Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 10, No. 4, pp. 206 –212, October-December 2022

p-ISSN: 2663-628X
e-ISSN: 2663-6298

ANALYSIS AND CLASSIFICATION OF AUTISM DATA USING MACHINE LEARNING ALGORITHMS

Sulav Adil Taher [a,*], and Masoud Muhammed Hassan [b]

[a] Dept. of Statistics, University of Duhok, Duhok, Kurdistan Region, Iraq – (sulav.adeel@gmail.com)
[b] Dept. of Computer Science, University of Zakho, Duhok, 42001, Kurdistan Region, Iraq

Received: *12 Oct. 2022* / ***Accepted:*** *17 Oct. 2022* / ***Published:*** *07 Nov., 2022* ***https://doi.org/10.25271/sjuoz.2022.10.4.1036***

ABSTRACT:

Autism is a neurodevelopmental disorder that affects children worldwide between the ages of 2 and 8 years. Children with autism have communication and social difficulties, and the current standardized clinical diagnosis of autism still relies on behaviour-based tests. The rapidly growing number of autistic patients in the Kurdistan Region of Iraq necessitates. However, such data are scarce, making extensive evaluations of autism screening procedures more difficult. For this purpose, the use of machine learning algorithms for this disease to assist health practitioners if formal clinical diagnosis should be pursued was investigated. Data from 515 patients were collected in Dohuk city related to autism screening for young children. Three classification algorithms, namely (DT, KNN, and ANN) were applied to diagnose and predict autism using various rating scales. Before applying the above classifiers, the newly obtained data set was in different ways undergo data reprocessing. Since our data is unbalanced with high dimensionality, we suggest combining SMOTE (Synthetic Minority Hyper sampling Technique) and PCA (Primary Component Analysis) to improve the performance of classification models. Experimental results showed that the combination of PCA and SMOTE methods improved classification performance. Moreover, ANN exceeded the other models in terms of accuracy and F1 score, suggesting that these classification methods could be used to diagnose autism in the future.

KEYWORDS: Autism, Supervised Learning, Classification, Artificial Neural Networks, PCA, SMOTE.

1. INTRODUCTION

Autism, also known as an autism spectrum disorder (ASD), is a developmental condition that manifests itself in early childhood. It may affect social, communication, relationship, and self-regulation abilities [1]. It usually takes within the first three years and lasts the remainder of the patient's life. The underlying cause of autism is uncertain, however, it is thought to be genetic, stress, inflammation, toxins, alcohol, air pollution, autoimmune illnesses, the child's emotional condition, as well as environmental factors during pregnancy and the first two years after birth [2]. There are four levels of autism (mild, moderate, severe, and very severe). Although autism among children has rapidly spread in recent years, there is not enough research addressing the causes and effects of this disease [3]. In this paper, we studied and analyzed autism data among children affected in the Dohuk governorate using Machine Learning (ML) methods. ML, as a branch of artificial intelligence, is the process of applying computers to real-world problems, to better analyze, train, and model data, and hence perform faster and provide better predictions [4]. For this purpose, a new dataset of autistic children in Duhok was created by collecting data from 515 cases in different centers. The primary objective of this study is to discover the factors that greatly affect the child's infection with this disease and to prevent other children from infection. Different ML algorithms and techniques are used to analyze data and predict outputsfrom specified inputs to predict and classify autistic levels.

ML algorithms are categorized into two main types, supervised (with labeled data) and unsupervised (unlabeled data) [4]. Classification, which is the main core of this study, is one of the most popular methods of supervised learning, which is based on predicting the output from a set of input variables [5]. Common classification algorithms are Decision Trees (DT), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) [6]. Unsupervised learning is used to analyze and aggregate unlabeled datasets to discover hidden patterns of data without having the output label [6]. Clustering (such as K-means) and dimensionality reduction (such as PCA), are the most common algorithms of unsupervised learning. In this paper, three classification algorithms (DT, KNN, ANN) are used along with different data preprocessing methods. As our data is highly dimensional and the classes are imbalanced, we applied PCA for feature extraction and SMOTE for rebalancing. The main contributions of this paper are as follows:

- To the best of our knowledge, this is the first study to address the usage of ML models to analysis the autism data in Duhok, Kurdistan region of Iraq.
- A new dataset of 515 autistic children in the Duhok is created by collecting real data from current cases.
- Three different classification algorithms (DT, KNN, ANN) are used to classify the autism data using different evaluation metrics.
- PCA and SMOTE methods were combined for reducing the dimensionality of data and balancing, before applying classifiers.

The remainder of the paper is organized as follows. In section 2, some relevant papers that used ML methods for Autism data are reviewed. Section 3 outlines the background of the topic under study, and the methodology. In section 4, we presented the main obtained results using different ML methods with different experiments. Finally, the study's primary conclusions are drawn in Section 5, along with some directions for future works.

* Corresponding author

2. RELATED WORK

Despite the global prevalence of autism in recent years, there is a limited number of research on this disease, and despite the usage of ML algorithms in many other fields, there are few in the field of autism. We have retrieved some examples of how ML algorithms have been used in the field of autism disease.

