Gaussian Projection Deep Extreme Clustering and Chebyshev Reflective Correlation based Outlier Detection

S. Rajalakshmi[1]
Research Scholar
Department of Computer Science
Periyar University, Salem, Tamilnadu, India

Dr. P. Madhubala[2]
Research Supervisor
Department of Computer Science
Periyar University, Salem, Tamilnadu, India

***Abstract*—Outlier detection or simply the task of point detection that are noticeably distinct and different from data sample is a predominant issue in deep learning. When a framework is constructed, these distinctive points can later lead to model training and compromise accurate predictions. Owing to this reason, it is paramount to recognize and eliminate them before constructing any supervised model and this is frequently the initial step when dealing with a deep learning issue. Over the recent few years, different numbers of outlier detector algorithms have been designed that ensure satisfactory results. However, their main disadvantages remain in the time and space complexity and unsupervised nature. In this work, a clustering-based outlier detection called, Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC) is proposed. First, Gaussian Random Projection-based Deep Extreme Learning-based Clustering model is designed. Here, by applying Gaussian Random Projection function to the Deep Extreme Learning obtains the relevant and robust clusters corresponding to the data points in a significant manner. Next, with the robust clusters, outlier detection time is said to be reduced to a greater extent. In addition, a novel Chebyshev Temporal and Reflective Correlation-based Outlier Detection model is proposed to detect outliers therefore achieving high outlier detection accuracy. The proposed approach is validated with the NIFTY-50 stock market dataset. The performance of the RPDEL-CRC method is evaluated by applying it to NIFTY-50 Stock Market dataset. Finally, we compare the results of the RPDEL-CRC method to the state-of-the-art outlier detection methods using outlier detection time, accuracy, error rate and false positive rate evaluation metrics.**

Keywords*—Outlier detection; clustering; Gaussian random projection; deep extreme learning; Chebyshev distance; temporal; reflective correlation

I. INTRODUCTION

Outliers are nothing but considered as data points or observations that plunge extraneous of an anticipated distribution or pattern and hence considered as the most-hottest topics as far as data mining is concerned. For example, if we were to perform data approximation with a Binomial distribution, then the outliers are the findings that do not emerge to go along with the pattern of a Binomial distribution. It discover anomalous data objects and are said to be of high use in several applications like detecting network intrusion, detecting fraudulent activities concerning credit card management, outlier detection in stock market to mention few. In the area of outlier detection, the ground truth is found to be seldom missing and hence machine learning techniques are extensively utilized in outlier detection research.

Most of the prevailing research works concentrates on outlier detection for categorical or numerical attribute data. A fuzzy rough set (FRSs) was proposed in [1] to detect outlier in mixed attribute data based on fuzzy rough granules. Initially, the granule outlier degree (GOD) was designed with the objective of characterizing the outlier degree of fuzzy rough granules via fuzzy approximation accuracy.

Followed by which, the outlier factor on the basis of fuzzy rough granules was designed by integrating GOD and respective weights to measure outlier degree of objects using fuzzy rough granules-based outlier detection (FRGOD) algorithm. With this both precision and recall were said to be improved. Despite improvement observed in terms of precision and recall, the time and space complexity were relatively high. To address on this aspect, Gaussian Random Projection-based Deep Extreme Learning-based Clustering model is first designed and then the outliers are detected. With this process, the time and space complexity involved in outlier detection will be reduced to a greater extent.

Iterative ensemble method with distance-based data filtering was proposed in [2] based on an iterative approach with the purpose of detecting outliers present in unlabeled data. The ensemble method was utilized in clustering the unlabeled data. Then, with the clustered data potential outliers were filtered in an iterative manner employing cluster membership threshold. This was performed in an iterative manner until Dunn index score for clustering was said to be maximized.

On the other hand, the distance-based data filtering eliminated the prospective outlier clusters from post-clustered data on the basis of the distance threshold utilizing the Euclidean distance measure from majority cluster as filtering factor. With this the improvement were found to be observed in terms of both precision and f-score value. Despite improvement observed in terms of both precision and f-score, by detecting possible outlier clusters based on weighted method, the false positive rate can be reduced to a greater extent. With this objective, Chebyshev Temporal and Reflective Correlation-based Outlier Detection model is

designed so that using Chebyshev distance based temporal factor obtains highly correlated data points, therefore reducing the false positive rate to a greater extent.

A. Objective and Contributions

The main objective of this research is to propose a novel cluster-based outlier detection method that performs clustering process and outlier detection separately in a significant manner. This clustering-based outlier detection method addresses the limitations of the earlier outlier detection methods by its multi-factor i.e., deep clustering and correlative outlier detection model. Further, the contributions of this paper include the following.

- To propose a novel Gaussian Random Projection-based Deep Extreme Learning-based Clustering algorithm to minimize a composite objective function, i.e., outlier detection time along with the improvement in error rate. The model minimizes the outlier detection time and reduces error rate during outlier detection via two different functions, Gaussian Random Projection and square gradient function.
- To design a new Chebyshev Temporal and Reflective Correlation-based Outlier Detection algorithm based on Chebyshev Temporal function and Reflective Correlative function that ensures accurate outlier detection.
- The proposed Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC) method has provided improved results for outlier detection time, accuracy, error and false positive rate as performance evaluation measures.

B. Organization of the Paper

The rest of the paper is organized as: The discussion about the obtainable modern outlier detection techniques is presented in Section II. In Section III, the proposed Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC) method has been discussed. In Section IV, Chebyshev temporal and reflective correlation-based outlier detection model is discussed. The discussion about the experimental setup and comparative analysis with an elaborate discussion is described in Section V and finally, the conclusions are presented in Section VI.

II. Related Works

The issue of outlier detection consists of detecting and eliminating malicious inferences from data. This problem is found to take place in several applications. However, outliers are frequently equipped by data stream that in turn influence the accuracy of data-based predictions. Hence, there arises an acute requirement to identify the outliers so as to enhance the data reliability.

