Long Term Evolution Anomaly Detection and Root Cause Analysis for Data Throughput Optimization

Simon Mbogo
School Computing and Informatics
University of Nairobi

Evans Miriti
School of Computing and Informatics
University of Nairobi

ABSTRACT
There is a growing demand for mobile network data services which is driven by high number of smartphones and web content search as multimedia. Network operators have tried to provide enough capacity and meet the data speeds that the customer needs. This has led to introduction of new technology and expansion of the mobile networks making them complex to manage. Detecting anomalies that affect data throughput and investigating the root causes in mobile networks is challenging as mobile environments are increasingly complex, heterogeneous, and evolving.

There is need to automate network management activities to improve network management processes and prevent revenue loss. Self-Organizing network is a standard introduced by third Generation Partnership Program (3GPP) to automate network management. However, the standard is still not fully developed.

This project focused on implementing an anomaly detection and root cause analysis model that helps in the process of data throughput optimization in Long-term evolution (LTE) networks. The model used Density Based Spatial Clustering of Applications with Noise (DBSCAN) for anomaly detection, K-Nearest Neighbour (KNN) for root cause analysis and real network performance data from a Kenyan operator.

Proposed anomaly detection model achieved a silhouette coefficient of 0.451 showing a good separation of existing clusters in the dataset and was able to detect anomalies with both positive and negative impact on data throughput. The root cause analysis model achieved an accuracy of 94.59% and was able to identify the root cause of detected anomalies that had a negative impact on data throughput.

Keywords
Density based spatial clustering of applications with noises (DBSCAN), K-Nearest neighbour (KNN), Long-term evolution (LTE), throughput

1. INTRODUCTION
Despite mobile network operators' race to create better and larger networks to meet the mobile data crunch, data demand is growing at a faster rate than the carriers can keep up with. As the mobile technology improves, customers use more data and the faster the speeds provided by the mobile operators, the more consumers swallow it up and demand more. That is great for carriers because data usage is monetized, but it presents major challenges in network maintenance and optimization. The additional revenue is great, but if speeds degrades, customers look for other service providers that can provide a more reliable connection [5]. As such, providing a reliable and fast network connection is becoming increasingly crucial to acquisition and retention of customers. However, the mobile network is complex and many anomalies affecting a network connection are difficult to detect, diagnose and resolve. Some network anomalies like transmission faults, traffic shifts, capacity constraints and network availability have a direct impact on revenue and customer experience and requires modern detection and root cause analysis approaches.

1.1 Related Work
Machine learning has been recognized as a powerful tool to bring intelligence into network and to achieve self-healing. Nevertheless, there have been major challenges identified regarding its practical applications for self-healing which includes data insufficiency, data imbalance, non-real-time response and cost insensitivity [1]. Other challenges are limitations of chosen algorithms.

One commonly used approach for anomaly detection and root cause analysis in mobile network is use of performance counters to predict the value of a metric and compare it with the real value to determine the root cause or detect an anomaly. In [3], random forest was used to detect LTE resource consumption anomalies and neural network for root cause analysis. He used random forest technique known as Gini importance or mean decrease impurity to rank performance counters used to predict resource consumption and quantile regression forest which is a modification of random forest to predict the resource consumption in LTE. Predicted resource consumption is compared with real consumption and based on the relationship, its determined whether there is an anomaly in the data. A self-organizing cellular radio access (SORA) network system, enhanced with deep learning was proposed in [1]. For anomaly detection, SORA predicts anomaly or normal cell state based on the current state of collected critical KPIs. Critical KPIs are identified with the help of an expert while prediction is based on CNN+convLSTM algorithms. After predicting an anomaly, a marked point is passed to the root cause analysis component which uses both supervised classification (auto-encoder+decision tree) and unsupervised clustering (auto-encoder+agglomerative clustering).

