EAI Endorsed Transactions
AI and Robotics
Research Article

Imbalanced Multiclass Medical Data Classification based on Learning Automata and Neural Network

Masoumeh Soleimani[1,*], Zahra Forouzanfar [2], Morteza Soltani[3], Majid Jafari Harandi[4]

[1]Department of Mathematics and Statistical Sciences, Clemson University, SC, USA
[2]Department of Computer Engineering, Dolatabad Branch, Islamic Azad University, Isfahan, Iran
[3]Department of Industrial Engineering, Clemson University, SC, USA
[4]Department of Mechanical Engineering, Khomeinishahr Branch, Islamic Azad University, Khomeinishahr, Iran

Abstract

Data classification in the real world is often faced with the challenge of data imbalance, where there is a significant difference in the number of instances among different classes. Dealing with imbalanced data is recognized as a challenging problem in data mining, as it involves identifying minority-class data with a high number of errors. Therefore, the selection of unique and appropriate features for classifying data with smaller classes poses a fundamental challenge in this research. Nowadays, due to the widespread presence of imbalanced medical data in many real-world problems, the processing of such data has gained attention from researchers. The objective of this research is to propose a method for classifying imbalanced medical data. In this paper, the hypothyroidism dataset from the UCI repository is used. In the feature selection stage, a support vector machine algorithm is used as a cost function, and the wrapper algorithm is employed as a search strategy to achieve an optimal subset of features. The proposed method achieves high accuracy, reaching 99.6% accuracy for data classification through the optimization of a neural network using learning automata.

Received on 04 July 2023; accepted on 21 July 2023; published on 24 July 2023

Keywords: Classification, Imbalanced Data, Neural Network, Learning Automata

doi:10.4108/airo.3526

1. Introduction

Classification of imbalanced data has gained significant attention due to its importance in fields such as medicine [1], robotics [2], and graph theory [3, 4]. A dataset is considered imbalanced if the distribution of classes or labels in the dataset is skewed or uneven. This class imbalance poses challenges to conventional classification algorithms, which are typically designed to work well with balanced datasets. Imbalanced data can lead to skewed decision boundaries and biased predictions, as classifiers tend to favor the majority class, resulting in a weak performance in minority-class instances. Imbalanced binary classification problems involve two classes such that instances of a class is significantly lower than the instances of the other class. The positive class, also known as the minority class, has fewer instances, while the negative class, or majority class, has a larger number of instances. Classical learning algorithms on imbalanced datasets tend to produce relatively optimal results as their final goal is to maximize accuracy. However, these algorithms often have lower performance on the minority class, which is typically the most important class in imbalanced data. The importance of minority instances is much higher compared to majority instances. Consequently, these algorithms are not suitable for processing imbalanced data [5]. In the past two decades, the rapid growth of data in the field of medicine has posed fundamental challenges [6]. With the advancements in biotechnology and data analysis methods, data mining techniques have been evolving. Nowadays, data mining and machine learning algorithms are utilized for analyzing and predicting large medical datasets. The primary goal of analyzing such data is to develop various prediction models in the medical domain. Disease diagnosis

*Corresponding author. Email: msoleim@.clemson.edu

is a significant focus in current research. The key feature of disease diagnosis models is quick decision-making by medical professionals and reducing errors in diagnosis while efficiently processing medical data [7]. Additionally, since medical data consists of a large number of features, researchers only consider subsets of these features, which are crucial for disease diagnosis. Therefore, the examination of these important features needs to be highly accurate [8]. Furthermore, attention must be given to handling minority instances. In this paper, a combination of a feature selection algorithm and an optimized neural network algorithm is used for classifying the instances.

In previous classification methods, data is assigned to the maximum class, resulting in high accuracy but lacking reliability. In this research, new combined methods at the data level will be introduced to address this issue. These methods aim to balance the imbalanced data distribution and improve the performance of classifiers when applied to such data.

The paper is structured as follows: In section 2, a review of previous works is presented. The third section introduces the proposed method and its associated components, detailing the functioning of the proposed method. The fourth section focuses on simulation and evaluation of the results. Finally, in the last section, a final conclusion of the conducted research is provided, along with future work and prospects.

2. Research Background

Numerous research studies have been conducted in the field of the textile industry, and we will now briefly review some of them. Yadao et al. (2019) focused on thyroid diseases. Data was gathered from GitHub, and a variety of data mining techniques, such as boosting, bagging, and ensemble methods, were employed. They achieved a ROC of 98.80% using the ensemble model [9].

Ahmad et al. [10] investigated thyroid diseases using classification data mining techniques. authors analyzed a dataset in three stages. In the initial phase, they minimized gain information, redundancy, and computational time. The second stage involved addressing missing values using the K-Nearest Neighbors (KNN) algorithm. Ultimately, in the final phase, they utilized a neuro-fuzzy inference system to establish an input-output relationship and eliminate the inputs, resulting in an impressive classification accuracy of 99.1%.

In [11], an approach proposed is evaluated on the data set that may be useful for imbalanced data sets.

Some papers examined the accuracy of distinguishing benign and malignant breast cancer through the utilization of diverse data mining techniques. [12]. After evaluations, the highest accuracy of 97.37% was observed [13].

