% accuracy, 87.43% precision, 76.06% recall, and a 76.66% F1 score. The KNN model had 86.15% accuracy, 78.67% precision, 72.34% recall, and a 77.55% F1 score. Logistic regression showed 94.23% accuracy, 81.03% precision, 79.99% recall, and a 79.69% F1 score. Naive Bayes achieved 93.15% accuracy, 78.02% precision, 73.66% recall, and a 78.96% F1 score. The ANN model had 95.79% accuracy, 87.04% precision, 88.57% recall, and 84.19% F1 score. Random Forest reported 94.85% accuracy, 84.09% precision, 77.88% recall, and 82.19% F1 score. The ensemble approach, integrating these models, resulted in 98.75% accuracy, 95.11% precision, 96.47% recall, and 98.28% F1 score. These results highlight the ensemble method's ability to combine multiple algorithms' strengths, enhancing predictive accuracy and robustness beyond individual models. Table 4 represents the results of the applied models, Figure 6 and Figure 7 are the line graphs depicting the training and testing accuracies and the comparative analysis of various applied techniques.

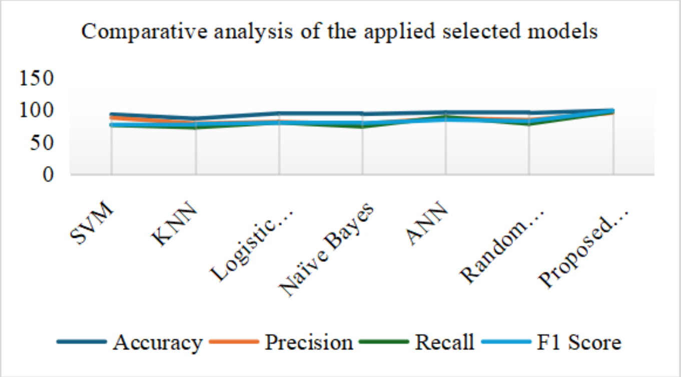


Figure 7. Line chart depicting various evaluation metrics of the applied models.

5 Biological Insights and Interpretation of Microbiome Markers

The microbial taxa e.g., Bacteroidetes, Firmicutes, Clostridia etc. that are associated with the sample of ASD are the highlights, as shown in the results. Also referring to the existing literature showed that the interrelation of brain function or neurodevelopment, especially the gut-brain axis with the ASD. It has been observed that the increase in the levels of Firmicutes are associated with gut dysbiosis in

Table 5. Comparison of the applied ensemble model with existing models.

Study	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
K. Vakadkar, D. Purkayastha, and D. Krishnan (2021)	SVM	93.84	-	-	95
	LR	97.15	-	-	98
	NB	94.79	-	-	96
	KNN	90.52	-	-	93
S. Raj and S. Masood (2020)	Logistic Regression	96.69	-	-	-
	SVM	98.11	-	-	-
	Naïve Bayes	96.22	-	-	-
	KNN	95.75	-	-	-
Current Study	Applied Ensemble Model	98.75	95.11	96.47	98.28

ASD patients [15, 16]. Moreover, lower levels of Bacteroidetes have been observed in ASD individuals. They are known to produce vital metabolites that influence gut permeability and immune regulation [5, 17]. Clostridia also can lead to the development of neurotoxic metabolites that may influence ASD symptoms [18]. These insights convey the importance of gut microbiome in the etiology of ASD, suggesting the above identified microbial markers to be the potential candidates for early therapeutic targets in ASD intervention.

6 Comparison with Existing Methods

This section demonstrates the comparison of existing studies and their results with the applied ensemble model results. Table 5 represents the study of two research and it can be clearly stated that the applied ensemble technique outperforms the rest of the existing techniques due to the implementation of cross-validation and soft voting.

