Mechanisms for Root Cause Analysis for Packet Loss in LTE networks implemented over on the Sea

Menglin Yao

May 6, 2022

Final year project thesis submitted in support of the degree of Master of Engineering in Electrical & Electronic Engineering

Department of Electrical and Electronic Engineering

University of Bristol

DECLARATION AND DISCLAIMER

Unless otherwise acknowledged, the content of this thesis is the original work of the author. None of the work in this thesis has been submitted by the author in support of an application for another degree or qualification at this or any other university or institute of learning.

The views in this document are those of the author and do not in any way represent those of the University.

The author confirms that the printed copy and electronic version of this thesis are identical.

Signed:

Meglin Yao

Dated:

05/05/2022

Abstract

With the advancement of renewable resource utilisation technology, large-scale offshore wind farms have been built over years. The staff on the maintenance vessels of these wind farms need telematics support, so a reliable communication network is required to be established in the power plant. In order to improve network performance, this project investigate the root cause of the data packet loss problem in the LTE network implemented over the Moray East offshore Wind Farm.

We propose a machine learning-based root cause analysis system. The system labels anomalies of the packet loss rate data and locates their root cause among other network metrics recorded within the same time period. Two subsystems are designed for the surveyed metrics that are taken from the databases of the CPE (Customer Premises Equipment) and cell station. The subsystem on the CPE end adopts single-feature novelty detection machines, which is trained on data correspond to normal packet loss rates and identifies anomalies in the unlabelled test data. The metrics on the cell level are analysed through the supervised decision tree algorithm, with the feature importance analysis provided by the algorithm, the root cause being defined as the metric that have the greatest impact on the packet loss. The cell level subsystem can also match recurring root cause with the pre-defined root cause models for a higher level of automation and efficiency.

The trained CPE end subsystem is verified to be able to accurately predict normal data and locate anomalies in the test data that could lead to packet loss. The cell level subsystem can comprehensively analyse the root cause of anomalous packet loss. Through simulated data, we also verified the attribute of automatic localisation of recurring root cause using this subsystem.

Acknowledgements

This project is an industrial placement supported by Vilicom UK Ltd. I would like to express my special thanks to CFO Mr. Rui Inacio and data scientist Mr. Abdelrahim Ahmad for providing data and technical support to the project.

To my supervisor, Professor Robert Piechocki, I would like to thank you for your enlightenment and encouragement. I really enjoyed the process of learning and innovating in the area of machine learning and network data analysis.

To my parents, thank you for your kindness and support as always, you have guided me to be a compassionate and honest person.

To Sherry, thank you for your sincerity and growing up with me.

To Peter, Sarang and Zhixian, you are my best friends, thank you for all the support and comfort when I was weak, your companionships are extremely meaningful and important to me.

Contents

Chapter 1 Introduction

Over the past two decades, with the emphasis on reducing carbon emissions, wind power technology has been well developed. Such advancements have promoted the utilisation of high-quality wind resources in the North Sea region, numerous offshore wind farms have been built during this period. One of these projects is the Moray East Wind Farm, which is located off the northeast coast of Scotland and covers an area of 295 km^2 with a plate output of 950MW [1]. Such a large-scale wind farm requires regular maintenance. During which, workers on the maintenance vessel need a secure and low-latency network connection to a private datacentre to acquire real-time information update and support. An offshore private LTE network over the wind farm is contracted by Vilicom UK Ltd. for such a purpose. The structure of this network is demonstrated in Figure 1.

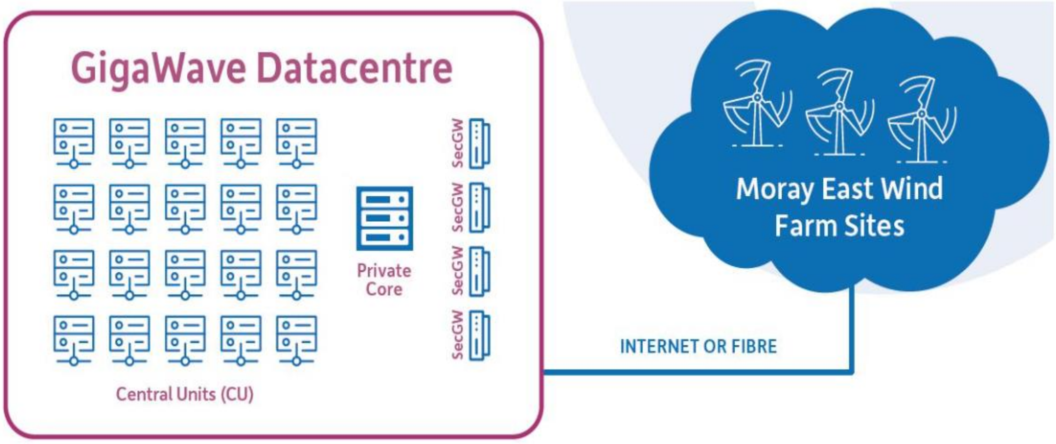


Figure 1: Private Network Connectivity for the Sea Vessels in Moray East Wind Farm [2]

Since the wireless communication environment on the sea is different from that on land, traditional land-based methods for assessing and diagnosing network performance may no longer be applicable, hence targeted analysis strategies need to be developed. This project focuses on the packet loss issue in this LTE network to conduct a root cause analysis. With the support of machine learning algorithms, patterns of performance metrics data related to this issue can be discovered, which provides critical information to derive the further root cause analysis of packet loss anomalies, and in some cases, the root cause can even be directly revealed.

1.1 Aims and Objectives

The research of this project is based on the databases provided by Vilicom. The purpose is to discover anomalies in the packet loss rate KPI data, then perform fault localisation through other relevant indicators corresponding to these anomalies. Specific goals are:

Goal 1: Identify the abnormal performance in the packet loss rate data, in order to classify and label the normal and abnormal data for machine learning algorithms.

Goal 2: Design the root cause analysis system with the support of machine learning algorithms to find the root cause of packet loss problems over time.

