IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 3, June, 2023, pp. 767-775
Journal homepage: http://iieta.org/journals/isi

An Enhanced Outlier Detection Approach for Multidimensional Datasets Using a Synergistic Firefly and Grey Wolf Optimization-Based Method

Manoharan Govindaraj[1*], Sivakumar Kaliappan[2], Ganesh Swaminathan[1]

[1] Department of Mathematics, Sathyabama Institute of Science and Technology, Chennai 600119, India
[2] Department of Mathematics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai 602105, India

Corresponding Author Email: vijimanoharan77@gmail.com

https://doi.org/10.18280/isi.280328

Received: 22 February 2023
Accepted: 3 May 2023

Keywords:
outlier detection, multivariate datasets, improved neural network, sun flower-based grey wolf optimization

ABSTRACT

Outlier identification and elimination are essential preprocessing steps for data analysis tasks such as clustering, classification, and regression. The accuracy of data analysis outcomes may be compromised if outliers are not adequately addressed. Detecting outliers is particularly challenging when they are characterized by unusual combinations of multiple attributes. Furthermore, the presence of outliers can impact various data processing activities, necessitating either the reduction of outlier influence or their complete removal. Outlier detection in multivariate data presents a complex process that becomes increasingly difficult when dealing with high-dimensional datasets. Consequently, this study focuses on the identification of such outliers in multivariate datasets using intelligent techniques. In the proposed approach, outliers are detected using an Improved Neural Network (INN), where the hidden neurons are tuned by a novel Synergistic Firefly-Grey Wolf Optimization (SF-GWO) algorithm. This algorithm combines the strengths of the Firefly Optimization (SFO) and Grey Wolf Optimization (GWO) techniques to maximize accuracy. The unique method results in enhanced classification model performance, reduced computation time, and increased classification accuracy. The proposed model has been evaluated and compared with well-established traditional techniques, demonstrating its effectiveness in addressing the challenges of outlier detection in multidimensional datasets.

1. INTRODUCTION

Outlier detection, a critical data analysis approach for identifying unusual data points in a dataset, has been employed in various fields such as finance, communication networks, medical, and environmental research [1]. Many of these applications involve categorical data. For instance, intrusion detection datasets often comprise intercepted packets with categorical attributes like "protocol" [2]. Although numerous outlier detection algorithms exist for numerical data, only a few traditional methods can handle categorical data. These techniques suffer from high time complexity and low detection accuracy [3].

The goal of outlier detection, also known as anomaly detection, is to identify data points with anomalous characteristics. As one of the most fundamental data analytics approaches, outlier detection is crucial since unrecognized outliers can negatively impact data analysis results and, consequently, analysis-driven decisions [4]. Outlier detection can be applied in two scenarios. If outliers are considered noise, they must be removed before analyzing the dataset for knowledge discovery. In contrast, when outliers are the main objectives (e.g., regular purchases in an online electronics store), they must be accurately identified [5]. Owing to its effectiveness, outlier detection has become a significant technique for various applications, including fraud detection in financial transactions, intrusion detection in communication networks, and disease diagnosis.

Numerous outlier detection strategies have been proposed in recent years [6]. Currently available methods can be grouped into three categories: distance-oriented, statistical distribution-oriented, and clustering-oriented approaches. Statistical distribution-oriented methods require the data points under investigation to follow a specific distribution [7], which can then be used to identify outliers. Distance-oriented approaches detect outliers by measuring the number of neighbors a data point has – an outlier is a data point with few neighbors. However, these methods require at least quadratic time complexity in relation to the number of data items, making them unsuitable for large datasets [8]. Clustering-oriented algorithms group data points into separate clusters to detect outliers. Most traditional outlier detection methods [9] are designed for numerical datasets, but many practical applications involve categorical data. For example, the detection of anomalous traffic in communication networks involves the analysis of a sequence of network packets [10], which often have multiple categorical attributes such as "protocol."

Detecting outliers in time series datasets is a common real-world challenge [11]. A time series dataset consists of a sequence of data points ordered chronologically. Anomaly detection in time series data is complex due to various factors. First, real-time detection is a crucial requirement in many intrusion detection and industrial monitoring systems for information security [12]. Second, high-dimensional datasets exhibit relationships between different data attributes. Third,

outliers are not always the highest or lowest values but rather those discordant with the underlying data, increasing computational complexity. Additionally, time series data exhibit a high degree of consistency and correlation [13], making it impossible to ensure that data points are uniformly and independently distributed. As a result, outlier detection techniques must address numerous complex requirements. Various researchers have proposed outlier detection methods for time series datasets, such as the FAR model, the INAR model, and the VAR methods, which are regression-based approaches that search for data points inconsistent with the theory [14]. Empirical likelihood-based methods, which incorporate simulated variables in each test and use likelihood ratios to detect outliers, have also been proposed. While these approaches are applicable in specific contexts, they also have some limitations. Some methods require prior knowledge of the data, which may be challenging in certain real-world situations. Furthermore, parameter selection significantly influences the quality and performance of the detection results [15]. Moreover, most algorithms are computationally intensive and do not apply to clustered outliers.

This paper contributes to the field by:

- Identifying outliers in multivariate datasets using intelligent technology.
- Detecting outliers in multivariate datasets using INN, with hidden neurons of NN tuned for accuracy maximization.
- Proposing a novel optimization algorithm called SF-GWO for enhancing the detection phases and comparing it with conventional models to demonstrate the superiority of the developed method.