A novel method to identify trajectory outlier group from large trajectory database using different types of algorithms was proposed in [3]. First, algorithms based on data mining were designed to identify the correlations between trajectory data and identify abnormal trajectories. Second, machine learning algorithms were applied to identify the group of trajectory outliers. Finally, convolution deep neural network were used to learn distinct different features to determine group of trajectory outliers, therefore enhancing runtime and accuracy performance.

Conventional outlier detection method however does not take into consideration the subset occurrence frequency and hence, the outliers being detected do not fit the definition of outliers. To address on this aspect, a two-phase minimal weighted rare pattern mining-based outlier detection method, called MWRPM-Outlier [4] was proposed to efficiently detect outliers based on the weight data stream.

A novel methodology to identify conjunct unusual human behaviors from large pedestrian data in smart cities was proposed in [5]. First, data mining was used followed by which convolution deep neural networks was explored that in turn identified distinct features to determine collective abnormal human behavior. With this both runtime and accuracy were said to be improved.

Despite several outlier detection algorithms are said to exist for scenarios necessitating numerical data, only a few prevailing methods can control categorical data. Moreover, the methods outlined for categorical data severely endure from two issues, low detection precision and high time complexity. Two novel outlier detection mechanisms for categorical data sets were proposed in [6]. First an entropy based method using Outlier Detection Tree (ODT) was designed followed by which second simple if-then rules were utilized for outlier detection. With these two integrated mechanisms both precision and computational complexity were improved to a greater extent.

Outlier detection has received paramount significance in the domain of data mining owing to the requirement to detect unusual events in different types of applications, to name a few being, fraud detection, intrusion detection and so on. Different types of outlier detection algorithms have been proposed in the recent past for utilization on static data sets employing a finite number of samples.

Probabilistic deep autoencoder was proposed in [7] with the objective of reconstructing measurements of power system that in turn can be employed in outlier detection. First, nonparametric distribution estimation method was utilized for obtaining information pertaining to uncertainty. Second, confidence intervals were acquired from estimated distribution and were further utilized as input. Finally, based on the multilayer encoding and decoding processes, the measurement intervals were reconstructed, with which outlier detection were made in an accurate manner.

Outlier detection methods employing machine learning are said to be receiving greater attention in the past few years in several domains. But, an ensemble of such outlier detection methods could improve the overall detection performance. An algorithm called, Average Selection and Ensemble of Candidates for Outlier Detection (DASEC-OD) was proposed in [8] for high dimensional data. A review of unsupervised outlier detection methods focusing on multi-dimensional data was investigated in [9].

With the exponential requirement in analyzing high speed data streams, the job of outlier detection becomes more

demanding as the conventional outlier detection method can no longer presume all data for processing. In [10], a Memory-efficient incremental Local Outlier (MiLOF) was proposed for large data streams, therefore ensuring accuracy to a greater extent.

Nowadays, there prevail very huge types of outlier detector methods that bestow satisfactory results. But their major disadvantage remains in their unsupervised characteristic in conjunction with the hyper parameters that has to be appropriately assigned for acquiring better performance.

An improved content-based outlier detection method was proposed in [11]. In [12], a novel supervised outlier estimator was designed. This was performed by pipelining an outlier detector in such a manner that the targets involved in the outlier detector were obtained in an optimal manner. However, these methods did not perform in a satisfactory manner in case of utilization of the complex datasets and hence suffer from noise introduced by outliers, specifically when the ratio of outlier was found to be high. To address this aspect, a framework called, Transformation Invariant AutoEncoder (TIAE) was proposed in [13] that in turn attained not only stability but also ensured high performance on outlier detection. A comprehensive review of outlier detection techniques were investigated in [14].

In several practical classification issues, a portion of outliers are said to exist in datasets that in turn would have heavy influence on the constructed model performance. A group method of data handing (GMDH) using neural network in outlier detection was proposed in [15]. A novel robust outlier detection method (RiLOF) based on Median of Nearest Neighborhood Absolute Deviation (MoNNAD) was designed in [16] that employed median of local absolute deviation of the samples to attain high detection performance.

Monitoring data including the significant information of monitored object forms the fundamentals for data mining and analysis. However, the data being monitored suffers from outlier pollution therefore causing negative influence on corresponding data processing. To address on this aspect, an outlier detection method based on stacked autoencoder (SAE) was proposed in [17]. The proposed SAE had the significant potentiality of feature extraction and heavily maintained the indigenous information of data to a greater extent.

Accuracy and time involved in outlier detection was not focused. To address this aspect, a Neighbor Entropy Local Outlier Factor was presented in [18] that with the aid of self organizing feature map not only improved accuracy but also reduced the execution time to a greater extent. Moreover, semantic information was focused on [19] for outlier detection employing meta path based outlier detection. Outlier detection based on the multivariable panel data was designed in [20] via correlation coefficient that in turn indicated high accuracy detection ability.

Motivated by the above mentioned techniques in this work, a novel cluster-based outlier detection method called, Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation is proposed (RPDEL-CRC). The elaborate description of RPDEL-CRC method is presented in the following sections.

III. Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC)

The proposed Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation method concentrates on the detection of outliers based on clustering. Methods designed based on cluster detect the outliers by placing data objects into distinct clusters. Here, the data objects in a data set are initially clustered. To design cluster-based outlier detection, the RPDEL-CRC method is split into two parts. Fig. 1 shows the block diagram of RPDEL-CRC method.

As illustrated in the figure below, the first part models robust cluster by means of Gaussian Random Projection-based Deep Extreme Learning. Here, the clustering based outlier detection initiates the outlier detection process by clustering the given input dataset, Nifty 50 Stock Market Data (2000 – 2021). Hence, to be more specific outliers are considered as data points within deviating clusters or the data points that deviate to the formed clusters. The second part uses the Chebyshev Temporal and Reflective Correlation-based Outlier Detection algorithm to detect outlier with minimum falsification. In this section, we first explain all prerequisites of the proposed method with a system model, and then finally we describe the proposed method.