Goal 3: Fully automate the system's process of detecting root cause.

Goal 4: Automatically intercept abnormal packet loss rate data segment in real-time and feed them into the trained system to investigate the root cause.

Goal 5: Apply the system on the metric data collected from the offshore LTE network and conduct root cause analysis with the designed system.

1.2 Delimitations

The investigation scope of this project is limited to the data currently provided by Vilicom UK Ltd. These data may not be representative of all normal or abnormal situations, and we only train and test the models based on the data we have so far. The machine learning algorithms involved in the project are currently provided by Scikit-Learn [3]. These algorithms are only used for feasibility studies.

1.3 Thesis Outline

All background information is presented in Chapter 1, along with a literature review. In Chapter 2, data selection and pre-processing methods are elaborated, and the design of the RCA algorithms is explained. There are two core RCA algorithms, which are validated in Chapters 4 and 5 respectively. The results of the RCA output by these two algorithms are also presented and discussed in these two chapters.

1.4 Nomenclature

LTE: Long Term Evolution Network

PLR: Packet Loss Rate

UE: User Equipment

CPE: Customer Premises Equipment

eNB: Evolved Node B, they are referred as cell in the RCA systems.

PDCP: Packet Data Convergence Protocol Layer

SDU: Service Data Unit

KPI: Key Performance Indicator, metrics traced in the eNodeB (cell) database and referred to as features in the machine learning algorithms.

KM: Key Metric, metrics traced in the CPE database and referred to as features in the machine learning algorithms.

RF: Radio Frequency

RCA: Root Cause Analysis

RC: Root Cause

LOF: Local Outlier Factor

SVM: Support Vector Machine

OCSVM: One Class Support Vector Machine

E-UTRAN: Evolved Universal Terrestrial Access Network

EPC: Evolved Packet Core

GSM: Global System for Mobiles

WCDMA: Wideband Code Division Multiple Access

Chapter 2 Background Research

The LTE network's architecture and advantages are first introduced in this chapter. Next, the mechanism of machine learning algorithms involved in this project are demonstrated, and the previous works that applied these algorithms are also mentioned. In the related work section, contributions on LTE network performance evaluation and systematic RCA are reviewed and discussed.

2.1 LTE

LTE is a 4G communication standard formulated by 3GPP. It is a pioneer in the commercial 4G standards. It reduces system complexity and increases spectral efficiency compared to previous generation communication technologies, such as GSM and WCDMA [4]. The LTE network is applied to the Moray East wind farm based on its additional benefits in this special case. Compared with 5G, LTE has a larger coverage area, which means that fewer base stations can be built within the same range to achieve full coverage. Although the bandwidth is inferior to 5G, the maximum bandwidth of 20MHz [5] is theoretically sufficient to support the usage in this private network. Compared with Wi-Fi, LTE has more reliable mobility management, and a licensed spectrum can protect the communication frequency band from violation to a certain extent [6]. Although the above features of LTE outperform other communication standard options, they are also vital factors leading to network instability in extreme conditions. Therefore, they are the focus of investigation in this project.

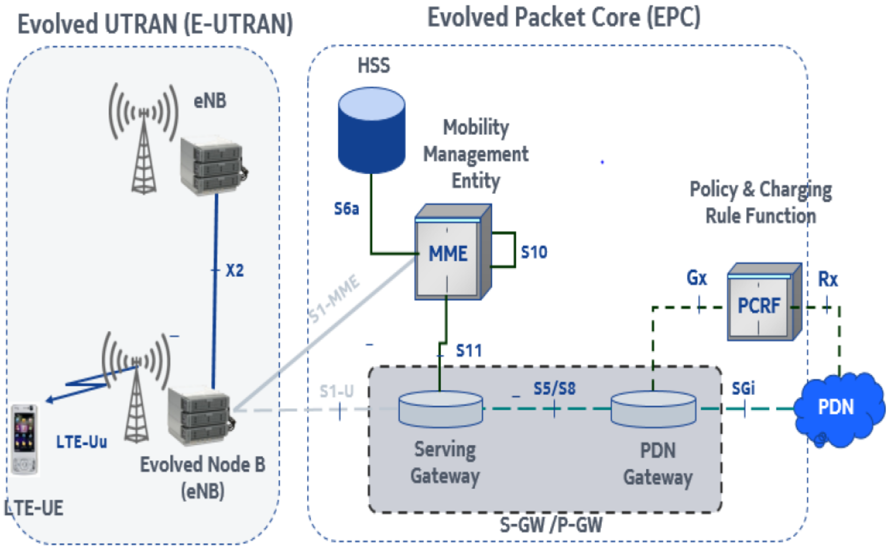


Figure 2: Overview of LTE Architecture [7]

The architecture of the LTE E-UTRAN and the EPC (Evolved Packet Core) is shown in Figure 2. The E-UTRAN part in the figure is the network implemented over the offshore wind farm. The eNBs are the radio base stations [8], which is also referred to as cells in this article. The representative of the UE in this offshore network is the CPE. This project mainly focuses on finding the root cause of the packet loss problem from the E-UTRAN. Each element in the E-UTRAN has metrics with data recorded in real-time. The eNB mainly tracks indicators related to network performance, namely KPIs. The key metrics (KM) in the CPE database primarily evaluate the environment in which the UE is located from different angles. The RCA system starts with these two databases and evaluates the relationship between these metrics and packet loss by mining patterns in metric data.

2.2 Machine Learning Algorithms

In this project, machine learning algorithms support RCA as predictors and data analysers. There are three categories of machine learning algorithms that are taken into consideration, namely unsupervised, supervised, and semi-supervised learning.

Unsupervised learning algorithms usually do not require the data to be labelled, and simply mine data patterns from input features. A classic application of unsupervised learning is outlier detection. When completing such tasks, unsupervised learning models can locate outliers in the data by analysing their deviation from the normal patterns of the data.

