Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EVALUATION OF MACHINE LEARNING TECHNIQUES FOR ANOMALY DETECTION ON HOURLY BASIS KPI

ANDIKA HAIRUMAN[1], GEDE PUTRA KUSUMA[2]

[1,2]Computer Science Department,
BINUS Graduate Program – Master of Computer Science
Bina Nusantara University,
Jakarta, Indonesia, 11480

E-mail: [1]andika.hairuman001@binus.ac.id, [2]inegara@binus.edu

ABSTRACT

The most common used method for anomaly detection in the mobile radio network is using the fixed threshold on hourly and daily basis key performance indicator (KPI) and consider all the hours to have the same trend. The issues with the fixed threshold are false and miss detection. This paper proposed hourly basis KPI anomaly detection using machine learning techniques such as supervised learning and outlier detection and to measure the performance of a specific hour because of traffic profile and user behavior differences. The dataset was collected from mobile radio network and the ground truth was determined and labeled by network performance expert. There were 6 selected KPIs with 12096 total data samples including data for training, validation, and testing. 13 machine learning algorithms, 1 statistical technique and 7 data scalers were evaluated. The best performing technique is extra tree algorithm using standard and quantile transformer scaler. With extra tree algorithm, there were 4 missing detections and 7 false detections from 2418 total samples from data testing resulting impressive 97.11% of average F1 score from all 6 KPIs and 3 KPIs are having 100% F1 score. The evaluation result is proof that extra tree algorithm is very suitable for anomaly detection on mobile network hourly basis KPI data and it can significantly reduce the false and miss detection, alongside some general notion of which algorithm is suited for a certain type of KPIs.

Keywords: *Anomaly Detection, Machine Learning, Extra Tree Algorithm, Key Performance Indicator, Mobile Network*

1. INTRODUCTION.

When mobile radio network performance deviated, mobile data and voice services will be affected. The impact scale from the deviation would be vary. Performance deviation could affect a particular user, a specific mobile device, a cluster, or an area, or worse affecting the users on nationwide level. Network performance is measured by Key Performance Indicator (KPI). KPI is required to be monitored and optimized to provide high quality services and to obtain a greater resource utilization [1]. KPI deviation or anomaly always occurs in the mobile network due to some reasons such as network software upgrade, network maintenance, network configuration change or due to mobile device firmware issue.

Continuous monitoring on mobile radio network performance and detecting performance deviation can substantially improve the response time in handling the network problems which could affect end-user experience. The challenge of developing a scalable and resilient monitoring system that can handle data in real-time and at a massive scale is nontrivial [2]. Evaluating network performance is requiring huge efforts and many operators opted to passively monitor the network performance [3]. This research is based on a study from mobile network operator in Southeast Asia. The operator opted to monitor the performance in hourly and daily basis because it is less complicated than real-time monitoring. As a best practice, the operator always doing hourly performance monitoring post major night activities such as software upgrade to detect deviation or anomaly.

The ability to detect KPI anomaly in hourly basis with the impact on the end-users is increasingly challenging due to the increase of mobile network complexity such as the existence of many radio access technologies, new mobile

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

devices, and implementation of new radio network functions. The current method to detect anomaly on hourly basis KPI is to use a fixed threshold for each KPI and treating all the hourly performance to have the same level of performance. Some KPI values can be worse in comparison with other operators in certain geographical areas [4]. In many years, operators have followed the same method for traffic profiling by measuring the network performance using traffic patterns of a normal and busy day and traffic profiling can be done using hours of a whole day [5]. However, the operator observed that different areas and different hours would have a different performance profile. The operator realized that there is a need to change the way to monitor the hourly performance.

As for the fixed threshold, it is not every effective in detecting KPI anomaly because the network performance is very dynamic and big changes in the network is happening frequently. Thresholding mechanisms on KPIs are used to monitor the mobile radio network health and rank base station to identify the offenders [6], but this method is having inaccurate result causing false and miss detection of the anomaly. False detection can trigger unnecessary actions and miss detection can directly affect the end-users. The main driving of this research is to find a better method in improving the number of false and miss detection. Figure 1 shows the example of an anomaly in the KPI time series with the fixed threshold.

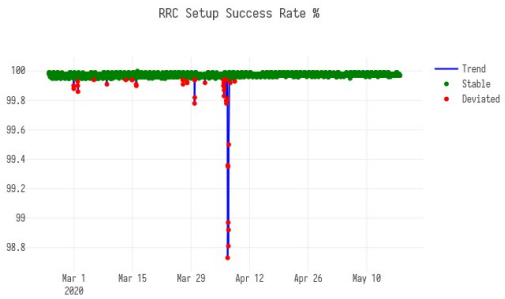

Figure 1: Illustration of an anomaly in hourly basis KPI with the common method using fixed threshold.

To overcome the shortcoming of fixed threshold, machine learning techniques offer the ability to better detect the KPI anomaly. Research conducted by Al Mamun et al. states that Supervised Machine Learning (SML) is applied to a set of long-term observation time series from a mobile network, and it has been shown that periodically collected KPIs can be analyzed by supervised ML and KNN algorithm is used to classify test data sets and able to solve major problems [7].

One of the key takeaways from research done by Omar et al. is the major difference between supervised and unsupervised learning is that supervised learning is done using a ground truth. It has the prior knowledge of what the output values for the data training and testing. This makes supervised learning the most suitable technique for anomaly detection on hourly basis KPI data [8].

There are many research topics about anomaly detection related with network traffic behavior, malicious attacks, and performance. This shows that the demand of finding the most suitable method for anomaly detection is still an interesting research topic. With the increase trend of automation and machine learning, there are many potential use cases of anomaly detection which can be addressed by new research and evaluation of machine learning techniques could provide an informative and useful inputs for the researcher.

Motivated by the mission of network performance management in innovating the method to better detect an anomaly and to reduce false and miss detection which could lead into efforts saving, this paper provided the model performance evaluation result and propose a method for hourly KPI monitoring. The evaluation result can help the operators to select the best machine learning algorithm and enable the organization to improve its method and detect the hourly basis anomaly of key performance indicators with the highest precision possible using the proven tested machine learning model.