Rahman et al. [7] provided a review study on machine learning methods using feature selection and classification for ASD. In this study, they showed issues concerning autism, using ML algorithms. The emphasis was on selecting the optimal features of autism data and improving classification while maintaining high accuracy. Therefore, by reducing data dimensionality and choosing the appropriate and essential features, their proposed method provided promising results in diagnosing ASD [7].

Zheng et al. [8] conducted another research on autism classification based on Logistic Regression (LR) model. In this work, they applied the LR model in the ASD diagnosis process along with the data feature engineering, and model training and testing. The authors concluded that their proposed model revealed that ML-based methods have the potential to help ASD diagnosis in practice [8].

Niu et al. [9] proposed a multichannel Deep Attention Neural Networks (DANN) for the classification of ASD using neuroimaging and personal characteristic data. They developed DANN by applying the state-of-the-art attention mechanism based on DL techniques for the automated diagnosis of ASD. They also used k-fold cross-validation (CV), and their experiments showed that their proposed model achieved an accuracy of (0.73), outperforming multiple peer ML models. They concluded that the leave-one-out CV experiments showed promise for the proposed model when applied to clinical data with unseen variations. Experiments using varying combinations of data modalities demonstrated the discriminative power of individual data modalities such as brain functional connectome and principal component data [9].

Arya et al. [10] conducted a study on Fusing Structural and Functional MRIs using Graph Convolutional Networks (GCNs) for autism classification. They utilized relational information from sMRI data as compared to phenotypic data together with fMRI data for autism classification using GCNs. The authors concluded that replacing the atlases with brain summaries makes the model more robust for new sites with the best-case improvement exceeding 18%. Unlike the previous works, they showed that the model can perform well even without subjectively picking samples from the full dataset. This implies that their model generalized well under scenarios of higher noise levels [10].

Raj et al. [11] used different ML and DL approaches for detecting and analysing ASD data. They used Naïve Bayes (NB), Support Vector Machines (SVM), LR, KNN, ANN, and Convolutional Neural Networks (CNN) for classification. Different performance evaluation measures were utilized to predict and analyze ASD data for children, adolescents, and adults. Three age groups were employed in non-clinical data collection. Their experimental results showed that SVM and CNN models have the same prediction accuracy of around 98.3% for the ASD Child dataset. The CNN model, on the other hand, was able to reach the greatest accuracy of 95% for the remaining two datasets, indicating that the CNN-based method is the best model for detecting ASD [11].

Abdullah et al. [12] also applied ML algorithms with LASSO regression for ASD Classification. Different methods were also used for selecting the most important features, and three classifiers (Random Forest (RF), LR, and K-NN) were compared and validated by K-fold CV. Their results showed that logistic regression had a maximum accuracy of 97.5%.

Tejwani et al. [13] proposed a method for autism classification using brain functional connectivity dynamics and machine learning methods. They used the temporal variability of functional connectivity for ASD classification in a large, multi-site, resting-state fMRI dataset. They concluded that ML models trained on brain region variability can yield up to 62% accuracy, which is comparable with classification accuracy obtained with static connectivity measures such as node strength [13].

3. BACKGROUND

This section provides general information about autism and the background of the machine learning methods used in this study, including data preprocessing, normalization, and the three classification algorithms used.

3.1 Autism

Autism is characterized by communication problems in the brain's remote and local networks. These networks influence brain development, and this illness manifests itself at an early age, with diagnoses ranging from 24 to 36 months [1]. Autistic children are classified into four categories: mild, moderate, severe, and very severe. They are all distinct and have varied behaviors, yet they all have an impact on social cohesiveness. Even when they are with their family, some autistic children prefer to be alone or far away, and some of them engage in repetitive acts such as waving their hands often and other strange gestures, and some of them suddenly lose their language and communication abilities. A multitude of factors, including social, health, psychological, genetic, and environmental influences on the growing brain, produce this condition [2], [10]. Autism manifests itself in a variety of ways. The patient, for example, appears to be extremely bashful in comparison to others and wants to be alone. As a result of mental abnormalities in their brain, individuals experience hopelessness and a lack of leadership. They also throw tantrums and injure themselves, their siblings, or their friends for no apparent cause. Verbal ratings are generally lower than performance ratings [14], [15].

3.2 Autism Treatment Methods

Although there are currently no pharmaceuticals to treat this illness, there are alternative ways to relieve or treat it, and the following are some of the techniques that professionals have discovered:

- Functional Analysis*:* This method involves diagnosing the child's surroundings in order to determine the causes and consequences of the child's behavior. An expert conducts interviews with their parents and keeps track of his\her living environment. Through these interviews, the elements that affect people are identified, and the things that affect them are treated [14].
- Choosing Target Behaviors: In this strategy, each individual has distinct behaviors, such as linguistic, auditory, or aggressive issues, among other things. The specialist identifies the specific behaviors that have a substantial influence on the autistic child and then uses specialized courses to change these behaviors [14], [16].
- Teaching Procedures*:* In this method, an expert educates the target child's behavior and then decides on various approaches or procedures to be used to treat them [16].

3.3. Data Preprocessing

Before using any machine learning techniques, the data must be preprocessed. This section provides a brief overview of the preprocessing approaches used in this study.

