Anomaly Detection on Pharmaceutical Data for Site Segmentation

Ammati VinayKumar[1], Nafisa Ali[2], S Praveen[3]

[1]Student, Department of Electronics and Communication, RV College of Engineering, Bangalore, India
[2]Student, Department of Electronics and Communication, RV College of Engineering, Bangalore, India
[3]Assistant Professor, Department of Electronics and Communication, RV College of Engineering, Bangalore, India

---

Abstract – *Anomalies, which were formerly thought to be only noisy data in statistics, have evolved into a significant subject that is being studied in a variety of subjects and application areas. Various algorithms specifically for outlier detection such as autoencoders, KNN, MCD etc. have been used with proper parameter setting. The parameters while using different algorithms for anomaly/ outlier detection are hyper- tuned according the given data set and even compared with respect to its performance in the form of scores. Two important algorithms namely, Isolation forest and Auto encoders have also been compared and inferences have been drawn out of it.*

Any Pharmaceutical company who must send their sales representatives to the customers or Doctors in this case, it needs to prioritize the customers in the market based on various factors to guide the sales representatives on whom to give more priority which will make their visit much more Efficient resulting in increase in total sales for the company. This is the reason for customer market segmentation, where the data us taken from various data sources and the Sites are segmented on basis of deciles formed using various factors from the data sources. This final segmentation of the sites then used by the sales representatives for their sales tactics like how many visits they should make to any site based on their segment.

Key Words: Data ingestion, univariate analysis, multivariate analysis, outlier detection algorithms, Data source, Segmentation, Decile

1. INTRODUCTION

Exceptions are outrageous qualities that digress from different perceptions on information, they may show an inconstancy in an estimation, exploratory mistakes or an oddity. As such, an anomaly is a perception that separates from a general example on an example. Exceptions or outliers can be of two sorts: univariate and multivariate. Univariate exceptions can be discovered when taking a gander at a circulation of qualities in a solitary component space. Multivariate anomalies can be found in a n-dimensional space (of n-highlights). Taking a gander at circulations in n-dimensional spaces can be exceptionally hard for the human cerebrum, that is the reason we need to prepare a model to do it for us. Anomalies can likewise come in various flavors, contingent upon the climate: point exceptions, logical anomalies, or aggregate anomalies. Aggregate anomalies can be subsets of oddities in information, for example, a sign that may show the disclosure of new wonders.

In any pharmaceutical company it is important to know the market dynamics that is how many doctors are prescribing their product for the patients. This is important because company hires sales representatives who give the product detailing to the doctors to explain the working of drug and it's uses, and side effects and they need to have a clarity on whom to approach more so that they get maximum sales. This is where Segmentation comes into place where prescription data is taken from various data vendors and then the Sites or a geographical area are segmented from Very High (VH) to Very Low (VL) based on the patient count they have. This is refreshed every year because every year new doctors join hospitals or clinics in a particular site and many doctors either retire from practice or move to new sites and it is important for the company to refresh the segmentation every year and have an idea of the customer market for the product, failing to do so the sales representatives may miss out on potential physicians or approach non relevant physicians. This may result in wastage of time and money of both sales representatives and company. So, that's why there is a need to find an optimal solution to perform segmentation of the sites based on different data sources.

2. Literature Survey

In paper [1], review of anomaly detection systems now in use was given, to confront several issues, including imbalanced detection rates, poor detection accuracy, and difficulties detecting anomalies in real-time high-speed networks. Therefore, hybrid machine learning algorithms were created to overcome these challenges by integrating numerous machine learning approaches, such as cascading various classifiers to increase the system's performance. Hybrid approaches are made up of two or more components, each of which performs a separate job, such as preprocessing, classification, clustering, and so on. Each component generates an output, which may be an interim result that is passed on to another component. The goal of hybrid approaches is to take the best of the different algorithms and mix them to overcome the constraints of each and increase the overall performance of the intrusion detection system. In order to create an intrusion detection system, the scientists employed a clustered form of self-organized map (SOM) neural networks.