2. RELATED WORKS.

Machine learning for anomaly detection has been studied and experimented by many academic research and industry. There are some finished and on-going works which could help to build the idea on improving the model.

Elmrabit et al. have evaluated the performance of 12 machine algorithms for anomaly detection which may be indicative of cyber-attacks [9]. Random Forest algorithm is the best performing algorithm among the techniques used in the study. Meanwhile, Naive Bayes algorithm has the lowest performance in terms of accuracy, precision, recall and AUC. The downside of their research is using accuracy as performance measurement. Accuracy is not a preferred indicator for anomaly detection specially if majority of data has no anomaly.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In 2022, Ahasan et al. studied supervised learning for anomaly detection on hourly and daily KPI data. Random Forest again becoming the best performing technique for hourly KPI data [10]. Accuracy is not the suitable matrix for evaluating machine learning algorithms for anomaly detection. The most popular and practiced one's are confusion matrix, precision, recall, F1-score, or AUC for evaluation. The Classification and Regression Trees (CART) algorithm achieved almost 100% accuracy and outperformed any other algorithms proposed by some of the literatures for heart disease prediction [11]. Ahasan et al. only pick 1 KPI while there are many KPIs in the network and having different profile such as different metric units, different trend and range. The result could have been different.

Amin et al. in their study mentioned that some key features and the best performing modelling techniques which improve the accuracy of heart disease prediction were selected. The best prediction model was created with the 9 significant parameters and with the Vote technique. The outcome of the benchmarking indicates that the classification model has produced a higher accuracy in prediction and performed better than the other studies [12]. As classification model shows a very good result, the author will use it as well to evaluate it against KPI data.

Ren et al. used the spectral residual (SR) model in the time-series anomaly detection and mix the SR and CNN model to achieve an outstanding performance. 81.1% F1-score is achieved when using combined SR and CNN model [13]. DeepAnT shows best AUCs score for most of the used data sets [14]. In novelty detection, OCSVM is considered the best method, but DeepAnT outperforms it. These results show that DeepAnT can find anomalies in multi-variant data set.

USAD (Unsupervised for Anomaly Detection) provides good performance in anomaly detection over multivariate time series while reducing training time by an average of 547 times with 69.01% F1 score using Orange Internal Dataset [15]. Unsupervised doesn't need a label and could save the efforts, the author will evaluate some of unsupervised techniques. OmniAnomaly provides interpretation based on the reconstruction probabilities of multivariate time series. The experiments are conducted on two public datasets from aerospace and a new server machine dataset from a service provider company. OmniAnomaly achieved 86% F1 score, significantly outperforming the best performing baseline method by 9% [16].

Multi-Scale Convolutional Recurrent Encoder-Decoder (MSCRED) able to detect all anomalies without any false positive and false negative. MSCRED performs best on all settings. The improvements over the best baseline range from 13.3% to 30.0%. MSCRED is much better than baseline methods as it can model both inter-sensor correlations [17]. Based on the research done by Schmidl et al, supervised learnings do not achieve superior or better results compared to semi supervised or even unsupervised approaches [18]. 71 algorithms and 976 datasets were used during evaluation. A clustering-based approach for anomaly detection in multivariate time series data performs very well in 3 datasets collected from US dollar exchange rate, EEG eye state, and air quality datasets [19]. Achieving 99% F1 score from 2 datasets.

Several anomaly detection techniques are tested by Geiger et al. and reported the best-suited one in their work. Their experimental results showed that an unsupervised anomaly detection approach built on Generative Adversarial Networks (TadGAN) outperformed all the baseline methods by having the highest averaged F1 score (70%) across all the datasets [20]. Adopt data augmentation for U-Net-DeW shows a very good result among other methods like ARIMA, SHESD and Donut. By adopting the loss adjustment and data augmentation, Gao et al. able to achieve F1 up to 81.2%, which is significantly better than the other state-of-the-art methods [21]. VAE-LSTM algorithm achieves 100% recall for all datasets, meaning no missed anomaly and the ability to detecting all types of anomalies. In one dataset, VAE-LSTM score 99.6% F1 score [22].

Based on the related works, the authors evaluate machine learning algorithms and data scaler for anomaly detection on hourly KPI data. Parameter optimization during the training would be required to be tested as needed. The datasets for KPI are not complicated and multi feature training is not required because individual KPI is not directly impacting other KPI.

3. RESEARCH METHODOLOGY.

The main goal of this paper is to evaluate the machine learning techniques for anomaly detection on hourly basis KPI by using the best performing scaler and to use datasets which divided into data training, data validation and data testing. The result will be concluded with confusion matrix to find the best performing technique. Figure 2 shows the conceptual framework of machine learning for

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

anomaly detection on hourly KPI data.

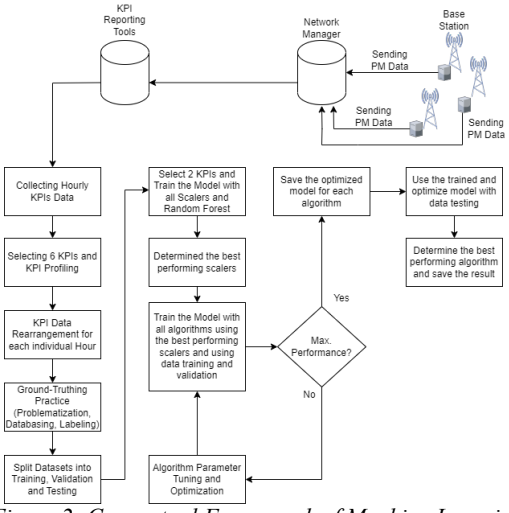

Figure 2: Conceptual Framework of Machine Learning for Anomaly Detection on Hourly Basis KPI Data.

To achieve the goal, the process started with PM data collection by network manager from all the base station. KPI reporting tools will download the raw PM files and to be processed and stored in the platform to produce the KPI report. Next step is to collect the hourly KPI data. There are many KPIs from the radio network, 6 KPIs will be selected. KPI profiling will have the information of KPI metric, number of anomalies in dataset, percentage of anomaly in dataset, normally high low or trend, standard deviation, median, minimum, and maximum value from data training and expected normal range of the data. After profiling completed, the KPI data will be separated by each individual hour. Dataset will be labeled by network performance expert and will follow the 3-dimensional conceptual morality guidance as ground-truthing practice. Dataset will then divide into training, validation, and testing.