3.3.1 Data Normalization: Because our autism data are measured at various scales; data normalization is utilized to ensure that all variables of the dataset contribute equally to model fitting. When the input data is of various dimensions and units, one method of data preparation that is often used in the field of machine learning is data normalization [17]. In this method, the data is transformed into a certain range, usually between (0, 1), or (-1, 1). In this way, a model works faster and produces less error [18]. The following data normalization methods are used:

Z-score: in this method, the input features are remeasured using the mean and standard deviation of the features [19] and the z-value of a standard distribution with a zero mean and variance of one [20]. The z-score formula is given as follows,

$$z_i = \frac{x_i - \mu}{s} \quad (1)$$

where μ and s are the mean and standard deviation of the feature x. Values of z_i will be between (-1 to 1) with mean = 0 and standard deviation = 1.

Min-Max Normalization: In this technique, the input data are rescaled using the minimum and maximum values for each feature to obtain new values, which often range between (0,1) [21].The min-max normalization formula is as follows,

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (2)$$

where x is the original feature, $\min(x)$ and $\max(x)$ are the minimum and maximum values of feature x.

3.3.2 Feature Extraction: Feature extraction is a method of extracting important features from data by reducing the dimensionality of the data to lower dimensions without losing a significant amount of information [22], [23]. There are various approaches for extracting features, Principal Component Analysis (PCA) is one of the most powerful methods in the literature, which provides effective results [24].

Principal Component Analysis (PCA): is a crucial feature extraction approach for reducing data dimensionality and extracting useful information [25]. It is regarded as the best feature extraction method since it employs original data and provides better performance for classifiers [26]. When dealing with high dimensional data, it is crucial to identify core significant features and to employ only the most important of these features. The removal of non-essential features has no effect on the accuracy performance of the findings [27]. This method was applied in this study to reduce the dimensionally of the feature with a view to improving the model performance.

3.3.3 Rebalancing Data with SMOTE: Imbalanced data happens when the number of samples in one class (called majority) is significantly greater than the other class (called minority), hence the distribution of classes is skewed [28]. In such cases, traditional ML methods often give low classification performance for unseen samples of the minority class. This is because the model tends to be strongly directed by the majority class. To tackle with this issue, which is the case in our newly created Autism dataset, imbalanced data should be balanced first using under-sampling or over-sampling methods. Resampling does over-sample by adding new cases to the minority class, or under-sample by removing existing cases from the majority class [29]. An oversampling method called SMOTE (Synthetic Minority Oversampling Technique) is employed for our datasets in this research, which is discussed below [30],[28].

SMOTE: It is one of the most common oversampling methods, which creates new synthetic samples from the existing minority class as follows.

$$x_{new} = x_i + rand(0,1) * (x_i - x_j) \quad (3)$$

where x_{new} is a new synthetic sample, x_j is a randomly chosen sample among the five nearest neighbours of x_i samples in the minority class based on the Euclidean distance, and $rand(0,1)$ is a random number between 0 and 1.

3.4 Data Description

We collected autism data and created a new dataset on children with this disease to find out what causes the most affected child with this disease and prevent other children from getting it. The dataset consisted of 515 cases, with one dependent variable (class) and 16 independent variables (age groups, gender, height, weight, number of healthy and unhealthy siblings, family financial status, etc.), and one dependent variable, autism level. The dataset includes cases distributed across four different categories of autism levels: 1, 2, 3, and 4, with the four categories denoting mild, moderate, severe, and very severe, respectively. However, autism cases were transferred into a binary class with two categories (0 and 1), where 0 represents the first two categories and 1 represents the last two categories. This data was collected over a period of nine years from (2013 to 2021). Details of these 17 features are listed in Table 2.

Table 2: Description of the features of the Autism dataset.

Variable	Name variable	Data Type	Range Value
x_1	Gender	Binary categorical	[1,2]
x_2	Address	Multi categorical	[1,4]
x_3	Age	Numerical	[2,12]
x_4	Weight	Numerical Continuous	[13,39]
x_5	Length	Numerical Continuous	[79,135]
x_6	Other diseases	Binary categorical	[1,2]
x_7	Type of other disease	Numerical Discrete	[0,5]
x_8	Family	Nominal	[1,4]
x_9	Number of siblings	Multi categorical	[0,9]
x_{10}	Unhealthy siblings	Multi categorical	[0,2]
x_{11}	Patient order in family	Multi categorical	[1,9]
x_{12}	Father education level	Multi categorical	[0,6]
x_{13}	Mother education level	Multi categorical	[0,6]
x_{14}	Father's job	Multi categorical	[0,4]
x_{15}	Mother's job	Multi categorical	[0,4]
x_{16}	Family economy	Numerical Continuous	[1,4]
y	Disease rate	Binary categorical	[0,1]

Figure 1 shows the distribution of cases across the two categories. Where x axis is the two classes (Autism level) and the y axis is the frequency (number of patients in each class). The percentage of class 0 is 31.45 (162 patients), and class 1 is 68.54% (353 patients). It is clear from figure 2 that the data is imbalanced, and therefore the SMOTE method was applied.

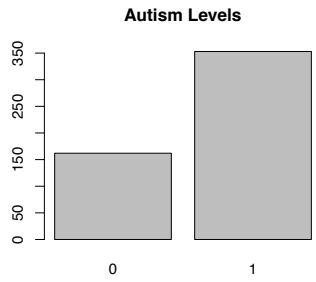

Fig. 1: Distribution of instances across classes (Autism rate).

