Evaluation of ECOD
Empirical Cumulative distribution based Outlier Detection

Abhiraj Ranjan

Indian Institute of Technology Kharagpur

Abstract—Outliers are data points that deviate from the general data distribution. Detection of such data points is what all Outlier Detection models do. The existing models that use unsupervised approaches suffer from 3 major things, especially while working with large datasets having high dimensions:
1.) High Computational Cost
2.) Complex Hyperparameter Tuning
3.) Limited Interpretability
ECOD[1] is a simple algorithm that addresses these issues, with a simple inspiration from the fact that outliers are typically 'rare' events that lie towards the tails of various distribution functions. However, the ECOD model does not know the distribution of the data *a priori*, instead it estimates the underlying distribution by calculating the empirical cumulative distribution per dimension of the data. It then uses these distribution to get the tail probabilities of every feature(dimension) for each data point. ECOD then calculates an Outlier Score[1] for every data point and if this score is beyond a certain threshold, it classifies them as Outliers.

Keywords—*outlier detection, anomaly detection, distributed learning, scalability, empirical cumulative distribution function.*

1. Introduction

Removal of outliers is a crucial preprocessing step in data analysis models, as the presence of outliers can significantly affect the results of statistical analyses. Hence, we need to detect those outliers, which is the goal of Outlier Detection(OD). OD has many applications in fraud detection, network intrusion detection, social media analysis, healthcare, and general machine learning applications. These application demand OD algorithms with:
1.)High detection accuracy
2.)Fast execution
3.)Easy interpretability

Other OD algorithms/models, especially the ones requiring density estimation and pairwise distance calculation suffer from both detection accuracy and runtime efficiency, as the number of data points and dimensions increase. Also, most of the OD models require hyperparameter tuning, which can be difficult in unsupervised setting

2. Problem Formulation and Challenges

We consider the following standard **unsupervised outlier detection (UOD) setup**[13]. We assume that we have n data points $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ that are sampled i.i.d. We collectively refer to this entire dataset by the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, which is formed by stacking the different data points' vectors as rows. Given $\mathbf{X}$, an OD model M assigns, for each data point X_i, an outlier score $O_i \in \mathbb{R}$ (higher means more likely an outlier).

These are a few notable challenges for UOD:
1.) Struggle against dimensionality: many unsupervised algorithms for OD lose accuracy or become computationally non-scalable when $\mathbf{X}$ is high dimensional, that is, the order of magnitude of d becomes comparable to, or larger than that of n, or number of data points n is large.
2.)Limited Interpretability: OD detection algorithms might require interpretability to some extent. For example, to detect whether a financial transaction is an outlier and is considered 'fraudulent', it could be helpful to provide some type of evidence on why a transaction is considered fraudulent or not. This capacity is often not present in many OD models.
3.)Hyperparameter Tuning: without access to any ground truth labels for which data points are outliers, model selection and hyperparameter tuning are challenging in existing OD algorithm.

3. ECOD: The Algorithm

3.1. Motivation and Broad-level Idea

The main motivation behind ECOD is that if the distribution is unimodal, then the outliers are present in tails of this distribution. Hence, this model, for each X_i, computes the probability of observing a point atleast as 'rare' as X_i in terms of tail probabilities.

Let $F : \mathbb{R}^d \to [0,1]$ be the joint CDF across all d dimensions/features. $X_1, X_2, \dots, X_n$ are assumed to be sampled i.i.d. from a distribution with joint CDF F.

For a vector $z \in \mathbb{R}^d$, we denote its j-th entry as $z^{(j)}$, e.g., we write the j-th entry of X_i as $X_i^{(j)}$. We use the random variable X to denote a generic random variable with the same distribution as each X_i.

Then, by the definition of a joint CDF, for any $x \in \mathbb{R}^d$,

$$F(x) = \mathbb{P}\big(X^{(1)} \le x^{(1)}, X^{(2)} \le x^{(2)}, \dots, X^{(d)} \le x^{(d)}\big)$$

$F(X_i)$ is a measure of how extreme X_i is looking at the left tail of the distribution for all the features. Hence $1 - F(X_i)$ is the measure of how extreme X_i is looking at the right tail of the distribution for all the features. Hence, if for any X_i, $F_{\mathbf{X}}(X_i)$ or $1 - F_{\mathbf{X}}(X_i)$ is extremely small, it suggests that X_i is a 'rare' event, hence a potential outlier.

An assumption we take is that the features are mutually independent. This implies

$$F(x) = \prod_{j=1}^{d} F^{(j)}(x^{(j)}) \quad \text{for } x \in \mathbb{R}^d,$$

where $F^{(j)} : \mathbb{R} \to [0,1]$ is the univariate CDF of j^{th} dimension: $F^{(j)}(z) = \mathbb{P}(X^{(j)} \le z) \quad \text{for } z \in \mathbb{R}$.

Using ECDF[2],

Left Tail ECDF:

$$\hat{F}^{(j)}_{\text{left}}(z) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i^{(j)} \le z\} \quad \text{for } z \in \mathbb{R},$$

Right Tail ECDF:

$$\hat{F}^{(j)}_{\text{right}}(z) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i^{(j)} \ge z\} \quad \text{for } z \in \mathbb{R}$$

For joint left and righttail ECDFs across all d dimensions,

$$\hat{F}_{\text{left}}(x) = \prod_{j=1}^{d} \hat{F}^{(j)}_{\text{left}}(x^{(j)}) \quad \text{and} \quad \hat{F}_{\text{right}}(x) = \prod_{j=1}^{d} \hat{F}^{(j)}_{\text{right}}(x^{(j)}) \quad \text{for } x \in \mathbb{R}^d.$$

The ECOD model consists of 2 main steps: computing each dimension's left and right tail ECDFs, then for every point X_i, aggregating its tail probabilities $\hat{F}_{\text{left}}(X_i^{(j)})$ and $\hat{F}_{\text{right}}(X_i^{(j)})$ to come up with an Outlier Score $O_i \in [0, \infty)$; higher means more likely an outlier. These scores are not probabilities, they are just meant to be used to compare among the data points.

3.2. Computing Outlier Score

The main idea here is to use Skewness to determine which tail to choose for which feature. It could be that in a specific dimension, being in the left tail could be considered more of an outlier, whereas in another dimension, we should instead consider the right tail. To automatically decide which tail we should consider for a particular dimension, we consider the skewness of the distribution of that dimension.

Hence to decide which tail to use in ECOD, skewness of the underlying distribution is considered, where the sample skewness coefficient of dimension j can be calculated as below:

$$\gamma_j = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i^{(j)} - \overline{X}^{(j)})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i^{(j)} - \overline{X}^{(j)})^2\right]^{3/2}},$$

where $\overline{X}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(j)}$. When $\gamma_j < 0$, we can thus consider points in the left tail to be more outlying. When $\gamma_j > 0$, we instead consider points in the right tail to be more outlying.