Another approach is to cluster cells based on the correlation between performance counters and fault causes. In [4], a system based on self-organizing map (SOM) to automatically identify LTE faults was proposed. The system has two phases which are design and exploitation phases. At design stage, the system automatically identifies several clusters that represents different faults. An approach based on the analysis of the statistical behavior of each cluster to obtain relevant clusters and label them properly was proposed. Cells were clustered based on the correlation between cell KPIs and fault causes, and performance evaluation was done using silhouette coefficient. At the first stage which is the system diagnosis, obtained clusters had to be labeled by an expert with the identified causes. In the second phase, exploitation phase,

SOM is used to classify the cell's state according to the behavior of its symptoms and subsequently identifying the fault cause. SOM was also used in [7] for both anomaly detection and root cause analysis for video-on-demand (VoD) cloud service. In [7], both supervised and unsupervised versions of Self Organizing Maps (SOM) were used and compared to ascertain which performs better than the other. Supervised learning approach which made use of available labelled data performed better than unsupervised learning approach. The advantage of using unsupervised learning approach is that it can detect previously unknown anomalies or identify previously unknown root causes.

One disadvantage of the two approaches used for anomaly detection and root cause analysis is that they require support of domain experts, who are scarce and expensive, in coming up with related counters and correlation counters to anomalies or root causes.

1.2 Proposed Solution

Proposed solution eliminates the need for an expert in anomaly detection phase by using the key metric to detect anomalies. This is done through use of a clustering algorithm, DBSCAN, and time series performance data which is organized in such a way that a change in trend or an outlier is detected marked as an anomaly. Once an anomaly is detected, a classification algorithm, KNN, is used in root cause analysis. In this phase, domain expert input is required to support in feature selection after which the model is trained with labelled data from a Kenyan operator. Visualizations are then generated to provide a clear view of anomalies and indicators of root cause and aid in resolving the anomalies.

1.3 PURPOSE

The goal of this study was to detect and diagnose the root cause of network anomalies in the radio access network.

2. METHODOLOGY

2.1 Data Collection

Long term evolution (LTE) historical performance data was extracted from a Kenyan Operator's database and shared folders after seeking necessary approvals.

2.2 Data Pre-processing

Python programming language was used in data pre-processing and model design. Anaconda distribution was used as a Python data science platform in this project.

2.2.1 Missing data

Base transceiver station (BTS) cell samples in the data with missing values, invalid entries like NA and negative numeric value were dropped.

2.2.2 *Data anonymization*

Anonymization was applied to direct identifiers of the chosen operator's cell names. The cell names of collected data was divided into two main parts, site name part and cell id part separated by a hyphen, and the two parts anonymized to form a cell name that had the format "siteX_Cell_Y".

2.2.3 Feature selection

Feature selection was done using domain knowledge of the problem at hand, which brought down the feature set to 6 features. These features are: throughput, data traffic, packet retransmission, downlink channel quality index (DL CQI), physical radio bearer (PRB) Utilization and Transmission radio network layer KPI. Throughput is the key metric for anomaly detection while the others are the main features that affect throughput in LTE.

2.2.4 Data normalization

Data normalization was done to avoid problems when training models with features of wildly varying ranges. Normalization was performed on data set used in root cause analysis. To normalize the feature data to the range [0,1], min-max normalization technique below was used.

$$zi = \frac{xi - \min(x)}{\max(x) - \min(x)}$$

Where x is x_1, x_2, x_n and zi is the ith normalized data.

2.3 Model Training

DBSCAN was used to detect anomalies and KNN used to identify the root cause of detected anomalies.

2.3.1 DBSCAN model training

Four weeks' network cluster data, 5426 rows of data, with throughput as the only metric was organized in such a way that each cell formed a data point while each day's average throughput formed a feature/dimension. With N number of cells in a network cluster (i.e. N time series) and four weeks daily average data, there was an N x 28 data array. This data was fitted to the model for training. Cells labeled as noise were considered anomalous. To determine epsilon, minimum distance between each data point to other data points was calculated. The minimum distance under which majority of the data points lie was chosen as the Epsilon. Chosen epsilon was used to calculate the number of neighbors for each data point and the value under which the neighbors were considered to be few was chosen as the minimum point parameter.

