Anomaly Detection on Hyperthyroidism Dataset

Mustafa Soydan
Physics Department
Univerity of Milan Bicocca
Milan/ITALY
m.soydan@campus.unimib.it

Kerem Erciyes
Physics Department
Univerity of Milan Bicocca
Milan/ITALY
k.erciyes@campus.unimib.it

***Abstract*—This study demonstrates an anomaly detection on hyperthyroidism dataset which has both categorical and continuous features. To detect anomalies, Nearest Neighbors, DBSCAN and Isolation Forest algorithms used and the results of each algorithm are compared.**

Keywords—anomaly detection, unsupervised learning, mixed data.

I. Introduction

Hyperthyroidism is a prevalent endocrine disorder characterized by the excessive production of thyroid hormones. The early detection and diagnosis of hyperthyroidism are crucial for effective treatment and management. However, the complexity of medical data, which often includes both continuous and categorical features, poses significant challenges for anomaly detection. Identifying anomalies in such datasets is essential as these anomalies can indicate potential cases of misdiagnosis, irregularities in patient data, or outliers that may lead to new medical insights.

The techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Nearest Neighbors, and Isolation Forest are applied for this purpose. Each of these methods has its unique advantages and limitations when applied to datasets with mixed data types. This study aims to apply and compare these three anomaly detection techniques on a hyperthyroidism dataset.

II. Methodology

A. Data Preprocessing

The unnecessary columns in the data are discarded. To apply Nearest Neighbors and DBSCAN Gower Matrix is calculated and used to work with binary and continuous data together. Before calculating the Gower Matrix, the continuous features scaled using standard scaler. To apply the Isolation Forest Algorithm, continuous features are scaled using standard scaler and categorical features are

B. Algorithms

One distance based approach, Nearest Neighbors (NN), one clustering based approach, DBSCAN, and Isolation Forest method is used for anomaly detection in the given dataset.

C. Evaluation of Results

To see the results are accurate, tSNE visualization method, metrics e.g. homogeneity, completeness, V-measure, Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI). The DBSCAN clustering results are evaluated with intra-inter clusters, silhouette score, and correlation between ideal similarity matrix and proximity matrix.

III. Data Preprocessing

A. Hyperthyroidism Dataset

Table 1: Hyperthyroidism Dataset

Variable	Range	Type
Dim_0	0.01-0.97	Object
Dim_1	0-1	Int64
Dim_2	0-1	Int64
Dim_3	0-1	Int64
Dim_4	0-1	Int64
Dim_5	0-1	Int64
Dim_6	0-1	Int64
Dim_7	0-1	Int64
Dim_8	0-1	Int64
Dim_9	0-1	Int64
Dim_10	0-1	Int64
Dim_11	0-1	Int64
Dim_12	0-1	Int64
Dim_13	0-1	Int64
Dim_14	0-1	Int64
Dim_15	0-1	Int64
Dim_16	0-0.53	Object
Dim_17	0.0005-0.18	Object
Dim_18	0.002-0.6	Object
Dim_19	0.017-0.233	Object
Dim_20	0.002-0.642	Object

The dataset used in this study comprises both continuous and categorical features as seen at Table 1. To ensure optimal performance of our anomaly detection algorithms, several preprocessing steps applied to the raw data.

As Figure 1 and Figure 2 represents the continuous features show diverse distribution patterns. Many features,

like Dim_16, Dim_17, Dim_18, and Dim_20, are right-skewed, with most values clustered at the lower end and a few outliers. Dim_0 and Dim_19 have more normal distributions, although Dim_19 still has some outliers.