Final aggregation step. Define $O_{left-only}(X_i)$ and $O_{right-only}(X_i)$ as the following:

$$O_{\text{left-only}}(X_i) := -\log \widehat{F}_{\text{left}}(X_i) = -\sum_{j=1}^{d} \log\left(\widehat{F}^{(j)}_{\text{left}}(X_i^{(j)})\right),$$

$$O_{\text{right-only}}(X_i) := -\log \widehat{F}_{\text{right}}(X_i) = -\sum_{j=1}^{d} \log\left(\widehat{F}^{(j)}_{\text{right}}(X_i^{(j)})\right).$$

Introducing skewing based 'Automatic' Outlier Score:

$$O_{\text{auto}}(X_i) = -\sum_{j=1}^{d}\left[\mathbb{1}\{\gamma_j < 0\}\log(\hat{F}^{(j)}_{\text{left}}(X_i^{(j)})) + \mathbb{1}\{\gamma_j \geq 0\}\log(\hat{F}^{(j)}_{\text{right}}(X_i^{(j)}))\right]$$

This $O_{auto}(X_i)$ function automatically picks left or right tail based on skewness of the distribution for each dimension.

Defining O_i as the outlier score for the i^{th} data point:

$$O_i = max[O_{left-only}(X_i), O_{right-only}(X_i), O_{auto}(X_i)]$$

4. Evaluation

The following tables and figures present observations and corresponding conclusions derived from an analysis of original ECOD paper and the outputs of benchmark models.

To evaluate ECOD's efficiency, comparative studies are conducted with established anomaly detection models such as KDE[3] , ABOD[4] , OCSVM[5] , LOF[6] , CBLOF[7] , HOBS[8] , kNN[9] , IForest[10], LODA[11] , and LUNAR[12] , all whose implementations have been adopted from PyOD[16]. Standard evaluation metrics such as ROC-AUC[14] and AUC-PR[15] are used in this evaluation process.

A comparison between ECOD, ECOD L, ECOD R is also made to show the significance of including skewness while calculating anomaly scores. ECOD L uses O_i as $O_{left-only}(X_i)$, while ECOD R uses O_i as $O_{right-only}(X_i)$.

Finally, an attempt is made to include the effect of feature correlation as a possible improvement strategy, and the observations are noted, and conclusions are drawn.

4.1. ECOD vs other SOTA models

Table 1. compares the ROC-AUC values for different models on various datasets. Figure 1. plots the histogram for ROC AUC scores across different models. Table 2. compares the ROC-AUC values for different models on various datasets. Figure 2. plots the histogram for ROC AUC scores across different models.

Table 3. compares the False Positives Count for different models on various datasets. Figure 3. plots the histogram for False Positives Counts across different models. Table 4. compares the False Negatives Count for different models on various datasets. Figure 4. plots the histogram for False Negatives Count across different models.

Algorithm	Cardio	Ionosphere	Mammography	Satimage-2	Statlog	Pima
ECOD	0.9713	0.8618	0.9110	0.9751	0.6736	0.6963
ABOD	0.9541	0.9438	0.5922	0.9959	0.8428	0.6576
OCSVM	0.9580	0.8469	0.8773	1.0000	1.0000	1.0000
KDE	0.9856	0.9538	0.8716	1.0000	1.0000	1.0000
LOF	0.9467	0.9574	0.8638	0.9945	0.8611	0.6866
CBLOF	0.9396	0.9815	0.8386	0.9981	0.8411	0.6595
HBOS	0.8751	0.7707	0.8524	0.9713	0.8690	0.7031
kNN	0.9388	0.9782	0.8809	0.9993	0.8876	0.7126
IForest	0.9471	0.9176	0.8805	0.9938	0.7806	0.7316
LODA	0.8784	0.8632	0.8829	0.9884	0.6958	0.6550
LUNAR	0.9328	0.9592	0.8708	0.9958	0.8326	0.7932

Table 1. ROC AUC scores for different models on various datasets

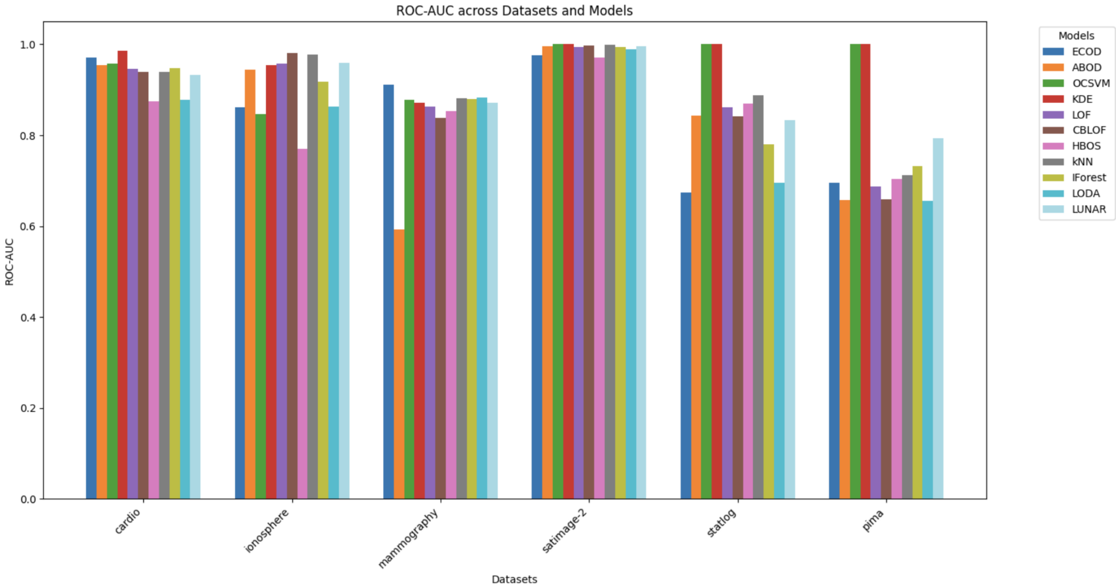

Figure 1. ROC-AUC Scores Across Different Models

Algorithm	Cardio	Ionosphere	Mammography	Satimage-2	Statlog	Pima
ECOD	0.7935	0.8102	0.5108	0.8546	0.6883	0.5697
ABOD	0.6611	0.9273	0.0264	0.9205	0.8061	0.5039
OCSVM	0.6678	0.8199	0.2526	1.0000	1.0000	1.0000
KDE	0.9561	0.9415	0.3282	1.0000	1.0000	1.0000
LOF	0.6509	0.9283	0.2431	0.9255	0.8374	0.5365
CBLOF	0.6139	0.9720	0.2421	0.9665	0.7902	0.5151
HBOS	0.5444	0.6105	0.1496	0.7343	0.8139	0.5415
kNN	0.6908	0.9679	0.3098	0.9762	0.8473	0.5678
IForest	0.6933	0.8733	0.2645	0.9188	0.7556	0.5686
LODA	0.5879	0.7707	0.2978	0.9063	0.7094	0.4650
LUNAR	0.6934	0.9510	0.4852	0.9646	0.8027	0.6341