A. System Model

Let ‘P ∈R^(m*n)’ represent a matrix with ‘m’ rows and ‘n’ columns of real numbers ‘P_ij∈R’. The matrix ‘P’ denotes a dataset ‘DS’ that includes the data for outlier analysis. The ‘n’ columns are called features and on the other hand, the ‘m’ are referred to as data points. Then, vector ‘〖DP〗_i∈R^n’ refers to the data point, which is a row in ‘P’. The matrix ‘P’ then consists of ‘m’ data points ‘DP={ 〖DP〗_1, DP_2,…., 〖DP〗_n }’. Then, with the aid of the outlier detection algorithm the outliers present in the dataset ‘DS’ are detected. Finally, the overall feature space represents the vector space defined by the given features that in turn estimates the characteristics of the examined occurrence or event. Inliers are detected in subsets of the overall feature space and referred to as normal regions or normal data points. To be more specific, inliers are considered to as the data points in the normal regions. On the other hand, an outlier is a data point that does not belong in the normal region.

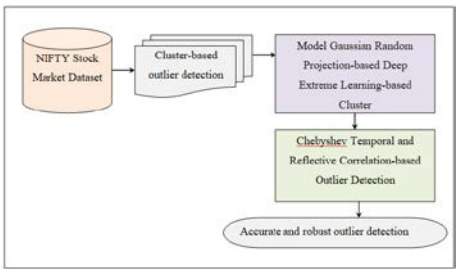

Fig. 1. Block Diagram of Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation Method.

B. Case Analysis of Outlier Detection Accuracy

To detect outliers based on the cluster in a given dataset, the data has to be initially clustered. In this paper, Gaussian Random Projection-based Deep Extreme Learning model is first employed for clustering. The objective behind the design of Gaussian Random Projection-based Deep Extreme Learning model remains in training feed forward network from a raw training data set with 'N' samples, '{P,Q}={P_i,Q_i}_(i=1,2,3, ⋯.,N)', with 'P_i ∈ R^d' and 'Q_i' represents 'M-dimensional' binary vector where one entry denotes '1' representing the cluster that 'P_i' belongs to. Fig. 2 shows the block diagram of Gaussian Random Projection-based Deep Extreme Learning-based Clustering model.

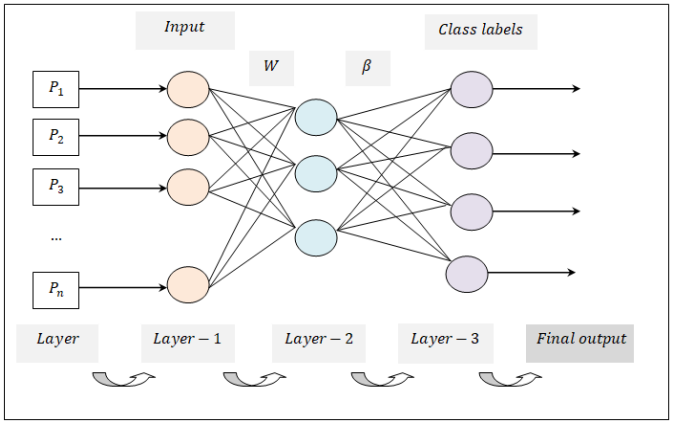

Fig. 2. Block Diagram of Gaussian Random Projection-based Deep Extreme Learning-based Clustering Model.

As shown in the above figure, the training process of GRP-DEL includes two steps. In (1), the hidden layer with 'K' nodes employing distinct numbers of neurons are constructed. Next, for the 'i-th' hidden layer node, a 'd-dimensional' vector 'x_j' and a metric 'y_j' are generated in an arbitrary manner. Then, for each input vector 'P_i', the pertinent output on the 'i-th' hidden layer node is obtained by utilizing Sigmoid activation function. This is mathematically stated as given below.

$$g(P_i, x_j, y_j) = \frac{1}{1+\exp\left[-\left(x_j^T * P_i + y_j\right)\right]} \quad (1)$$

Then, using the resultant value of the above Sigmoid activation function, the hidden layer outputs the matrix as given below.

$$H = \begin{bmatrix} g(P_1, x_1, y_1) & \cdots & g(P_1, x_K, y_K) \\ \cdots & \cdots & \cdots \\ g(P_N, x_1, y_1) & \cdots & g(P_i, x_K, y_K) \end{bmatrix}_{N*K} \quad (2)$$

In (2), an 'M-dimensional' binary vector 'α_j' represents the output weight that associates the 'j-th' hidden layer with the resultant output node. Here, a random projection is applied that states that if points associating the 'j-th' hidden layer in a vector space are of sufficiently high dimension, then the 'j-th' hidden layer may be projected into a lower-dimensional space in such a manner that it preserves the distances between points (therefore minimizing dimensionality). With original input vector being 'P_(K*M)', using a random 'K*d' matrix dimensional matrix 'R', then the projection of data on to a lower dimensional subspace is mathematically formulated as given below.

$$P_{K*M} = R_{K*d} P_{d*M} \quad (3)$$

Then, with the above lower dimensional subspace random projection dimensionality of set of points and output matrix 'Q' is mathematically stated as given below.

$$H.\alpha = Q \quad (4)$$

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_K \end{bmatrix}_{K*M} ; Q = \begin{bmatrix} Q_1 \\ Q_2 \\ \cdots \\ Q_N \end{bmatrix}_{N*M} \quad (5)$$

Next, with the resultant matrices 'H' and 'Q', the objective of GRP-DEL model remains in solving the output weights 'α' by reducing the losses of prediction errors, leading to the following equation.