2.3.2 KNN model training

Detected anomalous cells were passed to the last module for root cause analysis. 7066 data points acquired from a Kenyan operator were used to train the model. The features used in KNN model training were data traffic, packet retransmission, downlink channel quality index (DL CQI), physical radio bearer (PRB) Utilization and Transmission radio network layer KPIs. Training data was extracted from labelled performance data and divided into training (80%) and testing set (20%). Cross validation was used to determine the best value of k. KNN algorithm was applied to the training set and validated with the test set. A model was created using different values of k to predict the test data and then check the accuracy. The value of k that gave the maximum accuracy was chosen.

1.1 Performance Evaluation and Reporting the Performance

Silhouette coefficient was used for clustering phase and since previous information about the data set was available, accuracy was used for the root cause analysis phase.

3. RESULTS, DISCUSSIONS AND CONCLUSION

3.1 Results

3.1.1 Detecting anomalies using dbscan algorithm

Using technique explained in DBSCAN model training to choose parameters, two histograms were generated (see figures 1 and 2).

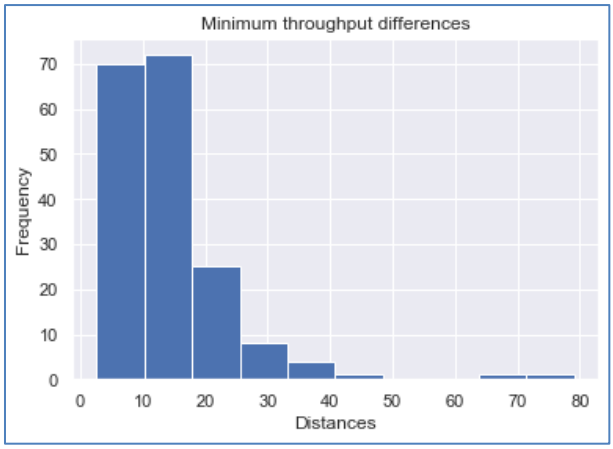

Figure 1: Minimum throughput differences

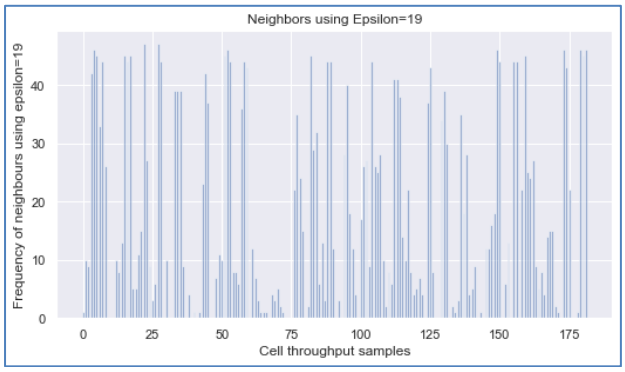

Figure 2: Neighbors using chosen epsilon

Figure 1 indicated that the majority of points lie within the first two bins. A matplotlib function was used to get the value of the left-hand edge of the third bin which was 17.88. The number of points that lie within each point's epsilon-neighborhood was then calculated. Figure 2 indicated that most points had more than ten (10) neighbors. Different values of epsilon within the neighborhood of 17.88 and different values of minimum points within the neighborhood of 10 were used as DBSCAN parameters and silhouette index calculated. The value Nineteen (19) as the epsilon and a value of ten (10) as the minimum points parameter produced the best value of silhouette index and were used as DBSCAN parameters.

Applying an epsilon of 19 and minimum points of 10 for the dataset, two clusters were generated, and 53 anomalous cells detected. The first cluster had a throughput range from 1Mbps to 20Mbps while the second cluster had a throughput range from 16Mbps to 34Mbps. The two clusters showed a separation between cells performance on average user throughput in terms of number of users, allocated channel bandwidth and channel quality. The cluster with higher throughput indicates cells with smaller number of users, larger bandwidth or good channel quality. On the other hand, cluster with lower throughput indicates cells with high number of users, smaller bandwidth or poor channel quality.

3.1.2 Evaluating the model

To evaluate the quality of the model's prediction, silhouette coefficient was calculated. Silhouette Coefficient ranges from -1 to +1. The larger the value is, the better the clustering effect is [6]. A silhouette coefficient of 0.451 was achieved which indicated a good separation between the two clusters. A cell that did not fit any of the two clusters was labelled as an anomaly.