The remainder of this paper is organized as follows: Section II presents a literature survey. Section III discusses outlier detection in multivariate datasets. Section IV describes the INN for outlier detection in multivariate datasets. Section V introduces the SF-GWO for outlier detection in multivariate datasets. Section VI reports the results, and Section VII concludes the paper.

2. LITERATURE REVIEW

In recent years, various outlier identification techniques for categorical data sets have been proposed, with some achieving remarkable improvements in computational complexity and detection accuracy. This literature review highlights several significant contributions to the field, demonstrating the advancements made in recent years.

In 2021, Du et al. [16] introduced two novel outlier identification techniques, namely the Outlier Detection Tree (ODT) and the Fast Outlier Detection Tree (FAST-ODT). ODT is a basic entropy-based strategy using a classification tree to categorize data objects as either normal or abnormal. FAST-ODT, an enhanced version, boasts reduced time complexity while maintaining exceptional detection accuracy. The results indicate that FAST-ODT outperforms existing methods in terms of both computational complexity and precision.

Lu et al. [17] presented the Outlier Detection for Categorical Attributes (ODCA) technique in 2018. This method is divided into three stages: data preparation, outlier ranking, and outlier analysis. Firstly, linear interpolation is employed to transform assembled outliers into isolated ones. Secondly, cross-correlation analysis is used to convert high-dimensional data sets into a one-dimensional cross-correlation function for isolated outlier detection. Finally, a multilevel Otsu's approach is utilized to adaptively determine rank thresholds and output anomalous samples at various levels. The authors conducted four experiments with multiple high-dimensional time series data sets and compared their method with other detection techniques. The results suggest that ODCA outperforms conventional approaches in terms of effectiveness and time complexity.

Degirmenci and Karal [18] proposed a novel method called RiLOF, which is based on the iLOF technique and overcomes traditional limitations. They introduced a new metric, the Median of Nearest Neighbors Absolute Deviation (MoNNAD), which combines the median and local absolute deviation of samples' LOF values. RiLOF can detect outliers in various data stream applications with consistent hyperparameters. Extensive tests on 15 real-world data sets demonstrated that RiLOF significantly outperforms 12 traditional competitors.

In 2017, Zhu et al. [19] developed FRIOD, an innovative interactive outlier detection method that incorporates deep human involvement to enhance detection performance and significantly simplify the detection process. The user-friendly interactive approach allows users to engage in the core outlier identification algorithm's primary stages, including location-aware distance thresholding, dense cell selection, and final top outlier validation. By incorporating human interaction, FRIOD optimizes the grid-oriented space partitioning, a crucial step, and can improve the quality of discovered outliers while making the detection process more effective and efficient.

Lastly, Yousef et al. [20] presented UN-AVOIDS in 2021, an unsupervised and nonparametric technique that provides invariant anomalous scores (normalized to [0, 1]) for both viewing and identification of outliers. The key feature of UN-AVOIDS is the transformation of data into a novel space called the Normalized Cumulative Distribution Function (NCDF), where both visualization and detection are performed. Outliers are easily visible in this space, resulting in high Area Under the Curve (AUC) scores for the anomaly detection algorithm. UN-AVOIDS outperformed three major anomaly detection techniques (IF, LOF, and FABOD) in terms of AUC when tested with both simulated and two newly released cybersecurity datasets.

In conclusion, the advancements in outlier identification techniques for categorical data sets have been remarkable, with novel methods demonstrating improved detection accuracy, computational complexity, and versatility. Further research in this area is expected to continue pushing the boundaries of what is possible in outlier detection.

In 2016, Salehi et al. [21] have presented a MiLOF detection technique for data streams, as well as a more customizable variant (MiLOF F), both of which contained a similar accuracy to Incremental LOF but are limited in memory. The outcomes reveal that both suggested techniques outperform Incremental LOF in terms of memory and temporal complexity while maintaining comparable accuracy. Furthermore, we demonstrated that MiLOF F was unaffected by variations in the count of data points, fundamental clusters, and dimensions in the data stream. These findings demonstrated that MiLOF/MiLOF F were well-applicable to application contexts having lessened memory (e.g., WSNs) and could handle high-volume data streams.

Lin and Wang [22] proposed a probabilistic deep autoencoder designed to reassemble power system data by

incorporating a nonparametric distribution estimation approach. This approach allowed for the inclusion of uncertainty information within the observed data. Confidence intervals, predicted by the method, were utilized as input for the initial layer of the neural network, while the reconstructed data facilitated the detection and replacement of outliers. The effectiveness of this technique was corroborated through simulated results. However, the probabilistic deep autoencoder necessitates an abundance of labeled data for training, which may be unattainable in certain real-world situations. Furthermore, the method may be ill-suited for managing high-dimensional data, a prevalent issue in numerous applications.

Li et al. [23] introduced a graph-based strategy, enabling an unsupervised approach to evaluate a minimal amount of labeled data. The semi-supervised system was extended to active outlier detection by incorporating a query method that selected top-ranked outliers. The semi-supervised outlier detection technique demonstrated performance comparable to that of the leading traditional methods, while the active outlier detection method surpassed them. This technique was tested across 12 real-world datasets. However, Li et al.'s graph-based strategy exhibits limitations in handling complex, high-dimensional datasets and may prove computationally taxing when processing vast quantities of data, potentially impacting performance.