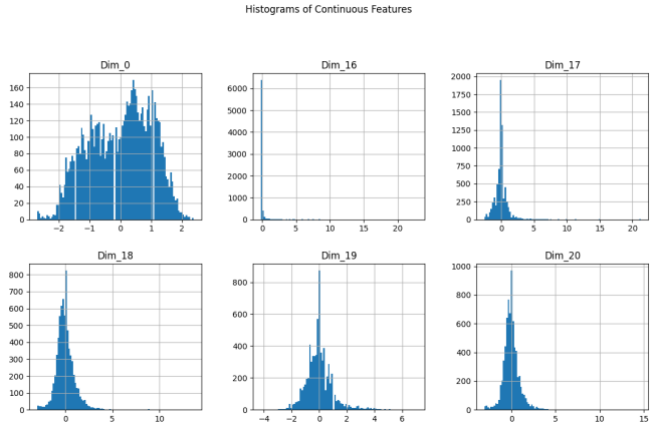

Figure 1: Histogram of Continuous Features

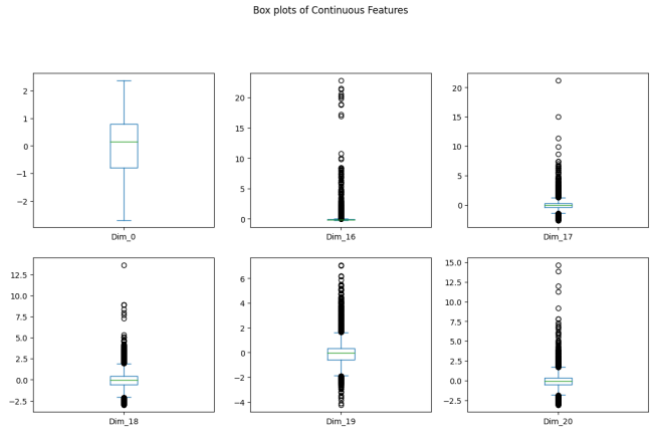

Figure 2: Boxplot of Continuous Features

Figure 3 represents the categorical features (Dim_1 to Dim_15) are binary. Their distributions vary, with some features balanced between categories and others skewed.

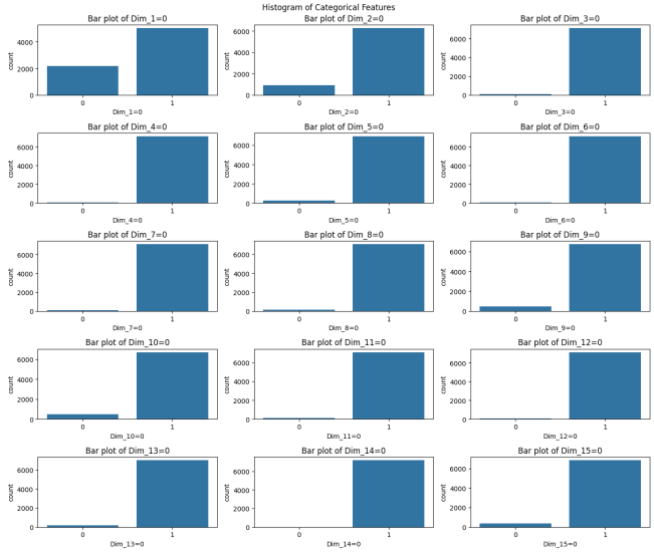

Figure 3: Histogram of Categorical Features

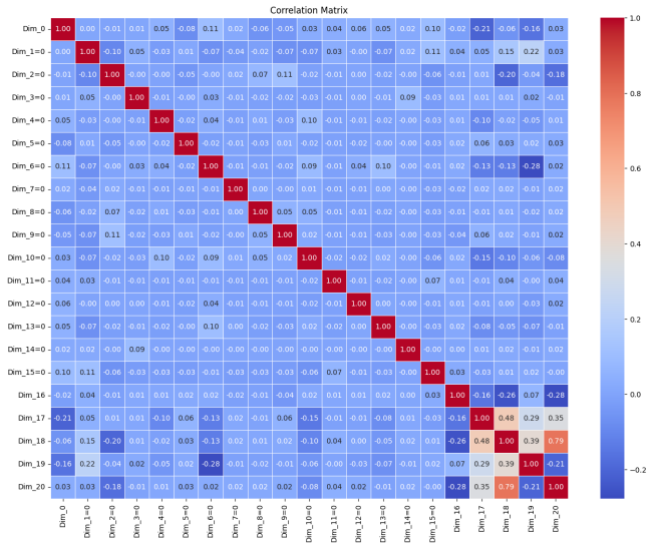

Figure 4: The Correlation Matrix of Hyperthyroidism Dataset

The correlation matrix shown in Figure 4 presents the relationships between features. Most features have low correlations, indicating weak linear relationships. Notable exceptions are Dim_18 and Dim_20, which show moderate correlations with each other and with Dim_17 and Dim_19. Since there are no correlation exceeding the score 0.8 it is decided that not to discard ant feature because of correlation.

B. Different Data Types

Initially, the data contained mixed types, with some features stored as objects and others as ‘int64’. Object-type features converted to numerical ‘float64’ type. For efficiency, continuous instances ‘int64’ were converted to ‘int8’.