Shanker et al. [14] investigated thyroid diseases. The data was gathered from the UCI repository, and various data mining algorithms. Following evaluations, a remarkable accuracy of 98.65% was attained. Samathe et al. [15] focused on analyzing a medical dataset using a vast array of machine learning algorithms. They employed EM and J48 algorithms and obtained an accuracy of 98.25% for thyroid diseases. These researchers specifically concentrated on thyroid dysfunction and the preservation of metabolic temperature for diagnosis. Paper [16] explored a medical dataset and constructed an intricate framework for prediction. They then obtained accuracy and sensitivity using numerous models with the training set. These researchers utilized support vector machines and neural networks as the complex classification framework. Prawin et al. [17] focused on predicting skin diseases using various data mining algorithms. They collected data and utilized Multi support vector machine, KNN, and simple Bayesian algorithms to achieve optimal performance. Ultimately, the highest accuracy of 97.4% was obtained for the Multi support vector machine algorithm. Wang et al. [18] investigated a medical dataset. They obtained the data from a hospital and utilized multiple-instance learning (MIL), convolutional neural networks (CNN), and EM algorithm for data mining. After evaluations, the highest accuracy of 93% was achieved. Some other studies examined thyroid disorders using a decision tree. The researchers examined challenges related to minority classes in the decision tree by employing the CART algorithm and obtained favorable outcomes in terms of the decision tree [19, 20].

Authors in [21] examined a medical dataset of thyroid disorders and utilized various classifiers, including Naive Bayes, decision tree, MLP, and RBF network. All classifiers were executed, and different results were obtained for classification accuracy, recall, false positive rate, true positive rate, F-measure, and ROC. Some others investigated the thyroid dataset using SVM, REP, and a combined framework model for classification and prediction. They achieved good results with this combined system (SVM, REP) in terms of accuracy and compared it to the decision tree, which assists in the diagnosis of thyroid diseases [22]. Gata et al. [23] examined a medical dataset of thyroid disorders and used a multivariate Bayesian algorithm with 7,200 data points and 21 attributes. They achieved the best accuracy of 97.77% for thyroid hypothyroidism. The predictive model helps physicians in diagnosing the disease. [24] presents a method to classify the imbalanced data. did research on drug-related diseases. They obtained the data from the skin and V.D unit of a Hospital in India and used two data mining techniques, support vector machine (SVM), and neural network. They achieved the highest

accuracy of 97.17%. Researchers also focused on the analysis of thyroid diseases. They collected the data from the UCI repository and used decision tree, k-nearest neighbors, and Naive Bayes as data mining techniques. After reducing the dataset, they achieved an accuracy of 97.53% for the three classifiers [25]. Razia et al. examined a classifier for thyroid disease. Diagnosing thyroid disease is challenging, but these researchers easily reached the desired point through a neural network model. The neural network assists in identifying disease attributes in a straightforward manner [26]. Gaykwad et al. conducted an empirical study of the rotating forest and achieved better performance in terms of classification accuracy. They analyzed misclassified data and the wrapper feature in various ways, obtaining the best result of 99.63% accuracy for the hypothyroidism dataset. Additionally, they evaluated the effectiveness of the rotating forest algorithm in diagnosing thyroid disorders [27].

3. Proposed Method

The first step in conducting this research is to access a standard dataset. For this purpose, the UCI standard dataset has been utilized. Then, the necessary preprocessing is performed on the dataset. Considering the large volume of the dataset, feature selection algorithms are used to reduce their dimensionality and select the best subsets for medical data classification [28]. The support vector machine (SVM) algorithm is used as the cost function, and the Fuzzy algorithm is employed as the search strategy in the feature selection stage. Finally, the optimized neural network algorithm is employed for the classification of samples. An illustration of the architecture of the proposed method is presented in Figure **??**.

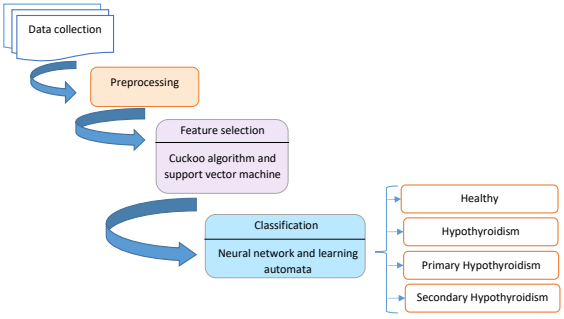

Figure 1. Made by REZA.

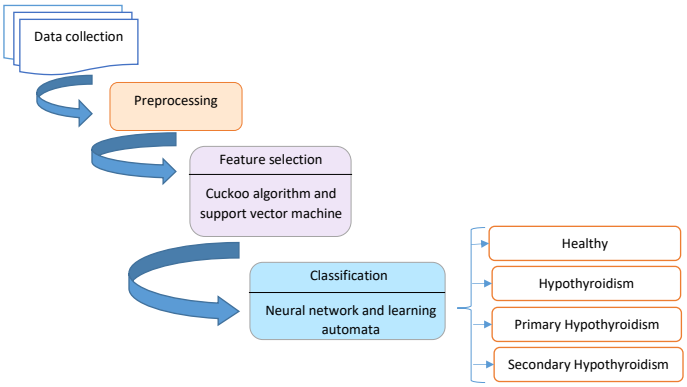

Figure 2. The process of the proposed method

3.1. Dataset

The dataset used for experimental purposes was downloaded from the University of California Irvine (UCI) repository [29]. This dataset consists of 3772 samples, with 3481 samples belonging to the healthy group, 194 samples belonging to the hypothyroidism category, 95 samples belonging to the primary hypothyroidism category, and 2 cases belonging to the secondary hypothyroidism category. The last feature of the dataset corresponds to the class of each sample, hence there are 29 features used for data classification. The dataset features are shown in Table 1.