Table 2. AUC PR scores for different models on various datasets

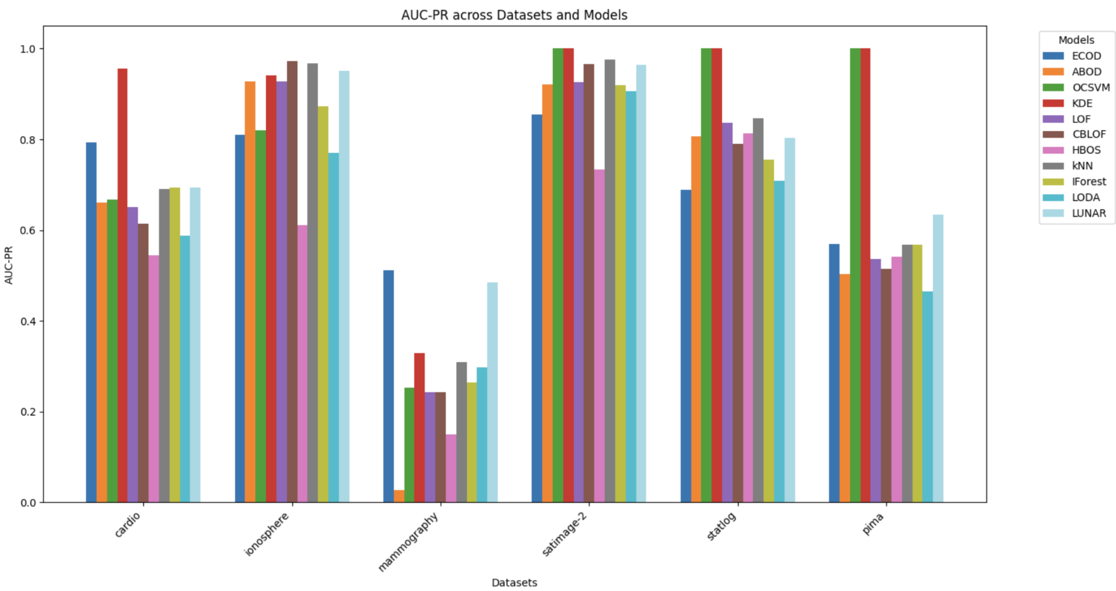

Figure 2. AUC PR Scores Across Different Models

Algorithm	Cardio	Ionosphere	Mammography	Satimage-2	Statlog	Pima
ECOD	165	17	1083	566	432	48
ABOD	178	29	0	804	638	50
OCSVM	166	23	1093	574	440	50
KDE	166	23	1093	0	0	0
LOF	143	19	992	495	383	49
CBLOF	166	23	1093	574	440	50
HBOS	166	23	1083	574	440	50
kNN	133	18	944	498	396	41
IForest	166	23	1093	574	440	50
LODA	165	23	1093	573	439	50
LUNAR	166	23	1093	574	440	50

Table 3. Number of False Positives for different models on various datasets

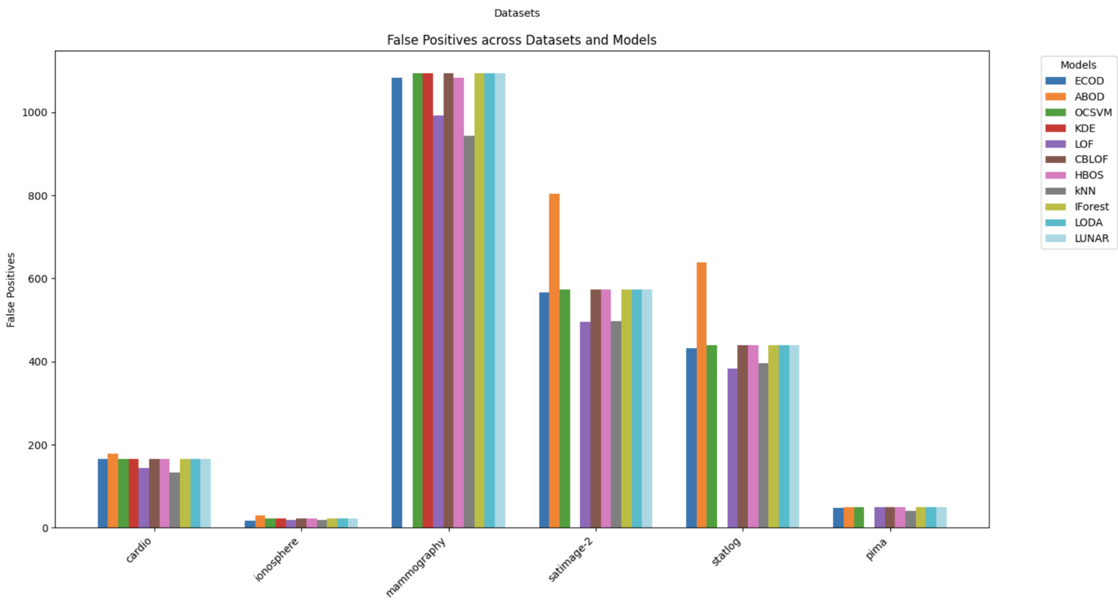

Figure 3. False Positives Count Across Different Models

Algorithm	Cardio	Ionosphere	Mammography	Satimage-2	Statlog	Pima
ECOD	10	51	48	5	942	181
ABOD	16	17	260	0	620	200
OCSVM	21	47	102	0	0	0
KDE	9	19	106	0	0	0
LOF	31	17	103	2	613	184
CBLOF	24	5	111	0	753	181
HBOS	71	90	110	5	688	190
kNN	45	12	90	0	589	186
IForest	34	36	97	1	790	184
LODA	67	55	90	2	880	222
LUNAR	38	13	112	1	685	161