Next step is to determine the best scaler by training the model using all selected scalers and 2 KPIs using Random Forest algorithm. At this level, the best performing scalers already determined. The next important thing is now to start training the model with all machine learning algorithms. The model will be trained until maximum performance is achieved. Parameter tuning and optimization will be performed in cycle to satisfy the performance target. Once maximum performance is achieved by the trained model. This trained model will be used to detect the anomaly on the data testing. The final step would be to evaluate and save the result and determine the best performing algorithm. The best performing algorithm expected to produce the result with very low false and miss detection. From the confusion matrix, at least 95% F1 is expected.

4. THEORY AND METHODS.

Recent studies concentrated on applying machine learning and statistical techniques for anomaly detection in a mobile network [23-27]. However, it is equally important to understand the method and how to pre-process the data before training the model.

4.1. Key Performance Indicator.

There are many KPIs in the mobile network. This paper discussed the specific KPIs for the LTE network [28]. The shortlisted KPIs will be used for anomaly detection testing with the proposed method. Each KPI represents the performance of the network and most of them can have a direct impact with the end-user.

Table 2 shows the list of the major LTE KPIs that are usually monitored by the mobile network operator.

Table 2: List of Major LTE KPIs

KPI	Metric	Normal Trend	Value Range
RRC Setup Success Rate	Percentage (%)	High	0-100
S1 Setup Success Rate	Percentage (%)	High	0-100
Initial ERAB Establishment Success Rate	Percentage (%)	High	0-100
Added ERAB Setup Success Rate	Percentage (%)	High	0-100
Intra-Freq Handover Success Rate	Percentage (%)	High	0-100
Inter-Freq Handover Success Rate	Percentage (%)	High	0-100
Intra LTE Inter Freq Cov. Trig. Success Rate	Percentage (%)	High	0-100
Intra LTE Inter Freq Lb. Success Rate	Percentage (%)	High	0-100
SRVCC Preparation Success Rate	Percentage (%)	High	0-100
DL Cell Throughput	Kilo bit per second	Trend	Trend
UL Cell Throughput	Kilo bit per second	Trend	Trend
PDCP UL Volume	Megabyte	Trend	Trend
PDCP DL Volume	Megabyte	Trend	Trend
VOIP Integrity	Percentage (%)	High	0-100
UL RLC ACK Success Rate	Percentage (%)	High	0-100
ERAB Drop Rate	Percentage (%)	Low	0-100
WCDMA Session Continuity Rate	Percentage (%)	Low	0-100
DL Latency	Millisecond	Low	Trend
UL Packet Loss Rate	Percentage (%)	Low	0-100

4.2. Issue with Fixed Threshold Detection

Ali et al. concluded a system can alter data inputs into meaningful and quantifiable anomaly scores. These scores are subsequently compared to a fixed detection threshold and categorized as either good or bad and a fixed threshold value cannot asssure good anomaly detection precision for such a time-varying input. Fixed threshold evaluates every group performance using same standards and because of this, it has issues and limitations.

The trend from fixed threshold will look like a normal trend and consistently overcomes the

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

threshold. It is very often generating bad anomaly detections as shown in Figure 3.

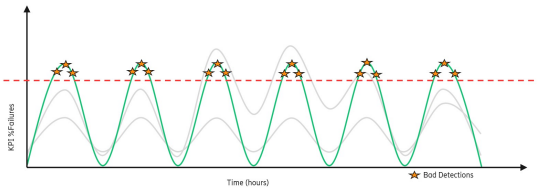

Figure 3: Fixed threshold generating bad anomaly detections.

Fixed threshold only able to partially detects the anomaly and missed anomaly detections can happen at any hours, not just busy or peak hour. Figure 4 shows how missed detections could happen during evaluation with fixed threshold. The worst part is that this miss detections are unable to detect the issue and operators may take long time to take actions in solving the deviated base station. Solution to this problem is needed and machine learning algorithm could be the answer.

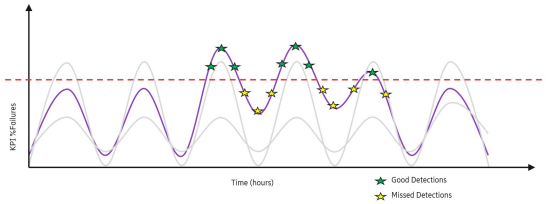

Figure 4: Fixed threshold missed detect the anomaly.

4.3. Scaler for Data Pre-processing.

Some data have very different scales and may have very high outliers. These types of data have different characteristics and it can deviate the performance of machine learning [29]. To solve this issue, data scaling is required. Scaling is a technique to normalize or standardize independent variables or features of data. Dataset often has multiple features with different degrees of magnitude, range and metric. This is a remarkable challenge as a machine learning algorithm is highly sensitive to these features. In this paper, the features contain KPIs and hour which need to be normalized before the model is trained. Feature scaling may significantly improve the performance of some machine learning algorithms but it may not work with some other algorithms.

Distance-based algorithms like K-means, KNN and SVM are the most impacted by the range of features. Machine learning that uses gradient descent as an optimization technique requires data to be scaled. For example linear regression, logistic regression and neural network. This is because the methods are using distances between data points to determine their similarity. On the other hand, tree-based algorithms are fairly insensitive to the scale of the features [30]. Figure 5 shows the difference between original data, the data after being normalized and the data after being standardized for illustration purposes.

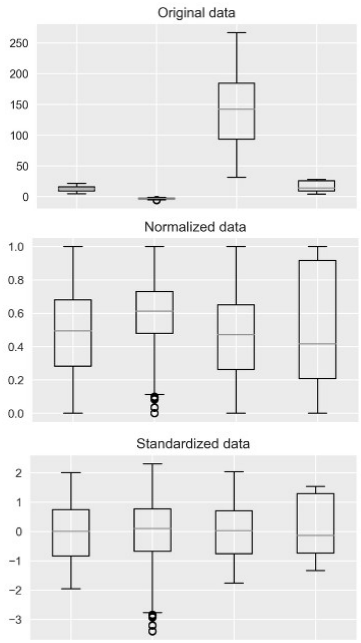

Figure 5: The difference between normalized and standardized data.

