1 Hyperparameter Tuning and Validation

This subsection describes the search methodologies for setting the optimal values of the parameters and their range with the analysis of their sensitivity.

For this, initially, the search methodology used for setting the optimal value of the parameter K is presented. It is used to calculate the nearest neighbors in the neighborhood K. Then, the search methodology used for setting the optimal value of the parameter M is presented. It is used to detect the global and local outliers such that data points having mutual neighbors greater than or equal to M are analysed for detecting the global outliers, and data points having mutual neighbors less than M are analysed for detecting the local outliers.

Finally, the search methodology used for setting the optimal value of the parameter $\mathcal{T}$ using the mean μ and standard deviation σ of the distance factors $Dis_{Factor}(d_i \in D)$ of each data point $d_i \in D$ belonging to the data set D, is presented. This threshold $\mathcal{T}$ is used for detecting the global outliers.

For setting these parameters, we analysed the 12 two-dimensional synthetic data sets [1] that are mentioned in subsection 4.1.

1.0.1 Search Methodology for Finding the Optimal Value of the Parameter K

In this subsection, we outline the methodology employed to determine the optimal value of the parameter (K). To achieve this, we computed the nearest neighbors for the various neighborhood sizes (K) within the range of 2 to 100, applying this approach to all 12 synthetic data sets as detailed in Table 1.

Table 1: Synthetic Data Sets

Dataset name	Instances	Outliers	Dimensions
SD01	1043	43	2
SD02	1000	85	2
SD03	1039	41	2
SD04	1641	45	2
SD05	876	77	2
SD06	1372	72	2
SD07	1037	37	2
SD08	2259	159	2
SD09	1034	36	2
SD10	2042	64	2
SD11	1020	26	2
SD12	1242	50	2

Table 2: Public Data Sets

Dataset name	Instances	Dimensions
PD1: Glass	214	9
PD2: Ionosphere	351	34
PD3: Cardiotocography	2126	23
PD4: Lymphography	148	18
PD5: Vowels	1455	12
PD6: Diabetes	768	8
PD7: Breast cancer	682	10
PD8: Wheat seeds	210	7
PD9: Iris	150	4
PD10: Haberman	306	3

For objective assessment, we have shown results in Table 3 for K, 2 to 25 for the synthetic data sets SD01 to SD04. The entry shown in this Table 3 represents the number of data points having no mutual neighbors (#DP). For subjective assessment, we have also shown the results in Fig. 1 for all 12 synthetic data sets.

Thus, from Table 3 and Fig. 1, it can be easily noticed that the value $K = 10$ serves as a transition point. Below the value of $K = 10$, there are a large number of data points having no mutual neighbors, and above the value of $K = 10$, the variation in the number of data points having no mutual neighbors becomes stable. Thus, this analysis shows the sensitivity of the parameter K below and above 10. Hence, the value 10 of the parameter K is fixed for our proposed methodology.

Further, the search methodology for finding the optimal value of the parameter M is discussed in the following subsection.

1.0.2 Search Methodology for Finding the Optimal Value of the Parameter *M*

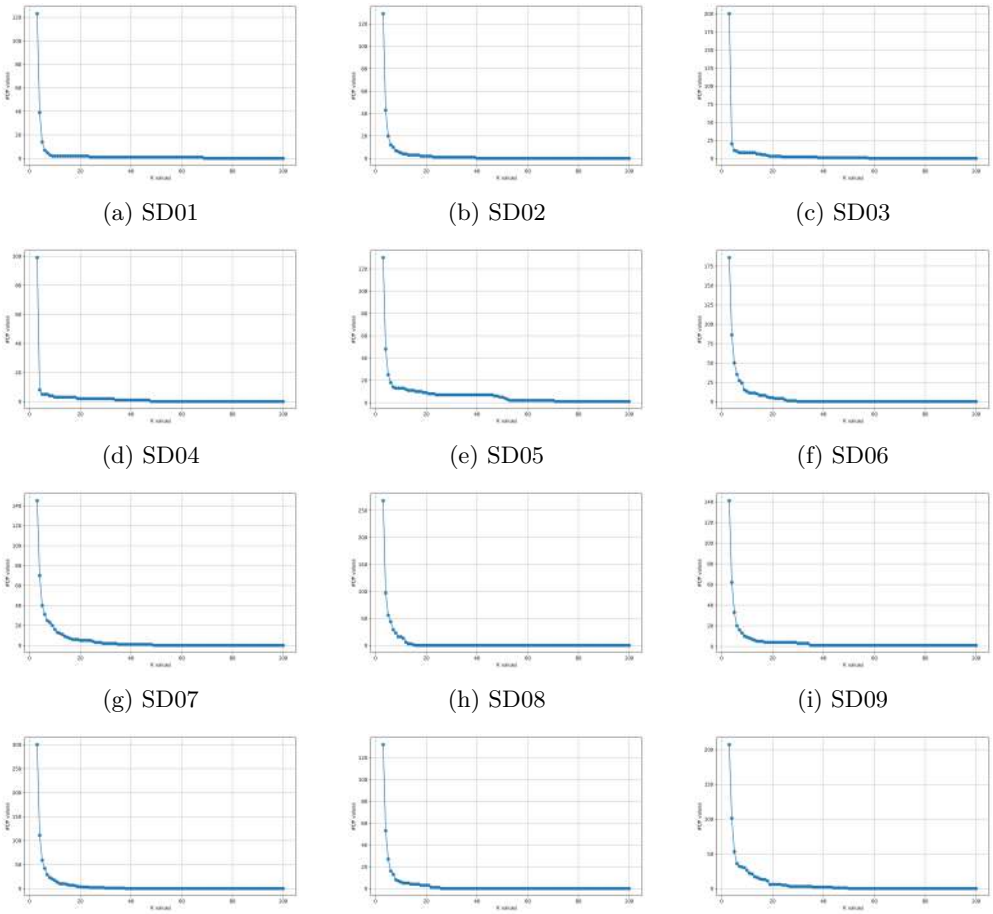
In this subsection, we present the search methodology for setting the parameter M, i.e., used for detecting the global and local outliers. For this, we have calculated the data points having mutual neighbors less than M for the synthetic data set SD01 for M = 3, 4, 5, and parameter $K = 10$ (this optimal value is fixed in the Subsection 1.0.1). We started the searching from $M = 3$, because for the detection of local outliers in our proposed methodology, we compare the estimated local density of the data point with the average estimated density of $(M - 1)$ nearest inliers and $(M - 1)$ nearest outliers of this data point. So, if $M = 2$, then $(M - 1) = 1$, but for calculating the average estimated density of inliers and outliers, the value of $(M - 1)$ needs to be at least 2. Therefore, we started searching from $M = 3$. The results for $M = 3$, 4, and 5 are shown in Fig. 2.

