Analyzing the Performance of Anomaly Detection Algorithms

Chiranjit Das[1], Akhtar Rasool[2], Aditya Dubey[3], Nilay Khare[4]
Department of Computer Science Engineering
Maulana Azad National Institute of Technology
Bhopal, India

***Abstract*—An outlier is a data observation that is considerably irregular from the rest of the dataset. The outlier present in the dataset may cause the integrity of the dataset. Implementing machine learning techniques in various real-world applications and applying those techniques to the healthcare-related dataset will completely change the particular field's present scenario. These applications can highlight the physiological data having anomalous behavior, which can ultimately lead to a fast and necessary response and help to gather more critical knowledge about the particular area. However, a broad amount of study is available about the performance of anomaly detection techniques applied to popular public datasets. But then again, have a minimal amount of analytical work on various supervised and unsupervised methods considering any physiological datasets. The breast cancer dataset is both a universal and numeric dataset. This paper utilized and analyzed four machine learning techniques and their capacity to distinguish anomalies in the breast cancer dataset.**

Keywords*—Anomaly; machine learning; outlier detection; minimum covariance determinant

I. Introduction

Anomalies can detect a fault in a system or a network, abnormalities in healthcare data, or fraud [1]. The fast growth in the amount, dimension, and complexity of data in the dataset has made it compulsory to automate the outlier detection in analytical processes. Those analytical results can then be used in decision-making processes by various algorithms to identify the health condition [2]. Description of these anomalous values can provide a better and new understanding of the dataset. Using different examples can visualize the whole phenomenon in a better way for the researchers. Particularly in healthcare, this could be used to develop good knowledge about the patient's health condition and complications, whereas in the case of predicted value, it does not focus only on the data description but also the future assumption about numerous states of a patient's health condition. That could help healthcare employees to treat the patient in an improved way by detecting the disease correctly. These outliers must not be considered an error or noise. Still, these are also important with respect to train the existing detection models and prepare to predict the new data added to the dataset.

Undoubtedly, the anomaly detection procedure depends on an effective definition of data point that must be considered an anomaly [3]. Anomaly detection approaches usually undertake two things, namely, (1) anomalies are rare and minor in amount, and (2) anomalies are dissimilar in some sense or another from other normal data. The presence of multiple kinds of outliers or anomalies further results in complicating the operational definition. According to the occurrence, data points can be subdivided into anomalous concerning neighboring data points (local anomalies) or the entire dataset (global anomalies). It is challenging to differentiate between classes from which the anomaly belongs. The dissimilarity is still convenient for reminding the data scientists that anomalies can differ from both the data point and each other. A vast variety of technology is present in the healthcare system. However, the application of both machine learning and physiological data-set still needs some attention and in its early stages. Adding to the already problematic and challenging work of health care specialists is the scarcity of employees in the public healthcare field. To enhance the massive burden of healthcare specialists, it requires analyzing the minor units of physiological data collected, and it also helps improve patient care facility. Faulty events in the healthcare sector and minor medical errors may increase the chances of accidental deaths. Anomaly detection is a vital task, especially in the healthcare sector, where errors are rare. This paper has tried to better understand a patient's safety by predicting anomalous behaviors in breast cancer data through different anomaly detection techniques and errors in predicting the disease. Whether anomaly detection techniques are instigated to identify that patient is sick, look after the health conditions through sensors that can be wearable, or support another patient care-related work, these processes are vital components of the automated patient care system.

Nevertheless, what is considered by the system as an anomaly is often challenging to define? In most cases, choosing the proper machine learning techniques to implement depends on both investigation and the variability of the outliers present in the dataset. For example, there exist numerous works where the neural network has been used in healthcare-related problems. At the same time, some researchers have utilized clustering approaches and the multi-layer perceptron [4], [5]. Undoubtedly, choosing the algorithm depends partly on both the type of dataset and the dataset's size. Classifying an ideal machine learning model [6] for a particular problem relies on either or not the ground truth labels for data exist. In labeled data, supervised methods are suitable, while unsupervised methods allow analyzing the unlabelled data.

Nevertheless, it must be momentarily stated that semi-supervised anomaly detection approaches provide yet another vital option comprised of datasets labeled and comparably large amounts of unlabelled datasets. While countless models can be effectively adjusted to process the anomaly detection problems, the present study focuses on the subsequent four approaches: Local Outlier Factor (LOF), Isolation Forests (IF), Minimum Covariance Determinant (MCD), and One-Class SVM. The reason to choose these approaches are (1) their extensive occurrence in the area of machine learning and anomaly detection works fundamentally, and (2) because they comprise both supervised and unsupervised techniques. In the following subdivided sections, the techniques are briefly described.