$$\alpha_i(n) = \alpha_i(n-1) - \beta_i(n) \frac{MA_i(n)}{\sqrt{G_i(n)}} \quad (6)$$

From the above (6), '〖MA〗_i (n)' symbolizes the moving average of feature or attribute 'i' at iteration 'n', with square gradient denoted by 'G_i (n)' and learning rate 'β_i (n)' respectively.

$$MA_i(n) = \gamma_n MA_i(n-1) + (1-\gamma_n) \quad (7)$$

$$G_i(n) = \theta_n G_i(n-1) + (1-\theta_n) \quad (8)$$

$$\beta_i(n) = \beta_i(n-1) \frac{\sqrt{(1-\theta_n)^n}}{(1-\gamma_n)^n} \quad (9)$$

From the above (7), (8) and (9) the factors 'γ_n' and 'θ_n' are utilized in fine tuning the decay rates of moving averages close to one (i.e., 'γ_n=0.85' and 'θ_n=0.9'). The pseudo code representation of Gaussian Random Projection-based Deep Extreme Learning-based Clustering is given below.

Algorithm 1: Gaussian Random Projection-based Deep Extreme Learning-based Clustering

Input: Dataset 'DS', data points '$DP = \{DP_1, DP_2, \ldots, DP_n\}$'
Output: obtain cluster 'Q_i' corresponding to 'P_i' in computationally efficient and precise manner
1: **Initialize** 'm' rows and 'n' columns
2: **Begin**
3: **For** each Dataset 'DS' with data points 'DP' and input vector 'P_i'
4: Obtain pertinent output on the '$i-th$' hidden layer employing Sigmoid activation function as in (1)
5: Obtain output matrix via hidden layer as in (2)
6: Evaluate Gaussian Random Projection as in (3)
7: Estimate hidden layer output and calculate the output matrix as in (4) and (5)
8: **Repeat** (training of neural networks)
9: Solve output weights by minimizing prediction loss error as in (6)
10: Treat each row of 'Q' as a point and cluster them into 'K' clusters
11: Estimate learning rates for cluster parameter as in (7), (8) and (9)
12: **Until** (first-order gradients for neural networks is arrived at)
13: **Return** 'Q'
14: **End for**
15: **End**

As given in the above Gaussian Random Projection-based Deep Extreme Learning-based Clustering algorithm, with the objective of reducing the outlier detection time along with the improvement in precision, two different functions are employed. First, by employing Gaussian Random Projection the dimensionality of data is said to be reduced by projecting original input space (i.e., the raw data) with the aid of a sparse random matrix. Second, by estimating the learning rate by means of square gradient minimizes the error involved during the process of clustering to a greater extent. As a result, with these two function, clusters are formed both in a computationally efficient and precise manner.

IV. Chebyshev Temporal and Reflective Correlation-based Outlier Detection Model

Outlier detection remains to be one of the primary step in data mining tasks. The motive behind the outlier detection strategy here is to identify the features or parameters that are counterfeit from several other features. Different types of outlier detection models are said to exist. In order to determine the perpetual temporal outliers, we obtain outliers based on distance measures by analyzing temporal values of the objects employing Chebyshev Temporal and Reflective Correlation-based Outlier Detection model. Fig. 3 shows the block diagram of Chebyshev Temporal and Reflective Correlation-based Outlier Detection model.

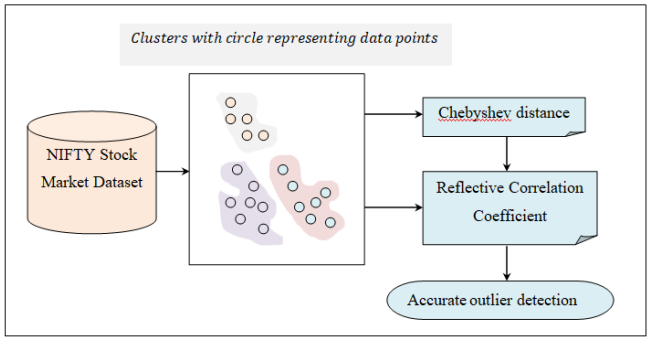

Fig. 3. Block Diagram of Chebyshev Temporal and Reflective Correlation-based Outlier Detection Model.

As shown in the above figure, with the obtained clusters for the given dataset 'DS', in a 'd-dimensional' vector, with data point denoted by 'DP={DP[1],DP[2],…DP[d] }' at time instance 'T', distance between two points ' ⟦DP⟧ _i' and 'DP_j' employing Chebyshev distance is mathematically, expressed as given below.

$$Dis\left(DP_i, DP_j\right) = Max\left(\left|DP_i - DP_j\right|\right) \quad (10)$$

From the above (10), by employing the Chebyshev distance measure 'Dis (⟦DP⟧ _i,DP_j)', the greatest difference between two vectors (i.e., the data points) along any coordinate dimension (i.e., the cluster) is evaluated based on the maximum 'Max(| ⟦DP⟧ _i- ⟦DP⟧ _j |)' distance along one axis. To be more specific, based on the principle of chessboard distance as the minimum number of moves required by a king to go from one square to another is utilized; by means of Chebyshev distance, the overall outlier detection accuracy is said to be improved. Next, with the assumption of 'm' clusters ' ⟦Cl⟧ _1, ⟦Cl⟧ _2,…, ⟦Cl⟧ _m' the centroid data point 'CP' is then measured as given below.

$$Cl_i CP\ [i] = \frac{\sum_{P \in Cl_i} DP[i]}{|Cl_i|} \quad (11)$$

With the above determination of the centroid (11), with the assumption that in a cluster 'Cl' most of the normal data points are hardly encircling the centroid data point of cluster 'Cl', the abnormal data points or the outlier are those generally farther from the centroid data point. Then the updated weight of a centroid data point is mathematically stated as given below.