Thus, from Fig. 2, we can observe that for $M = 4$ and $M = 5$, the data points located at the center of the dense region are marked as the data points having mutual

Table 3: Number of data points having no mutual neighbors (#DP) for different data sets

K values	#DP for			
	SD01	SD02	SD03	SD04
2	123	129	200	99
3	39	43	20	8
4	14	20	11	5
5	7	12	10	5
6	5	10	8	5
7	3	7	8	4
8	2	6	8	4
9	2	5	8	3
10	2	4	8	3
11	2	4	8	3
12	2	3	8	3
13	2	3	6	3
14	2	3	6	3
15	2	3	5	3
16	2	3	5	3
17	2	2	4	3
18	2	2	3	2
19	2	2	3	2
20	2	2	3	2
21	2	2	3	2
22	2	1	3	2
23	1	1	2	2
24	1	1	2	2
25	1	1	2	2

neighbors less than M, and for detecting the local outliers, these data points are the potential outliers. It implies that these data points will be analyzed to check whether these are outliers or inliers. However, the data points located at the center of the dense region cannot be outliers. Hence, setting the value of parameter $M = 4$ and 5 is incorrect. Thus, this analysis shows the sensitivity of the parameter M after the value 3. Hence, the value 3 of the parameter M is fixed for our proposed methodology. Furthermore, the results shown in Fig. 3 for the data sets SD02 to SD12 support the optimal value of the parameter M as 3.



(a) SD01 (b) SD02 (c) SD03

(d) SD04 (e) SD05 (f) SD06

(g) SD07 (h) SD08 (i) SD09

(j) SD10 (k) SD11 (l) SD12

Fig. 1: Results of the Number of Data Points having no Mutual Neighbors (#DP) for Synthetic Data Sets for different values of K.

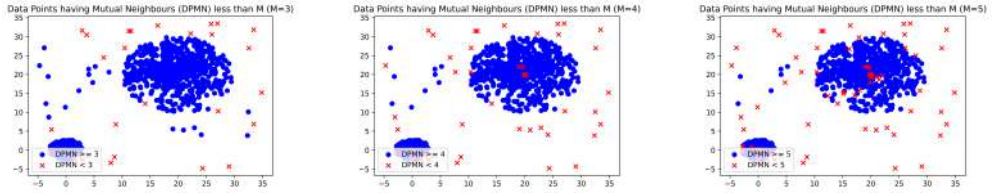
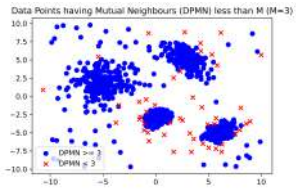
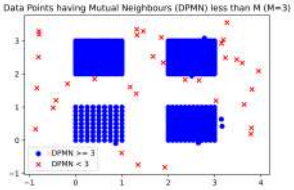


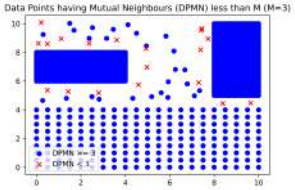
Fig. 2: Data points having Mutual Neighbors less than different values of M.