Figure 2 shows the frequency distribution for each feature in our Autism dataset. Where x axis is the category of each variable and the y axis is the frequency (number of cases).

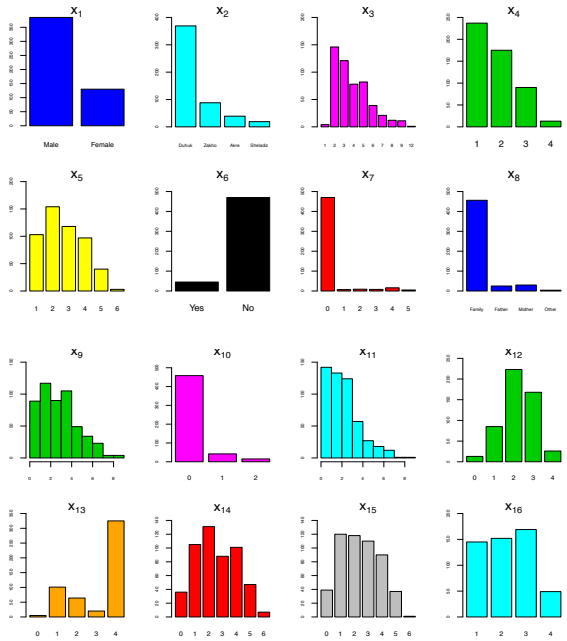

Fig. 2: Frequency distribution of each independent feature associated with the Autism dataset.

Figure 3 clearly shows the distribution of each feature. For example, the variable x_1, which represents the child's gender, shows that 74 % of the impacted cases are boys and 26 % are girls, indicating that boys are more affected than girls. Children who live in the city are influenced by 72%, while other children are affected by 28%, according to the variable x_2, which reflects the child's address. This suggests that the child's injury is influenced by the environment or place of residence. Autism does not cause another condition in the ratio of 91%, while 0.09% indicates that this disease affects a child who is infected with another disease. The variable x_{10}, which refers to the number of siblings of the affected child, shows that 89% of the affected children have no affected siblings, whilst 8% have one affected sibling, and only 3% are having more than one affected sibling.

3.5 Classification Algorithms

One of the most important tasks in supervised learning classification. It is the process of recognizing patterns, concepts, and other objects in order to better comprehend them and classify them based on incoming data [5]. Classification can help uncover abnormalities when developing a learning model from prior data [4]. There are various classification algorithms, each of which builds a prediction model in a different way. Following that, the model is developed in two stages: training and testing [5]. The performance of the classification algorithm is determined by evaluating the confusion matrix's performance [31]. The three classification algorithms used in this study are as follows.

3.5.1 Decision Tree (DT): It is a supervised learning algorithm that uses a tree structure to represent decision sets in order to classify data according to various data qualities, with each branch of the tree representing a new decision output [32]. It has a great ability to make predictive models [33]. It can be used to create guesses regarding category variable names [34]. Each branch might be relegated to the training sample category [4]. The decision tree is formulated as follows:

$$Entropy(D) = -\sum_{i=1}^{k} p_i \log(p_i) \ ,i: 1,2,3,..,K \quad (4)$$

Where $Entropy(D)$ is the amount of information required for classification ,p_i is the probability of each feature affecting the model output, and k represents the number of features.

3.5.2 K-Nearest Neighbors (KNN): It is a nonparametric supervised learning method used for classification. Analyzing a category can be one of the most important judgments when there is almost no previous information on knowledge appropriation [4]. Because K-NN stores rather than learn from the training information, it is known as a lazy learning algorithm. For classifying the new dataset, it uses the Euclidean distance to calculate the distance between the new point and the previously stored training points. The new point will be given to the class with the nearest neighbors [5]. To find the nearest neighbor, the Euclidean distance function d_i is used as follows.

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^2} \quad (5)$$

where x_i, y_i are variable for input data.

3.5.3 Artificial Neural Network (ANN*):* It is one of the most effective and precise supervised learning algorithms. It works by simulating the human brain's neuron structure, which is made up of three layers: input, hidden, and output. ANN is a sophisticated adaptive system capable of changing its internal structure in response to information it receives and dealing with nonlinear challenges. It is accomplished by adjusting the weight of the network. Each connection has a particular amount of weight attached to it. A weight, which is a number, controls the signal between two neurons. To improve the outcome, reverse the error between the actual and expected values layer by layer, adjusting the weight of each layer until convergence is obtained [35], [36]. Figure 3 represents the main structure of ANN.