Contrary to unsupervised learning, supervised learning makes judgments on new data based on empirical data. It has high requirements on the quality of training data. In addition to the need for training data to be labelled, the class sample size also needs to be balanced, and the training data needs to cover all the cases under investigation.

Semi-supervised learning algorithms are in between the above two types, their requirements for training data are not as strict as supervised learning, and new data can be analysed with only a small amount or incompletely labelled training data through pattern learning.

Decision Trees

Decision Tree is a supervised learning algorithm. Among the mainstream algorithms, the CART (Classification and Regression Tree) was developed by Breiman *et al.* in 1984 [9]. ID3 was proposed in 1986 by Quinlan [10]. Later in 1993, Quinlan improved the algorithm to C4.5 [11]. A decision tree classifier model organises data by categories in the target labels through automatic rule-setting.

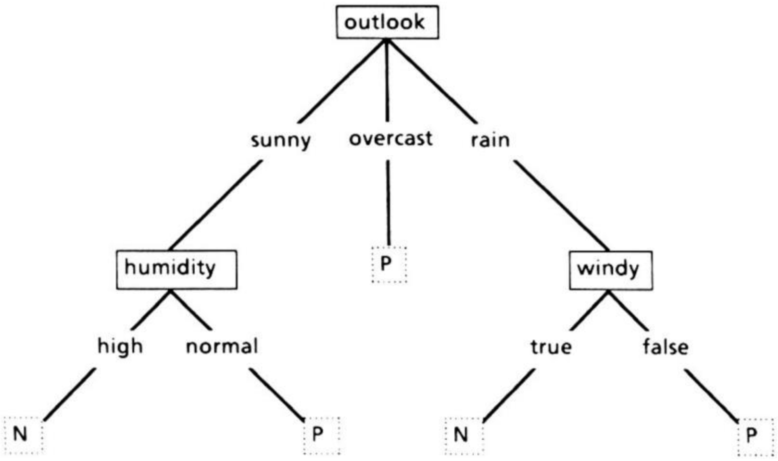


Figure 3: A Decision Tree for Golf Player's Attendance Decision [10]

Figure 3 illustrates a simple decision tree model. The model predicts golf players' attendance decisions based on environmental factors [10]. In the model, the decision of whether attend or not is the target, represented by the N and P labels, respectively. The outlook, humidity and windy conditions are features that support the model.

On the root node of the model, all categories of data are mixed, then the algorithm splits the node based on the feature and threshold that can bring maximum impurity decrease (information gain) according to the evaluation criterion, so that the amount of data biases to a particular category on the child nodes. This operation is performed on the child nodes repeatedly. Such recursion continues until the data purity on a child node meet the data purity standard. Different versions of decision trees have different data impurity evaluation indicators and information gain measurements. For instance, The ID3 Tree uses mutual information $I(P;C)$ to measure information gain G:

$$G = I(P;C) = H(P) - H(P|C) \tag{1}$$

Where $H(P)$ is the data entropy on the parent node and $H(P|C)$ is the sum of all children nodes' entropy weighted by sample fraction.

The CART algorithm uses the Gini index as the impurity evaluation standard. The Gini index on a single node with label classes $\{C_1, C_2, C_3, \dots, C_K\}$ is calculated by:

$$Gini(P) = 1 - \sum_{i=1}^{K} \left(\frac{|N_i|}{|D_i|}\right)^2 = 1 - \sum_{i=1}^{K} {p_i}^2 \tag{2}$$

Where N_i is the number of samples under C_i on node P, and D_i is the total number of samples under C_i.

Since the decision-making process of the algorithm is highly transparent, and the attributes such as feature importance and model quality evaluation can provide numerous amounts of information about the input data, the decision tree algorithm is still one of the most popular data mining tools. With the aid of visualised decision tree, a risk factor analysis for pressure ulcer in Korean long-term care system was performed by Moon and Lee in 2017 [12]. In the project of oral cancer prognosis through different data mining methods, Tseng *et al.* [13] concluded that compared with artificial neural networks, the data analysis results displayed by decision trees are more understandable for medical staff, provided that the performances of both techniques are better than traditional statistical methods.

Schölkopf One-Class SVM

In 1999, Schölkopf et al. [14] proposed a semi-supervised one-class classification variation of SVM [15], namely one-class support vector machine (OCSVM), for the purpose of addressing novelty detection problems. The goal of such an algorithm is to detect the test data that lies outside the one class domain, which is created by the single class training data. When training the model with data $\{x_1, x_2, \dots, x_n\}$ where $x_i \in \mathbb{R}^d$, the idea is to map the samples located in the original d dimensional space to a higher dimensional feature space *F* corresponding to the kernel function:

$$K(x, y) = dot\left(\Phi(x_i), \Phi(x_j)\right) \tag{3}$$

There are a variety of kernel functions to choose from, and they are discussed in detail in Bounsiar and Madden's work [16]. Among them, the most representative Gaussian RBF (radial basis function) kernel deserves a separate mention:

$$K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right), \quad \sigma \geq 0 \tag{4}$$

Then with the aid of the support vectors, a hyperplane is created in *F* space, which separates most samples from the origin with the maximum margin. When data x is input, it is considered normal if x lies inside the margin defined by the hyperplane, otherwise it is considered abnormal. The decision function is:

$$f(x) = sgn\left(\sum_i \alpha_i K(x_i, x) - b\right) \tag{5}$$

In which an x_i with a non-zero α_i is a support vector. The one class margin is created by referencing the support vectors. The α_i is defined by the solution of:

$$\min_{\alpha} \ \frac{1}{2}\sum_{ij} \alpha_i \alpha_j K(x_i, x_j) \quad subject\ to \quad 0 \leq \alpha_i \leq \frac{1}{\nu d}, \qquad \sum_i \alpha_i = 1 \tag{6}$$

It can be seen that the choice of parameter ν is critical to this process. It defines the upper limit of *outlier proportion* that is allowed in the training data and the lower limit of sample fraction that is treated as support vectors [17].

