Identification of Outliers Based on Sensitivity of Data

Prof. Shiva Sumanth Reddy[1], Amogh S Bharadwaj[2], Chandrakanth N Murthy[3,] H Vishwajit[4] and Harsh V Challa[5]

[1]*Asst. Professor, Department of CSE, Dayananda Sagar Academy of Technology and Management, Bangalore, India*

[2,3,4,5]*Student, Department of CSE, Dayananda Sagar Academy of Technology and Management, Bangalore, India*

***Abstract*—Outlier detection is examined and can be applied in many fields and domains. The Outliers occur because of the manual errors which are most often made by humans while data entry, fraudulent error when a malicious practice occurs, system behavior, by natural deviations in datasets or instrument error. The detection of outliers has been used over many years, to identify outliers and analyses outliers where required. Sometimes the outliers need to be removed, sometimes we use it to identify Outliers. The main challenge in outlier detection is to work on data that is highly Sensitive data. Sometimes the data is so sensitive that the outlier data coincides with the normal data, this usually occurs in the domain of malicious activities. The proposed system assists to clean data in less time and great accuracy. Our paper focuses on outlier detection which can be applied to various domains that have time series data. This shows critical review on various approaches to detect outliers and provides the most accurate technique for particular type of data.**

***Index Terms*—Anova Test, Data Preprocessing, Data Analytics, Dataset Sensitivity, Box plot, Mean Absolute Deviation, Isolation Forest, Local Outlier Factor.**

I. INTRODUCTION

An Outlier is data entry that deviates drastically from the normal data. The outliers are based on their sensitivity, if the outliers coincide with normal data, it is highly sensitive and if the outliers deviate from normal data, it is moderately sensitive. Detecting outliers may be used in a broad range of fields, including system defect detection, intrusion detection, credit card fraud detection, insurance fraud detection, and monitoring for enemy activity in the military.

Outliers are basically classified based on their occurrences namely Point Outlier, Contextual Outlier and Collective Outlier. When a single data instance is sitting outside the normal range of the dataset. For instance, if a patient's height was recorded but one digit was left off, the dataset would include a point outlier. When the data instances are drastically different from the dataset, but only within a specific context. For example, in a dataset of Bangalore temperatures over time. A temperature reading less than fifteen degrees in winter is normal but the same dataset reading below fifteen in summer is an outlier. When a series of data instances drastically drift from the normal form of the dataset. For Example, using a time series to highlight seasonal or daily variations in subscribes and unsubscribes to an email marketing list. The level of subscribed users might be labelled as a collective outlier if it remained constant for several weeks without variation. It is common for individual users to unsubscribe and for new users to subscribe, therefore a static figure would be considered an oddity.

TYPES OF DATASSETS

1. Cross-Sectional data

When several people are observed at the same time, cross-sectional data is obtained. Cross-sectional data may include observations made several times, although in such circumstances, time itself has little bearing on the study. An example of cross-sectional data is the test scores of students in a particular year. Another example of cross-sectional data is the GDP of nations in a specific year. Another illustration of cross-sectional data is data used for customer turnover analysis. It should be noted that the exam scores of students and the GDP of countries are cross sectional datasets because all observations were made within a single year. In both instances, the cross-sectional data essentially serves as a picture of the world at a particular moment in time. But data on customers for churn research may be gathered over longer periods of time, including years and months. Though time may not be a significant factor for analytic purposes, But customer turnover data may be gathered at several

periods in time, it may still be regarded as a cross-sectional dataset.

The following graph, which uses the example of military spending as a proportion of GDP for 85 nations in 2010. We guarantee the data's cross-sectional character by using data from a single year. To illustrate various statistical characteristics of the military spending data.

The figure clearly shows that military spending has a large peak at around 1.0% and is slightly left-skewed. Near 6.0% and 8.0%, a few tinier peaks may also be seen.