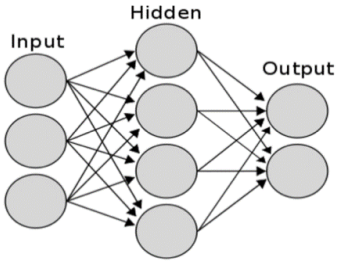

Fig. 3: Simple Model of ANN with Multilayers

3.6 Performance Evaluation

We assessed the suggested predictive model using various commonly used assessment criteria in the literature to check the efficacy of the classification algorithms. Accuracy, sensitivity, specificity, precision, and F-Score are the evaluation metrics employed. A confusion matrix is used to determine the rating scale or the accuracy of the three classification algorithms (DT, KNN, and ANN) utilized. The confusion matrix is a useful tool for comparing actual values to those predicted by the model, and it may also be used to assess the quality of classifiers from

various classes [37]. There are four variables in this matrix: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Where TP refers to cases that are predicted as positive and are actually positive, FP are cases that are predicted as negative and are actually positive, TN refers to cases that are predicted as positive but are actually negative, and FN is cases that are predicted as negative and in fact, are negative [38]. The confusion matrix is defined as follows:

Table 1. Confusion Matrix

	Predicted Class		
Actual Class		Yes	No
	Yes	**TP**	**FP**
	No	**FN**	**TN**

The evaluation metrics are calculated from the confusion matrix as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Specificity = \frac{TN}{TN+FP} \quad (7)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (8)$$

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$F_1 = 2 * \frac{Precision * Sensitivity}{Sensitivity+Precision} \quad (10)$$

4. PROPOSED METHOD

In this paper, we investigated using supervised machine learning algorithms to analyse and classify autism data in Dohuk, Kurdistan region Iraq. Different ML methods were investigated in this study, including three classifiers (DT, KNN, ANN), rescaling data by normalization to ensure that all variables contribute equally to the model fitting, extracting the important features by reducing the dimensionality of the data, and the data is balanced using SMOTE method. Figure 4 represents the proposed method as follows.

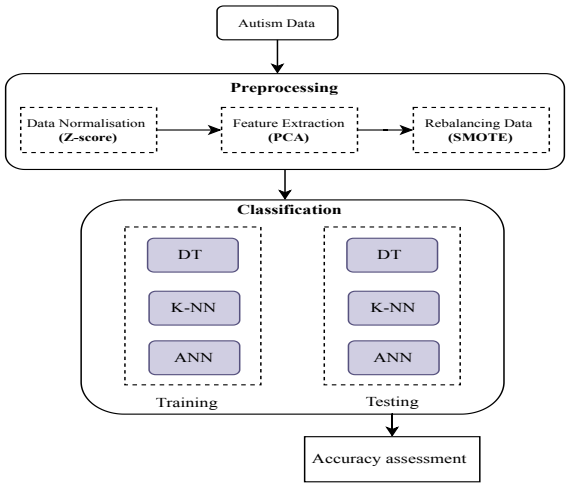

Fig. 4: Flow diagram of the proposed method.

5. RESULTS AND DISCUSSION

In this section, we present our results starting with the experimental setup, describing our data, descriptive graphs, and evaluation of classification algorithm results using original data, and preprocessed data.

5.1 Experimental Setup

Different experiments were carried out to analyze and classify autism data. Three different machine learning algorithms (KNN, DT, ANN) were used in our investigation with the following setup. In all algorithms used, the data were divided into two parts: testing and training using different test-training ratios. Based on different experiments on trial and error, we chose an 80%-20% ratio, since it gave the best performance for almost all experiments. In the KNN algorithm, based on the trial-error, the value of the nearest neighbor K is chosen to be 3, and the Euclidean distance was used as a distance metric. In the Decision Tree, the ID3 algorithm was used, which contains one of the most popular decision tree algorithms for binary classification. Entropy concepts and acquired information were used to measure how well the training samples were separated. In the ANN, different structures are used for our ANN model. The obtained results are based on the ANN model consisting of 4 layers: an input layer, 2 hidden layers, and an output layer. For the first hidden layer, we had 7 neurons, while in the second hidden layer we had 3 neurons. Backpropagation is also used with the Stochastic Gradient Descent (SGD) for optimizing our model weights, and sigmoid as an activation function. Furthermore, before applying classification algorithms, we preprocessed our data using normalization for rescaling, PCA for dimensionality reduction and SMOTE for rebalancing data. Finally, we compared the performance of all classifiers using different evaluation metrics.

5.2 Classification Results

In this section, we show the performance of the three classification algorithms used (DT, KNN, and ANN) for our dataset to classify children with autism levels. We first transformed the data using different normalization methods, and then applied the three classification algorithms. All test results are obtained from a computer equipped with a CPU i7-8550U @ 1.80 GHz 1.99 GHz, 4 GB of RAM, and a 64-bit Windows 10 operating system. All data analyses were handled using an R programming language. The data was also divided into two parts, the test and training section.

Table 3 reports results based on different evaluation metrics: Accuracy (Acc.), Sensitivity (Sen.), Precision (Pre.), and F_1 score for the three classification algorithms used.

Table 3. Comparing the performance of the six classification algorithms using the Autism dataset.