II. Related Works

A. Local Outlier Factor (LOF)

The LOF technique was first presented by Breunig et al. as per the name recommends [18]. The technique primarily evaluates the separation between each of the data points concerning its neighborhood. This technique partially depends on the k-nearest neighborhood of the data point. Though other approaches classify outliers in a binary manner, LOF can also generate the degree of outliers of a data point. This local outlier density calculation is summarized with k-nearest neighbors. Data points in comparatively lower density zones are classified as outliers or assign a higher degree of outliers.

B. Isolation Forest (IF)

Isolation forest is an unsupervised type of machine learning technique for outlier detection implemented based on isolating anomalies instead of the most popular techniques of profiling normal points [19]. In this type of detection technique, the outlier is isolated by randomly partitioning them by selecting a random feature at a time. As a result, the outliers are more manageable to isolate than normal data. However, the node with the regular data needs time to isolate and requires many partitions.

C. Minimum Covariance Determinant

The input variables in the Gaussian distribution can deal with normal statistical methods to identify the outliers or anomalies. For example, suppose a dataset has two input variables. Minimum Covariance Determinant (MCD) is a detection technique that deals with those datasets to form the multi-dimensional Gaussian distribution [20]. Information of this distribution can be cast off to detect entries far from the distribution. Thus, the Minimum Covariance Determinant (MCD) technique is a vastly robust detector of multivariate location and scatter.

D. One-Class SVM (OC-SVM)

One-class SVM is a kind of unsupervised learning technique formally based on the classical Support Vector Machine (SVM) [21]. It is an algorithm where the training is only done with regular data or the dataset's negative data. Generally, this algorithm classifies the boundary with these training data subsets and can classify data point that does not belong to that boundary.

Fig. 1 and 2 shows the example features of various anomaly detection techniques on a 2D dataset. For every dataset, 15% of samples are produced as uniform random noise. This ratio is the value assigned to the nu parameter of One-Class SVM and the contamination boundary for other anomaly detection techniques. Considering a manually generated dataset, it tries to visualize the different outlier detection models for better understanding.

Outliers are generally the data points that do not comply with (other data points) general behaviors of the dataset. If the outliers are graphically represented, they fall typically somewhat apart from the regions consisting of normal data points. Outliers typically show variability in the dataset, errors present in measurement, or a novelty.

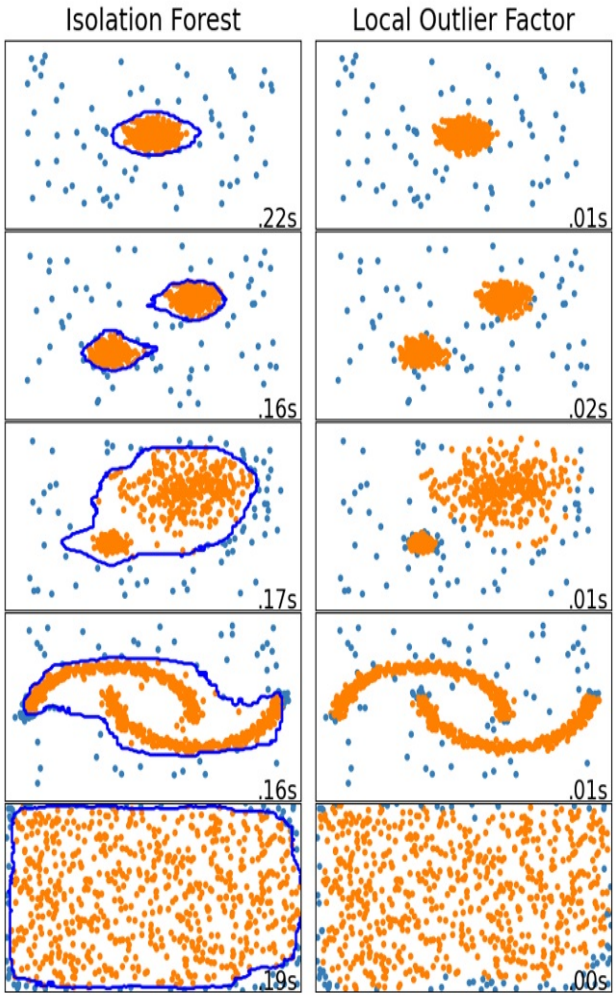

Fig. 1. Scatter Plot Showing the Graphical Representation of Outliers using Isolation Forest and LOF.