C. Continuous Features

Given the non-uniform distribution of continuous features, standard scaling applied. The ‘StandardScaler’ function from scikit-learn library employed to transform these features, ensuring they have a mean of 0 and a standard deviation of 1. Since the distribution of instances was not uniform scaling was necessary.

D. Categorical Features

Categorical features were processed differently based on the algorithm requirements. For Isolation Forest method, we applied one-hot encoding using the ‘OneHotEncoder’ function from scikit-learn library. This transformation converts categorical variables into a series of binary features, facilitating their use in algorithms that require numerical input.

E. Gower Distance Matrix

To accommodate the mixed types of data in Nearest Neighbors (NN) and DBSCAN algorithms, we utilized the Gower distance metric. The Gower distance is particularly suited for mixed-type data, as it can handle both continuous and categorical variables. This distance metric computes a similarity matrix, which serves as input for our clustering and classification algorithms. Figure 5 represent computed gower distance matrix with a colorbar.

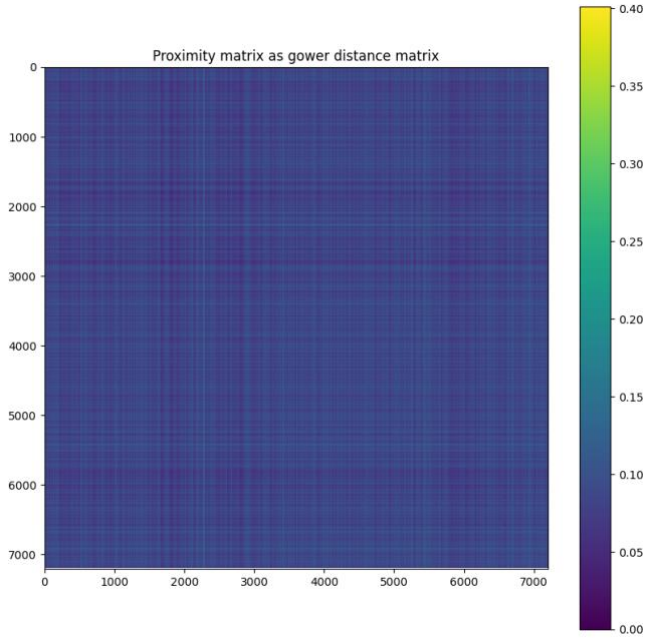

Figure 5: Gower Distance Matrix

IV. Algorithms

A. *Nearest Neighbours*

The nearest neighbors algorithm for anomaly detection works by identifying points in a dataset that deviate significantly from the rest. The process begins with the extraction of features from a dataset of normal instances during the training phase. Each instance's k-nearest neighbors are identified in the feature space, and distances to these neighbors are calculated. During the inference phase, a test instance is evaluated by measuring the distances to its nearest neighbors among the training instances. An anomaly score is derived based on these distances, with the assumption that anomalous points will lie in low-density regions of the feature space, indicating greater distances from their nearest neighbors compared to normal points. This method is robust and effective for both semi-supervised and unsupervised anomaly detection scenarios, offering significant advantages in detecting anomalies even in high-dimensional spaces (Doe, Smith, & Brown, 2023).

1) *Hyperparameter Selection with Knee Point*

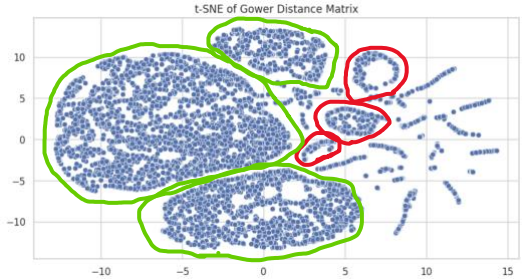

Figure 6: tSNE Plot of Gower Distance Matrix (Red Marked Areas Represents Smaller Dense Areas, Green Marked Areas Represents Larger Dense Areas)

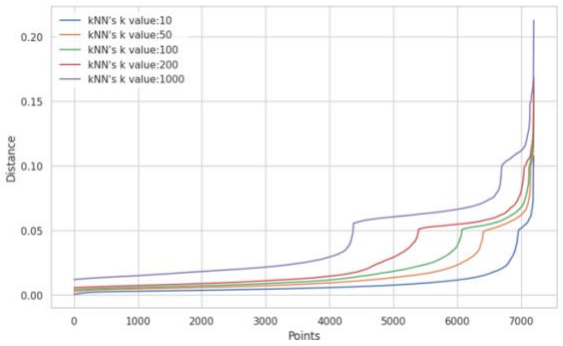

Figure 7: Distance-Points Plot of Different k Values

Figure 7 represents the distances for different k values of kNN. Figure 2 implies that there are different dense areas in the data which marked at Figure 6.

