European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 4, 2020

An Ensemble Framework Based Outlier Detection System in High Dimensional Data

[1]N Jayanthi, [2]Dr Burra Vijaya Babu, [3]Dr N Sambasiva Rao
[1]Research Scholar, [2,3]Professor
[1,2] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[3] Vardhaman College of Engineering Hyderabad.
[1]jneelampalli.phd@gmail.com,[2] vijay_gemini@kluniversity.in, [3]snandam@gmail.com.

Abstract
Machine learning based outlier detection methods are widely used in various domains. However, an ensemble of such detection methods could leverage detection performance. The existing ensemble methods made up of multiple unsupervised learning algorithms lack in ideal strategy for choosing right candidates as constituent detectors. It resulted in mediocrity in model stability and accuracy. To overcome this problem, in this paper, we propose an ensemble framework based outlier detection system in high dimensional data. It has ideal mechanism for effectively choosing base outlier detectors. Out of many candidate outlier detectors, the ones that yield highest performance are combined. An algorithm named Average Selection and Ensemble of Candidates for Outlier Detection (ASEC-OD). Many real world datasets are used for empirical study. The results of experiments revealed that the proposed framework outperforms many existing methods.

Keywords –Machine learning, outlier detection, ideal selection of candidates for outlier ensemble, outlier detection system

1. INTRODUCTION

Outlier detection methods discover anomalous data objects in the given dataset. They are very useful in different applications like network intrusion detection ,credit card fraud detection , video surveillance[1] and social networks[2] to mention few. Outliers can be predicted[3] or detected. In the area of outlier detection, the ground truth is often missing and this is the reason why unsupervised machine learning is widely used in outlier detection research[4,5]. Literature is rich in outlier detection methods in various fields few are present in [6,7,8]. For instance, the subspace method for outlier detection is used in [9], [10] and [11]. These methods are famous as they exploit localized regions to detect abnormalities in high-dimensional data. It is also found that reverse nearest neighbour based methods are good for outlier detection as explored in [12] and [13]. It is understood that dimensionality reduction has its positive impact on increasing performance of outlier detection methods. It is evident in the research insights of [14], [15], [16] and [17]. When dimensions are reduced, it improves accuracy in outlier detection as it reduces search space. Towards this end, feature extraction method is generally used as in [18].

Different methods used for outlier detection can be combined in order to have better performance. This approach is known as ensemble learning. It is widely used approach for leveraging outlier detection recognition as discussed in [19], [18], [20], [21], [15], [22] and [23]. Many ensemble methods do follow unsupervised methods such as the ones explored in [24] and [12]. The problem with the existing ensemble outlier detection methods is that they are not using ideal strategy to identify better constituent detectors leading to mediocrity in performance. This problem is overcome this in this paper by proposing a framework that carefully selects constituent detectors using an ideal selection approach. Our contributions in this paper are as follows.

1. We propose an ensemble framework based outlier detection system in high dimensional data for leveraging performance by exploiting knowledge of multiple detectors.
2. We propose an algorithm named Average Selection and Ensemble of Candidates for Outlier Detection (ASEC-OD) to overcome mediocrity problem of ensemble learning.

3. We built a prototype application to evaluate the proposed framework by comparing it with multiple state of the art outlier detection methods.

The remainder of the paper is structured as follows. Section 2 reviews literature on ensemble outlier detection methods. Section 3 presents the proposed framework. Section 4 provides results of empirical study. Section 5 concludes the paper gives directions for future scope of the research.

2. RELATED WORK

This section reviews literature on outlier ensembles. Aggarwal [19] explored different kinds of outlier ensembles. They analysed them in the context of both clustering and classification. They found many common combination functions such as maximum function, averaging function, damped averaging and pruned averaging. With respect to unsupervised methods, they found intermediate evaluation, diversity and consensus issues. Different methods associated with classification used for ensemble analysis are boosting, bagging, random forests, model averaging and bucket of models. They envisaged that ensemble outlier methods provide more useful business intelligence (BI). Aggarwal [9] discussed about several kinds of technique used for outlier detection based on subspace method. They include rarity-based, unbiased and aggregation-based. They opined that outlier detection is the most difficult problem out of all problems related to subspaces. Chakraborty et al. [18] explored outlier type scenarios with ensemble and deep feature learning approaches combined. Features are extracted using stacked autoencoders. Majority voting is used in order to detect outliers. In future, they intend to focus on multiple outlier type scenarios. Zhang et al. [20]proposed a genetic framework known as Locality Sensitive Hashing (LSH) forest. It is an ensemble method with fast tree isolation. They intended to improve it to support low-latency anomaly detection. Trittenbach and Bohm [10] proposed an algorithm based on dimension based subspace search to discover outliers. They also proposed a heuristic known as Greedy Maximum Deviation (GMD). It is based on certain criteria known as runtime, number of subspaces, robustness and search-result quality.

Zhang et al. [11] studied high-dimensional data streams in order to discover anomalies based on sliding window. Their method includes data preparation, feature normalization, reference set derivation, relevant subspace selection, computing local outlier scores and determination of control limit. It performs model training offline and actual anomaly detection online. In [21] foundation is made on outlier ensemble methods that exploit trade-off between bias and variance. The framework is widely used in order to solve generalization error in classification methods. It minimizes the reducible generalization error in ensembles. It is understood that a high bias detector is less sensitive while high variance detector is more sensitive to data variations. Therefore, it is important to control both bias and variance leading to reduction in generalization error.

Dang et al. [14] proposed an algorithm that uses discriminative features in order to discover outliers. In the process, it reduces dimensionality and improves performance. They found that their method works better with ensemble approach. Chen et al. [25] considered Intelligent Transportation System (ITS) as case study for outlier detection. They studied different outlier detection methods useful for ITS. They followed different strategies such as partition, ensemble learning and average LOF for better recognition of outliers. In future, they intend to use different outlier methods combined for better results. Liu et al. [24] proposed a novel outlier detection method known as Single-Objective Generative Adversarial Active Learning (SO-GAAL). Gupta et al. [26] made review of outlier detection methods that work on temporal data. Their study includes different methods that work on temporal data such as time series data, data streams, distributed data, spatio-temporal data and network data. Domingues et al. [27] studied different outlier detection categories known as probabilistic methods, distance based methods, neighbour based methods, information theory based methods, neural networks, isolation methods and domain based methods.

