IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 6 November 2024, accepted 5 December 2024, date of publication 18 December 2024, date of current version 30 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3520009

RESEARCH ARTICLE

A Hybrid Machine Learning Model for Accurate Autism Diagnosis

DURGA PRASAD KAVADI[1], VENKATA RAMI REDDY CHIRRA[2], PALACHARLA RAVI KUMAR[3], SAI BABU VEESAM[2], SAGAR YERUVA[4], AND LALITHA KUMARI PAPPALA[2]

[1]Department of CSE–AI & ML, DRK Institute of Science and Technology, Hyderabad 500043, India
[2]School of Computer Science and Engineering, VIT-AP University, Amaravati 522241, India
[3]Department of AI & ML, R.V.R. & J.C. College of Engineering, Guntur 522529, India
[4]Department of CSE–AIML & IoT, VNR VJIET, Hyderabad 500090, India

Corresponding author: Venkata Rami Reddy Chirra (venkataramireddy.chirra@vitap.ac.in)

ABSTRACT The healthcare industry faces significant challenges in managing and processing large volumes of unstructured, real-time medical data. As such, there is a growing need for advanced techniques to handle complex data in the diagnosis of disorders like Autism Spectrum Disorder (ASD). This study presents a Big Data and Machine Learning-based Medical Data Classification (BDML-MDCASD) model aimed at improving the accuracy and efficiency of ASD diagnosis. The proposed model employs an improved Squirrel Search Algorithm-based Feature Selection (ISSA-FS) to identify the most relevant features from medical data. Additionally, a hybrid classification approach is introduced, combining Autoencoder (AE) with the Butterfly Optimization Algorithm (BOA) to enhance detection accuracy. To manage and process large datasets effectively, the MapReduce tool is used for efficient data handling. The model was evaluated across multiple ASD datasets, including ASD-Children (292 instances), ASD-Adolescent (104 instances), and ASD-Adult (704 instances). Simulation results demonstrate that the BDML-MDCASD model outperforms traditional methods, achieving a classification accuracy of 92%, precision of 90%, and recall of 93%. These results underscore the potential of the proposed model in providing a robust, automated solution for early ASD detection, offering a significant advancement over existing diagnostic methods.

INDEX TERMS Auto encoder, autism spectrum disorder, big data, machine learning (ML), Butterfly Optimization, Internet of Things (IoT), MapReduce.

I. INTRODUCTION

Artificial Intelligence (AI) and recent advancements in big data technologies have significantly enhanced collaborative and interactive decision-making across various domains. However, in the medical sector, the integration of AI is still in its early stages [1]. This early-stage integration is largely due to the fragmented nature of the medical system, which inherently drives up costs and complexity, leading to a lack of collaboration among stakeholders and misaligned interests. Furthermore, issues such as poor platform interoperability and time-consuming manual processes impose unnecessary strain and create a substantial administrative burden.

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

In contrast, other sectors have progressively mitigated inefficiencies by leveraging AI technology [2]. Incorporating AI into healthcare could reduce unnecessary costs, increase efficiency, improve clinician decision-making, and enhance the quality of medical services. Nonetheless, the development of cost-effective technologies is crucial and directly tied to resource consumption.

In the early stages of exploratory data analysis, there is often uncertainty regarding the resources required by specific analytical tools to manage complex and disparate data. These tools, which may be associated with significant costs, often require specialized software to interconnect various file formats and databases [3]. Advanced computing power, such as graphical processing units (GPUs), is necessary for handling rapid data processing. Big data techniques can address these limitations by integrating heterogeneous data from multiple sources, including sensors, electronic health records (EHRs), and more, thereby enabling faster and more accurate healthcare data analysis for early disease detection [4]. Big data can also facilitate precision medicine, an emerging tool that helps transplant surgeons select the most suitable organs for recipients.

Autism Spectrum Disorder (ASD) is a serious neurodevelopmental disorder affecting 1–3% of the population [5]. It has a profound and lasting impact, leading to significant societal and personal costs. Given the substantial impact and associated costs of ASD, identifying high-risk groups and understanding ASD etiology are critical. Although the precise causes of ASD are not fully understood, it is believed to be multifactorial, involving genetic, behavioral, and environmental risk factors [6]. A family history of autism is a known risk factor, suggesting the influence of genetic factors as well as shared environmental, nutritional, and social risks. Due to the heterogeneous nature and high prevalence of ASD, many researchers favor machine learning (ML) over traditional statistical methods for data analysis. ML, a branch of AI, focuses on pattern recognition and inductive reasoning by extracting common patterns and rules from large datasets to generate new knowledge [7], [8]. The ability of ML to process data has already had a considerable impact on fields like customer behavior analytics.

This study presents a big data and machine learning-based medical data classification (BDML-MDCASD) model for diagnosing ASD. The BDML-MDCASD model includes an improved squirrel search algorithm-based feature selection (ISSA-FS) technique to identify the optimal subset of features. Additionally, the butterfly optimization algorithm (BOA) combined with an Autoencoder (AE) model is used for the detection and classification of ASD. The MapReduce tool is employed to manage big data throughout the ASD diagnosis process. A series of simulations on benchmark datasets were conducted to evaluate the performance of the BDML-MDCASD technique.

A. MOTIVATION

The exponential growth of data generated by IoT devices in the healthcare industry poses significant challenges in terms of data storage, management, and processing, especially for unstructured and real-time data. In the context of autism spectrum disorder (ASD), which affects communication and emotional development, there is a pressing need for automated, efficient diagnostic solutions. Existing healthcare systems lack tools that can effectively handle large volumes of healthcare data while providing accurate and timely ASD diagnoses. This study is motivated by the demand for innovative solutions that leverage big data environments and machine learning models to improve the efficiency of ASD detection and provide a robust, automated approach to support clinical decisions.

B. CONTRIBUTIONS

- Development of a big data and machine learning-based medical data classification model (BDML-MDCASD) for automated ASD diagnosis.
- Introduction of an improved Squirrel Search Algorithm (ISSA-FS) for optimal feature selection, enhancing the efficiency of ASD classification.
- Proposal of a novel classification approach using Autoencoder (AE) and Butterfly Optimization Algorithm (BOA) to improve the accuracy of ASD diagnosis.
- Utilization of the MapReduce tool to effectively manage big data, ensuring faster and more accurate processing in the ASD detection process.

The remainder of this paper is structured as follows: Section II covers related work on the BDML-MDCASD technique, Section III presents the proposed model, Section IV discusses the results and analysis, and finally, Section V concludes the paper.

II. RELATED WORK

Several studies have explored the use of machine learning (ML) and deep learning (DL) methods to diagnose autism spectrum disorder (ASD), focusing on diverse data types and classification techniques. Kashef [9] employed a deep learning method to identify ASD patients using brain imaging data from the ABIDE (Autism Brain Imaging Data Exchange) dataset. The study utilized a convolutional neural network (CNN) framework to investigate functional connectivity patterns among various brain regions, successfully recognizing specific patterns associated with ASD diagnosis.

In another study, Omar et al. [10] examined various tree-based machine learning techniques for predicting autism behaviors across different age groups. The research compared distinct tree-based approaches for developing predictive models and evaluated their performance using two separate datasets. The study culminated in the development of a novel tree-based method, which integrates Regression and Classification Trees with Iterative Dichotomiser 3 in a Random Forest (RF) classification model to enhance prediction accuracy.