The authors will explore the performance of each scaler to achieve the highest result possible using the scikit-learn framework. The proposed scalers to be evaluated are Min Max Scaler, Max Abs Scaler, Standard Scaler, Robust Scaler, Normalizer, Quantile and Power Transformer.

4.4. Machine Learning for Anomaly Detection.

Machine learning is already being used in our day to day lives even though we may not be aware of it. Machine learning answers the question of how to model computers that improve automatically through experience [31]. One of the use cases of machine learning is to replicate what human does. Machine learning is being used to understand the data, to make the data make sense.

There are few machine learning techniques such as supervised learning, unsupervised learning, reinforcement learning and deep learning [32]. Supervised learning is the technique where learning

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

required a labeled training set. Unsupervised learning is the method of discovering patterns in unlabeled data. Reinforcement learning is learning based on feedback. In this paper, selected supervised and unsupervised learning will be tested and tune to achieve the best result possible in detecting hourly basis KPI anomaly with the proposed method. Table 3 shows the advantage of selected machine learning algorithms. The authors will test these selected algorithms.

Table 3: Advantage of Selected Machine Learning Algorithms.

Algorithm	Learning	Type	Advantage
KNN	Supervised	Classification	Easy implementation and robust with the search space
Random Forest	Supervised	Classification	Stable and robust to outliers and can manage them automatically
Decision Tree	Supervised	Classification	Does not require scaling and identify all possible results
Extreme Gradient Boost	Supervised	Classification	Works well with any data size and complicated data
SVM	Supervised	Classification	Good accuracy and use less memory
ADA Boost	Supervised	Classification	Easy to use, fewer tweaks and improve the accuracy of weak classifier
Bagging	Supervised	Classification	Reduce the variance, avoid overfitting
Extra Tree	Supervised	Classification	Faster and higher performance in a noisy feature
Gradient Boost	Supervised	Classification	Very good predictive accuracy with lots of flexibility
Histogram Gradient Boost	Supervised	Classification	It is experimental with improved speed
Local Outlier Factor	Unsupervised	Outlier Detection	Effective with the samples having distinct underlying densities
One Class SVM	Unsupervised	Outlier Detection	Effective unsupervised method and accurate when processing a big dataset
Isolation Forest	Unsupervised	Outlier Detection	More efficient on high-dimensional datasets

4.5. Statistical Technique for Anomaly Detection.

Seasonal Hybrid Extreme Studentized Deviate (SH-ESD), built upon the Generalized ESD and was specifically designed to detect anomalies. It can be used to detect both local and global anomalies. This two-step process allows this technique to detect global anomalies that extend outside the expected seasonal minimum and maximum and also detect local anomalies that would otherwise be masked by the seasonality. Figure 6 shows the anomalies detected in raw data using SH-ESD.

SH-ESD is a simple and robust statistic technique which is based on a method that can effectively detect anomalies in a mobile network. The algorithm was developed with very small processing power and limited storage, some cases should use statistical technique instead of more computationally expensive machine learning techniques, but this is depending on the dataset size and how frequent the use cases are. Based on SH-ESD testing using the data acquired from the actual live mobile network, it showed an improved detection capability compared with the existing basic failure detection methods. The authors tested the algorithm to predict the improved method.

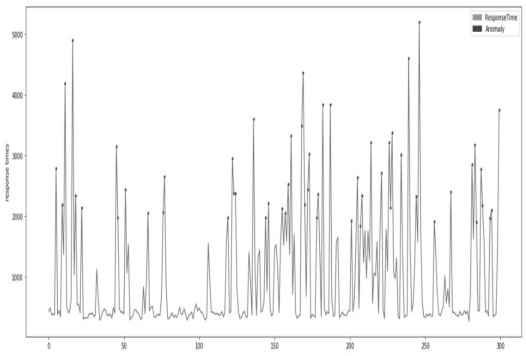

Figure 6: Anomalies present in the raw data and detected by SH-ESD.

4.6. Confusion Matrix.

This evaluation technique is frequently used to evaluate the performance of the tested model [33]. The confusion matrix will compare the ground truth with the result predicted by the trained model or by any methods. The results from the confusion matrix are accuracy, precision, recall and F1 score. It also provides the number of true positive, false positive, false negative and true negative. In this paper, true positive is showing the true anomaly, false positive represent false detection and false negative represent miss detection. The main objective is to find the best precision and recall, so the F1 score will be the criteria to measure the highest performance.

5. PROPOSED METHODS.

This paper proposed an improved method with the evaluation based on the individual hour in detecting hourly basis KPI anomaly using data scaler and machine learning for mobile network performance monitoring to have the highest precision possible.

5.1. Performance Evaluation based on Individual Hour.

The most used method is to evaluate the performance for the whole hours in the time series, but this method will trigger false and miss detection as each hour has a different traffic profile. Figure 7 shows how the current method detects the anomaly for the whole hours.

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

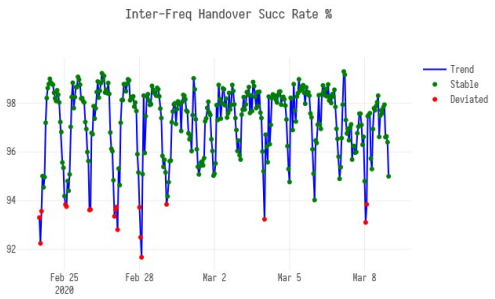

Figure 7: Example of performance evaluation based on the whole hours in the time series.

This paper proposed a performance evaluation based on the individual hour. The KPI performance of each individual hour will be compared with the same hour for a certain period. Figure 8 shows the KPI anomaly detection with the performance evaluation based on individual hour and there is a different pattern for each hour. The benefit of this method is to improve the number of true anomalies and reduce false and miss detection and it takes consideration of what is normal to that specific time of the hour.