OCSVMs are classic semi-supervised learning algorithms. In binary classification tasks, when we have sufficient knowledge about one state of an event, but little about the other state, OCSVM can be used to detect the occurrence of anomalies. Therefore, it is often used in the field of network security to identify intrusion events. Zhang *et al.* [18] applied OCSVM to network anomaly detection, the results showed that the OCSVM give more effective predictions *i.e.*, fewer false alarms than the traditional rule setting methods. On the other hand, in [19], Winter *et al.* trained an OCSVM with only anomalous data and applied it to a lightweight network to detect in-class anomalies. They also obtained satisfying results with such a method, and concluded that the OCSVM mitigates typical drawbacks brought by the traditional anomaly detection techniques.

Local Outlier Factor

The LOF (Local Outlier Factor) is an unsupervised density-based anomaly detection algorithm proposed by Breunig *et al.* [20]. The algorithm judges whether a point is an outlier by comparing the sample density in its k-neighbourhood with the density in these samples' (neighbours) k-neighbourhood. A point is considered anomalous when its neighbourhood density is significantly lower than that of its neighbours.

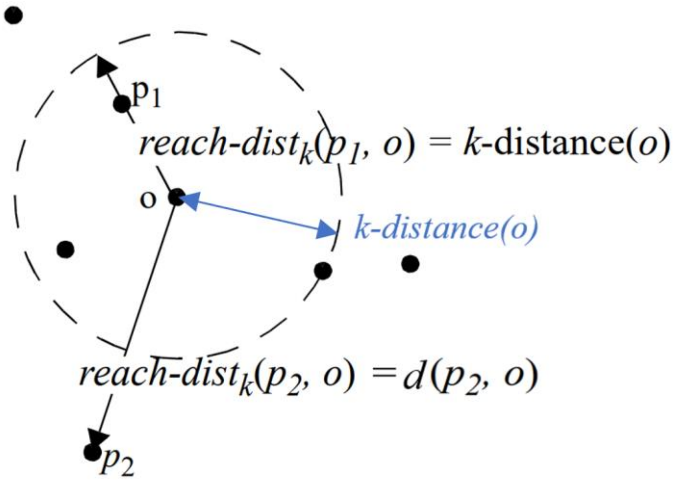

p1
reach-dist_k(p1, o) = k-distance(o)
o
k-distance(o)
reach-dist_k(p2, o) = d(p2, o)
p2

Figure 4: The Definition of Reachable Distance [20]

Figure 4 shows the method for judging the reachable distance from point *p* to *o*, the figure is developed from the original demonstration presented in [20]. It can also be expressed as:

$$reach - dist_k(p, o) = \max\{k - distance(o), d(p, o)\} \tag{7}$$

Next, the local reachable density (LRD) in the neighbourhood of a point $N_{MinPts}(p)$ can be defined by the reciprocal of the mean reachable distance from its neighbours:

$$LRD_{MinPts}(p) = \frac{1}{\left(\frac{\sum_{o \in N_{MinPts}(p)} reach - dist_{MinPts}(p, o)}{|N_{MinPts}(p)|}\right)} \tag{8}$$

Where p's neighbourhood N is defined to must contain at least $MinPts$ points. An example of $N_{MinPts}(p)$ is illustrated in Figure 5.

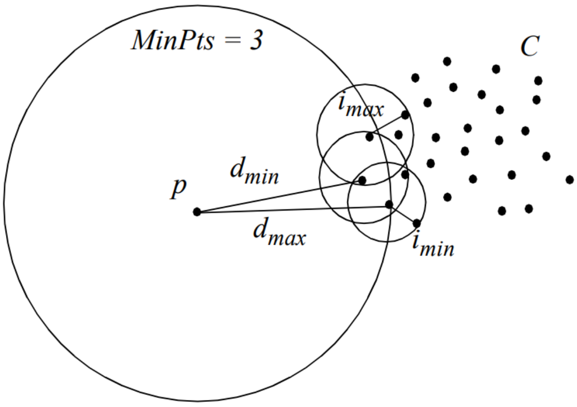


Figure 5: Neighbourhood of Point *p* [20]

Finally, the decision function which is the density comparison between the survey point and the neighbour points can be expressed as:

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{LRD_{MinPts}(o)}{LRD_{MinPts}(p)}}{|N_{MinPts}(p)|} \tag{9}$$

Where the hard decision criterion is:

$$f(p) = \begin{cases} Same\ Class & LOF(p) = 1 \\ Inlier & LOF(p) < 1 \\ Outlier & LOF(p) > 1 \end{cases} \tag{10}$$

The point *p* in Figure 5 is likely to be determined as an outlier according to (7), (8) and (9) since the density in its neighbourhood defined by mean reachable distance is smaller than its neighbours'.

LOF can be used for both anomaly detection and novelty detection. Lee *et al.* [21] combined LOF novelty detection with independent component analysis in their work and used it for process monitoring in the industry. The results showed that the superiority of LOF is that it is not restricted by the data distribution of the detected indicators, and can achieve the expected anomaly detection rate. Budiarto *et al.* compared the performance of OCSVM, LOF and K-Means on outlier detection on the drug use dataset in [22], and concluded that OCSVM surpasses the other two detection methods by a slight advantage.

2.3 Related Works

2.3.1 LTE Performance Analysis

Miyim *et al.* evaluated a local LTE network in Nigeria in [23]. The evaluation indicators include packet loss rate, jitter, delay and throughput. They simulated three cases successively. In the case of increasing the number of UEs from 10 to 200, all performance of the network deteriorated in proportion. The same situation also occurs in the experiment in which the distance between UE and eNB is increased from 0.1 to 2 km. The experimental results of UE’s movement speed showed that the delay and jitter only fluctuate in a certain speed range, but the packet loss rate is still proportional to the speed. The authors attribute the cause to the Doppler shift effect and link it to changes in frequency. These results are valuable observations, proving that in an LTE network, coverage, movement speed and the number of UE can all affect the packet loss rate.