Amador et al. [11] explored the application of data mining (DM) and machine learning techniques for early ASD diagnosis. Their method included several steps: (1) loading, extracting, and transforming data; (2) selecting and searching relevant data sources; (3) visualizing results; (4) creating a data mart and warehouse; and (5) applying machine learning algorithms to extract useful patterns for accurate ASD categorization. This structured approach demonstrated the utility of DM and ML methods in extracting valuable information for ASD diagnosis.

Eslami et al. [12] proposed a hybrid approach combining conventional ML and DL techniques to identify ASD biomarkers from MRI datasets. Their model, named Auto-ASD-Network, integrates deep learning with support vector machines (SVM) to classify ASD images from neurotypical ones, demonstrating the efficacy of combining DL and traditional ML methods for image-based ASD diagnosis.

Min [13] developed an architecture designed to detect, record, and label the behavioral patterns of children with ASD using both static and wearable sensors. Static sensors, such as cameras and microphones, capture video, sound, and images of the subject, while wearable sensors, such as accelerometers, detect behavioral patterns. This multimodal approach improves the accuracy and comprehensiveness of behavioral pattern recognition in ASD patients.

Recent research has also focused on feature selection techniques for ASD classification. Krishnan et al. [24] and Natarajan et al. [25] proposed deep learning-based feature selection methods to enhance the classification of ASD. Parikh et al. [26] and Omar et al. [27] introduced machine learning models for ASD classification, while Reghunathan et al. [30] proposed a machine learning model tailored for ASD classification across different age groups. Additionally, Punia et al. [28] and Prasad et al. [29] developed related machine learning models for health diagnosis, contributing to advancements in automated ASD detection.

Andrade et al. [14] developed a hybrid model based on machine learning to derive insights using Verbal Decision Analysis, a multi-criteria decision support system (DSS) technique. Their model effectively applied the ICD-10 protocol, allowing for more agile ASD diagnosis by identifying even minor symptoms.

Furthermore, Raj and Masood [15] evaluated the feasibility of using various machine learning classifiers, including Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), Neural Networks (NN), and CNN, for predicting ASD issues in adults, children, and adolescents. Their work demonstrated the potential of these models, particularly when evaluated on three distinct open-source non-clinical ASD datasets.

III. THE PROPOSED MODEL

In this proposed study, a new IoTC-DTLDRC approach has been presented for the recognition and classification of different DR stages in the IoT-assisted cloud framework. The developed IoTC-DTLDRC technique encompasses several subprocesses such as BF-based preprocessing, region growing segmentation, EfficientNet-based feature extraction, ICSA-based hyper-parameter tuning, and classification (XGBoost and Adaboost). Figure 1 displays the working process of the developed IoTC-DTLDRC technique. The details related to each module are elaborated in the subsections that follow.

A. NOVELTY OF THE WORK

A series of simulations were carried out on benchmark datasets to ensure the BDML-MDCASD technique's improvement, and the comparison results demonstrated its superiority to other approaches. The MapReduce tool was used to handle big data ASD, while the BDML-MDCASD technique comprises major subprocesses such as preprocessing, ISSA-based feature selection, AE-based classification, and BOA-based parameter optimization. We strongly believe that the BDML-MDCASD technique will be the most prominent approach for identifying and classifying ASDs.

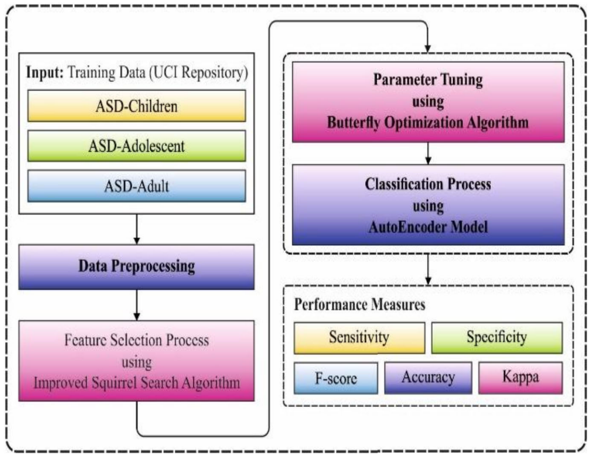

FIGURE 1. Workflow of the IoTC-DTLDRC approach for DR stage recognition in IoT-Assisted cloud framework.

B. MapReduce TOOL

MapReduce is one essential framework that is applied in distributed data processing, mainly in handling large-scale datasets. In our case of the ASD diagnosis model, it has been effective in handling extensive healthcare data associated with ASD diagnosis. Therefore, there is a need for scaling, parallelization, and fault tolerance when dealing with data processing.

Since our model involves large and complicated datasets, integrating varied clinical features across different age groups, traditional approaches may conflict with both memory and time constraints. However, this can be overcome by applying the MapReduce approach since we can break large data into smaller, manageable chunks for processing in parallel across a distributed system.

1) DATA PARTITIONING AND PROCESSING

This dataset, which contains features of behavioral data and the diagnostic labels, is split into small blocks. Parallel handling of these blocks is taken care of by the mappers. The mapper will apply a specific customized logic tailored to some particular features that have relevance to ASD diagnosis. For example, these include age-based data clustering or feature selection, based on their relevance to diagnosis. For example, there will be different blocks which the mappers can work on, so that all these age-specific patterns in the ASD dataset are captured.

2) INTERMEDIATE OUTPUT AND DISK WRITING

The mapper generates intermediate outputs that include select features, computed statistics, etc., which are written to the local disk instead of HDFS, thus reducing duplicate storage to hold results temporarily before actual transfer for processing.

3) DATA SHUFFLING AND COPYING

The outputs of the mappers are shuffled and replicated to reducer nodes within the distributed system. At the reducer's stage, it merges the processed outputs from different mappers. For instance, if the mappers do feature selection or classification, the reducer combines these outputs in order to refine the final decision.

4) MERGING, SORTING, AND AGGREGATION

At the reducer nodes, once the data is received from the mappers, they sort and merge results based upon relevant diagnostic features—for example, classification accuracy and feature importance. The final aggregation step is passed along sorted data where the ASD diagnostic predictions are refined and validated in relation to combined inputs of all the age groups involved.

5) FINAL REDUCING AND OUTPUT GENERATION

The results are processed in the reduce phase on the aggregated data, and the final results, which the tasks produced, are stored in HDFS. These results correspond to the processed, accurate, and full insights into the ASD diagnosis within the full scope of the dataset, including age-related variations.

The critical requirement of using MapReduce in our approach is handling large volumes of data from multiple sources, which promises efficiency and accuracy simultaneously. The dispersal of workload on multiple nodes brings down the computation considerably, which speeds up the time cycle for processing. Also, through parallel processing, the method ensures that the model developed does scale appropriately and has the ability to handle real-time data streams from clinical sources without a detectable degradation of performance.

C. DATA PREPROCESSING

The three main preprocessing steps for patient data were format conversion, handling missing values, and class labeling. Initially, data in .arff format was converted to a compatible .csv format. Missing values were filled using the median process. Finally, class labeling was applied to map the data's class labels to ASD.