In paper [2], we use a Systematic Literature Review (SLR) to examine machine learning models that identify abnormalities in their applications. The models are examined from four perspectives: anomaly detection applications, machine learning methodologies, performance measures for ML models, and anomaly detection categorization. In our evaluation, we found 290 research publications published between 2000 and 2020 that explore machine learning algorithms for anomaly detection. We provide 43 various applications of anomaly detection discovered in the chosen research papers after reviewing the chosen research publications. Furthermore, we discover 29 different machine learning models that are utilized to detect abnormalities. Finally, they give 22 distinct datasets used in anomaly detection trials, as well as a variety of additional datasets. Furthermore, we discover that researchers have used unsupervised anomaly detection more than other categorization anomaly detection techniques. Anomaly detection using machine learning models is a promising topic of study, and several ML models have been deployed by academics.

In paper [3], the findings of the ML-based anomaly detection methods for the datasets under investigation are presented in this paper. The dataset ds1 comprises all of the sensor data, while ds2 only includes data from the LIT-101 sensor, which is under assault. The findings reveal that all of the algorithms perform better in the case of ds1, which represents the derived dataset characterized by all of the sensor readings, with the exception of the AE. This implies that the assault on the LIT-101 sensor has an impact on the status of the other sensors, aiding in the detection of malicious activity. Furthermore, on ds1, RF and KNN both get a detection score of 1.0, indicating that their detection methods are quite effective on this data. SVM also does well on this dataset, failing to identify just one element, indicating that the data can be separated from the hyperplane. OCSVM is a discrete method that performs better than supervised algorithms but not as well as AE. The result emphasizes the data's divisibility by the hyperplane, which is capable of better categorizing the data than the RE-based technique.

3. Design and implementation of anomaly detection model and site segmentation

The type of the incoming data is an important part of any anomaly detection approach. The term" input" refers to a collection of data instances (also known as" objects,"" records,"" points," "vectors," "patterns," "events," "cases," "samples," "observations," and "entities"). Collection of attributes (also known as variable, characteristic, feature, field, dimension) may be used to characterize each data object. Different sorts of characteristics, such as binary, category, and continuous, may be used. Each data instance may have just one (univariate) or numerous (multivariate) characteristics (multivariate). In multivariate data instances, all characteristics may be of the same type or may be a mix of various data types. The data provided is of a clinical study which contains multiple domains. The domains have been classified as adverse effects, lab test, medical history, comorbidities, demographics, vital signs, tumor response etc.

Clinical data usually contains the patient history and his/her related test values which can be useful in many applications as such.

- Data set Adverse effects Adverse Effects (AE) lets us know what the adverse effects are the patient have undergone after intake of any medicine or after the lab test.
- Data set Lab test LB (LB) lets us know about various lab results for a patient and for a test.
- Data set medical history Medical History (MH) lets us know about patient's previous medical history, ex. whether the patient was suffering from the disease from a long time or not.
- Data set demographics Demographics (DM) will entirely let us know about the location of the patient and that of his country/ region of stay.
- Data set tumor response Tumor Response (RS) will give information about types of responses registered over a period of cycles for a patient. The usefulness of anomaly detection algorithms is determined on the

nature of the characteristics. The type of characteristics would dictate the distance metric to be utilized in closest neighbor-based approaches. The pairwise distance between instances is often presented in the form of a distance (or similarity) matrix, rather than the actual data. in which no link between data instances is assumed. The raw data in XML format is parsed with an XML parser and each data is converted into its relative data type.