The project [24] specifically aimed at researching the packet loss caused by the user's movement in the mobile network. The researchers took measurements on trains in Norway to investigate the relationship between packet loss and movement speeds, cell handovers, and service and communication technology switching. The results showed that the packet loss rate of the LTE network during cell handovers is significantly better than that of the 3G technologies when the user is driving at a speed below 50 km/h, but when the driving speed exceeds 50 km/h, the packet loss in LTE is severer than that of 3G. Despite this, the packet loss caused by LTE cell handover is still serious. When the UE moving speed is lower than 10, 10% of the data packets are lost. As the speed rises above 50 km/h, the packet loss rate caused by the handover reaches 67%. This study showed that although the mobility management system of the LTE network optimises the data loss problem caused by handovers, the packet loss still exists and may affect the user experience.

Masum and Babu's Thesis [25] evaluated the performance of VOIP (Voice Over IP) in LTE networks. The research on the relationship between network congestion and packet loss in this project deserves to be highlighted. The researchers tested an LTE network with 20MHz bandwidth by simulating extremely high VOIP traffic load, and the results showed that when only VOIP traffic was used, the congested network had little effect on the packet loss rate. However, when VOIP and FTP (File Transfer Protocol) traffic jam the network together, the packet loss rate increases significantly compared to the former case.

The above projects are all network performance analyses based on terrestrial communication. These works provide us with factors that may lead to packet loss in general. In the special case of maritime communication, there is also contributions on the simulation and evaluation of the radio communication environment that are worth to mention. Wang *et al.* conducted a thorough study on the maritime wireless communication channel in [26]. Through mathematical modelling, they suggested that when the UE and the base station are far away, although there is no shadowing effect, the weak multipath caused by the reflection of the sea surface is also destructive to the direct path. Huo *et al.* conducted a supplementary study [27] on this work, specifically exploring the impact of ocean waves on wireless communication paths. However, the research target of this study is a buoy-based communication base station. In this project, the base station is located far above sea level, so it is less affected by sea waves factor. These studies all belong to the evaluation category of the electromagnetic environment. In this project, there will be special indicators to evaluate the RF conditions in an all-around way.

2.3.2 Systematic RCA Projects

In project [28], Forsberg designed an automatic RCA system based on unsupervised anomaly detection for a microservice cluster. The system applies the clustering algorithm to learn the normal data pattern of each metric separately and detects the abnormality in the new data. The detection result is then compared with the pre-defined RC (root cause) instances to determine which type of RC the abnormality belongs to. The system achieved automation and can efficiently locate the RC of the problem, which has certain inspiration for the design of this project. However, due to its heavy reliance on empirical data and RC instances, its scope of investigation is limited to pre-defined RC types, and no judgment can be made on unseen anomalies.

Alvarez proposed an Autoencoder-based LTE network RCA system in [29]. Different from Forsberg's work, this project packs all indicators into an Autoencoder model for research, and detects abnormalities based on RE (reconstruction errors). It is proposed in this paper that the RE of each feature can be calculated separately and treated as features' importance to localise the RC, but no specific method is mentioned. Since the error caused by a certain feature will be propagated through the entire Autoencoder model, only the average RE of the whole set of features can be obtained without adding additional steps. This is mentioned in a similar project [30], for which the author provides a solution. They calculate feature importance through biasing input data and define the feature with most deviated data as RC when an anomaly occurs.

In a complementary study [31], we found that black-box model interpretation methods such as permutation importance also can solve the feature selection problem. For example, the permutation importance technique shuffles the data input to the model and measures the importance of features based on the impact of this operation on the prediction results. The most important feature can be used as RC.

Josefsson proposed an RCA system for the cloud infrastructure [32] that applies Kohonen's SOM (Self-Organizing Map) [33]. The SOM is originally designed to be an unsupervised single-layer neural network, the author also tried to build a two layered SOM in this work. The methods and results in the article prove that the SOM algorithm itself has both anomaly detection and fault localisation functions.

The application of SOM is very popular in the RCA field, it can also be found in [34] and [35]. The system performance of the former work was compared in Josefsson's work with his results, and he demonstrated the superiority of his system. The latter project applied SOM to design the fault localisation part of a self-healing LTE network. The KPIs from the LTE physical layer is analysed in this project.

Demetgul [36] studied the performance of different types of Decision Trees and SVMs with different kernels on the RCA tasks, and the results demonstrated that both algorithms are competent for the RCA task. However, this project was carried out in a supervised style, which means that the researcher had a well-covered training data. This is not suitable for the specific situation of our project, but it provided some references for model selection.

As a supplement, Sole *et al.* conducted a comprehensive analysis of more RCA systems [37], they are evaluated through system functionality, operating efficiency, and complexity. This article also inspired the research direction of this project.

Chapter 3 Methodology

In this chapter, the RCA framework system is first presented, then each part of the framework system is discussed separately. The data selection and labelling techniques are introduced. The two core RCA subsystems are highlighted, and the algorithm choices and supporting reasons in the system design process are explained.