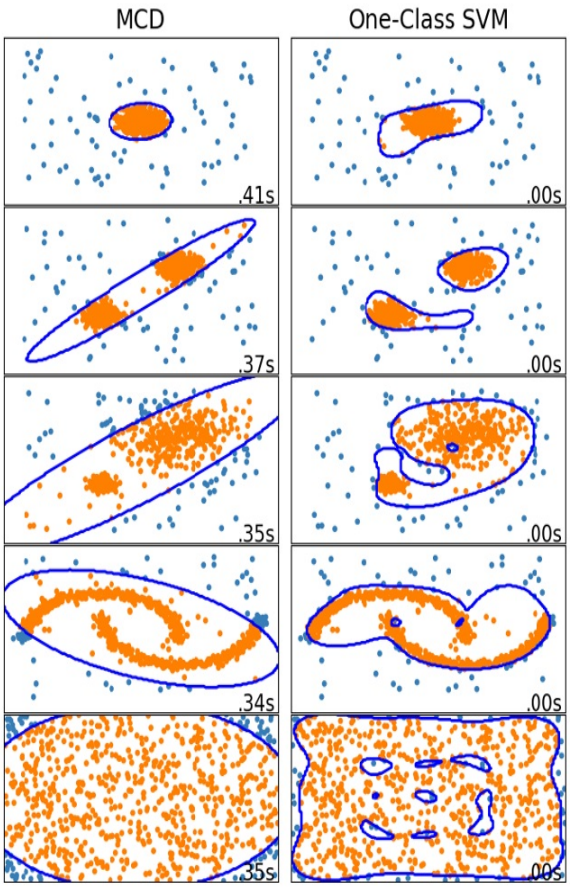

Fig. 2. Scatter Plot Showing the Graphical Representation of Outliers using MCD and One-Class SVM.

Random forest was a detection technique implemented by Leo Breiman [7]. It is an unsupervised learning technique and uses a Gini index classification of the anomalous data point. Many of the trees are generated randomly in this approach, selecting features, and then the most popular one is voted out. Outlier detection was implemented with machine learning techniques by Hadi et al. [8]. It is a supervised technique based on the regression analysis model trained by the training subset and predicts the outliers' test subset. Accordingly, Ramaswamy et al. come with a partition-based supervised technique capable of handling the numeric dataset [9]. It is a proximity-based algorithm. Pawlak et al. use the roughest methodology to derive the outlier from the numeric dataset [10]. It is a supervised detection algorithm based on the model of soft computing. Peterovskiy et al. proposed a fuzzy approach based on kernel function, a supervised learning technique that requires no separate training data [11]. It is capable of handling the numeric type of dataset. The class represents the degree of membership is used to classify the outliers from the normal data point. A trendy outlier detection technique known as Local Outlier Factor (LOF) was introduced by Kriegel et al. [12]. It is a density-based detection technique. It is a supervised learning technique that can process datasets with both numeric and categorical data.

Benjamin Nachman et al. introduced an approach to find outliers using density estimation [13]. A density-based outlier detection technique uses conditional probability density to classify outliers in the dataset containing categorical and numerical data.

Bo Tang et al. introduced an outlier detection algorithm based on the Local Density of data [14]. It is a density-based detection technique that calculates outlier in a dataset containing both numerical and categorical data. Duan et al. proposed (LDBSCAN) a clustering-based model and used Euclidian distance as a proximity measure [15]. It is capable of the handle only numeric datasets. Kriegel has discussed many subspace-based detection algorithms that consider the subspace of the feature set [16]. All these works are generally based on a subset of the different existed detection techniques. Rank Based Detection Algorithm (RBDA) (2011) was introduced by H. Huang et al. [17]. It is a rank-based model that uses the density of the data points to identify the outliers. Our paper will provide an analysis of four popular detection techniques with different analytical measures.

III. Materials and Model Implemented

A. Approaches Involved

Outlier detection is significant and essential in data mining, which can be used to preprocess the dataset, increasing the final result's accuracy. If the database is vast, then it always consists of some abnormal dataset. These data points are known as outliers which are irrelevant from the regular data points. Therefore, these data points must be removed from standard data in data mining. Hence, outlier detection, also known as anomaly detection, is required to classify the outliers to improve the data quality.