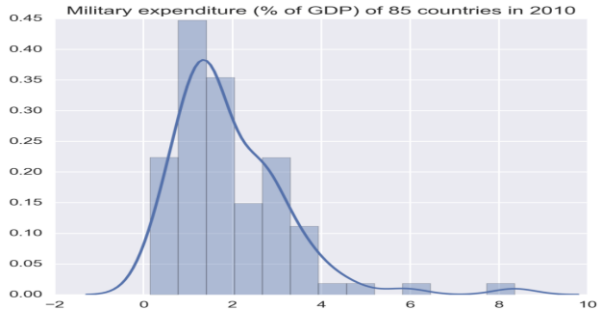

Figure 1: Military Expenditure for 85 countries in 2010

2. Time-Series Data

A time-series data set consists of information or observations acquired across a range of regular or erratic time intervals. A time series is often a collection of data points that were collected at consistently spaced-out intervals. The number of data points that are recorded may be counted every hour, every day, every week, every month, every quarter, or every year. Applications of time series can be found in business, finance, or statistical fields. The daily closing value of stock indices like the NASDAQ or SENSEX is a widely popular example of time series data. Time series are also often used in econometrics, signal processing, pattern identification, sales and demand forecasting, weather forecasting, and earthquake prediction.

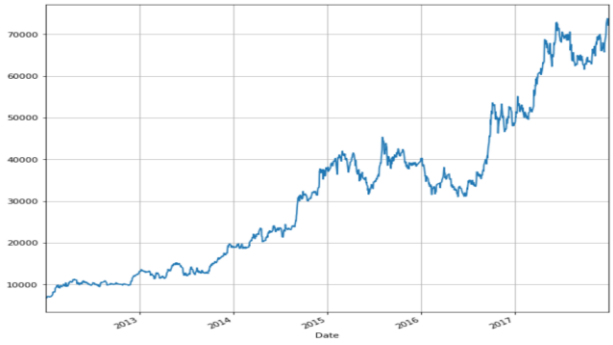

Figure 2: Senex value from 2012 to 2018

The given figure illustrates a time series dataset for change in the value of Sensex for a fixed period of a year

3. Panel Data

Data that is observed on various cross sections across time is known as panel data, often known as longitudinal data. It is an assortment of measurements acquired from many sources, accumulated over regular time periods, and arranged chronologically. Several examples of the types of groups that can be included in panel data series are countries, organisations, people, or demographic categories. Similar to time series data, panel data also comprises observations that were sequentially acquired at predetermined intervals. Panel data involves observations made across a group of individuals, just like cross-sectional data does.

Person	Year	Income	Age	Sex
1	2013	20,000	23	F
1	2014	25,000	24	F
1	2015	27,500	25	F
2	2013	35,000	27	M
2	2014	42,500	28	M
2	2015	50,000	29	M

Table 1: Income of two people from 2013-2015

The information (the characteristics of income, age, and sex) gathered over the course of 3 years for various individuals is depicted in the table above. It displays the information gathered over a three-year period for two individuals (persons 1 and 2). (2013, 2014, and 2015). This is a typical Panel dataset table.

II. PROPOSED SYSTEM

The proposed system has models related to data collection, data pre-processing, defining the sensitivity of data, applying appropriate techniques to detect outliers and assessing the result produced by each of those algorithms in order to derive a conclusion. Understanding and modelling the available data are the core components of the outlier detection system.

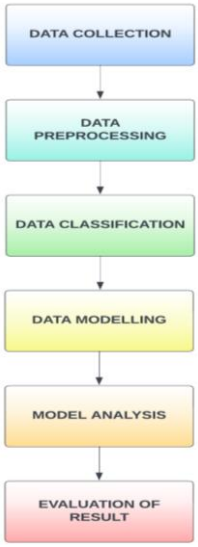

Figure 3: Methodology

1. DATA COLLECTION

Data collection is an important aspect in developing a proposed system. It is a systematic approach to accurately collect the information from various sources, the main driver to it is the quality of the data. This is a process of gathering, storing, visualising and analysing data from various sources. In this system we are more engrossed in time-series datasets. The data collection process has had to change and grow with time. Time Series data is a set of data points observed through repeated measurements over time, For example electrical activity in the brain over time, heartbeats per minute in a day, stock prices, annual retail sales and many more. Time series data can be of two types: measurements gathered in regular time intervals and irregular time intervals. Heartbeat measured per minute is measured in regular intervals and annual retail sales are measured in irregular intervals.