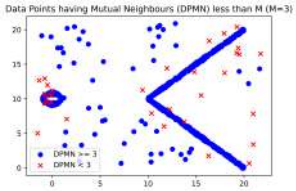
(a) SD02



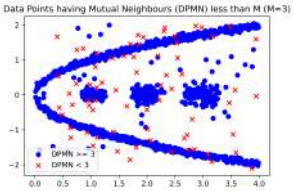
(b) SD03



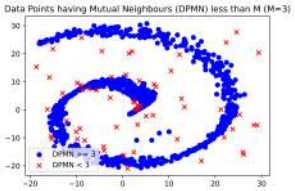
(c) SD04



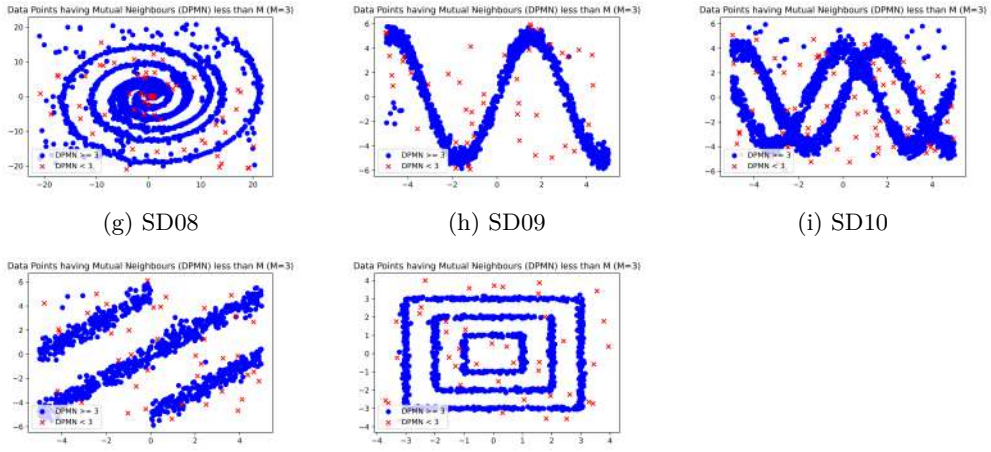
(d) SD05



(e) SD06



(f) SD07



(g) SD08 (h) SD09 (i) SD10

(j) SD11 (k) SD12

Fig. 3: Result for the Data Points having Mutual Neighbours less than M ($M = 3$) for Synthetic Data Sets.

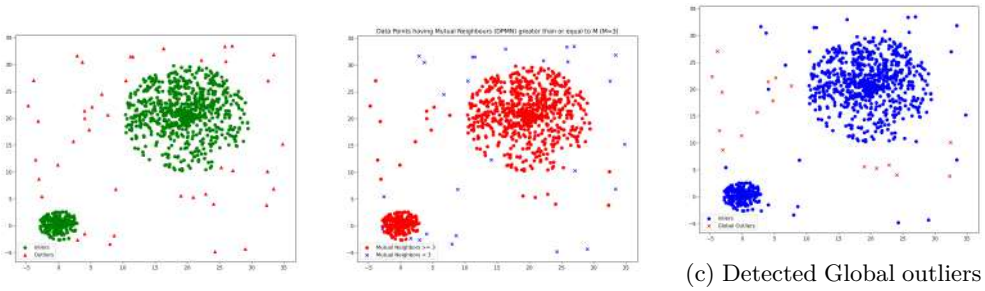
Further, the search methodology for finding the optimal value of the parameter $\mathcal{T}$ is discussed in the following subsection.

1.0.3 Search Methodology for Finding the Optimal Value of the Parameter $\mathcal{T}$

In this subsection, we present the search methodology for setting the optimal value of the parameter $\mathcal{T}$ using the mean μ and standard deviation σ of the distance factors $Dis_{Factor}(d_i \in D)$ of each data point $d_i \in D$ belonging to the data set D, i.e., established as a threshold for detecting the global outliers. For this, first we identify the data points having mutual neighbors greater than or equal to M, and using this threshold $\mathcal{T}$, we detect the outliers, which are the actual outliers.

The reason for establishing the threshold $\mathcal{T}$ using the mean μ and standard deviation σ is that the unsupervised outlier detection methodologies were largely grounded in statistical analysis and assumptions like the 3σ rule, and the data points that lie beyond the $\mu \pm 3\sigma$ limit are identified as outliers [2].

Further, to check the correctness of this $\mu+3\sigma$ limit as a threshold $\mathcal{T}$, we executed the proposed methodology on the data set SD01 for the parameters $K = 10$, and $M = 3$ (these optimal values are fixed in the Subsections 1.0.1, and 1.0.2) and found that few data points, which are actual outliers, are not detected (i.e., shown in Fig. 4).