Radovanovic et al. [12] explored outlier detection methods and found that the proposition “reverse nearest neighbour discovery helps in detection of outliers” is true. They used Influenced outlierness (INFLO) and Local Outlier Factor (LOF) for evaluation and that the proposition is validated. Gogoi et al. [28] explored outlier detection methods used to identify network anomalies. Feature bagging [15] is a method proposed

to overcome curse of dimensionality and leverage precision in clustering. Koufakou and Georgiopoulos [16] proposed a fast outlier detection strategy in distributed datasets with multiple dimensions. Gao et al. [13] proposed a framework known as LEAP for outlier detection. It is based on reverse nearest neighbour concept besides two optimization principles such as minimal probing and lifespan-aware prioritization. In future they intend to know its scalability. The ensemble method in [22] is capable of picking useful detectors an eliminate underperforming ones in an iterative process to obtain impressive results in multi-dimensional data.

Cassisi et al. [17] proposed an outlier detection method based on space stratification by enhancing density-based clustering that includes dimensionality reduction. For the detector combinations, data locality is not considered often. Instead of data locality, most of the existing ensemble methods such as the one in [23] used all training samples. Unbiased distance-based method [29], angle-based time approximation method [30], outlier detection method based on direct density ratio estimation [31], SVDD based outlier detection [32] and hybrid outlier detection framework [33] are other important approaches found in the literature. From the review of literature, it is understood that the existing ensemble methods made up of multiple unsupervised learning algorithms lack in ideal strategy for choosing right candidates as constituent detectors. It resulted in mediocrity in model stability and accuracy. This problem is overcome in this paper.

3. ENSEMBLE FRAMEWORK BASED OUTLIER DETECTION SYSTEM

A framework is proposed for effective ensemble of outlier detection candidates. It takes high-dimensional dataset as input and pre-processes it to divide into training data (60%) and testing data (40%). A set of base detectors used with sampling based diversity and also averaging based final outlier detectors. For each detector, training data is given as input in order to compute training outlier score. This will result in an outlier score matrix that is used to generate global ground truth by averaging the scores. This global ground truth is used while generating local ground truth later on. For each instance of the test data, kNN ensemble is used to define a local region. Based on the local region, local ground truth is generated. While generating the local ground truth, the global ground truth is used for better results. Afterwards, for each candidate detector outlier score is computed. Then Pearson correlation is computed based on the generated local ground truth and also the computed local outlier score. For a group of subset of detectors, averaging method is used to return competent detectors that finally produce outlier detection results. The proposed ensemble framework based outlier detection system is as shown in Figure 1.

