



International Journal of Computational Biology and Drug Design

ISSN online: 1756-0764 - ISSN print: 1756-0756
<https://www.inderscience.com/ijcbdd>

AdaCluCSL: an approach for autism spectrum disorder prediction using adaptive clustering smote and cost-sensitive learning

M. Kavitha, M. Kasthuri

DOI: [10.1504/IJCBDD.2024.10067421](https://doi.org/10.1504/IJCBDD.2024.10067421)

Article History:

Received:	21 November 2023
Last revised:	21 May 2024
Accepted:	23 August 2024
Published online:	09 May 2025

AdaCluCSL: an approach for autism spectrum disorder prediction using adaptive clustering smote and cost-sensitive learning

M. Kavitha* and M. Kasthuri

Department of Computer Science,
Bishop Heber College,
Affiliated to Bharathidasan University,
Tiruchirappalli-620017, Tamil Nadu, India
Email: kavitha.ca@bhc.edu.in
Email: kasthuri.ca@bhc.edu.in

*Corresponding author

Abstract: Machine learning struggles to predict autism spectrum disorder (ASD) due to real-world datasets' underlying class imbalance. Adaptive cluster SMOTE and cost-sensitive learning (AdaCluCSL) is a new approach that improves ASD prediction accuracy. Combining adaptive k-means clustering SMOTE with cost-sensitive learning achieves this. Because positive ASD samples are few, conventional classification methods sometimes have poor predictive accuracy. AdaCluCSL uses adaptive clustering to find complex clusters in the underrepresented class and address this disparity. It then creates synthetic samples using SMOTE with cluster-specific weights. Then, the enriched samples are used with the original cases to train a cost-sensitive classifier that allocates additional costs to minority class misclassifications. Experimental tests on benchmark ASD datasets show that AdaCluCSL improves ASD prediction accuracy. Precision and recall rates have improved significantly, outperforming traditional methods. The AdaCluCSL algorithm can advance ASD prediction. It solves uneven data distribution issues, making early diagnosis and intervention more reliable.

Keywords: prediction of autism spectrum disorder (ASD); K-means cluster; SMOTE; synthetic minority oversampling technique; CSL; cost-sensitive learning; imbalance classification; adaptive cluster SMOTE; cost-sensitive learning (AdaCluCSL).

Reference to this paper should be made as follows: Kavitha, M. and Kasthuri, M. (2025) 'AdaCluCSL: an approach for autism spectrum disorder prediction using adaptive clustering smote and cost-sensitive learning', *Int. J. Computational Biology and Drug Design*, Vol. 16, No. 3, pp.185–211.

Biographical notes: M. Kavitha is working as an Assistant Professor in the Department of Computer Applications at Bishop Heber College, Tiruchirappalli, Tamil Nadu, India. She graduated with a Master of Computer Science from Bharathidasan University in 2003. She completed her Master of Philosophy in Computer Science in September 2011 at Bharathidasan University, Tiruchirappalli. She is pursuing a doctorate of Philosophy in Computer Science under the broad topic, machine learning framework for early disease detection: an integrated approach to handling imbalanced data and feature selection. Her area of interest is data mining, specifically focusing on

classification technology. She is passionate about wards research and publishes research papers in various international journals. She has authored two books which are underlying her research area.

M. Kasthuri is an Associate Professor in Computer Science at Bishop Heber College, Tiruchirappalli, Tamil Nadu, India. In June 2017, she received her Decorate of Philosophy in Computer Science from Bharathidasan University in Tiruchirappalli. She completed the NET exam in December 2019. She has published research papers in national and international research forums based on stemming, web mining, image processing, IoT, deep learning, and machine learning concepts. She had completed many minor research projects, including UGC-sponsored amounts. She authored TEN Books, which are useful to CBSE School Students. She is also a NET exam trainer for Computer Science.

1 Introduction

ASD is a complex neurodevelopmental condition characterised by various challenges related to social communication, repetitive behaviours, and restricted interests. Its prevalence has been increasing globally, highlighting the urgent need for effective early diagnosis and intervention (Baio, 2018). The diagnostic process for ASD typically involves comprehensive assessments by clinicians, often relying on behavioural observations and standardised tools (Chunduri et al., 2024). However, the subjective nature of these assessments can lead to variations in diagnosis, highlighting the potential for computational approaches to provide more objective and standardised diagnostic tools (Al-Taei et al., 2023). The growing availability of large-scale datasets, advancements in data processing techniques, and the power of machine learning have offered new avenues for enhancing ASD diagnosis and prediction (Anand et al., 2023). Machine learning algorithms, driven by patterns and insights extracted from data, hold the potential to uncover subtle features and relationships that might elude human observers (Esteva and Kuprel, 2019). Researchers have explored diverse data sources, including behavioural data, neuroimaging, genetic markers, and electronic health records, to identify reliable biomarkers for ASD (Abrol and Modgil, 2021).

However, ASD prediction through machine learning is hindered by the inherent challenge of imbalanced datasets (Elayaraja et al., 2024). The scarcity of positive ASD cases, in contrast to the larger pool of negative cases, can lead to biased model training and reduced predictive performance (Zhang et al., 2021). Conventional classification algorithms prioritise the majority class, resulting in suboptimal sensitivity (recall) for the minority class (ASD cases) (Chunduri et al., 2024). This limitation necessitates the development of specialised techniques that can effectively handle imbalanced data to improve ASD prediction accuracy (Anand et al., 2024).

Over the years, researchers have explored diverse strategies to combat this issue, with techniques such as SMOTE (Synthetic Minority Over-sampling Technique), SVM-SMOTE, adaptive SMOTE, and borderline SMOTE garnering considerable attention (Ghozali, 2022). The advent of SMOTE introduced an innovative approach to synthesising artificial instances for the minority class, effectively balancing the class

distribution and enhancing the predictive capacity of models (Obeta et al., 2024). The subsequent developments, including SVM-SMOTE combined SMOTE with Support Vector Machines, adaptive SMOTE that dynamically adjusted oversampling based on data distribution, and borderline SMOTE that focused on challenging borderline instances, have all contributed to the evolution of addressing class imbalance (Jani et al., 2024). While SMOTE variants have shown promising results, a unique potential in k-means-cluster SMOTE remains largely untapped (Karthikeyan et al., 2024). This method enriches the synthetic data generation by fostering a nuanced understanding of the underlying data clusters (Kumar et al., 2024). By aligning the generation process with the clusters, k-means-cluster SMOTE capitalises on the intrinsic data structure, thereby augmenting the quality and diversity of synthetic instances (Gharban and Al-Shaeli, 2021).

Despite its potential, adopting k-means-cluster SMOTE remains relatively sparse in the existing literature (Maraju, 2024). Most research endeavours have centred around the more established SMOTE variations, overlooking the synergistic possibilities that emerge when clustering techniques are integrated into oversampling (Ghozali et al., 2022). This research seeks to bridge this gap by exploring the uncharted territory of k-means-cluster SMOTE and its interplay with cost-sensitive learning, a tandem approach that intertwines oversampling with classification, preserving class ratios while considering the varying misclassification costs (Gandhi et al., 2024).

Design of “AdaCluCSL: approach for autism spectrum disorder (ASD) prediction using adaptive clustering SMOTE and cost-sensitive learning”: The data imbalance problem of ASD prediction datasets has been emphasised (Paramasivan et al., 2024). In an era where accurate, early detection of ASD has become saliently indispensable, this approach integrates adaptive clustering SMOTE and cost-sensitive learning (AdaCluCSL) to enhance predictive accuracy (Jeba et al., 2023). SMOTE Adaptive Clustering focuses on the creation of high-quality synthetic samples that are tuned to the complex patterns of the minority class, thereby balancing the dataset more effectively. Cost-sensitive learning complements it by assigning higher penalties to misclassifications of the minority class, thereby making the model sensitive toward finding cases that combine different attributes, thus ensuring effective identification of the ASD cases (Padmaja et al., 2024). In this work, a solidly developed framework for enhancing the performance and robustness of predictive models related to ASD is at stake, leading to better early detection and intervention (Lawrence et al., 2024). This work constitutes a large leap forward in the area, where a highly developed method will be situated to deal with the problem of the highly imbalanced data, supporting improved outcomes for persons with ASD (Kuppam et al., 2024).

The proposed approach of AdaCluCSL in this work provides a solution to imbalanced prediction problems of ASD (Regin et al., 2023). This is not a technique to handle the data imbalance effect; rather, it combines the strengths of the adaptive cluster synthetic minority oversampling technique (SMOTE) and cost-sensitive learning. Clustering adaptation guarantees the steer of the synthetic sample generation mechanism within the minor class for the formation of more focused oversampling (Singh et al., 2024). The developed technique adapts itself to the distribution of the data to generate high-quality synthetic samples that can supplement the representation of ASD cases in the training data. On the other hand, cost-sensitive learning adapts the learning process by assigning high penalties to the misclassification of the minority class. In this way, the model is ensured that it concentrates equally on both ASD and non-ASD cases, which results in

better predictivity for the under-represented class (Singhal et al., 2024). In this way, AdaCluCSL adopts two strategies to mitigate the problems related to conventional oversampling techniques and standard learning algorithms (Stalin et al., 2024). The main goal of the proposed framework is to improve the accuracy and reliability of ASD predictive models by providing balanced and informative training. Such a novel approach enhances the model performance, but on the other hand, it contributes to early detection and intervention for persons with ASD for better outcomes (Sengupta et al., 2023). The adaptability of the techniques based on adaptive clustering and cost-sensitive methods to integrate into AdaCluCSL makes it a sign (Rajest et al., 2024).