3.1.3 Detected anomalies

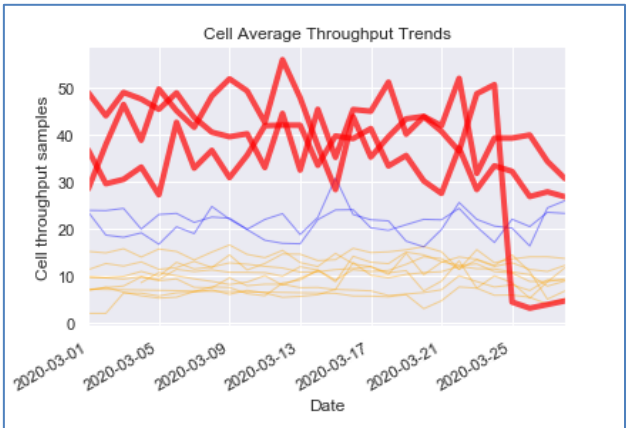

Figure 3: Detected anomalies 1

Figures 3 shows cells with anomalies (in red). Two of the cells were outliers with a very high throughput compared to other cells. Unlike the cells in the two clusters, the two cells' throughput varied with large values from one day to the next. The third cell had a step reduction in throughput on 25th March which was due to a fault on site.

3.1.4 Root cause analysis

Labelled data acquired from a Kenyan operator was distributed as in figure 4 below.

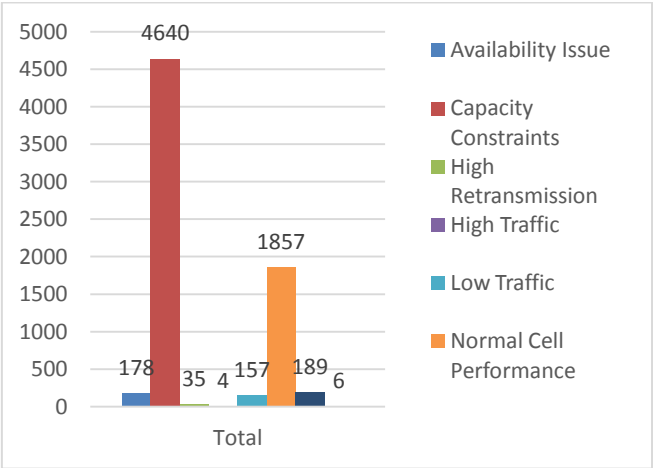

Figure 4: Labelled data distribution

Figure 4 shows that majority of the data points in acquired labelled data had a label of capacity constraints. This is indicative of the high data demand in the network leading to capacity constraints. This was followed by normal cell performance, poor channel quality, availability issue, low traffic, high retransmission, transmission fault and high traffic in that order. Data points labelled normal cell performance were less than those labelled capacity constraints. The reason is because thresholds are used to identify the cells that require further investigations by Engineers in manual optimization process. Data points labelled transmission fault were few indicating a stable network with few transmission issues.

Labelled data was divided into two sets i.e. training set and test set (80:20 ratio). The only parameter in K-nearest neighbor, parameter k, was determined through cross-validation. Six different values of k were used, and accuracy calculated.

Table 1: Cross-validation to determine k

k Parameter Value	Achieved Accuracy (%)
3	96.4084
5	99.9616
7	99.733
9	99.943
11	99.8654
13	96.2354

Table 1 above shows the result of using different values of k. The value of k = 5, that resulted in the best accuracy was chosen.

3.1.4.1 Performance evaluation for root cause analysis model

The training data was shuffled and then split into 5 and 10 groups for a 5-Fold and 10-Fold cross validation. Each unique group was used as hold data while the other groups were used as training data. This ensured that data used for training and testing were non-overlapping and thereby reporting unbiased test results. The model was then fit with the training data and evaluated using the test data generating tables 2 and 3.