1) Local Outlier Factor (LOF): LOF is a type of score that tells either a data point is an outlier or not [18]. Initially, let us start with the basic introduction of a parameter k, determined by calculating the distance between a point, say, p, and its kth nearest neighbors. This detection algorithm works by considering the neighborhood density of a specific data point to its k nearest neighbor. This density of the data points is then considered to determine the relative densities. Now choosing the correct value of k is a bit tricky, for selecting a too-small value of k it will focus more on local data, i.e., concentrated more on neighborhood points giving erroneous result in terms of noisy data, and for a large value of k it will completely miss out the local outliers. Breunig et al. introduced a LOF detection algorithm to understand the algorithm more precisely, which consists of three important terms discussed below:

a) k-Distance: The term k-distance can be defined as a distance is between the data point from its kth neighborhood data point. Let us assume k has a value of 4; k-distance will be the distance from a point to its 4th nearest neighbor data point.

Reachability Distance: To determine the reachability distance, the k-distance has been used. The reachability distance is a measure to calculate the maximum distance between two data points and the k-distance value of the second point. Mathematically,

$$reach\_dist = max\{-distance(p,q), distance(p,q)\} \quad (1)$$

Suppose p is a k-nearest neighbor of point q, then reach_distance (p, q) can be defined as k-distance of q or the distance between these two points.

b) Local Reachability Distance (lrd): This value of reachability distance calculates another key concept called local reachability distance (lrd). To determine the value of lrd first step, calculate all the reach_distance of k-nearest neighborhood points and find the mean of these values; use the inverse of this mean will be the lrd. Here, for densities, the higher the distance between the points lesser will be the density. Local reachability distance can be represented as:

$$lrd(p) = 1/\left(\frac{mean(reach\_distance(p,n))}{k}\right) \quad (2)$$

Parameter lrd(p) represents the distance from a point to its nearest neighbors. Each point's lrd and the neighbor's lrd are then related. Through this, the ratio for each point is determined, and calculates the average. The LOF is determined by calculating the average ratio of lrd of a particular point by its neighbor's lrd. Lastly, the density of that data point is correlated with the density of its nearest neighbors. Less density of a data point than its neighbors indicates that point as an outlier.

2) Isolation forest: A feature is chosen randomly at once in an isolation forest from a subset of the dataset. It then isolates each data point until they are entirely isolated from one another [19]. This algorithm is based on the idea that anomalous data points are easier to isolate than normal data points. The detection algorithm enables building all the possible collections of isolation trees (Itree). Those Itrees are made considering all random subsets of data points and assigning each Itree with an outlier score that sums up an outlier score for each dataset's data points. The most important part of this algorithm is building the Itree. The process that comprises choosing a random subset of the whole dataset and then selecting one random feature at a time separates each point until they are entirely isolated. It is required to conduct a binary search operation to predict a new data point's anomalous behavior for that point. That operation must follow top to bottom order. Then assign the anomaly score to that data by calculating the total path length to reach the data point. Then determine the collective outlier score of the data point from the outlier score of individual Itree. The required mathematical equation to compute the outlier score is given by:

$$S(p,m) = 2^{\frac{-E(h(p))}{c(m)}} \quad (3)$$

Where S (p, m) is the outlier score of the data point p and m is the sample size, h(p) gives the average depth to reach the point p from the Itree. c(m) is the average value of h(p). Now, if outlier score, i.e., when E(h(p)) << c(m) indicates the point to be an anomalous point and if outlier score, when indicates a regular point. The points with a value close to 1 are outliers, and values close to 0.5 are normal.

The whole detection process can be sub-divided into two stages- the training stage and the evaluation stage. In the training stage, the building of Itree is done by recursively separating every data point up to a precise height is reached, which is approximately the average depth of Itree. Then, each test data point's outlier score is set from the expected path distance E(h(p)) in the evaluating stage. Finally, the value of E(h(p)) is determined by passing the test data point through Itree, a part of an isolation forest.

3) Minimum Covariance Determinant (MCD): Robust estimation of the multivariate mean ($\hat{\mu}$) and covariance ($\hat{\Sigma}$) in MCD, searching data subset from the h data points is concluded in MCD [20]. The robust estimate must comply with a minimum determinate value the covariance matrix. where h data point must lie between $\frac{k+p+1}{2}$ to k. The MCD is based on Mahalanobis Distance (MD) [21] and given by:

$$m^2 = (x - \hat{\mu})^T \hat{\Sigma}^{-1} (x - \hat{\mu}) \quad (4)$$

When the determinate covariance is equal to zero, then the inverse of the covariance ($\hat{\Sigma}$) will not exist, and also, the value of m will be undefined in the case of (h < p). So, it must be necessary to have more observation than the number of variables to determine the value of m for the dataset. It is complex and costly to implement exact MCD instead; fast-MCD is very popular in practice. Fast-MCD starts with randomly choosing a subset from observation equal to p+1 from dataset x [22]. The subset x_0 of real dataset x, simultaneously calculating their Mahalanobis Distance m_0 for k number of observations with subsequent mean and variance $\hat{\mu}$ and $\hat{\Sigma}$, respectively. Then the number of h observations are separated from a whole new subset of the dataset x_i having the lowest value of MD m_0. The value of h is determined by-

$$h = \left\lfloor 2\left\lfloor\frac{k+p+1}{2}\right\rfloor - k + 2\alpha\left(n - \left\lfloor\frac{k+p+1}{2}\right\rfloor\right)\right\rfloor \quad (5)$$

The value of the parameter α always lies between 0.5 and 1, representing the desirable robustness. Equating a lower value of α may cause in increasing the robustness but also cost inefficiency and resulting in a potentially large outlier set. Then all x_i and m_i is computed for all k number of observations each time choosing a new subset of h, unless the subset at the present iteration is the same as the previous iteration, the process will repeat. At that point, the local minimum value of determinate of covariance is attained. This algorithm repeats itself up to a maximum number of the subset of x. x_{Mcd} is then defined by a subset of h observation with the lowest value of covariance determinate. The robust estimation is calculated by:

$$\hat{\mu}_{Mcd} = \frac{1}{h}\sum_{i=1}^{h} x_{Mcd\ i} \quad (6)$$

$$\hat{\Sigma}_{Mcd} = c_0 \frac{1}{h}\sum_{i=1}^{h} (x_{Mcd_i} - \hat{\mu}_{Mcd})\,(x_{Mcd_i} - \hat{\mu}_{Mcd})^T \quad (7)$$

To handle correctly comparatively small sample of a scalar consistency factor (c_0) is introduced in the equation of estimation, following the estimate of m_{Mcd} . Then the threshold value of m_{Mcd} is calculated, which indicates the data point as an outlier when it falls beyond that threshold. A popular approach is followed to do so introduced by Dovoedo and Chakraborty [23]. In this approach, the first step is to transform the value of each x_{Mcd} to Robust Mahalanobis

Distance (r_{Mcd})[20] outliers, to limit the distance distribution in the range of zero to one.

$$r_{Mcd} = 1 - \frac{1}{1+m_{Mcd}} \quad (8)$$

Then on the second step, some regular multivariate sample of k observation is simulated with $\hat{\mu}_{Mcd}$ and $\hat{\Sigma}_{Mcd}$. Finally, the outliers of those simulated k observations are determined and $\hat{\epsilon}_{Mcd}$ percentile is being used from these simulations to limit the outliers. The user-defined value $\hat{\epsilon}_{Mcd}$ lies between zero to one. More reliably, the detect outlier is needed to implement a robust estimation to resist a possible outlier set. The robust MCD estimator is more helpful to deal with the multivariate dataset. The distance called robust distance [20] is used to flag anomalies. The robust distance is given by-

$$RD^2 = (x - \hat{\mu}_{Mcd})^T \hat{\Sigma}^{-1}_{Mcd}(x - \hat{\mu}_{Mcd}) \quad (9)$$

It is more reliable to define a precise hypersphere capable of covering all regular data points, and the points out of this distance from the origin are termed outliers.

4) One-Class Support Vector Machine (OCS): A One-Class SVM is a type of unsupervised machine learning technique [24]. The One-Class SVM is only trained with regular points or negative examples. It automatically develops l0earning boundaries of those negative points and can successfully estimate the classification of any data that belongs outside that defined boundaries and identify them as outliers. The training stage of any unsupervised machine learning technique is challenging, and so with the One-Class SVM. Crucial parameter υ in this detection technique controls the amount of outlier or contamination one user expects to identifies as outliers. The gamma is another parameter used to calculate the smoothing of boundary lines.

In One-Class, SVM defines an optimized boundary used to differentiate regular data from the anomalous data points in higher dimensionality by maximizing the difference between regular and anomalous points. For example, in One-Class SVM, a hyperplane [25] is defined as a separate anomalous and regular data point from the origin.