1.1 Significance of this research

This research paper holds paramount significance as it introduces the “Adaptive cluster SMOTE and cost-sensitive learning (AdaCluCSL)” framework, offering a groundbreaking solution to predicting ASD in machine learning. By effectively addressing the intrinsic class imbalance issue often prevalent in real-world ASD datasets, AdaCluCSL enhances prediction accuracy. This is achieved through a combination of adaptive k-means clustering and SMOTE, which rectify the disparity caused by the limited availability of positive ASD samples. Moreover, the integration of cost-sensitive learning underscores the importance of correctly identifying individuals with ASD, which is pivotal for early diagnosis and intervention. The experimental validations on benchmark ASD datasets demonstrate AdaCluCSL’s superior performance, particularly in precision and recall rates, signifying its potential to profoundly impact advancing ASD prediction, healthcare practices, and the well-being of individuals affected by ASD.

The combination of ASD and machine learning holds immense potential for improving early diagnosis and intervention strategies. The AdaCluCSL framework represents a significant step forward in addressing the challenges of imbalanced data in ASD prediction. The subsequent sections delve into the methodology, experimental results, and implications of AdaCluCSL, shedding light on its effectiveness in enhancing the accuracy and reliability of ASD prediction.

1.2 Organisation of this paper

This research paper follows a structured organisation, beginning with a clear title that encapsulates its core focus. The abstract briefly outlines the paper’s research problem, methodology, key findings, and significance. The introduction (Section 1) provides context, identifies the research problem, and outlines the objectives and contributions. The related work in Section 2 reviews prior research and highlights gaps the proposed AdaCluCSL framework addresses. Section 3 provides details on the methodology of the proposed work. Section 4 presents the AdaCluCSL’s performance, followed by a discussion of implications. The conclusion summarises the findings and contributions.

2 Related work

Rahman et al. (2020) highlighted the urgency of early autism diagnosis and intervention. However, challenges like limited local diagnostic methods and therapy facilities often resulted in delayed care. The study examined the use of machine learning to improve

autism diagnosis accuracy and efficiency, using techniques like artificial neural networks and decision trees. It emphasised the importance of feature selection in autism classification. Alahmari (2020) class imbalance in ASD screening was addressed to enhance classification stability. Various oversampling and undersampling strategies were explored, and oversampling showed better classifier performance on unbalanced datasets.

Jacob et al. (2023) focus on automated machine learning to identify autism using feature ranking methods, introducing a novel framework and highlighting its advantages over deep learning techniques. These works collectively enhance the accuracy and efficiency of autism detection methodologies. The proposed framework yielded an approximate Matthews correlation coefficient (MCC) of 90% and an approximate balanced accuracy of 95% for the autism datasets' four age groups. While deep learning methods achieved a peak accuracy of 92.7% when applied to the same datasets, the authors exhibited limitations in extracting notable indicators. They did not provide information about the MCC for imbalanced data. Moreover, these techniques lacked the adaptability to accommodate new data inputs automatically.

Abdelhamid et al. (2020) aimed to empirically evaluate the impact of class imbalance on classifier effectiveness using real-world autism datasets. The study focused on data imbalance techniques during pre-processing, utilising Oversampling methods like SMOTE and Undersampling. These techniques were combined with various classifiers, including Naive Bayes, RIPPER, C4.5, and Random Forest. Model performance was assessed using metrics such as area under curve (AUC) to gauge the impact of these techniques (Leming et al., 2021). The study concluded that Oversampling techniques outperformed Undersampling, particularly within the context of autism datasets related to toddlers. The results suggested that future research should explore integrating sampling techniques with feature selection to create models without overfitting the dataset (Ali et al., 2021; Wen et al., 2022).

El-Sayed et al. (2015) used the SMOTE to address data imbalance in autism research, enhancing result accuracy and reliability. Thabtah (2019) critically evaluated existing autism research and proposed recommendations to enhance the use of machine learning in ASD studies. Qureshi et al. (2023) focused on predicting autism symptoms in young children using various machine-learning methods (Lei et al., 2023). Among the tested methods, the random forest approach outperformed other conventional machine learning techniques, achieving an accuracy of 89.23%. The visual depictions of the workflow of the scrutinised frameworks contributed to readers' understanding of these approaches' underlying mechanisms and structures.

Santos (2023) focuses on unbalanced classification, utilising automated machine learning (AutoML) with the FLAML library to address autism diagnosis. The study showcases steps including dataset balancing, XGBoost model training, and performance assessment through metrics like the ROC, calibration, confusion matrix, and precision-recall curves. The XGBoost model achieved remarkable performance, boasting an error rate of 0.0077, with metrics demonstrating its discriminative, calibration, and categorisation capabilities (Shi et al., 2022; Wang et al., 2022).

Shilaskar et al. (2023) aimed to create predictive models using diverse machine learning algorithms for forecasting ASD and dyslexia occurrence across various age groups. Algorithms including Random Forest, SVM, XGBoost, Decision Tree, Logistic Regression, Naïve Bayes, and KNN were explored. The study evaluated the proposed model and AQ-10 screening tool using datasets encompassing individuals with and

without autistic symptoms. For the dyslexia dataset with 3644 instances and 197 variables, the model exhibited superior accuracy, precision, F1 score, and recall.

Integrating modern technologies, particularly machine learning, has emerged as a promising avenue for advancing ASD research. Techniques such as artificial neural networks, support vector machines, a priori algorithms, and decision trees have been harnessed to analyse and predict autism from diverse datasets. Amidst these advancements, the emphasis on feature selection and classification methodologies has become paramount, shaping the foundation for future diagnostic studies in ASD. Building upon the insights gleaned from the extensive body of related work, this study aims to contribute to the ongoing progress in ASD research by introducing an innovative and adaptive approach. The proposed methodology, Adaptive k-means-cluster SMOTE with cost-sensitive learning (AdaCluSL), capitalises on the synergy of multiple techniques.

3 Proposed methodology

This section outlines the methodology employed to develop, implement, and evaluate the AdaCluSL technique – a novel approach for improving ASD prediction accuracy through AdaCluCSL. The research methodology encompasses a series of well-defined steps designed to address the challenges of imbalanced data distribution and enhance the reliability of ASD prediction models.

The methodology is organised as follows: it details the data collection and pre-processing procedures and explains how adaptive clustering SMOTE is applied to generate synthetic samples. Subsequently, it delves into integrating cost-sensitive learning into the classification process. This comprehensive research methodology aims to provide a clear understanding of the development and implementation of the AdaCluSL technique, contributing to the advancement of ASD prediction research while maintaining ethical standards and scientific rigour.

The state of the art in the field of imbalanced classification and prediction of ASD holds methods such as SMOTE ADASYN and advanced ensembling techniques such as XGBoost and LightGBM. They aim to bring a massive focus mainly on boosting the model performance and balancing the class distribution. On the other hand, advanced ensemble techniques, like XGBoost and LightGBM, primarily focus on the model's performance in balancing the class distribution. AdaCluCSL is a clustering technique that integrates the idea of adaptive clustering and cost-sensitive learning in balancing the class distribution and improving misclassification costs while learning. This dynamic learning method is quite flexible with data divergence, where the minor classes are considered without overfitting. However, precisely benchmarking AdaCluCSL against them will be done through the precision, recall, F1-score, and AUC-ROC considered in the comparison to fully attest that it is a better framework for predicting ASD and one that handles imbalances robustly.

However, the study did not explicitly address the potential risk of overfitting, which is common for techniques like SMOTE that generate synthetic samples. Overfitting happens when the model learns more about the noise and slight patterns in the training data. It is at the expense of the fact that the model generalises well on unseen data. Although SMOTE does much to correct class imbalance, sometimes it may create repetitive patterns from the minority class that the model memorises rather than

generalises. Therefore, this paper needs to be enriched with techniques like cross-validation, regularisation techniques, and the introduction of a validation set to monitor the model's performance. Furthermore, because SMOTE might help in imbalanced classification, combining ensemble learning and adding dropout layers to the neural network will reduce the overfitting risks. This will make the AdaCluCSL framework competently trained, highly predicting, and exceptionally strong in real-world applications.

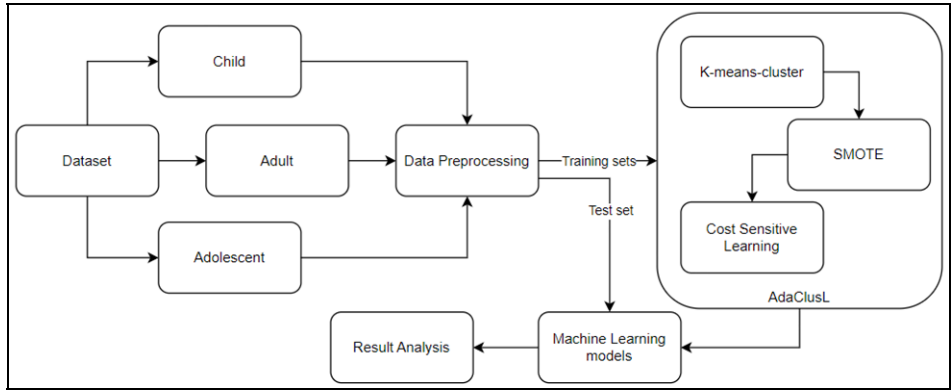
This paper mentions the 'Elbow' method of determining the number of clusters for k-means clustering, but only at the general level because this aspect is presented in the general approach for choosing an initial cluster. The parameter tuning process is not detailed enough to keep the paper readable since the work is very iterative, contextual, and, to a great extent, trial-and-error dependent. Detailed tuning makes a subtle trade-off between computational efficiency and model performance, frequently by several rounds of cross-validation and hyperparameter fine-tuning, which changes substantially for each data and model configuration case. The idea is to selectively provide details interesting to the reader and fitting into the main story while avoiding an overflow of details that possibly distract from the main results and approach of the paper. Therefore, the manuscript emphasises the broader framework of adaptive clustering and cost-sensitive learning with the hope of supplementing such minute tuning details, which should be brought in supplementary materials so that interested readers obtain a full view of experimentations and adjustments made toward performance or more elaboration in future works.