Table 4. Number of False Negatives for different models on various datasets

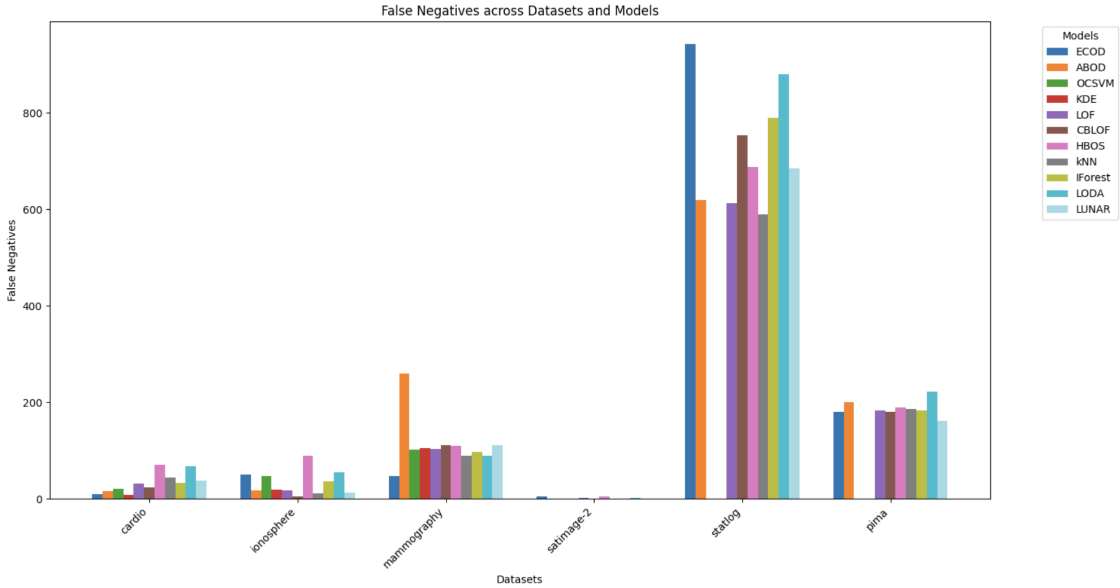

Figure 4. False Negatives Count Across Different Models

4.2. Analysis on the Dataset: mammography

To visualize the effect of various OD models, along with ECOD, an attempt to plot the outliers and the normal instances on a 2-D plane has been made. Since the mammography dataset has 6 dimensions, we have used t-SNE[17] to compress it to 2 dimensions. Figure 5. shows the 2-D plot of the mammography dataset:

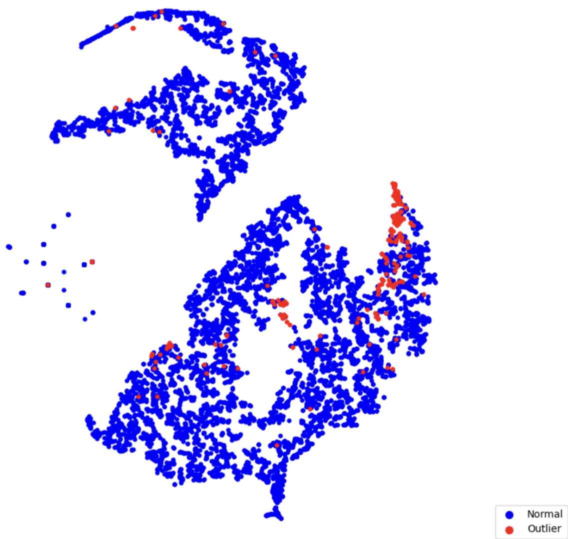

Figure 5. mammography dataset compressed to 2 dimensions

Figure 6. shows the 2D plots of predictions of the evaluated models on the dataset.

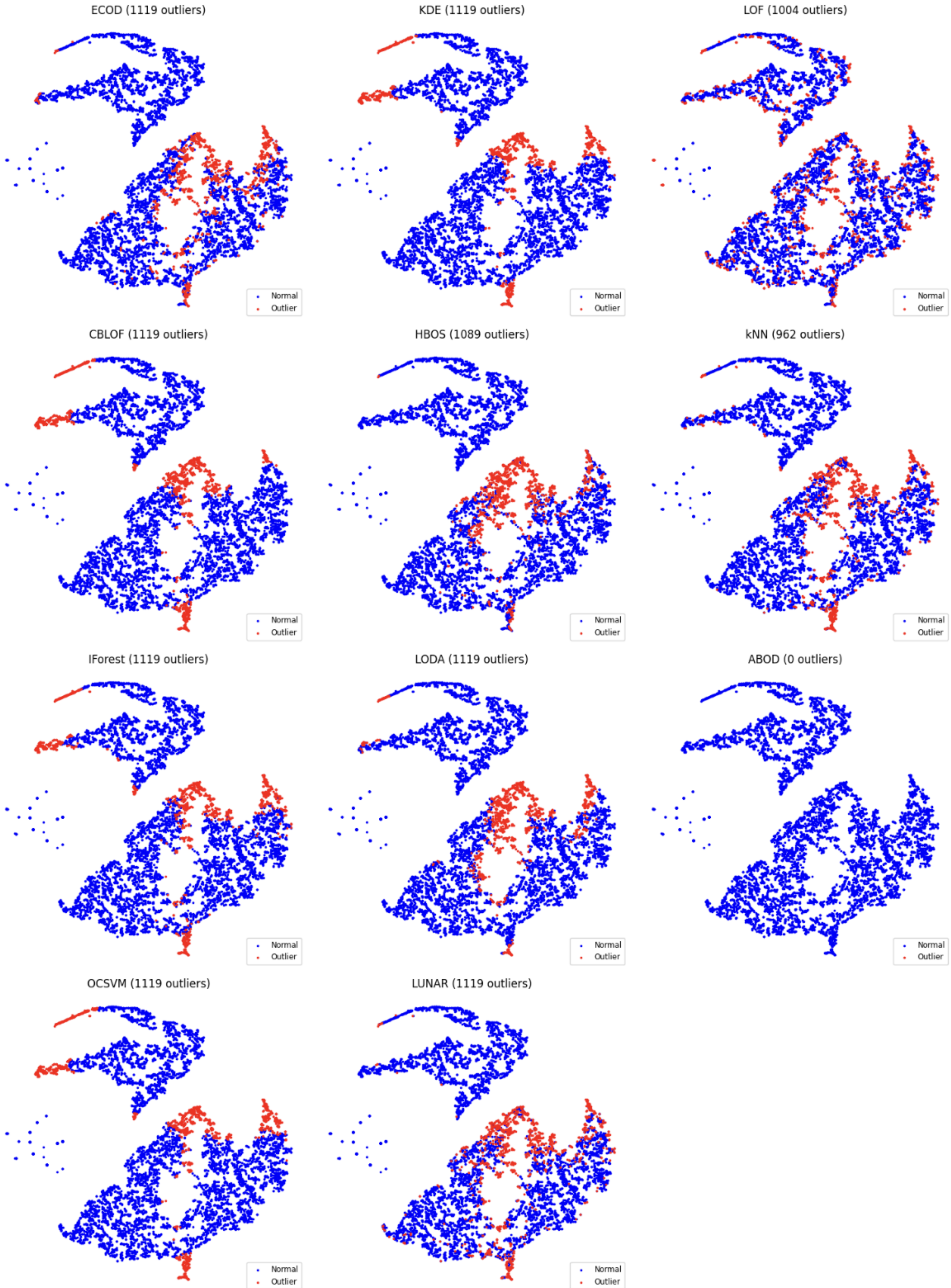

Figure 6. 2D plot of predictions of models over mammography dataset

4.3. Skewness and Outlier Scores

To show the importance of considering Skewness while calculating O_i $\forall i \in \{1, 2, \dots, n\}$, we do a comparison between different possible variants of ECOD, namely ECOD$_L$[1], ECOD$_R$[1], ECOD$_B$[1].
1.) ECOD$_L$ considers only the left tail for computing O_i.
2.) ECOD$_R$ considers only the right tail for computing O_i.
3.) ECOD$_B$ considers the average of both, left tail and right tail, for computing O_i.