The objective function to generate a hyperplane for One-Class SVM is given by-

$$min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}(\xi_i - \rho) \quad (10)$$

$$\text{Subject to: } (w \cdot \phi(x_i)) \geq \rho - \xi_i, \xi \geq 0 \quad (11)$$

Where w is weight vector, ξ is represented slack variables, ρ is the distance between origin and hyperplane, $\phi(x_i) \rightarrow F$ is feature space mapping for input dataset x. The regularization parameter is represented by ρ, which controls the boundary around regular data and leaves a fraction of data classified as anomalous. The main objective is to attain a hyperplane with less distance from the origin, which calculates the optimized value of w & r so that some miss-classification allows for few data points. Moreover, the hyperplane-based One-Class SVM has a low-performance ability. So, to overcome this, another approach is developed where a hypersphere is constructed aiming to separate regular and anomalous point by a sphere of radius R, center C, and feature space F [26]. the objective function is given by-

$$min_{R,\xi,C} R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \quad (12)$$

$$\text{Subject to: } \|\phi(x_i) - C\|^2 \leq R^2 + \xi_i \quad (13)$$

Here the main issue is to reduce the value of R of hypersphere and to compute this. It has to determine the value R & C so that most data points have a comparative Euclidean distance to radius R so that some miss-classifications are allowed.

B. Model Implemented

To analyze the different outlier detection techniques, make a baseline algorithm performance and compare it with other algorithm's performance. We developed the baseline performance by fitting a linear regression for the given dataset. The evaluation of method performance is then carried out by training a particular algorithm on the training data subset. The Mean Absolute Error, Root Mean Square Error, and model score [26] are determined based on the algorithm's prediction on the test dataset. Then we fit the dataset with each of the detection algorithms after defining them. The previously fitted algorithm predicts the anomalous and non-anomalous data points. The anomalous data points are dropped from the training part of the dataset. Then the remaining data points are fitted on the detection algorithm, and finally, the prediction on the test part of the dataset is analyzed.

This is the analytical model use to analyze breast cancer data. Four different outlier detection techniques are used to detect outliers. The various analytical measures are used to analyze the predicted outcomes. Then a classifier is called a decision tree classifier. The proposed algorithm is given below:

Input: Breast cancer dataset.

Output: Analysis and classification of data.

Step 1: Prepare the breast cancer dataset.

Step 2: Prepare a Base performance using the linear regression technique.

Step 3: Apply different detection methods to identify outliers.

Step 4: Analyze the results with different analytical measures.

Step 5: ROC value and the precision rank of different detection are measured.

Step 6: Out of the above outlier detection, the best one is used to detect the outlier from the breast cancer dataset, and using the confusion matrix as a classifier, the prediction of breast cancer has been made.

IV. Result and Discussion

A. Data Description

The breast cancer dataset is considered is downloaded from the data world repositories. The breast cancer dataset contains information about the women who are likely to have cancer. The dataset consists of a total of 11 attributes, namely sample id, clump thickness (1-10), uniformity of cells (1-10),

marginal adhesion (1-10), single epithelial cell size (1-10), bare nuclei (1-10), bland chromatin (1-10), normal nucleoli (1-10), mitoses (1-10) and finally the class attribute. The dataset consists of 698 instances. The dataset also comprises 10 feature attributes and one class attribute. The original dataset is split into training and test dataset. The training dataset is 70% of the total dataset, i.e., 488 instances, and the test data is 30% of the entire dataset, i.e., 210 cases.

B. Analytical Measures

The different detection algorithms' performance is evaluated by some analytical measures like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Model Score of each detection algorithm [27],[28],[29]. The measurements are described in the following Table I.

TABLE I. ANALYTICAL MEASURES

Evaluating Measures	Description	Formula
Mean Absolute Error (MAE)	It helps to evaluate the mean error in the prediction set, not considering its direction. It indicates the error between prediction and actual observations.	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Root Mean Square Error (RMSE)	It is a type of quadratic grading method that uses to determine the average magnitude of the error.	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Model score	It takes input as a feature matrix and probable target values. Then both predictions made on the feature matrix are compared with the target to attain a score.	mdl.fit (X_train, y_train) mdl.score(X_test, y_test)
Precision	It is the ratio of the true anomalies present with the total number of anomaly candidates.	Tp / (Tp + Fp) where Tp is the number of true positives and Fp is the number of false-positive cases.

Both MAE and RMSE are the average error of the prediction made by the detection algorithms. This error measure works on negatively oriented scoring, which implies the lowest values are better values. For considering larger errors, RMSE is much useful. The more excellent value of RMSE does not mean the larger value of variance in the errors. Using the MAE value, we can always bound the values of RMSE.