Wang et al. [24] developed the RODA algorithm, capable of handling both single and multiple query processing. A novel outlier estimation method was proposed for single query processing, and the R-tree index was expanded to reduce the retrieval space by prioritizing data points with high outlier degrees. The algorithm was designed to explore the sharing mechanism among multiple queries in depth for multiple query processing. Experimental results indicated that RODA enhanced operational efficiency and held significant practical applicability. However, the RODA algorithm requires a pre-processing step to construct an R-tree index, which may be infeasible for dynamic or constantly changing data. Additionally, the method may struggle with high-dimensional data, as performance could deteriorate as data dimensionality increases.

Yu et al. [25] suggested transferring knowledge from labeled source data to support unsupervised outlier detection in a target dataset. The source and target data were combined for joint clustering and outlier identification, utilizing the source data clustering algorithm as a constraint to maximize the use of source knowledge. A K-means algorithm was employed to address the problem using an augmented matrix. This algorithm was found to be a dependable approach with a precise mathematical definition and theoretical convergence guarantee. However, Yu et al. [25] 's transfer learning-based method assumes that the labeled source data accurately represents the target data, which may not always be the case. Moreover, the approach may be computationally intensive when handling large, high-dimensional datasets, potentially affecting its scalability.

In conclusion, the literature review presented herein outlines the developments and limitations of various outlier identification techniques for categorical data sets. The experimental results and comparisons conducted across multiple real-world datasets demonstrate the effectiveness and potential improvements of the proposed methods in terms of outlier detection and cluster validity measures. Further research is necessary to address the challenges associated with high-dimensional data, computational complexity, and the availability of labeled data for training.

3. OUTLIER DETECTION IN MULTIVARIATE DATASETS

Outliers can have a detrimental impact on data analysis outcomes by skewing the results and distorting the underlying patterns or trends present in the data. This can lead to incorrect conclusions and flawed decision-making processes, which can have serious consequences in various fields such as finance, healthcare, and environmental research. For example, in finance, the presence of outliers in financial data can lead to inaccurate risk assessments and investment decisions. In healthcare, outliers in medical data can lead to misdiagnosis or ineffective treatment plans. In environmental research, outliers in data on pollutant levels can lead to incorrect assessments of environmental risks and inadequate regulatory measures. Furthermore, outliers can also result in inefficient data analysis processes. For instance, if outliers are not identified and removed from the dataset, data analysis techniques that rely on statistical assumptions such as normality or homogeneity of variance may not be applicable, resulting in the need for more complex and time-consuming analyses. Consider, for example, [26, 27] for compositional data consisting of multivariate characteristics with the goal of evaluating potential influence to a whole. As a result, instead of the variables itself, the ratios of the variables include the important information of compositional data. This has consequences for outlier identification algorithms, which must be tailored to this sort of data. For example, statistical analysis of the chemical composition of rock or the mineral content of geological samples may be of interest in the area of geochemistry, and knowledge concerning the availability of outliers may be useful to the researcher.

Compositional data are distinguished from merely restricted data by two extra conditions, despite the fact that they are always defined by a constant sum constraint. Consequently, the knowledge in the variables must be independent of the scale invariance (unit scale), and on the other extreme, the outcomes of sub compositions should be compatible with the entire composition's findings (sub compositional coherence). Furthermore, compositional data do not conform to Euclidean geometry, but instead generate their unique geometry on the plain, which is known as the Aitchison geometry. The variables in this region can only have values scaling from 0 to a defined constant (e.g., 100 in the instance of percentages).

Considering that the raw values associated with a composition are reliant on one another, another weakness of the comparative nature of the variables is its skewed covariance architecture [28]. In fact, a rise in single variable in one observation may result in a reduction in another. The closed nature and inherent interrelatedness of compositional data, overall, impede the effective use of typical statistical approaches for data analysis consider [29, 30].

To address these flaws, the log ratio transformations for data from simplex to real space were devised, allowing the compositional data points to be expressed in ortho normal coordinates [31]. In the E-part simplex space, every composition y describes a random vector with strictly positive elements.

$$T^E = \{y = (y_1, \cdots, y_E) \epsilon S^E : y_j > 0, j = 1, \cdots, E, \sum_{j=1}^{E} y_j = l\} \quad (1)$$

Here, l represents a fixed constant once again. A part shows an element of a composition that must not be zero, because only the ratios among the parts are useful in compositional data analysis. At this point, it should be mentioned that suitable techniques are available in the scenario of zero parts $y_j=0$ for $j \in \{1, \cdots, E\}$ in Eq. (1) induced, for instance, by metrics beneath a specific missing information or detection limit, view [32, 33], and in the high-dimensional scenario [34].

By transferring the actual information from the limited simplex space T^E, to the Euclidean real space, S^{E-1}, the logratio transformation technique provides for preprocessing. The modified data can then be used to adjust normal statistical processes for data analysis, resulting in better outcomes. The clr, the alr, and the ilr transformations define the primary family members of log ratio transformations emphasized in the literature. Only the final two are isometric, but they're completely bijections. Aitchison [26] presented both the alr as well as clr transformations, which were later supplanted by transformation [35]. The alr does not preserve distance, and the clr, while isometric, produces a unique covariance architecture. In the case of (E-1)-dimensional hyperplane covered by clr coefficients, the ilr-transformed shows the assessment of an orthonormal basis. The ilr transformation defines a bijective and isometric mapping, using the formula ilr: $T^E \rightarrow S^{E-1}$. One among the selected basis's suggestions is:

$$a = ilr(y) = (a_1, \cdots, a_{E-1}) \quad (2)$$

$$a_k = \sqrt{\frac{E-k}{E-k+1}} ln \frac{y_k}{\sqrt[E-k]{\prod_{l=k+1}^{E} y_l}} \quad for \quad k = 1, \cdots, E-1 \quad (3)$$

As a result of this definition, the non-collinear data point a in the (E-1)-dimensional hyperplane has become the expression of $y \in T^E$. Because one portion of the composition is chosen as the pivot, the recommended ilr coordinates are known to as pivot (log ratio) positions (in this scenario y_1). The pivot is not selected randomly in applications because only the pivot can be read directly with respect to its overall dominance relative to the rest of the arrangement. Because y_1 is not implicated in any one the remaining coordinates, the equivalent coordinate a_1, communicates complete relevant data regarding component y_1 in the composition [27]. This is especially helpful with respect to interpretation, since a_1 can now be interpreted with respect to y_1.