The parsing of the xml file and converting the updated data types into excel sheet is called data ingestion. The labels attached to a data instance indicate whether it is normal or abnormal. Furthermore, anomalous behavior is often dynamic in character; for example, new forms of anomalies may emerge for which no labelled training data exists. Anomalies might lead to catastrophic occurrences in some contexts, such as air traffic safety, and so will be very infrequent. Data is either undergoing label encoding or one hot encoding. We replace the categorical value with a numeric value between 0 and the number of classes minus 1 in Python label encoding. We utilize (0, 1, 2, 3, and 4) if the categorical variable value has five unique classes uncommon. In one hot encoding, we generate a new column (also referred to as a dummy variable) for each category of a feature to indicate whether a given row corresponds to that category.

Once done, we implement the algorithms such as isolation forest (iForest), HBOS, MCD, KNN, Autoencoders on the label/ one hot encoded data. Flow between univariate and multivariate analysis For univariate analysis, we go for the domains which are having no correlation with the other domains and each column is independent of other and that all the spare dependent columns have been reduce or eliminated. The domains on which the univariate analysis will be applied are : Adverse Effects (AE), Lab test (LB)

We formulate our manual analysis in the form of code.

1. Mark the data points which are null in nature. These can be termed as anomalous as they can be termed as missing data.
2. Develop a code to find the variance of each lab test (or any other test) of a subject and make a separate column for it on which the outlier detector algorithm will work.
3. Develop a code to find the results value lying under a certain specified range.
4. Find the percentage change in the values of each test of a particular subject i.e. records and the highest one should be anomalous.

The next step for site segmentation, we use the final output refreshed data source after removing the anomalies for the product, analyse them and get useful insights from them and summarize the data source.

- Create an entire Site Universe based on data from data sources, where all the unique Sites from all the data sources are added after they pass the eligibility criteria which is the Doctors in a particular site are currently active in practice. After this the sites in the data source will be given deciles based on the patient count.
- These Deciles are then converted to Segments using decile to segment mapping.
- Create deciles based on the patient count or sales based on the type of data source
the Account belongs to. Decile is giving the Account a value from 1 to 10 based on it's patient count where 10 indicates the most valuable and 1 indicates least. To calculate the decile for an Account the same procedure is followed as for the HCP segmentation.

4. Application of outlier detection algorithms

We have applied various algorithms where we run the algorithms on all the manually stipulated columns which were calculated with respect to variance, missing data etc.

The algorithms are:

- K- Nearest Neighbors (KNN) algorithm with contamination varying from 1% to 10% and n_neighbours = 5
- Histogram-based outlier detection (HBOS) algorithm with contamination varying from 1% to 10%, n_bins = 10
- Maximum covariance determinant (MCD) algorithm with contamination varying from 1% to 10%
- Isolation forest (iForest) algorithm with contamination varying from 1% to 15% and n estimators=100

- Autoencoder algorithm with contamination varying from 1% to 15% and hidden neurons changing from [3,1,3] to [8,3,8]

For multivariate analysis, we go for the domains which have multiple attributes which can affect each other or are incomplete without each other.

These algorithms would generate anomaly score and anomaly flag as an output and a general note is that the higher the anomaly score, the more chances of the data point to be anomalous.
Outliers are identified by large differences between input and reconstructed data. These algorithms would generate anomaly score and anomaly flag as an output and a general note is that the higher the anomaly score, the more chances of the data point to be anomalous.

5. Results

5.1 Experimental results for univariate analysis

The univariate analysis has been performed on data sets in which certain anomaly occurrence have been notified by manual analysis.

Below figure depicts the results when the variance calculation was done for Lab Results (LB) data sets.