Algorithm	Cardio	Ionosphere	Mammography	Satimage-2	Statlog	Pima
$ECOD$	0.9713	0.8618	0.9110	0.9751	0.6736	0.6963
$ECOD_L$	0.9249	0.8592	0.2277	0.9571	0.8088	0.2453
$ECOD_R$	0.7804	0.5158	0.9079	0.8863	0.5000	0.7759

Table 5. ROC AUC scores comparison among $ECOD, ECOD_L, ECOD_R, ECOD_B$

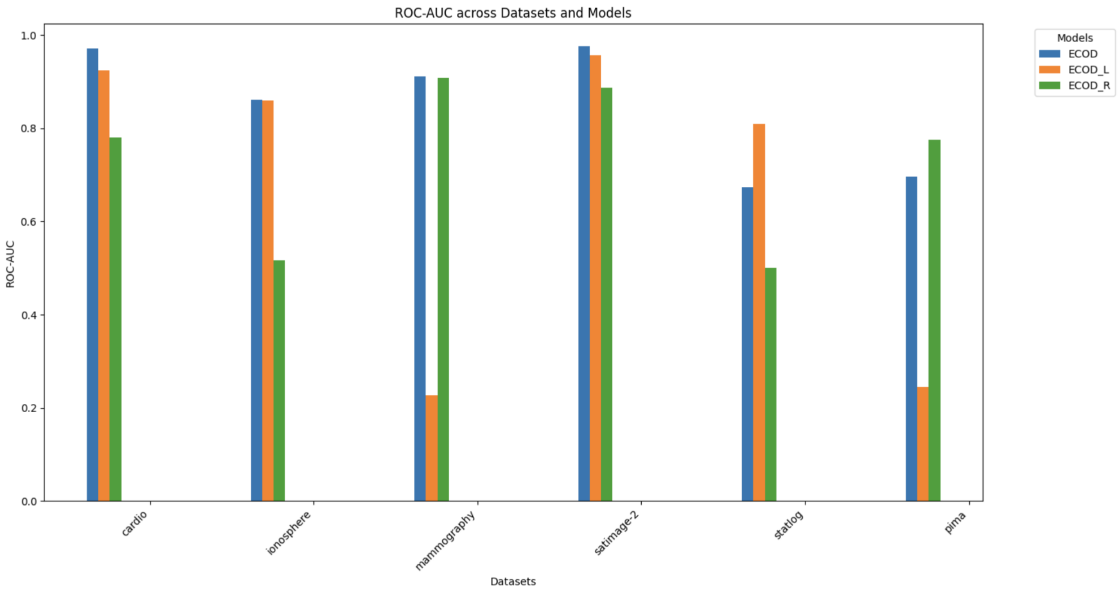

Figure 7. ROC-AUC Scores comparison among $ECOD, ECOD_L, ECOD_R, ECOD_B$

Algorithm	Cardio	Ionosphere	Mammography	Satimage-2	Statlog	Pima
$ECOD$	0.7935	0.8102	0.5108	0.8546	0.6883	0.5697
$ECOD_L$	0.6260	0.7946	0.0139	0.3428	0.7441	0.2343
$ECOD_R$	0.2413	0.4578	0.5042	0.0907	0.3860	0.6402

Table 6. AUC PR scores comparison among $ECOD, ECOD_L, ECOD_R, ECOD_B$

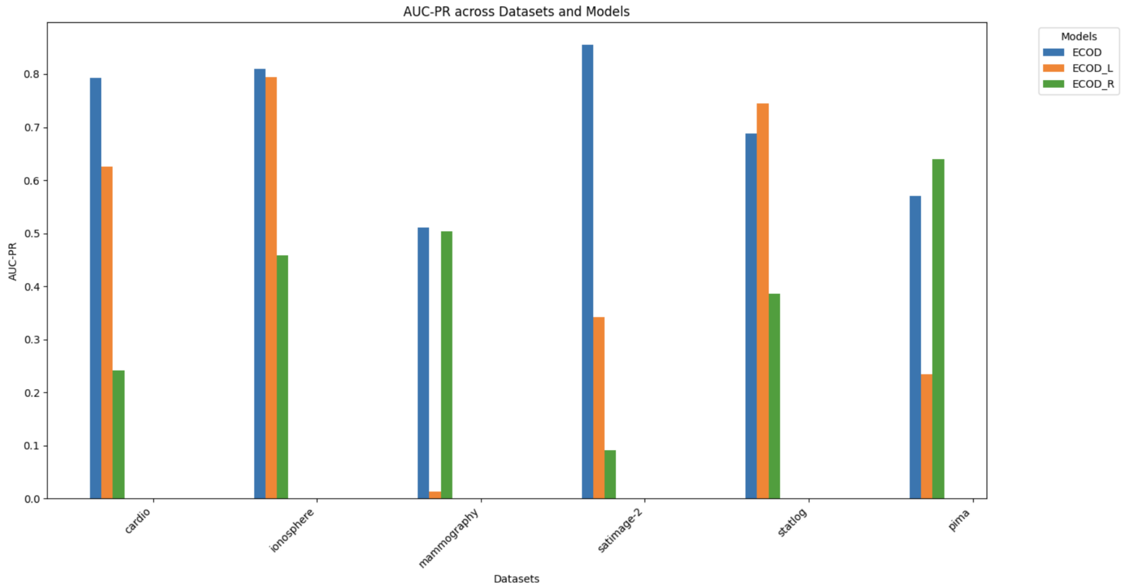

Figure 8. AUC PR comparison among $ECOD, ECOD_L, ECOD_R, ECOD_B$

Algorithm	Cardio	Ionosphere	Mammography	Satimage-2	Statlog	Pima
$ECOD$	165	17	1083	566	432	48
$ECOD_L$	162	18	0	571	438	47
$ECOD_R$	161	19	1086	569	436	48