- The RMSE error values are either greater than or equal to MAE, i.e., RMSE ≥ MAE.
- When the prediction errors are calculated from the same test dataset, then the difference in MAE and RMSE will be the highest. In that respect, the total squared error will always equal ($MAE^2 \times n$), where n is the data in the test dataset. Thus, RMSE is less than or equal to the absolute squared error's square root, i.e., RMSE ≤ [$MAE^2 \times n$].

Finally, RMSE is more beneficial for penalizing a large number of errors, but both are equally important.

TABLE II. EVALUATING RESULTS OF OUTLIER DETECTION TECHNIQUES

Analytical Measures	Anomaly Detection Approaches			
	Isolation Forest	MCD	OC-SVM	LOF
ROC (Training dataset)	0.9833	0.9864	0.9838	0.8863
ROC (Test dataset)	0.9804	0.9832	0.88	0.8123
Precision	0.8888	0.9028	0.8880	0.9164
Model Score	0.802	0.835	0.814	
Computation Time	0.3	0.3	0.3	<0.1

C. Result Analysis

The evaluating results are tabulated in Table II, representing the ROC and the precision of the different algorithms and the algorithm's model score.

Fig. 3 and 4 represent the MAE and RMSE of the different outlier detection techniques. The figure shows the performance analysis of different detection algorithms; we can derive that the best performing algorithm is Minimum Covariance Determinant (MCD). It has lower MAE and RMSE values of 0.241 and 0.371 than other detection algorithms. It also has a higher prediction accuracy rate of 84%.

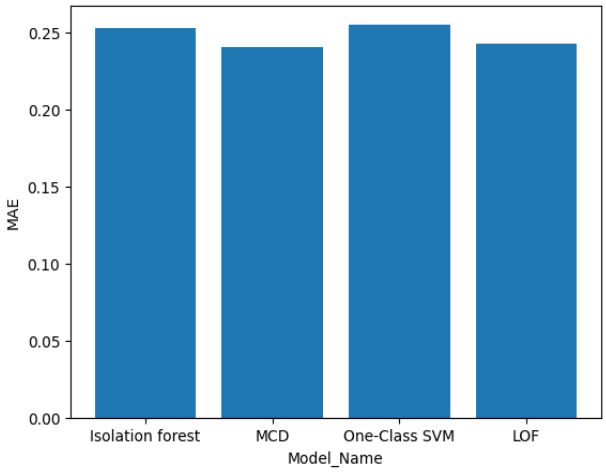

Fig. 3. Mean Absolute Error of Different Detection Algorithm.

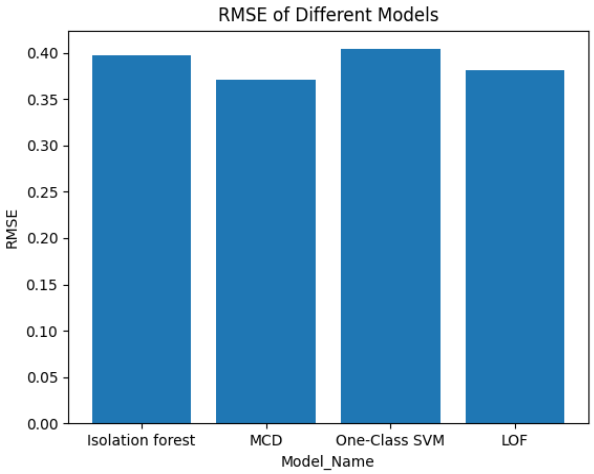

Fig. 4. Root Mean Square Error of the Different Detection Algorithm.

V. Conclusion

This paper discussed the details of the four most popular and versatile outlier detection algorithms. It is a performance analysis-based study where various analytical measures are considered to derive a final result. The breast cancer dataset was considered in our research from multiple datasets available in the real world. The detection algorithms considered are also very diverse. The MAE and RMSE are used for analyzing the performance of different algorithms. Also, the model score of all the algorithms is calculated. The returned value of the model score is measured as these algorithms' accuracy to predict the test dataset after training the model with the training dataset. The slight variance in both the error values can make a lot of difference in analyzing algorithms' performance with the particular dataset. The inclusive conclusion observed that the algorithm that produced a minor error in predicting the test dataset is Minimum Covariance Determinant (MCD) with MAE and RMSE of 0.241 and 0.371. MCD also has the highest accuracy among the other algorithm.

References

[1] C. Romero, and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, e. 1355, Jan. 2020.

[2] A. Dubey, and A. Rasool, "Clustering-Based Hybrid Approach for Multivariate Missing Data Imputation," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 11, pp. 710-714, 2020.