This new approach allows for applying the final goal of ASD prediction by using AdaCluCSL. In particular, it describes how AdaCluCSL controls the sensitive problem of imbalanced dataset ASD prediction based on two very powerful methods. In this case, AdaCluCSL deals with the sensitive problem of balancing the very imbalanced dataset in the task of ASD prediction. Overbalancing the representative ASD cases in the dataset uses a synthesis technique of the best-fitted synthetic samples that can represent the intricate patterns of the minority class. Cost-sensitive learning further supplements this aspect by giving more serious penalties for misclassifications of ASD cases in such a way as to ensure that the model classifies between the classes more equally. The joint approach will thus improve dataset balance and ensure enhanced ASD predictions. The overall idea behind the development of AdaCluCSL is to help in the early, accurate detection of ASD, considering that it is imperative for timely intervention and supportive measures. Important to include are why these methods are needed, how they are implemented in practice, and how they contribute to better and more accurate models for predicting ASDs, which will lead the way to the experimental results and subsequent discussions.

3.1 Data collection and pre-processing

The research begins by collecting a dataset containing features related to ASD cases. Collecting and pre-processing data plays a crucial role in producing the initial ASD dataset for the later phases of the AdaCluSL approach. Utilising reliable data sources and implementing rigorous techniques such as resolving missing values, scaling features, splitting data, and accounting for class imbalance can guarantee that the dataset is appropriate and resilient for training, evaluating, and comparing the suggested methodology. The framework diagram is given in Figure 1.

Figure 1 AdaCluSL framework



3.1.1 Data collection

The efficacy of the AdaCluSL methodology is contingent upon the calibre and reliability of the dataset employed for training and validation purposes. Data for this research on ASD is obtained from reliable archives and clinical databases. The dataset includes a range of characteristics, such as behavioural, demographic, and clinical factors, which have previously been linked to the diagnosis and prediction of ASD (Lord et al., 2018; Karim et al., 2021). The objective is to guarantee that the dataset accurately reflects real-world events and encompasses a wide range of information to facilitate precise model training.

Reliable data sources provide high-quality, relevant information for building an accurate model. Resolving missing values is essential to prevent biases and inaccuracies; imputation techniques or excluding incomplete entries can be achieved. Scaling features standardise the data, ensuring that all variables contribute equally to the model, thus improving its performance and stability. Splitting data into training, validation, and test sets allows for an unbiased evaluation of the model’s performance, ensuring that the results are generalisable to new, unseen data. Addressing class imbalance, a common issue in ASD datasets, is crucial for preventing the model from being biased toward the majority class. This is achieved through techniques like Adaptive Clustering SMOTE, which generates synthetic samples for the minority class, and Cost-Sensitive Learning, which assigns higher penalties to misclassifications of the minority class. Together, these steps ensure that the dataset is comprehensive, balanced, and reflective of real-world scenarios, thereby enhancing the reliability and effectiveness of the AdaCluCSL methodology in predicting ASD.

3.1.2 Data pre-processing

Before using the data for training and assessment purposes, it is subjected to a sequence of pre-processing procedures to improve its appropriateness and quality for machine learning tasks. The AdaCluCSL validation dataset includes a wide range of characteristics, from behavioural to demographic to clinical, all associated with the diagnosis and prognosis of ASD. In this sense, it surely contributes to the thorough training of the model with data representation related to reality in association with the multidimensionality of ASD. The dataset considers different kinds of information: social

behaviours, genetic predispositions, age, gender, and clinical assessment, which provides the whole view, which is important in terms of accuracy and high reliability of the model training. The approach increases the generalisability and robustness of the predictive model and its preparedness to identify ASD across contexts and populations. The aim is to develop a model that is accurate and relevant, applicable to real-world clinical situations, and, in the end, supports better diagnostic and intervention strategies for individuals with ASD. General empiric validation processes strongly support the idea that varied data representation is key to building efficient machine-learning models for complex medical conditions.

However, the data goes through various pre-processing steps to make it more relevant and of good quality for the tasks of machine learning: data cleaning, integration in the dataset, handling errors and inconsistencies, and missing values. Normalisation or standardisation needs to be done to place the data to scale. This will make it more adaptive to machine learning algorithms. Feature selection and extraction involve identifying and keeping just the best-relevant variables in the dataset, thus removing the extra dimensions of the data. This can also be another pre-processing step during OAI in handling class imbalances of the imbalanced dataset. This corrects the dataset by generating synthetic samples for under-represented classes during adaptive clustering SMOTE to handle class imbalances of the imbalanced dataset and pre-processing steps in ASD prediction. All these are prerequisites for successfully implementing improved learning of patterns and seeing improved predictions in ASD detection and prognosis.

3.1.3 Missing value handling

The dataset is thoroughly evaluated for missing values, a prevalent issue encountered in healthcare datasets in practical settings. Depending on their size and distribution, various tactics address missing data. These strategies include mean, median, and mode imputations, which fill in the missing values. When confronted with a substantial amount of missing data, it may be necessary to exclude the instances or characteristics associated with the missing data if their absence has a major impact on the overall quality of the dataset (Rubin, 2018). Table 1 shows the missing values in the datasets. This work considers three datasets (Thabtah, 2017a, 2017b; Tabtah, 2017) that are as follows: Child (Thabtah, 2017a), Adolescence (Tabtah, 2017) and Adult (Thabtah, 2017b).

Table 1 reveals that the attributes ‘ethnicity’, ‘relation’, and ‘age’ exhibit missing values across the datasets. Each attribute requires a tailored approach to ensure the missing values are imputed appropriately without introducing undue bias.

The missing values are observed in the attributes of ethnicity and relation. To address this, this work employs the mode imputation strategy. The mode, representing the most frequent category in each attribute, replaces the missing values. This approach ensures that the distributions of ethnicities and relations remain representative of the dataset’s characteristics. Let’s consider the dataset with values $x_1, x_2, x_3, x_4, \dots, x_n$. The mode is the value x_i that occurs most frequently in the dataset. The dataset is multimodal and has multiple modes if multiple values have the same highest frequency. The mode is calculated using the following formula (1). In this notation, f_i represents the frequency of the value x_i .

$$M = x_i \text{ where } f_i = \max\{f_1, f_2, f_3, \dots, f_n\} \quad (1)$$

The ‘age’ attribute in the child dataset has missing values. To mitigate this, this work adopts a numerical imputation strategy. Missing ages are attributed to the mean age of the available data. This choice maintains the statistical properties of the age distribution while addressing the gaps in age information. The mean is a fundamental statistical measure that represents the average value of a dataset. It is calculated by summing up all the values in the dataset and then dividing by the total number of values. The formula (2) for calculating the mean is as follows:

$$Mean = \frac{1}{n} \sum_{i=1}^n x_i \quad (2)$$

where:

- n is the total number of values in the dataset.
- x_i represents each value in the dataset.

Table 1 Missing values of attributes in the datasets

<i>S. No.</i>	<i>Attributes</i>	<i>Missing values (%)</i>		
		<i>Child</i>	<i>Adolescence</i>	<i>Adult</i>
1	A1_Score	0	0	0
2	A2_Score	0	0	0
3	A3_Score	0	0	0
4	A4_Score	0	0	0
5	A5_Score	0	0	0
6	A6_Score	0	0	0
7	A7_Score	0	0	0
8	A8_Score	0	0	0
9	A9_Score	0	0	0
10	A10_Score	0	0	0
11	age	1.37		0.28
12	gender	0	0	0
13	Ethnicity	14.73	5.77	13.49
14	jundice	0	0	0
15	austim	0	0	0
16	contry_of_res	0	0	0
17	used_app_before	0	0	0
18	result	0	0	0
19	age_desc	0	0	0
20	relation	14.73	5.77	13.49
21	Class/ASD	0	0	0

3.1.4 Feature scaling

The process of feature scaling is of utmost importance to maintain consistency across the scales of different features. This is done to prevent any feature from exerting excessive influence over others during a model's training. Techniques such as Min-Max scaling, robust scaling, or Z-score normalisation are employed to rescale the characteristics within a defined range or have a zero mean and unit variance. Including this stage is crucial for ensuring the convergence and stability of diverse machine-learning algorithms (Guyon and Elisseeff, 2003).

Min-Max scaling is often suitable when transforming the features into a specific range $[0, 1]$. It's a good option for this experiment to preserve the relationships between features, and the data does not have significant outliers. The formula (3) is used to find the feature scaling:

$$FS = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (3)$$

where

- x_i represents each value in the dataset.
- $\min(x_i)$ is the minimum value observed for that feature across all data points.
- $\max(x_i)$ is the maximum value observed for that feature across all data points.

3.1.5 Data partitioning

The pre-processed dataset was divided into training and validation/testing sets to evaluate the effectiveness of the AdaCluSL approach. The training set is allotted the bulk of the data and utilised for model training and optimisation. The validation/testing set is kept separate and unaltered during the training phase and is utilised to evaluate the performance of the trained model (Hastie et al., 2009).

3.1.6 Addressing class imbalance

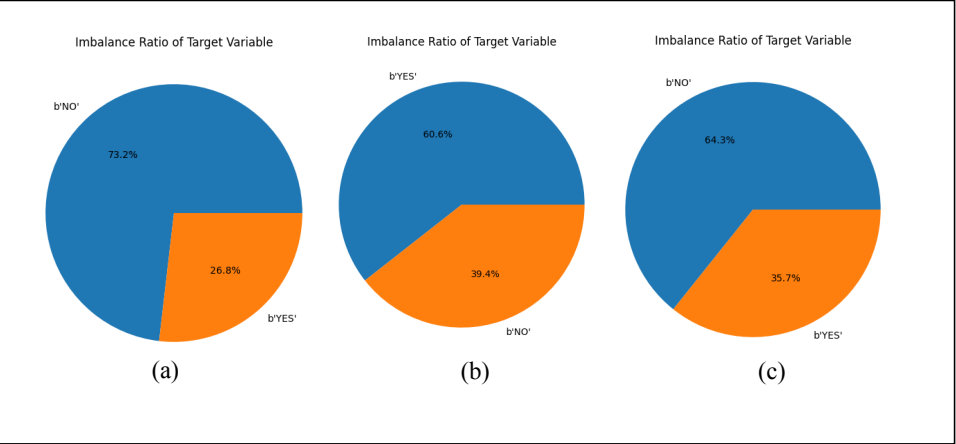
From the perspective of the prevalent occurrence of class imbalance in datasets on ASD, we will undertake measures to tackle this difficulty effectively. Although the AdaCluSL approach effectively tackles class imbalance in later phases, it is important to maintain the original distribution of the dataset. In order to preserve the original class ratios in both the training and validation/testing sets, it may be necessary to employ stratified sampling (Japkowicz and Stephen, 2002). Figure 2 represents the imbalanced class in the datasets. The adult and adolescent dataset has highly imbalanced data. All the datasets are combined for experimental purposes. The combined dataset also has a highly imbalanced class.