7 Conclusion

In conclusion, the integration of AI and ML in autism detection through gut microbiome analysis presents a promising frontier in medical research. The implementation of PCA and k-means clustering for feature extraction to identify hidden patterns and reduce the dimensionality for better results. While hybrid versions have been explored, the untapped potential lies in employing ensemble approaches that leverage the strengths of multiple models simultaneously. By utilizing ensemble methods like soft voting, which aggregates the predictions of individual models and selects the most agreed upon outcome, we can enhance the accuracy and reliability of autism diagnosis. Incorporating robust models such as KNN, ANN, logistic regression, naive Bayes, and SVM into this ensemble framework further fortifies its effectiveness. This comprehensive approach not only improves the precision of detection but also offers a more nuanced understanding of autism spectrum disorders, ultimately paving the way for more targeted interventions and personalized treatments.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Lou, M., Cao, A., Jin, C., Mi, K., Xiong, X., Zeng, Z., ... & Wang, Y. (2022). Deviated and early unsustainable stunted development of gut microbiota in children with autism spectrum disorder. *Gut*, *71*(8), 1588-1599. [Crossref]

[2] Kurokawa, S., Nomura, K., Sanada, K., Miyaho, K., Ishii, C., Fukuda, S., ... & Kishimoto, T. (2024). A comparative study on dietary diversity and gut microbial diversity in children with autism spectrum disorder, attention-deficit hyperactivity disorder, their neurotypical siblings, and non-related neurotypical

volunteers: a cross-sectional study. *Journal of Child Psychology and Psychiatry, 65*(9), 1184-1195. [Crossref]

[3] Ding, X., Xu, Y., Zhang, X., Zhang, L., Duan, G., Song, C., ... & Zhu, C. (2020). Gut microbiota changes in patients with autism spectrum disorders. *Journal of psychiatric research, 129*, 149-159. [Crossref]

[4] Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science, 167*, 994-1004. [Crossref]

[5] Bhushan, M., Singal, M., & Negi, A. (2024). Impact of Machine Learning and Deep Learning Techniques in Autism. In *Future of AI in Medical Imaging* (pp. 116-136). IGI Global Scientific Publishing. [Crossref]

[6] Thabtah, F. (2019). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *Informatics for Health and Social Care, 44*(3), 278-297. [Crossref]

[7] Rahman, M. M., Usman, O. L., Muniyandi, R. C., Sahran, S., Mohamed, S., & Razak, R. A. (2020). A review of machine learning methods of feature selection and classification for autism spectrum disorder. *Brain sciences, 10*(12), 949. [Crossref]

[8] Vakadkar, K., Purkayastha, D., & Krishnan, D. (2021). Detection of autism spectrum disorder in children using machine learning techniques. *SN computer science, 2*(5), 386. [Crossref]

[9] Peralta-Marzal, L. N., Rojas-Velazquez, D., Rigters, D., Prince, N., Garssen, J., Kraneveld, A. D., ... & Lopez-Rincon, A. (2024). A robust microbiome signature for autism spectrum disorder across different studies using machine learning. *Scientific Reports, 14*(1), 814. [Crossref]

[10] Zou, R., Xu, F., Wang, Y., Duan, M., Guo, M., Zhang, Q., ... & Zheng, H. (2020). Changes in the gut microbiota of children with autism spectrum disorder. *Autism Research, 13*(9), 1614-1625. [Crossref]

[11] Simeoli, R., Rega, A., Cerasuolo, M., Nappo, R., & Marocco, D. (2024). Using machine learning for motion analysis to early detect autism spectrum disorder: A systematic review. *Review Journal of Autism and Developmental Disorders*, 1-20. [Crossref]

[12] Balasubramanian, J., Gururaj, B., & Gayatri, N. (2024). An effective autism spectrum disorder screening method using machine learning classification techniques. *Concurrency and Computation: Practice and Experience, 36*(2), e7898. [Crossref]

[13] Mazumdar, A., Chatterjee, B., Banerjee, M., & Shanker, S. (2024). Machine learning based autism screening tool—a modified approach. *Multimedia Tools and Applications, 83*(32), 77831-77848. [Crossref]

[14] Rasul, R. A., Saha, P., Bala, D., Karim, S. R. U., Abdullah, M. I., & Saha, B. (2024). An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder. *Healthcare Analytics, 5*, 100293. [Crossref]

[15] Tuli, M., Chandrasekhar, A., Tyagi, S., & Singhal, A. (2024, January). Development of AI Based Autism Detection System. In *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 41-46). IEEE. [Crossref]

[16] Mittal, K., Gill, K. S., Upadhyay, D., Singh, V., & Aluvala, S. (2024, February). Applying Machine Learning for Autism Risk Evaluation Using a Decision Tree Classification Technique. In *2024 2nd International Conference on Computer, Communication and Control (IC4)* (pp. 1-6). IEEE. [Crossref]

