Anomaly Detection in Hypothyroidism Dataset Using DBSCAN and LOF

Oktay Kurt
Department of Artificial Intelligence for Science and Technology
University of Milano-Bicocca
Email: o.kurt@campus.unimib.it

***Abstract*—This report presents an in-depth analysis of a dataset related to hypothyroidism, containing 21 attributes (15 binary and 6 continuous) and 7,200 data objects. The analysis involves data preprocessing, visualization, and anomaly detection using DBSCAN, Local Outlier Factor (LOF), and a combined method of DBSCAN and LOF. The results are evaluated using various clustering performance metrics, and visualizations are provided to illustrate the findings.**

GITHUB REPOSITORY

The code for this project is available at: https://github.com/oktaykurt/Anomaly-Detection-with-DBSCAN-and-LOF

***Index Terms*—Anomaly Detection, DBSCAN, Local Outlier Factor, Hypothyroidism, Gower Distance, t-SNE**

I. INTRODUCTION

Hypothyroidism is a medical condition characterized by an underactive thyroid gland. Detecting anomalies in medical datasets is crucial for identifying unusual patterns that could indicate potential health issues. This study employs unsupervised learning techniques to detect anomalies in a hypothyroidism dataset, focusing on the identification of outliers that may represent misdiagnosed cases or rare variations of the disease.

II. METHODOLOGY

The methodology section encompasses data preprocessing, visualization, and the implementation of clustering algorithms to detect anomalies. The techniques used include Gower distance for handling mixed data types, t-SNE for dimensionality reduction, and evaluation metrics like the silhouette score, Davies-Bouldin index, and Calinski-Harabasz index for assessing clustering quality.

A. Data Preprocessing

Data preprocessing is a critical step to ensure the dataset is clean and suitable for analysis. The steps include:

1) **Loading the Dataset:** The dataset is loaded from a CSV file with attributes separated by semicolons and decimals marked by commas.
2) **Dropping Unnecessary Columns:** The last two columns, which are empty due to the way the CSV file is read, are dropped as they are not required for analysis.
3) **Identifying and Converting Column Types:** Binary columns are identified and converted to integer type, while continuous columns are converted to float type. The 'Row' column, which serves as an index, is also dropped.
4) **Handling Missing Values:** The dataset is checked for missing values. Fortunately, no missing values were found, ensuring a complete dataset for analysis.

Normalization and Scaling:

- **Gower Distance:** Gower distance, implemented from sklearn, applies max-min scaling to continuous values. Therefore, no additional normalization was applied to the continuous features. After discussing and considering standard scaling for continuous features, it was decided not to use it as it could reduce the separation between normal and anomaly points.
- **Binary Features:** For binary features, which only contain 0 and 1, no normalization or hot encoding was applied. The binary labels are inherently ready for analysis.

B. Visual Exploration

Visual exploration helps in understanding the distribution and relationships of the features within the dataset:

1) **Histograms of Continuous Features:** These provide a visual summary of the distribution of continuous features, helping identify patterns such as normality, skewness, and the presence of multiple modes.

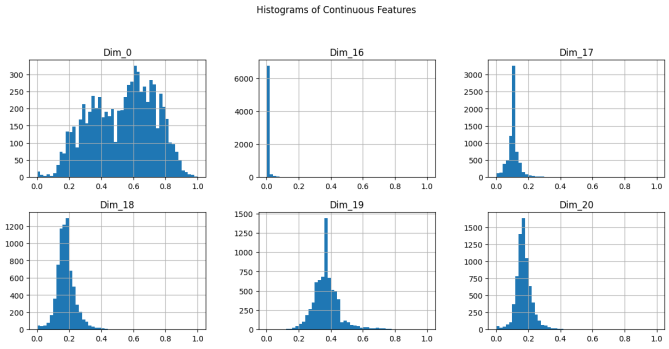

Fig. 1. Histograms of Continuous Features

2) **Box Plots of Continuous Features:** Box plots highlight the spread and skewness of continuous data and identify potential outliers.