To capture small dense areas ‘n_neighbours’ parameter should be lower e.g. 10, 50. Small dense regions that can be seen in the Figure 1 are likely be considered as clusters. These areas are considered as outliers because they have at least 10 or 50 nearby points. This approach may result in many small clusters, capturing the fine-grained structure of the data but also including many minor variations.

To focus on larger dense areas ‘n_neighbours’ parameter should be higher e.g. 200, 1000. Only the largest dense regions in the Figure 6 are considered as clusters. Smaller dense areas with fewer than 1000 points are considered as outliers. This approach focuses on major clusters, treating smaller dense regions as noise, which may be useful if interested in identifying major trends and not minor variations.

The parameter ‘n_neighbours’ be selected depend on the specific goals for the clustering analysis. For this study, ‘n_neighbours’ parameter chosen as 100.

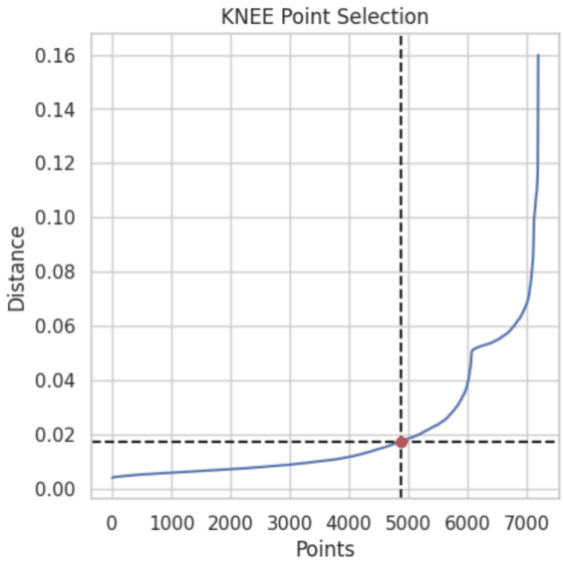

Figure 8: KNEE Point Selection

The knee point is chosen by using the KneeLocator algorithm, which identifies a point of maximum curvature on a plot of the distance values. This algorithm considers the data points as part of a convex curve that is increasing and uses polynomial interpolation to find the point where the curve bends the most sharply. This point, known as the knee,

represents the optimal balance or threshold, in this case, the best estimated epsilon (eps) value for clustering.

As Figure 8 shows, estimated best eps value is 0.02 which means that in Figure 8, the instances have distance values above 0.02 are detected as anomalies.

B. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an effective algorithm for anomaly detection due to its ability to identify clusters of varying shapes and sizes, and to handle noise effectively. It works by grouping together points that are closely packed, marking points that are isolated as noise, and distinguishing between dense regions in the data. DBSCAN requires two parameters: epsilon (the maximum distance between two points for them to be considered in the same neighborhood) and 'minPts' (the minimum number of points required to form a dense region). These properties make it particularly useful for identifying anomalies in datasets with irregular patterns and noise (Thi Lan, & Yoon, 2023).

1) Hyperparameter Selection

To use DBSCAN algorithm for the dataset, Scikit-learn library is used. The hyperparameters of the DBSCAN algorithm are 'eps', 'min_samples' and 'metric'.

'eps' parameter is selected as 0.02, found by the help of knee point.

'min_samples' parameter is selected as 100 which determined by the 'n_neighbours' at Nearest Neighbors with the help of knee point.

'metric' parameters is selected as 'precomputed' to use Gower distance matrix.

C. Isolation Forest

The Isolation Forest (iForest) algorithm is a powerful tool for anomaly detection, leveraging the principle that anomalies are more easily isolated in a dataset. By constructing an ensemble of isolation trees, where each tree randomly selects features and split values, iForest isolates data points and calculates anomaly scores based on path lengths from the root to the isolated points, with shorter paths indicating anomalies. This method is particularly effective for high-dimensional and large-scale datasets due to its scalability and linear time complexity (Xu, Zhang, Zhang, & Guo, 2022).