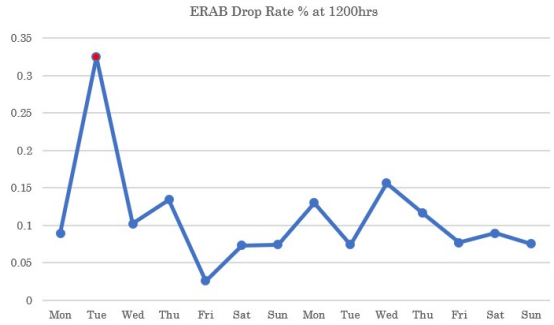

Figure 8: Example of performance evaluation based on the individual hour.

5.2. Shortlist KPIs.

Before collecting the data, KPIs need to be shortlisted for testing. Network performance expert has determined 6 LTE KPIs which represent the major indicator for network performance and has a distinctive characteristic. The shortlisted KPIs are RRC Success Rate (%), IFHO Success Rate (%), DL Throughput (kbps), UL Volume (MB), ERAB Drop Rate (%) and DL Latency (ms).

5.3. Determine Data Ratio for Training, Validation and Testing.

The authors preferred 60% of the collected data for training, 20% for validation and 20% for testing. Data training and validation will be used for parameter tuning during the experiment. Table 4 shows the data distribution. Once the model achieved the highest performance result, then it will be used to predict the result from data testing and every model will be evaluated using a confusing matrix function.

Table 4: Datasets distribution.

Data	Ratio Distribution	Per Hour Per KPI	Total Hours per KPI	Total Hours for 6 KPIs
Training	60%	50	1210	7260
Validation	20%	17	403	2418
Testing	20%	17	403	2418
Total	100%	84	2016	12096

5.4. Data Collection and Labeling.

The authors collected 2016 hours of hourly data from an actual live mobile network. Based on the data ratio, the first 1210 hours will be used for training (60%), the next 403 hours will be used for validation (20%) and the last 403 hours will be used for testing (20%). The collected data will be normalized before being labeled by the expert using the proposed method, evaluation based on the individual hour. Once the data has been normalized, the expert will proceed to label the anomaly for each KPI. Table 5 shows the summary of shortlisted KPIs profile after being labeled such as metric, number of anomalies, normally high or low, standard deviation, minimum and maximum value, and range.

Table 5: Profile of KPIs Dataset.

KPI	Metric Unit	Number of Anomalies in Dataset	Percentage of Anomaly in Dataset	Normally High, Low or Trend	Standard Deviation	Median	Min and Max Value from Data Training
RRCSR	%	305	15.12%	High	0.05	99.97%	98.73%-100%
IFHOSR	%	357	17.70%	High	2.22	97.99%	75.19%-99.65%
DLTHP	Kbps	441	21.87%	Trend	12193	16977	2047-47109
ULVOL	MB	505	25.04%	Trend	385135	881494	233595-3864037
ERABDR	%	459	22.76%	Low	0.05	0.11	0.02%-0.36%
DLLAT	ms	448	22.22%	Low	1.21	7.4	4.74-11.84

Imbalance data is expected for any case study related with machine learning for anomaly detection. Studiawan et al. performed study in 2020 about imbalance data and results from their study demonstrate that by taking data imbalance into consideration, there is an improvement in the method performance in terms of precision and recall scores [34].

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

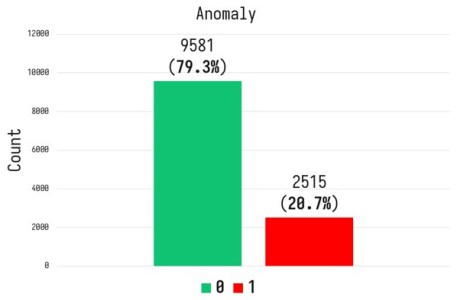

Figure 9: Imbalance Ratio of Hourly KPI Data

For this research, imbalance ratio is around 20.7% between anomalies and non-anomaly as shown in Figure 9. This imbalance ratio expected to produce a very good performance.

5.5. Evaluate Data Scaler Performance.

This step is very important to select the optimal data scaler for training the model. For data scaler testing, the authors will use the random forest to test some known scalers such as Min-Max Scaler, Max Abs Scaler, Standard Scaler, Robust Scaler, Normalizer, Quantile Transformer and Power Transformer. DL cell throughput and ERAB drop rate KPI will be tested and will use only data training and validation. Table 6 and 7 shows the scaler testing result and then the authors concluded which scaler to be used for training.

Table 6: Scaler Performance on KPI ERAB Drop Rate.

KPI	Scaler	Parameter	Algorithm	Precision	Recall	F1 Score
ERAB Drop Rate	Standard Scaler	Default	Random Forest	100.00	100.00	100.00
ERAB Drop Rate	Robust Scaler	Default	Random Forest	100.00	100.00	100.00
ERAB Drop Rate	Max Abs Scaler	Default	Random Forest	100.00	100.00	100.00
ERAB Drop Rate	Min Max Scaler	Default	Random Forest	100.00	100.00	100.00
ERAB Drop Rate	Quantile Transformer	n_quantiles=1210, random_state=100	Random Forest	100.00	100.00	100.00
ERAB Drop Rate	Power Transformer	Default	Random Forest	93.10	100.00	96.42
ERAB Drop Rate	Normalizer	Default	Random Forest	79.16	70.37	74.50

Table 7: Scaler Performance on KPI DL Throughput.