Time series data plays an important part in our daily lives as it treats time as the primary axis which is used for statistical analysis, forecasting and monitoring systems.

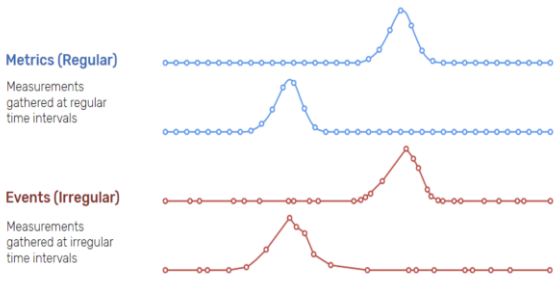

Figure 4: Classification of Time series data

Time series data are of two types univariate and multivariate. Univariate time series consists of a single observed value in regular or irregular time interval. Consider an example that shows the industrial index with respect to time

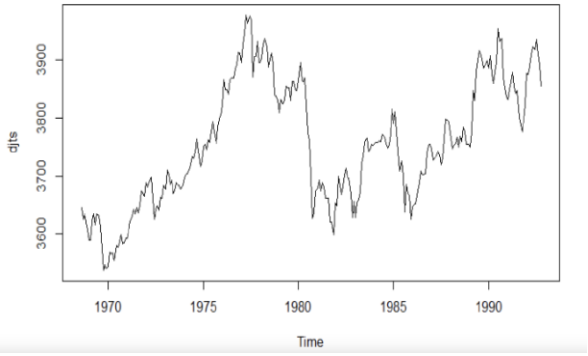

Figure 5: Univariate time-series data

This is univariate time series data that shows the industrial index value for different years from 1970 to 1990.

Multivariate time series data has data points in the datasets that have more than one time-dependent variable. If there are 2 variables that are dependent on time it is called Bivariate.

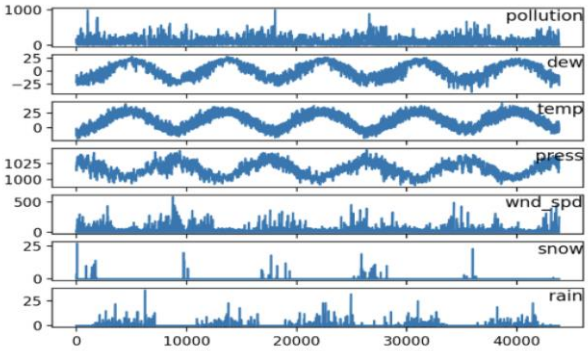

Figure 6: Multivariate time - series data

In this example we see that the air pollution is derived from all these variables and the time axis is based on the primary axis. Each variable has dependency on each other and all together can be used to predict air pollution.

2. DATA PRE-PROCESSING

Data pre-processing includes procedures we follow to change or encode data so that the computer can rapidly decode it. The primary requirement of a model to be robust and precise is that the algorithm should be able to quickly understand the characteristics of the data. When data mining methods are used to this abnormal data, the outputs may not be of greatest quality of data, Since the

patterns may not be successfully identified. Hence, data processing is pivotal to raising the general level of data quality. Here the proposed prepossessing model used is ANOVA Test.

An ANOVA test is a form of demographic test used to determine if there is a statistically significant difference between two or more groups by testing for variance-based differences in means between them. ANOVA divides the unusual variable into two or more groups. For instance, one group may be employed as a control group and not be expected to have an impact on the dependent variable, whereas the other would be expected to do so.

The audacity for any parametric test applies to the ANOVA test are:

- The samples taken from the population should have a *normal distribution*.
- *Case independence*: The sample instances have to be separate from one another.
- The variance should be *homogeneous*, which implies that it should be about equal across all groups.

3. DATA CLASSIFICATION

Data is classified based on the sensitivity. It can be classified as highly sensitive, low sensitive and medium sensitive. Here the word sensitive basically means the confidentiality and the integrity of the data.

- *Highly sensitive data*: They are those type of data which when destroyed in an unauthorised manner, will have a catastrophic effect on the organisation
- *Medium sensitive data*: They are intended for only the purpose of internal use.
- *Low sensitivity data*: These kinds of data are intended for the use of public. These are low sensitive

4. DATA MODELLING

The proposed model aims in detecting the abnormal observations or the outliers in various kinds of data sets. The input given to the model is the dataset and the model works on what kind of outlier detection algorithm is best suitable for what kind of the dataset based on the sensitivity of the data which is simultaneously calculated by various statistical metrics such as the variance, standard deviation and covariance.