	Measurements				
Data Type	**Algorithms**	**Acc.**	**Sen.**	**Pre.**	**F_1**
Original	DT	78.4	47.58	71.08	57.0
	KNN	71.84	34.21	76.47	47.27
	ANN	92.96	85.48	90.60	87.96
Normalized	DT	78.43	47.54	71.09	56.97
	KNN	95.15	86.84	99.9	92.91
	ANN	93.84	87.25	90.12	88.66
Normalized + PCA	DT	62.47	62.5	64.7	63.4
	KNN	99.03	98.39	98.39	98.40
	ANN	100	100	100	100
Normalized + PCA + SMOTE	DT	81.99	77.84	86.45	81.92
	KNN	94.71	99.72	91.03	91.18
	ANN	98.94	92.7	98.54	95.53

For the original data used, results in Table 3 show that the highest accuracy was obtained in the ANN classifier with an accuracy of 92.26%, followed by the DT algorithm with an accuracy of 78.4%, but the lowest accuracy was obtained in KNN with an accuracy of 71.84%. Similarly, the ANN had the highest sensitivity with 85.48% followed by the DT with 47.58%, and the lowest sensitivity was obtained by KNN with 34.21%. In terms of precision, ANN had the best precision with 90.60%, followed by KNN with 76.47%, while DT had the lowest precision with 71.08%. In terms of the F_1 score, which is

especially valuable for unbalanced data, the ANN classifier obtained an F_1 score of 87.96%, outperforming DT and KNN, with 57.0% and 47.27, respectively.

For the normalized data in Table 3, the highest accuracy results were obtained in KNN with an accuracy of 95.15%, followed by ANN with an accuracy of 93.84%, but the lowest accuracy was obtained in DT with 78.43%. This indicates that the data normalization has improved the model performances. The highest sensitivity results were obtained in ANN with 87.25%, followed by KNN with 86.84%, but the lowest sensitivity was obtained in DT with 47.54%. The highest precision was obtained in KNN with 99.9%, followed by ANN with 90.12%, while the lowest precision was obtained in DT with 71.09%. In the same token, the highest F_1 score for normalized data was obtained in KNN with 92.1% (which shows a significant improvement), followed by ANN with 88.66%, and the lowest F_1 score was obtained in DT with 56.97%.

On the other hand, when PCA was applied to the data, the performances of the classifiers were significantly improved. The highest results were obtained in the ANN with 100% of accuracy, sensitivity, precision, and F_1, but the lowest accuracy was obtained in the DT at 62%. The results of the KNN classifier were also improved significantly, with accuracy, sensitivity, precision, and F_1 of 95.15%, 98.39%, 98.39%, and 98.44, respectively. However, the results of the DT classifier were not promising with the PCA method.

Results presented in Table 3 also show that when the SMOTE method was applied with classification algorithms, there was a significant improvement in the results, especially for the F_1 score. The highest F_1 was obtained in the ANN with 95.53%, while the lowest F_1 was obtained in the DT with 81.92%. Similarly, the highest accuracy results were obtained in ANN with an accuracy rate of 98.94%, but the lowest accuracy was obtained in DT with an accuracy rate of 81.99%. The highest sensitivity results were obtained in KNN with 99.72%, and the lowest sensitivity was obtained in DT with 77.84%. The highest precision results were obtained in ANN with 98.54%, while the lowest precision was obtained in DT with 86.45%. This indicates that when using SMOTE, the classification performances were improved.

Figure 5 displays the accuracy of each classification algorithm for our autism dataset using original data, normalized data, (normalized + PCA), and (normalized + PCA + SMOTE).

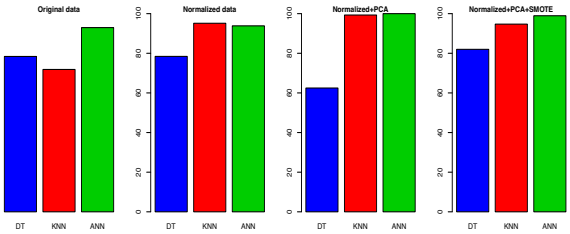

Fig.5: Classification accuracy for Autism dataset using original data and normalized data and PCA,SMOTE.

Figure 3 shows that the DT algorithm has an improvement with preprocessing methods used. It has the same accuracy of 78.4% when using original and normalized data, but it was reduced to 62.5% using PCA, and improved to 81.99% when using SMOTE method. On the other hand, the accuracy of the KNN algorithm was 71.84% when using the original data, but it improved to 95.15% when the data were normalized, and further improved to 98.39% and 94.71% using PCA and SMOTE. Similarly, the accuracy of the ANN algorithm was 92.96% with the original data, but it increased to 93.84% when the data used was normalized, and much improved to 100% when using PCA, and also improved to 98.94% when using SMOTE. Therefore, by comparing the results of Table 3 and Figure 4, we can clearly see that there is a significant improvement in the performance of the classification models, especially for the KNN and ANN algorithms. As a result, it appears that the data preprocessing methods (rescaling with normalization, feature extraction with PCA, and rebalancing with SMOTE) were powerful to achieve better results for the classification algorithms. Therefore, we conclude that our proposed method (Normalization + PCA + SMOTE) was effective.

6. CONCLUSION

The growing number of autism cases worldwide, as well as the economic and sociological implications of this disease, underline the critical need for developing simple and accurate diagnostic technologies for this disease. For this purpose, in this paper, we investigated using supervised machine learning algorithms to analysis and classify autism data in Dohuk, Kurdistan region Iraq. Different ML methods were investigated in this study, including three classifiers (DT, KNN, and ANN), rescaling with normalization, dimensionality reduction with PCA, and oversampling with SMOTE method. Experimental results showed that the performance of the classification algorithms varies based on the data preprocessing (normalization, PCA, and SMOTE) methods used. However, the best classification results for the original data were obtained with the ANN classifier, which had very good results relative to an accuracy of 92.96% and an F_1 score of 87.96% compared to the other two classifiers. On the other hand, when using data normalization, PCA, and SMOTE methods, the performances of the classification models are greatly improved with an accuracy of 98.94% and an F_1 score of 95.53%. Furthermore, the experimental results on our newly created autism dataset showed that the ANN classifier outperformed others in terms of accuracy and F_1 score. Therefore, we conclude that these algorithms can be used by physicians to diagnose autism. In future works, we will apply deep neural networks to further check the performance of our dataset.