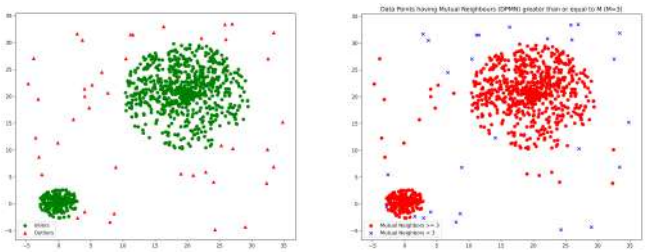


(a) Actual Data (b) DPMN $\geq M$ (c) Detected Global outliers using $\mu + 3\sigma$

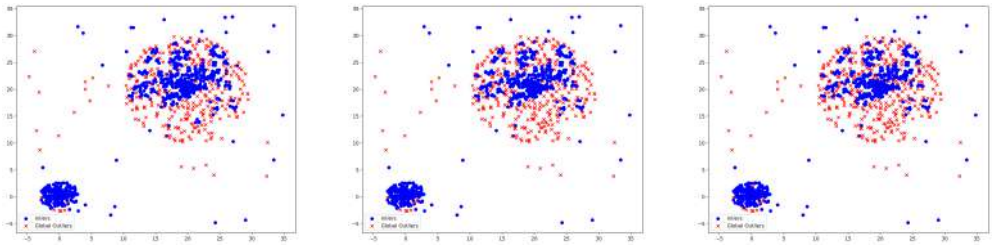
Fig. 4: Results of detected Global Outliers by using the threshold $\mathcal{T} = \mu + 3\sigma$.

Therefore, we set the $\mathcal{T}$ as $\mu + 3\frac{\sigma}{\sqrt{n}}$, with the reason that σ represents dispersion of data points around the mean, and $\frac{\sigma}{\sqrt{n}}$ represents dispersion of the sample mean. Hence, dividing σ by $\sqrt{n}$ reflects the principle that as the sample size increases, the estimation of the population mean becomes more precise.

Further, we analyzed the proposed methodology on the same data set SD01, using the parameters $K = 10$ and $M = 3$ for threshold $\mathcal{T} = \mu + 3\frac{\sigma}{\sqrt{n}}$, and then for $\mathcal{T} = \mu + 2\frac{\sigma}{\sqrt{n}}$, $\mu + \frac{\sigma}{\sqrt{n}}$. For $\mathcal{T} = \mu + 3\frac{\sigma}{\sqrt{n}}$, the center data points are also detected as outliers. Similarly for $\mu + 2\frac{\sigma}{\sqrt{n}}$ and $\mu + \frac{\sigma}{\sqrt{n}}$, results are not proper (i.e., shown in Fig. 5).



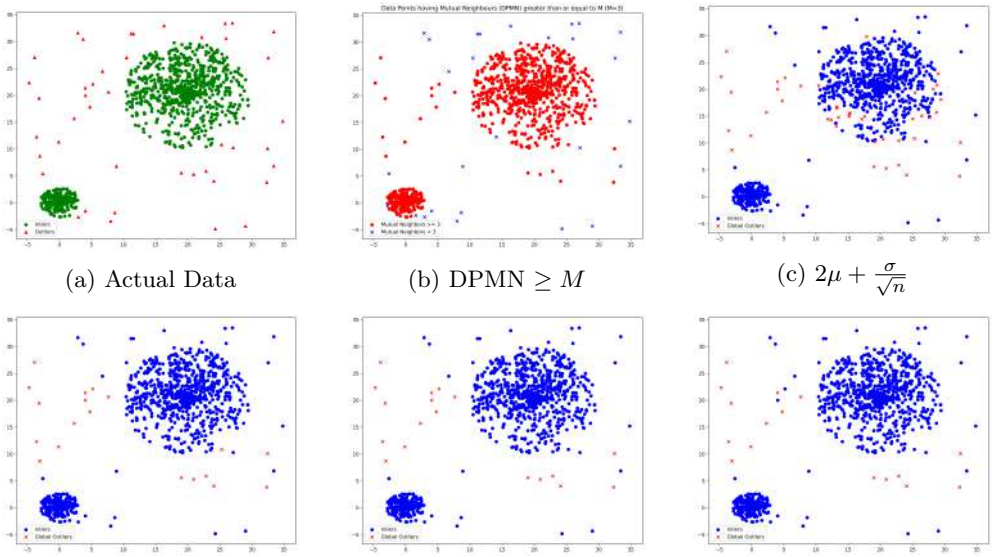
(a) Actual Data (b) DPMN $\geq M$



(c) $\mu + 3\frac{\sigma}{\sqrt{n}}$ (d) $\mu + 2\frac{\sigma}{\sqrt{n}}$ (e) $\mu + 1\frac{\sigma}{\sqrt{n}}$

Fig. 5: Results of different values of c for threshold $\mathcal{T} = \mu + c \cdot \frac{\sigma}{\sqrt{n}}$.