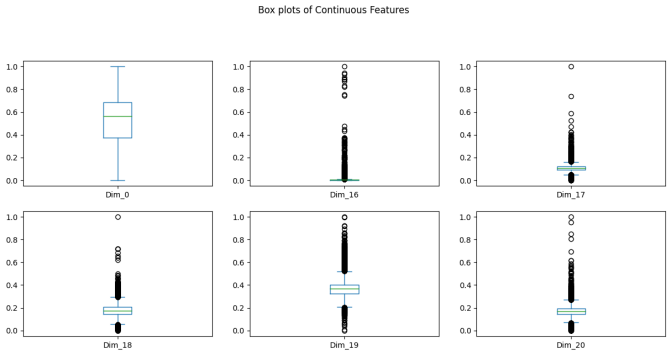

Fig. 2. Box Plots of Continuous Features

3) **Bar Plots of Binary Features:** These plots show the frequency of binary attributes, indicating the balance or imbalance between different classes. The binary columns are dominated by value of "1".

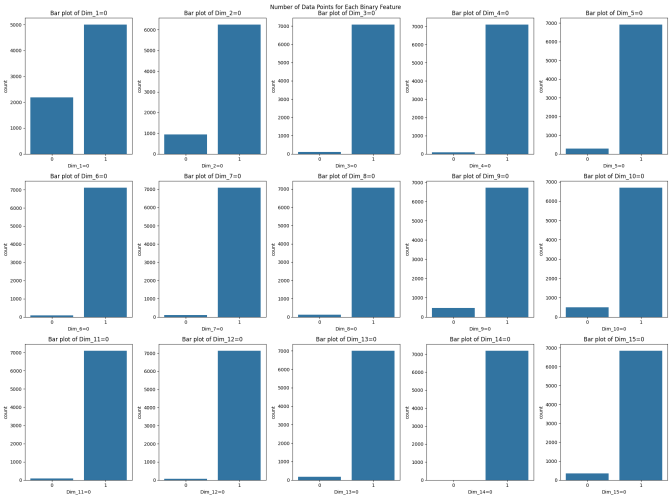

Fig. 3. Bar Plots of Binary Features

4) **Correlation Matrix Heatmap:** The heatmap visualizes the correlation between features, indicating which features are positively or negatively correlated. Notably, there are no significant correlations between the majority of the features, with the exception of Dim18 and Dim20.To ensure the reduction of potential anomalous points, neither Dim18 nor Dim20 has been excluded. This approach preserves the integrity and robustness of the dataset.

C. Anomaly Detection

Anomaly detection is performed using DBSCAN, LOF, and a combined method of DBSCAN and LOF, with Gower distance utilized to handle mixed data types (both binary and continuous).