REFERENCES

[1] W. Jamal, S. Das, I. A. Oprescu, K. Maharatna, F. Apicella, and F. Sicca, “Classification of autism spectrum disorder using supervised learning of brain connectivity measures extracted from synchrostates,” J. Neural Eng., vol. 11, no. 4, pp. 1–27, 2014, doi: 10.1088/1741-2560/11/4/046019.

[2] P. Adak, S. Sinha, and N. Banerjee, “An Association Study of Gamma-Aminobutyric Acid Type A Receptor Variants and Susceptibility to Autism Spectrum Disorders,” J. Autism Dev. Disord., vol. 51, no. 11, pp. 4043–4053, 2021, doi: 10.1007/s10803-020-04865-x.

[3] C. Chen, L. Geng, and S. Zhou, “Design and implementation of bank CRM system based on decision tree algorithm,” Neural Comput. Appl., vol. 33, no. 14, pp. 8237–8247, 2021, doi: 10.1007/s00521-020-04959-8.

[4] G. A. A. MULLA, Y. DEMİR, and M. HASSAN, “Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data,” Bitlis Eren Üniversitesi Fen Bilim. Derg., vol. 10, no. 3, pp. 858–869, 2021, doi: 10.17798/bitlisfen.939733.

[5] E. M. Senan et al., “Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques,” J. Healthc. Eng., vol. 2021, 2021, doi: 10.1155/2021/1004767.

[6] M. M. Hassan, N. Njmh Amiri, M. Muhammed Hassan, and N. Amiri, “Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms Bayesian Deep Learning View project Technical SCIENCE View project Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms,” no. October, 2019, [Online]. Available: https://www.researchgate.net/publication/336672231.

