Anomaly Detection for Hyperthyroidism

Michele Banfi
869294
Milano-Bicocca
m.banfi3397@campus.unimib.it

Annalisa Di Pasquali
858848
Milano-Bicocca
a.dipasquali@campus.unimib.it

***Abstract*— This paper explores the application of multiple anomaly detection algorithms on a given mixed-type dataset, containing both categorical and continuous variables, to identify and classify anomalies without prior knowledge of their nature. Applying these types of algorithms can speed up the disease recognition process and opens up the possibility for earlier medical intervention enhancing the general efficiency of medical treatments.**

The results demonstrate the comparative advantages and limitations of each method, providing insights into their applicability for mixed-type medical datasets.

I. Introduction

Anomaly detection algorithms are a powerful tools to find observations that deviate significantly from the majority of the data and do not conform to a well defined notion of normal behavior.

Four distinct methods are utilized: Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Isolation Forest, Principal Component Analysis (PCA), and the K-Nearest Neighbor (KNN).

The structure of the paper is as follows: in Section II the dataset is presented, providing a detailed description and analyzing the mixed-type nature of the data through box plots and histograms. Section III delves into the metrics used to handle mixed data types: Gower's metric. Section IV covers the implementation of the anomaly detection algorithms and in section V the results of these algorithms are reported. Section VI discusses the overall results, incorporating t-SNE for visualization. Finally, Section VII provides the conclusions of the study.
Notice that the outlier percentage is set at 5% of the overall dataset to ensure a consistent basis for comparison for all algorithms used.

Keywords: Mixed-type Dataset, Anomaly Detection, Gower Distance, DBSCAN, Isolation Forest, PCA, Knee Method, T-SNE.

II. Dataset Description

The given dataset consists of 7200 observations described by 21 features, both continuous and categorical attributes. Preprocessing involves the removal of the last two columns to ensure data integrity and relevance to the anomaly detection task. In the context of this study, the absence of information regarding the origin of the dataset, precluded the execution of feature selection procedures, in this way the preservation of all available attributes to maintain data integrity and minimize the risk of information loss is ensured.

Table 1 presents the head of the dataset, showing the firsts of rows and a subset of the columns. This sample helps illustrate the nature and distribution of the data.

Dim_0	Dim_1=0	Dim_2=0	...	Dim_20
0.75	1.0	0.0	...	0.225
0.239583	1.0	1.0	...	0.165625
0.479167	1.0	1.0	...	0.11875
0.65625	0.0	1.0	...	0.129688
0.229167	1.0	1.0	...	0.235938

Table 1: Dataset head

For binary columns identified by the suffix '=0', the count of each unique value is examined to assess the balance and the distributions of these binary attributes. Below, histograms illustrating the distribution of values are reported:

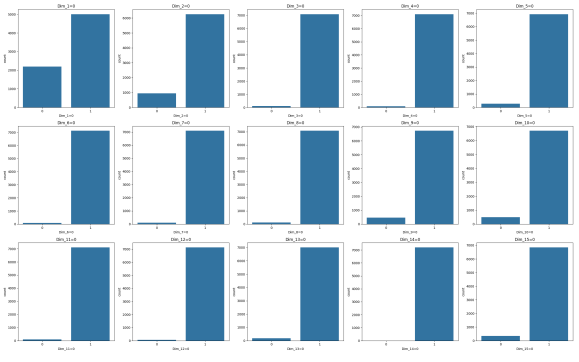

Figure 1: **Histograms for categorical features**: binary columns that describe the presence of the categories `0` and `1` in the categorical features

Most of the attributes present unbalanced frequencies for the binary classes. for example:

- `Dim_2`: the category `0` is significantly less frequent than `1`.
- `Dim_4`: the category `0` has a much lower count compared to `1`.
- `Dim_9`: the category `0` shows a low frequency compared to `1`
- `Dim_14`: the category `0` is notably less frequent.