The inbuilt outlier detection algorithms such as the Mean Absolute Deviation (2 standard and 3 standard), 2 sigma and 3 sigma deviation, Boxplot and Adjusted Boxplot are used as the basis for the proposed model.

Based on the accuracy of the algorithm for the particular dataset, it is said that particular algorithm is best suited for that dataset.

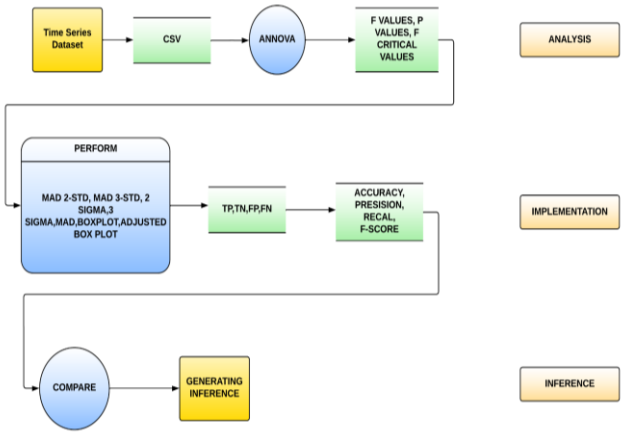

Figure 7: Data Flow Diagram

5. MODEL ANALYSIS

Analysis of the model is done by means of confusion matrix. A confusion matrix is a table that is often used to evaluate the performance of a machine learning model. It is also known as an error matrix. The matrix provides a detailed breakdown of correct and incorrect predictions made by a classifier or model, compared to the actual or expected outcomes.

The confusion matrix consists of four different categories:

- *True positives (TP):* The amount of accurate and positive predictions the model produced.
- *False positives (FP):* The amount of positive predictions that the model made that were inaccurate.
- *True negatives (TN):* The amount of accurate negative predictions the model made.
- *False negatives (FN):* The amount of the model's negative predictions were inaccurate.

The rows in the confusion matrix correspond to the actual values of the target variable, while the columns correspond to the predicted values. The values in the matrix are usually represented as counts or percentages. Accuracy, precision, recall, and F1 score are a few examples of performance measures that may be calculated for a model using a confusion matrix. These indicators may be used to assess a model's overall efficacy and pinpoint areas that need improvement.

6. EVALUATION OF RESULTS

The accuracy of each algorithm with different types of sensitivity of the dataset are compared side by side. The result are to be illustrated as a table of performance of each of the dataset with one particular algorithm in order to find the method with is most suitable for that type if dataset sensitivity.

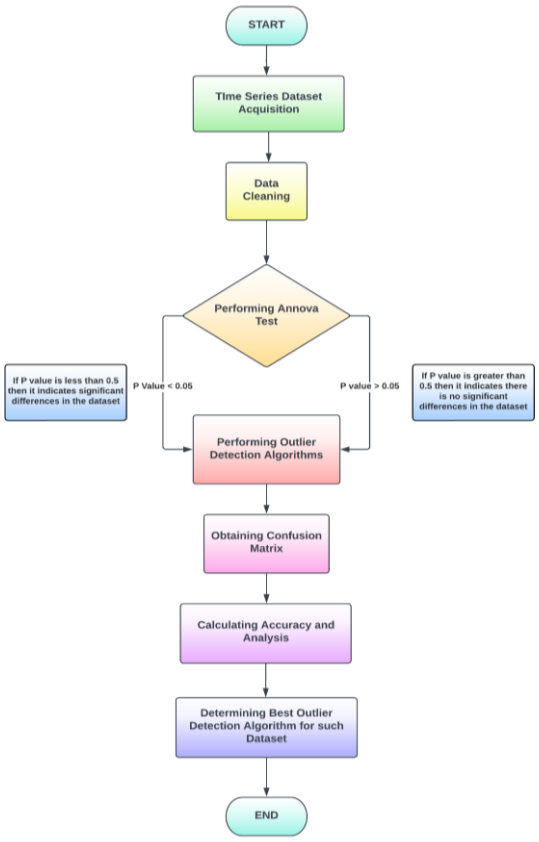

Figure 8: Flowchart

III. METHODOLOGY

The methodology involves preprocessing the dataset, quantifying data sensitivity, applying outlier detection techniques, evaluating performance, and analyzing the impact of data sensitivity on outlier identification.