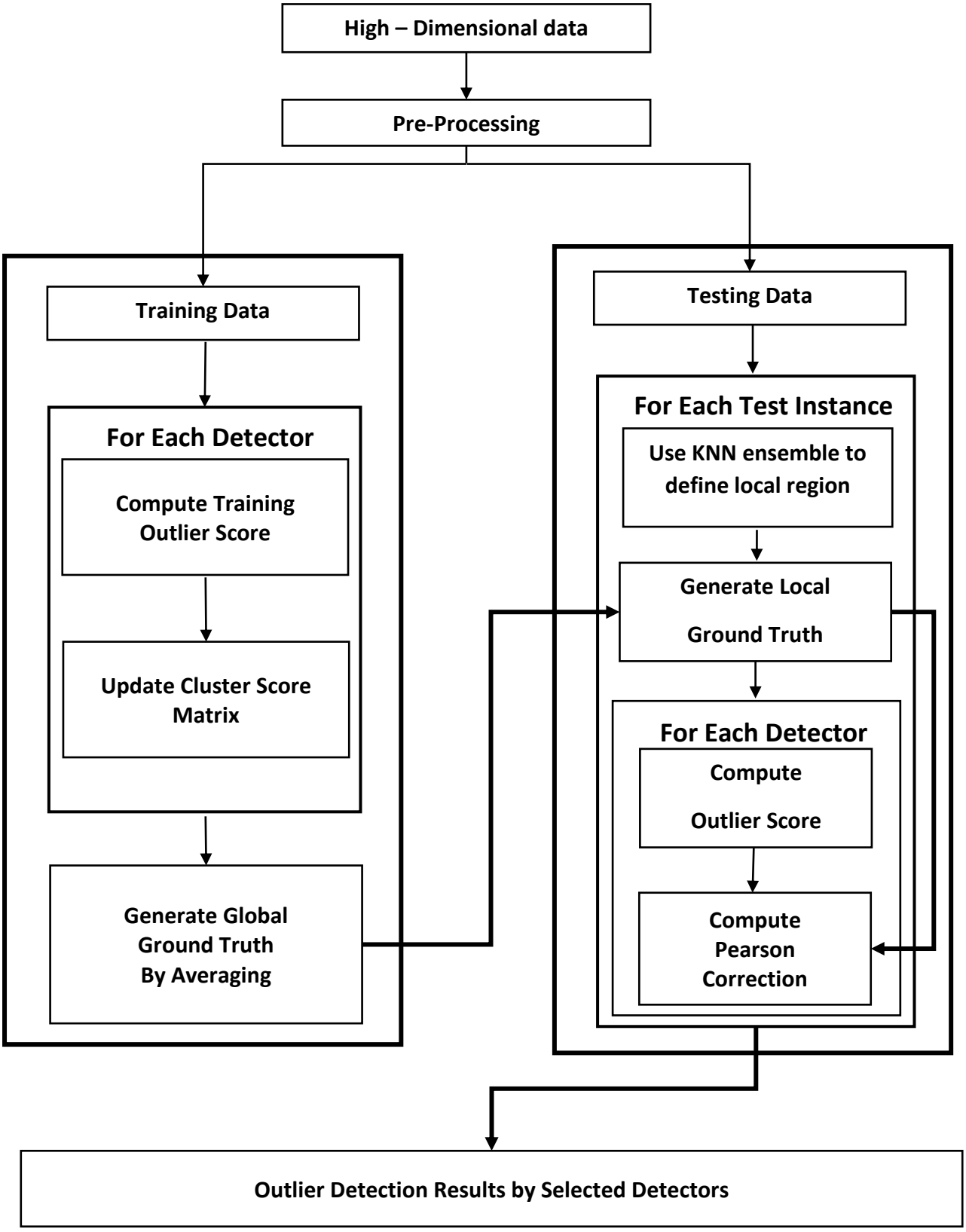


Figure 1: Ensemble framework based outlier detection system

The framework has required procedures to be followed on both training data and testing data in order to have better outlier detection based on ensemble approach that exploits ideal section of candidate detectors besides using a strategy for local region definition. The following sub sections provide details of algorithm design and pseudocode of the proposed algorithm.

3.1 Algorithm Design

Candidate detector diversity is maintained either with distinct hyper parameters of same detector or multiple heterogeneous candidate detectors. A set of LOF detectors are used with distinct MinPts as candidate detectors. All the candidate detectors are trained with the training set denoted as X_{train}. It results in an inference gained in the form of computed outlier score. The score vector is denoted as $C_r(.)$ where r is the index of candidate detector. The final results (computed outlier scores of all candidate detectors) are finally maintained in an outlier score matrix as in Eq. 1.

$$O(X_{train}) = [C_1(X_{train}), \dots C_r(X_{train})] \in R^{n \times R} \quad (1)$$

Where C denotes candidate detector and $R^{n \times R}$ denotes training set while $O(X_{train})$ denotes outlier score matrix. As the proposed framework finds detector competency (without actually ground truth labels as done in supervised learning approaches). The generation of global ground truth with aggregation parameter $\emptyset$ is shown in Eq. 2.

$$global\_ground\_truth = \emptyset\big(O(X_{train})\big) \in R^{n \times 1} \quad (2)$$

It is an important observation here is that the global ground truth is generated purely based on data and it is used for ideal detector selection process. After generating ground truth with aggregation (averaging is used), the system is ready to work on the test data. One way to find local region is to use KNN [34,35,36].Each test instance is used to find its nearest neighbours (k nearest objects in the training set). This is achieved with KNN ensemble method as in Eq. 3.

$$\varphi_j = \left\{ x_i \mid x_i \in X_{train},\ x_i \in KNN_{ens}^{(j)} \right\} \quad (3)$$

Where φ_j represents local region, $KNN_{ens}^{(j)}$ denotes the nearest neighbours of test instance based on ensemble criteria. For each instance, after finding local region, local ground truth is generated as in Eq. 4.

$$global\_ground\_truth^{\varphi_j} = global\_ground\_truth_{\,x_i} \big| \, x_i \in \varphi_j \} \in R^{|\varphi_j| \times 1} \quad (4)$$

Where $|\varphi_j|$ represents cardinality of φ_j and the local training outlier scores can be obtained from the outlier score matrix. Therefore, it results in Eq. 5.

$$O(\varphi_j) = \left[C_1(\varphi_j), \dots C_r(\varphi_j) \right] \in R^{|\varphi_j| \times R} \quad (5)$$

In order to estimate competency of detectors in actual outlier detection on test data, there is an iterative process needed Pearson correlation is used to identify and select outliers that are used in averaging approach to produce final outlier detection results.

3.2 Average Selection and Ensemble of Candidates for Outlier Detection

An algorithm named Average Selection and Ensemble of Candidates for Outlier Detection (ASEC-OD) is proposed and implemented. This algorithm is based on the formulations made in Section 3.1. It takes a set of candidate detectors as input along with training data, testing data and number of neighbours to be considered in finding local regions. After intended processing, the algorithm returns the best set of candidate detectors that produce results of outlier detection.

Algorithm: Average Selection and Ensemble of Candidates for Outlier Detection
Inputs: Candidate detector set C, training data T, test data T2, number of neighbours n
Output: Outlier scores of T2

1. Start
2. Initialize outlier score matrix M
3. Initialize final outlier score vector F
4. For each candidate detector c in C
5. outlier_score = computeScore(T)
6. M = updateMatrix(M)
7. End For
8. global_ground_truth = generateGlobalGroundTruth(M)
9. For each test instance t2 in T2
10. local_region = computeLocalRegion(t2, n, T)
11. local_ground_truth = generateLocalGroundTruth(local_region, global_ground_truth)
12. For each candidate detector c in C
13. outlier_score = computeScore(local_region, T)
14. pearson_correlation = computeCorrelation(outlier_score, local_ground_truth)
15. End For
16. C = getDetectors(pearson_correlation)
17. outlier_score = getFinalOutlierScore(C, t2)
18. Add outlier_score to F
19. End For
20. Return F
21. End

Algorithm 1: Average selection and ensemble of candidates for outlier detection

As presented in Algorithm 1, Step 2 and Step 3 initialize outlier score matrix M and final outlier score vector F respectively. They are used in the later stages to get updated with actual scores. Step 4 through Step 7 is an iterative process used to train candidate detectors and update the matrix M with outlier score produced by the detector. In Step 8, global ground truth is computed based on M. Step 9 starts an iterative process that includes an inner iterative process from Step 12 through Step 15. The outer iterative process ends at Step 19. In Step 10 local region is computed based on number of nearest neighbours, test instance t2 and training set T. In Step 11, local ground truth is computed based on local region and global ground truth. In Step 13, outlier score is computed for each detector based on local region and training set T. Afterwards, Pearson correlation is computed as in Step 14 based on outlier score and local ground truth. In Step 16, all detectors that are competent are identified. They are used to find final outlier score for given instance with averaging method. Then that outlier score is stored in vector F. Finally, in Step 20, the final outlier score vector F is returned which has outlier scores of all instances found in T2.