3.2. Preprocessing

Superior data quality yields exceptional outcomes and diminishes data mining expenses. The data available in the database may contain duplicate entries; therefore, data cleaning is performed by examining the data. A combination of data is used to eliminate missing values and remove redundant data, as the presence of low-quality and redundant data can render the analysis ineffective.

3.3. Feature Selection in the Proposed Method

Due to the need for examining a large volume of high-dimensional medical data in data classification systems, processing each sample with all features may be time-consuming or even infeasible. Dimensionality reduction for such data can improve the efficiency of malware detection systems in terms of speed and accuracy. Therefore, selecting the best features that represent all the data and eliminate redundant features is a crucial aspect of these systems. In the feature selection stage of the proposed method, the support vector machine algorithm is used as the cost function, and the cuckoo algorithm is employed as the optimization strategy. Figure **??** illustrates the flow of feature selection using the cuckoo algorithm

Table 1. Dataset Features

1	Age	2	Query Hyperthyroid	3	TT4 measured
4	Sex	5	Lithium	6	TT4
7	On thyroxine	8	Goiter	9	T4U measured
10	Query on thyroxine	11	Tumor	12	T4U
13	On antythyroid	14	Hypopitutory	15	FTI Measured
16	Sick	17	Psych	18	FTI
19	Pregnant	20	Tsh measured	21	TBG Measured
22	Thyroid surgery	23	TSH	24	TBG
25	T131 treatment	26	T3 measured	27	Referal source
28	Query Hypothyroid	29	T3	30	Class

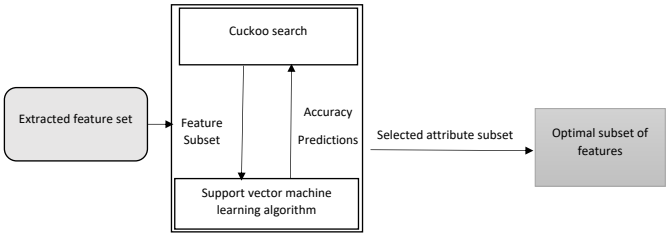

Figure 3. Flow of Feature Selection Based on Wrapper

and support vector machine based on the wrapper approach.

The problem of feature selection does not have a deterministic solution in general, and so far, no exact method has been proposed to solve this problem. Metaheuristic algorithms are one example of exploratory approaches that perform well in feature selection problems. Therefore, in the proposed method, a solution is presented to improve the malware detection system by optimally selecting its features using a metaheuristic algorithm called the cuckoo Optimization Algorithm. The cuckoo Optimization Algorithm is inherently an algorithm used to maximize its objective function, meaning that the cuckoos move towards a point that offers the greatest benefit and attractiveness for their survival. The stages of the feature selection method are illustrated in Figure 4.

Displaying the residence of Cuckoo

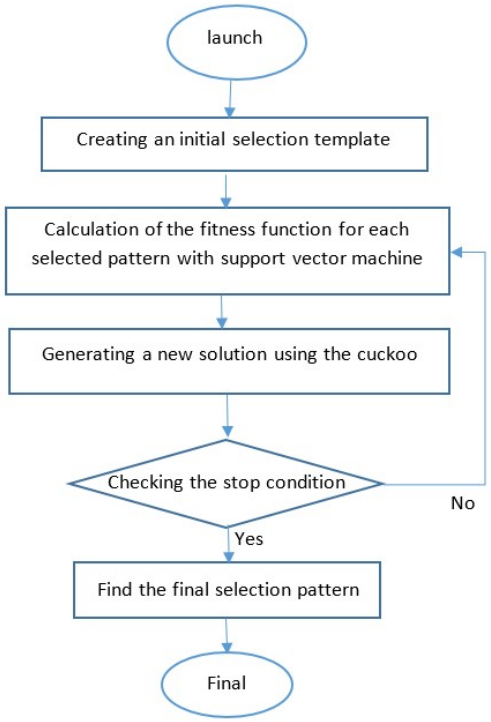

Figure 4. Feature Selection Framework

In the optimization algorithm, the residence of Cuckoo is represented by a solution vector. In N-dimensional problems, the residence is an array of size $1 \times N$, which denotes the current residence of Cuckoo. This array is expressed by the following equation:

$$habitat = (x_1, x_2, \ldots, x_n) \quad Eq1$$

In this method, each residence is represented as a binary string of numbers. When the value of a variable is

equal to 1, it indicates the selection of a feature, and if the value is 0, it indicates that the feature is not selected. Figure 5 represents a feature as a residence in the proposed method. Essentially, each residence corresponds to one of the possible subsets of features.

Classification accuracy is considered as the profit for each residence in this research. To achieve this, the Support Vector Machine (SVM) algorithm has been used.