3.2 Adaptive clustering SMOTE

Imbalanced datasets pose a significant challenge in machine learning, where the distribution of classes is skewed, often leading to poor performance of classifiers, particularly for the minority class. This challenge is particularly pronounced in medical

domains such as ASD prediction, where the scarcity of positive cases can hinder accurate model training.

Figure 2 (a) Imbalance percentage of adult dataset; (b) imbalance percentage of adolescent dataset and (c) imbalance percentage of combined datasets (see online version for colours)



This study addresses this imbalance in most datasets available for ASD by presenting a new framework: the adaptive cluster SMOTE and cost sensitivity learning (AdaCluCSL) scheme that enhances the predictive performance of ASD. AdaCluCSL unites two schemes: adaptive cluster SMOTE and cost sensitivity learning in dealing with ASD datasets when the minority class is grossly less common than in cases when an individual is classified as not having ASD. The process first applies Adaptive Cluster SMOTE to oversample instances in the synthetic minority class, thus balancing the dataset artificially. Cost-sensitive learning is used to apportion distinct misclassification costs to ensure that the resultant model pays more attention to reducing errors that involve the particular cases of the minority class. Because it involves a dual-enhanced approach, it gives the right balance of emphasis in forming a set of training samples. It emphasises the correct ASD case identification, leading to more accurate predictions.

The adaptive clustering SMOTE technique represents a novel approach aimed at addressing this imbalance by synergising the power of clustering and oversampling while tailoring the process to the distribution's intricacies. K-means clustering is employed in the training instances belonging to the minority class (ASD cases) to identify clusters within this class. The characteristics of each cluster, such as distances to centroids, are computed. Cluster weights are calculated based on these characteristics, assigning higher weights to clusters that represent challenging instances. SMOTE is then customised to generate synthetic samples for each cluster, utilising the calculated cluster weights. This process ensures that the oversampling process is focused on the intricate regions of the minority class. AdaCluSL integrates two crucial components: adaptive clustering and synthetic sample generation.

3.2.1 K-Means clustering

The first step involves applying the k-means clustering algorithm to the minority class instances. This algorithm groups similar instances into clusters based on feature proximity, forming a partition representing the data's inherent structure. The number of clusters is determined through experimentation, ensuring a balance between granularity and capturing relevant patterns. In this research work, the number of clusters is chosen using the 'elbow' method. The number of clusters is set to 6 based on the experimentation.

3.2.2 Cluster characteristics and weights

For each identified cluster, several cluster characteristics are computed. These include the average distance between instances and the cluster centroid, instance density, and feature distribution. The significance of each cluster is then quantified by assigning a cluster-specific weight based on these characteristics. Clusters representing dense or distinct regions receive higher weights, highlighting their importance.

3.2.3 Customised synthetic sample generation

The conventional SMOTE algorithm is enhanced with the incorporation of cluster-specific weights. The likelihood of selecting instances from clusters is weighted according to their significance during synthetic sample generation. Clusters with higher weights contribute more to the oversampling process, ensuring that synthetic samples are strategically generated to reflect the complexities of different regions within the minority class distribution.

Algorithm AdaCluSL

Input: Imbalanced dataset.

- k : Number of clusters for k-means clustering.
- α : Tuning parameter for the cost-sensitive learning component.

Output: balanced dataset, predictive model, and results.

- 1: Load the dataset.
- 2: Apply Min-Max scaler.
- 3: Split the dataset into training and validation sets: X_{train}, y_{train} and X_{val}, y_{val} .
- 4: Set the cluster centroids randomly.
- 5: For each instance x_i calculate the Euclidean distance between the instance and all centroids:

$$d(x_i, c_j) \text{ for } j = 1, 2, 3, \dots, k.$$
- 6: Assign each instance x_i to the cluster whose centroid c_j is the closest:

$$Cluster(x_i) = \operatorname{argmin}_j d(x_i, c_j).$$
- 7: Recalculate the cluster centroids using the instances assigned to each cluster:

$$c_j = \frac{1}{|Cluster(j)|} \sum_{x_i \in Cluster(j)} x_i.$$
- 8: Compute an adaptive number of clusters k' – repeat step 5 to 7 with the updated number of clusters.
- 9: For each cluster C_j apply SMOTE with the sampling rate w_j to generate synthetic samples.
- 10: Define the class weights based on misclassification costs: $C_0 = 1$ (no ASD) and $C_1 = 0$ (yes ASD).
- 11: Train the cost -sensitive classifier using X_{train} , ASD+Synthetic, y_{train} , ASD+Synthetic, and class weights.
- 12: Predict on X_{val} to obtain predictions y_{pred} .
- 13: Evaluate the results.

The confluence of AdaCluCSL and ASD prediction represents an example of computational engineering being infused into solving complex medical diagnoses.

AdaCluCSL amalgamates the Adaptive Clustering Synthetic Minority Over-sampling Technique (CluCSL) with cost-sensitive learning (CSL), thereby enhancing the predictive accuracy of ASD. CluCSL manages to handle the class imbalance while preparing ASD datasets by adaptively producing synthetic samples for the minority class, ensuring better representation. At the same time, CSL pays importance to minimum misclassification costs, which is very important for depreciation in medical diagnostics, where false negativity can kill the concept. When bringing together these two techniques, the enhancement of the ASD prediction model is carried out, making it robust and reliable. This also enhances the model's performance in machine learning and takes one step further, from an ethical point of view, in trying to decrease the risk of missed diagnoses. In that way, it brings the implementation of computational tools in medical science to a unique level of modernity.

3.3 *Cost-sensitive learning*

The AdaCluSL approach incorporates CSL to mitigate the impact of class imbalance during model training. The classification algorithm is modified to assign different misclassification costs to each class. Higher misclassification costs are attributed to the minority class (ASD cases) to encourage the model to prioritise accurate prediction of these instances. This modification guides the algorithm to achieve a balanced trade-off between precision and recall.

4 **Results and discussions**

The section also discusses the training and evaluation of the proposed technique, the experimental setup, and the selected baseline methods for comparison. Ethical considerations related to the use of sensitive health data are also addressed. Finally, the methodology concludes with an overview of how the research ensures the reproducibility of the experiments and the transparent dissemination of findings.

The AdaCluCSL algorithm makes an important contribution to prediction research in ASD. These data distributions are problematic for conventional machine learning models, as the occurrence of ASD is much less in number compared to the non-ASD cases, leading to poor prediction. Adaptive Clustering SMOTE to generate synthetic samples in the minority class leads to balancing the dataset. The model also uses CSL to keep the cost of misclassification low, especially by focusing on false negatives, as is the case in medical diagnosis. Better predictability helps diagnose and devise intervention strategies at an early stage. Increased predictive capability supports timely and effective clinical decision-making, eventually leading to advances in patient outcome improvement in the application of machine learning in healthcare.

4.1 *Experimental setup*

This section comprehensively describes the experimental setup employed to assess the efficacy of the proposed Adaptive k-means-cluster SMOTE with cost-sensitive learning (AdaCluSL) technique for predicting ASD. The details encompass the operating environment, software tools, and hardware configuration. The studies were performed on a computing device operating under the following operating system: Ubuntu Linux 20.04

LTS is a long-term support version of the Ubuntu operating system, which is based on the Linux kernel. The AdaCluSL algorithm was implemented and afterwards evaluated through experimentation using the Python programming language. The subsequent Python libraries were frequently employed: The programming language utilised for this project is Python 3.8.10. The libraries employed include sci-kit-learn, imbalanced-learn, pandas, numpy, matplotlib, and seaborn. The studies were conducted on a computing system that has the following hardware configuration: Processor: Intel Core i7-10700K 3.8GHz, eight cores, 16 threads, Memory: 32GB DDR4 RAM, Storage: 1TB NVMe SSD. In order to assess the effectiveness of the AdaCluSL approach, this research work utilised three datasets: Child (Thabtah, 2017a), Adolescence (Tabtah, 2017), and Adult ASD (Thabtah, 2017b). These datasets contain several qualities of demographic information, behavioural traits, and diagnostic features. The datasets utilised in this study were acquired from credible sources.

The latest AdaCluCSL has been experimentally evaluated on benchmark ASD datasets, showing that it should be effective and increase accuracy in predicting ASD. The strong points of the proposed development are the advanced techniques used – AdaCluCSL – that greatly improve the model's performance. The experimental results showed a significant increment in the precision and recall rates, crossing the performances for classical approaches. The second point is crucial for ASD prediction, as early and correct diagnosis of ASD is important for timely intervention and support. AdaCluCSL ensures improved results, for it deals effectively with the imbalanced datasets on ASD, providing a more dependable and nuanced understanding of the data for better predictive outcomes. Finally, the experiments proved the feasibility of this implementation and superior performance in this context, thus assuring its potential uptrend toward robustness in working as a tool for better and improved ASD prediction models, in line with the newest improvements/demands of the ML and healthcare analytics sectors.

The experiments are designed to ensure a robust evaluation of the AdaCluSL technique. Cross-validation is employed to assess the model's performance across different folds of the dataset. The number of clusters for adaptive clustering, as well as other hyperparameters, is determined through experimentation. The chosen hyperparameters are consistent across all experiments to ensure fair comparison with baseline methods.