Therefore, we analyzed the results (shown in Fig. 6). for $\mathcal{T} = 2\mu + \frac{\sigma}{\sqrt{n}}$, $3\mu + \frac{\sigma}{\sqrt{n}}$, $4\mu + \frac{\sigma}{\sqrt{n}}$, and $5\mu + \frac{\sigma}{\sqrt{n}}$.



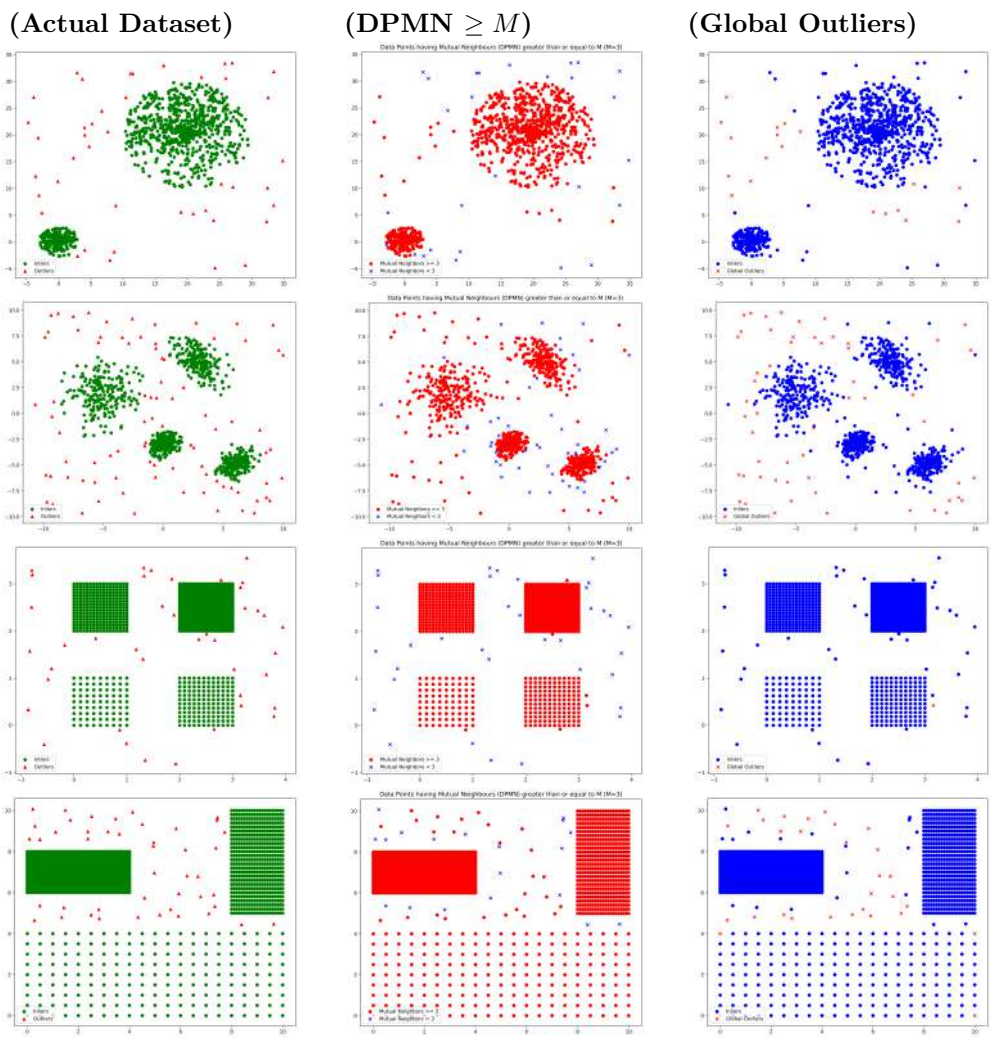
(a) Actual Data (b) DPMN $\geq M$ (c) $2\mu + \frac{\sigma}{\sqrt{n}}$

(d) $3\mu + \frac{\sigma}{\sqrt{n}}$ (e) $4\mu + \frac{\sigma}{\sqrt{n}}$ (f) $5\mu + \frac{\sigma}{\sqrt{n}}$