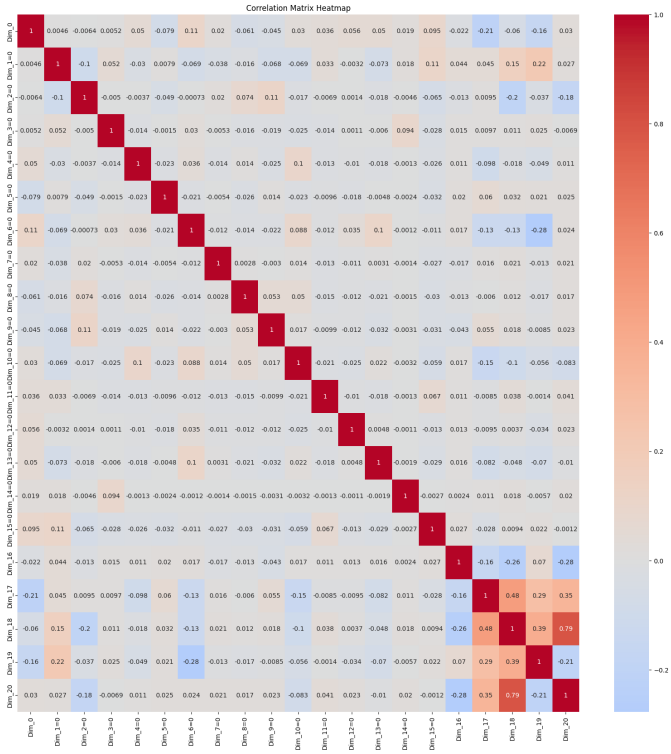

Fig. 4. Correlation Matrix Heatmap

Gower Distance: Gower distance is a metric designed to handle mixed data types. It computes the distance between each pair of samples by considering binary and continuous attributes separately, then combining them into a single distance metric. This is particularly useful for medical datasets with diverse data types.

t-SNE for Dimensionality Reduction: t-SNE (t-distributed Stochastic Neighbor Embedding) is used to reduce the high-dimensional Gower distance matrix to two dimensions for visualization. t-SNE emphasizes preserving local structures, making it ideal for visualizing clusters in complex datasets.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN is a clustering algorithm that groups points closely packed together, marking points in low-density regions as outliers. It requires two parameters: *eps* (the maximum distance between two samples for them to be considered as in the same neighborhood) and *min_samples* (the number of samples in a neighborhood for a point to be considered a core point).

Random Search for Best Parameters: A random search is conducted over 100 iterations to find the best parameters for DBSCAN. The parameters *eps* and *min_samples* are varied within the following intervals:

- *eps = random.uniform(0.01, 0.08)*
- *min_samples = random.randint(4, 40)*

Evaluation of DBSCAN Clustering:

- **Silhouette Score:** Indicates that the clusters are

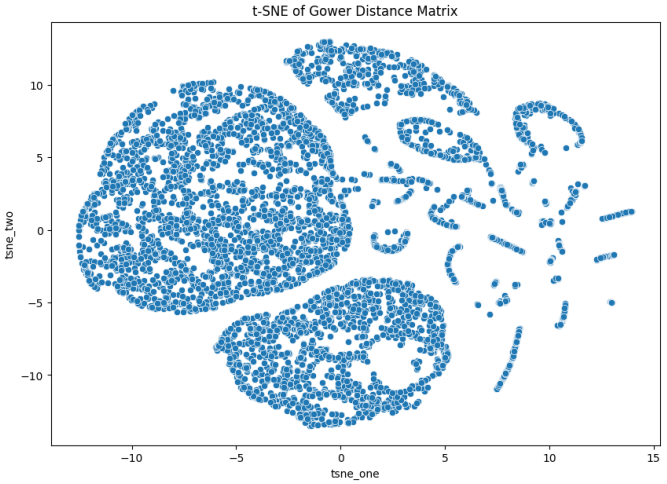

Fig. 5. t-SNE of Gower Distance Matrix

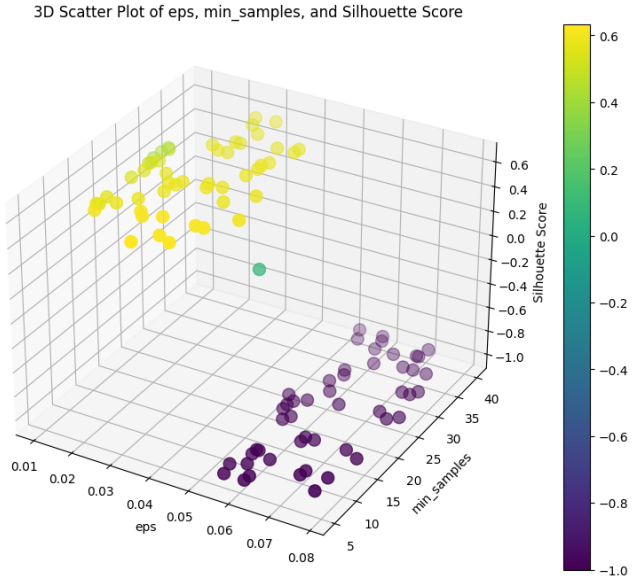

Fig. 6. 3D Scatter Plot of eps, min_samples, and Silhouette Score

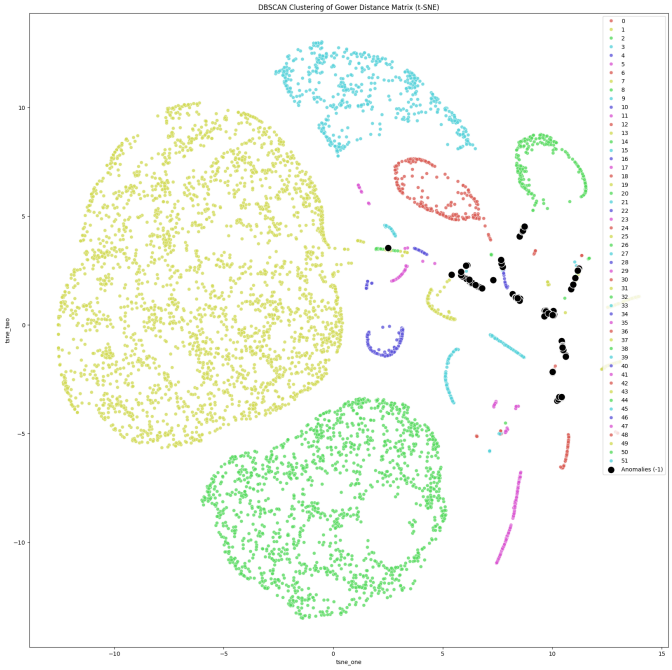

Fig. 7. DBSCAN Clustering (t-SNE)

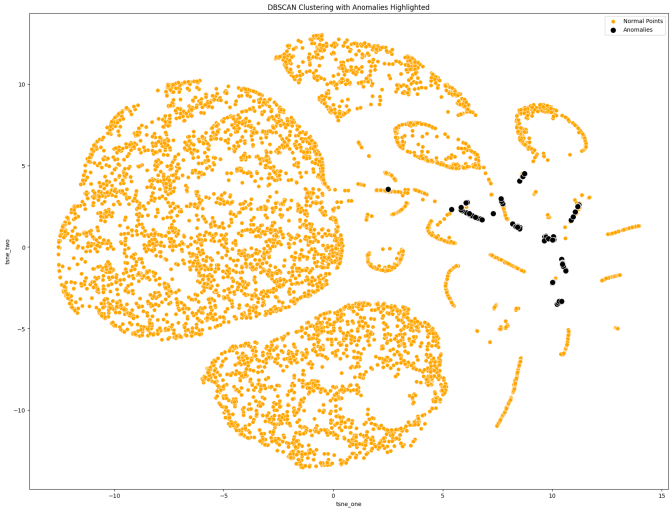

Fig. 8. DBSCAN Clustering with Anomalies Highlighted (t-SNE)

well-defined with a score of *0.6344*.

- **Davies-Bouldin Index:** A low value of *1.0238* suggests that the clusters are compact and well-separated.
- **Calinski-Harabasz Index:** A high value of *1469.37* confirms the clustering quality.

Top 5 Parameters for DBSCAN:

TABLE I
TOP 5 PARAMETERS FOR DBSCAN BASED ON SILHOUETTE SCORE

Rank	eps	min_samples	Silhouette Score	DB Index	CH Index
1	0.0337	5	0.6344	1.0238	1469.37
2	0.0407	7	0.6310	1.0751	1542.19
3	0.0371	8	0.6255	1.0675	1632.31
4	0.0433	12	0.6181	1.0619	1712.37
5	0.0413	12	0.6181	1.0619	1712.37

D. LOF Clustering

LOF Implementation: LOF is applied with multiple iterations (250 in this case) to identify robust anomalies. The threshold for considering a point as an anomaly is set at 90% of the iterations. Specifically, a data point is considered an anomaly if it is identified as an anomaly in at least 225 out of 250 iterations. The parameter ranges used in the random search are:

- *n_neighbors = random.randint(3, 50)*
- *contamination = random.uniform(0.005, 0.4)*

Evaluation of LOF Clustering: LOF effectively identifies anomalies by comparing the density of a point to that of its

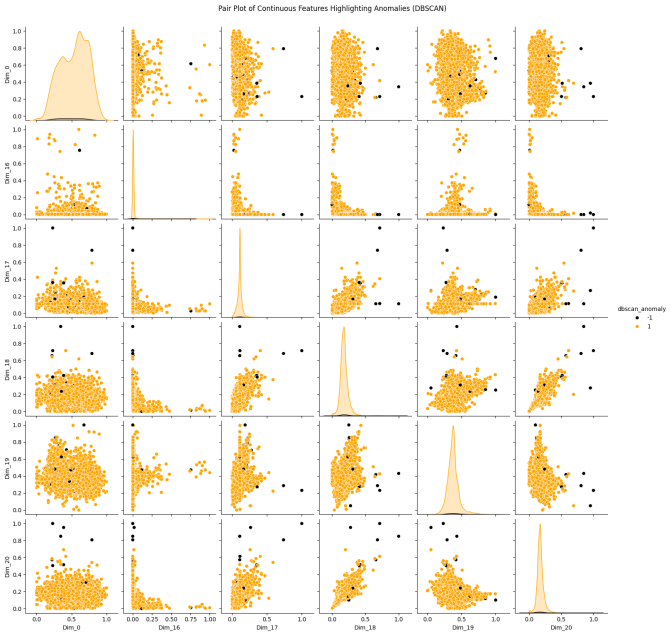

Fig. 9. DBSCAN Pair Plot of Continuous Features Highlighting Anomalies

neighbors. Points with significantly lower density are marked as outliers.

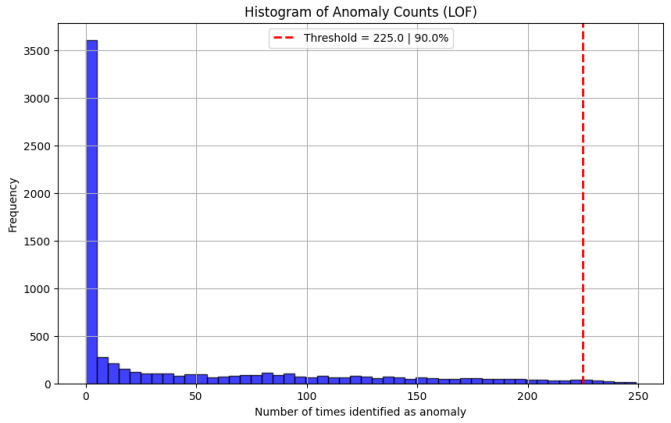

Fig. 10. Histogram of Anomaly Counts (LOF)

Analysis of LOF Clustering Results: In the t-SNE visualization of LOF clustering, some points appear to be mislabeled as anomalies despite being located near large clusters. However, in general, the LOF model successfully captures the anomalies, as indicated by the separation of outliers from the main data points in the pair plot and clustering visualizations.

Comparison of Outliers Detected by DBSCAN and LOF: The plot below shows the anomalies detected by both methods, allowing for a direct comparison. Areas of agreement and discrepancy are highlighted, providing insights into the robustness of the anomaly detection.