4. IMPROVED NEURAL NETWORK FOR THE OUTLIER DETECTION IN MULTIVARIATE DATASETS

NNs [36] are made up of small computational units called "neurons," which are linked to one another through weight connectors. These units then compute the weighted sum of the inputs and determine the output utilizing activation functions or squashing. There are three primary elements of a neural method:

- The input signal y_1 coupled to neuron l is amplified by synaptic weight $\omega\omega_{lj}$ at synapses, or joining links, which contain strength or weight.
- An adder that adds the weighted inputs together.
- A neuron's output is produced by an activation function. It's also known as a squashing function since it reduces (limits) the output signal's intensity range to a predefined limit.

Based on whether the bias c_l is negative or positive, it contains the impact of boosting or reducing the net input of the activation function. The output of neuron l may be expressed quantitatively as:

$$z_l = \varphi\left(\sum_{j=1}^{n} y_j \cdot \omega_{lj} + c_l\right) \quad (4)$$

The input signals are shown by $y_1, y_2, y_3, \cdots, y_n$. The weights of neuron are represented by $\omega_{l1}, \omega_{l2}, \omega_{l3}, \cdots, \omega_{ln}$. The bias is given by c_l. φ defines the function of activation. To better understand the impact of bias on neuron function, the output supplied in Eq. (4) is split into two phases, the initial of which contains the weighted inputs and the total, which is presented as T_l:

$$T_l = \sum_{j=1}^{n} y_j \cdot \omega_{lj} \quad (5)$$

The output of the adder is thus shown in Eq. (6):

$$w_l = T_l + c_l \quad (6)$$

Here, the neuron's output will be:

$$z_l = \varphi(w_l) \quad (7)$$

The connection among the adder output and weighted input will be altered on the basis of the bias value [37].

Improved Neural Network

The INN is used for detecting the outliers in the multivariate datasets. The major objective of the introduced outlier detection in the multivariate datasets is to optimize the hidden neurons of NN with the intention of accuracy maximization as below.

$$fit = \underbrace{argmax}_{HN_{NN}} (Accuracy) \quad (8)$$

Here, the fitness or the objective function is given by *fit*, and hidden neurons of NN is given by HN_{NN} respectively. Accuracy is the accurate prediction count divided by the total prediction count as below.

$$Accuracy = \frac{TU_P + TU_N}{TU_P + TU_N + FU_P + FU_N} \quad (9)$$

In the above equation, true positive is TU_P, false negative is FU_N, true negative is TU_N, and false positive is FU_P respectively.

5. PROPOSED OPTIMIZATION FOR THE OUTLIER DETECTION IN MULTIVARIATE DATASETS

The proposed SF-GWO is used for enhancing the detection phase of the outlier detection model through the optimization of the hidden neurons of NN with the consideration of accuracy maximization. GWO [38] is influenced by the grey wolves. It is composed of four wolves such as alpha, beta,

delta, and omega. The main phases included are getting nearer to the prey, prey harassing, and prey attacking. The GWO is better in the case of challenging search spaces. But it has the drawback in the case of multi objective optimization problems. Hence, to overcome its limitations, SFO is integrated into it and the so formed algorithm is referred as SF-GWO. This SF-GWO can handle all forms of multi objective constrained problems.

In the case of multi-modal problems, the SFO [39] approach defines a population-oriented iterative heuristic global optimization algorithm. Pollination and root velocity are words used by SFO to provide robustness.

In the proposed SF-GWO, the algorithm is modelled through the random concept. If ra≥0.5, then the update is by SFO.

$$\overrightarrow{Y_{J+1}} = \vec{Y_J} + e_j \times \vec{t_J} \tag{10}$$

Here, the new individual is Y_{j+1}, sunflower individual is Y_j, step of the sunflower is e_j, and the direction is $\vec{t_j}$ respectively.

Otherwise, if ra<0.5, then the update is by GWO.

$$\overrightarrow{Y_{J+1}} = \frac{\vec{Y_1}+\vec{Y_2}+\vec{Y_3}}{3} \tag{11}$$

Here, the new position is $\overrightarrow{Y_{J+1}}$, alpha position is $\vec{Y_1}$, beta position is $\vec{Y_2}$, and delta position is $\vec{Y_3}$ respectively. The pseudo code of SF-GWO is in Algorithm 1.

Algorithm 1: Proposed SF-GWO

Start
Parameter initialization
Population initialization
Fitness computation
While $iter <$ max $\_iter$
If $ra \geq 0.5$
$\overrightarrow{Y_{J+1}} = \vec{Y_J} + e_j \times \vec{t_J}$
else
$\overrightarrow{Y_{J+1}} = \frac{\vec{Y_1} + \vec{Y_2} + \vec{Y_3}}{3}$
iter=iter+1
Stop

6. RESULTS DISCUSSION

The suggested outlier detection methodology has been empirically validated, and the findings are discussed in this section. The suggested method was run on an i3 processor with 8GB RAM and MATLAB 14.1 loaded. In the case of single outliers, two outlier datasets were evaluated, while in the case of multiple outliers, one dataset was used. The datasets utilized in the introduced method experiments are described in Tables 1 and 2.