Summary statistics such as mean, standard deviation, minimum, maximum, and quartile values are computed for continuos attributes to understand the central tendency and spread of the data. The key components of a box plot include the interquartile range (IQR), whiskers extending to 1.5 times the IQR, and data points beyond these whiskers, which are generally considered outliers. The results are reported in the following box plots:

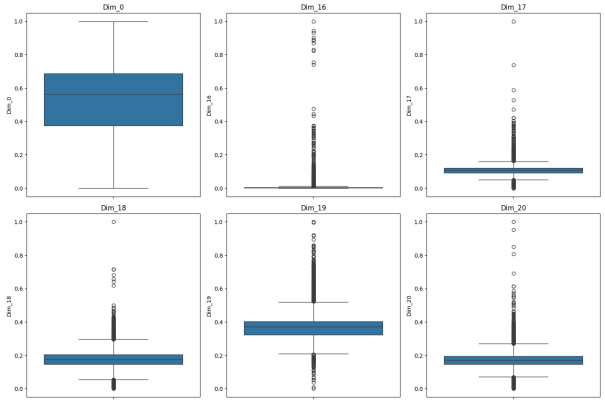

Figure 2: **Box Plots**: Summary statistics for continuous features

Histograms provide additional insight into the distribution of data points.

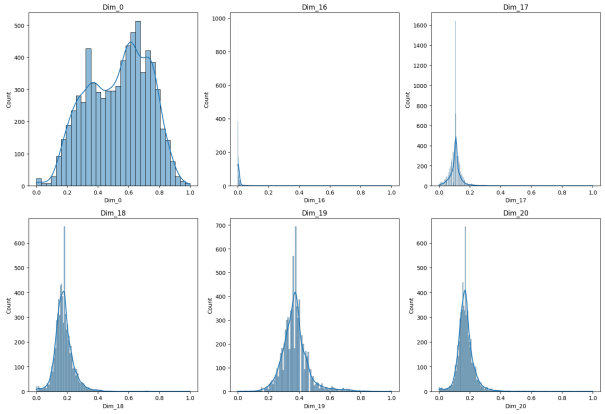

Figure 3: **Distribution**: distributions for continuous features

The combination of box plots and histograms with distribution curves provides a comprehensive view of outliers in the continuous variables, in particular for:

- `Dim_0`: no potential outliers detected (the distribution appears bimodal with no significant outliers).
- `Dim_16`: multiple potential outliers identified above the upper whisker (highly skewed distribution with a peak near 0, indicating many potential outliers).
- `Dim_17`: several potential outliers detected above the upper whisker (skewed distribution with a concentration near 0, supporting the presence of outliers).
- `Dim_18`: potential outliers found both above the upper whisker and below the lower whisker (skewed distribution with a peak near 0.2, indicating potential outliers).
- `Dim_19`: numerous potential outliers observed above the upper whisker and below the lower whisker (skewed distribution with a concentration around 0.4, highlighting potential outliers).
- `Dim_20`: multiple potential outliers noted above the upper whisker (skewed distribution with a peak near 0.2, indicating potential outliers).

Even if considering each attribute at a time could be helpful to identify potential outliers, it may not capture the full complexity of the data. Therefore, all dimensions must be taken into account simultaneously to effectively detect anomalies. To address the multidimensionality and consider the entire dataset's structure, advanced anomaly detection techniques are implemented: these methods allow the identification of anomalies that may not be evident when examining individual attributes separately.

The subsequent sections detail the approach to handling mixed data types and the implementation of the chosen anomaly detection algorithms.

III. Preprocessing

The dataset presents challenges due to the presence of mixed data types, comprising both categorical and continuous attributes. Since we want to detect anomalies in the dataset and the algorithms need a distance metric, normal distances can't be employed like Euclidean or Minkowski (they are used with continuos variables only) so in the following paragraph Gower's distance is introduced.