$$W(DP) = \frac{Dis(DP, Cl_i CP\ [i]\ )}{\sum_{S \in Neigh(DP)} Dis(S, Cl_i CP\ [i]\ )} \quad (12)$$

From the above (12), the weight of data point 'DP' in cluster ' ⟦Cl⟧ _i' is estimated based on the neighbors of 'Neigh(DP)' in ' ⟦Cl⟧ _i'. Finally, to reflect the robustness and direction of linear correlation between two data points and minimizing the false positive cases, Reflective Correlation Coefficient is applied. RCC function is employed to evaluate the amount of dependency between two distributions of normalized scores 'G_i^Norm (DP),G_j^Norm (DP)' and is mathematically stated as given below.

$$RCC\left(DP_i, DP_j\right) = \frac{W(DP_i)W(DP_j)}{\sqrt{(\sum DP_i)^2 (DP_j)^2}} \quad (13)$$

From the above (13), reflective correlation coefficient 'RCC' is obtained based on the weighted data points 'W(⟦DP⟧ _i)' and 'W(⟦DP⟧ _j)' respectively. The final form of the objective function for minimizing the false positive cases of the 'j-th detector' is mathematically stated as given below.

$$Res_j = \left[G_j^{Norm}(DP)\right] - \left[G_j^{Norm}(DP_O)\right] \quad (14)$$

From the above (14) 'G_j^Norm' forms the normalized score function of the 'j-th detector', 'DP' denoting the data points with contaminated dataset and ' ⟦DP⟧ _O' denoting the outliers. The pseudo code representation of Chebyshev Temporal and Reflective Correlation-based Outlier Detection is given.

Algorithm 2 Chebyshev Temporal and Reflective Correlation-based Outlier Detection

Input: Dataset 'DS', data points '$DP = \{DP_1, DP_2, \ldots, DP_n\}$'
Output: Accurate Outlier Detection
1: **Initialize** time instance 'T'
2: **Begin**
3: **For** each Dataset 'DS' with data points 'DP' and cluster 'Q_i'
4: Evaluate distance between data points 'DP_i' and 'DP_j' as in (10)
5: **For** each cluster 'Q_i'
6: Evaluate centroid data point as in (11)
7: Evaluate weight of data point as in (12)
8: Estimate Reflective Correlation Coefficient as in (13)
9: Obtain final form of the objective function of the '$j-th\ detector$' as in (14)
10: **Return** (outliers 'DP_O')
11: **End for**
12: **End for**
13: **End**

As given in the above Chebyshev Temporal and Reflective Correlation-based Outlier Detection algorithm, with the objective of improving the outlier detection accuracy with minimum falsification, two different functions are employed. First with the obtained clusters based on the data points, Chebyshev distance function is applied to estimate the difference between two data points along any cluster. Based on the minimum number of positioning between clusters, according to time, results are obtained, therefore ensuring outlier detection accuracy. Second by employing the Reflective Correlation Coefficient function dependency between two distributions or data points are obtained therefore reducing the false positive rate to a greater extent. Finally, the outliers are obtained.

V. Experimental Setup

In this section, experimental analysis of the Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC) method for outlier detection in data mining is presented. In this section, the performance of the proposed RPDEL-CRC is compared with the state-of-the-art methods, fuzzy rough granules-based outlier detection (FRGOD) [1] and Iterative ensemble method with distance-based data filtering [2] using NIFT-50 Stock Market Dataset (https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market-data). Simulations are performed in R Programming language. Fair comparison between proposed RPDEL-CRC method and existing fuzzy rough granules-based outlier detection (FRGOD) [1] and Iterative ensemble method with distance-based data filtering [2] are made for evaluating different parameters like, outlier detection time, outlier detection accuracy, false positive rate and error for different iterations.

A. Case Analysis of Outlier Detection Time

The first metric significant for cluster based outlier detection is the time consumed in detecting the outlier. To be more specific, outlier detection time refers to the time consumed in detecting the outliers. Lower the outlier detection time more efficient the method is said to be because earlier the time consumed in detecting the outlier is more efficient the method is. The outlier detection time is mathematically stated as given below.

$$OD_{time} = \sum_{i=1}^{n} Samples_i * Time\left[Res_j\right] \quad (15)$$

From the above (15), the outlier detection time ‘〖OD〗_time’ is measured based on the samples involved in the simulation process ‘〖Samples〗_i’ and the time consumed in detecting the outliers ‘Time [〖Res〗_j]’. It is measured in terms of milliseconds (ms). Table I given below shows the results of outlier detection time observed for three different methods, RPDEL-CRC, FRGOD [1] and Iterative ensemble method with distance-based data filtering [2].

Fig. 4 illustrated above shows the outlier detection time with respect to 50000 different numbers of samples obtained at different intervals from different companies stock values between years 2007 and 2021. These curves are plotted with increasing cardinality of training samples ranging between 5000 and 50000. With the increasing cardinality, the number of samples involved in analysis for outlier detection increases and therefore an increase in the outlier detection time is observed. However, simulations with 5000 samples observed ‘250ms’ for detecting outliers with respect to single stock sample using RPDEL-CRC, ‘350ms’ for detecting outliers with respect to single stock sample using [1] and ‘450ms’ for detecting outliers with respect to single stock sample using [2]. From this analysis it is inferred that the outlier detection time using RPDEL-CRC is comparatively lesser than [1] and [2]. The reason behind is the incorporation of Gaussian Random Projection-based Deep Extreme Learning-based Clustering model. By applying this model, dimensionality of data or data points are said to be reduced using Gaussian Random Projection based on projecting original input space (i.e., the raw data) with the aid of a sparse random matrix. With this, data points considered to be outliers are obtained that in turn assist in detecting the outliers altogether. Therefore, the outlier detection time using RPDEL-CRC method is found to be reduced by 20% compared to [1] and 37% compared to [2].