Table 7. Number of False Positives comparison among $ECOD, ECOD_L, ECOD_R, ECOD_B$

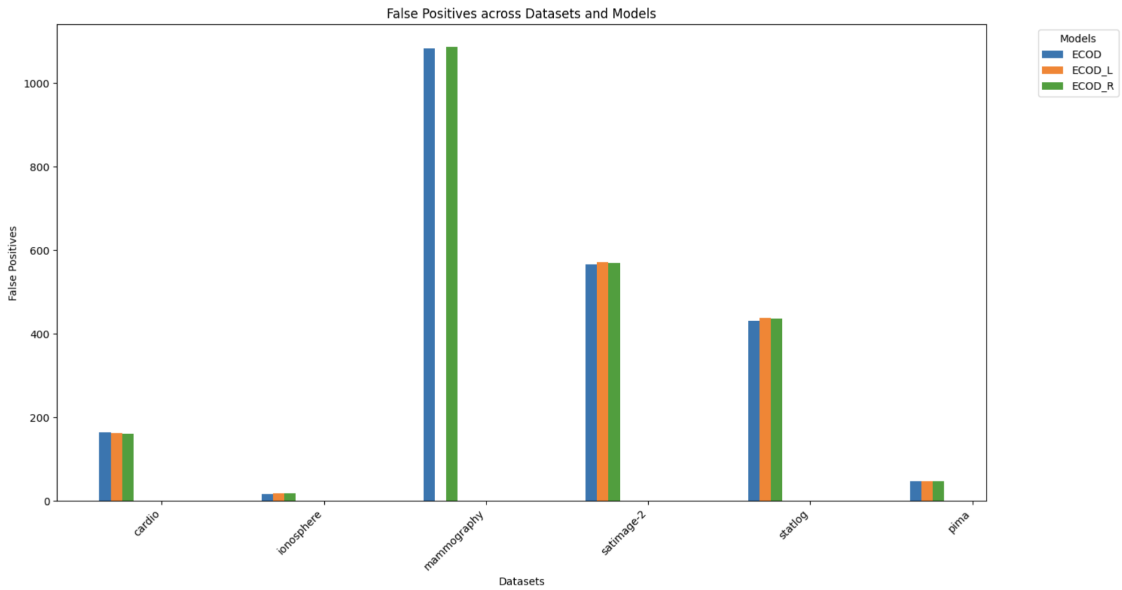

Figure 9. False Positives Count comparison among $ECOD, ECOD_L, ECOD_R, ECOD_B$

Algorithm	Cardio	Ionosphere	Mammography	Satimage-2	Statlog	Pima
$ECOD$	10	51	48	5	942	181
$ECOD_L$	43	52	260	10	922	266
$ECOD_R$	120	98	52	28	1597	163

Table 8. Number of False Negatives comparison among $ECOD, ECOD_L, ECOD_R, ECOD_B$

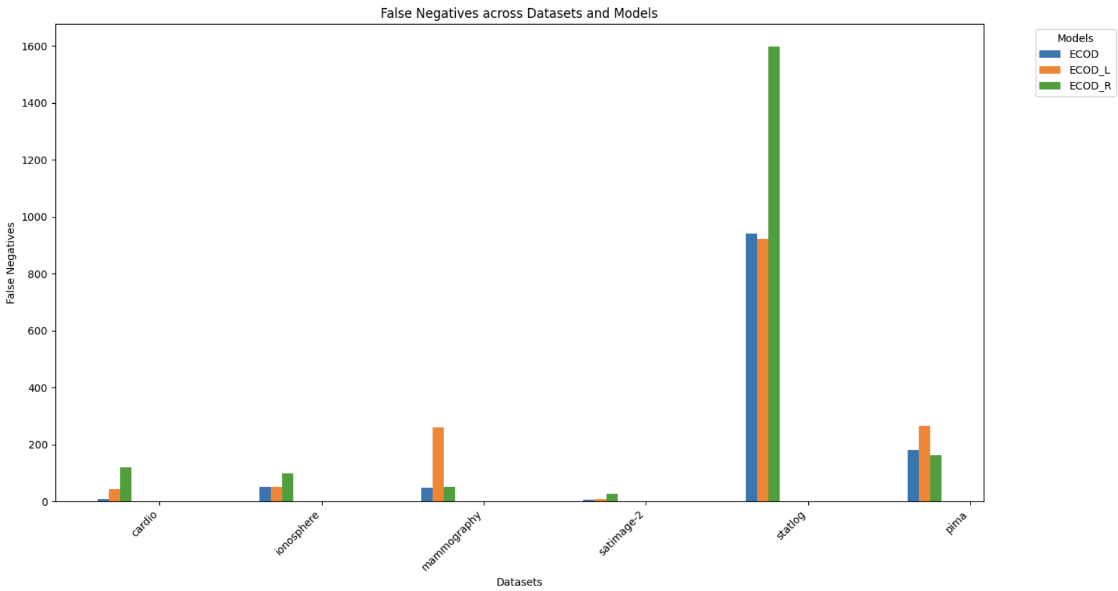

Figure 10. False Negatives Count comparison among $ECOD, ECOD_L, ECOD_R, ECOD_B$

From the histograms(Figures 7-10), we conclude that ECOD performs much better over all the datasets than its variants that don't consider skewness of the distribution.