4. RESULTS AND DISCUSSION

Experiments are made with a prototype application built using Python data science platform. Different outlier datasets collected from [37]. The datasets used for empirical study are known as Cardio, Letter, Arrhythmia, Mnist, Satellite, Thyroid and Stamps. The dataset details are shown in Table 1 in terms of name of dataset, number of points, number of dimensions, number of outliers and % of outliers.

Dataset	Pts	Dim	Outliers	%Outlier
Arrhythmia	452	274	66	14.60
Cardio	1831	21	176	9.61
Letter	1600	32	100	6.25
MNIST	7603	100	700	9.21
Satellite	6435	36	2036	31.64
Stamps	340	9	31	9.12
Thyroid	3772	6	93	2.47

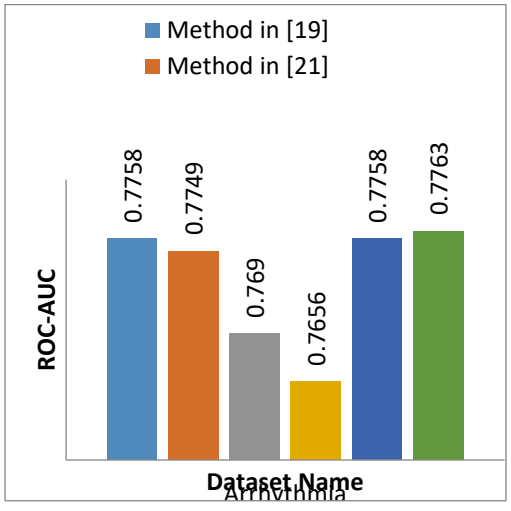
Table 1: Details of outlier datasets used for experiments

The proposed framework and its underlying algorithm is evaluated by comparing results with many existing methods found in [19], [21], [15], [22] and [23]. Two performance metrics such as mean average precision (mAP) and area under the receiver operating characteristic (ROC-AUC). The results of performance are the average of 30 independent experiments.

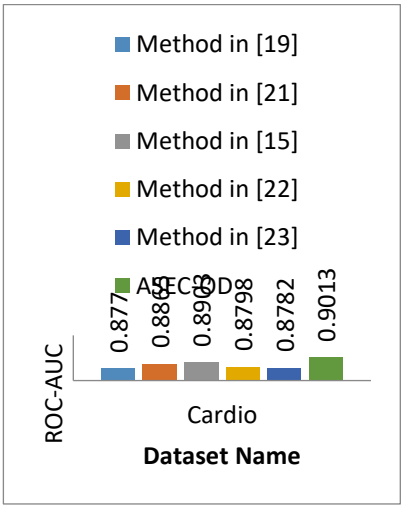
Dataset	ROC-AUC					
	Method in [19]	Method in [21]	Method in [15]	Method in [22]	Method in [23]	ASEC-OD
Arrhythmia	0.7758	0.7749	0.7690	0.7656	0.7758	0.7763
Cardio	0.8770	0.8865	0.8903	0.8798	0.8782	0.9013
Letter	0.7925	0.8031	0.8300	0.8434	0.7908	0.7867
MNIST	0.8557	0.8588	0.8553	0.8349	0.8563	0.8633
Satellite	0.5881	0.5992	0.6220	0.6258	0.5876	0.6015
Stamps	0.8946	0.8927	0.8763	0.8559	0.8953	0.8985
Thyroid	0.9656	0.9647	0.9510	0.9385	0.9665	0.9700

Table 2: Performance evaluation with ROC-AUC

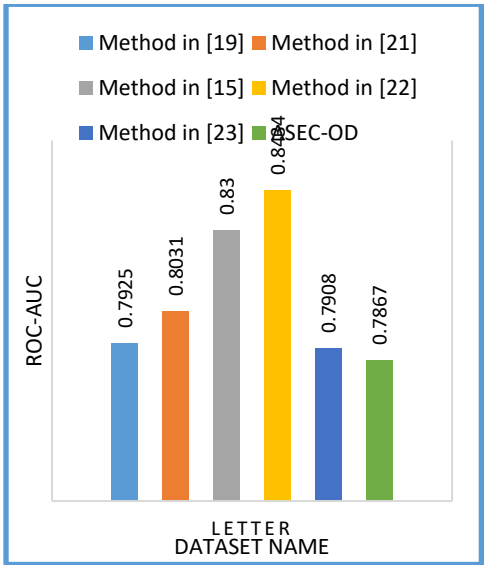
As presented in Table 2, ROC-AUC performance of the ASEC-OD is compared with state of the art methods for all the seven high dimensional datasets.



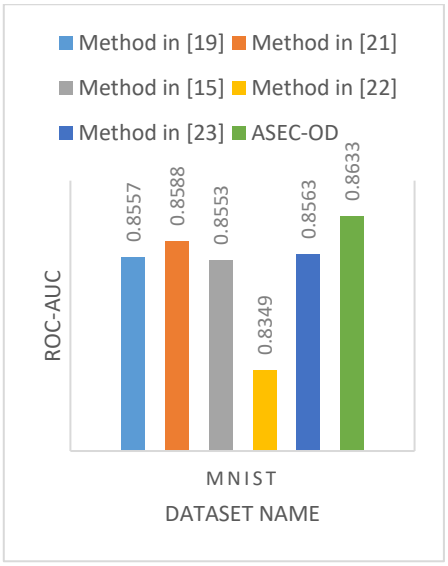
(a)



(b)



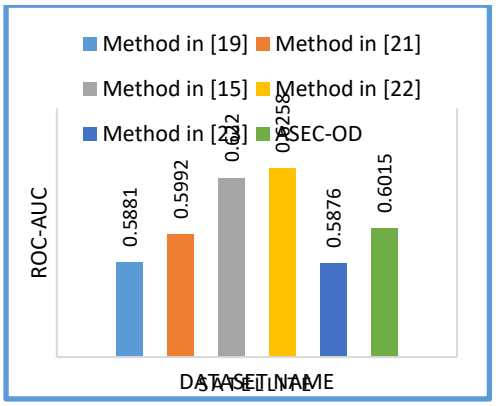
(c)