Table 2: 5-fold cross validation

Folds	Accuracy (%)
Fold 1	94.34
Fold 2	94.76
Fold 3	94.62
Fold 4	94.62
Average Accuracy	**94.59**
Standard Deviation	**0.15**

Table 3: 10-fold cross validation

Folds	Accuracy (%)
Fold 1	94.48
Fold 2	94.9
Fold 3	94.62
Fold 4	94.19
Fold 5	94.48
Fold 6	94.9
Fold 7	94.48
Fold 8	94.33
Fold 9	94.05
Fold 10	95.04
Average Accuracy	**94.55**
Standard Deviation	**0.31**

Applying the 5-fold cross validation, the model had an accuracy of 94.59% while the 10-fold cross validation had an accuracy of 94.99%. This shows good quality of the model predictions. The standard deviation for 5-fold cross validation was 0.15% while that of 10-fold cross validation was 0.31%. Both had a low standard deviation indicating stable model performance and thus, a good model performance.

3.1.4.2 Applying the model on new data

After choosing five (5) as the value of k, anomalous cells performance data was used as new data in the root cause analysis module. Out of the fifty-three anomalous cells, forty-six (46) cells were predicted to have normal performance while seven (7) were predicted to have negative impact anomalies. For all the cells predicted to have a negative impact anomaly, visualizations were created showing the KPI which indicates the cause of anomaly. The thick line in visualizations shows the throughput trend line while the thin line shows the trend of cause of anomaly.

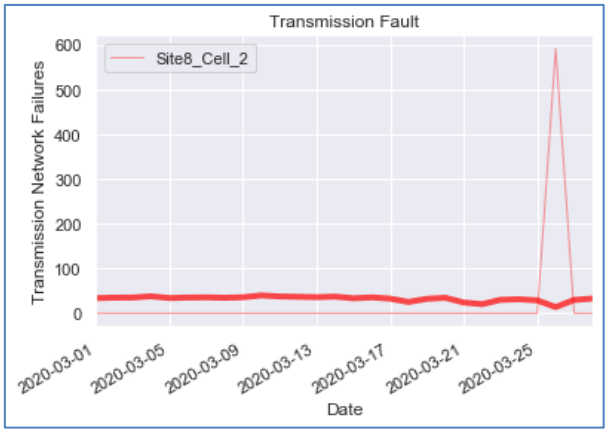

Figure 5: Transmission network failure

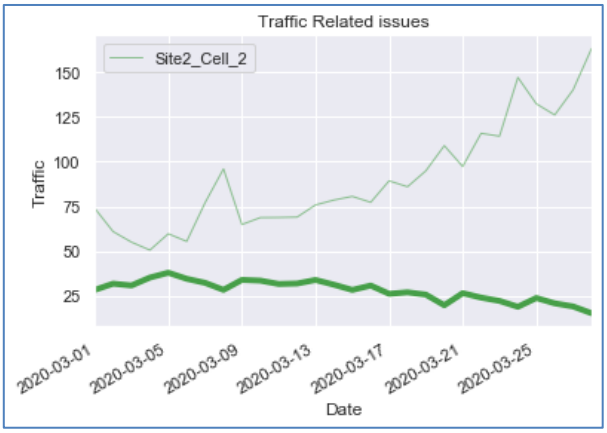

Figure 6: Traffic related anomalies

Figure 4 shows a cell with a spike in transmission network link failures which caused a degradation in user throughput. Figure 5 shows a cell which experienced an increase in traffic from 14th March causing a degradation in average user throughput.