• Finding new solutions and Leaping Flight Feature selection based on the Leaping Flight search is used to find new solutions. Some of the new solutions need to be generated by the leaping process alongside the optimal solution, which increases local search.

• Cost function The cost function of the Cuckoo optimization algorithm receives a matrix of populations or residence locations, which is essentially a subset of features. It calculates and returns the cost, which is the classification accuracy. For each row of the population matrix sent to the cost function, the cost of that row is calculated and stored in a column-wise matrix. Each row of the population matrix represents a candidate solution for the problem, which is sent to the cost function. Therefore, the cost vector has the same size as the number of rows in the population matrix that is sent to the cost function. The cost function calculates its cost using the SVM classifier function and returns it to the main program. In the genetic algorithms and particle swarm optimization algorithms, the cost function takes only one population and returns a cost for that population. Thus, the accuracy and convergence speed of Cuckoo are higher than genetic algorithms and particle swarm optimization algorithms. f(x) represents the cost function, and it is measured by the accuracy of the SVM algorithm. The cost function f(x) is defined as follows:

$$f(x) = Accuracy(x) \quad Eq2$$

Accuracy(x) is the accuracy of the SVM classification on the testing data using the selected subset of features from the training data indicated by x. C represents the number of samples correctly classified by the SVM classifier, and t is the total number of samples in the test set. The classification accuracy of the SVM is obtained using:

$$Accuracy(x) = (c/t) \times 100 \quad Eq3$$

• Cuckoo Algorithm and its Parameters The Cuckoo optimization algorithm is inherently an algorithm used to maximize its objective function. This means that the Cuckoos move towards a point that provides them with the highest profit and the highest fitness for their survival. The code snippet of the Cuckoo algorithm used for feature selection in this thesis is as follows:

(i) Generate an initial population of Cuckoos;

(ii) Calculate the fitness function for this initial population;

(iii) Start a while loop until a specified condition is met;
- Select a Cuckoo randomly (e.g., i) and replace its solution using Lévy flights.
- Calculate the fitness function for this new solution.
- Select another solution randomly (e.g., j).
- If Fitness(j) is lower than the new solution, replace it with j.
- Remove a portion of the worst solutions (pa) and replace them with the new solutions.
- Keep the best solutions.
- Transfer the best solution to the next generation.

(iv) End of the loop;

In this paper, the Cuckoo algorithm is used as the search strategy, and the SVM algorithm serves as the cost function. The parameters of the Cuckoo algorithm for feature selection in the dataset are presented in Table 2.

3.4. Classification in the Proposed Method

The proposed method utilizes neural networks for classification. The performance of neural networks is heavily dependent on their structural parameters. The number of neurons, the number of hidden layers, and the connectivity between neurons are prominent examples of these parameters. In the proposed method, a structure optimization approach is presented to achieve optimal performance. This approach is based on the Automata Learning algorithm.

In the proposed method, the structural parameters of the neural network are considered independent variables, and the network's performance function is optimized using an automated learning optimization method. A fixed structure is used for the learning automata to adjust the Momentum factor and the steepness parameter. The fixed structure is a pentuple structure $< \alpha, \phi, \beta, F, G >$ where $\alpha = (\alpha_1, ..., \alpha_r)$ represents the set of actions, $\phi = (\phi_1, ..., \phi_s)$ represents the set of states, and $\beta = 0, 1$ represents the set of input signals, with $\beta = 1$ indicating a penalty and $\beta = 0$ indicating a reward. $F : \phi \times \beta \longrightarrow$, ϕ is the mapping function, and $G : \phi \longrightarrow \alpha$ is the output mapping.

The chosen action in our project acts as the input from the environment, leading to the generation of an unpredictable response denoted as $\beta(n)$ at time n. The response $\beta(n)$ takes on values of either 0 or 1 and represents the feedback received from the environment in response to the automaton's action. The environment imposes a penalty on the automaton, with the probability of penalty given by $\phi(n)$, which

F_1	F_2	F_3	...	F_{n-1}	F_{n-2}
1	0	1	...	1	0

Figure 5. An example of feature representation in the proposed method

Table 2. Feature selection parameters

Parameter	*Value*
Population size	100
Egg size	6.2
Maximum number of iterations	100
λ	1.5

is contingent on the chosen action. Based on the received response $\beta(n)$, the automaton's state $\phi(n)$ is updated, and a new action is selected for the next time step $(n+1)$. Initially, the value of c_i is unknown, and our objective is to achieve convergence through the interaction between the automaton and the environment. The ultimate goal is to identify an action that minimizes the expected penalty response. The neural network serves as the environment for the learning automaton. Based on the error received from the neural network, the learning automaton adjusts the parameter values of the Backpropagation algorithm. The automaton's actions correspond to the values of the convergence coefficient (or the learning rate), and the automaton's input is a function of the output error of the neural network.