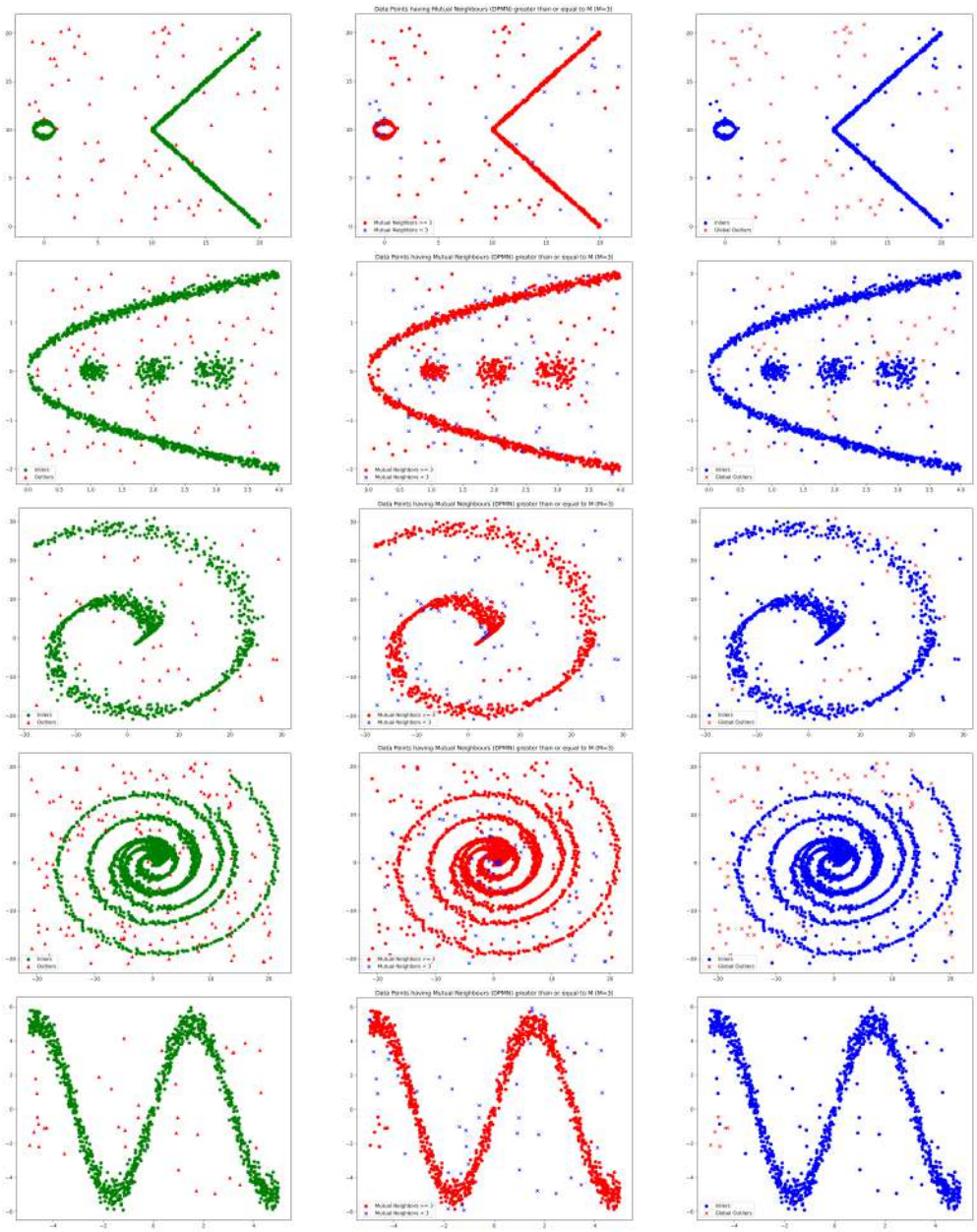
Fig. 6: Results of different values of c for $c \cdot \mu + \frac{\sigma}{\sqrt{n}}$.

From Fig. 6, it is clear that when we use $\mathcal{T} = 2\mu + \frac{\sigma}{\sqrt{n}}$, then some of the data points of dense clusters are detected as outliers and many of the actual outliers are

missed, and for $\mathcal{T} = 3\mu + \frac{\sigma}{\sqrt{n}}$, it correctly detects the outliers. For $4\mu + \frac{\sigma}{\sqrt{n}}$ and $5\mu + \frac{\sigma}{\sqrt{n}}$, it also correctly detects the outliers; therefore, we set the optimal value of the parameter $\mathcal{T}$ as $3\mu + \frac{\sigma}{\sqrt{n}}$. Furthermore, the results shown in Fig. 7 for the data sets SD02 to SD12 support the optimal value of the parameter $\mathcal{T}$ as $3\mu + \frac{\sigma}{\sqrt{n}}$.



(contd...)



(contd...)