KPI	Scaler	Parameter	Algorithm	Precision	Recall	F1 Score
DL Cell Throughput	Max Abs Scaler	Default	Random Forest	79.48	100.00	88.57
DL Cell Throughput	Min Max Scaler	Default	Random Forest	79.48	100.00	88.57
DL Cell Throughput	Power Transformer	Default	Random Forest	79.48	100.00	88.57
DL Cell Throughput	Quantile Transformer	n_quantiles=1210, random_state=100	Random Forest	90.26	98.38	88.40
DL Cell Throughput	Standard Scaler	Default	Random Forest	78.48	100.00	87.94
DL Cell Throughput	Robust Scaler	Default	Random Forest	78.48	100.00	87.94
DL Cell Throughput	Normalizer	Default	Random Forest	29.50	58.06	39.13

Based on the scaler test result, DL cell throughput KPI has the best result from quantile transformer and the metric for this KPI is kbps, not a percentage. The authors decided to proceed with quantile transformer for any other KPIs which is not categorized as a percentage metric such as UL volume and DL latency. ERAB drop rate has the best result from most of the scalers except normalizer and power transformer. The authors decided to proceed with the standard scale for the rest of the percentage KPIs like the RRC success rate and the Inter Frequency HO success rate. As the authors stated above, some scalers may not affect some of the machine learning algorithms. There is an opportunity for future research, a comprehensive scaler test can provide a broader view about the impact on the result from a trained model.

5.6. Evaluate Machine Learning Model.

The next step is to train the model based on data training and validation and extensive parameter tuning has been done to find the best result. The authors performed the experiment using a scikit-learn library except for xgboost. Histogram gradient boost still under experimental research but it is added to see if the algorithm can show some good results. The authors only use data training and validation to evaluate the performance of the trained model.

Data testing will only be tested after the authors concluded the model is selected for testing. Table 8 shows some of key parameters which will be tested, and the optimal parameters will be used for final testing. There are many parameters to be optimized, the authors may not test all the possible parameters.

Table 8: Parameters to Optimized

Parameter	Description
contamination	the proportion of outliers in the data set between 0 and 0.5
criterion	the function to measure the quality of a split
kernel	specify the kernel to precompute the kernel matrix
max_depth	The maximum depth of the tree
metric	metric to use for distance computation
n_estimators	number of estimators
n_neighbors	number of neighbors
weights	weight function used in prediction

The pros of machine learning using performance evaluation based on individual hour are no threshold is determined and detections will be made based on its own hourly historic trend, and it takes into consideration what is normal to that specific time of the day.

The cons of supervised learning are requiring effort in labelling the data and it really needs so much time. While unsupervised, the result might not be good because it cannot tell what is good or bad.

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. MAIN RESULTS.

6.1. Unsupervised Learning Testing Result.

Isolation Forest is the best performer on data testing with 24.18% F1 score. This is expected as per the result on data validation. However, it doesn't meet the expectation to have at least 95% F1 score for anomaly detection based on hourly KPI monitoring. LOF performed slightly better compared with data validation, and One Class SVM performed worse than data validation result. Based on this result, unsupervised learning is not suitable for KPI hourly monitoring. The testing results are shown in Table 9 below.

Table 9: Performance of Unsupervised Learning on Data Testing.

Unsupervised Algorithm	Precision	Recall	F1	TP	FP	FN
Isolation Forest (Un)	14.19	81.82	24.18	162	980	36
LOF (Un)	12.70	47.98	20.08	95	653	103
One Class SVM RBF (Un)	6.27	36.87	10.72	73	1091	125

During data validation, it has been observed that the recall performance is better than precision. Scatter plot in Figure 10 below shows the observation.

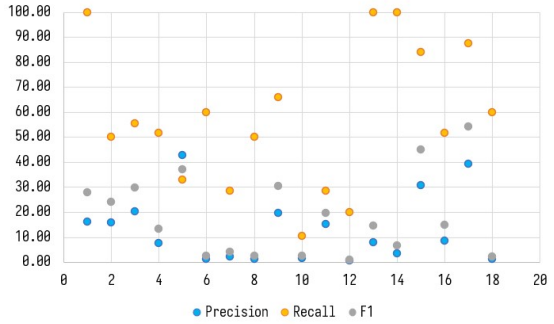

Figure 10: Scatter Plot Performance of Unsupervised Learning on Data Validation

6.2. Supervised Learning Testing Result.

In total, there are 3 supervised techniques having F1 score exceeding 95%. Extra tree algorithm is the best performing technique based on the data testing result. There are only 4 missing detections and 7 false detections from 2418 samples of data testing resulting impressive 97.24%

F1 score from all 6 KPIs and 3 KPIs are having 100% F1 score. Extra tree algorithm is proven to work well with the proposed method of measuring hourly KPI. There are 3 KPIs with a 100% F1 score. 2 KPIs with F1 score around 92% and 1 KPI with an F1 score around 98%. KNN is having the best precision score 97.37%, but KNN recall performance is less than 95%.

The lowest performance technique is SVM RBF with 31.25% F1 score. Based on this final result, supervised learning is very suitable for anomaly detection on hourly KPI data. Table 10 shows the performance of supervised learning on data testing.

Table 10: Performance of Supervised Learning on Data Testing.

Algorithm	Precision	Recall	F1	TP	FP	FN
Extra Tree Classifier	96.52	97.98	97.24	194	7	4
KNN	97.37	93.43	95.36	185	5	13
GB Classifier	94.09	96.46	95.26	191	12	7
Hist GB Classifier	92.72	96.46	94.55	191	15	7
Random Forest	92.42	92.42	92.42	183	15	15
XG Boost	90.77	89.39	90.08	177	18	21
Bagging Classifier	88.61	90.40	89.50	179	23	19
ADA Boost	84.62	88.89	86.70	176	32	22
Decision Tree	81.06	92.93	86.59	184	43	14
SVM RBF	40.98	25.25	31.25	50	72	148

6.3. Comparison Between Best Performing ML Technique and Statistical Technique.

SH-ESD doesn't bring a good result compared with machine learning technique. The main reason likely due to the sensitivity with the scaling and like unsupervised learning, both methods cannot tell the correct or incorrect one, or good one, bad one. It can detect the outliers, even a very good KPI can be considered as outliers by these methods. Both methods could produce the better result if it is combined with ruled-based policy.

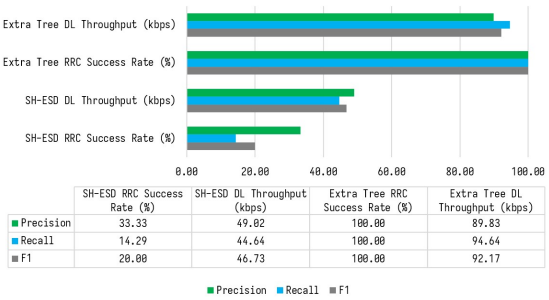

Figure 11: Performance Comparison between Extra Tree and SH-ESD on Data Testing.