A. *Gower's distance*

Unlike traditional distance metrics which assume homogeneity of variable types, the main peculiarity of Gower's distance is the sensitivity to the variable types present in the dataset. It recognizes that different types of variables require distinct measures of similarity or dissimilarity. For example, Gower's distance utilizes appropriate measures such as

Manhattan distance for continuous variables, and matching coefficients or Jaccard's coefficients for categorical variables. Hence, by quantifying the dissimilarities between observations, Gower's distance enables the detection of anomalous instances also across an heterogeneous dataset.
Below is reported the mathematical formula to calculate Gower's Distance:

$$d_{ij} = \frac{\sum_{k=1}^{v} d_{ijk}\delta_{ijk}}{\sum_{k=1}^{v} \delta_{ijk}} \quad (1)$$

where:

- v is the total number of variables;
- δ_{ijk} is a binary indicator that equals 1 if neither observation x_i nor observation x_j has missing values for variable k, and 0 otherwise;
- d_{ijk} is the distance between i and j with respect to the variable k.

The distance d_{ijk} is calculated differently based on the variable type:

- Numerical variables: $d_{ijk} = \frac{|x_{ik}-x_{jk}|}{R_k}$
  where: x_{ik} and x_{jk} are the values of variable k for observations i and j and R_k is the range of variable k.
- Categorical variables : $\begin{cases} 0 & \text{if } x_{ik}=x_{jt} \\ 1 & \text{if } x_{ik}\neq x_{jt} \end{cases}$ where: x_{ik} and x_{jt} are the categories for variable k of observation i and j.

The overall distance d_{ij} is a weighted average of the normalized distances d_{ijk} for all the variables, with weights based on data availability, in the formula above represented by δ_{ijk}; this factor describes the possibility of making comparisons, hence $\delta_{ijk} = 1$ when variable k can be compared for i and j, and 0 otherwise. When $\delta_{ijk} = 0$ for all attributes, d_{ij} is undefined, while when all comparisons are possible $\sum_{k=1}^{v} \delta_{ijk} = v$, the total number of attributes.
For more details on the Gower distance, see Gower's original paper. [1]

By using the Gower distance, the mixed-type nature of the dataset is adequately addressed, allowing for a more accurate and meaningful anomaly detection process.

The next sections covers the implementation of the selected anomaly detection algorithms, leveraging this preprocessing step.

IV. Anomaly Detection

There are different families of anomaly detection algorithms. For example density-based methods, such as DBSCAN, identify anomalies by examining the density of data points, flagging those in low-density regions as potential outliers. Proximity-based methods like k-nearest neighbors, detect anomalies by measuring the distance between data points and their neighbors, identifying those that are far from others. Statistical methods use statistical models to determine the probability of data points, labeling those with low probabilities as anomalies. Informational theoretical methods utilize concepts from information theory, such as entropy and mutual information, to identify anomalies by detecting irregularities in the data distribution.
Each of these methods offers unique advantages but they may also struggle in some cases: the type of data they are dealing with and the task they are asked to solve play a crucial role. Understanding these factors is essential for selecting and implementing the most appropriate anomaly detection technique for a given dataset and problem. In this study DBSCAN, Isolation Forest, PCA and KNN are chosen to be the best algorithms to be implemented, taking into account the structure of the dataset.

A. DBSCAN [2]

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that is particularly effective for identifying clusters of varying shapes and sizes but also detecting outliers. The DBSCAN method is implemented using the Gower's distance matrix and it requires two parameters:

- `eps = 0.04` → the maximum distance between two samples for one to be considered as in the neighborhood of the other;
- `min_samples = 10` → the number of samples (or total weight) in a neighborhood for a point to be considered as a core point[1].

These parameters were initially determined through visual inspection using the knee method and the calculation of the silhouette score (the knee method will be presented afterwards). Subsequently, they were adjusted to conform with the expected proportion of outliers in the dataset.