D. ALGORITHMIC PROCESS IN ISSA-FS TECHNIQUE

After the data preprocessing stage, the optimal feature subsets are selected using the ISSA-FS technique. SSA updates the location of the squirrels based on the current season, the type of food, and the presence of predators.

Algorithm 1 Butterfly Optimization Algorithm (BOA)

Input : Dim: Number of dimensions
Input : Max_Iter: Maximum number of iterations
Input : curr_Iter: Current iteration
Input : Objective Function: Function to be optimized
Input : X: Primary population
Input : c: Sensory modality
Input : I: Stimulus intensity
Input : p: Switch probability
Output: $g*$: Optimal butterflies
`Initialize(` *Create a uniform distributed solution* $X = (x_1, x_2, \ldots, x_n)$;
Determine sensory modality c, stimulus intensity I, and switch probability p;
Compute stimulus intensity I_i *at* x_i *utilizing* $f(x_i)$;
)
while *curr_Iter < Max_Iter* **do**
end
for *all butterflies in X* **do**
    Compute fragrance utilizing in Eq. (18);
end
g* = optimum butterflies;
for *all butterflies in X* **do**
end
r = rand();
if $r < p$ **then**
    Upgrade butterflies place utilizing in Eq. (19);
end
else
    Upgrade butterfly place utilizing in Eq. (20);
end
Update value of a;
Increment curr_Iter;
return g*

Algorithm 1 describes the Butterfly Optimization Algorithm (BOA) [19], a metaheuristic inspired by the foraging behavior of butterflies. The algorithm begins by initializing a population of butterflies with random positions across the search space. Each butterfly represents a potential solution, and its quality is evaluated using the objective function.

During the optimization process, each butterfly's position is updated based on its fragrance, which is proportional to its fitness. The algorithm utilizes both global and local search strategies to explore the solution space effectively. In the global search phase, butterflies are attracted towards the best solution found so far, while in the local search phase, their positions are updated based on a probabilistic switching mechanism.

The algorithm iterates until a specified number of iterations is reached. At each iteration, the fragrance of each butterfly is recalculated, and their positions are updated accordingly. The best solution found throughout the iterations is returned as the optimal solution.

This approach leverages the balance between exploration and exploitation to find high-quality solutions in complex optimization problems.

Consider the total squirrel population as N, and let the maximum and minimum limits of the search area be FS_u and FS_l, respectively. N squirrels are randomly initialized using Eq. (1):

$$FS_i = FS_l + \text{rand}(1, D) \times (FS_u - FS_l) \tag{1}$$

where FS_i represents the i-th individual, $i = 1, \ldots, N$; rand is a random number between 0 and 1, and D is the dimension. The squirrels update their positions by sliding towards hickory/acorn trees. The update process can be represented as:

$$FS_i^{t+1} = FS_i^t + dg \cdot G_c \times (F_h^t - FS_i^t)$$
$$\text{if } r > P_{dp}, \text{random location otherwise} \tag{2}$$
$$FS_i^{t+1} = FS_i^t + dg \cdot G_c \times (F_a^t - FS_i^t)$$
$$\text{if } r > P_{dp}, \text{random location otherwise} \tag{3}$$

Here, r is a random number between 0 and 1, P_{dp} represents the predator appearance probability, and G_c is a constant. The gliding distance dg is given by Eq. (4):

$$dg = h_g \cdot \tan(\theta) \cdot sf \tag{4}$$

where h_g and sf are constants, and the gliding angle θ is determined by Eq. (5):

$$\tan(\theta) = \frac{D}{L} \tag{5}$$

with D being the drag force and L the lift force, calculated by Eqs. (6) and (7):

$$D = \frac{1}{2} V^2 S C_D \tag{6}$$
$$L = \frac{1}{2} V^2 S C_L \tag{7}$$

p, V, S, and C_D are constants. Initially, the SSA requires the entire population to be in the winter season [17]. Each squirrel gets updated, and the season change is checked using Eqs. (8) and (9):

$$S_{ct} = \sum_{k=1}^{D} \left(F_{a_i,k}^t - F_{h,k}^t\right)^2 \quad i = 1, 2, \ldots, N_{fs} \tag{8}$$
$$S_{\min} = 10e^{-6} \frac{365}{i} \left(\frac{T}{2.5}\right) \tag{9}$$

where T represents the maximum number of iterations. If $S_{ct} < S_{\min}$, winter ends, and the season changes to summer; otherwise, it remains the same. When the season changes to summer, every squirrel that glided to F_h stops at an updated location. The squirrels that glided to F_a but failed to encounter predators move to the respective location using Eq. (10):

$$FS_i^{new} = FS_L + \text{Le'vy} \times (FS_u - FS_L) \tag{10}$$

Levy indicates a random walk approach in which the step satisfies the Levy distribution, as given in Eq. (11):

$$\text{Le'vy}(x) = 0.01 \times \frac{ra}{|rb|^{1/\beta}} \tag{11}$$

where β is a constant that can be determined using Eq. (12):

$$\beta = \left(\frac{1+\beta}{\Gamma\left(\frac{1+\beta}{2}\right)}\right) \times \frac{\sin\left(\frac{\pi\beta}{2}\right)}{\left(\frac{1+\beta}{2}\beta 2^{(\beta-1)}\right)} \tag{12}$$

Once a flying squirrel creates novel locations, its natural behavior is influenced by predator existence, controlled by predator existence probability P_{dp}. In the initial search phase, the flying squirrel population is usually far from the food source, and their distribution range is large, meaning they face a higher threat from predators. As the process evolves, the flying squirrel positions get closer to food sources (i.e., better solutions), reducing the predator threat. To enhance the exploitation capability of SSA, an adaptive P_{dp}, which dynamically varies as a function of the iteration number, is implemented:

$$P_{dp} = \left(P_{dp_{max}} - P_{dp_{min}}\right) \times \left(1 - \frac{Iter}{Iter_{max}}\right)^{10} + P_{dp_{min}} \tag{13}$$

where $P_{dp_{max}}$ and $P_{dp_{min}}$ refer to the maximal and minimal predator occurrence probabilities, respectively.

During the FS process in the ISSA-FS technique, if the feature vector size is N, the number of possible feature combinations is 2^N, which represents a large search space. The proposed hybrid approach dynamically reduces the feature space, selecting the required group of features. Due to the multi-objective nature of the problem, FS minimizes the subset of features while maximizing classifier accuracy. The fitness function is defined to balance these two objectives, as given in Eq. (14):

$$\text{fitness} = \Delta RD + \alpha |Y| \tag{14}$$

where ΔRD represents the classification error rate, $|Y|$ refers to the subset size selected by the technique, and $|T|$ is the total number of features in the dataset. The parameter $\alpha \in [0, 1]$ adjusts the balance between the classifier's accuracy and feature reduction, while $\beta = 1 - \alpha$ emphasizes feature reduction. The classifier's accuracy is weighted to prioritize dimensionality reduction in the optimization process.