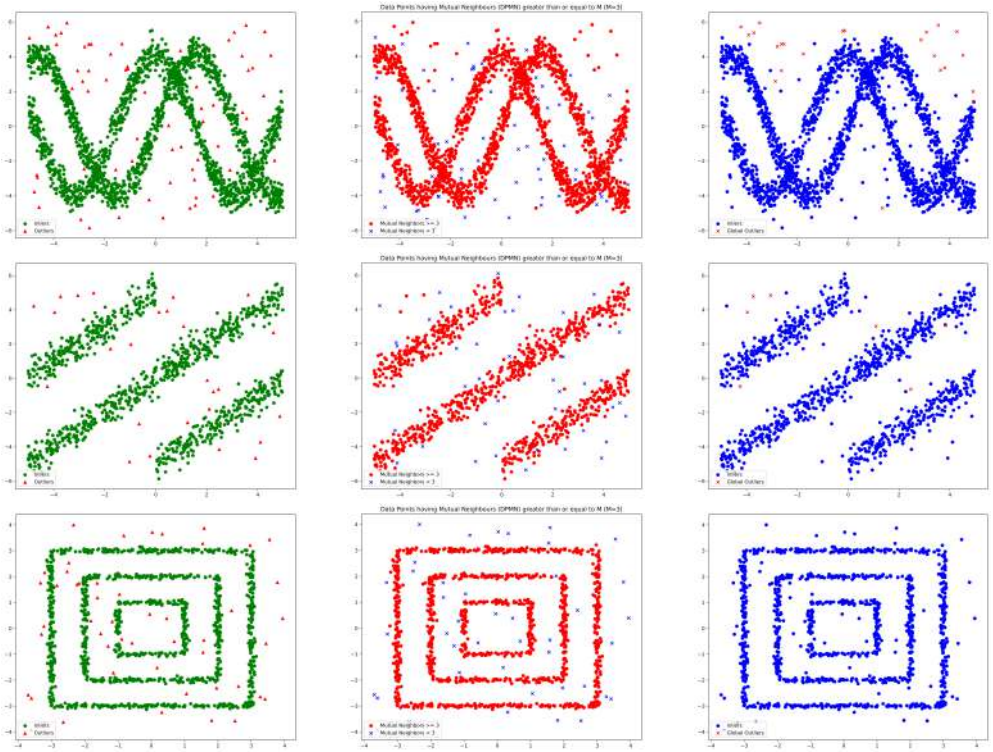


Fig. 7: Results of synthetic data sets of detected Global Outliers by using threshold $\mathcal{T} = 3\mu + \frac{\sigma}{\sqrt{n}}$.

1.0.4 Validation of Parameters K, M, and $\mathcal{T}$

This subsection validates that the optimal values of parameters $K = 10$, $M = 3$, and $\mathcal{T} = 3\mu + \dfrac{\sigma}{\sqrt{n}}$ (obtained in subsections 1.0.1, 1.0.2, and 1.0.3) correctly detect the outliers for the proposed methodology. For this, we used the performance metric Area Under the Receiver Operating Characteristic Curve (AUC) [3].

For validating these parameters, we fix the parameters $M = 3$, and $\mathcal{T} = 3\mu + \dfrac{\sigma}{\sqrt{n}}$, and we started varying the parameter K from 2 to 70 for the proposed methodology for all the synthetic data sets mentioned in Table 1, and then we calculated the average AUC. The result is shown in Fig. 8. The reason for fixing the parameters $M = 3$, and $\mathcal{T} = 3\mu + \dfrac{\sigma}{\sqrt{n}}$ and varying the parameter K for calculating the average AUC is that the value of the parameters $M = 3$, and $\mathcal{T} = 3\mu + \dfrac{\sigma}{\sqrt{n}}$ depends on K in the proposed methodology. In the proposed methodology, first we compute the

nearest neighbors in the neighborhood K, then the mutual neighbor, M^{th} mutual neighbor distance, and M^{th} nearest neighbor distance, and using these we establish the threshold $\mathcal{T}$. Moreover, it is discussed in Preliminaries: Mutual Neighbour that the proper selection of parameter K is needed to obtain the good results. Also, using the concept of mutual neighbor with the optimal value of parameter K reduces the computational complexity for large data sets [4]. Thus, we vary the parameter K for validating its optimal value.

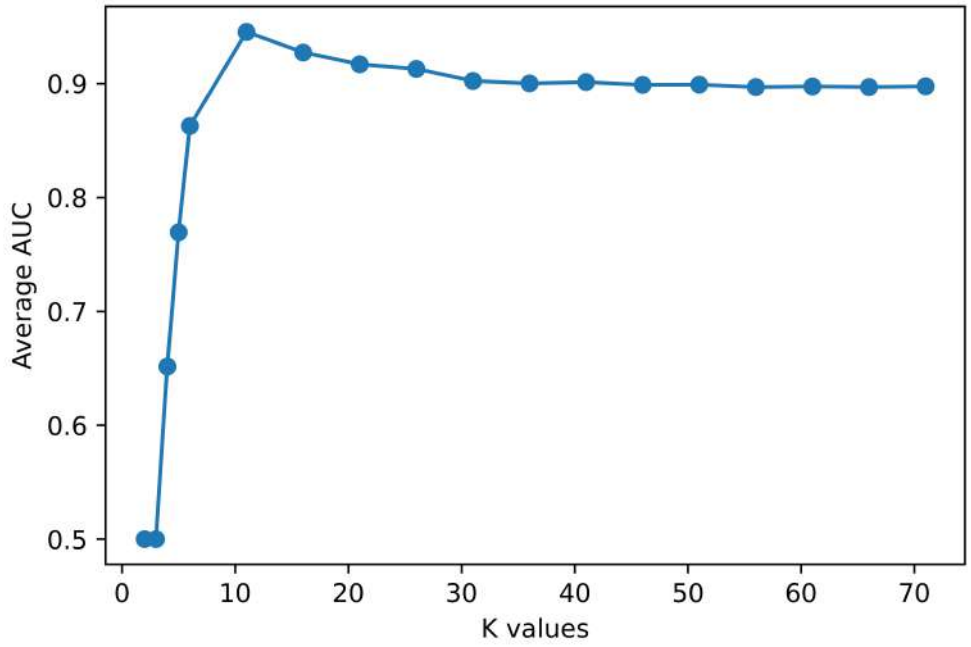


Fig. 8: Average AUC values for different values of K.