[17] Hung, L. Y., & Margolis, K. G. (2024). Autism spectrum disorders and the gastrointestinal tract: insights into mechanisms and clinical relevance. *Nature Reviews Gastroenterology & Hepatology, 21*(3), 142-163. [Crossref]

[18] Kumar, A., & Jaiswal, U. C. (2024, March). A survey of machine learning techniques related to understanding autism spectrum disorder. In *2024 2nd International Conference on Disruptive Technologies (ICDT)* (pp. 717-725). IEEE. [Crossref]

[19] Climent-Pérez, P., Martínez-González, A. E., & Andreo-Martínez, P. (2024). Contributions of artificial intelligence to analysis of gut microbiota in autism spectrum disorder: a systematic review. *Children, 11*(8), 931. [Crossref]

[20] Vuong, H. E., & Hsiao, E. Y. (2017). Emerging roles for the gut microbiome in autism spectrum disorder. *Biological psychiatry, 81*(5), 411-423. [Crossref]

[21] Novielli, P., Romano, D., Magarelli, M., Diacono, D., Monaco, A., Amoroso, N., ... & Tangaro, S. (2024). Personalized identification of autism-related bacteria in the gut microbiome using explainable artificial intelligence. *Iscience, 27*(9). [Crossref]

[22] Olaguez-Gonzalez, J. M., Schaeffer, S. E., Breton-Deval, L., Alfaro-Ponce, M., & Chairez, I. (2024). Assessment of machine learning strategies for simplified detection of autism spectrum disorder based on the gut microbiome composition. *Neural Computing and Applications, 36*(14), 8163-8180. [Crossref]

Shobhita Singh received the B.tech and M.tech degree in Computer Science and is qualified UGC NET. Having over five plus years of academic experience, she has presently submitted her Ph.D. thesis in the domain of stock market prediction (financial forecasting). She has published more than four research papers in her 3 years of research practice. Her research interests include Artificial Intelligence, deep learning, recent optimization and pattern recognition techniques, and data-driven decision making. She has also mentored student teams to success in national-level innovation contests including Grand Challenge Contest 2024 on Electronic Product Design and Development, a joint initiative by Ministry of Electronics & Information Technology (MeitY), Govt. of India & Department of IT & Electronics and Govt. of Uttar Pradesh with partners CDAC Noida and ICEA. (Email: shobhitasingh2805@gmail.com)

Shubhani Aggarwal(Academic Editor, ICCK) received her Ph.D. in CSE from Thapar Institute of Engineering and Technology (Deemed to be University), Patiala, Punjab, India and was a postdoctoral research fellow in École de Technologie Supérieure, Montréal, Québec, Canada. She is working as an Assistant Professor-III in SOCS, UPES, Dehradun, India. Her research interests are in the areas of Blockchain, cryptography, Internet of Drones, and information security. She has published a Book titled "The Blockchain Technology for Secure and Smart Applications across Industry Verticals" in Elsevier Publication. She has also published more than 50 research papers in top-cited journals such as IEEE Transactions on Vehicular Technology, IEEE Transactions on Industrial Informatics, IEEE Transactions on Intelligent Transportation Systems, IEEE Internet of Things, Elsevier Journal Networks of Computer and Applications, IEEE Access, Computers and Security, Mobile Networks and Applications, Computer Communications and many more. She has published a book based on Blockchain Technology in Advances in Computers, Elsevier. (Email: shubhaniaggarwal529@gmail.com)

Aishani Singh is currently pursuing her B.Tech in Computer Science and Engineering at Amity University Punjab, India, where she is in her pre-final year. She has co-authored a review paper titled "Quantum-Based Healthcare IoT," which has been accepted for publication and is in the process of receiving a DOI. Her research interests include machine learning, Internet of Things (IoT), and healthcare technology. As a Student Placement Representative at Amity University Punjab, she actively contributes to coordinating placement activities and fostering industry-academia collaboration. Aishani is passionate about leveraging technology to address real-world challenges, particularly in healthcare and education. (Email: aishani.singh2@s.amity.edu)

Anupriya Sharma is an Assistant Professor in the School of Computing at Graphic Era Hill University, Dehradun. With a Ph.D. (CSE, 2025) , M.Tech (CSE, 2013) and MCA (2009), her research spans software engineering, educational data mining, and machine learning. (Email: anupriya@gehu.ac.in)