Table 1. Data description (single outlier)

Datasets	No. of samples	No of features	Outlier (%)
PenDigits [40]	6870	16	2.27
MNIST [41]	7603	100	9.2

Table 2. Data description (multiple outlier)

Datasets	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Wine quality [42]	0.41 %	3.32%	29.74%	44.9%	17.96%	3.57%	0.1%

In the suggested method experiments, parameters like true negative rate and false negative rate are employed to calculate the detection rate. The parameters are written as follows:

$$FNR = \frac{FU_N}{TU_P + FU_N} \tag{12}$$

$$TNR = \frac{TU_N}{FU_P + TU_N} \tag{13}$$

The detection rate is calculated using Eqns. (12) and (13).

$$Detection\ rate = (1 - FNR) \times TN \tag{14}$$

6.1 Confusion matrix analysis

Figure 1 depicts the suggested method's confusion matrix, which is dependent on the TU_P, TU_N, FU_P, and FU_N parameters.

Actual class	Predicted class	
	Inliers	Outliers
Inliers	TU_P	FU_N
Outliers	FU_P	TU_N

Figure 1. Confusion matrix analysis

6.2 Detection rate analysis

The Figure 2 and Table 3 demonstrate the suggested method's detection performance for three data sets, against the suggested method's higher performance; it is compared to traditional methods. The suggested SF-GWO detects outliers better than the standard method, as shown in Figures. The word actual refers to the dataset's outlier proportion. In terms of real outliers in the dataset, the suggested method has a higher detection rate. The suggested SF-GWO outperforms the conventional methods in terms of detection performance. Table 4 shows the detection rate.

Table 3. Detection rate analysis (Single outlier)

	Datasets									
	PenDigits					MNIST				
Detection rate (%)	Actual	MHMM [43]	HSMM [44]	RKHS [45]	SF-GWO	Actual	MHMM [43]	HSMM [44]	RKHS [45]	SF-GWO
	3	2.6	2.4	2.7	2.8	9.5	9.1	8.9	9.0	9.3

Table 4. Detection rate analysis (Multiple outlier)

Classes	Detection rate (%)	
Class 1	Actual	1
	MHMM [43]	0.4
	HSMM [44]	0.7
	RKHS [45]	0.5
	SF-GWO	0.9
Class 2	Actual	4
	MHMM [43]	3.4
	HSMM [44]	3.7
	RKHS [45]	3.2
	SF-GWO	3.8
Class 3	Actual	30
	MHMM [43]	24
	HSMM [44]	27
	RKHS [45]	25
	SF-GWO	28
Class 4	Actual	45
	MHMM [43]	42
	HSMM [44]	43
	RKHS [45]	41
	SF-GWO	44
Class 5	Actual	18
	MHMM [43]	14
	HSMM [44]	15
	RKHS [45]	12
	SF-GWO	17
Class 6	Actual	4
	MHMM [43]	3.3
	HSMM [44]	3.4
	RKHS [45]	3.2
	SF-GWO	3.7
Class 7	Actual	2
	MHMM [43]	1.5
	HSMM [44]	1.4
	RKHS [45]	1.7
	SF-GWO	1.9

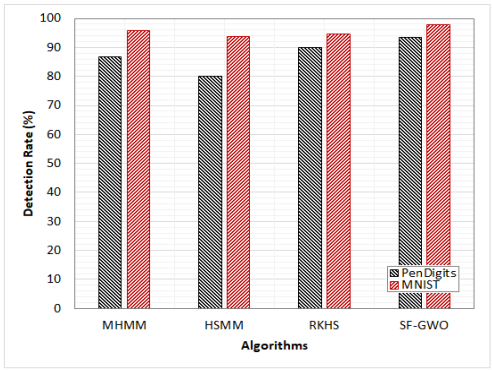

Figure 2. Detection rate analysis (Single outlier)

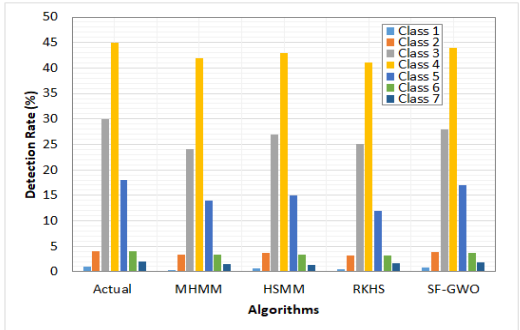

Figure 3. Detection rate analysis (Multiple outlier)

6.3 Accuracy analysis

Figure 3 shows the detection accuracy of the introduced method versus the suggested SF-GWO. For both the single outlier dataset as well as the multiple outlier dataset, the suggested method outperforms the traditional methods in terms of detection accuracy. For all three datasets, the suggested method's average detection accuracy is 97.2%, which is pretty satisfactory. The average of the traditional methods is 93.5 percent, which is 4% lower than the suggested approach. To evaluate the performance gain in classification methods owing to outlier identification, the introduced method is tested alongside a classifier approach. Figure 4 and Table 5 show the classification efficiency of existing approach in comparison to the suggested method. There are two types of performance evaluations: with outlier identification and without outlier detection. The variances in classification accuracy and enhancements in calculation time are studied as a result of this. Figure 5 and Table 6 show that without outlier detection methods, the classification method in NN, LR, and DT methods is lowered for all three data sets.