B. Isolation Forest [3]

Isolation Forest is particularly effective for anomaly detection tasks in high dimensional datasets; unlike clustering-based approaches, Isolation Forest specifically targets anomalies by constructing an ensemble of random trees, following the principle that anomalies, being few and different, are more likely to be isolated earlier in the process than normal points.

In particular, a subset of the dataset is sampled to *build trees in the forest*, for each tree then the data is recursively partitioned by selecting random features and a random split value between minimum and maximum values of the selected feature. The process iterates until each observation is isolated: anomalies are isolated more quickly and therefore

[1] A core point is a point that has at least `min_samples` points within its `eps` radius (including itself). Core points are central to the clusters.

have shorter path lengths at the end of the iterations.
The following parameters are used in the algorithm:

- `contamination = 0.05` → The amount of contamination of the data set, i.e. the proportion of outliers in the dataset;
- `n_estimatorsint = 100` → default value of based estimators (tress) in the ensemble (forest);
- `max_features = 1.0` → default number of features to be drawn to train each base estimator.

Also in this case, the contamination parameter is set according to the proportion of outliers in the dataset.

C. PCA [4]

Principal component analysis (PCA) is a linear dimensionality reduction technique. The data is linearly transformed in a new coordinate system such that the directions (principal components) capturing the largest variation in the data can be easily identified. In particular, the first principal component captures the maximum variance in the data, the second principal component (orthogonal to the first) captures the maximum remaining variance, and so on. The new coordinates are linear combinations of the original variables.

PCA algorithm is based on the computation of the covariance matrix. This matrix quantifies the pairwise variability between variables, then the principal components are derived from the eigenvectors of this matrix and the corresponding eigenvalues indicate the amount of variance captured by each principal component. Subsequently the eigenvalues are ranked in descending order and among them the top q (where q will be the number of principal component) are selected.

In order to detect the anomalies each data object is projected into the lower dimensional space; y_i denotes the $i\text{-}th$ component and it has the following χ^2 distribution:

$$\sum_{i}^{q} \frac{y_i^2}{\lambda_i} \quad q < n \tag{2}$$

Then a data object is anomalous, if for a given significance level α

$$\sum_{i}^{q} \frac{y_i^2}{\lambda_i} > \chi_q^2(\alpha) \tag{3}$$

PCA is performed with the following hyper-parameter:

- `n_components = 10` → number of components to keep.
- `alpha = 0.05` → a threshold is set equal to the $95th$ percentile of the χ^2.

D. KNN

The KNN is a technique that belongs to the family of distance-based anomaly detection algorithm, which focus on identifying anomalies by measuring the distances between data and using these distances to detect points that deviate significantly from the norm. With the implementation of the Knee Locator method a knee point is computed and represents the threshold where distances between an observation and its $n\text{-}th$ (n is an hyper-parameter) neighbor star to increase, suggesting the presence of outliers for distance's values larger than this point.

As the previous case KNN is performed using the Gower's distance. The cutoff value is computed and then adjusted basing on visual inspection of the knee point to ensure accurate identification of anomalies, in particular its value is set equal to `1.2`

The other parameter used is:

- `n_neighbors = 10` → this is the parameter that the KNN use and it is the number of neighbors;

Each of the four methods employed in this analysis produce a binary output consisting of two labels:

- `0` for inlying observation, indicating that an observation is considered normal;
- `1` for outlying observation, indicating that an observation is identified as anomaly.

The use of these labels provides a clear categorization of the data points and allows for a comparison between the used algorithms, which will be the topic of the next section.

V. Results of the algorithms

In this section is presented the percentage of outliers detected by each of the four methods applied in this study, the following results summarize the performance of each method, even in terms of their sensitivity in detecting anomalies within the given dataset.

Method	Num. Outliers	Percentage
DBSCAN	241	3%
Isolation Forest	288	4%
PCA	443	6%
KNN	369	5%

Table 2: Comparison of the Results for each Outliers Detection Methods

The reported results highlight various numbers of outliers detected by each method due to the their different nature. In particular:

- DBSCAN detects the fewest outliers percentage (with 39 clusters produced). One reason could be the fact that it is a density based approach, so it is sensitive to the choice of parameters used such as epsilon and the minimum number of samples.