1. TWO SIGMA ALGORITHM

This algorithm is used for detecting outliers based on normal distribution. One dimensional space means that you detect outliers based on one variable. Data within two standard deviations of the mean are referred to statistically as values of 2-sigma. The upper and lower control limits in statistical quality control charts are defined using two-sigma limits. With these 2-sigma limits. Control chart is used to determine whether a dataset contains a controlled or uncontrolled variation.

2. THREE SIGMA ALGORITHM

This algorithm is based on normal distribution, It has 68 to 95 percentage of values that lie within an interval around mean, Data within three standard deviations of the mean are referred to statistically as values of 3-sigma. The upper and lower control limits in statistical quality control charts are defined using three-sigma limits. Control chart is used to determine if a dataset's variations inside the bounds are controlled or uncontrolled.

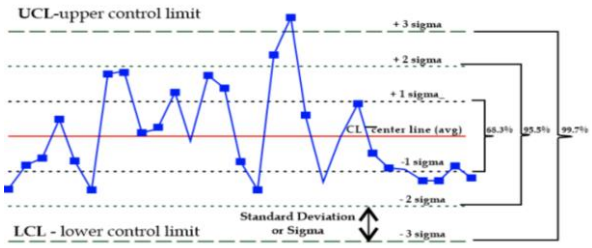

Figure 9: Illustration of 3 Sigma and 2 Sigma Algorithm

3. BOXPLOT ALGORITHM

Boxplot algorithm is the algorithm used for detecting the outliers in a univariate data. It enables a simpler way for the visualisation of the data. It also enables simpler comparison of characteristics of data within the categories.

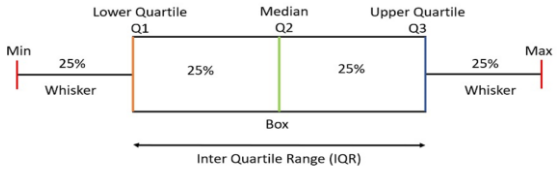

Figure 10: Illustration of Box Plot Algorithm

4. ADJUSTED BOXPLOT ALGORITHM

Boxplot and Adjusted Boxplot are almost the same, apart from the exponential function which came to control the skewness of the distribution. An adjusted boxplot algorithm is based on the upper quartiles and lower quartiles, along with an estimator called the Medcouple. Data can be skewed, meaning it tends to have a long tail on one side or other. We use skewness as a measure of symmetry. On the right side, you can see a plot with the positive skew, the right tail is longer. On the left side you can see a negative skew plot, the left tail is longer.

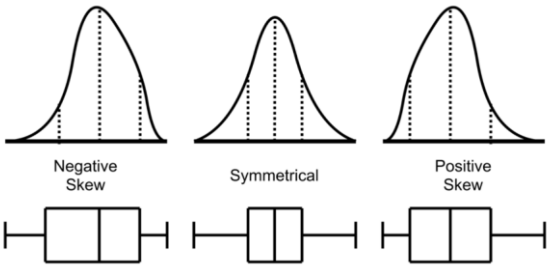

Figure 11: Illustration of Left Skew and Right Skew

5. MEAN ABSOLUTE DEVIATION ALGORITHM (1,2 STD)

Median Absolute Deviation is a measure of statistical dispersion. We know that Mean and Median are two measures of central tendency, but Median is a robust measure that is not sensitive to extreme values. Three Sigma is sensitive to extreme values. The MAD is a robust measure. Let's divide the MAD calculation into three parts. First part is we calculate median of the samp, then subtract median from every observation and take absolute value and the third part is to take the median of the just computed n values.