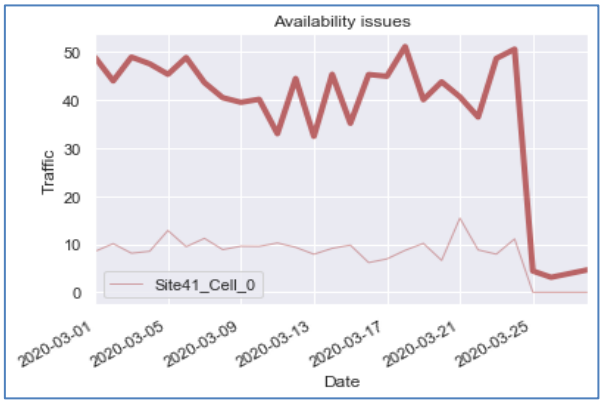

Figure 7: Availability related anomaly

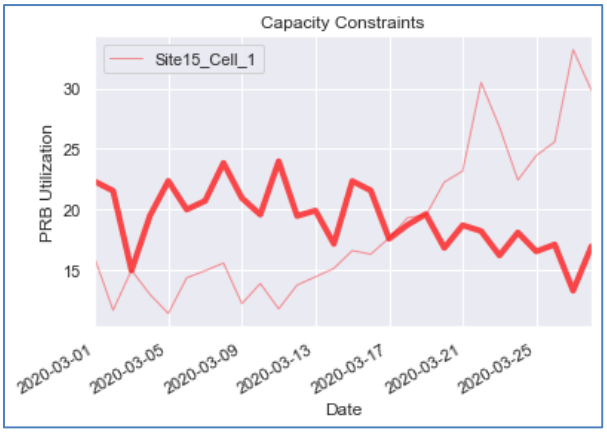

Figure 8: Capacity Constraints anomaly

Figure 6 shows a cell that experienced an availability problem on 25th March causing a step reduction in average user throughput. Figure 7 shows a cell whose physical resource block (PRB) utilization increased from 14th March causing a degradation in average user throughput.

3.2 Discussions

Use of machine learning has been identified as a great tool to introduce intelligence to network management and to develop self/automated healing tools. This project implemented a model to detect anomalies and identify the root cause of those anomalies. The model was able to detect anomalies affecting data throughput and diagnose the root cause that resulted in the degradation in data throughput.

The proposed model anomaly detection adopted an unsupervised learning algorithm similar to approaches used in [4] and [7]. However, the implementation used in this model is different. While [4] system worked by clustering cells based on the correlation between cell KPIs and fault causes, proposed model clustered cells based on the target metric trends in time dimension. Both approaches can detect known and unknown anomalies. However, in [4] model, obtained anomalous clusters had to be labeled with identified causes by experts, who are scarce and expensive. To ensure that a fault cause is not misdiagnosed, [4] used silhouette index and percentile-based approach for border root causes. Average silhouette index for each diagnosis was calculated and used to evaluate the quality of each diagnosis. The root cause with a higher silhouette index was selected. This approach led to reduction of diagnosing error. To eliminate the need for an expert in the training phase of the proposed model, DBSCAN was used to cluster cells based on the key metric trends and any cell that didn't fall in any cluster was labelled as anomalous. A silhouette index of 0.451 was achieved in proposed model showing a good quality of obtained clusters. The proposed model anomaly detection function performed as expected and was able to detect anomalies with both positive and negative impact on data throughput. It was also able to detect anomalies that affect user experience without the data throughput crossing set thresholds used in manual anomaly detection methods.

Root cause analysis in the proposed model adopted supervised learning algorithm (KNN). The model takes advantage of available labelled data to learn the pattern of anomalies based on cell KPIs. Once the model is trained, it is used to diagnose the root cause of detected anomalies. Past research has shown good performance of supervised learning approach for root cause analysis. In [7], both supervised and unsupervised approaches are used in root cause analysis model and on comparing results from both approaches, supervised approach performed better. In [3], supervised learning is also used in anomaly detection and root cause analysis, although data labelling was done internally in his system. Comparison of seven supervised learning algorithms are done in [2] for anomaly detection and KNN had the best performance of 97% accuracy. Other algorithms, multiple layer perceptron, Naïve Bayes, QDA, random forest, AdaBoost and IDE, had an accuracy ranging from 83% to 95%. The proposed model used historical data labelled by Engineers in their network optimization activities which helped to speed up root cause analysis. The model had a performance of 94.59% accuracy. The use of supervised learning in root cause analysis is limited in that it relies on previously known causes and can misdiagnose unknown anomalies. However, this can be remedied by visualizations generated after root cause analysis which can help Engineers identify such cases. Anomaly detection and root cause analysis models were linked up in the application phase whereby the output of anomaly detection was used as new data for the root cause analysis model after training the model.