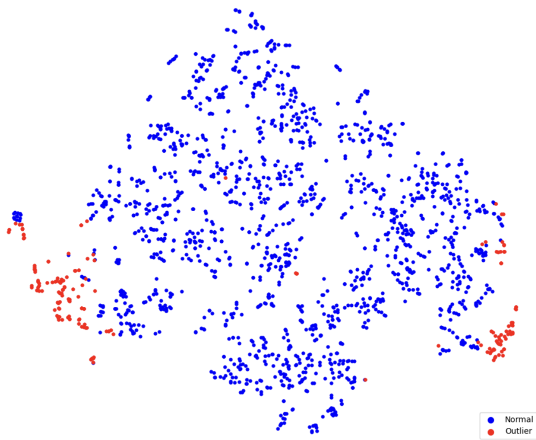

Figure 11. cardio dataset compressed to 2 dimensions using t-SNE

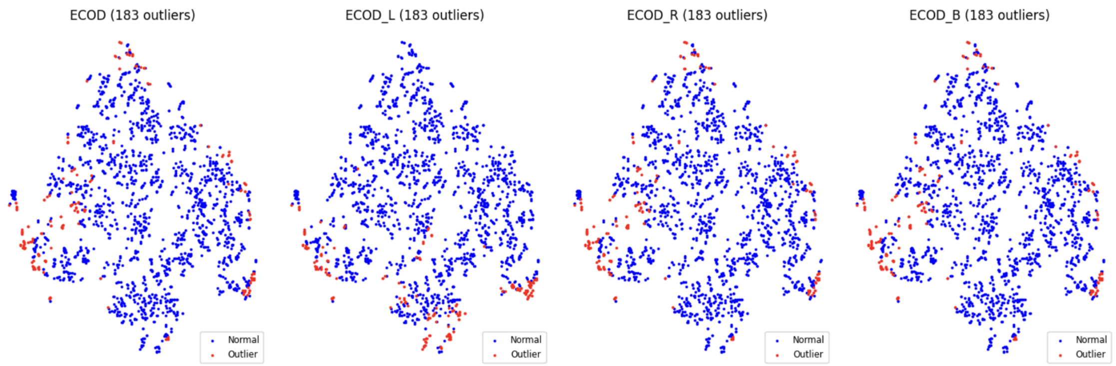

Figure 12. Predictions on cardio dataset by ECOD, $ECOD_L$, $ECOD_R$, $ECOD_B$

From Figure 11. and Figure 12., it can be observed that $ECOD_L$ predicts the anomalies correctly only if they lie on the left tail extremes, while it misses right tail anomalies. Similarly $ECOD_R$ misses right tail extremes. $ECOD_B$ only predicts half of both the tails, and is hence not correct. ECOD seems to work better than its counterparts.

5. Correlation in ECOD: A Failed Attempt at Improvement

5.1. The Idea

The ECOD model operates under the fundamental assumption that features are mutually independent. However, this assumption may not always hold, making it essential to explore a more general approach that accounts for feature correlations. To quantify this, I define a correlation score C_i:

$$C_i := \sum_{j=1}^{d} |C_{ij}|$$

where C_{ij} is the correlation between i^{th} and j^{th} feature. The core idea is that a feature with a high correlation score plays a more significant role, as even a slight change in its value influences many other features. This could make certain feature combinations 'rare,' potentially signaling anomalies. Thus, assigning greater weight to more 'important' features may enhance the model's effectiveness. Lets name this customized ECOD model $ECOD_{CUSTOM}$.

5.2. Experimentation and Results

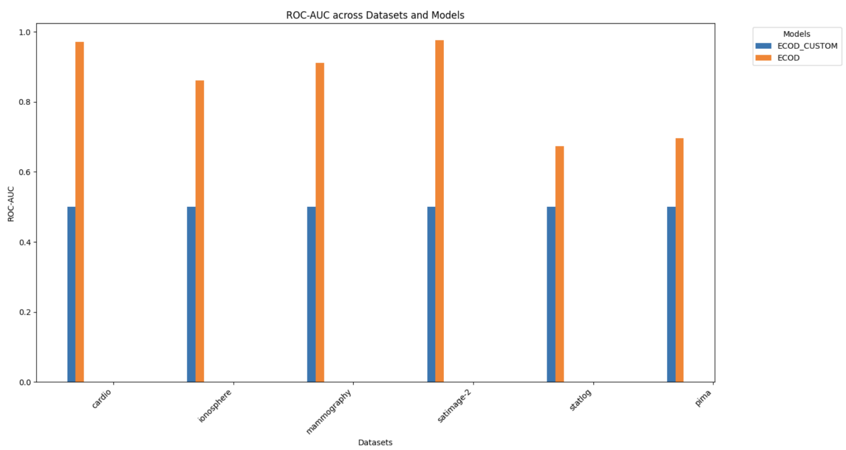

Figure 13. ROC-AUC comparison between ECOD and $ECOD_{CUSTOM}$

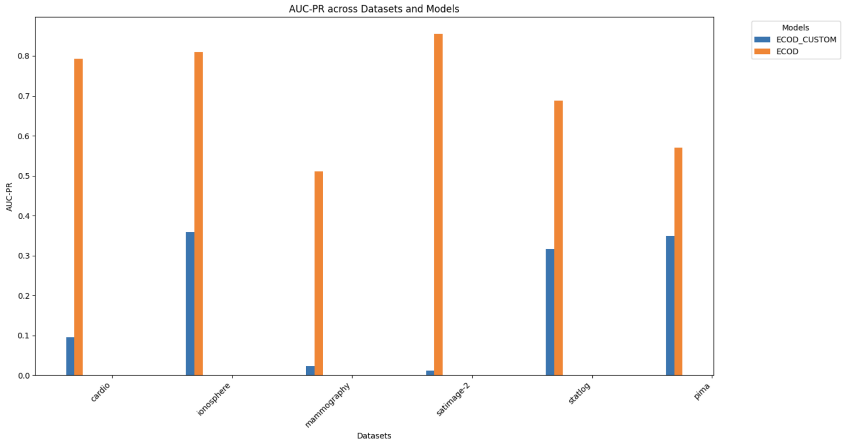

Figure 14. AUC-PR comparison between ECOD and $ECOD_{CUSTOM}$

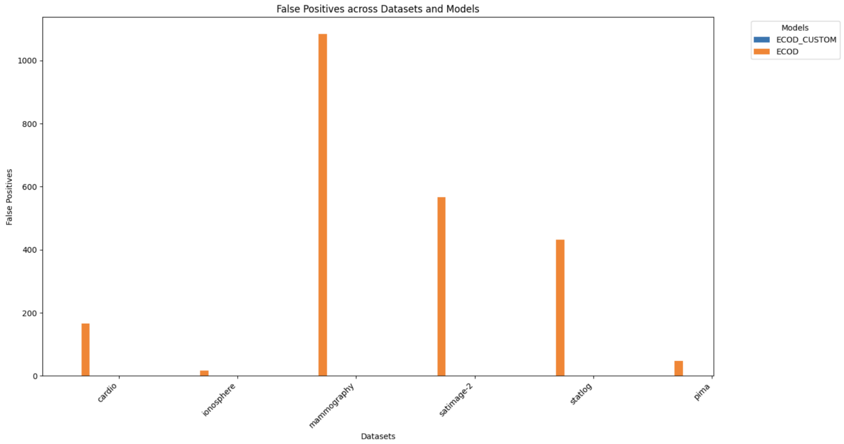

Figure 15. False Positives Count comparison between ECOD and $ECOD_{CUSTOM}$

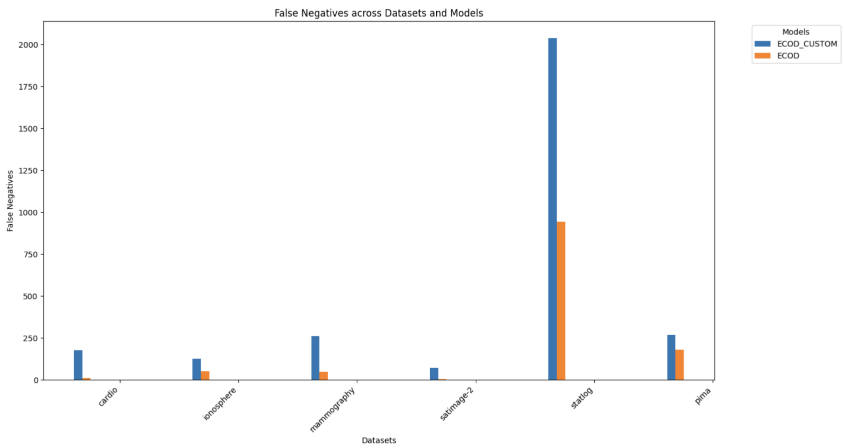

Figure 16. False Negatives Count comparison between ECOD and $ECOD_{CUSTOM}$

From Figures 13-16, it is evident that the proposed Idea of weighing dimensions differently based on their 'importance' fails.