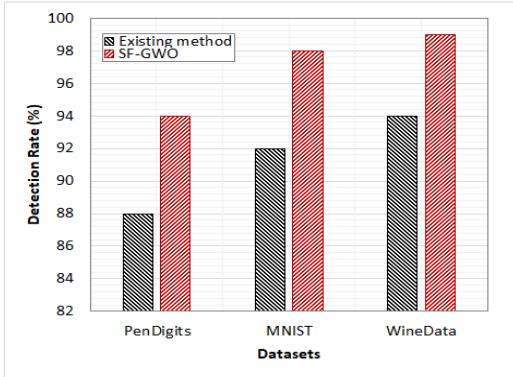

Figure 4. Detection accuracy analysis

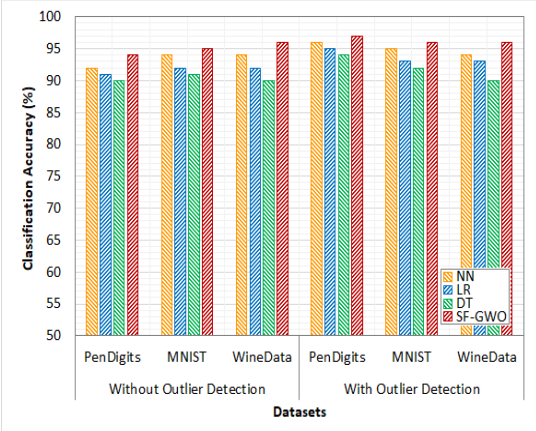

Figure 5. Classification accuracy analysis

Table 5. Detection accuracy analysis

Data sets	Detection accuracy (%)	
Pen Digits	Existing method	88
	SF-GWO	94
MNIST	Existing method	92
	SF-GWO	98
WineData	Existing method	94
	SF-GWO	99

Table 6. Classification accuracy analysis

	Classification accuracy (%)					
	Without outlier detection			With outlier detection		
	PenDigits	MNIST	WineData	PenDigits	MNIST	WineData
NN [36]	92	94	94	96	95	94
LR [46]	91	92	92	95	93	93
DT [47]	90	91	90	94	92	90
SF-GWO	94	95	96	97	96	96

6.4 Computational time analysis

Figure 6 and Table 7 show a comparison of classification method calculation times. Outliers are effectively recognized using the suggested outlier detection, resulting in a reduction in computation time when categorizing data for classification methods. On the other hand, for classification methods without outlier identification, the calculation time increases, this has an impact on classification accuracy:

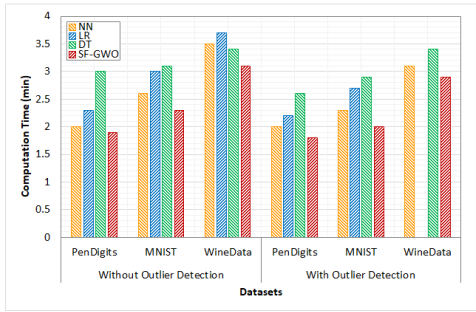

Figure 6. Computational time analysis

Table 7. Computational time analysis

	Classification accuracy (%)					
	Without outlier detection			With outlier detection		
	PenDigits	MNIST	WineData	PenDigits	MNIST	WineData
NN [36]	92	94	94	96	95	94
LR [46]	91	92	92	95	93	93
DT [47]	90	91	90	94	92	90
SF-GWO	94	95	96	97	96	96

7. CONCLUSIONS

In conclusion, this research proposed an intelligent technology for identifying outliers in multivariate datasets using an INN. The proposed model employed a novel SF-GWO algorithm to fine-tune the hidden neurons of the NN, resulting in improved classification model performance and increased accuracy. Moreover, the proposed model significantly reduced computation time, making it more practical and efficient for real-world applications. The proposed model was compared to well-known conventional technologies in the experimental analysis, and the results demonstrated the superiority of the proposed model in terms of accuracy and efficiency. The proposed model's potential to detect outliers in various fields, including finance, healthcare, and engineering, was also discussed. However, there are limitations to this study. For instance, the proposed model's performance could be affected by the size and quality of the dataset used, and further research is needed to investigate the model's robustness in real-world applications.

Overall, this research provides a valuable contribution to the field of outlier detection and presents a promising approach for detecting outliers in multivariate datasets using an intelligent and efficient algorithm.

REFERENCES

[1]= Luan, F., Lv, J., Cao, K. (2016). A fast outlier detection= for categorical datasets. In 2016 12th International= Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, China, pp. 1130-1135. https://doi.org/10.1109/FSKD.2016.7603337

[2] Suri, N.N.R., Murty, M.N., Athithan, G. (2014). A ranking-based algorithm for detection of outliers in categorical data. International Journal of Hybrid Intelligent Systems, 11(1): 1-11. https://doi.org/10.3233/HIS-130179

[3] Sun, Z., Du, H., Ye, Q., Liu, C., Kibenge, P.L., Huang, H., Li, Y. (2019). Outlier detection forest for large-scale categorical data sets. In Computational Data and Social Networks: 8th International Conference, CSoNet 2019, Ho Chi Minh City, Vietnam, pp. 45-56. https://doi.org/10.1007/978-3-030-34980-6_4