From Fig. 8, it can be easily noticed that the average AUC improves as parameter K increases up to 10; thereafter, it starts reducing. The peak average AUC value is observed at value 10 of the parameter K, validating that the optimal values of parameters are $K = 10$, $M = 3$, and $\mathcal{T} = 3\mu + \frac{\sigma}{\sqrt{n}}$ for correctly detecting the outliers for the proposed methodology.

1.1 Parameters sensitivity test between K and M using grid search

We have conducted a comprehensive grid search-based sensitivity analysis for the hyper parameters K (number of nearest neighbors) and M (minimum mutual neighbors), which directly influence the outlier detection performance of the proposed methodology.

1.1.1 Grid Search Implementation:

We systematically varied K and M across a defined range using grid search to evaluate the Area Under the ROC Curve (AUC) for each combination. The search space included values of $K \in \{5, 7, 9, \ldots, 20\}$ and $M \in \{1, 2, 3, \ldots, 20\}$.

1.1.2 Visualization:

The Fig. 9 presents a heatmap illustrating the sensitivity of the AUC (Area Under the ROC Curve) score to variations in the hyperparameters K (number of nearest neighbors) and M (the data points having mutual neighbors less than M). The vertical axis represents different values of K, while the horizontal axis corresponds to values of M. Each cell shows the AUC score obtained from grid search for a specific (K, M) pair, with a color gradient from purple (low AUC) to yellow (high AUC). The results indicate that detection performance is highly influenced by both parameters. This suggests that moderate values of both parameters strike a balance between including sufficient neighbors and ensuring reliable mutual connections. Conversely, very low or very high values of either parameter result in degraded performance. This visualization demonstrates the importance of systematic tuning and supports the choice of final parameter values used in our model.

These visualizations illustrate the AUC performance across the grid of K and M, allowing us to:
1) Clearly observe the regions of optimal performance.
2) Identify performance degradation when the parameters deviate significantly from the optimal range.
3) Justify the selected hyperparameter values based on consistent high performance across multiple datasets.

Based on this rigorous analysis, we selected the K and M values that yielded the highest AUC scores, while also considering generalization across datasets to avoid overfitting to any specific dataset.

References

[1] Dataset: https://github.com/AIHIsora/DCROD/tree/main/dataset. https://github.com/AIHIsora/DCROD/tree/main/dataset

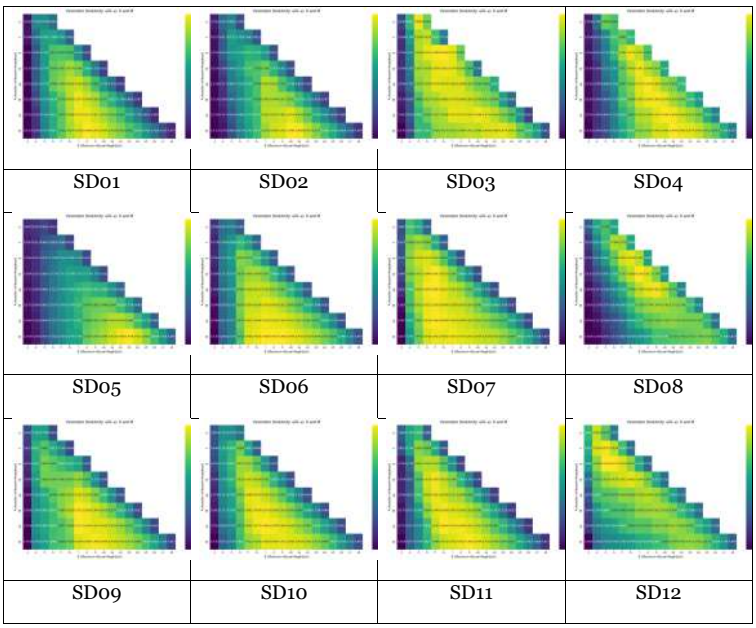


Fig. 9: Heatmap for synthetic datasets.

[2] Li, K., Gao, X., Fu, S., Diao, X., Ye, P., Xue, B., Yu, J., Huang, Z.: Robust outlier detection based on the changing rate of directed density ratio. Expert Systems with Applications, 117988 (2022)

[3] Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine learning, 203-231 (2001)

[4] Zhu, Q., Feng, J., Huang, J.: Natural neighbor: A self-adaptive neighborhood method without parameter k. Pattern recognition letters 80, 30-36 (2016)