- Isolation Forest identify a percentage of outliers equal to 4%, especially, it is less sensitive to parameter choices compared to DBSCAN.
- PCA detected the highest percentage of outliers (6%): this result reflects the fact that the algorithm captures anomalies that deviate from the principal components, so it is more sensitive to variation in the data.
- KNN detected the expected number of anomalies equal to the 5% of the data. The handcrafting of the threshold value needs to be taken into account in the goodness of the results.

VI. Results

The results discussed above allow to assign a probability to each observation of being an outlier. This is achieved by computing the module of the average of the labels generated by each of the four methods:

$$p_i = \left| \frac{\sum_{k=1}^{n} l_k}{n} \right| \tag{4}$$

where n is equal to four (number of methods applied), and l_k is the label assigned by the k-th method to the i-*th* observation .

In order to better visualize such results t-Distributed Stochastic Neighbor Embedding (t-SNE) is used.

t-SNE is a dimensionality reduction technique that visualizes high-dimensional data by giving each data point a location in lower dimensional space with respect to the original one. t-SNE is capable of capturing much of the local structure of the high-dimensional data and also revealing global structure such as the presence of clusters at several scales that may be hidden in the higher-dimensional space.

For more details see [5]

In particular in Figure 4 the results are reported as follows: in yellow observations classified as non-outliers, in orange observation classified as outliers by one method only, in pink observations classified as outliers by two methods, in purple observation classified as outliers by three methods and in blue observations classified as outliers by all four methods.

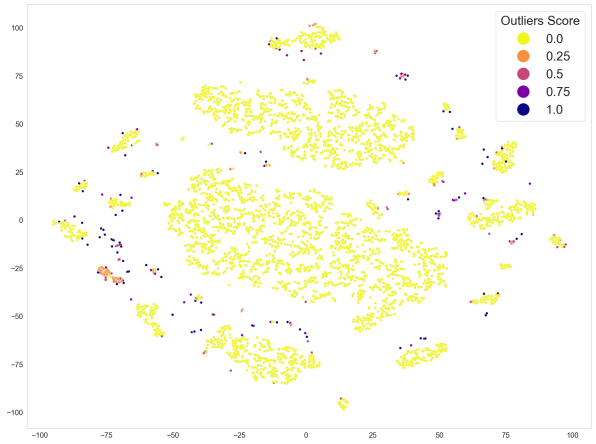

Figure 4: **t-SNE**: plot of the dataset with outlier percentage computed for each observation

The Rand score between different outlier detection methods is calculated, resulting in a diagonal matrix. This matrix indicates a high level of agreement across all the methods, with a Rand score above 0.9 for all algorithms. Specifically, the Rand score is calculated as follows:

$$\text{Rand Score} = \frac{\text{number of agreeing pairs}}{\text{number of pairs}} \tag{5}$$

and in the case of this study unbalanced binary data are present, being the initial portion of outliers set to 5%. As a consequence, the rest of the data is not considered outliers according to all the algorithms.

Due to this imbalance, the Rand score remains high because it includes pairs of data points that are not flagged as outliers by the algorithms.

A different coefficient is needed in order to face this issue, so the Jaccard coefficient is introduced.

The Jaccard coefficient between two sets is defined as the intersection over the union, the formula between the sets A and B is the following:

- M_{00} represents the total number of attributes where A and B both have a value of 0.
- M_{01} represents the total number of attributes where the attribute of A is 0 and the attribute of B is 1.
- M_{10} represents the total number of attributes where the attribute of A is 1 and the attribute of B is 0.
- M_{11} represents the total number of attributes where A and B both have a value of 1.

$$J = \frac{M_{11}}{M_{11} + M_{01} + M_{10}} \tag{6}$$

In this stay is used -1 instead of 1 as label for outliers, while 0 stends for inliers.