[7] M. M. Rahman, O. L. Usman, R. C. Muniyandi, S. Sahran, S. Mohamed, and R. A. Razak, "A review of machine learning methods of feature selection and classification for autism spectrum disorder," Brain Sci., vol. 10, no. 12, pp. 1–23, 2020, doi: 10.3390/brainsci10120949.
[8] Y. Zheng, T. Deng, and Y. Wang, "Autism Classification Based on Logistic Regression Model," 2021 IEEE 2nd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. ICBAIE 2021, no. Icbaie, pp. 579–582, 2021, doi: 10.1109/ICBAIE52039.2021.9389914.
[9] K. Niu et al., "Multichannel Deep Attention Neural Networks for the Classification of Autism Spectrum Disorder Using Neuroimaging and Personal Characteristic Data," Complexity, vol. 2020, 2020, doi: 10.1155/2020/1357853.
[10] D. Arya et al., "Fusing Structural and Functional MRIs using Graph Convolutional Networks for Autism Classification," Proc. Mach. Learn. Res., vol. 121, pp. 1–17, 2020, [Online]. Available: https://proceedings.mlr.press/v121/arya20a.html.
[11] S. Raj and S. Masood, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," Procedia Comput. Sci., vol. 167, no. 2019, pp. 994–1004, 2020, doi: 10.1016/j.procs.2020.03.399.
[12] A. A. Abdullah, S. Rijal, and S. R. Dash, "Evaluation on Machine Learning Algorithms for Classification of Autism Spectrum Disorder (ASD)," J. Phys. Conf. Ser., vol. 1372, no. 1, 2019, doi: 10.1088/1742-6596/1372/1/012052.
[13] R. Tejwani, A. Liska, H. You, J. Reinen, and P. Das, "Autism Classification Using Brain Functional Connectivity Dynamics and Machine Learning," no. December 2017, 2017, [Online]. Available: http://arxiv.org/abs/1712.08041.
[14] S. J. Rogers et al., "A Multisite Randomized Controlled Trial Comparing the Effects of Intervention Intensity and Intervention Style on Outcomes for Young Children With Autism," J. Am. Acad. Child Adolesc. Psychiatry, vol. 60, no. 6, pp. 710–722, 2021, doi: 10.1016/j.jaac.2020.06.013.
[15] J. B. McCauley, R. Elias, and C. Lord, "Trajectories of co-occurring psychopathology symptoms in autism from late childhood to adulthood," Dev. Psychopathol., vol. 32, no. 4, pp. 1287–1302, 2020, doi: 10.1017/S0954579420000826.
[16] M. L. Matson, S. Mahan, and J. L. Matson, "Parent training: A review of methods for children with autism spectrum disorders," Res. Autism Spectr. Disord., vol. 3, no. 4, pp. 868–875, 2009, doi: 10.1016/j.rasd.2009.02.003.
[17] X. Wei, L. Zhang, H. Q. Yang, L. Zhang, and Y. P. Yao, "Machine learning for pore-water pressure time-series prediction: Application of recurrent neural networks," Geosci. Front., vol. 12, no. 1, pp. 453–467, 2021, doi: 10.1016/j.gsf.2020.04.011.
[18] O. D. Madeeh and H. S. Abdullah, "An Efficient Prediction Model based on Machine Learning Techniques for Prediction of the Stock Market," J. Phys. Conf. Ser., vol. 1804, no. 1, 2021, doi: 10.1088/1742-6596/1804/1/012008.
[19] M. M. Hassan, N. Njmh Amiri, M. Muhammed Hassan, and N. Amiri, "c Technical SCIENCE ⚙ View project Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms," no. October 2019, [Online]. Available: https://www.researchgate.net/publication/336672231.
[20] A. B. Al-Ghamdi, S. Kamel, and M. Khayyat, "Evaluation of Artificial Neural Networks Performance Using Various Normalization Methods for Water Demand Forecasting," Proc. - 2021 IEEE 4th Natl. Comput. Coll. Conf. NCCC 2021, 2021, doi: 10.1109/NCCC49330.2021.9428856.
[21] K. H. Abdulkareem et al., "Realizing an Effective COVID-19 Diagnosis System Based on Machine Learning and IoT in Smart Hospital Environment," IEEE Internet Things J., vol. 8, no. 21, pp. 15919–15928, 2021, doi: 10.1109/JIOT.2021.3050775.
[22] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," J. Appl. Sci. Technol. Trends, vol. 1, no. 2, pp. 56–70, 2020, doi: 10.38094/jastt1224.
[23] K. Kaplan, Y. Kaya, M. Kuncan, M. R. Minaz, and H. M. Ertunç, "An improved feature extraction method using texture analysis with LBP for bearing fault diagnosis," Appl. Soft Comput. J., vol. 87, 2020, doi: 10.1016/j.asoc.2019.106019.
[24] C. L. Chowdhary and D. P. Acharjya, "Segmentation and Feature Extraction in Medical Imaging: A Systematic Review," Procedia Comput. Sci., vol. 167, no. 2019, pp. 26–36, 2020, doi: 10.1016/j.procs.2020.03.179.
[25] M. Asadur Rahman, M. Foisal Hossain, M. Hossain, and R. Ahmmed, "Employing PCA and t-statistical approach for feature extraction and classification of emotion from multichannel EEG signal," Egypt. Informatics J., vol. 21, no. 1, pp. 23–35, 2020, doi: 10.1016/j.eij.2019.10.002.
[26] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Informatics Med. Unlocked, vol. 19, p. 100330, 2020, doi: 10.1016/j.imu.2020.100330.
[27] M. Pouyap, L. Bitjoka, E. Mfoumou, and D. Toko, "Improved Bearing Fault Diagnosis by Feature Extraction Based on GLCM, Fusion of Selection Methods, and Multiclass-Naïve Bayes Classification," J. Signal Inf. Process., vol. 12, no. 04, pp. 71–85, 2021, doi: 10.4236/jsip.2021.124004.
[28] A. J. Mohammed, "Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE
Method," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 3, pp. 3161–3172, 2020, doi: 10.30534/ijatcse/2020/104932020.
[29] M. M. Hassan, A. S. Eesa, A. J. Mohammed, and W. K. Arabo, "Oversampling method based on gaussian distribution and K-means clustering," Comput. Mater. Contin., vol. 69, no. 1, pp. 451–469, 2021, doi: 10.32604/cmc.2021.018280.
[30] A. D. Sappa and F. Dornaika, An Edge-Based Approach to Motion Detection Conference, vol. 11538, no. May. 2019.
[31] A. A. Salih and A. M. Abdulazeez, "Evaluation of Classification Algorithms for Intrusion Detection System: A Review," J. Soft Comput. Data Min., vol. 02, no. 01, pp. 31–40, 2021, doi: 10.30880/jscdm.2021.02.01.004.
[32] G. Chen and J. Chen, "A novel wrapper method for feature selection and its applications," Neurocomputing, vol. 159, no. 1, pp. 219–226, 2015, doi: 10.1016/j.neucom.2015.01.070.
[33] A. Ghosh and R. Maiti, "Soil erosion susceptibility assessment using logistic regression, decision tree and random forest: study on the Mayurakshi river basin of Eastern India," Environ. Earth Sci., vol. 80, no. 8, pp. 1–16, 2021, doi: 10.1007/s12665-021-09631-5.
[34] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.
[35] Z. Yang, C. Kong, Y. Wang, X. Rong, and L. Wei, "Fault diagnosis of mine asynchronous motor based on MEEMD energy entropy and ANN," Comput. Electr. Eng., vol. 92, no. March, p. 107070, 2021, doi: 10.1016/j.compeleceng.2021.107070.
[36] R. Bala and D. Kumar, "Classification Using ANN: A Review," Int. J. Comput. Intell. Res., vol. 13, no. 7, pp. 1811–1820, 2017, [Online]. Available: http://www.ripublication.com.
[37] N. Yulias et al., "JurnalMantik," vol. 4, no. 4, pp. 2599–2603, 2021.
[38] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," Int. J. Inf. Technol., vol. 13, no. 4, pp. 1503–1511, 2021, doi: 10.1007/s41870-020-00430-y.