At the start of each epoch in the neural network algorithm, the learning automata chooses one of its actions. In fixed-structure automata, the action selection is determined based on the average of the output function G. The selected action is used in the neural network algorithm for the corresponding time step. The response of the environment to the learning automata is a function of the mean square error (MSE), which will be described later. In the kth time step, the mean square error in the past W time steps is calculated based on Equation (4). W is referred to as the window size.

$$MSE_W(k) = \frac{1}{w}\sum_{m=1}^{w} MSE(k-m) \quad Eq4$$

Here, $MSE(n)$ and $MSE_W(k)$ represent the mean square error in the nth time step and the mean square error in the past W time steps, respectively. $MSE(k)$ is compared with $MSE_W(k)$ and the received penalty from the environment (when $MSE_W(k) - MSE(k)$ is less than a defined threshold) and the received reward (for other cases). The response of the environment (input to the learning automata) can be formulated as follows:

$$\beta(n) = \begin{cases} 0 & \text{if } MSE_W(n) - MSE(n) \leq 0 \\ 1 & \text{if } MSE_W(n) - MSE(n) > 0 \end{cases} \quad Eq5$$

At the beginning of the first time step, the action of the learning automata is randomly chosen from the set of allowable actions. In this thesis, the learning automata is responsible for adjusting the parameters of the neural networks. In each iteration of the algorithm, an output from the training set is given to the neural networks. Then, the responses of the networks are calculated, and the weights are adjusted. Weight adjustment is performed at the end of each time step. The magnitude of the adjustment is proportional to the neural network parameters. By using the learning automata as an adaptation technique, the search for optimal values of the neural network parameters is performed in a probabilistic space rather than the parameter space, resulting in the algorithm's ability to find global optima.

4. Simulation and Results Evaluation

There are various criteria for evaluating machine learning algorithms, which are usually chosen based on the problem conditions, and it is not possible to generalize a specific criterion to all algorithms and data. One of the evaluation criteria for a classifier is its accuracy on the test set. This criterion is a logical measure of performance as it indicates the success rate of a model against samples whose classes are unknown. Table 3 represents the confusion matrix used in the mentioned criteria.

The classification rate is obtained using equation (6), which states that the most important values that

Table 3. Confusion Matrix

		Predicted Class	
		Normal	**Abnormal**
Real Class	**Normal**	TP	FN
	Abnormal	FP	TN

should be maximized in a binary classification problem are TP (True Positive) and TN (True Negative) . TP and TN represent correct classifications. False Positive (FP) occurs when a sample that is actually negative is predicted as positive. False Negative (FN) occurs when a sample that is actually positive is predicted as negative. Additionally, in multi-class problems, the values placed on the main diagonal of this matrix, which are in the numerator, should be maximized.

$$Accuracy = \frac{(TN + TP)}{(TN + FN + TP + FP)} \quad Eq6$$

In malware detection systems, accuracy and recall are commonly used metrics. Accuracy indicates the number of correctly classified samples among the retrieved samples. Recall indicates the number of correctly classified samples divided by the relevant samples.

$$precision = \frac{TP}{TP + FP} \; Eq7$$

$$recall = \frac{TP}{TP + FN} \; Eq8$$

According to the calculations performed for accuracy and recall metrics, the weighted F-Measure can be calculated at this stage. F-Measure is an appropriate parameter for evaluating the quality of classification and serves as a weighted average between precision and recall. For an ideal classification algorithm, this measure is equal to 1, and in the worst case, it is equal to 0. This parameter is calculated based on equation (9).

$$F = 2 \times \frac{(precision.recall)}{(precision + recall)} \quad Eq9$$

The ROC curve allows for visual comparison of a set of classifiers, and multiple points in the ROC space are noteworthy. The point at the bottom-left (0,0) represents a strategy where no positive classifiers are generated. The opposite strategy, which generates positive classifiers without any condition, is represented by the top-right point (1,1). The point (0,1) represents perfect and flawless classification. Generally, a point in the ROC space is better than another if it is located further northwest [19].

AUC represents the area under the ROC curve, and the higher the value of this number for a classifier, the better the final performance of the classifier is evaluated. The ROC curve is a method for evaluating the performance of classifiers. Essentially, ROC curves are two-dimensional curves where the true positive rate is plotted on the Y-axis, and the false negative rate is plotted on the X-axis. In other words, an ROC curve depicts the relative trade-off between benefits and costs. Another common method is statistical tests that provide a certain level of confidence regarding the validity and non-randomness. In this thesis, the mean square error (MSE) is used.

Mean square error is an evaluation measurement system for models. MSE for a model, considering the examined set, is the mean squared difference between the predicted values and the actual values for all the samples in the examined set. Prediction error is the difference between the true value and the predicted value for a sample.

The MSE is calculated using equation (10):

$$MSE = \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n} \quad Eq10$$

Where n is the total number of test data, P_i represents the measured values, and O_i represents the classifier values.

4.1. Database Specifications

To test and evaluate the proposed method for classifying imbalanced data, a standard dataset related to hypothyroidism is used. This dataset has been downloaded from the UCI repository. The specifications of this dataset are provided in Table 4.

4.2. Feature Selection

Data dimension reduction is performed through feature selection and feature extraction. During dimension reduction, feature extraction methods lose the original meaning of the data. Therefore, feature selection methods are used to preserve the meaning of the data. These methods preserve the meaning of the original dataset during dimension reduction. In this thesis, the algorithm is used as the search strategy, and the Support Vector Machine algorithm is used as the cost function. This algorithm is based on an optimization algorithm and is used for feature selection in the dataset. The results obtained from this feature selection method show that the algorithm has a high convergence speed and is capable of searching the problem space effectively, being able to find the smallest acceptable subset. The output of the MATLAB software shows that the selected features are 1, 2, 3, 7, 9, 14, 16, 18, 20, 21, 24, 25, 28.