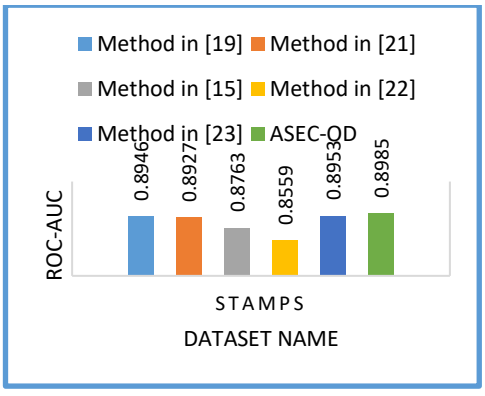


(d)

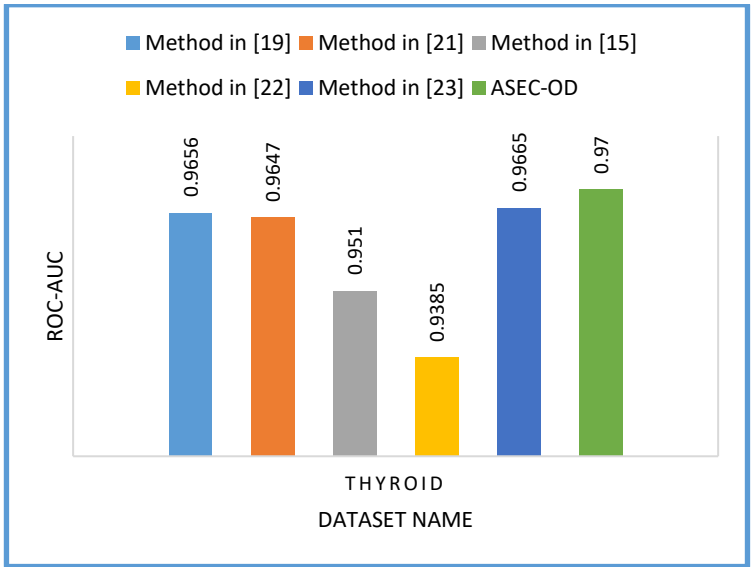
Figure 2: Performance comparison with ROC-AUC for datasets a) Arrhythmia b) Cardio c) Letter d) Mnist

As presented in Figure 2, ROC-AUC performance is compared for four datasets such as Arrhythmia, Cardio, Letter and Mnist. The dataset name is taken in horizontal axis and vertical axis shows value of ROC-AUC. Ground truth generation methods are used in order to gain advantage from variance and also reduce bias. However, the ground truth depends on the quality of data available. In case of Arrhythmia dataset, the ROC-AUC of the proposed method is 0.7763 which is higher than all the existing methods. The least performance for this dataset is exhibited by the method in [22] with 0.7656. With respect to Cario dataset also, the ASEC-OD outperformed all other methods with ROC-AUC value 0.9013. In this case, the least performance is shown by the method in [19] with ROC-AUC value 0.877. With regard to Letter dataset, highest performance is shown by the method in [22] with 0.8434 and the least performance is exhibited by ASEC-OD. This is due to the irrelevant features found in the dataset as the proposed method depends on quality of instance for generating ground truth. With Mnist dataset the performance of ASEC-OD is higher than all other methods with 0.8633. The least performing method in terms of ROC-AUC is the method in [22] with 0.8349.





(a) (b)



(c)

Figure 3: Performance comparison with ROC-AUC for datasets a) Satellite b) Stamps c) Thyroid

As presented in Figure 3, ROC-AUC performance is compared for three datasets such as Satellite, Stamps and Thyroid. The dataset name is taken in horizontal axis and vertical axis shows value of ROC-AUC. Ground truth generation methods are used in order to gain advantage from variance and also reduce bias. However, the ground truth depends on the quality of data available. This is the reason for performance differences for the proposed method. In case of Satellite dataset, the ROC-AUC of the proposed method is 0.6015 which is higher than many existing methods in [19], [21] and [23]. The least performance for this dataset is exhibited by the method in [23] with 0.5876. With respect to Stamps dataset also, the ASEC-OD outperformed all other methods with ROC-AUC value 0.8985. In this case, the least performance is shown by the method in [22] with ROC-AUC value 0.8559. With regard to Thyroid dataset, highest performance is shown by the ASEC-OD with 0.97 and the least performance is exhibited by the method in [22]. From the results of Figure 2 and Figure 3, it is observed that the proposed method outperformed all the existing methods for most of the datasets.

Dataset	**mAP**					
	Method in [19]	**Method in [21]**	**Method in [15]**	**Method in [22]**	**Method in [23]**	**ASEC-OD**
Arrhythmia	0.3766	0.3769	0.3722	0.3690	0.3766	0.3796
Cardio	0.3516	0.3708	0.3864	0.3666	0.3535	0.4117
Letter	0.2388	0.2473	0.2867	0.3160	0.2372	0.2407
MNIST	0.3911	0.3941	0.3896	0.3701	0.3918	0.3979
Satellite	0.4047	0.4139	0.4352	0.4385	0.4047	0.4196
Stamps	0.3694	0.3660	0.3387	0.3144	0.3706	0.3779

Thyroid	0.4045	0.4123	0.3488	0.2850	0.4130	0.4651

Table 3: Performance evaluation with mean average precision

As presented in Table 2, performance of the ASEC-OD is compared with state of the art methods for all the seven high dimensional datasets in terms of mean average precision.