TABLE I. Tabulation for Outlier Detection Time

Samples	Outlier detection time (ms)		
	RPDEL-CRC	***FRGOD***	***Iterative ensemble method with distance-based data filtering***
5000	250	350	450
10000	295	395	555
15000	355	435	635
20000	410	485	680
25000	435	535	745
30000	485	625	795
35000	525	685	835
40000	595	745	890
45000	685	800	920
50000	735	835	955

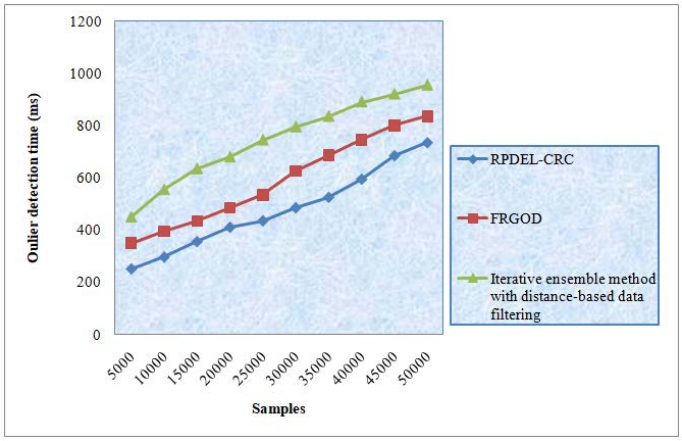

Fig. 4. Graphical Representation of Outlier Detection Time.

B. Case Analysis of Outlier Detection Accuracy

The second parameter of significance for cluster based outlier detection is the accuracy rate. In other words, the outlier detection accuracy is measured based on percentage ratio

between the samples involved in simulation process '〖Samples〗_i' and the actual samples accurately detected with outliers '〖Samples〗_ODC'. This is mathematically stated as given below.

$$OD_{acc} = \sum_{i=1}^{n} \frac{Samples_{ODC}}{Samples_i} * 100 \quad (16)$$

Table II given shows the results of outlier detection accuracy observed for three different methods, RPDEL-CRC, FRGOD [1] and Iterative ensemble method with distance-based data filtering [2].

Fig. 5 illustrates the outlier detection accuracy for 50000 different stock samples obtained from the NIFTY-50 stock dataset at different time instances. From the figure it is inferred that the outlier detection accuracy is found to be inversely proportional to the stock samples involved in the simulation process. In other words, increasing the stock samples for detecting the outlier causes an increase in the overall data points involved in the process and this in turn minimizes the outlier detection accuracy. However, sample simulations performed with 5000 samples 4845 samples were accurately detected with outliers as it is using RPDEL-CRC, 4755 samples using [1] and 4695 samples using [2]. With this the overall accuracy using the three methods were found to be 96.90%, 95.1% and 93.9% respectively. The overall accuracy was found to be improved using RPDEL-CRC upon comparison with [1] and [2]. The reason behind the outlier detection accuracy improvement was owing to the application of Chebyshev distance function. By applying this distance function, the difference between two data points along any cluster was first evaluated. Then, on the basis of the minimum number of positioning between clusters, according to time, results were obtained, i.e., outliers were detected, therefore ensuring outlier detection accuracy. This in turn improved the outlier detection accuracy using RPDEL-CRC method by 3% compared to [1] and 7% compared to [2].

C. Case Analysis of False Positive Rate

False positive rate is measured as the ratio between the numbers of negative events (i.e., negative outliers) wrongly categorized as positive (i.e., outliers) and the total number of actual negative events (i.e., actual outliers). This is mathematically stated as given below.

$$FPR = \frac{FP}{FP+TN} \quad (17)$$

From the above (17), the false positive rate 'FPR' is measured based on the false positive samples 'FP' (i.e., actually the data are not outliers) and the true negative samples 'TN' (i.e., outliers detected as outliers) respectively. Table III given shows the results of false positive rate observed for three different methods, RPDEL-CRC, FRGOD and Iterative ensemble method with distance-based data filtering [2].

TABLE II. TABULATION FOR OUTLIER DETECTION ACCURACY

Samples	Outlier detection accuracy (%)		
	RPDEL-CRC	***FRGOD***	***Iterative ensemble method with distance-based data filtering***
5000	96.9	95.1	93.9
10000	95.35	94.35	91.15
15000	94.15	92.85	90.35
20000	94.05	91.55	88.15
25000	93.85	91	88
30000	93.25	90.85	87.35
35000	93	90.25	86
40000	92.55	89.85	85.25
45000	92.15	89.15	84.35
50000	92	89	83

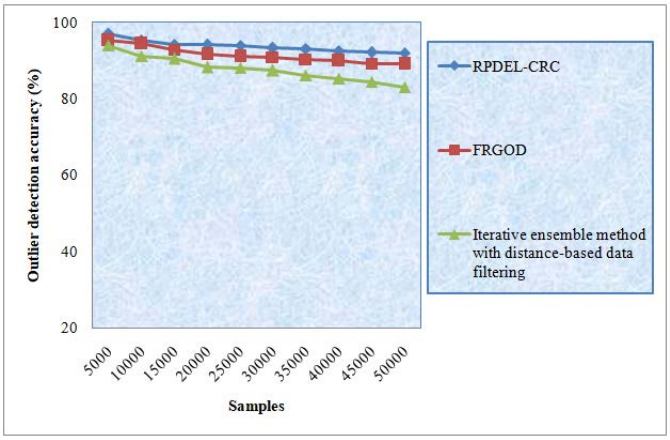

Fig. 5. Graphical Representation of Outlier Detection Accuracy.