Isolation Forest algorithm is not suitable to work with Gower distance matrix because it requires direct access to feature values for its random partitioning strategy, which is incompatible with the pairwise distance information provided by a Gower distance matrix. Data preprocessing for the İsolation Forest made with scaling the continuous features with standard scaler and categorical features encoded with one hot encoding.

1) Hyperparameter Selection

'n_estimators' is selected as 100 a standard value for number of trees. Increasing this improves performance but at the cost of computational time.

'max_samples' is selected as 'auto' which uses the smaller value between 256 and the total number of samples, which is generally sufficient for good performance.

'contamination' is selected as 'auto' assumes ratio of anomalies. This should be adjusted based on domain knowledge.

'max_features' is selected as 1.0 considers all features for each split, which is useful given the mixed type of features in this dataset.

'random_state' is selected as 42.

V. Results and Discussion

A. Nearest Neighbors

The number of anomalies detected by Nearest Neighbors method is 1609.

B. DBSCAN

DBSCAN found 9 clusters including noise. The number of anomalies detected by DBSCAN is 1499.

1) tSNE

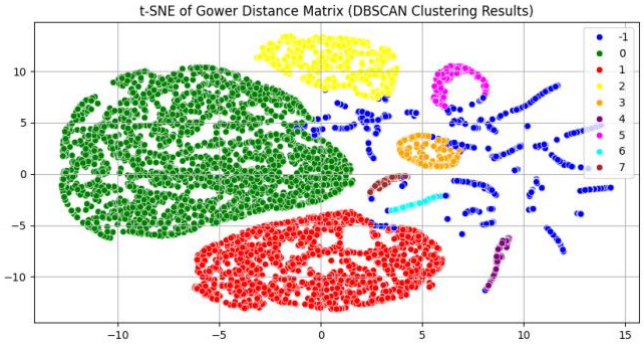

Figure 9: tSNE Visualization of DBSCAN

The tSNE visualization of DBSCAN clustering is shown at Figure 9. In Figure 9, labels '-1' represent anomalies.

DBSCAN algorithm identified several distinct clusters and noise points in the dataset. Large dense areas are captured. Separate points count as anomalies.

2) Intra Inter Cluster Distances

Table 2: Intra Distances

Cluster	*Intra-Distance*
0	1.83
1	1.65
2	1.78
3	1.81
4	1.45
5	1.66
6	1.36
7	1.34

Table 3: Inter Cluster Distances

Cluster	0	1	2	3	4	5	6	7
0	0.0	1.11	1.38	1.07	1.58	1.11	1.24	1.22
1	1.11	0.0	1.89	1.61	1.24	1.51	1.63	1.58
2	1.38	1.89	0.0	1.63	2.08	1.91	1.80	1.90
3	1.07	1.61	1.63	0.0	1.88	1.63	1.62	1.70
4	1.58	1.24	2.08	1.88	0.0	2.06	1.01	2.18
5	1.11	1.51	1.91	1.63	2.06	0.0	1.83	1.46
6	1.24	1.63	1.80	1.62	1.01	1.83	0.0	1.97
7	1.22	1.58	1.90	1.70	2.18	1.46	1.97	0.0

The Table 1 represents that intra distances in the clusters. The Table 2 represents inter cluster distances between clusters.

The intra-cluster distances indicate that most clusters are reasonably compact, with clusters 6 and 7 being the most tightly clustered. It can be said from the inter-cluster distances, the cluster distances are proper and well separated, some clusters such as clusters 0, 1, 3, and 5, are not. Suggesting potential overlap and less distinct boundaries.

3) Silhouette Score

With DBSCAN clustering's optimal number of clusters equal to 9, the silhouette score is found 0.421. Since the silhouette score is greater than 0 but not close to 1, it indicates moderate clustering performance.

4) Correlation Between Ideal Similarity Matrix and Proximity Matrix

The cross-correlation between the ideal similarity matrix and the proximity matrix 0.47. After removing 1499 outliers (i.e., 20.8% of the entire dataset), cross-correlation result is obtained as 0.80. As seen, when noise points are removed there is a significant increase in cross-correlation. This supports that noise points are found by DBSCAN tend to be outliers.

C. Isolation Forest

The number of anomalies detected by Isolation Forest method is 360.

As seen in Figure 10, in the categorical attributes plot points (1, 1) seen as normal, point (0, 0) seen as anomaly. The points (1,0) and (0,1) are uncertain. This results are expected because as can be seen in the histogram of the data, the values of categorical variables were mostly 1 and 0's were sparse. For continuous attributes, Isolation Forest identifies the points that are far from dense areas as anomalies.