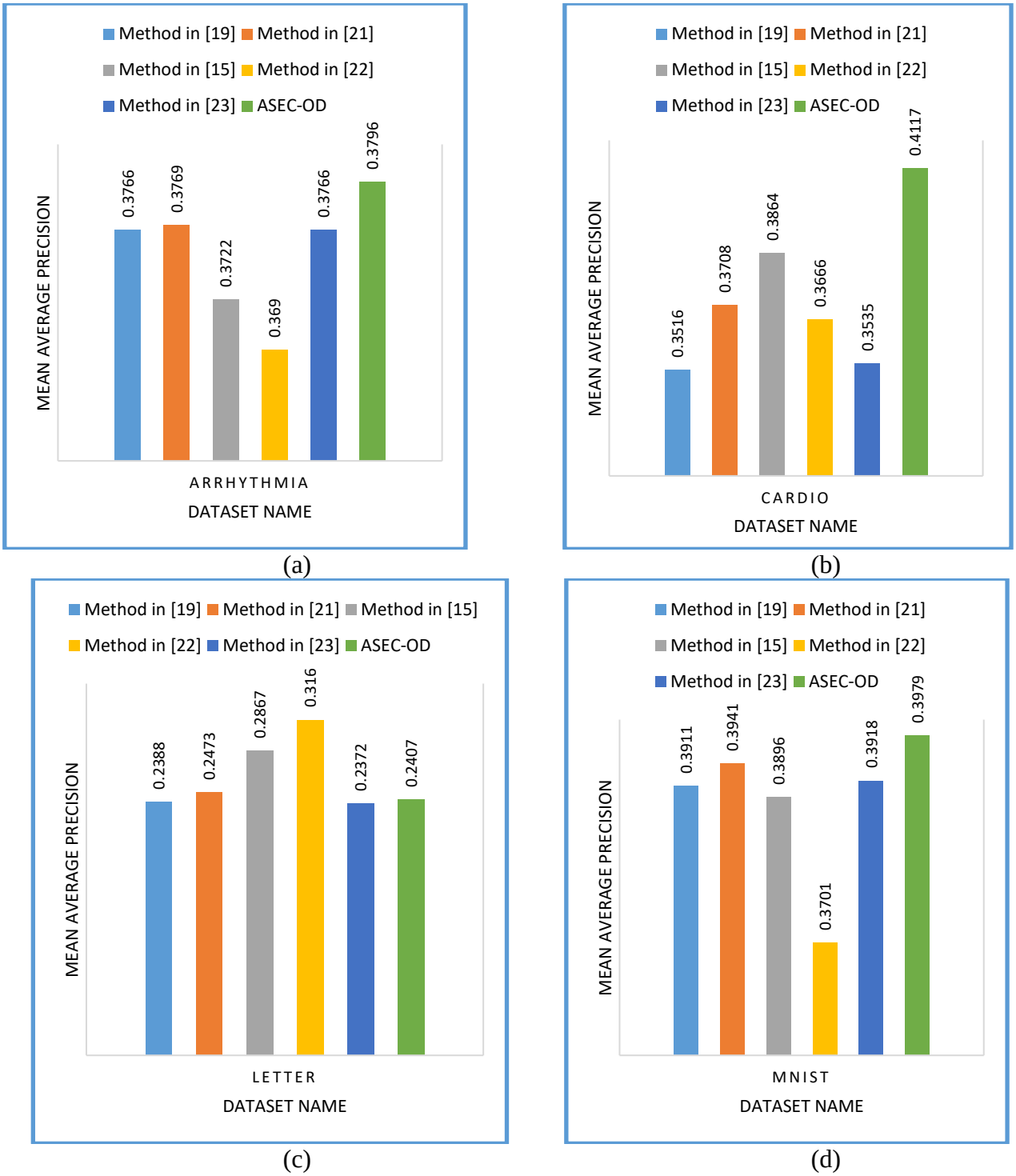


Figure 4: Performance comparison with mean average precision for datasets a) Arrhythmia b) Cardio c) Letter d) Mnist

As presented in Figure 4, performance is compared for four datasets such as Arrhythmia, Cardio, Letter and Mnist in terms of mean average precision. The dataset name is taken in horizontal axis and vertical

axis shows value of mean average precision. Ground truth generation methods are used in order to gain advantage from variance and also reduce bias. However, the ground truth depends on the quality of data available. This has led to performance differences for each dataset with respect to ASEC-OD. In case of Arrhythmia dataset, the mean average precision of the proposed method is 0.3796 which is higher than all the existing methods. The least performance for this dataset is exhibited by the method in [22] with 0.369. With respect to Cardio dataset also, the ASEC-OD outperformed all other methods with mean average precision value 0.4117. In this case, the least performance is shown by the method in [19] with mean average precision value 0.3516. With regard to Letter dataset, highest performance is shown by the method in [22] with 0.316 and the least performance is exhibited by the method in [23]. The ASEC-OD showed better performance over the methods in [19] and [23] only. This is due to the irrelevant features found in the dataset as the proposed method depends on quality of instance for generating ground truth. With Mnist dataset the performance of ASEC-OD is higher than all other methods with 0.3979. The least performing method in terms of mean average precision is the method in [22] with 0.3701.