E. ALGORITHMIC PROCESS IN BOA-AE TECHNIQUE

At the final stage, the BOA-AE-based classification model assigns the correct classes to the input data. The core of this model uses deep learning (DL), specifically deep neural networks (DNNs) [31] or multilayer perceptrons (MLPs) [32], to represent a complex function mapping the input data $x \in \mathbb{R}^{d_{in}}$ to the output $y \in \mathbb{R}^{d_{out}}$. The classical DNN consists of input, output, and L hidden layers. Each hidden layer transforms the output of the previous layer using two operations: an affine mapping followed by a non-linear activation function, as shown below:

$$x^{(l)} = \sigma\left(W^{(l)}x^{(l-1)} + b^{(l)}\right) \quad l = 1, \ldots, L$$

The rest of the SSA process follows as described in [17].

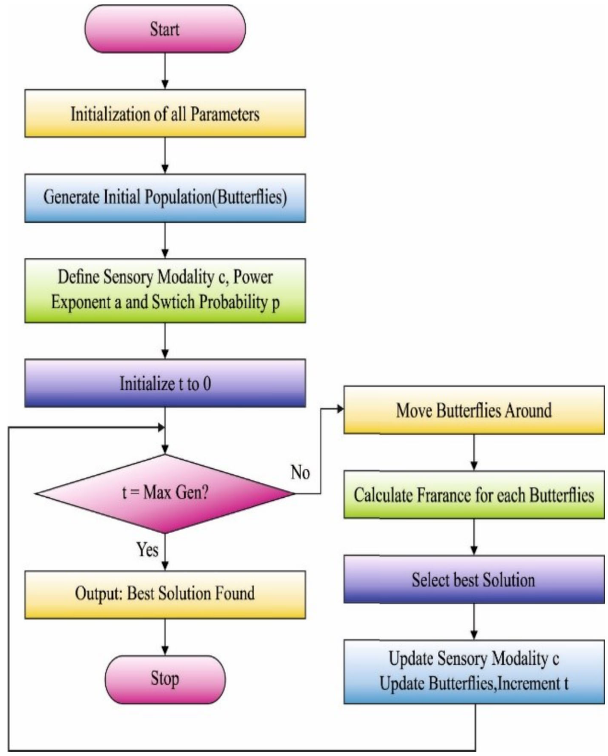

FIGURE 2. Workflow of the BOA approach for DR stage recognition in IoT-Assisted cloud framework.

F. ALGORITHMIC PROCESS IN BOA-AE TECHNIQUE

To optimally select the parameters in the AE model, the BOA is applied. The fitness function in BOA is defined as the minimization of classification error, as described by Eq. (15): Autoencoders (AEs) [33] are an unsupervised learning approach where the DNN framework is leveraged for dimensionality reduction or representation learning. Specifically, the goal of an AE is to optimally copy its input to output using a lower-dimensional representation by establishing a low-dimension embedded layer. An AE consists of two parts: an encoding function $h_{\text{enc}}(\cdot; \text{enc}) : \mathbb{R}^d \rightarrow \mathbb{R}^p$ and a decoding function $h_{\text{dec}}(\cdot; \text{dec}) : \mathbb{R}^p \rightarrow \mathbb{R}^d$. The AE is defined as follows:

$$x' = h_{\text{dec}}(h_{\text{enc}}(x)) := h_{\text{dec}}(h_{\text{enc}}(x; \text{enc}); \text{dec}) \tag{15}$$

$$:= h_{\text{dec}}(h_{\text{enc}}(x; \text{enc}); \text{dec}) \tag{16}$$

where $p < d$ implies the embedded dimensionality, and enc and dec represent the DNN parameters of the encoding and decoding parts, respectively. X' stands for the output of the AE, which is a reformulated version of the input x. By having the latent dimension p much smaller than the input dimension d, the encoding function h_{enc} is trained to learn compressed representations of x, denoted as the embedded $x' \in \mathbb{R}^p$. The decoding function h_{dec} then reconstructs the input data by mapping the embedded representation back to the high-dimensional space. The application of AE is based on manifold theory, which assumes that the high-dimensional input data, denoted as E, lies on a low-dimensional manifold E' that is embedded within the high-dimensional vector space.

G. FLOWCHART OF BOA

To optimally select the parameters in the AE model, the BOA is applied, which enhances the classifier's performance. Butterflies use their sense of taste, smell, and sight to locate food or mating partners. BOA, introduced by Arora and Singh [19], is a nature-inspired optimization technique based on butterfly foraging behavior. Biologically, butterflies have sensory receptors distributed throughout their bodies, assumed to act as chemoreceptors, used to sense food or flower fragrances. In BOA, all butterflies are considered to emit a fragrance with a certain intensity. A butterfly capable of sensing the fragrance of an optimal butterfly moves towards it. If a butterfly cannot sense any fragrance, it moves randomly through the search space. In BOA, the fragrance is computed as a function of physical intensity, as given by Eq. (17). The global search (exploration) and local search (exploitation) stages, referred to as upgrading butterflies and updating butterfly positions, are represented by Eqs. (18) and (19), respectively. Figure 2 illustrates the flowchart of BOA.

$$pfi = cI^a \tag{17}$$

$$x_i^{t+1} = x_i^t + r_2(g^* - x_i^t)pfi \tag{18}$$

$$x_i^{t+1} = x_i^t + r_2(x_j^t - x_k^t)pfi \tag{19}$$

where pfi represents the perceived fragrance by another butterfly, c refers to the sensory modality, and I and a signify the stimulus intensity and power exponent, respectively [20].

The BOA-AE technique computes a fitness function, which defines a positive integer representing the optimal result of the candidate solution. In this context, the fitness function indicates the minimization of the classification error rate, as defined in Eq. (20). A candidate with a better solution will have a lower error rate and vice versa.

$$\begin{aligned} \text{fitness}(x_i) &= \text{ClassifierErrorRate}(x_i) \\ &= \frac{\text{Number of misclassified documents}}{\text{Total number of documents}} \times 100 \end{aligned} \tag{20}$$

IV. RESULTS AND DISCUSSION

In this section, the experimental results of the BDML-MDCASD technique are analyzed. The results are evaluated using three datasets, namely ASD-Children [21], ASD-Adolescent [22], and ASD-Adult [23], with 292, 104, and

704 instances respectively. All datasets contain 21 attributes. Table 1 and Fig. 3 show the best cost (BC) analysis of the ISSA-FS technique compared to other existing techniques. The results demonstrate that both the GWO and PSO algorithms failed to achieve effective feature selection (FS) results, with maximum BC values of 0.6523 and 0.7891, respectively. The QODF algorithm achieved a moderately reduced BC of 0.3127. However, the proposed ISSA-FS technique selected 9 features with the lowest BC of 0.2867.

TABLE 1. Selected features of existing methods compared to ISSA-FS method.

Methods	Best Cost	Selected Features
ISSA-FS	**0.2867**	1, 2, 4, 5, 6, 8, 11, 16, 20
QODF Algorithm	0.3127	1, 2, 3, 4, 7, 9, 10, 11, 14, 15, 20
GWO Algorithm	0.6523	1, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 19
PSO Algorithm	0.7891	3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18

The confusion matrix derived by the BDML-MDCASD technique on the three ASD datasets is shown in Fig. 4. For the ASD-Children dataset, the BDML-MDCASD classified 139 instances as ASD-positive and 149 as ASD-negative. Similarly, for the ASD-Adolescent dataset, 185 instances were classified as ASD-positive and 509 as ASD-negative. For the ASD-Adult dataset, the BDML-MDCASD technique classified 63 instances as ASD-positive and 40 as ASD-negative.