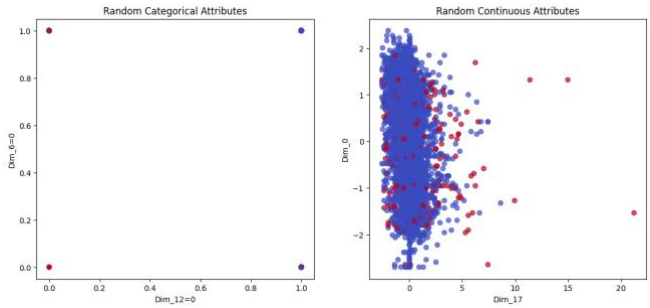

Figure 10: Scatterplot of Random Two Attributes (Anomalies are Red, Normal are Blue)

D. Compare The Methods

In Figure 11, overlapping anomalies found by three different methods marked with points in first 100 samples.

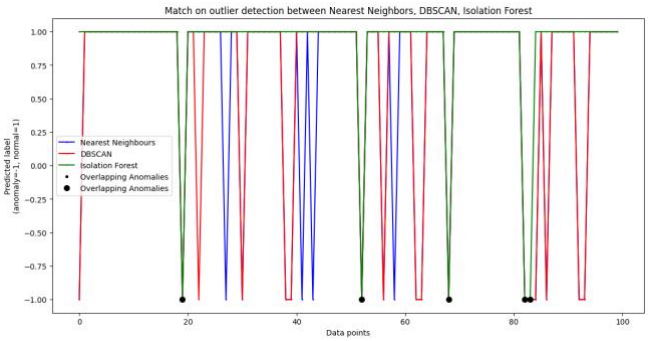

Figure 11: Overlapping Anomalies in First 100 Instances

In total 357 anomalies overlapped. Since the anomalies detected by Isolation Forest overlaps almost 100% with the other methods, it is chosen as best anomaly detection algorithm for the given dataset.

The anomalies found by Nearest Neighbors and DBSCAN methods are 92% same and the anomalies at least found by two of the methods %99.8 same as overlapped anomalies of Nearest Neighbors and DBSCAN, it is proved that two methods verifies each other and the idea of Isolation Forest is the best method supported.

The Anomaly Verification Metrics:

- Homogeneity: Measure of how much each cluster contains only data points that are members of a single class.
- Completeness: Measure of how well all members of a given class are assigned to the same cluster.
- V-measure: The harmonic mean of homogeneity and completeness.
- Adjusted Rand Index: It measures the similarity between the two clustering by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clustering
- Adjusted Mutual Information: It measures the agreement of the two assignments, ignoring permutations and with chance normalization.

The Scores of NN vs DBSCAN Anomalies:

- Homogeneity: 0.655
- Completeness: 0.680
- V-measure: 0.668
- Adjusted Rand Index: 0.806
- Adjusted Mutual Information: 0.667

These results suggests that while there are some differences between two algorithms in detecting anomalies, they generally agree. Especially, the higher ARI score indicates a strong alignment of anomaly results.

These metrics are just used for DBSCAN and NN because there are significant difference in the number of anomalies found by Isolation Forest and the other methods.

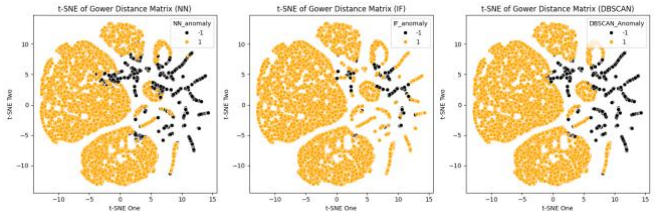

Figure 12: tSNE Plot

Accoring to Figure 12, the anomalies detected by Isolation Forest are also detected by the other two methods. In addition, the results of NN and DBSCAN seem quite similar.

REFERENCES

[1] Lan, D. T., & Yoon, S. (2023). Trajectory Clustering-Based anomaly detection in indoor human movement. *Sensors*, *23*(6), 3318. https://doi.org/10.3390/s23063318

[2] Nizan, O., & Tal, A. (2023, May 28). *K-NNN: Nearest neighbors of neighbors for anomaly detection*. arXiv.org. https://arxiv.org/abs/2305.17695

[3] Xu, H., Pang, G., Wang, Y., & Wang, Y. (2022). Deep isolation forest for anomaly detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2206.06602