[4] Jung, K.M. (2018). Fast entropy attribute value frequency algorithm to detect outliers for categorical data. In Proceedings of the 2018 International Conference on Big Data and Education, pp. 63-66. https://doi.org/10.1145/3206157.3206172

[5] Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G.F., Clermont, G. (2013). Outlier detection for patient monitoring and alerting. Journal of Biomedical Informatics, 46(1): 47-55. https://doi.org/10.1016/j.jbi.2012.08.004

[6] Xie, Z., Li, X., Wu, W., Zhang, X. (2016). An improved outlier detection algorithm to medical insurance. In Intelligent Data Engineering and Automated Learning–IDEAL 2016: 17th International Conference, Yangzhou, China, pp. 436-445. https://doi.org/10.1007/978-3-319-46257-8_47

[7] Li, T., Xiao, N.F. (2015). RETRACTED: Novel heuristic

dual-ant clustering algorithm for network intrusion outliers detection. Optik-Int. J. Light Electron Opt., 126: 494-497. https://doi.org/10.1016/j.ijleo.2014.08.036

[8] Lee, I.H., Mahmood, M.T. (2017). Adaptive outlier elimination in image registration using genetic programming. Information Sciences, 421: 204-217. https://doi.org/10.1016/j.ins.2017.08.098

[9] Abid, A., Masmoudi, A., Kachouri, A., Mahfoudhi, A. (2017). Outlier detection in wireless sensor networks based on OPTICS method for events and errors identification. Wireless Personal Communications, 97: 1503-1515. https://doi.org/10.1007/s11277-017-4583-7

[10] Gil, P., Martins, H., Januário, F. (2019). Outliers detection methods in wireless sensor networks. Artificial Intelligence Review, 52: 2411-2436. https://doi.org/10.1007/s10462-018-9618-2

[11] Bellini, T. (2016). The forward search interactive outlier detection in cointegrated VAR analysis. Advances in Data Analysis and Classification, 10: 351-373. https://doi.org/10.1007/s11634-015-0216-8

[12] Zhang, L., Wang, D., Gao, R., Li, P., Zhang, W., Mao, J., Zhang, Q. (2016). Improvement on enhanced Monte-Carlo outlier detection method. Chemometrics and Intelligent Laboratory Systems, 151: 89-94. https://doi.org/10.1016/j.chemolab.2015.12.006

[13] Li, L., Li, T. (2012). A case study on regression model based outlier detection. In Advances in Information Technology and Industry Applications, pp. 661-669. https://doi.org/10.1007/978-3-642-26001-8

[14] Kaneko, H. (2018). Automatic outlier sample detection based on regression analysis and repeated ensemble learning. Chemometrics and Intelligent Laboratory Systems, 177: 74-82. https://doi.org/10.1016/j.chemolab.2018.04.015

[15] Kontaki, M., Gounaris, A., Papadopoulos, A.N., Tsichlas, K., Manolopoulos, Y. (2016). Efficient and flexible algorithms for monitoring distance-based outliers over data streams. Information Systems, 55: 37-53. https://doi.org/10.1016/j.is.2015.07.006

[16] Du, H., Ye, Q., Sun, Z., Liu, C., Xu, W. (2020). FAST-ODT: A lightweight outlier detection scheme for categorical data sets. IEEE Transactions on Network Science and Engineering, 8(1): 13-24. https://doi.org/10.1109/TNSE.2020.3022869

[17] Lu, H., Liu, Y., Fei, Z., Guan, C. (2018). An outlier detection algorithm based on cross-correlation analysis for time series dataset. IEEE Access, 6: 53593-53610. https://doi.org/10.1109/ACCESS.2018.2870151

[18] Degirmenci, A., Karal, O. (2021). Robust incremental outlier detection approach based on a new metric in data streams. IEEE Access, 9: 160347-160360. https://doi.org/10.1109/ACCESS.2021.3131402

[19] Zhu, X., Zhang, J., Li, H., Fournier-Viger, P., Lin, J.C.W., Chang, L. (2017). FRIOD: A deeply integrated feature-rich interactive system for effective and efficient outlier detection. IEEE Access, 5: 25682-25695. https://doi.org/10.1109/ACCESS.2017.2771237

[20] Yousef, W.A., Traoré, I., Briguglio, W. (2021). UN-AVOIDS: unsupervised and nonparametric approach for visualizing outliers and invariant detection scoring. IEEE Transactions on Information Forensics and Security, 16: 5195-5210. https://doi.org/10.1109/TIFS.2021.3125608

[21] Salehi, M., Leckie, C., Bezdek, J.C., Vaithianathan, T., Zhang, X. (2016). Fast memory efficient local outlier detection in data streams. IEEE Transactions on Knowledge and Data Engineering, 28(12): 3246-3260. https://doi.org/10.1109/TKDE.2016.2597833

[22] Lin, Y., Wang, J. (2019). Probabilistic deep autoencoder for power system measurement outlier detection and reconstruction. IEEE Transactions on Smart Grid, 11(2): 1796-1798. https://doi.org/10.1109/TSG.2019.2937043

[23] Li, Y., Wang, Y., Ma, X., Qian, C., Li, X. (2019). A graph-based method for active outlier detection with limited expert feedback. IEEE Access, 7: 152267-152277. https://doi.org/10.1109/ACCESS.2019.2947736

[24] Wang, X., Li, J., Bai, M., Ma, Q. (2021). RODA: A fast outlier detection algorithm supporting multi-queries. IEEE Access, 9: 43271-43284. https://doi.org/10.1109/ACCESS.2021.3058660