3.1 Overview of the Root Cause Analysis System

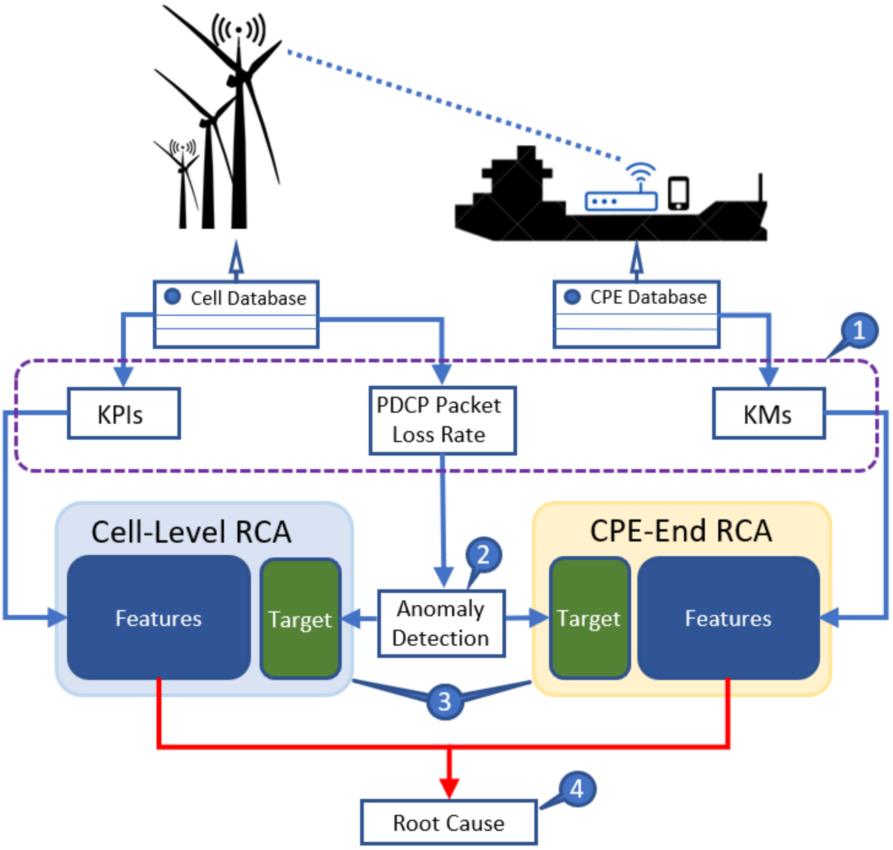


Figure 6: An Overview of the RCA System

An overview of the system for root cause analysis on packet loss is presented in Figure 6. Such a system operates as steps numbered in the graph:

1. The data of network KPI (key performance indicators) and KM (key metrics) are selected from the databases of the cells (eNBs) and CPE respectively.
2. The packet loss phenomenon in this network is identified and labelled through anomaly detection techniques on the PLR (packet loss rate) KPI data.

3. Two RCA subsystems with machine learning cores for corresponding databases are employed. The labelled PLR data are entered into the subsystems as the target variable and the other interested KPIs and KMs are treated as input features.
4. The subsystems evaluate the connection between other metrics and PLR in the same period. Consequently, the factors that are most likely to lead to packet loss can be observed by these systems as root causes.

3.2 Data Selection and Cleansing

The data involved in the study are acquired from two network elements' databases provided by Vilicom, they are also demonstrated in Figure 6. The cell databases provide the PDCP layer data flow and delivery performance metrics that are recorded from the CPU at the eNB base stations. The CPE database mainly stores the KMs that evaluates the user's surrounding physical environment. The RCA subsystems proposed in this project are built for these two network elements respectively. The interested KPIs and KMs are discussed in this section. The data are raw when they are drawn from the databases, so cleansing is required before further operations can be performed with them.

3.2.1 Cell-Level

A total of nine eNB cell stations are built on top of the wind turbines in this offshore LTE network. When they are connected to the CPE, all data traffic between CPE and the data centre or the external network passes through these base stations. Every five minutes, a set of KPI data is added to the cell database. Within the five minutes window, the methods of data statistics for each KPI are different, the specific method is determined by the nature of the KPI. The metrics selected for the cell-level RCA algorithm are listed in the Appendix 2 along with their units, recording methods, description and specifications provided by 3GPP.

All eNB database KPIs involved in this project are PDCP layer measurements. Due to the advantages of the eNBs' location in the topology of this network, the KPI data related to network traffic captured here can accurately reflect the actual performance [6]. Therefore, the PLR data as the analysis target is obtained from this database. For the same reason, the other KPIs selected for root cause analysis in Appendix 2 are all data delivery and usage related performance metrics. On the other hand, since these KPIs are obtained at the same level with PLR, their relationship with the target is unmediated (one-to-one), which means less risk of error in the results caused by other unknown factors in the network.

The original database contains a large amount of missing or incompletely recorded data, which need to be further cleaned before they can be put into the system for training and testing. Due to the essence of RCA algorithms in this project is to trace the root causes by analysing and mining unlabelled data, the processing of missing data should be more careful, operations such as changing null values in the original dataset to zero may cause the algorithms being wrongly trained. Therefore, data rows that contain missing values are excluded from the selected data and will not be further considered in this project.

3.2.2 CPE Data

CPE is a network device at the user end, which provides LAN services for users by connecting to eNB cells. The database on this device records the KMs (Key Metrics) that are critical to RCA on packet loss. In other words, from these metrics we have the best chance of tracing the root cause of the packet loss issue.

The data sampling frequency in this database is every 10 seconds. Such a high-resolution dataset is more conducive to detailed exploration. Nevertheless, this also means a huge gap in sampling frequency with the target, which is the PLR data that is taken on 5 minutes average. This situation greatly reduces the feasibility of performing RCA on the data from the two network elements through the same algorithm, and lays the foundation for the RCA algorithm established for CPE-end.

These KMs and related information are listed in Appendix 3. They can be divided into three main categories, the KMs that fall under each category, the status and reasons for which they are considered in this project are listed below:

- **Physical Environment**

  GPS Coordinates: The *longitude* and *latitude* data are provided as two separate features. They are fully considered for the coverage investigation. A possible situation is that when the user moves from one cell's coverage area to another cell's, the CPE is still connected to the previous cell. In this scenario, the coverage of the previous cell at the current location could be extremely poor, such a weak connection may lead to a high packet loss.

  Device Temperature: Not Considered. Perhaps under extreme device temperature conditions, the CPE may malfunction and cause a loss of connection. However, when this happens, the device must be issuing a warning at first, and it is not practical to further identify it as a root cause. On the other hand, in the provided period of data, such a thing never happened, so it cannot be studied and analysed with the existing data.