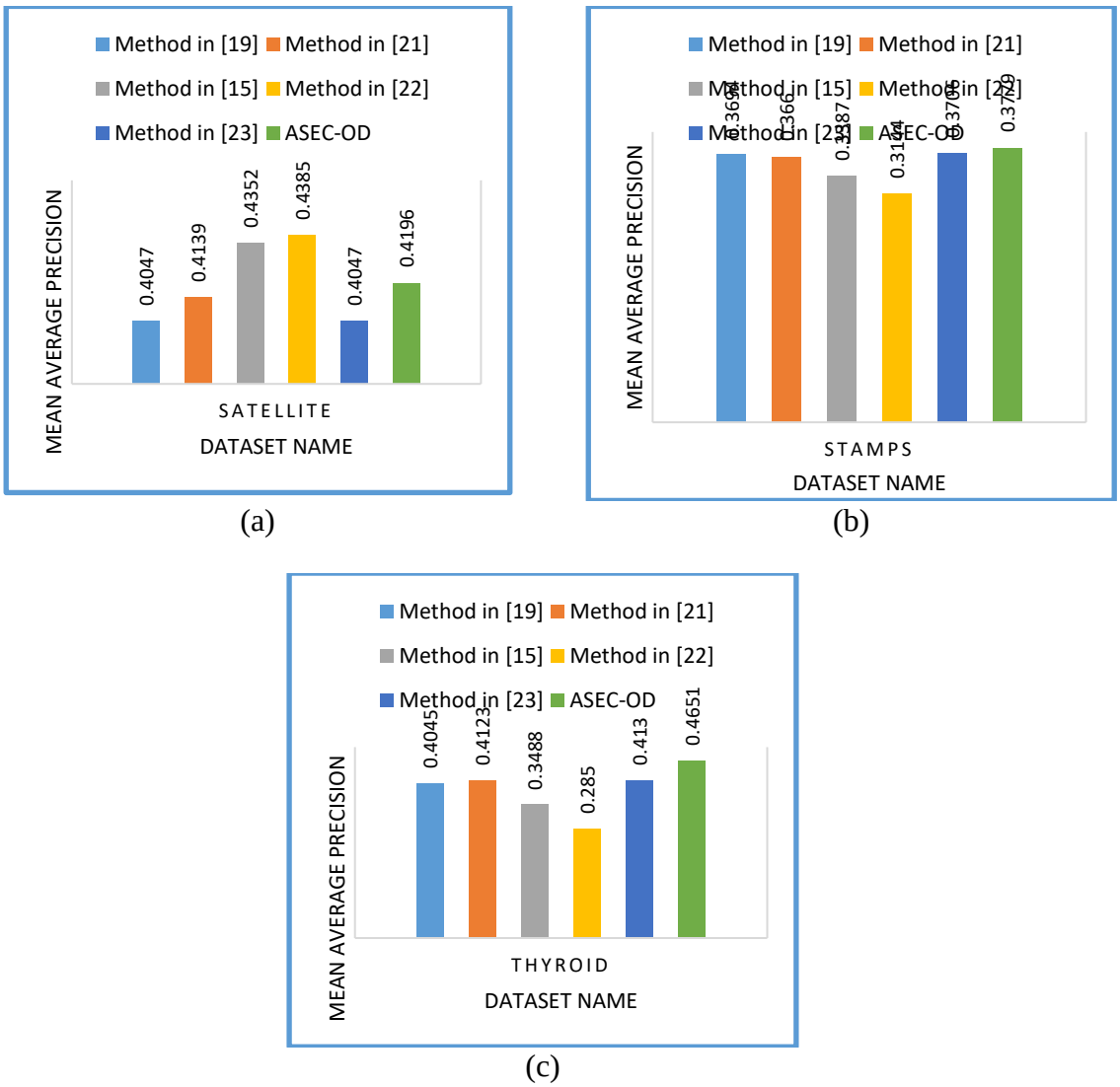


Figure 5: Performance comparison with mean average precision for datasets a) Satellite b) Stamps c) Thyroid

As presented in Figure 5, mean average precision performance is compared for three datasets such as Satellite, Stamps and Thyroid. The dataset name is taken in horizontal axis and vertical axis shows value

of mean average precision. Ground truth generation methods are used in order to gain advantage from variance and also reduce bias. However, the ground truth depends on the quality of data available. This is the reason for performance differences for the proposed method. In case of Satellite dataset, the

mean average precision of the proposed method is 0.4196 which is higher than many existing methods in [19], [21] and [23]. The least performance for this dataset is exhibited by the methods in [19] and [23] with 0.4047. With respect to Stamps dataset also, the ASEC-OD outperformed all other methods with mean average precision value 0.3779. In this case, the least performance is shown by the method in [22] with mean average precision value 0.3144. With regard to Thyroid dataset, highest performance is shown by the ASEC-OD with 0.4651 and the least performance is exhibited by the method in [22] with 0.285. From the results of Figure 4 and Figure 5, it is observed that the proposed method outperformed all the existing methods for most of the datasets.

5. CONCLUSION AND FUTURE WORK

In this work, we proposed an ensemble based outlier detection framework that exploits unsupervised learning methods. The existing ensemble approaches suffer from mediocrity in selecting ideal base selectors that are part of ensemble. This problem is overcome in this paper with an ideal selection of constituent candidate outlier detectors. An algorithm named Average Selection and Ensemble of Candidates for Outlier Detection (ASEC-OD) is proposed and implemented to realize the ensemble framework. Many real world datasets are used for empirical study. The results of experiments revealed that the proposed framework outperforms many existing methods. However, the proposed algorithm has certain limitations. First, defining local region is based on fining nearest neighbours. It may result in deteriorated performance when irrelevant features are found in the dataset. Second, the ground truth generation is made based on averaging. However, it could be improved with combination of heterogeneous base detectors and supervised learning. In our future research we improve the framework to overcome aforementioned limitations.

6. References

[1] Kotkar , V.A & Sucharita, V.2019," Praticle filtering based optical flow computing model for crowd anomaly detection using Gaussian mixture model", International Journal of Recent Technology and Engineering vol7, o 6, pp, 583-591.

[2] Ketan Anand, Jay Kumar, Kunal Anand, "Anomaly detection in online social network: A survey", 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)

[3] Prasad, Y.A.S. & Krishna, G.R. 2019, "Filter based hybrid decision tree construction model for high dimensional anomaly classification", International Journal of Recent Technology and Engineering vol. 7, no. 6, pp, 1200-1207.

[4] Y. Vijay Bhaskar Reddy, Dr. L.S.S. Reddy, Dr. S.Sai Satya Naryana Reddy, "Comparative Study of Density-Based Clustering Algorithms", 2017, Vol 8, International Journal of Civil Engineering and Technology (IJCIET).

[5] Babu, B.S., Prasanna, P.L. & Vidyullatha P. 2018, "Customer data clustering using density based algorithm", International Journal of Engineering and Technology(UAE), Vol 7, no 2, pp, 35-38.

[6] Vijay A. Kotkar and V Sucharita, " A comparative Analysis of Machine Learning Based Anomaly Detection Techniques in Video Surveillance", 2017 Vol 12, Journal of Engineering and Applied Sciences

[7] Lavanya, K., Reddy, L.S.S. & Eswara Reddy, B. 2019, A Study of High-Dimensional Data Imputation Using Additive LASSO Regression Model.