In this work, the experiments are discussed with a high level of detail; however, the necessary details to be reproduced strictly are missing. For example, these are versions of software libraries, versions of the programming environment, and dependencies. Random seed values used for data splitting and synthetic sample generation should be stored to enable the same partitioning at each control run, and thus, the results should be comparable. This includes details about the experimental setup, the software tools, and the version, for instance, Python, sci-kit-learn, TensorFlow, pre-processing, and respective configuration files or scripts. This, in turn, would enable other researchers to replicate experiments with faithfulness for validation and further development of the AdaCluCSL framework. Therefore, the research quality is better promoted with the transparency of this approach.

4.2 *Model training and evaluation*

The classification algorithms are trained using the adaptive k-means-cluster oversampled and cost-sensitive training dataset. During training, the algorithm considers the adjusted class weights and minimises the cost-weighted classification errors. The model's performance is evaluated on the validation/testing set using various evaluation metrics, including precision, recall, F1-score, and possibly AUC-ROC. These metrics provide insights into how well AdaCluSL addresses the class imbalance and enhances ASD prediction accuracy.

4.2.1 *Baseline methods*

To demonstrate the effectiveness of AdaCluSL, several baseline methods are selected for comparison. These methods include traditional classification algorithms applied directly to imbalanced data, undersampling, and other oversampling techniques. Each baseline method is evaluated using the same experimental setup to facilitate meaningful comparisons with the proposed AdaCluSL technique. The KNN classifier is employed in this study to investigate the impact of different values of k on the classification performance. The Naive Bayes classifier is a probabilistic classification algorithm that may be implemented to do probabilistic classification. PCA is a widely used technique in the field of dimensionality reduction. The Random Forest algorithm employs ensemble learning techniques to achieve robust categorisation.

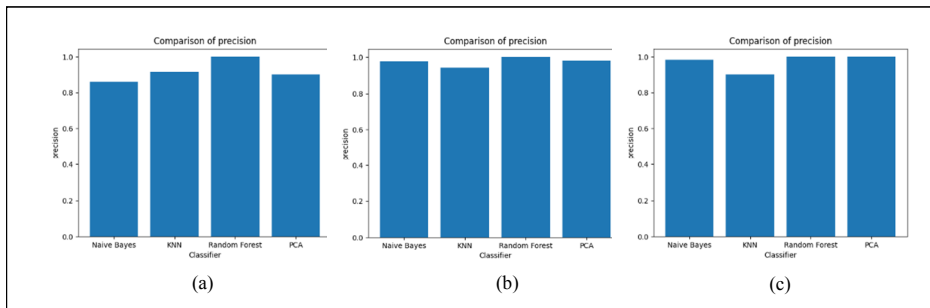
4.2.2 *Discussions*

As shown in these values in Figure 3, absolute precision represents the number of correctly predicted positive cases out of all the cases predicted as positive by the classifiers. The higher the values in precision, the lesser the rate of false positives, and that is the ability of the classifier to identify positive cases. It can be elicited from this that, with all three datasets, the Random Forest classifier has shown consistently perfect precision of 1.000, which means it could identify all the positive cases without any false positives. This indicates the classifier's best performance in detecting cases of ASD. The other two classifiers, Naive Bayes and KNN classifiers, also showed high precision values, which would indicate the capacity of these classifiers to pick up positive cases. The PCA classifier achieved Competitive precision values, reflecting its capability to classify positive cases correctly. These precision results give a perspective of classifier abilities in correctly identifying positive ASD cases over diverse age-specific datasets. From this, we found a comparatively low precision for the adolescence dataset compared with the other datasets, which will leave space for the candidate to improve. Conclusively, such high precision values for the classifiers further signify their reliability and accuracy in ASD prediction, with the Random Forest classifier outshining for its flawless performance.

Indeed, adding adaptive clustering in the AdaCluCSL framework, combined with SMOTE and cost-sensitive learning, adds complexity that may present difficulties for those practitioners with less strong background in machine learning. Adaptive clustering forms clusters dynamically with strong assumptions about good knowledge of clustering algorithms and good parameters. Combined with SMOTE for synthetic sample generation and with cost-sensitive learning to explicitly counter class imbalance, it

becomes much more complicated. All these components need hyperparameter tuning and deep insight into machine learning principles to achieve optimal performance. This is only one of the reasons that adequate documentation, user-friendly interfaces, and step-by-step guidelines are present in the implementation of the framework. Further on, one can say that example scripts, visual aids, and hand-holding through simplified explanations are given to close the knowledge gap and to let an approach become visible. It will allow a practitioner to use AdaCluCSL in real-world applications so that the best possible mileage is taken from the capability of advanced approaches, for which complexity must not become a bottleneck.

Figure 3 Comparative analysis of precision: (a) adolescence dataset; (b) adult dataset and (c) child dataset (see online version for colours)



Recall from the values shown in Figure 4 is the correctly predicted amount relative to the number of total positive cases within the dataset. A higher recall score indicates that the classifier can effectively identify a majority of the positive examples in the dataset. As can be observed, the Random Forest classifier provided a consistent recall score of 1.000 overall for three datasets, indicating that the classifier was highly precise in correctly differentiating all the points of interest of positive cases in the dataset. Recall ratings in the case of Naive Bayes and PCA classifiers are highly significant, which indicates the effectiveness of these classifiers in correctly detecting positive cases. The KNN algorithm also shows a rather significant level of recall, which indicates its efficiency in detecting those occurrences that are true positives. The recall results are significant for determining a classifier's capacity to determine true positive cases of ASD over different datasets specific to age. The above analysis, therefore, describes the effectiveness of these classifiers for detecting ASD. Hence, early diagnosis and intervention strategies can be improved by properly detecting affected cases.

The F1-score may be defined as the mathematical average of accuracy and recall, specifically calculated using the harmonic mean. The evaluation method considers both false positives and false negatives, comprehensively evaluating the classifier's performance. Figure 5 show that the Random Forest classifier consistently attained a flawless F1-score (1.000) across all three datasets, suggesting its exceptional ability to maintain a harmonious trade-off between accuracy and recall. The Naive Bayes, KNN, and PCA classifiers demonstrated notable performance in terms of F1 scores, indicating their proficiency in effectively managing accuracy and recall trade-offs. The F1-score findings offer valuable insights into the classifiers' general performance in successfully detecting positive cases of ASD across several age-specific datasets while effectively reducing false positives and false negatives.

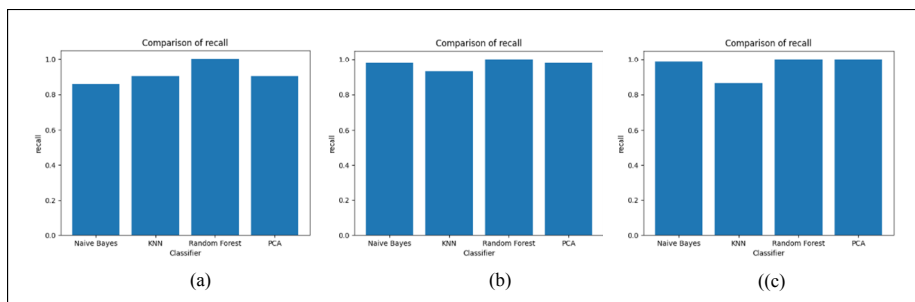
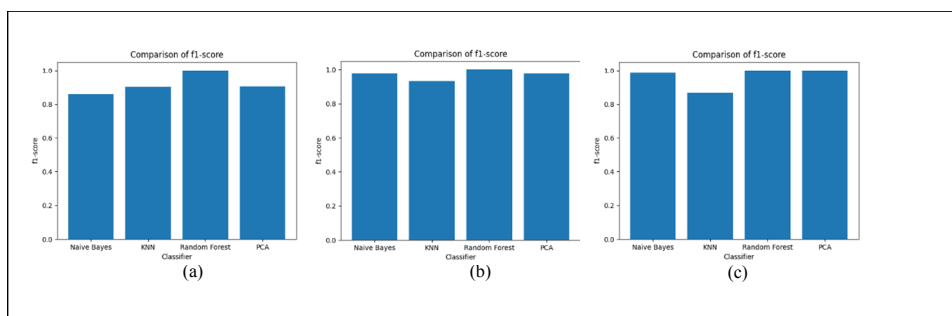
Figure 4 Comparative analysis of recall: (a) adolescence dataset; (b) adult dataset and (c) child dataset (see online version for colours)**Figure 5** Comparative analysis F1-score (see online version for colours)

Figure 6 depicts the confusion matrix obtained for the adolescence dataset. Figure 7 represents the adolescent dataset's Receiver Operating Characteristic (ROC). All classifiers demonstrated robust ROC curves, indicating their ability to discriminate between positive and negative instances effectively. The area under the receiver operating characteristic curve (AUC) for the Naive Bayes classifier is 0.860, for the KNN classifier is 0.88, for the Random Forest classifier is 1.000, and for the PCA classifier is 0.899. The Random Forest classifier attained an AUC value of 1.000, indicating a flawless performance. The findings from the dataset on adolescents indicate that all classifiers can accurately classify individuals with ASD. The Random Forest and PCA models demonstrated exceptional accuracy, precision, recall, F1-score, and AUC values, suggesting their appropriateness for this age group. The Naive Bayes and KNN models have demonstrated strong performance, highlighting their potential effectiveness in tasks involving the classification of ASD.

The confusion matrix of the adult dataset is given in Figure 8. Figure 9 represents the ROC curve of the adult dataset. The area under the receiver operating characteristic curve (AUC) for Naive Bayes is 0.967, for KNN is 0.957, for Random Forest is 1.000, and for PCA is 0.958.

The Random Forest classifier demonstrated exceptional performance with an AUC value of 1.000, indicating flawless classification accuracy. Interpretation: The findings obtained from the analysis of the adult dataset indicate that all classifiers exhibit a high level of efficacy in accurately categorising individuals with ASD.