SH-ESD used a usual mean or with mean absolute deviation. 2 KPIs are selected for the comparison based on extra tree result, 1 KPI is selected because it has 100% F1 performance from data testing, which is RRC SR KPI and the other 1 KPI is DL cell throughput as it has 92.17% F1

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

score.

SH-ESD F1 scores on these 2 KPIs are less than 50%. Based on this result, supervised learning is very superior compared with SH-ESD.

The comparison between extra tree and SH-ESD is explain in Figure 11. Detection with SH-ESD also using the approach where each hour having a different trend.

6.4. Comparison Between Best Performing ML Technique and Fixed Threshold.

With the fixed threshold, there are very high numbers of false and miss detection as shown in Table 11 below and this is expected. 692 false detection and 109 miss detection from fixed threshold with all hours treated with the same trend. When fixed threshold is used with a method where each hour has different trend, the detection is improved specially the recall performance which achieved 94.44% score. With machine learning using extra tree, the number of false and miss detection is further reduced to 7 false detection and 4 miss detection using Extra Tree. Based on this comparison, it concludes the machine learning technique with supervised method able to outperform fixed threshold for hourly KPI monitoring when using an improved monitoring method where each hour has individual trend. For comparison purpose, the fixed thresholds are calculated using mean.

Table 11: Performance Comparison between Fixed Threshold and Extra Tree.

Anomaly Detection Method	Hourly Monitoring Method	Precision	Recall	F1	TP	FP	FN
Fixed Threshold	All Hours Same Trend	11.40	44.95	18.18	89	692	109
Fixed Threshold	Each Hour Different Trend	33.27	94.44	49.21	187	375	11
Machine Learning with Extra Tree	Each Hour Different Trend	96.52	97.98	97.24	194	7	4

In the chart shown in Figure 12, 13 and 14 below, it is very clear the difference on how anomaly is poorly detected with fixed threshold and how machine learning can have a better result in detecting anomaly. Chart below is the representation of one of the 6 KPIs.

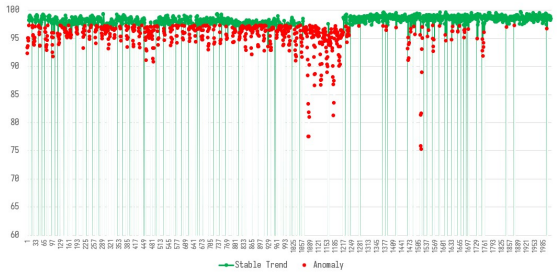

Figure 12: Anomaly Detection with Fixed Threshold on Dataset All Hours Same Trend.

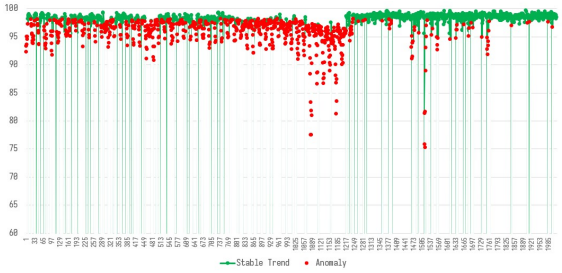

Figure 13: Anomaly Detection with Fixed Threshold on Dataset Each Hour has Different Trend.

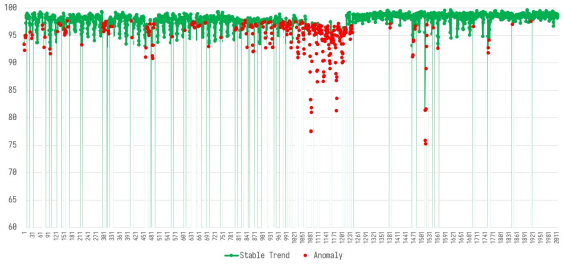

Figure 14: Anomaly Detection with Machine Learning using Extra Tree on Dataset Each Hour has Different Trend.

7. CONCLUSIONS.

Hourly KPI monitoring can be improved by comparing the performance of each individual hour and machine learning technique outperform the fixed threshold significantly. Number of false detections (false positive) is significantly reduced when using supervised learning for anomaly detection on hourly KPI data. Both are using the same approach of comparing the performance of individual hour. The number of miss detection (false negative) and good detection (true positive) is slightly improved with machine learning using extra tree. The precision performance of machine learning using extra tree outperforms the fixed threshold. This also concludes hourly KPI monitoring by treating all hours to have the same trend is not good in detecting anomaly and monitoring the performance of individual hour is significantly improved the anomaly detection.

Based on the evaluation result, supervised learning is very suitable for the anomaly detection on hourly KPI data and there are 3 supervised algorithms which having a superior F1 score above 95%, extra tree, KNN and GB. Extra tree is the best algorithm based on this evaluation and achieved 97.24% F1 score. The worst performer from machine learning technique is SVM RBF. Unsupervised and SH-ESD is not performing well to detect anomaly on hourly KPI data.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Mobile network is frequently change and upgraded. The trained model probably will not be effective in detecting anomaly after a new software upgrade. It requires to train the model every major update in the network. A new method is required so a model can be used at least for one year duration. Generating a dataset which could forecast the next one-year performance and adding it into the training data could potentially be the solution.

Training a model requires an effort while unsupervised learning doesn't require much effort. A combination between unsupervised learning or statistical technique like SH-ESD, z-score, histogram with rule-based policy method might have a better result and efficiency than supervised learning to detect anomaly.

AUTHOR CONTRIBUTIONS

Both authors contributed equally.

REFERENCES

[1] Krasniqi, F., Maraj, A., & Blaka, E. (2018, September). Performance analysis of mobile 4G/LTE networks. In 2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM) (pp. 1-5). IEEE.

[2] Zhao, S., Chandrashekar, M., Lee, Y., & Medhi, D. (2015, March). Real-time network anomaly detection system using machine learning. In 2015 11th international conference on the design of reliable communication networks (drcn) (pp. 267-270). IEEE.