[25] Yu, W., Ding, Z., Hu, C., Liu, H. (2019). Knowledge reused outlier detection. IEEE Access, 7: 43763-43772. https://doi.org/10.1109/ACCESS.2019.2906644

[26] Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2): 139-160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

[27] Filzmoser, P., Hron, K., Templ, M. (2018). Applied compositional data analysis. Cham: Springer. https://doi.org/10.1007/978-3-319-96422-5

[28] Pawlowsky-Glahn, V., Buccianti, A. (2011). Compositional data analysis. Chichester: Wiley. https://doi.org/10.1002/9781119976462

[29] Filzmoser, P., Hron, K. (2008). Outlier detection for compositional data using robust methods. Mathematical Geosciences, 40: 233-248. https://doi.org/10.1007/s11004-007-9141-5

[30] Filzmoser, P., Hron, K., Reimann, C. (2009). Principal component analysis for compositional data with outliers. Environmetrics: The Official Journal of the International Environmetrics Society, 20(6): 621-632. https://doi.org/10.1002/env.966

[31] Hron, K., Templ, M., Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. Computational Statistics & Data Analysis, 54(12): 3095-3107. https://doi.org/10.1016/j.csda.2009.11.023

[32] Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology, 35: 253-278. https://doi.org/10.1023/A:1023866030544

[33] Templ, M., Hron, K., Filzmoser, P. (2017). Exploratory tools for outlier detection in compositional data with structural zeros. Journal of Applied Statistics, 44(4): 734-752. https://doi.org/10.1080/02664763.2016.1182135

[34] Templ, M., Hron, K., Filzmoser, P., Gardlo, A. (2016). Imputation of rounded zeros for high-dimensional compositional data. Chemometrics and Intelligent Laboratory Systems, 155: 183-190. https://doi.org/10.1016/j.chemolab.2016.04.011

[35] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35(3): 279-300. https://doi.org/10.1023/A:1023818214614

[36] Zayegh, A., Al Bassam, N. (2018). Neural network principles and applications. Digital Systems.

http://dx.doi.org/10.5772/intechopen.80416
[37] Hu, Y.H., Hwang, J.N. (Eds.). (2002). Handbook of neural network signal processing. Boca Raton: CRC Press.
[38] Mirjalili, S., Mirjalili, S.M., Lewis, A. (2014). Grey wolf optimizer. Advances in Engineering Software, 69: 46-61. https://doi.org/10.1016/j.advengsoft.2013.12.007
[39] Gomes, G.F., da Cunha, S.S., Ancelotti, A.C. (2019). A sunflower optimization (SFO) algorithm applied to damage identification on laminated composite plates. Engineering with Computers, 35: 619-626. https://doi.org/10.1007/s00366-018-0620-8
[40] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, 2019.
[41] Aggarwal, C.C., Sathe, S. (2015). Theoretical foundations and algorithms for outlier ensembles. Acm Sigkdd Explorations Newsletter, 17(1): 24-47. https://doi.org/10.1145/2830544.2830549
[42] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4): 547-553. https://doi.org/10.1016/j.dss.2009.05.016
[43] Manoharan, G., Sivakumar, K. (2022). A modified hidden markov model for outlier detection in multivariate datasets. International Journal of Engineering Systems Modelling and Simulation, 1(1): 1. https://doi.org/10.1504/ijesms.2022.10046736
[44] Manoharan, G., Sivakumar, K. (2022). An enhanced Hidden Semi-Markov model for outlier detection in multivariate datasets. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-7. https://doi.org/10.3233/JIFS-213374
[45] Govindaraj, M., Kaliappan, S., Swaminathan, G. (2022). Outlier detection of functional data using reproducing kernel Hilbert space. Instrumentation Mesure Métrologie, 21(4): 145-150. https://doi.org/10.18280/i2m.210404
[46] Cao, Z., Liu, L., Markowitch, O. (2017). Comment on "highly efficient linear regression outsourcing to a cloud". IEEE Transactions on Cloud Computing, 7(3): 893-893. https://doi.org/10.1109/TCC.2017.2709299
[47] Mitrofanov, S., Semenkin, E. (2021). An approach to training decision trees with the relearning of nodes. In 2021 International Conference on Information Technologies (InfoTech), Varna, Bulgaria, pp. 1-5. https://doi.org/10.1109/InfoTech52438.2021.9548520

NOMENCLATURE

NN	Improved Neural Network
FAR	Functional coefficient Auto-Regressive
SF-GWO	Sun Flower-based Grey Wolf Optimization
alr	additive logratio
ODT	Outlier Detection Tree
ELM	Extreme Learning Machine
SFO	Sun Flower Optimization
INAR	INteger Auto-Regressive
ODCA	Outlier Detection method based on Cross-correlation Analysis
RiLOF	Robust outlier detection method
FRIOD	Feature-Rich Interactive Outlier Detection
GWO	Grey Wolf Optimization
VAR	Vector Auto-Regression
NCDF	Neighborhood Cumulative Density Function
PSO	Particle Swarm Optimization
MoNNAD	Median of Nearest Neighborhood Absolute Deviation
AUC	Area Under the ROC Curve
MiLOF	Memory-efficient incremental Local outlier
clr	centred logratio
NCDF	Neighborhood Cumulative Density Function
PSO	Particle Swarm Optimization
RODA	R-tree based Outlier Detection Algorithm
VAAD	Visualization Aided Anomaly Detection
WSNs	Wireless Sensor Networks
ilr	isometric logratio