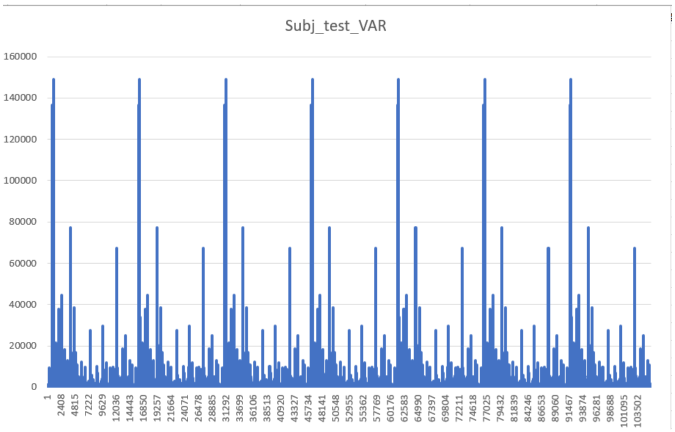


Fig -1: Variance calculation was done for Lab Results

The spikes for each subject ID indicate there are abnormalities or a huge variation between two visits/ cycles/ lab results of a subject. The higher the standard deviation, the higher the score of it being the anomaly. Threshold is a certain level above which the certain subject corresponding a test is decided by the algorithms which we use on them.

Similar process was also conducted for all the datasets.

5.2 Experimental results for multivariate analysis

The data set used for multivariate analysis is Tumor response (RS) as it had multiple columns related to each other and feature engineering was done on top of that.

We get the output of tumor response using the Isolation Forest (Iforest) by keeping the contamination as 0.15 i.e. 15%.

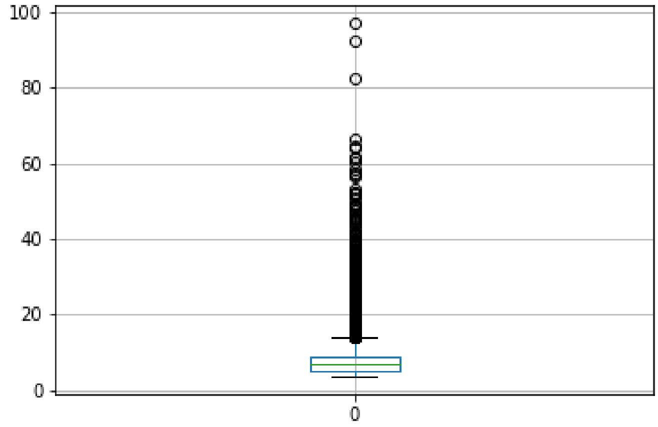


Fig-2: Box plot for Model anomaly scores

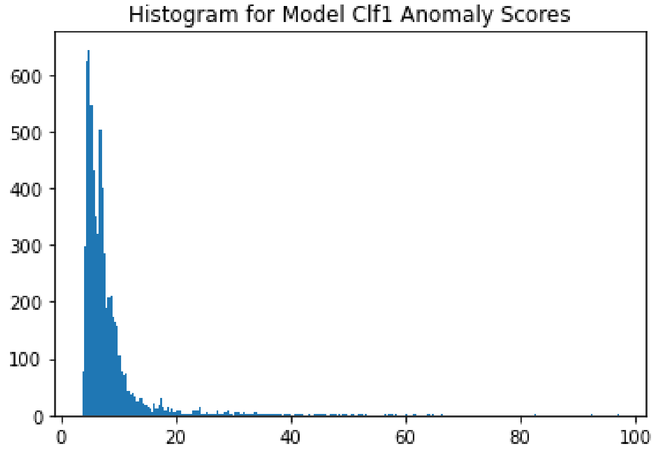


Fig-3: Histogram for Model Anomaly Scores

1. Performance Comparison between Auto Encoders and Isolation forest and inferences drawn

The tumor response and adverse effects was also tested with multiple variations in the columns, by taking few combinations from the base data and few from the featured engineered outputs. By taking into account the cross variations in the combinations, auto encoder algorithm was being compared with that of the isolation forest. The hidden neurons were the only factor which was supposed to be changed apart from the contamination factor as we are in search of the correct quantity of anomalies which is compatible with the quantity of data too.

Here are some of the histograms and box plots depicting the performance of Auto encoder predicting the threshold.