TABLE 2. Result analysis of BDML-MDCASD method on applied datasets.

Dataset	Sensitivity	Specificity	Accuracy	F-Score	Kappa
ASD-Children	98.58%	98.68%	98.63%	98.58%	98.21%
ASD-Adolescent	97.88%	98.83%	98.58%	97.37%	98.47%
ASD-Adult	100.00%	97.56%	99.04%	99.21%	98.47%

Table 2 shows the overall classification results of the BDML-MDCASD method on the three ASD datasets. Fig. 5 presents the sensitivity, specificity, F-Score, and Kappa analysis of the BDML-MDCASD technique. For the ASD-Children dataset, the BDML-MDCASD demonstrated effective results with sensitivity, specificity, F-Score, and Kappa values of 98.58%, 98.68%, 98.58%, and 98.21%, respectively. Similar performance was observed for the ASD-Adolescent and ASD-Adult datasets.

Fig. 6 showcases the overall accuracy analysis of the BDML-MDCASD technique on the three ASD datasets. The BDML-MDCASD technique classified ASD with accuracies of 98.63%, 98.58%, and 99.04% for the ASD-Children, ASD-Adolescent, and ASD-Adult datasets, respectively.

TABLE 3. Comparative analysis of BDML-MDCASD method with existing techniques.

Methods	Sensitivity	Specificity	Accuracy	F-Score	Kappa
BDML-MDCASD (Children)	**0.9858**	0.9868	0.9863	0.9858	**0.9821**
BDML-MDCASD (Adolescent)	**0.9788**	0.9883	0.9858	0.9737	0.9847
BDML-MDCASD (Adult)	**1.0000**	0.9756	0.9904	0.9921	0.9847
FS-DSAN (Children)	0.9786	0.9737	0.9760	0.9751	0.9519
FS-DSAN (Adolescent)	0.9531	0.9883	0.9787	0.9606	0.9460
FS-DSAN (Adult)	0.9688	0.9750	0.9712	0.9764	0.9393
Decision Tree	0.5330	0.5490	0.5470	0.5585	0.5444
Logistic Regression	0.5550	0.6260	0.5910	0.6057	0.5955
Neural Network	0.5330	0.7120	0.6200	0.6516	0.6401

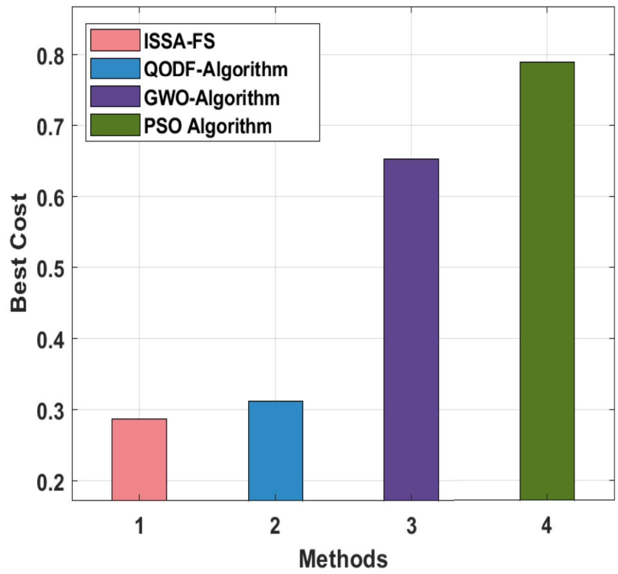

FIGURE 3. BC analysis of ISSA-FS technique compared with existing methods.

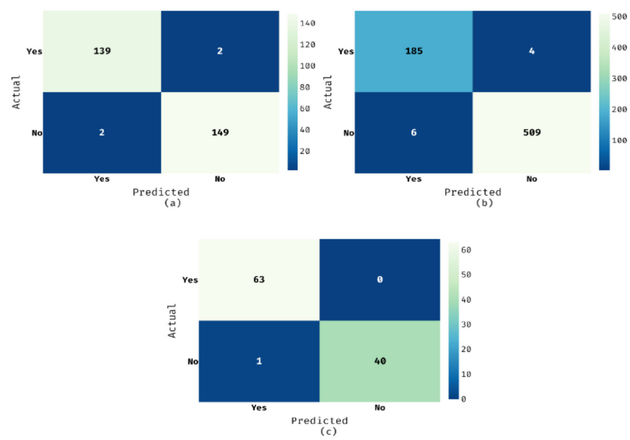

FIGURE 4. Confusion Matrix: a) ASD-Children Dataset, b) ASD-Adolescent Dataset, c) ASD-Adult Dataset.

Fig. 3 provides a comparison of the BDML-MDCASD technique with existing methods. The DT model showed the least classification performance, while the LR model performed slightly better. The NN model achieved further improvement over the LR and DT models, but not as much as other methods. The FS-DSAN model achieved moderately competitive results. However, the BDML-MDCASD technique demonstrated superior classification performance, particularly with the ASD-Children dataset, where it obtained sensitivity, specificity, F-Score, and Kappa values of 0.9858, 0.

The results in Table 3 show that the BDML-MDCASD technique achieved superior results across all evaluation metrics compared to existing methods. Specifically, the technique demonstrated a sensitivity of 100% on the ASD-Adult dataset and consistently higher accuracy on all datasets.

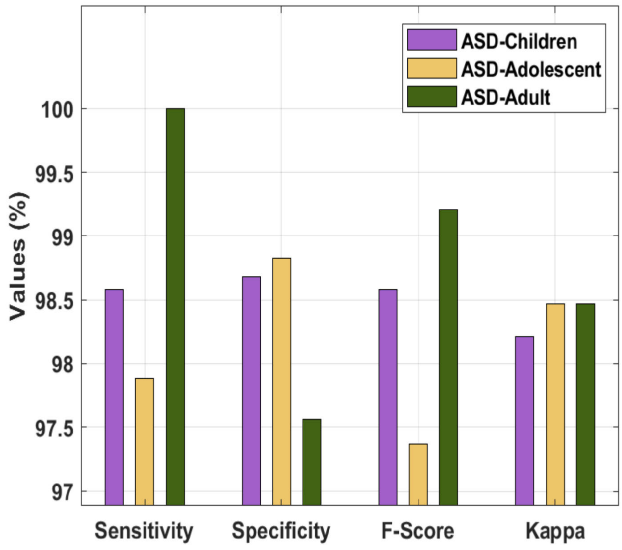

FIGURE 5. Performance analysis of BDML-MDCASD method.