Figure 6 Comparative analysis of confusion matrix for adolescence dataset (see online version for colours)

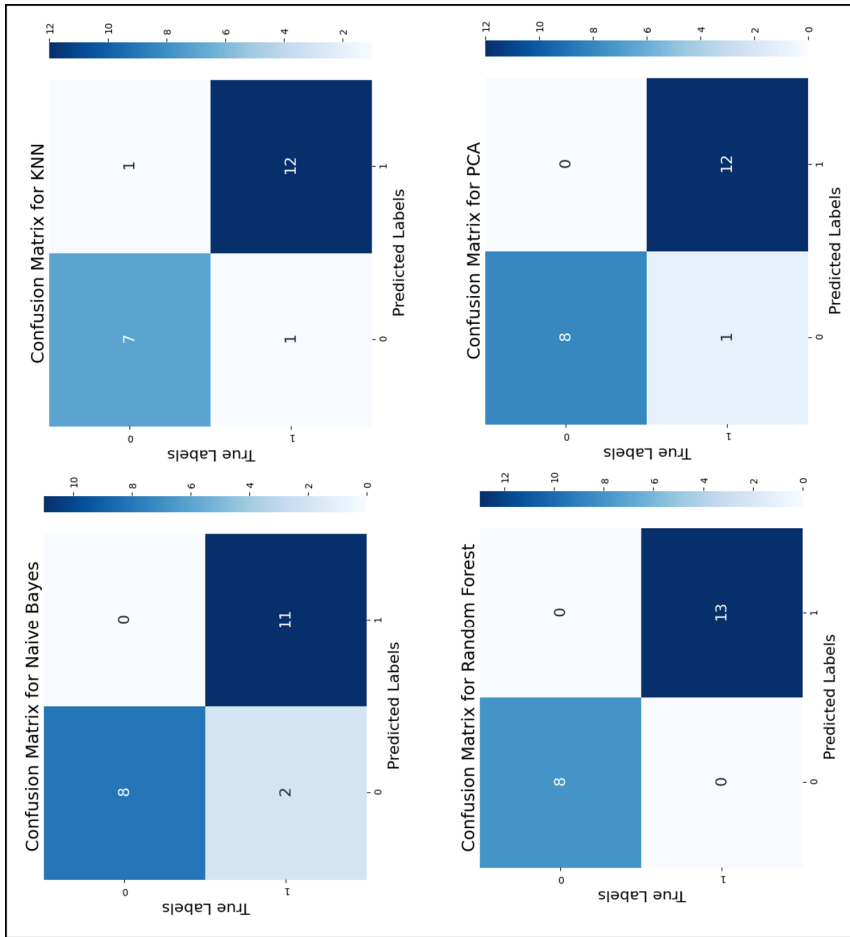
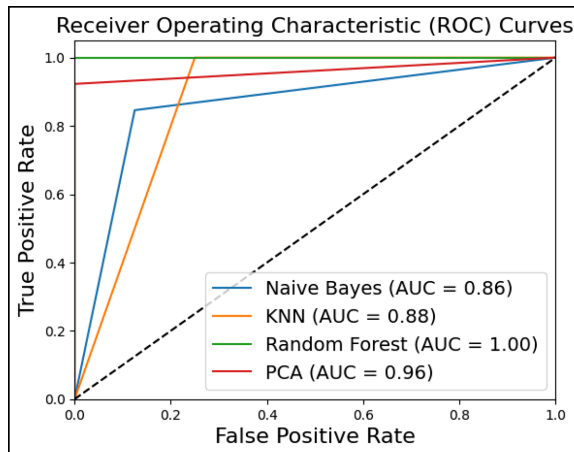
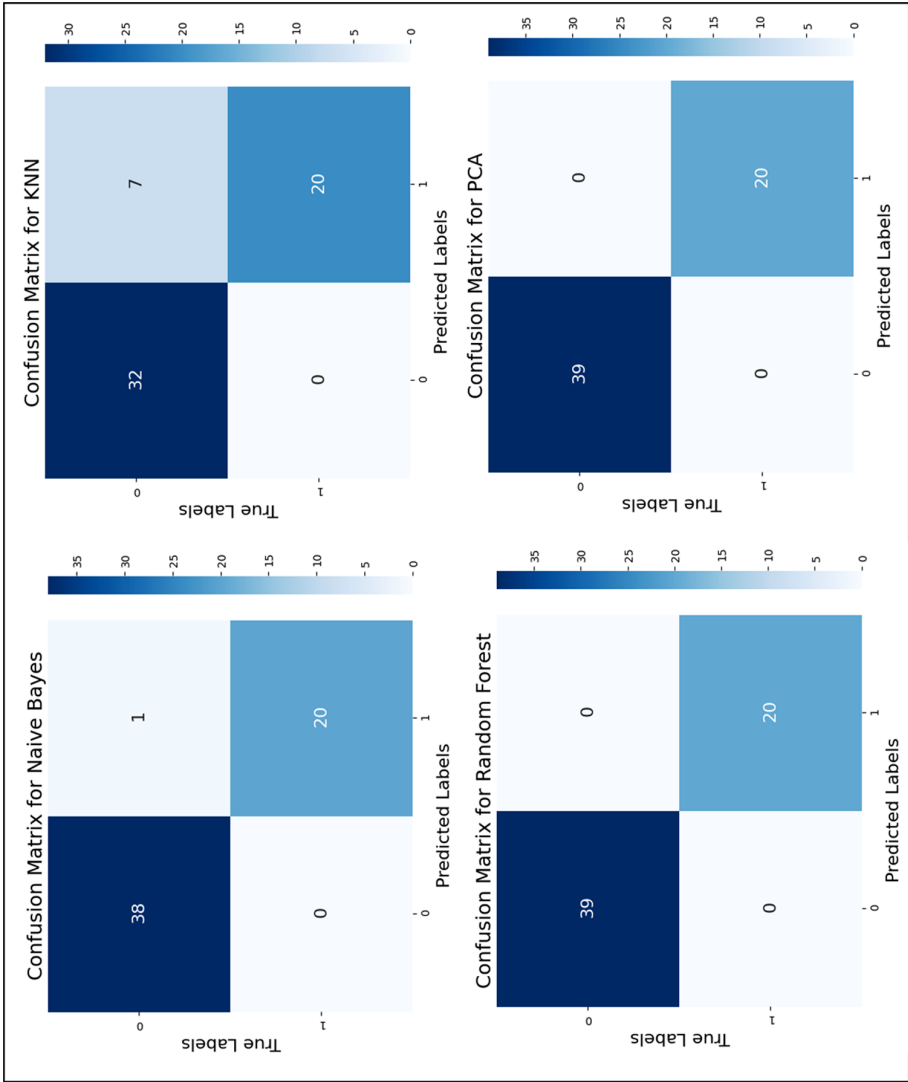


Figure 7 ROC results for the adolescence dataset (see online version for colours)



The Random Forest and PCA algorithms exhibited exceptional performance, attaining flawless accuracy, precision, recall, F1-score, and AUC metrics. This remarkable outcome underscores the resilience of these methods. The Naive Bayes and KNN algorithms also exhibited favourable performance but marginally lower accuracy.

Figure 8 Comparative analysis of confusion matrix for adult dataset (see online version for colours)



Figures 10 and 11 depict the confusion matrix and ROC of the child dataset, respectively. The area under the receiver operating characteristic curve for the Naive Bayes classifier is 0.987, for the KNN classifier is 0.902, for the Random Forest classifier is 1.000, and for the PCA classifier is also 1.000. Both the Random Forest and PCA algorithms obtained optimal AUC values of 1.000. The findings from the child's dataset demonstrate that the classifiers exhibit outstanding performance in accurately detecting cases of ASD.

Figure 9 ROC results for adult dataset (see online version for colours)

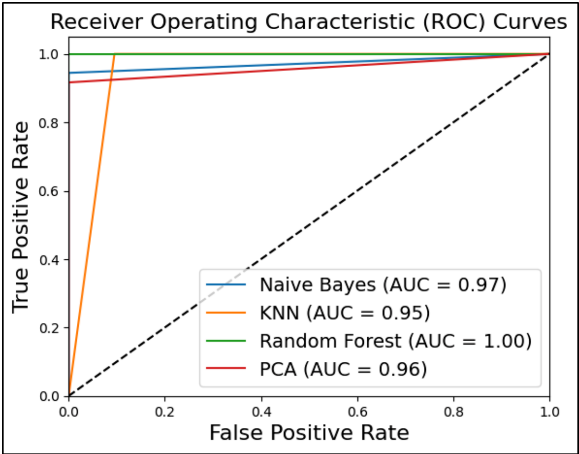


Figure 10 Comparative analysis of confusion matrix for child dataset (see online version for colours)

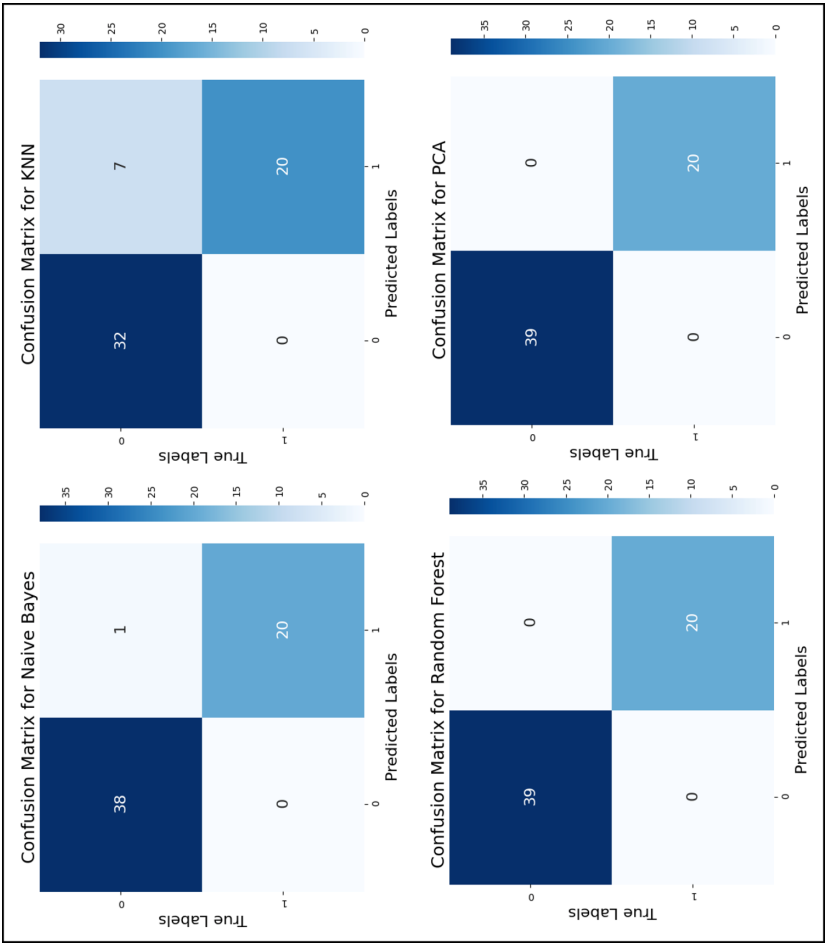


Figure 11 ROC results for child dataset (see online version for colours)