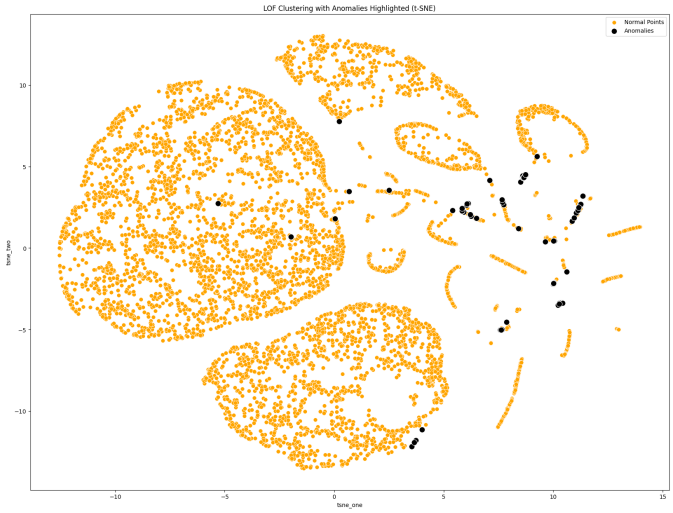

Fig. 11. LOF Clustering with Anomalies Highlighted (t-SNE)

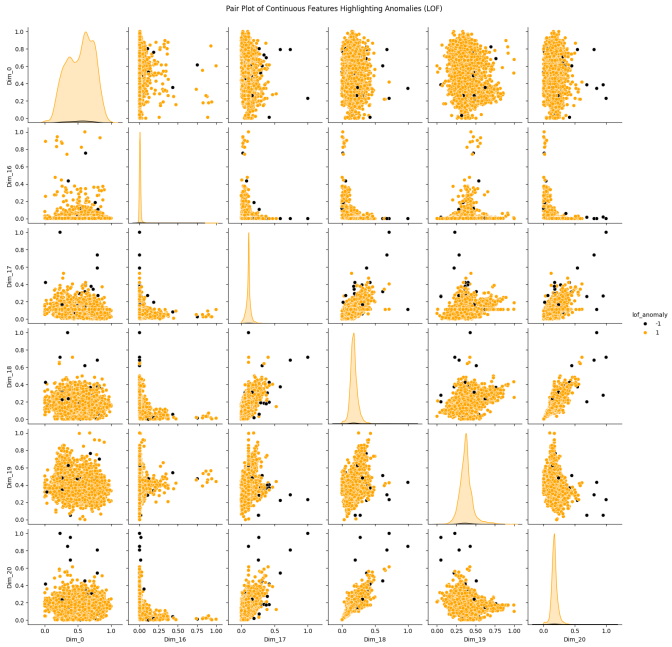

Fig. 12. LOF Pair Plot of Continuous Features Highlighting Anomalies

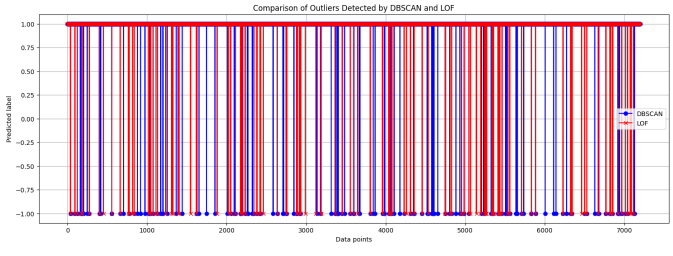

Fig. 14. Comparison of Outliers Detected by DBSCAN and LOF

Number of Anomalies Detected by Each Technique: The bar plot below shows the number of anomalies detected by DBSCAN and LOF. DBSCAN detected 158 anomalies,

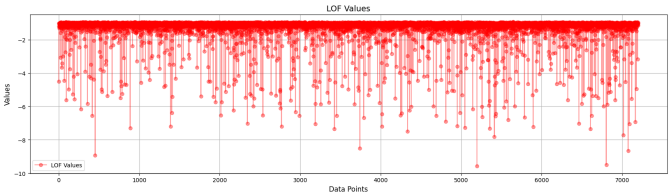

Fig. 13. LOF Values

while LOF detected 117 anomalies.