Figure 12: Illustration of MAD Algorithm

IV. LITERAURE SURVEY

[1] In their research paper, Yen-Cheng Lu, Chih-Wei Wu, Chang-Tien Lu, and Alexander Lerch conducted a comparative study of six outlier detection algorithms applied to a Music Genre Recognition dataset. The purpose was to evaluate the ability of the outlier detection algorithms in identifying the presence of corrupted or mislabelled music genre data records. The study's findings were used to improve the dataset by applying various pre-processing techniques to remove any noise. To extract features, the block-wise analysis method was used, and the K-means algorithm was implemented for unsupervised clustering. To define outlier scores, the KNN-KNN method was used, which determines the distance of each data point to its kth nearest neighbour.

[2] Teodora Mecheva's research paper examines outlier detection in a traffic data set using three techniques: Local Outlier Factor (LOF), Isolation Forest algorithm, and Support Vector Machine. The traffic data is obtained from virtual detectors installed in road cameras, and a subset is created based on working day hours for each day. The three outlier identification techniques are applied to each subset, and the article also focuses on using these techniques to purify the data set. Efficiency is measured by comparing the coefficient of variation of the raw data and the purified data. The methodology for data purification can be improved by expanding the dataset size.

[3] S P Maniraj, Aditya Saini, and Swarna Deep Sarkar Shadab Ahmed conducted a study on credit card fraud detection using outlier detection. The authors focused on demonstrating how to model a dataset using machine learning algorithms with a credit card dataset. The problem involves modelling fraudulent transactions based on previously detected fraudulent transactions, allowing new transactions to be evaluated for fraud. The study implemented Local Outlier Factor and Isolation Forest algorithms for this purpose, and the project was designed to allow for the integration of various algorithms. The proposed system was implemented in Python, and the Local Outlier Factor algorithm achieved an accuracy of 97%, while the Isolation Forest algorithm achieved an accuracy of approximately 76%.

[4] Alva Presbitero, Rick Quax, and Peter M. A. Sloot conducted a project on anomaly detection in clinical data of patients undergoing heart surgery, with the aim of detecting and estimating the physical conditions of patients to determine their health status. Anomalies in this type of dataset can hinder the ability to diagnose patients correctly, and these anomalies are detected by changes in patterns that indicate transitions from a healthy state to an unhealthy state. The study utilized algorithms such as Isolation Forest Algorithm and Local Outlier factor algorithms to detect anomalies in the dataset. However, the paper does not provide a technique for distinguishing critical patients from non-critical ones, which limits the ability to identify important patients using simple outlier identification methods. Instead, a broader strategy, such as the surprise technique, is necessary.

[5] The research paper titled "Graph-based Anomaly Detection and Description" by Leman Akoglu, Hanghang Tong, and Danai Koutra aims to provide an organised overview of state-of-the-art techniques for detecting

outliers in data represented by graphs. The paper presents a general framework that categorises these techniques based on various factors such as supervised vs. unsupervised, static vs. dynamic, and attributed vs. plain graphs. Graphs represent the interdependent nature of data attributes, and the paths connecting data objects can capture long-range correlations effectively. In addition to detecting anomalies, the paper also focuses on post-detection analysis and sense-making of the detected anomalies by providing a survey of tools and techniques. The paper also utilises unsupervised clustering techniques such as K-means and agglomerative clustering for anomaly detection

[6] Varun Chandola conducted a survey on Outlier Detection, which is divided into three main sections. The first section deals with the identification of the problem statement, the second section deals with the identification of various aspects that contribute to the exact formulation of the problem, and the third section deals with the applications that utilise outlier detection. The paper emphasises that each data instance can be described using a set of attributes, and the main challenge in outlier detection is to identify the best set of attributes that can calculate the efficiency of the algorithm based on parameters such as accuracy. The outliers are given numerical representation and categorised based on their level of importance, which helps to prioritise them.

[7] The paper titled "Outlier Detection for Patient Monitoring and Alerting" was authored by Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F. Cooper and Gilles Clermont. The aim of the study was to investigate the impact of anomalies on medical datasets, which could potentially have harmful effects on patients' health. The study proposed an outlier-based monitoring and alerting method to improve the overall clinical coverage of patient monitoring. The study used data from 4,486 post-cardiac surgery patients and a subset of 222 warnings produced by the proposed method to demonstrate a positive correlation between statistical outliers and clinically significant warnings. The results suggest that outlier-based alerting has potential clinical applications.