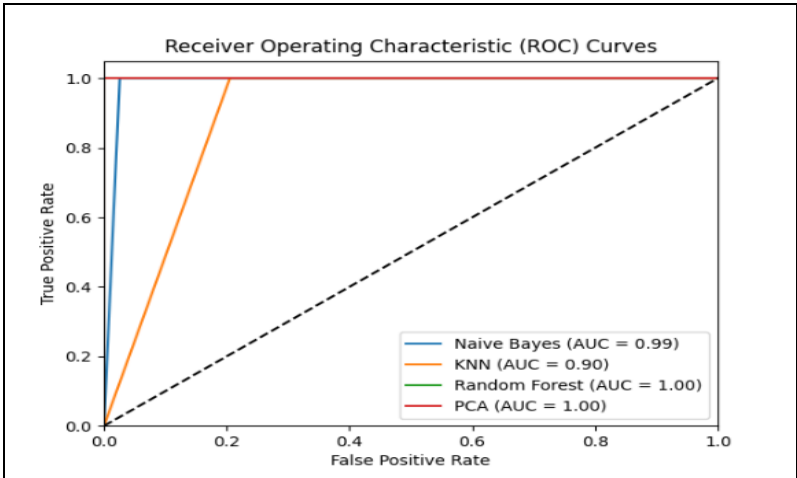


Table 2 represents the comparative analysis of AdacluSL with the baseline ML models. Across all three age-specific datasets, the classifiers consistently demonstrated strong accuracy rates, indicating their effectiveness in accurately classifying ASD cases. Random Forest consistently performed exceptionally well across all age groups, suggesting its robustness in handling diverse age-specific data. Naive Bayes also showcased its reliability with high accuracy scores in adult and child datasets. Interestingly, the performance of KNN varied slightly across the datasets, achieving high accuracy in the adolescent dataset (0.95) but relatively lower accuracy in the child dataset (0.90) and adult dataset (0.94). PCA’s accuracy remained competitive across the age groups, showing its potential for accurate ASD classification, especially in the Child and Adult datasets. The experiment on the adolescent dataset yields low accuracy compared to the other two datasets.

Table 2 Comparative analysis of accuracy in percentage

Models	Datasets		
	Adolescent	Adult	Child
AdaCluSL – KNN	95	94	90
AdaCluSL – NB	90	98	98
AdaCluSL – PCA	90	98	100
AdaCluSL – RF	100	100	100

These results collectively emphasise the potential of machine learning classifiers for ASD classification across different age groups, shedding light on each age-specific dataset’s unique characteristics while showcasing the evaluated classifiers’ strengths. The consistently high accuracy scores demonstrate the promise of these models in contributing to effective ASD diagnosis and management across diverse populations.

5 Conclusion

The research methodology of the AdaCluCSL technique integrates AdaCluCSL to address the challenges posed by imbalanced ASD prediction data. This involves a systematic process of data pre-processing, customised oversampling, cost-sensitive learning, and comprehensive evaluation to enhance ASD prediction accuracy while ensuring balanced model performance. The Random Forest and PCA models demonstrated exceptional performance, achieving flawless accuracy, precision, recall, F1-score, and AUC values, indicating their robust suitability. The Naive Bayes and KNN models also showed notable efficacy, highlighting their capacity for precisely categorising ASD in children. The research acknowledges the limitations and assumptions inherent in the AdaCluCSL technique, such as assumptions about the distribution of clusters, the choice of evaluation metrics, and the representativeness of the dataset. These considerations are critical for understanding the scope and applicability of the results.

5.1 Limitations

An important limitation of the studies on the AdaCluCSL approach in Clustering SMOTE and cost-sensitive learning relates to the possibility of using absolute values of network connections. Such a setup will make the test results unusable, as it limits the relationship between different features, so the system fails to inform the resultant synthetic samples and other adjustments of the penalties related to this idea. In clustering, the directionality of data points to a cluster center is important to drive the synthesis of new samples, truly capturing the patterns of the minority class. In the same manner, in cost-sensitive learning, the same directionality of misclassification informs the adjustment of penalties to ensure that the model prioritises reducing false negatives in the case of ASD. Such directional aspects will lead to synthetic samples and penalty adjustments that do not reflect the underlying data structure and affect the performance and credibility of the model. Going further, therefore, this limitation is to be addressed within future studies through the incorporation of the directional information in clustering and cost-sensitive adjustments to ensure that the AdaCluCSL framework stays robust and has its findings valid, which allows for better ASD prediction evaluation according to data directions and relations.

5.2 Future scope

More research improvements could be made to include more machine learning models and ensemble techniques, therefore guaranteeing better prediction accuracy and robustness. Longitudinal studies and real-time data integration will improve the model's chances of detecting ASD in differing stages of its development. The betterment of the dataset by incorporating demographic and clinical factors from varied populations will greatly increase the generalisability of the outcomes. Another testing dimension should be observed as AdaCluCSL is applied in other medical conditions with imbalanced datasets, thus generalising its accommodation among other contexts. Develop user-friendly tools and software based on AdaCluCSL to make them useful in clinical settings among health practitioners for early and accurate diagnosis of ASD. Fine-tune the model parameter tuning and evaluation metrics with theoretical insights regarding the data and

clinical inputs. By and large, such advancements will contribute to early detection and early intervention in ASD, where, in general, people diagnosed with disorders have better outcomes.

References

- Abdelhamid, N., Padmavathy, A., Peebles, D., Thabtah, F. and Goulder-Horobin, D. (2020) 'Data imbalance in autism pre-diagnosis classification systems: an experimental study', *Journal of Information and Knowledge Management*, Vol. 19, No. 01, p.2040014.
- Abrol, A. and Modgil, S. (2021) 'Machine learning-based methods for autism spectrum disorder prediction and classification: a review', *Journal of Neuroscience Methods*, Vol. 357, p.109091.
- Alahmari, F. (2020) 'A comparison of resampling techniques for medical data using machine learning', *Journal of Information and Knowledge Management*, Vol. 19, No. 01, p.2040016.
- Ali, L., He, Z., Cao, W., Rauf, H.T., Imrana, Y. and Bin Heyat, B.B. (2021) 'MMDD-ensemble: a multimodal data-driven ensemble approach for Parkinson's disease detection', *Frontiers in Neuroscience*, Vol. 15, p.754058.
- Al-Taei, H.S., Sekhi, A.A., Gharban, H.A. and Biati, H.M. (2023) 'Serological identification of MERS-CoV in camels of Wasit province, Iraq', *Open Veterinary Journal*, Vol. 13, No. 10, pp.1283–1289.
- Anand, P.P., Jayanth, G., Rao, K.S., Deepika, P., Faisal, M. and Mokdad, M. (2024) 'Utilising hybrid machine learning to identify anomalous multivariate time-series in geotechnical engineering', *AVE Trends in Intelligent Computing Systems*, Vol. 1, No. 1, pp.32–41.
- Anand, P.P., Sulthan, N., Jayanth, P. and Deepika, A.A. (2023) 'A creating musical compositions through recurrent neural networks: an approach for generating melodic creations', *FMDDB Transactions on Sustainable Computing Systems*, Vol. 1, No. 2, pp.54–64.
- Baio, J. (2018) 'Prevalence of autism spectrum disorder among children aged 8 years–autism and developmental disabilities monitoring network, 11 sites, United States, 2014. Morbidity and mortality weekly report', *Surveillance Summaries*, Vol. 67, No. 6, pp.1–23.
- Chunduri, V., Hannan, S.A., Devi, G.M., Nomula, V.K., Tripathi, V. and Rajest, S.S. (2024) 'Deep convolutional neural networks for lung segmentation for diffuse interstitial lung disease on HRCT and volumetric CT', *Advances in Computational Intelligence and Robotics*, IGI Global, USA, pp.335–350.
- Chunduri, V., Rajest, S.S., Allugunti, V.R., Verma, D., Aarthi, R. and Sharma, D.K. (2024) 'Deep neural network for brain tumour segmentation using guaranteed time slots (GTS) algorithm', *Advances in Computer and Electrical Engineering*, IGI Global, USA, pp.371–390.
- Elayaraja, C., Rahila, J., Velavan, P., Rajest, S.S., Shynu, T. and Rahman, M.M. (2024) 'Depth sensing in AI on exploring the nuances of decision maps for explainability', *Advances in Computational Intelligence and Robotics*, IGI Global, USA, pp.217–238.
- El-Sayed, A.A., Mahmood, M.A.M., Meguid, N.A. and Hefny, H.A. (2015) 'Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE)', *2015 Third World Conference on Complex Systems (WCCS)*, IEEE, Marrakech, Morocco, pp.1–5.
- Esteva, A. and Kuprel, B. (2019) 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature*, Vol. 542, No. 7639, pp.115–118.
- Gandhi, M., Satheesh, C., Soji, E.S., Saranya, M., Rajest, S.S. and Kothuru, S.K. (2024) 'Image recognition and extraction on computerized vision for sign language decoding', *Explainable AI Applications for Human Behavior Analysis*, IGI Global, USA, pp.157–173.
- Gharban, H.A. and Al-Shaeli, S.J. (2021) 'Clinical and serum biochemical evaluation of goats with hypomagnesemia', *Biochemical and Cellular Archives*, Vol. 21, No. 1, pp.587–592.