Mobile networks function relatively well most of the running time with a low chance of failure or network degradation. Therefore, the amount of normal data is often much more than abnormal data which in turn generates imbalanced data for classification problems. To deal with this problem in the proposed model, under-sampling technique was used whereby normal data was randomized and a certain amount was removed. Randomization was done to prevent loss of some important information in majority classes, which is a problem of using under-sampling technique. To deal with imbalanced data problem, [1] and [8] proposed the use of over-sampling minority class data was increased. Over-sampling, on the other hand, may result in over-fitting due to the duplicating operations of minority class samples.

3.2.1 Achievements

This project was able to automate the process of detecting and diagnosing the root cause of those anomalies that affect LTE data throughput. The model if integrated in the network can create efficiency in network optimization through offloading the manual work of detecting and diagnosing LTE data throughput anomalies from the network Engineers. Whereas the manual process focuses on identifying anomalies by using thresholds, this model can detect anomalies that affect user experience without the data throughput crossing set thresholds. An example is a transmission fault that degrades average base station cell throughput from 20Mbps to 10Mbps. Such a degradation may not be picked through manual process which focuses on identifying cells with less than 1Mbps on average user throughput (worst cells), yet the degradation affects experience of users served by that base station. Once such an anomaly is detection, the root cause module can diagnose the cause and generate a visualization after which the network Engineer will focus on resolving the transmission fault. With such a model, network issues can be picked proactively and resolved before a customer complains thus improving network promoter score.

4. CONCLUSION AND RECOMMENDATIONS

In this project, a machine learning model that aims to detect anomalies affecting data throughput in LTE and root cause analysis of anomalies affecting data throughput negatively was proposed. The model makes use of real network KPIs and labelled root cause data from a Kenyan operator, unsupervised

learning for anomaly detection and supervised learning to diagnose network anomalies. The model emulates the manual process followed by radio network Engineers to detect and diagnose network anomalies.

Results show the effectiveness of the proposed anomaly detection and root cause analysis model. It was able to detect anomalies which had both positive and negative impact on data throughput. From the results, those data points that had very high throughput and didn't fall in any cluster were outliers and considered as positive anomalies. Anomalies with negative impact were those that caused a reduction in data throughputs like the one caused by capacity constraints.

To improve the quality of predictions for the root cause analysis module, there is need to sensitize network Engineers on the need to properly recording root cause of new anomalies in the network. With both positive and negative impact anomalies' labelled data, the model will be able to diagnose both type of anomalies in future.

The next steps for future research direction would be to use reinforcement learning to resolve anomalies that do not need physical intervention on site through configuration robots in the network. This will allow the module to work autonomously for the self-healing functionality.

5. REFERENCES

[1] Zhang, T., Zhu, K., & Hossain, E. (2019). Data-Driven Machine Learning Techniques for Self-healing in Cellular Wireless Networks: Challenges and Solutions. Retrieved from https://arxiv.org

[2] Kostas, K. (2018). Anomaly Detection in Networks Using Machine Learning. Retrieved from https://www.researchgate.net

[3] Alvarez, S. L. (2018). Anomaly Detection and Root Cause Analysis for LTE Radio Base Stations. Retrieved from https://pdfs.semanticscholar.org.

[4] Andrades, G., Muñoz, P., Serrano, I., & Barco R, (2016). Automatic root cause analysis for LTE networks based on unsupervised techniques. Retrieved from https://www.researchgate.net

[5] Mobolize. (2014). Can Mobile Networks Keep Up with Data Demand? Retrieved from https://www.mobolize.com/2014/09/30/can-mobile-networks-keep-up-with-demand/

[6] Duan, H., Wei, Y., Liu, P. & Yin, H. (2019). A Novel Ensemble Framework Based on K-Means and Resampling for Imbalanced Data. Retrieved from https://doi.org/10.3390/app10051684

[7] Josefsson, T. (2017). Root-cause analysis through machine learning in the cloud. Retrieved from http://www.diva-portal.org.

[8] Wang, Y., Zhu, K., Sun, M., & Deng, Y. (2019). An Ensemble Learning Approach for Fault Diagnosis in Self-Organizing Heterogeneous Networks. Retrieved from https://ieeexplore.ieee.org