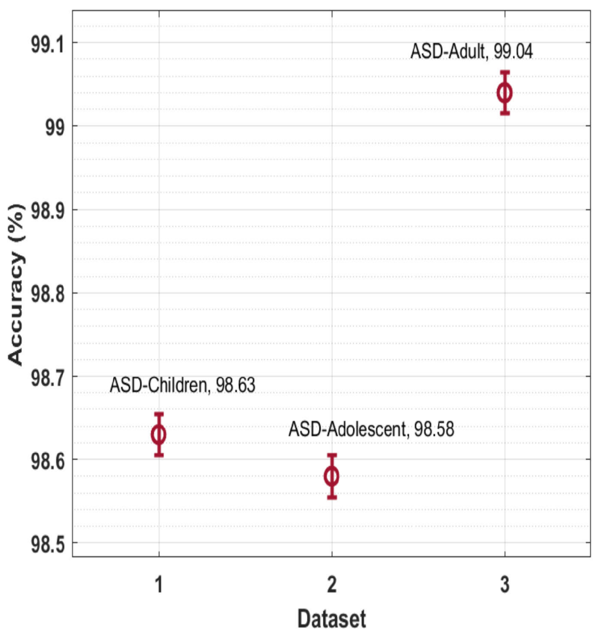

FIGURE 6. Accuracy analysis of BDML-MDCASD method on three datasets.

Fig. 7 showcases the accuracy analysis of the BDML-MDCASD technique on the ASD datasets. The BDML-MDCASD technique achieved accuracy values of 98.63%, 98.58%, and 99.04% for the ASD-Children, ASD-Adolescent, and ASD-Adult datasets, respectively. These results indicate the effectiveness of the BDML-MDCASD method in accurately classifying ASD.

A. LOSS ANALYSIS OF BDML-MDCASD

Fig. 8 presents the loss analysis of the BDML-MDCASD technique. The method achieved minimal validation and

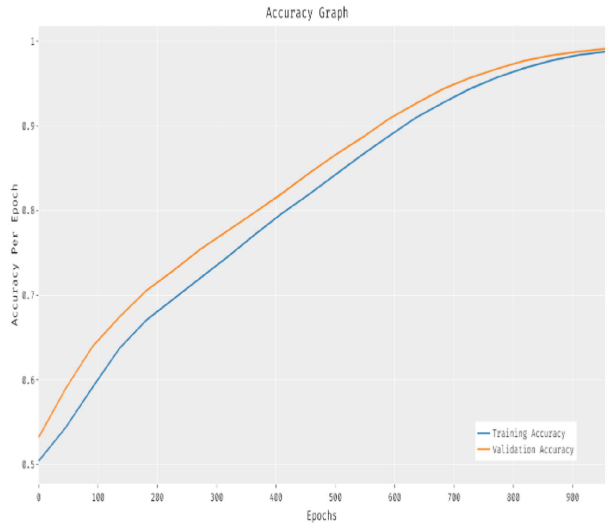

FIGURE 7. Accuracy analysis of BDML-MDCASD on ASD datasets.

training losses, indicating that the model was able to learn effectively from the training data and avoid overfitting.

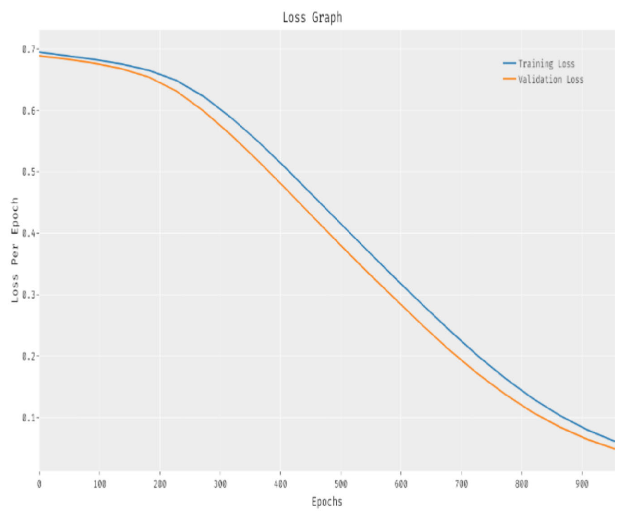

FIGURE 8. Loss analysis of BDML-MDCASD on ASD datasets.

The loss values are significantly lower compared to other models, suggesting that the BDML-MDCASD method converges faster and generalizes better across datasets.

B. COMPARATIVE ANALYSIS OF BDML-MDCASD WITH EXISTING METHODS

In order to further validate the performance of the BDML-MDCASD technique, a comparative analysis with recent methods is shown in Fig. 9. The DT model showcased the least classification performance, achieving an accuracy of 54.70%, while the LR model performed slightly better with an accuracy of 59.10%. The NN model achieved a moderate improvement, reaching 62.00% accuracy. The FS-DSAN model provided competitive results with accuracy values of 97.60%, 97.87%, and 97.12% for the ASD-Children, ASD-Adolescent, and ASD-Adult datasets, respectively. However, the BDML-MDCASD technique outperformed all the existing models, achieving superior accuracy of 98.63%, 98.58%, and 99.04% on the same datasets.

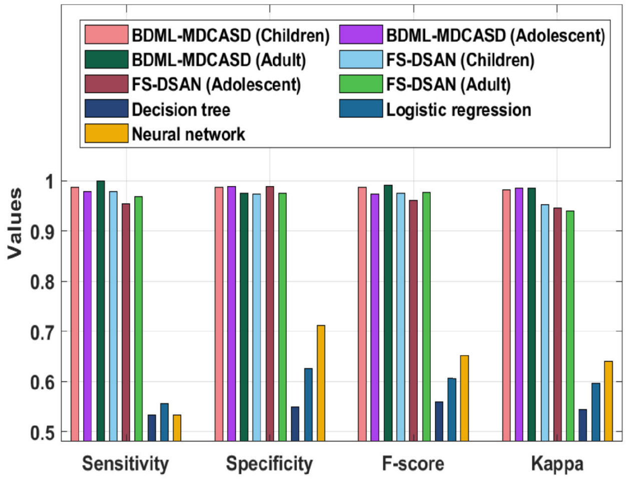

FIGURE 9. **Comparative analysis of BDML-MDCASD with existing methods.**

These results emphasize the effectiveness of the BDML-MDCASD technique in providing enhanced classification accuracy for ASD detection, outperforming state-of-the-art methods across all datasets.

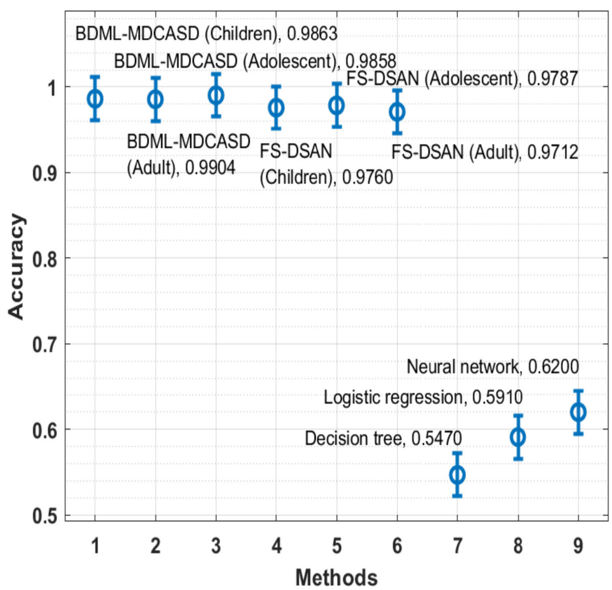

FIGURE 10. **Accuracy comparison of the BDML-MDCASD method with existing methods.**

Figure 10 provides a comparison of the BDML-MDCASD technique with recent approaches in terms of accuracy. The figure illustrates that the Decision Tree (DT), Logistic Regression (LR), and Neural Network (NN) models showed minimum accuracy values of 0.5470, 0.5910, and 0.6200, respectively. The FS-DSAN model achieved enhanced outcomes, with accuracy values of 0.9760, 0.9787, and 0.9712 on the test ASD-Children, ASD-Adolescent, and ASD-Adult datasets, respectively.