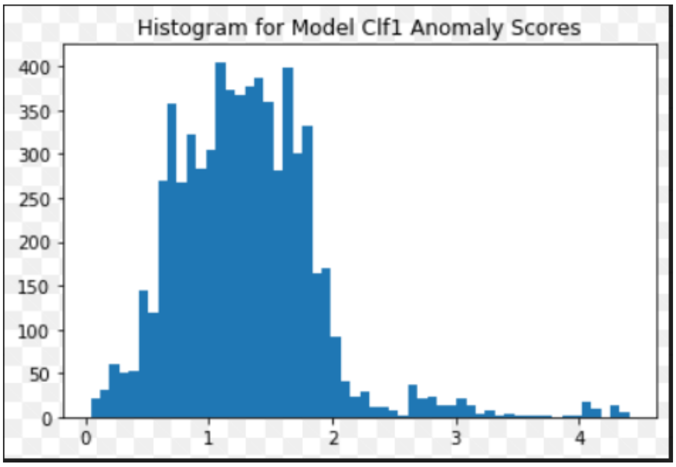


Fig-5: Histogram depiction for Autoencoder model architecture with hidden neurons [8,3,8] and contamination = 10%

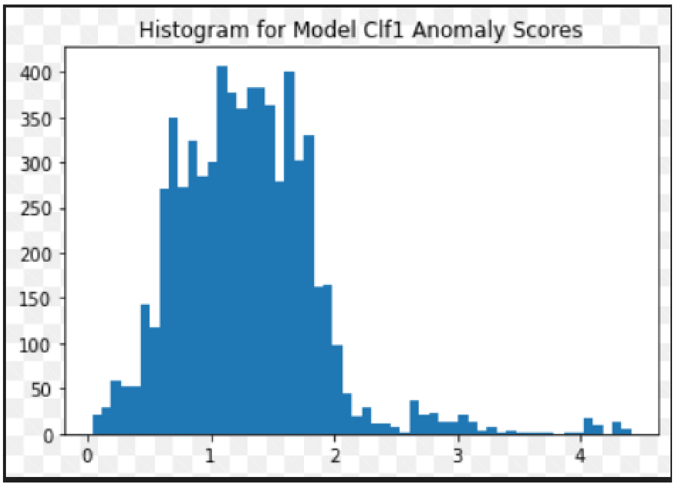


Fig-4: Histogram depiction for Autoencoder model architecture with hidden neurons [8,3,8] and contamination = 15%

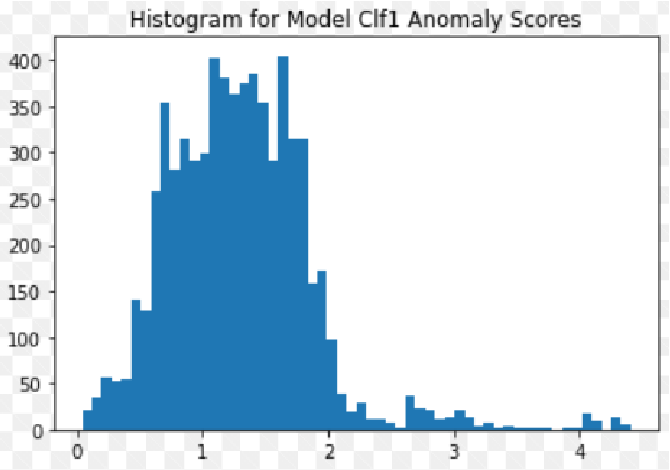


Fig-5: Histogram depiction for Autoencoder model architecture with hidden neurons [3,1,3] and contamination = 10