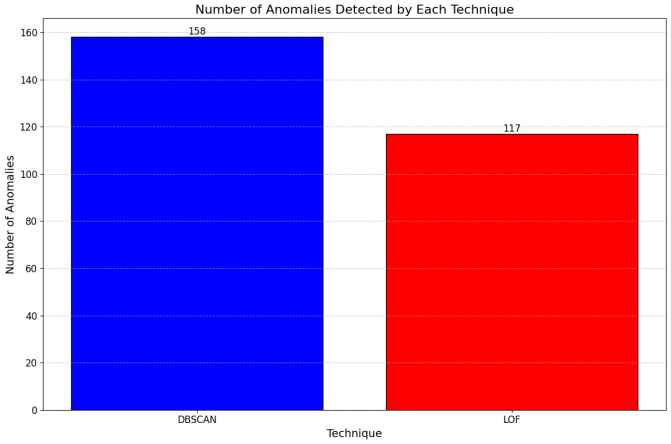

Fig. 15. Number of Anomalies Detected by Each Technique

Adjusted Rand Index: The adjusted Rand index between DBSCAN and LOF is **0.5048**. This index measures the agreement between the two clustering methods in identifying anomalies. A value of 0.5048 indicates moderate agreement, suggesting that while there is some overlap in the anomalies detected by both methods, each method also identifies unique outliers. This reinforces the value of using multiple techniques for a comprehensive anomaly detection strategy.

E. Combined DBSCAN and LOF Method

Random Search for Best Combined Parameters: A random search over 500 iterations was performed to find the optimal parameters for DBSCAN and LOF, focusing on the intersection of outliers detected by both methods. The best parameters were found to be *eps=0.0499*, *min_samples=13* for DBSCAN, and *n_neighbors=14*, *contamination=0.0395* for LOF.

Evaluation Metrics: The combined approach achieved an adjusted Rand index of *0.6322*, indicating a high level of agreement between DBSCAN and LOF in identifying outliers.

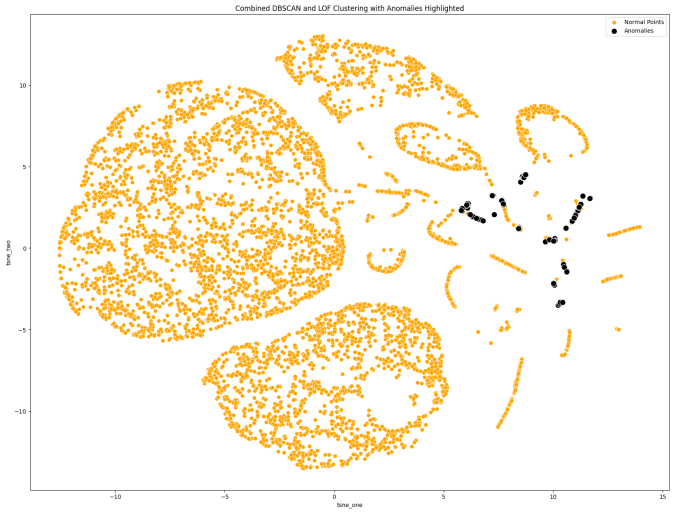

Fig. 16. Combined DBSCAN and LOF Clustering with Anomalies Highlighted (t-SNE)

Fig. 17. Combined DBSCAN and LOF Pair Plot of Continuous Features Highlighting Anomalies

III. Discussion

The analysis demonstrates the effectiveness of DBSCAN and LOF in detecting anomalies in the hypothyroidism dataset. Both methods show good performance as indicated by the evaluation metrics. The combined approach further enhances anomaly detection by leveraging the strengths of both methods. However, visual interpretations from the pair plot of continuous features and the clustering with anomalies highlighted in the t-SNE graph indicate that LOF performs better in separating anomalies from normal data points.

IV. Conclusion

This study successfully applies DBSCAN, LOF, and a combined approach for anomaly detection in a hypothyroidism dataset. The visualizations and performance metrics indicate that these methods are effective in identifying unusual patterns. Based on the visual interpretation, LOF was selected as the preferred method for anomaly detection due to its better performance in separating anomalies from normal data points. Future work could involve exploring other clustering algorithms and improving parameter optimization techniques.

V. References

References

[1] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[2] L. Van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.

[3] J. C. Gower, "A general coefficient of similarity and some of its properties," Biometrics, vol. 27, pp. 857-874, 1971.