  Movement Speed: Partly Considered. The rapid movement of the user across the windfarm area means that handovers between cells occur more frequently, and each handover may occur the scenario described in the device's GPS coordinates section. Moreover, the impact of user mobility on network performance is a big issue in the industry, and engineers are constantly trying to deal with handover failures, which means that the related technologies are still not perfect yet, so this factor cannot be ignored. In order to avoid the RCA system declaring extremely low moving speeds or stationary scenario as abnormal behaviours, the system does not take them into account in the final assessment, and only studies the cases above a certain speed level.

- **Electromagnetic Environment**

  All indicators related to the electromagnetic environment are fully considered, and the aspects they represent are shown in the Appendix 3. They are *SINR*, *RSSI*, *RSRP* and *RSRQ*. They assess the RF environment in which the user is located from different perspectives.

  The reason why these indicators are all analysed is the specificity of the offshore environment. Unlike on land, the law enforcement on telecommunication spectrum protection in the open sea is weaker, it is possible that the channels of this network are occupied by unknown signal sources, which could lead to a strong interference and cause the failure of transmission. On the other hand, the extreme weather conditions that may occur at sea could corrupt the communication link, which can also be reflected by the quality of the RF environment. From the above factors, it is not difficult to conclude that the electromagnetic environment needs to be evaluated from multiple perspectives to cover as many scenarios, the abnormality of these indicators has the potential of being the root causes of the data packet loss.

- **Data Usage**

  Data usage is measured separately by upload (*Tx*), download (*Rx*), and total bandwidth usage (*Usage*). Different from the measurement at the cell level, the CPE only measures the usage of the UEs that are connected to it, instead of that in the entire network. These data also have certain reference value. For example, when the connected user experiences data packet loss, the system can analyse these data to determine whether a consequential rise of PLR is caused by the users' overload usage in the CPE LAN.

  However, usage data are not fully considered on CPE devices. Since this network is only used by the maintenance personnel of the wind farm, and they do not use the network a lot when they are resting at night, the consequential decrease in the network usage at night is predictable. Therefore, the usage-related data during late night is temporarily not considered in this project.

After these data are selected, they also need to be cleaned in the same way as the cell-level operation, that is, to remove data rows containing null values. It is worth mentioning that missing data is rare in this database, so a more specific analysis can be done over the timeline.

3.3 Labelling on Packet Loss Rate

This project analyses the packet loss problem, which is a study with a clear target orientation. Therefore, the RCA systems are built based on either semi-supervised or supervised learning algorithms to predict the target *i.e.*, the PLR. In the case of binary classification, both types of algorithms require the target data to be labelled by class. The anomaly detection algorithms executed on the PLR data are designed to achieve this purpose. The labelling technique is divided into automatic and manual, they are applied in different scenarios. The criteria of these methods and the reasons for choosing them for specific cases will be discussed in this section.

3.3.1 Automatic Anomaly Detection

Automatic detection algorithms separate normal and abnormal phenomena by finding deviants from the input data. In this project, its advantage is reflected in the aspect of dynamic criteria setting. Maybe the evaluator can set a hard criterion to classify the data abnormality in general, but such a standard may not be suitable for some specific situations. For example, due to some unknown reasons, the PLR in a certain day has been maintained at around 2%, and the standard is manually set to be 1% on PLR, this will halt the RCA systems from functioning on any of these data, because almost all data are classified as anomalous, and the RCA of this situation may be out of scope. Luckily, automatic unsupervised learning methods such as isolation forests can overcome this restriction, even if the average PLR is slightly higher than the standard, these algorithms can still find the real anomalies from such a dataset. Therefore, it is an excellent method for labelling data that are going to be tested on the systems for tracing root causes.

3.3.2 Manual Labelling

Although the benefits brought by automatic labelling technique can help better adapt to various situations of test data, it still has drawbacks if used globally. They are revealed when it is used to label the training dataset. The anomaly detection algorithms are not perfect, they certainly will be making inaccurate predictions. Since the outlier detection models are unsupervised, the number of misclassified samples in the selected training dataset is uncontrollable. Unfortunately, the RCA system is designed in a cascaded style, which means that when training the model, the risk of error in the early stages should be avoided as much as possible, otherwise these errors will be carried into each subsequent stage and amplified, eventually leading to the failure of the determination.

The other issue is the inevitable overfitting. Automatic detection for a specific segment of data is optimal, but this also means that it is not general enough. Its prediction results are strongly depending on the sample space, class balance and other attributes of the specific dataset. Nonetheless, the model is expected to be trained generalised enough to make accurate decisions about a certain range of different situations, so a specialised training dataset is unwanted.

Accordingly, when producing the training dataset, it is necessary to manually set a threshold for anomaly determination in order to strictly meet the training requirements:

$$\delta(p) \equiv \begin{cases} 0 & p \leq \lambda \\ 1 & p > \lambda \end{cases} \tag{11}$$

- δ is the classification label, 0 is normal and 1 is abnormal.
- p is the PLR data.
- λ is the threshold manually set on PLR.

Based on the threshold, the PLR data lower than it will be marked as normal. When it rises above the bar, it will be labelled as abnormal.

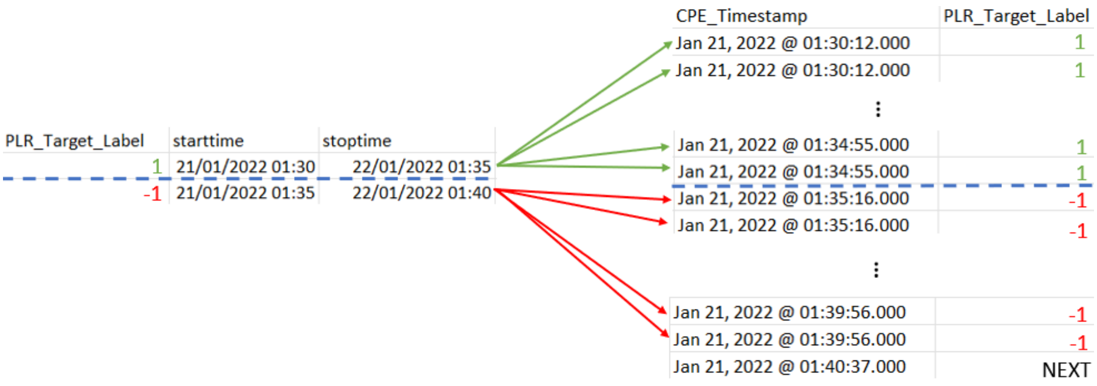
3.4 Algorithm 1: CPE-End Root Cause Analysis

In this section, a root cause analysis system design for KMs from the CPE database is presented. First, the problems and challenges of conducting RCA on the CPE end are deliberated first, then their solutions are proposed. Then the choice of machine learning model is discussed. Finally, the mechanism of the originally designed RCA system is demonstrated and explained.