[8] The paper titled "Anomaly Detection for Cyber-Security Based on Convolution Neural Network: A Survey " was conducted by Montdher Alabadi and Yuksel Celik of Karabuk University in Turkey. The aim of the project is to explore different outlier detection methods for various domains, with a focus on distance-based and density-based approaches. The distance-based approach calculates the distance between data objects and uses this information to identify outliers, while the density-based approach identifies anomalies that are located in the neighbourhood of data objects. A function called "outlier factor" is used to quantify the amount of distance between a given object and others in the dataset. By identifying anomalies, the study hopes to detect traffic attacks based on deviations from established profiles of data abnormality.

[9] The paper "Outlier Detection in Network Traffic Monitoring" by Marcin Michalak, Łukasz Wawrowski, Marek, Rafał Kurianowicz, Artur Kozłowski, and Andrzej Białas discusses the results obtained from analysing real-world network traffic data collected at an institute. The study focuses on two variables, namely the number and size of data packets, making the dataset bivariate and requiring the detection of outliers in two-dimensional space. The experiments were conducted in two modes, one to learn the data in the best possible way, and the other to investigate the practical applications of the methods in real-time scenarios.

V. RESULTS

In this research we have taken a total of three dataset of different sensitivities and have test a total of six algorithms. Dataset one has a p-value from the anova test of 0.000699 making it the moderately sensitive dataset in the test followed by dataset two with p-value of 0.003557 of making it most sensitive dataset and then dataset three with p-value of 7.5E-52 making it the least sensitive dataset of this research.

	EXAMPLE	ALGORITHMS	TP	FP	FN	TN
0	Car Engine	MAD-2std	48.0	5.0	6.0	11.0
1	None	MAD-3std	52.0	1.0	13.0	4.0
2	None	2Sigma dev	52.0	1.0	14.0	3.0
3	None	3Sigma dev	53.0	0.0	17.0	0.0
4	None	BoxPlot	51.0	2.0	13.0	4.0
5	None	Adj-Boxplot	46.0	7.0	14.0	3.0
6	Heart Rate	MAD-2std	56.0	2.0	1.0	16.0
7	None	MAD-3std	58.0	0.0	6.0	11.0
8	None	2Sigma dev	58.0	0.0	12.0	5.0
9	None	3Sigma dev	58.0	0.0	15.0	2.0
10	None	BoxPlot	58.0	0.0	1.0	16.0
11	None	Adj-Boxplot	58.0	0.0	2.0	15.0
12	Electricity Consumption	MAD-2std	67.0	4.0	1.0	18.0
13	None	MAD-3std	71.0	0.0	2.0	17.0
14	None	2Sigma dev	71.0	0.0	17.0	2.0
15	None	3Sigma dev	71.0	0.0	19.0	0.0
16	None	BoxPlot	71.0	0.0	4.0	15.0
17	None	Adj-Boxplot	56.0	15.0	19.0	0.0

	ALGORITHMS	ACCURACY	PRESISION	RECALL	F-SCORE(β=2)
0	MAD-2std	0.842857	0.905660	0.888889	0.897196
1	MAD-3std	0.800000	0.961132	0.800000	0.861356
2	2Sigma dev	0.785714	0.961132	0.787879	0.873950
3	3Sigma dev	0.757143	1.000000	0.757143	0.861789
4	BoxPlot	0.785714	0.962264	0.796875	0.871795
5	Adj-Boxplot	0.700000	0.867925	0.766667	0.814159
6	MAD-2std	0.960000	0.965517	0.982456	0.973913
7	MAD-3std	0.920000	1.000000	0.906250	0.950820
8	2Sigma dev	0.840000	1.000000	0.828571	0.906250
9	3Sigma dev	0.800000	1.000000	0.794521	0.885496
10	BoxPlot	0.986667	1.000000	0.983051	0.991453
11	Adj-Boxplot	0.973333	1.000000	0.966667	0.983051
12	MAD-2std	0.944444	0.943662	0.985294	0.964029
13	MAD-3std	0.977778	1.000000	0.972603	0.986111
14	2Sigma dev	0.811111	1.000000	0.806818	0.893082
15	3Sigma dev	0.788889	1.000000	0.788889	0.881988
16	BoxPlot	0.955556	1.000000	0.946667	0.972603
17	Adj-Boxplot	0.622222	0.788732	0.746667	0.767123