As the simulation results indicate, after applying the feature selection algorithm, the number of features is reduced from 29 to 13.

Table 4. Dataset Specifications

Class	*Numberofsamples*
Healthy	3481
Hypothyroidism	194
Primary Hypothyroidism	95
Secondary Hypothyroidism	2

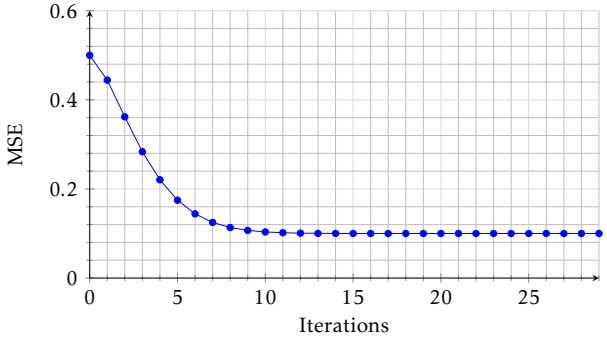

Figure 6. The classification algorithm for 29 iterations

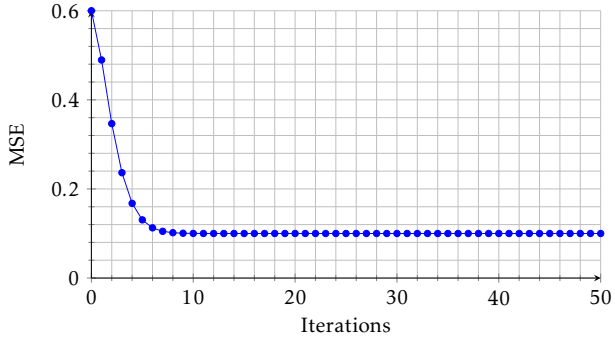

Figure 7. The classification algorithm for 50 iterations

4.3. Classification using Neural Network and Learning Automata

In this paper, the learning automata method is used to find the optimal weights for the neural network. The cost function is equal to the mean squared error (MSE) resulting from the neural network classification. The more optimization is performed, the lower the MSE of the neural network becomes, resulting in higher classification accuracy.

The online output of the intermediate stages of the algorithm is shown in the Figure 6. It is evident that by performing iterations using the learning automata, better weights are obtained, and consequently, the error of the network decreases over time.

The final graph for 50 iterations is displayed in Figure 7 below.

Table 5 is the final output of the implementation:

The first line represents the MSE error and classification accuracy of the neural network combined with the learning automaton. The second line represents the MSE error and classification accuracy of the neural network trained using the MATLAB toolbox.

Based on the provided information, the neural network achieved an accuracy of 98%, while the classification accuracy improved to 99.6% when using the weights obtained from the learning automaton algorithm.

4.4. Comparison of Results

In this section, we discuss the comparison between the proposed method with and without the feature selection algorithm. The results, as shown in Figure **??**, demonstrate a significant improvement when the proposed method is combined with the feature selection algorithm.

To conduct a more comprehensive evaluation of the feature selection algorithm, we employed particle swarm optimization and genetics as search strategies. Table 7 presents the accuracy of the selection algorithm in the proposed method, outperforming the other two methods. The cuckoo algorithm exhibited better convergence compared to the particle swarm optimization and genetics algorithms.

In Figure 9, a comparison between the accuracy and MSE (Mean Squared Error) is presented for the backpropagation neural network algorithm and the optimized neural network algorithm when combined with the feature selection algorithm using support vector machine and FaHD optimization.

Now comparison and evaluation of the proposed method with the approaches presented in references [14] and [15] will be addressed. The important parameter used for comparison is accuracy. Accuracy is the most well-known and widely used performance metric for classification algorithms, indicating the percentage of correctly classified instances from the

Table 5. Results obtained using MATLAB

Best MSE= 0.02	Best Accuracy=99.60 %
Toolbox MSE= 0.06	Toolbox Accuracy=98.00 %

Table 6. Comparison of feature selection methods

Feature selection method	*Classification rate*	*Precision*	*Recall*	*F – Measure*	*ROC*
PSO-SVM	96.06	94	95.6	0.947	0.936
GA-SVM	95.8	93.2	93.8	0.934	0.923
CO-SVM	99.6	99.2	99	0.99	0.997

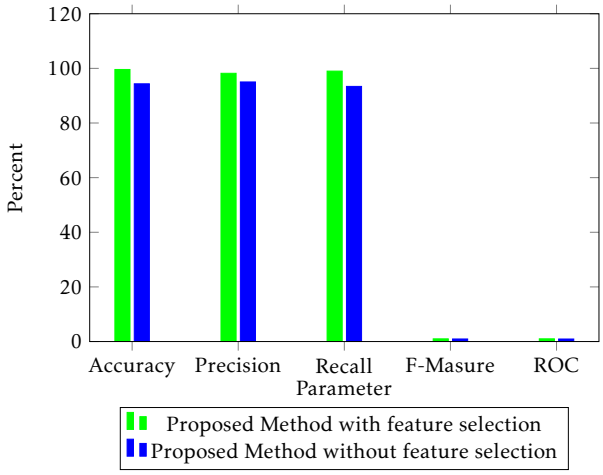

Figure 8. Comparison of the proposed method using feature selection and without feature selection