[3] A. Boukerche, L. Zheng, and O. Alfandi, "Outlier Detection: Methods, Models, and Classification," ACM Computing Surveys (CSUR), vol. 53, no. 3, pp. 1-37, Jun. 2020.

[4] D. Xu, and Y. Tian, "A comprehensive survey of clustering algorithms," Annals of Data Science, vol. 2, no 2, pp. 165-193, Aug. 2015.

[5] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S.Y. Philip, "A comprehensive survey on graph neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4-24, Jan. 2021.

[6] M. Fatima, and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9 no. 1, p.1-16, 2017.

[7] L. Breiman, Random forests. Machine learning, vol. 45, no. 1, pp. 5-32, Oct. 2001.

[8] A. S. Hadi, "A new measure of overall potential influence in linear regression," Computational Statistics & Data Analysis, vol. 14, pp. 1–27, 1992.

[9] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," ACM SIGMOD Record, vol. 29, pp. 427–438, 2000.

[10] Z. Pawlak, J. Grzymala-Busse, and W. Ziarko, "Rough sets", Communications of the ACM, Vol. 38, pp. 88–95, 1995.

[11] M. I. Petrovskiy, "Outlier detection algorithms in data mining systems," Programming and Computer Software, Vol. 29, pp. 228–237, 2003.

[12] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," ACM SIGMOD, Vol. 29, pp. 93–104, 2000.

[13] B. Nachman, and D. Shih, "Anomaly detection with density estimation," Physical Review D, vol. 101, no. 7, pp.075042-1-16, Apr. 2020.

[14] B. Tang, and H. He, 2017, "A local density-based approach for outlier detection," Neurocomputing, vol. 241, pp.171-180, 2017.

[15] L. Duan, L. Xu, Y. Liu, and J. Lee, "Cluster-based outlier detection," Annals of Operations Research, Vol. 168, pp. 151–168, 2009.

[16] H. P. Kriegel, P. Kröger, E. Schubert, A. Zimek, (2, "Outlier detection in axis-parallel subspaces of high dimensional data," In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 831-838, 2009.

[17] H. Huang, K. Mehrotra, C. K. Mohan, "Rank-based outlier detection," Journal of Statistical Computation and Simulation, Vol. 83, No. 3, pp. 518-531, 2011.

[18] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," ACM SIGMOD, Vol. 29, pp. 93–104, 2000.

[19] A. Mensi, and M. Bicego, "A novel anomaly score for isolation forests," in proc. International Conference on Image Analysis and Processing, Springer, University of Verona, Verona, Italy, Sep. 2019, pp. 152-163.

[20] M. Hubert, and M. Debruyne, "Minimum covariance determinant," Wiley interdisciplinary reviews: Computational Statistics, Vol. 2, No. 1, pp. 36-43, 2010.

[21] C. Leys, O. Klein, Y. Dominicy, and C. Ley, "Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance," Journal of Experimental Social Psychology, Vol. 74, pp.150-156, Jan. 2018.

[22] M. Hubert, M. Debruyne, and P. J. Rousseeuw, "Minimum covariance determinant and extensions," Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 10, No. 3, e. 1421, 2017.

[23] Y. H. Dovoedo, and S. Chakraborti, "Outlier detection for multivariate skew-normal data: a comparative study," J Stat Comput Simul, Vol. 83, pp. 773–83, 2011.

[24] S. Amraee, A. Vafaei, K. Jamshidi, and P. Adibi, "Abnormal event detection in crowded scenes using one-class SVM," Signal, Image, and Video Processing, vol. 12, no. 6, pp. 1115-1123, Mar. 2018.

[25] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," Pattern Recognition, Vol. 58, pp. 121-134, 2016.

[26] Q. Li, "Covariance modelling with hypersphere decomposition method and modified hypersphere decomposition method," The University of Manchester, Manchester, United Kingdom, 2018.

[27] A. Dubey, and A. Rasool, "Time-Series Missing Value Prediction: Algorithms and Applications," International Conference on Information, Communication and Computing Technology ICICCT, pp. 21-36, 2020.

[28] A. Dubey, and A. Rasool, "Local Similarity-Based Approach for Multivariate Missing Data Imputation." International Journal of Advanced Science and Technology, Vol. 29, No. 06, pp. 9208 – 9215, 2020.

[29] A. Dubey and A. Rasool, "Data Mining based Handling Missing Data," International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud), pp. 483-489, 2019.