TABLE III. TABULATION FOR FALSE POSITIVE RATE

Samples	False positive rate (%)		
	RPDEL-CRC	***FRGOD***	***Iterative ensemble method with distance-based data filtering***
5000	0.007	0.015	0.025
10000	0.015	0.018	0.026
15000	0.018	0.025	0.028
20000	0.02	0.028	0.033
25000	0.022	0.031	0.035
30000	0.025	0.032	0.036
35000	0.027	0.035	0.038
40000	0.029	0.037	0.042
45000	0.035	0.039	0.044
50000	0.038	0.042	0.048

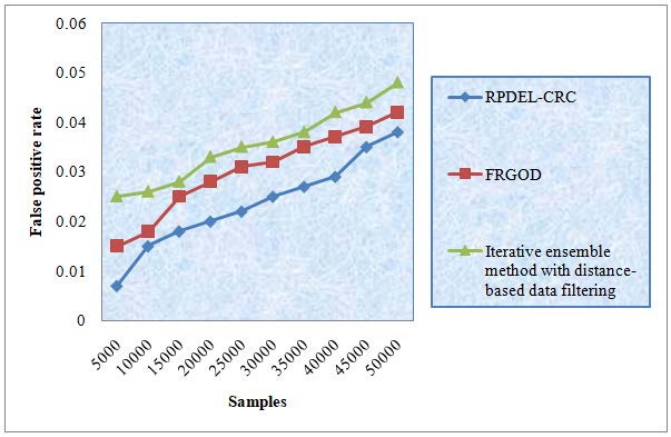

Fig. 6. Graphical Representation of False Positive Rate.

Fig. 6 given depicts false positive rate for different stock samples. From the figure, it is inferred that the false positive rate also increases with the increase in the number of stock samples involved in the simulation and hence the false positive rate is found to be directly proportional to the stock samples or samples. However, simulations conducted for 5000 samples show a false positive rate of 0.007 using RPDEL-CRC, 0.015 using FRGOD [1] and 0.025 using Iterative ensemble method with distance-based data filtering [2]. From this, it is observed that the false positive rate is comparatively lesser using RPDEL-CRC when compared to [1] and [2]. The reason behind is the application of Chebyshev Temporal and Reflective Correlation-based Outlier Detection algorithm. By applying this algorithm, dependency between two distributions of data points or data are separated. First, according to different weight of data points, i.e., based on the neighbors or data points in cluster, updated weight of a centroid data point is obtained. Next, with the identified updated weight of a centroid data point, outliers are detected based on the linear correlation between data points. Hence, by applying different updated weight of a centroid data point for each cluster, false positive rates are significantly reduced using RPDEL-CRC method by 24% compared to [1] and 36% compared to [2].

D. Case Analysis of Error Rate

Finally, the error rate involved in outlier detection is discussed in this section. The error rate is one of the significant parameters involved in the outlier detection process. This is owing to the reason that while clustering certain data points are said to be misplaced in the adjoining clusters, therefore resulting in error. This error rate is mathematically stated as given below.

$$ER = \left(\frac{V_{actual} - V_{expected}}{V_{expected}}\right) * \% \quad (18)$$

From the above (18), the error rate 'ER' is measured based on the actual value 'V_actual' or the actual data point positioning and the expected value 'V_expected' or the expected data positioning. It is measured in terms of percentage (%). Finally, Table IV lists the error rate obtained using the (18).

Finally, Fig. 7 illustrates the error rate observed during the process of outlier detection. From the figure, an increasing trend is found to be observed using all the three methods, RPDEL-CRC, FRGOD [1] and Iterative ensemble method with distance-based data filtering [2] increasing the stock samples. This is due to the reason that with the increase in the stock samples provided as input obtained during different time instances from different companies, first, clusters are performed. While performing the clustering based on data points certain data points due to temporal instances cause a small shift in the positioning of clusters. This in turn results in the deviation and therefore error. However, simulations conducted with 5000 samples with actual data positioning observed to be 53, the expected data positioning using the three methods were observed to be 48, 45 and 43. Hence, the error rate were found to be 9.4%, 15.09% and 18.86% respectively using the three methods, therefore reducing the error with RPDEL-CRC method. The reason behind the minimization of error using RPDEL-CRC method was due to the application of Gaussian Random Projection-based Deep Extreme Learning-based Clustering algorithm. By applying this algorithm, the learning rate for solving the output weights were estimated by means of square gradient. As a result, the error rate using RPDEL-CRC was said to be reduced by 28% compared to [1] and 45% compared to [2].

TABLE IV. TABULATION FOR ERROR RATE

Samples	Error rate (%)		
	RPDEL-CRC	*FRGOD*	*Iterative ensemble method with distance-based data filtering*
5000	9.4	15.09	18.86
10000	9.75	16.15	21.32
15000	10.35	16.35	22
20000	10.85	17.25	22.85
25000	11.35	17.85	24.35
30000	13.15	18.35	24.85
35000	15.25	20	25
40000	17.35	21.35	28
45000	19.55	22.45	29.35
50000	21.25	23	30

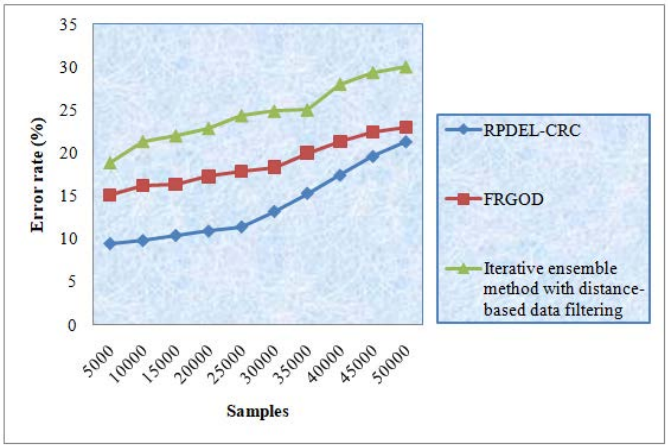

Fig. 7. Graphical Representation of Error Rate.