Figure 13: Result containing Most Accurate Algorithm obtaining from Confusion Matrix

We found that among the six algorithms the Boxplot algorithm worked the best in detection outliers in the most sensitive dataset with a f-score of 0.991453, for the moderately sensitive dataset we found that the MAS-2STD algorithm worked the best in identifying the outlier with a f-score of 0.897196 and lastly the MAD-3STD algorithm performed the best in detection of outliers in the least sensitive dataset with a f-score of 0.986111.

VI. CONCLUSION

In conclusion, identifying outliers in data is an important task in various domains such as healthcare, cyber-security, and network traffic monitoring. The models and approaches discussed in the research papers highlight the need for sensitivity analysis and the selection of appropriate algorithms to effectively detect outliers.

Additionally, post-detection analysis and sense-making tools and techniques have been explored to provide a better understanding of the outliers and their potential impact. Overall, the research in this area is ongoing and will continue to play a vital role in ensuring the quality and reliability of data analysis in various fields.

VII. FUTURE ENHANCEMENT

To optimise the efficiency and effectiveness of identifying the most appropriate algorithm for outlier detection, one potential strategy is to incorporate a diverse range of datasets with varying levels of sensitivity into the proposed system. By leveraging multiple datasets with different characteristics, the system can better understand the nuances of each dataset, and make more informed decisions about which algorithm is best suited for outlier detection.

Moreover, an automated model can be developed to assist users in selecting the most appropriate algorithm for detecting outliers in their specific dataset. This automated model would be designed to take in the user's dataset, and use advanced analytics and machine learning techniques to evaluate the dataset's unique characteristics and identify the optimal algorithm for outlier detection. With continuous learning and refinement, this automated model can provide increasingly accurate recommendations over time, resulting in more efficient and effective outlier detection for the user.

Overall, this approach has the potential to improve the accuracy and efficiency of outlier detection in various domains, including finance, healthcare, and environmental monitoring, among others.

REFERENCES

[1] Yen-Cheng Lu, Chih-Wei Wu, Chang-Tien Lu, Alexander Lerch. Automatic Outlier Detection in Music Genre Datasets. KNN-KNN implementation. August 2016 DOI:10.1085/9.1093854 PMID: 356087659.

[2] Teodora mecheva outlier detection in traffic data set. 17th international conference on concentrator photovoltaic systems (cpv-17) DOI:10.1063/5.0093554

[3]S P Maniraj, Aditya Saini, Swarna Deep Sarkar Shadab Ahmed. Credit card fraud detection using outlier detection October 2021. International Journal of Scientific and Technology DOI:10.17577/IJERTV8IS090031

[4] Alva Presbitero, Rick Quax, Peter M. A. Sloot . Anomaly Detection in Clinical Data of Patients Undergoing Heart Surgery. December 2017 DOI:10.1016/j.procs.2017.05.002

[5] Leman Akoglu, Hanghang Tong, Danai Koutra Graph-based Anomaly Detection and Description: A Survey. June 2022 DOI: 10.18280/ts.390327

[6] Outlier Detection A Survey by Varun Chandola, University of Minnesota Arindam Banerjee, University of Minnesota and Vipin Kumar, University of Minnesota.

[7] Outlier Detection for Patient Monitoring and Alerting by Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F. Cooper and Gilles Clermont.

[8] Anomaly Detection for Cyber-Security Based on Convolution Neural Network: A survey by Montdher

Alabadi Computer Engineering Department, Karabuk University, Karabuk, Turkey and Yuksel Celik Computer Engineering Department, Karabuk University, Karabuk, Turkey.
[9] Outlier Detection in Network Traffic Monitoring by Marcin Michalak, Łukasz Wawrowski, Marek, Rafał Kurianowicz, Artur Kozłowski and Andrzej Białas Research Network Łukasiewicz, Institute of Innovative Technologies EMAG, ul. Leopolda 31, 40–189 Katowice, Poland