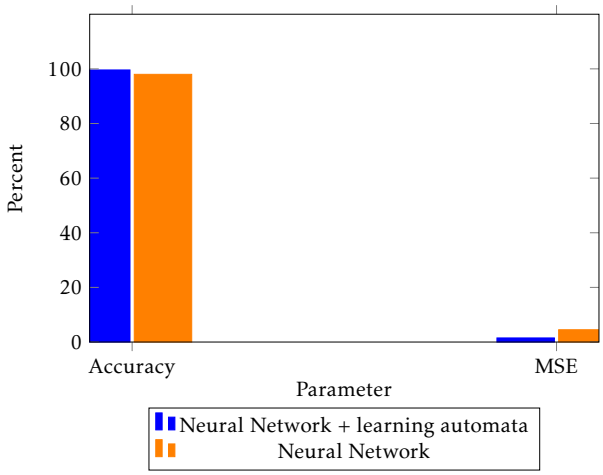

Figure 9. Comparing the performance of a neural network and an optimized neural network with learning automata

total test dataset. As shown in Figure 9, the accuracy of the proposed method is 99.6%, while the accuracies of the compared methods are 98.65% and 98.25%, respectively. This indicates an improvement in accuracy in the proposed method, as it utilizes the Automatic Learning Automata algorithm for optimizing the neural network weights.

Finally, the proposed method is compared with several common machine learning algorithms. The results of this comparison are shown in Table 7.

5. Conclusion

One of the current scientific challenges is dealing with imbalanced datasets. Such these sets can be seen in graph theory [30]. In such datasets, one of the classes has a significantly smaller number of instances compared to others (the minority class) and is also more important. Classifying a large number of instances into one class (the majority class) poses difficulties for conventional classification methods. These methods tend to assign data to the majority class, resulting in high accuracy but low reliability. In this study, new hybrid methods at the data level are introduced to address this problem. These methods balance the imbalanced distribution of data, leading to better performance of classifiers on them.

To evaluate the proposed method, a low-quality thyroid dataset from the UCI database is utilized. For this purpose, a dataset is selected for training the model and used accordingly. Effective features for classification are selected using the Rough Set and Support Vector Machine algorithms. In the feature selection stage, a combined Support Vector Machine algorithm is used as the cost function, and the Rough Set algorithm is used as the search strategy. Autómata Learning and Neural Network are employed for data classification. Ultimately, our proposed method

Table 7. Comparison of the proposed method with common machine learning algorithms

Algorithm	*Classificationrate*	*Precision*	*Recall*	*F – Measure*	*ROC*
Simple bayesian	78.3	78	79.2	0.78	0.812
Decision tree	88.07	88.2	90	0.89	0.89
Support vector machine	85.4	85.6	84.5	0.85	0.842
Suggested method	99.2	99.2	99	0.99	0.997

achieves an accuracy of over 99.6%. It is a high-accuracy approach that has been implemented in MATLAB simulation.

As for future work, it is suggested to explore other optimization algorithms, such as Genetic Algorithms for feature selection. Additionally, utilizing neural networks as the cost function can be beneficial. This thesis proposes a novel method based on an optimized algorithm, and in the future, it can be combined with sampling-based approaches. Moreover, the graph-based approaches may provide alternative techniques to address the class imbalance by leveraging the inherent structure and relationships present in the data. They can complement traditional resampling methods and machine learning algorithms, helping to improve the performance and handling of imbalanced datasets [31, 32]. Also, robotics can complement these approaches by actively participating in data collection, augmentation [33], and learning processes to enhance the handling of imbalanced data.

References

[1] Harouni M, Baghmaleki HY. Color image segmentation metrics. Encyclopedia of Image Processing. 2018 Nov 8;95:10-21.

[2] Moshayedi AJ, Khan AS, Shuxin Y, Kuan G, Jiandong H, Soleimani M, Razi A. E-Nose design and structures from statistical analysis to application in robotic: a compressive review. EAI Endorsed Transactions on AI and Robotics. 2023 Apr 20;2(1):e1.

[3] Mahmudi F, Soleimani M. Some results on Maximal Graph of a Commutative Ring, 2016.

[4] Hoorfar H, Bagheri A. Geometric Embedding of Path and Cycle Graphs in Pseudo-convex Polygons. arXiv preprint arXiv:1708.01457. 2017 Aug 4.

[5] Jia, F., Lei, Y., Lu, N., Xing, S. (2018). Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. Mechanical Systems and Signal Processing, 110, 349-367.

[6] Harouni M, Karimi M, Rafieipour S. Precise segmentation techniques in various medical images. Artificial Intelligence and Internet of Things. 2021 Aug 25:117-66.

[7] Karimi M, Harouni M, Rafieipour S. Automated medical image analysis in digital mammography. InArtificial Intelligence and Internet of Things 2021 Aug 25 (pp. 85-116). CRC Press.

[8] García, S., Zhang, Z. L., Altalhi, A., Alshomrani, S., Herrera, F. (2018). Dynamic ensemble selection for multi-class imbalanced datasets. Information Sciences, 445, 22-37.

[9] Verma AK, Pal S, Kumar S (2019) Classification of skin disease using ensemble data mining techniques. Asian Pac J Cancer Prev 20(6):1887–1894.