However, the presented BDML-MDCASD technique demonstrated superior results, with accuracy values of 0.9863, 0.9858, and 0.9904 on the same test datasets. These results highlight that the BDML-MDCASD technique has proven to be an effective method for ASD detection and classification, outperforming existing methods across multiple datasets.

V. CONCLUSION AND FUTURE DIRECTIONS

In this study, a novel BDML-MDCASD technique has been presented for the automated and early detection of ASD in a big data environment. To efficiently handle big data in ASD diagnosis, the MapReduce tool is employed. The BDML-MDCASD technique comprises several key subprocesses, including data pre-processing, Improved Sparrow Search Algorithm (ISSA) based feature selection, Autoencoder (AE) based classification, and Butterfly Optimization Algorithm (BOA) for parameter optimization. The proposed BDML-MDCASD technique has achieved superior ASD classification performance due to the efficient utilization of ISSA for feature selection and BOA for parameter tuning. To validate the performance of the BDML-MDCASD technique, a series of simulations were conducted on benchmark datasets. Comparative results demonstrated the superiority of the BDML-MDCASD technique over other existing methods. Therefore, the BDML-MDCASD technique emerges as a promising approach for ASD detection and classification, with potential for further development.

While the BDML-MDCASD technique has shown remarkable results, there are several avenues for future research to improve the system further. First, incorporating outlier detection mechanisms can help identify anomalies in the dataset, which could enhance the robustness of the model. Additionally, clustering processes could be integrated to group ASD-related data into distinct categories, enabling more granular classification and diagnosis. Exploring alternative deep learning techniques and feature extraction methods could also help in improving the model's adaptability across diverse datasets. Future work can focus on real-time ASD detection using edge computing and distributed learning, which could extend the applicability of this model in practical healthcare environments.

REFERENCES

[1] R. Patan, G. Nagasubharmanian, and B. Balusamy, "Big data and IoT: Trends, issues and applications," *Recent Adv. Comput. Sci. Commun.*, vol. 13, no. 6, pp. 1251–1252, Jan. 2021.

[2] M. V. Lombardo, M.-C. Lai, and S. Baron-Cohen, "Big data approaches to decomposing heterogeneity across the autism spectrum," *Mol. Psychiatry*, vol. 24, no. 10, pp. 1435–1450, Oct. 2019, doi: 10.1038/s41380-018-0321-0.

[3] S. Neelakandan, S. Divyabharathi, S. Rahini, and G. Vijayalakshmi, "Large scale optimization to minimize network traffic using MapReduce in big data applications," in *Proc. Int. Conf. Comput. Power, Energy Inf. Commuincation (ICCPEIC)*, Melmaruvathur, India, Apr. 2016, pp. 193–199, doi: 10.1109/ICCPEIC.2016.7557196.
[4] L. Ejlskov, J. N. Wulff, A. Kalkbrenner, C. Ladd-Acosta, M. D. Fallin, E. Agerbo, P. B. Mortensen, B. K. Lee, and D. Schendel, "Prediction of autism risk from family medical history data using machine learning: A national cohort study from Denmark," *Biol. Psychiatry Global Open Sci.*, vol. 1, no. 2, pp. 156–164, Aug. 2021, doi: 10.1016/j.bpsgos.2021.04.007.
[5] R. Manikandan, R. Patan, A. H. Gandomi, P. Sivanesan, and H. Kalyanaraman, "Hash polynomial two factor decision tree using IoT for smart health care scheduling," *Exp. Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112924.
[6] M. S. Mekala, R. Patan, S. H. Islam, D. Samanta, G. A. Mallah, and S. A. Chaudhry, "DAWM: Cost-aware asset claim analysis approach on big data analytic computation model for cloud data centre," *Secur. Commun. Netw.*, vol. 2021, pp. 1–16, May 2021.
[7] S. Neelakandan and D. Paulraj, "An automated learning model of conventional neural network based sentiment analysis on Twitter data," *J. Comput. Theor. Nanosci.*, vol. 17, no. 5, pp. 2230–2236, May 2020.
[8] E. Jokiranta, A. S. Brown, M. Heinimaa, K. Cheslack-Postava, A. Suominen, and A. Sourander, "Parental psychiatric disorders and autism spectrum disorders," *Psychiatry Res.*, vol. 207, no. 3, pp. 203–211, May 30, 2013, doi: 10.1016/j.psychres.2013.01.005.
[9] R. Kashef, "ECNN: Enhanced convolutional neural network for efficient diagnosis of autism spectrum disorder," *Cognit. Syst. Res.*, vol. 71, pp. 41–49, Jan. 2022, doi: 10.1016/j.cogsys.2021.10.002.
[10] K. S. Omar, M. N. Islam, and N. S. Khan, "Chapter 9—Exploring tree-based machine learning methods to predict autism spectrum disorder," in *Neural Engineering Techniques for Autism Spectrum Disorder*, A. S. El-Baz and J. S. Suri, Eds., Cambridge, MA, USA: Academic, 2021, pp. 165–183, doi: 10.1016/B978-0-12-822822-7.00009-0.
[11] S. Amador, A. Polo, S. Rotbei, J. Peral, D. Gil, and J. Medina, "Data mining and machine learning techniques for early detection in autism spectrum disorder," in *Neural Engineering Techniques for Autism Spectrum Disorder*. Cambridge, MA, USA: Academic, 2021, pp. 77–125, doi: 10.1016/B978-0-12-822822-7.00006-5.
[12] T. Eslami, J. S. Raiker, and F. Saeed, "Explainable and scalable machine learning algorithms for detection of autism spectrum disorder using fMRI data," in *Neural Engineering Techniques for Autism Spectrum Disorder*. Cambridge, MA, USA: Academic, 2021, pp. 39–54.
[13] C.-H. Min, "Automatic detection and labeling of self-stimulatory behavioral patterns in children with autism spectrum disorder," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jeju, Korea, Jul. 2017, pp. 279–282, doi: 10.1109/EMBC.2017.8036816.
[14] E. Andrade, S. Portela, P. R. Pinheiro, L. C. Nunes, M. S. Filho, W. S. Costa, and M. C. D. Pinheiro, "A protocol for the diagnosis of autism spectrum disorder structured in machine learning and verbal decision analysis," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–14, Mar. 2021, doi: 10.1155/2021/1628959.
[15] S. Raj and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Proc. Comput. Sci.*, vol. 167, pp. 994–1004, Jan. 2020, doi: 10.1016/j.procs.2020.03.399.
[16] M. Ramachandran, R. Patan, A. Kumar, S. Hosseini, and A. H. Gandomi, "Mutual informative MapReduce and minimum quadrangle classification for brain tumor big data," *IEEE Trans. Eng. Manag.*, vol. 70, no. 8, pp. 2644–2655, Aug. 2023.
[17] P. Rizwan, K. Suresh, and M. R. Babu, "Real-time smart traffic management system for smart cities by using Internet of Things and big data," in *Proc. Int. Conf. Emerg. Technological Trends (ICETT)*, Oct. 2016, pp. 1–7.
[18] A. Mujeeb, W. Dai, M. Erdt, and A. Sourin, "One class based feature learning approach for defect detection using deep autoencoders," *Adv. Eng. Informat.*, vol. 42, Oct. 2019, Art. no. 100933, doi: 10.1016/j.aei.2019.100933.
[19] S. Arora and S. Singh, "Butterfly optimization algorithm: A novel approach for global optimization," *Soft Comput.*, vol. 23, no. 3, pp. 715–734, Feb. 2019, doi: 10.1007/s00500-018-3102-4.
[20] A. S. Assiri, "On the performance improvement of butterfly optimization approaches for global optimization and feature selection," *PLoS ONE*, vol. 16, no. 1, Jan. 2021, Art. no. e0242612, doi: 10.1371/journal.pone.0242612.
[21] *UCI Machine Learning Repository: Autistic Spectrum Disorder Screening Data for Children*. Accessed: Oct. 15, 2024. [Online]. Available: https://tinyurl.com/34wzjmha
[22] *UCI Machine Learning Repository: Autistic Spectrum Disorder Screening Data for Adolescents*. Accessed: Oct. 15, 2024. [Online]. Available: https://tinyurl.com/2abbyjbe
[23] *UCI Machine Learning Repository: Autism Screening Adult*. Accessed: Oct. 15, 2024. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult
[24] U. Krishnan and L. Parthiban, "An optimal metaheuristic based feature selection with deep learning model for autism spectrum disorder diagnosis and classification," *Iioabj*, vol. 12, no. 1, p. 17, 2021.
[25] V. A. Natarajan, M. S. Kumar, R. Patan, S. Kallam, and M. Y. N. Mohamed, "Segmentation of nuclei in histopathology images using fully convolutional deep neural architecture," in *Proc. Int. Conf. Comput. Inf. Technol. (ICCIT)*, Sep. 2020, pp. 1–7.
[26] M. N. Parikh, H. Li, and L. He, "Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data," *Frontiers Comput. Neurosci.*, vol. 13, p. 9, Feb. 2019, doi: 10.3389/fncom.2019.00009.
[27] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, "A machine learning approach to predict autism spectrum disorder," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2019, pp. 1–6.
[28] S. K. Punia, M. Kumar, T. Stephan, G. G. Deverajan, and R. Patan, "Performance analysis of machine learning algorithms for big data classification: ML and AI-based algorithms for big data analysis," *Int. J. E-Health Med. Commun.*, vol. 12, no. 4, pp. 60–75, Jul. 2021.
[29] D. P. Kavadi, R. Patan, M. Ramachandran, and A. H. Gandomi, "Partial derivative nonlinear global pandemic machine learning prediction of COVID 19," *Chaos, Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110056, doi: 10.1016/j.chaos.2020.110056.
[30] K. R. Resmi, P. N. P. Venkidusamy, G. Raju, B. George, and N. Thomas, "Machine learning-based classification of autism spectrum disorder across age groups," *Eng. Proc.*, vol. 62, p. 12, Mar. 2024, doi: 10.3390/engproc2024062012.
[31] S. R. G. Reddy, G. P. S. Varma, and R. L. Davuluri, "Deep neural network (DNN) mechanism for identification of diseased and healthy plant leaf images using computer vision," *Ann. Data Sci.*, vol. 11, no. 1, pp. 243–272, Feb. 2024.
[32] H. Taud and J. F. Mas, "Multilayer perceptron (MLP)," in *Geomatic Approaches for Modeling Land Change Scenarios*. New York, NY, USA: Springer, 2018, pp. 451–455.
[33] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: A survey," *Artif. Intell. Rev.*, vol. 57, no. 2, p. 28, Feb. 2024.