VI. Conclusion

In spite of the fact that there has been an improvement in outlier detection, nevertheless mushrooming outliers are still found in its disastrous intents. In this day and age, it has become a big ultimatum that the behavior of outliers has to be monitored in many data mining tasks. In this paper, we proposed a new outlier detection method, called, Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC). The main contributions of our proposed RPDEL-CRC method is to the field of outlier detection that reduces the outlier detection time, error and false positive rate involved with maximum accuracy. The proposed method reduces the outlier detection time and error for operating the outlier detection via Gaussian Random Projection-based Deep Extreme Learning model that initially performs the clustering process by means of Gaussian Random Projection function. Next, with behavior grouping and clustering using Deep Extreme Learning, false positive rate is reduced in a timely manner. Third, the actual outlier detection process based on the clustering results is performed by means of Chebyshev Temporal and Reflective Correlation-based Outlier Detection model. Simulations were performed to evaluate the performance of RPDEL-CRC, FRGOD and Iterative ensemble method with distance-based data filtering method. Simulation results revealed that the proposed RPDEL-CRC method outperforms, FRGOD and Iterative ensemble method with distance-based data filtering method implementations, in terms of outlier detection time, accuracy, error rate and false positive rate.

References

[1] Zhong Yuan, Hongmei Chen, Tianrui Li, Binbin Sang, and Shu Wang, "Outlier Detection Based on Fuzzy Rough Granules in Mixed Attribute Data," IEEE Transactions on Cybernetics, May 2021 [fuzzy rough granules-based outlier detection (FRGOD)].

[2] Bodhan Chakraborty, Agneet Chaterjee, Samir Malakar, and Ram Sarkar, "An iterative approach to unsupervised outlier detection using ensemble method and distance-based data filtering," Springer Complex & Intelligent Systems, Feb 2022, [Iterative ensemble method with distance-based data filtering].

[3] Asma Belhadi, Youcef Djenouri, Djamel Djenouri, Tomasz Michalak, and Jerry Chun-Wei Lin, "Deep Learning Versus Traditional Solutions for Group Trajectory Outliers," IEEE Transactions on Cybernetics, Dec 2020.

[4] Saihua Cai, Ruizhi Sun, Shangbo Hao, Sicong Li, and Gang Yuan, "An Efficient Outlier Detection Approach on Weighted Data Stream Based on Minimal Rare Pattern Mi," IEEE Xplore, Oct 2019.

[5] Asma Belhadi, Youcef Djenouri, Gautam Srivastava, Djamel Djenouri, Jerry Chun-Wei Lin, and Giancarlo Fortino, "Deep learning for pedestrian collective behavior analysis in smart cities: A model of group trajectory outlier detection," Information Fusion, Elsevier, Feb 2021.

[6] Hongwei Du, Qiang Ye, Zhipeng Sun, Chuang Liu, and Wen Xu, "FAST-ODT: A Lightweight Outlier Detection Scheme for Categorical Data Sets," IEEE Transactions of Network Science and Engineering, Oct 2020].

[7] You Lin, and Jianhui Wang, "Probabilistic Deep Autoencoder for Power System Measurement Outlier Detection and Reconstruction," IEEE Transactions on Smart Grid, Jul 2019.

[8] N. Jayanthi, Burra Vijaya Babu, and N. Sambasiva Rao, "An ensemble framework based outlier detection system in high dimensional data using Tree Technique," Materials Today: Proceedings, Elsevier, Nov 2020.

[9] Atiq ur Rehman, and Samir Brahim Belhaouari, "Unsupervised outlier detection in multidimensional data," Journal of Big Data, Springer, Jun 2021.

[10] Mahsa Salehi, Christopher Leckie, James C. Bezdek, Tharshan Vaithianathan and Xuyun Zhang, "Fast Memory Efficient Local Outlier Detection in Data Streams," IEEE Transactions on Knowledge and Data Engineering, Nov 2016.

[11] Huiping Li, BinWang, and Xin Xie, "An improved content-based outlier detection method for ICS intrusion detection," EURASIP Journal onWireless Communications and Networking, Springer, Aug 2020.

[12] Ángela Fernández, Juan Bella, and José R. Dorronsoro, "Supervised outlier detection for classification and regression," Neurocomputing, Springer, Feb 2022.

[13] Zhen Cheng, En Zhu, Siqi Wang, Pei Zhang, and Wang Li, "Unsupervised Outlier Detection via Transformation Invariant Autoe," IEEE Access, Mar 2021.

[14] Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad, "Progress in Outlier Detection Techniques: A Survey," IEEE Access, Aug 2019.

[15] Xie Ling, Jia Yanlin, Xiao Jin, Gu Xin, and Huang Jing, "GMDH-Based Outlier Detection Model in Classification Problems," Journal of Systems Science and Complexity, Springer, Feb 2020.

[16] Ali Degirmenci and Omer Karal, "Robust Incremental Outlier Detection Approach Based on a New Metric in Data Streams," IEEE Access, Nov 2021.

[17] Fangyi Wan, Gaodeng Guo, Chunlin Zhang, Qing Goo, and Jie Liu, "Outlier Detection for Monitoring Data Using Stacked Autoencoder," IEEE Access, Nov 2019.

[18] Ping Yang, Dan Wang, Zhuojun Wei, Xiaolin Du, and Tong Li, "An Outlier Detection Approach Based on Improved Self-Organizing Feature Map Clustering Algorithm," IEEE Access, Aug 2019.

[19] Lu Liu, and Shang Wang, "Meta-path-based outlier detection in heterogeneous information network," Frontiers of Computer Science, Springer, Nov 2019.

[20] Jintao Song, Shengfei Zhang, Fei Tong, Jie Yang, Zhiquan Zeng, and Shuai Yuan, "Outlier Detection Based on Multivariable Panel Data and K-Means Clustering for Dam Deformation Monitoring Data," Advances in Civil Engineering, Hindawi, Dec 2021.