[8] N Jayanthi, B Vijay Babu, N Sambasiva Rao," Advanced Techniques for Outlier Detection in High Dimensional Data", National conference on Recent Advancements in computer science CONRACS 2019. pg 169.

[9] Charu C. Aggarwal. (2013). HIGH-DIMENSIONAL OUTLIER DETECTION: THE SUBSPACE METHOD. *Springer*, p1-76.

[10] Holger Trittenbach ,Klemens B¨ohm. (2019). Dimension-based Subspace Search for Outlier Detection. *Springer*, p1-16.
[11] Liangwei Zhang, Jing Lin,Ramin Karim. (2016). Sliding Window-Based Fault Detection From High-Dimensional Data Streams. *IEEE*, p1-15.
[12] Milos Radovanovi c, Alexandros Nanopoulos, and Mirjana Ivanovi c. (2014). Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection. *IEEE*, p1-14.
[13] Lei Cao , Di Yangt , Qingyang Wang, Yanwei Yu , Jiayuan Wang, Elke A. Rundensteiner. (2014). Scalable Distance-Based Outlier Detection over High-Volume Data Streams . *IEEE*, p1-12.
[14] Xuan Hong Dang, Ira Assent,Raymond T. Ng,Arthur Zimek, Erich Schubert. (2014). Discriminative Features for Identifying and Interpreting Outliers. *IEEE*, p1-12.
[15] A. Lazarevic and V. Kumar, Feature bagging for outlier detection, ACM SIGKDD, (2005), p. 157.
[16] Anna Koufakou ,· Michael Georgiopoulos. (2010). A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes . *Springer*, p1-32.
[17] Carmelo Cassisi, Alfredo Ferro n , Rosalba Giugno, Giuseppe Pigola, Alfredo Pulvirenti. (2013). Enhancing density-based clustering: Parameter reduction and outlier detection. *Elsevier*, p317-330.
[18] Debasrita Chakrabortya , Vaasudev Narayananb , Ashish Ghosha. (2019). Integration of deep feature extraction and ensemble learning for outlier detection. *Elsevier*, p161-171.
[19] Charu C. Aggarwal. (2013). Outlier Ensembles. *ACM*. 14 (2), p1-10.
[20] Xuyun Zhang,Wanchun Dou,Qiang He,Rui Zhou,Christopher Leckie,Ramamohanarao Kotagiri,Zoran Salcic. (2017). LSHiForest: A Generic Framework for Fast Tree Isolation based Ensemble Anomaly Analysis. *IEEE*, p1-13.
[21] C. C. Aggarwal and S. Sathe, Theoretical Foundations and Algorithms for Outlier Ensembles, ACM SIGKDD Explorations, 17 (2015), pp. 24-47.
[22] S. Rayana, W. Zhong, and L. Akoglu, Sequential ensemble learning for outlier detection: A bias-variance perspective, ICDM, (2017), pp. 1167-1172.
[23] A. Zimek, R. J. G. B. Campello, and J. o. r. Sander, Ensembles for unsupervised outlier detection: Challenges and research questions, ACM SIGKDD Explorations, 15 (2014), pp. 11-22.
[24] [24] Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang and Xiangnan He. (2019). Generative Adversarial Active Learning for Unsupervised Outlier Detection. *IEEE*, p1-13.
[25] Shuyan Chen , Wei Wang , Henk van Zuylen . (2010). A comparison of outlier detection algorithms for ITS data. *Elsevier*, p1169-1178.
[26] Manish Gupta, Jing Gao,Charu C. Aggarwal,Jiawei Han. (2014). Outlier Detection for Temporal Data: A Survey. *IEEE*. 26 (9), p1-18.
[27] R´emi Dominguesa , Maurizio Filipponea , Pietro Michiardia , Jihane Zouaouib. (2018). A comparative evaluation of outlier detection algorithms: experiments and analyses. *Elsevier*, p1-26.
[28] Prasanta Gogoi1, D K Bhattacharyya, B Borah and Jugal K Kalita. (2011). A Survey of Outlier Detection Methods in Network Anomaly Identification. *IEEE*, p1-19.
[29] Hoang Vu Nguyen1 , Vivekanand Gopalkrishnan , and Ira Assent. (2011). An Unbiased Distance-based Outlier Detection Approach for High-dimensional Data. *Springer*, p1-15.
[30] Ninh Pham,Rasmus Pagh. (2012). A Near-linear Time Approximation Algorithm for Angle-based Outlier Detection in High-dimensional Data. *ACM*, p1-9.
[31] Shohei Hido,Yuta Tsuboi,Hisashi Kashima,Masashi Sugiyama,Takafumi Kanamori. (2011). Statistical Outlier Detection Using Direct Density Ratio Estimation. *Springer*. 26 (2), p1-35.
[32] Bo Liu, Yanshan Xiao , Longbing Cao, Zhifeng Hao , Feiqi Deng. (2013). SVDD-based outlier detection on uncertain data. *Springer*, p1-23.
[33] N Jayanthi, Dr Burra Vijaya Babu, Dr N Sambasiva Rao, "Hybrid Machine Learning based Framework for Outlier Detection in High Dimensional Data", 2020 .Vol 7, Journal of Critical Reviews.
[34] Ghuge, c A., Ruikar, s.D.& Prakash, V C. 2018, "Support vector regression and extended nearest neighbour for video object retrieval", Evolutionary Intelligence.

[35] Kodali, s., Dabbiru, M., Thirumala Rao, B, & Kartheek Chandra Patnaik, U, 2018, A k-NN-based approach using MapReduce for meta-path classification in heterogeneous information networks.
[36] Nallamala, S,H., Mishra, P, & Koneru, S,V, 2019, " Pedagogy and reduction of K-nn algorithm for filtering samples in the breast cancer treatment", International Journal of Scientific and Technology Research Vol. 8, no.11, pp.2168-2173.
[37] Outlier Detection Datasets (2020). Retrieved from http://odds.cs.stonybrook.edu/