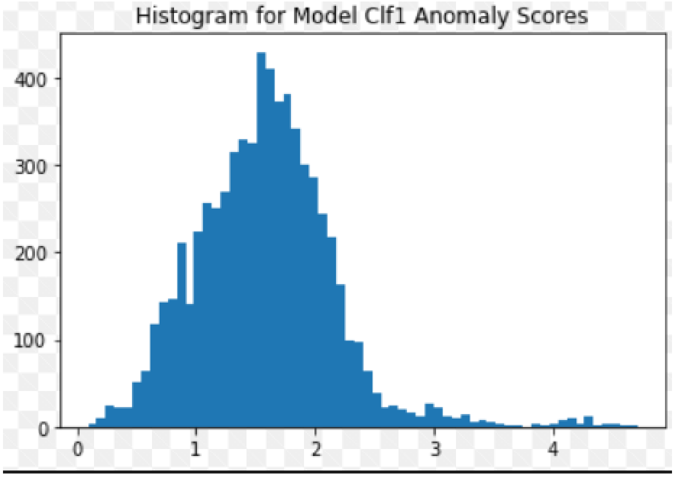


Fig-6: Histogram depiction for Isolation forest model architecture with contamination = 10%

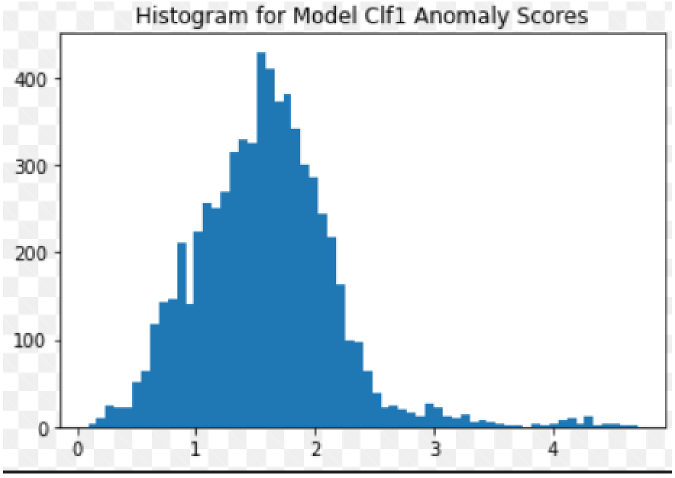


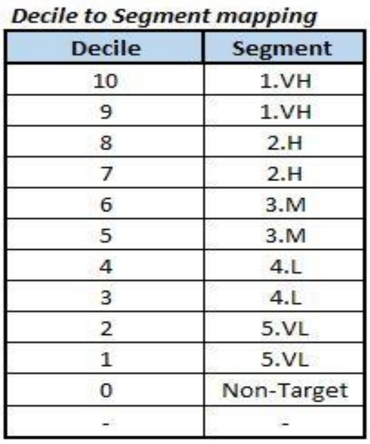
Fig-7: Histogram depiction for Isolation forest model architecture with contamination = 15%

With the variations in the contamination factor in iForest and variations of contamination factor and hidden neurons, we notice that with the same contamination factor, auto encoder performs with accuracy greater than 90% if working on large dataset. Isolation forest on the other hand needs more features and less base columns to find a pattern to establish and therefore have a greater control over finding the anomalies with the developed pattern.

For creating Decile for any site in the data source, first take the sum of the patient count of all the sites who have patient count less than that of site we are considering at present and divide it by the total patient count of all the site and at last multiply this number by 10 to get a number between 1 to 10.This is the process used to get decile.

Decile of Site A= (Sum of patients in site A with patient count less than that of site A/sum of the total patient count) *10

These Deciles are converted into Segments based on the below Decile to segment convert table.



Decile to Segment mapping

Decile	Segment
10	1.VH
9	1.VH
8	2.H
7	2.H
6	3.M
5	3.M
4	4.L
3	4.L
2	5.VL
1	5.VL
0	Non-Target
-	-

Figure -8: Decile to segment Conversion.

So, based on the table above deciles are finally converted to Segments and then assigned to the doctors where VH means Very High ,H means High , M means Medium , L mean Low and VL means Very Low.