[10] Ahmad W, Huang L, Ahmad A, Shah F, Iqbal A, Saeed A (2017) Thyroid diseases forecasting using a hybrid decision support system based on ANFIS, k-NN and information gain method.J Appl Environ Biol Sci 7(10):78–85.

[11] Khan AR, Doosti F, Karimi M, Harouni M, Tariq U, Fati SM, Ali Bahaj S. Authentication through gender classification from iris images using support vector machine. Microscopy research and technique. 2021 Nov;84(11):2666-76.

[12] Soleimani M, Mahmudi F, Naderi MH. On the Maximal Graph of a Commutative Ring. Mathematics Interdisciplinary Research. 2021 Jul 2.

[13] Chaurasia V, Pal S, Tiwari BB (2018) Prediction of benign and malignant breast cancer using data mining techniques. J Algorithm Comput Technol 12(2):119–126.

[14] Moshayedi, A. J., Gharpure, D. C. (2012, May). Development of position monitoring system for studying performance of wind tracking algorithms. In ROBOTIK 2012; 7th German Conference on Robotics (pp. 1-4). VDE.

[15] Sumathi A, Nithya G, Meganathan S (2018) Classification of thyroid disease using data mining techniques. Int J Pure Appl Math 119(12):13881–13890

[16] Moshayedi, A. J., Li, J., Sina, N., Chen, X., Liao, L., Gheisari, M., X. Xie,(2022). Simulation and validation of optimized pid controller in agv (automated guided vehicles) model using pso and bas algorithms. Computational Intelligence and Neuroscience, 2022.

[17] Pravin SR, Jafar OA (2017) Prediction of skin disease using data mining techniques. IJARCCE 6(7):313–318.

[18] . Wang J, Li S, Song W, Qin H, Zhang B, Hao A (2018) Learning from weakly-labeled clinical data for automatic thyroid nodule classification in ultrasound images. In: 2018 25th IEEE international conference on image processing (ICIP), IEEE, pp 3114–3118.

[19] Sivasakthivel A, Shrivakshan GT (2017) A comparative study of diagnosing thyroid diseases using classification algorithm. Int JAdv Res Comput Sci Softw Eng 7(8):181–184.
[20] Moshayedi A. J., Gharpure, D. C. (2017, May). Evaluation of bio inspired Mokhtar: Odor localization system. In 2017 18th international carpathian control conference (ICCC) (pp. 527-532). IEEE.
[21] IoniN I., IoniN´L. (2016) Prediction of thyroid disease using data mining techniques. Brain Broad Res Artif Intell Neurosci 7(3):115–127.
[22] Rathi M, Pareek V (2016) Disease prediction tool: an integrated hybrid data mining approach for healthcare. IRACST Int J Comput Sci Inf Technol Secur (IJCSITS) 6(6):32–40
[23] K. Geetha, Baboo CSS (2016) Efficient thyroid disease classification using differential evolution with SVM. J Theoret Appl Inf Technol 88(3):410–420.
[24] Soleimani M, Mirshahzadeh AS. Multi-class Classification of Imbalanced Intelligent Data using Deep Neural Network. EAI Endorsed Transactions on AI and Robotics. 2023 Jul 12;2.
[25] Moshayedi, A. J., Chen, Z. Y., Liao, L., Li, S. (2022). Sunfa Ata Zuyan machine learning models for moon phase detection: algorithm, prototype and performance comparison. TELKOMNIKA (Telecommunication Computing Electronics and Control), 20(1), 129-140.
[26] Razia S, Narasingarao MR, Sridhar GR (2015) Decision support system for prediction of thyroid disease—a comparison of multilayer perceptron neural network and radial function neural network. J Theoret Appl Inf Technol 80(3):544–551.
[27] Gaikwad S, Pise N (2014) An experimental study on hypothyroid using rotation forest. Int J Data Min Knowl Manag Process (IJDKP) 4(6):31. https://doi.org/10.5121/ijdkp.2014.4603.
[28] Hoorfar H, Bagheri A. Minimum hidden guarding of histogram polygons. arXiv preprint arXiv:1708.05815. 2017 Aug 19.
[29] Soleimani M, Naderi MH, Ashrafi AR. Tensor product of the power graph of some finite rings. Facta Universitatis, Series: Mathematics and Informatics. 2019 Mar 13:101-22.
[30] Soleimani M, Mahmudi F, Naderi MH. Some results on the maximal graph of commutative rings. Advanced Studies: Euro-Tbilisi Mathematical Journal. 2023 Mar;16(supp1):21-6.
[31] Mahmudi F, Soleimani M, Naderi MH. Some Properties of the Maximal Graph of a Commutative Ring. Southeast Asian Bulletin of Mathematics. 2019 Jul 1;43(4).
[32] Hoorfar H, Bagheri A. A New Optimal Algorithm for Computing the Visibility Area of a simple Polygon from a Viewpoint. arXiv preprint arXiv:1803.10184. 2018 Mar 27.
[33] Moshayedi A. J., Hosseini M. S. , F. Rezaee. (2019). WiFi based massager device with NodeMCU through arduino interpreter. Journal of Simulation and Analysis of Novel Technologies in Mechanical Engineering, 11(1), 73-79.