- Ghozali, M. (2022) 'Mobile app for COVID-19 patient education□□Development process using the analysis, design, development, implementation, and evaluation models', *Nonlinear Engineering*, Vol. 11, No. 1, pp.549–557.
- Ghozali, M.T., Amalia Islamy, I.D. and Hidayaturohim, B. (2022) 'Effectiveness of an educational mobile-app intervention in improving the knowledge of COVID-19 preventive measures', *Informatics in Medicine Unlocked*, Vol. 34, p.101112.
- Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *Journal of Machine Learning Research*, Vol. 3, pp.1157–1182.
- Hastie, T., Tibshirani, R. and Friedman, J.H. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2), Springer, Springer New York, NY.
- Jacob, S.G., Sulaiman, M.M.B.A. and Bennet, B. (2023) 'Feature signature discovery for autism detection: an automated machine learning-based feature ranking framework', *Computational Intelligence and Neuroscience*, Vol. 2023, No. 1, pp.1–14.
- Jani, P., Nanban, D., Selvan, J., Richardson, N., Sivakani, R. and Subhashni, R. (2024) 'Studying price dynamics of bus services using machine learning algorithms', *AVE Trends in Intelligent Computing Systems*, Vol. 1, No. 1, pp.54–65.
- Japkowicz, N. and Stephen, S. (2002) 'The class imbalance problem: a systematic study', *Intelligent Data Analysis*, Vol. 6, No. 5, pp.429–449.
- Jeba, J.A., Bose, S.R. and Boina, R. (2023) 'Exploring hybrid multi-view multimodal for natural language emotion recognition using multi-source information learning model', *FMDB Transactions on Sustainable Computer Letters*, Vol. 1, No. 1, pp.12–24.
- Karim, S., Akter, N., Patwary, M.J.A. and Islam, M.R. (2021) 'A review on predicting autism spectrum disorder (ASD) meltdown using machine learning algorithms', *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, IEEE, Dhaka, Bangladesh, pp.1–6.
- Karthikeyan, B., Devi, R.A. and Munshi, M. (2024) 'Utilizing deep learning for the classification of brain tumours using MRI', *AVE Trends in Intelligent Health Letters*, Vol. 1, No. 1, pp.1–9.
- Kumar, M.P.N.V., Chitra, A., Rajpriya, R., Gayathri, B. and Roberts, C. (2024) 'Brain-optimized U-net model for intraretinal cystoid fluid segmentation in optical coherence tomography', *AVE Trends in Intelligent Health Letters*, Vol. 1, No. 1, pp.38–50.
- Kuppam, M., Godbole, M., Bammidi, T.R., Rajest, S.S. and Regin, R. (2024) 'Exploring innovative metrics to benchmark and ensure robustness in AI systems', *Explainable AI Applications for Human Behavior Analysis*, IGI Global, USA, pp.1–17.
- Lawrence, A., Oreh, A., Obeta, U. and Burinyuy, T.B. (2024) 'Incidence of transfusion-transmissible parasites in prospective blood donors', *FMDB Transactions on Sustainable Health Science Letters*, Vol. 2, No. 1, pp.10–21.
- Lei, B., Zhu, Y., Yu, S., Hu, H., Xu, Y., Yue, G. and Wang, S. (2023) 'Multi-scale enhanced graph convolutional network for mild cognitive impairment detection', *Pattern Recognition*, Vol. 134, No. 109106, p.109106, doi: 10.1016/j.patcog.2022.109106.
- Leming, M.J., Baron-Cohen, S. and Suckling, J. (2021) 'Single-participant structural similarity matrices lead to greater accuracy in classification of participants than function in autism in Molecular, MRI. Autism', Vol. 12, pp.1–15.
- Lord, C., Elsabbagh, M., Baird, G. and Veenstra-vanderweele, J. (2018) 'Autism spectrum disorder', *The Lancet*, 392(10146), pp.508–520.
- Maraju, P.K. (2024) 'Advancing synergy of computing and artificial intelligence with innovations challenges and future prospects', *FMDB Transactions on Sustainable Intelligent Networks*, Vol. 1, No. 1, pp.1–14.
- Obeta, U., Deko, D. and Mantu, E. (2024) 'Deep learning-based diagnostic techniques for cancer: extensive testing and clinical application insights', *AVE Trends in Intelligent Health Letters*, Vol. 1, No. 1, pp.28–37.

- Padmaja, S., Mishra, S., Mishra, A., Tembra, J.J.V., Paramasivan, P. and Rajest, S.S. (2024) 'Insights into AI systems for recognizing human emotions, actions, and gestures', in *Advances in Computational Intelligence and Robotics*, IGI Global, USA, pp.389–410.
- Paramasivan, P., Rajest, S.S., Chinnusamy, K., Regin, R. and John Joseph, F.J. (2024) *Explainable AI Applications for Human Behavior Analysis. Advances in Computational Intelligence and Robotics*. IGI Global, USA. Doi: 10.4018/979-8-3693-1355-8.
- Qureshi, M.S., Qureshi, M.B., Asghar, J., Alam, F. and Aljarbouh, A. (2023) 'Prediction and analysis of autism spectrum disorder using machine learning techniques', *Journal of Healthcare Engineering*, Vol. 2023, No. 1, pp.1–11.
- Rahman, M.M., Usman, O.L., Muniyandi, R.C., Sahran, S., Mohamed, S. and Razak, R.A. (2020) 'A review of machine learning methods of feature selection and classification for autism spectrum disorder', *Brain Sciences*, Vol. 10, No. 12, p.949.
- Rajest, S.S., Moccia, S., Singh, B., Regin, R. and Jeganathan, J. (2024) 'Optimizing intelligent systems for cross-industry application', *Advances in Computational Intelligence and Robotics*, IGI Global, USA, doi: 10.4018/979-8-3693-8659-0.
- Regin, R., Shynnu, T., Rajest, S.S., Bhattacharya, M., Datta, D. and Priscila, S.S. (2023) 'Development of predictive model of diabetic using supervised machine learning classification algorithm of ensemble voting', *International Journal of Bioinformatics Research and Applications*, Vol. 19, No. 3, pp.151–169, doi: 10.1504/ijbra.2023.10057044.
- Rubin, D.B. (2018) 'Multiple imputation', *Flexible Imputation of Missing Data*, 2nd ed., Chapman and Hall/CRC, pp.29–62.
- Santos, D. (2023) 'FLAML-boosted XGBoost model for autism diagnosis: a comprehensive performance evaluation', *Qeios*, <https://doi.org/10.32388/A7N97Y>
- Sengupta, S., Datta, D., Rajest, S.S., Paramasivan, P., Shynnu, T. and Regin, R. (2023) 'Development of rough-TOPSIS algorithm as hybrid MCDM and its implementation to predict diabetes', *International Journal of Bioinformatics Research and Applications*, Vol. 19, No. 4, pp.252–279.
- Shi, D., Zhang, H., Wang, G., Wang, S., Yao, X., Li, Y. and Ren, K. (2022) 'Machine learning for detecting Parkinson's disease by resting-state functional magnetic resonance imaging: a multicenter radiomics analysis', *Frontiers in Aging Neuroscience*, Vol. 14, p.806828, doi: 10.3389/fnagi.2022.806828.
- Shilaskar, S., Bhatlawande, S., Deshmukh, S. and Dhande, H. (2023) 'Prediction of autism and dyslexia using machine learning and clinical data balancing', *2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS)*, IEEE, Kochi, India, pp.1–11.
- Singh, K., Visuwasam, L.M.M., Rajasekaran, G., Regin, R., Rajest, S.S. and Shynnu, T. (2024) 'Innovations in skeleton-based movement recognition bridging AI and human kinetics', *Explainable AI Applications for Human Behavior Analysis*. IGI Global, USA, pp.125–141.
- Singhal, S., Kotagiri, A., Samayamantri, L.S. and Rajest, S.S. (2024) 'Interpretable machine learning models for human action and emotion deciphering', *Advances in Computer and Electrical Engineering*, IGI Global, USA, pp.449–468.
- Stalin, G.K., Bayoumi, M.M. and Jeganathan, J. (2024) 'Knowledge and attitude regarding traction among patients with fracture', *FMD Transactions on Sustainable Health Science Letters*, Vol. 2, No. 1, pp.22–30.
- Tabtah, F. (2017) 'Autistic spectrum disorder screening data for adolescents', *UCI Machine Learning Repository*, <https://doi.org/10.24432/C5V89T>
- Thabtah, F. (2017a) 'Autistic spectrum disorder screening data for children', *UCI Machine Learning Repository*, <https://doi.org/10.24432/C5659W>
- Thabtah, F. (2017b) 'Autism screening adult', *UCI Machine Learning Repository*, <https://doi.org/10.24432/C5F019>
- Thabtah, F. (2019) 'Machine learning in autistic spectrum disorder behavioral research: a review and ways forward', *Informatics for Health and Social Care*, Vol. 44, No. 3, pp.278–297.

- Wang, L., Feng, Q., Ge, X., Chen, F., Yu, B., Chen, B. and Ding, Z. (2022) ‘Textural features reflecting local activity of the hippocampus improve the diagnosis of Alzheimer’s disease and amnesic mild cognitive impairment: a radiomics study based on functional magnetic resonance imaging’, *Frontiers in Neuroscience*, Vol. 16, p.970245, doi: 10.3389/fnins.2022.970245.
- Wen, G., Cao, P., Bao, H., Yang, W., Zheng, T. and Zaiane, O. (2022) ‘MVS-GCN: a prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis’, *Computers in Biology and Medicine*, Vol. 142, p.105239.
- Zhang, J., Xie, J. and Ma, X. (2021) ‘An imbalanced classification survey’, *ACM Computing Surveys (CSUR)*, Vol. 54, No. 1, pp.1–35.