We further analyze $ECOD_{CUSTOM}$ on cardio dataset. Firstly, the correlation matrix(containing Pearson correlation coefficients[18]) for cardio is calculated.

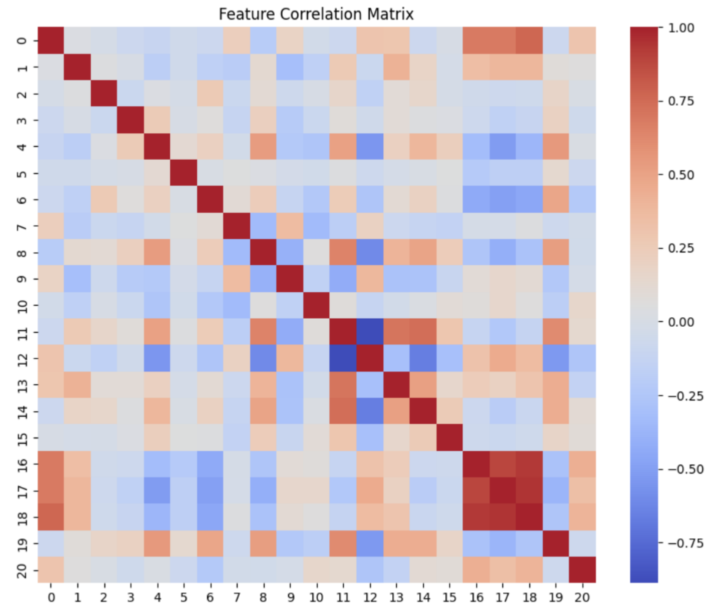

Figure 17. Correlation matrix with Pearson Correlation Coefficients

We compress the cardio dataset to 2 dimensions using t-SNE and plot the predictions of ECOD and $ECOD_{CUSTOM}$.

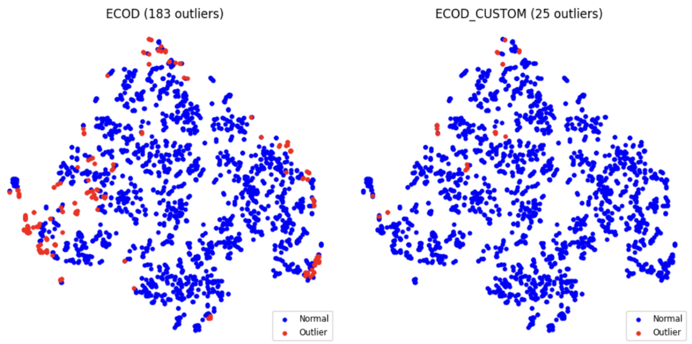

Figure 18. Predictions on cardio by ECOD and $ECOD_{CUSTOM}$

It is clear that $ECOD_{CUSTOM}$ failed to detect many anomalies.(See Figure 11. for cardio dataset's plot in 2D)

5.3. Potential Cause of Failure

Increasing the importance of highly correlated features can lead to missing out on valuable information provided by uncorrelated features. When two features are highly correlated, obtaining information from one feature does not add much new insight, as the second feature is largely predictable. This results in redundancy if the weights of such correlated features are increased. Additionally, assigning lower weights to uncorrelated features may cause the loss of potentially important information, as these features contribute unique and unpredictable insights.

6. Conclusion

This section presents the key takeaways from the evaluation of ECOD.

1. **Nonparametric and Unsupervised**: ECOD is a nonparametric, unsupervised learning model. It does not require hyperparameter tuning, which is often challenging in the unsupervised domain.

2. **Optimal Runtime Complexity**: The model has a time complexity of $O(nd)$, which is the most efficient among similar methods.
3. **Independence Assumption**: ECOD operates under the assumption that the features of the dataset are mutually independent.
4. **Performance Across Different Dimensionalities**: ECOD performs well on datasets with a small to medium number of dimensions. However, in high-dimensional settings, where data tends to become sparse, it struggles to converge effectively.
5. **Comparison with Variants**: ECOD outperforms its counterparts ($ECOD_L$, $ECOD_R$, $ECOD_B$) as it effectively leverages the skewness of data, which its variants do not utilize.
6. **Exploration of Feature Correlation**: Since ECOD assumes feature independence, an alternative approach was explored by incorporating feature correlations. However, the results were not satisfactory. Despite this, the analysis provided valuable insights into highly correlated features, leaving an open question for future research.

7. References

[1] Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., & Chen, G. H. (2022). "ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions." *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 12181-12193.
[2] Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., & Chen, G. H. (2022). "ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions." *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 12181-12193.
[3] Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
[4] Kriegel, H. P., Schubert, M., & Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 444-452).
[5] Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., & Williamson, R.C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443-1471.
[6] Breunig, M.M., Kriegel, H.P., Ng, R.T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 93-104.
[7] He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10), 1641-1650.
[8] Hobbs, J.R. (1978). Resolving pronoun references. Lingua, 44(4), 311-338.
[9] Fix, E., & Hodges, J.L. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.
[10] Liu, F.T., Ting, K.M., & Zhou, Z.H. (2008). Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, 413-422.
[11] Pevný, T. (2016). Loda: Lightweight on-line detector of anomalies. Machine Learning, 102(2), 275-304.
[12] Goodge, A., Hooi, B., Ng, S.K., & Ng, W.S. (2022). Unifying Local Outlier Detection Methods via Graph Neural Networks. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), 6737-6745.
[13] Hawkins, D.M. (1980). Identification of Outliers. Chapman and Hall, London.
[14] Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.
[15] Boyd, K., Eng, K.H., & Page, C.D. (2013). Area Under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In Proceedings of the 2013 IEEE International Conference on Data Mining, 1-10.
[16] Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD: A Python Toolbox for Scalable Outlier Detection. Journal of Machine Learning Research, 20(96), 1-7.
[17] Maaten, L.v.d., & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9, 2579-2605.
[18] Pearson, K. (1895). Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58, 240-242.