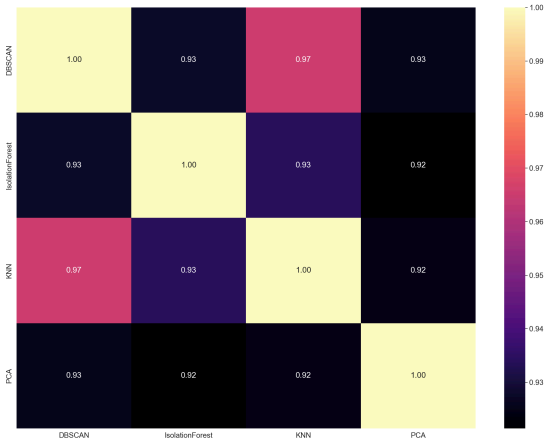

Figure 5: **Rand index matrix**: it displays the pairwise rand index scores between different anomaly detection methods applied. Higher values indicate greater agreement between methods in identifying similar observations as outliers. The lowest observed rand index score is 0.92, demonstrating a high level of consistency among the methods.

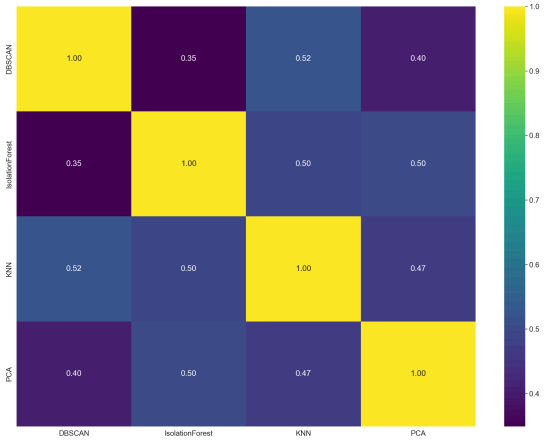

Figure 6: **Jaccard index matrix**: it displays the pairwise Jaccard index scores between different anomaly detection methods applied. Higher values indicate greater agreement between methods in identifying similar observations as outliers.

Notably, the values obtained with the Jaccard coefficient are lower than those obtained with the Rand Score because only the observations classified as outliers by at least one method are considered.

It is noted that 6,611 observations are not flagged as outliers by any method, corresponding to 91.82% of the dataset, indicating that these observations are inliers.

To get more insight into the performance of the methods, only the observations flagged as outliers by at least one method are considered in the following analysis.

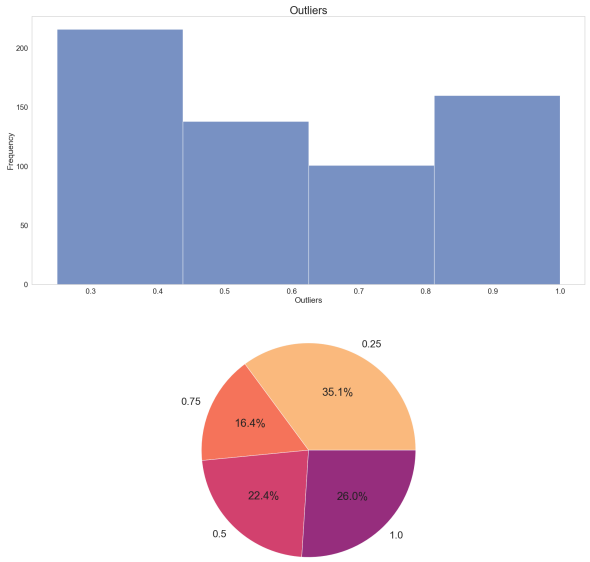

Figure 7: **Outliers frequency** w.r.t. the number of algorithms that flagged an observation as outlier