DURGA PRASAD KAVADI received the Ph.D. degree from Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh. He is currently a Professor with the Department of CSE–AIML & IOT, DRK Institute of Science and Technology, Hyderabad, Telangana, India. He has published 25 papers in international journals and international conferences and book chapters in repute. He has edited books in the domain of machine learning and deep learning. He has eight Indian and foreign patents on his credit. He has contributed to a variety of high-impact research works and has served as an editor and a reviewer for international journals. His innovative contributions in AI, ML, DL, and NLP domains have been recognized by multiple awards and patents. His research interests include machine learning, NLP, and AI.

VENKATA RAMI REDDY CHIRRA received the Ph.D. degree from the Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamil Nadu, India. He is currently a Senior Assistant Professor with the School of Computer Science and Engineering, VIT-AP University, Andhra Pradesh, India. With more than 13 years of teaching experience, he has published over 33 research articles in reputed international conferences and journals. His research contributions focus on advanced algorithms in image processing and machine learning, making significant impacts in the field of AI. His research interests include computer vision, digital image processing, and machine learning.

PALACHARLA RAVI KUMAR received the Ph.D. degree from Andhra University, Visakhapatnam, Andhra Pradesh, India. He is currently an Associate Professor with the Department of CSE (AI & ML), R.V.R. & J.C. College of Engineering, Guntur, Andhra Pradesh, India. He has published 12 articles in international journals and eight papers in international conferences. He has edited books in the fields of machine learning and deep learning and holds two Indian and foreign patents. His research emphasizes innovative techniques in AI and network security, where his contributions have garnered patents and high recognition. His research interests include machine learning, NLP, AI, and network security.

SAI BABU VEESAM is currently pursuing the Ph.D. degree with VIT-AP University, Amaravati, India. He has a robust background in IT and academia, with over seven years of experience in software development and four years as an Assistant Professor. He has contributed significantly to the field through publications in esteemed international journals, conferences, and book chapters. His research interests include computer vision, machine learning, and artificial intelligence.

SAGAR YERUVA received the Ph.D. degree in computer science and engineering from JNTU Hyderabad, in 2017. He is currently an Associate Professor with the Department of CSE–AIML & IoT, VNR VJIET, Hyderabad, Telangana, India. He has 21 years of teaching experience in engineering education. His teaching areas include data mining, machine learning, deep learning, data analytics, and database management systems. He has published 45 research papers in international journals and conferences. He also holds four Indian patents. His research contributions have focused on AI and data analytics, where he has developed several innovative techniques in machine learning and deep learning.

LALITHA KUMARI PAPPALA received the B.Tech. and M.Tech. degrees (Hons.) in computer science and engineering from JNTU Kakinada, Andhra Pradesh, India, and the Ph.D. degree from the National Institute of Technology Warangal, Telangana, India. She is an Assistant Professor (Sr. Gd-1) with the School of Computer Science and Engineering, VIT-AP University, India. She has about 18 years of teaching experience. She has published more than 20 research papers in refereed international journals and conferences. Her research interests include machine learning, deep learning, and image processing. She has received several best paper awards for her research contributions in international conferences. She actively participates in research related to AI and image processing.