[3] Suzuki, M., Plessis, Q., Kitahara, T., & Ano, S. (2016, March). Monitoring communication quality degradation in LTE network using statistics of state transition. In 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA) (pp. 33-38). IEEE.

[4] Yigit, I. O., Ayhan, G., Zeydan, E., Kalyoncu, F., & Etemoglu, C. O. (2017, November). A performance comparison platform of mobile network operators. In 2017 8th international conference on the network of the future (NOF) (pp. 144-146). IEEE.

[5] Ozovehe, A., Okereke, O. U., & Anene, E. C. (2015). Literature survey of traffic analysis and congestion modeling in mobile network. IOSR Journal of Electronics and Communication Engineering Volume, 10(6), 31-35.

[6] Leontiadis, I., Serra, J., Finamore, A., Dimopoulos, G., & Papagiannaki, K. (2017, April). The good, the bad, and the KPIs: how to combine performance metrics to better capture underperforming sectors in mobile networks. In 2017 IEEE 33rd International Conference on Data Engineering (ICDE) (pp. 297-308). IEEE.

[7] Al Mamun, S. A., & Valimaki, J. (2018). Anomaly detection and classification in cellular networks using automatic labeling technique for applying supervised learning. Procedia Computer Science, 140, 186-195.

[8] Omar, S., Ngadi, A., & Jebur, H. H. (2013). Machine learning techniques for anomaly detection: an overview. International Journal of Computer Applications, 79(2).

[9] Elmrabit, N., Zhou, F., Li, F., & Zhou, H. (2020). Evaluation of Machine Learning Algorithms for Anomaly Detection. 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security).

[10] Ahasan, M. R., Haque, M. S., & Alam, M. G. (2022, July). Supervised Learning based Mobile Network Anomaly Detection from Key Performance Indicator (KPI) Data. In 2022 IEEE Region 10 Symposium (TENSYMP) (pp. 1-6). IEEE.

[11] Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. Technologies, 9(3), 52.

[12] Amin, M. S. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.

[13] Ren, H., Xu, B., Wang, Y., Yi, C., Huang, C., Kou, X., & Zhang, Q. (2019, July). Time-series anomaly detection service at microsoft. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, (pp. 3009-3017).

[14] Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S. (2018). DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. Ieee Access, 7, 1991-2005.

[15] Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020, August). Usad: Unsupervised anomaly detection on

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

multivariate time series. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, (pp. 3395-3404).

[16] Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., & Pei, D. (2019, July). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, (pp. 2828-2837).

[17] Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., & Chawla, N. V. (2019, July). A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In Proceedings of the AAAI conference on artificial intelligence, 33, pp. 1409-1416.

[18] Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: a comprehensive evaluation. Proceedings of the VLDB Endowment, 15(9), 1779-1797.

[19] Li, J., Izakian, H., Pedrycz, W., & Jamal, I. (2021). Clustering-based anomaly detection in multivariate time series data. Applied Soft Computing, 100, 106919.

[20] Geiger, A., Liu, D., Alnegheimish, S., Cuesta-Infante, A., & Veeramachaneni, K. (2020, December). TadGAN: Time series anomaly detection using generative adversarial networks. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 33-43). IEEE.

[21] Gao, J., Song, X., Wen, Q., Wang, P., Sun, L., & Xu, H. (2020). Robusttad: Robust time series anomaly detection via decomposition and convolutional neural networks. arXiv preprint arXiv:2002.09545.

[22] Lin, S., Clark, R., Birke, R., Schönborn, S., Trigoni, N., & Roberts, S. (2020, May). Anomaly detection for time series using vae-lstm hybrid model. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4322-4326). Ieee.

[23] Novaczki, S., An Improved Anomaly Detection and Diagnosis Framework for Mobile Network Operators, in 9th International Conference on the Design of Reliable Communication Networks (DRCN), pp. 234-241, 2013.

[24] Ciocarlie, G.F., Lindqvist, U., Novaczki, S. & Sanneck, H., Detecting Anomalies in Cellular Networks Using Ensemble Method, in 9th International Conference on Network and Service Management (CNSM) Zürich, pp. 171-174, 2013.

[25] Chernov, S., Cochez, M. & Ristaniemi, T., Anomaly Detection Algorithms for the Sleeping Cell Detection in LTE Networks, in IEEE Vehicular Technology Conference (VTC Spring), pp. 1-5, 2015.

[26] Pablo, M., Barco, R., Serrano, I. & Gómez-Andrades, A., Correlation-Based Time-series Analysis for Cell Degradation Detection in SON, IEEE Communications Letters, 20(2), pp. 396-399, 2016.

[27] Novaczki, S. & Szilagyi, P., Radio Channel Degradation Detection and Diagnosis Based on Statistical Analysis, in 73rd IEEE Vehicular Technology Conference (VTC Spring), pp. 3158-3159, 2011.

[28] Kreher, R., & Gaenger, K. (2015). Key Performance Indicators and Measurements for LTE Radio Network Optimization.

[29] Raghav, R., Lemaitre, G., & Unterthiner, T. (2018). Compare the effect of different scalers on data with outliers. Retrieved, 9(2), 2019.

[30] Bhandari, A. (2020). Feature scaling for machine learning: Understanding the difference between normalization vs. standardization. Analytics Vidhya.

[31] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[32] Stinis, P. (2019). Enforcing constraints for time series prediction in supervised, unsupervised and reinforcement learning. arXiv preprint arXiv:1905.07501.

[33] Sammut, C., & Webb, G. I. (2017). Encyclopedia of machine learning and data mining. Springer Publishing Company, Incorporated.

[34] Studiawan, H., & Sohel, F. (2020, July). Performance evaluation of anomaly detection in imbalanced system log data. In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4) (pp. 239-246). IEEE.

[35] Ali, M. Q., Al-Shaer, E., Khan, H., & Khayam, S. A. (2013). Automated anomaly detector adaptation using adaptive threshold tuning. ACM Transactions on Information and System Security (TISSEC), 15(4), 1-30.