Segmentation of Sites			
Segment	Number of Sites	Number of Patients	Patients per Site
1.VH	5	804	161
2.H	10	805	81
3.M	14	725	52
4.L	20	703	35
5.VL	63	726	12

Table-1: Final sites Segmentation

We could see that the in the segmentation results the Sites from the higher segment cover the same patient percentage as the other segments but the number of Sites in that segment is very low compared to other segments. Thus, indicating patient per site is very high for them, which makes them high priority sites for the sales representatives.

2. CONCLUSIONS

Various forms of outlier identification methods were evaluated on the preprocessed data-set in this study, and we concluded the anomalous spots in the data using the score and the histogram. This paper presents the findings of an experimental study of many conventional outlier detection methodologies. To begin, we look at two statistical techniques for identifying outliers: linear regression and deep learning techniques. The results of the experiments reveal that the deep learning technique beats the liner regression approach when it comes to recognizing outlier data. With the aid of grid search techniques, we experimented with hyper tuning several parameters, and the best parameters were considered to locate the proper group of outliers. We also provide a comprehensive and extended description of numerous outlier detection algorithms in this paper. As a consequence of this effort, we have a better understanding of the many avenues of research into outlier analysis.

Based on the data source we got, the sites were given Deciles based on the patient count they had. These deciles were calculated based on decile formula which like percentile. Then, these deciles were converted to segments based on decile to segmentation mapping table.

Future scope

More research is needed in most outlier detection-based systems. Therefore, in addition to the previously specified future work in each area, the following additional research gaps must be addressed:

- More study is required to fully characterize and correlate some of these methods to real-world data, particularly in very large and high-dimensional databases, where first-hand methods for estimating data densities are valuable. In high-dimensional data sets, the curse of dimensionality and distance concentration are still unsolved challenges.

These data, sources were not giving the exact count for the patients. So, more efficient ways to retrieve customer data for the product could make segmentation process more accurate than it is now.

REFERENCES

[1] F. Angiulli, S. Basta, and C. Pizzuti, “Distance-based detection and prediction of outliers,” IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 2, pp. 145–160, 2006. doi: 10.1109/TKDE.2006.29.

[2] H. Wang, M. J. Bah, and M. Hammad, “Progress in outlier detection techniques: A survey,” IEEE Access, vol. 7, pp. 107 964–108 000, 2019. doi: 10.1109/ACCESS. 2019.2932769.

[3] S. B. Wankhede, “Anomaly detection using machine learning techniques,” in 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1–3. doi: 10.1109/I2CT45611.2019.9033532.

[4] G. R. Jidiga and P. Sammulal, “Anomaly detection using machine learning with a case study,” in 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, 2014, pp. 1060–1065. doi: 10.1109/ICACCCT. 2014.7019260.

[5] Y. Wang, B. Xue, L. Wang, H.-C. Li, L.-C. Lee, C. Yu, M. Song, S. Li, and C.-I. Chang, “Iterative anomaly detection,” in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017, pp. 586–589. doi: 10.1109/IGARSS. 2017.8127021.

[6] K. Zhang, X. Kang, and S. Li, “Isolation forest for anomaly detection in hyperspectral images,” in IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 437–440. doi: 10.1109/IGARSS.2019.8899812

[7] Agnetis, Alessandro , Messina, Vincenzina Pranzo, Marco. (2010). Call planning in European pharmaceutical sales force management, Ima Journal of Management Mathematics - IMA J MANAG MATH. 21. 267-280. 10.1093/imaman/dpp019.

[8] Ahmadi, Pooria, Samsami, F. (2010). Pharmaceutical market segmentation using GA K-means, European Journal of Economics, Finance and Administrative Sciences. 72-82. Giaquinta, Diane. (2003). The emerging specialty pharmaceutical market segment, Managed care interface. 16. 35-6. Krykavskyy, Yevhen, Kosar, Nataliia Pytulak, Nataliia. (2020). Marketing research of pharmaceutical market trends.