In particular the histogram and the pie chart display the distribution of the outlier probabilities for each observation and the proportion of observations classified as outliers by different numbers of methods, respectively. The highest bar in the histogram is observed around the value 0.2 indicating that a significant number of observations are classified as outliers by only one method, the percentage associated in this case is 37.7%, which indicates some disagreement or variability in the detection methods. The subsequent highest values are around 0.7 and 1.0, showing that there is a great number of observation classified as outliers by three or all four methods. By looking at the pie chart it can be stated that a considerable portion of outliers is identified by three or all four methods (14,7% and 17.0% respectively) which underline a high confidence in these observations to be true outliers. Moreover, there are fewer observations in the middle range, from 0.4 to 0.6, and just a small percentage of observations is associate in this case, which is 8.4%

A sharp threshold is chosen in order to return a portion of the original data which is considered outlying data by the majority of the implemented methods. In particular the correspondent value is $(0.6, 1]$, which means that any observation with a probability of being an outlier grater than 0.6 is classified as an outlier (in other words, an observation is asked to be detected as an outlier by at least three methods for being actually consider anomaly). As a result 3.08% of the total amount of observations in the dataset is flagged as outliers.

The t-SNE plot reported visualizes the dataset after applying this threshold, categorizing the data into the two required classes: inliers and outliers.

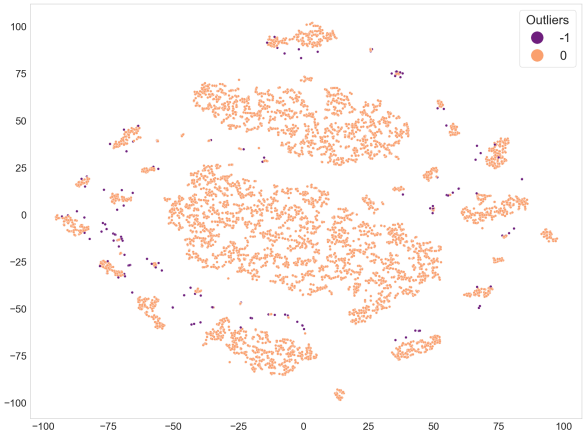

Figure 8: **t-SNE with sharp threshold**: visualization of the dataset after that each observation is categorized in outliers or inliers according to the sharp threshold.

The visualization provided with the t-SNE method helps understanding the spatial distribution of the outliers within the dataset. From the plot is clear that the observation flagged as outliers based on the sharp threshold (purple dots) are *deviating* from what can be consider a sort of *clustering structure* of the rest of the dataset, so the inliers observations. (orange dots).

From the analysis performed in this section it can be stated that even if the methods that are used are based on different assumptions and belong to different families of anomalies detection algorithm, a substantial number of observation are identified as outliers by multiple of these methods. The high level of agreement suggests that a robust identification of the anomalies has been performed enhancing the consistency of the results.

VII. Conclusions

In this study are presented the results of applying various anomaly detection algorithms to a real, mixed-type dataset. Given that the dataset pertains to the medical domain, specifically hyperthyroidism, the results must be interpreted with caution and an understanding of the relevant medical context is needed. This study demonstrates the use of multiple anomaly detection algorithms, assigning a probability of being an outlier to each observation, at first for each algorithm, and then averaging these probabilities in order to obtain a more sharp result. The high level of method's agreement and in the identification of anomalies demonstrate the effectiveness of the applied techniques.

In this study anomaly detection is performed without knowing the significance of the attributes, the implication of their corresponding values and the context of the dataset. In order to improve the quality and the meaningfulness of the results obtained a deeper knowledge of the problem assigned is necessary, in particular considering its medical foundation.

We declare that this report is the result of our own work and has not been copied from any other source. All references and sources used have been properly cited.

References

[1] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, vol. 27, pp. 857–871, 1971.

[2] M. Ester, H.-P. Kriegel, J. S. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1996, pp. 226–231.

[3] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422.

[4] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, p. 417–418, 1933.

[5] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008, [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html