3.4.1 The Sampling Frequency Inconsistency Issue

As mentioned in section 3.2.2, there is a significant difference in the sampling rate between the metrics in the CPE database and PLR. Such differences must be eliminated if the RCA on CPE-end is to be done by any machine learning algorithms other than unsupervised learning, and since the analysis of the data in this case has packet loss as an explicit target, unsupervised learning is not an optimal choice.

The most intuitive way to eliminate inconsistencies in the sampling frequency of CPE metrics and target is to broadcast low resolution targets to high resolution KM data by their timestamps. The PLR data is recorded on five-minute average, and after labelled by the anomaly detection algorithm, a PLR sample corresponds to a class label that represents normal or abnormal performance, this column of data is referred as the target in the system. If the target and CPE metrics are made consistent by broadcasting in terms of timestamps, for one PLR label, the target value of all CPE metric samples within that PLR's five-minute measurement window will be assigned to that label. An example of target label broadcasting is demonstrated in Figure 7.



PLR_Target_Label	starttime	stoptime
1	21/01/2022 01:30	22/01/2022 01:35
-1	21/01/2022 01:35	22/01/2022 01:40

CPE_Timestamp	PLR_Target_Label
Jan 21, 2022 @ 01:30:12.000	1
Jan 21, 2022 @ 01:30:12.000	1
⋮	
Jan 21, 2022 @ 01:34:55.000	1
Jan 21, 2022 @ 01:34:55.000	1
Jan 21, 2022 @ 01:35:16.000	-1
Jan 21, 2022 @ 01:35:16.000	-1
⋮	
Jan 21, 2022 @ 01:39:56.000	-1
Jan 21, 2022 @ 01:39:56.000	-1
Jan 21, 2022 @ 01:40:37.000	NEXT

Figure 7: The Broadcasting Method of Target-Features Data Resolution Conversion

However, such an approach to data resolution conversion will rise a problem, which is uncertainty in the CPE metrics data with abnormal targets. Within the five minutes covered by an averaged PLR sample, numerous situations could occur. For example, the PLR suddenly rises to 40% at a timestamp, lasts for one minute and then returns to normal. Such a situation will result in an average PLR of 8% in a five-minute period even when all other samples are 0%. This result may have exceeded the classification standard we manually set when selecting the training data, hence all CPE samples within this period would be labelled as anomalies. This is clearly unfair for the majority of data within those five minutes, such a portion of incorrectly labelled targets is bound to cause inaccurate predictions.

On the other hand, since the normality threshold of PLR is set close to the lower limit of the sample value's range, which is 0%, an averaged PLR classified as normal can represent that the PLR corresponding to the majority of the CPE KM samples within the five minutes interval is normal. Therefore, the broadcasting method can be carried with a much less consequential risk of uncertainty when the target data is labelled as normal.

There is one type of machine learning algorithm that is specifically designed to solve similar problems, and that is the novelty detection machines. Novelty detection algorithms are based on semi-supervised learning. When training the model, only normal class data is input, and the model analyses and learns from the input data patterns, then it will determine the normality of the test data based on their deviation from the normal patterns. Such algorithms inspired the design of the RCA system at the CPE end, and eventually became a part of it. When training the RCA system, only the CPE-end KM data corresponding to normal PLRs are filtered out as whitelist data and entered the system for training. When testing, the CPE data no longer needs to be filtered by the target label. The RCA system automatically determines whether a KM value is consistent with that KM's data pattern when the PLR is normal. If it is severely deviated from the behaviour that leads to a normal PLR, it is identified as an outlier that may cause packet loss to occur. In this way, the risk of uncertainty associated with broadcasting anomalous labels can be minimised.

3.4.2 Model Selection

In this system, the approach of entering all interested KMs into a single model is not applied, because even if the occurrence of packet loss anomalies can be accurately predicted, the specific feature that causes the fault still cannot be directly located. Although we could go one step further to find the root cause through model interpretation, this extra step could expose the system to a greater risk of error in what is already a two-tier cascade system. Therefore, we make predictions by means of one model corresponding to a single feature, in this way, the root cause can be directly located from the output.

Eventually, the OCSVM model based on RBF kernel is used to investigate GPS coordinate data. The reason is that based on the limited data, the distance or density-based algorithms such as LOF may identify test samples located between normal sample clusters as abnormal [38], but such locations are often not in the poor coverage area when investigating a single cell. The OCSVM is an algorithm based on decision margin, the ellipse-shaped decision boundary established by the RBF kernel can create a larger normal decision region, and it can also be used for soft-decision [14].

The rest of the KMs are all scalars, and they are all analysed through the LOF model. These KMs have high sample density in the normal regions, so they are suitable for the LOF algorithm that is based on normal sample density.

Theoretically, to eliminate the uncertainty risk, only when the PLR of both uplink and downlink is 0% can it be selected as whitelist data. However, it is mentioned in the design of these algorithms that a small amount of anomalous data should be included in the normal training dataset when performing novelty detection, in order to obtain accurate detection results [20] [14]. The algorithms will recognise and label these contaminations during training. Based on this, instead of strictly setting it to 0% when choosing a normal PLR criterion, we look for a value that is slightly higher. This will deliberately cause a small portion of CPE data to be incorrectly assigned to normal PLR labels, the anomalous KM data causing these abnormal PLRs become outliers in the training data accordingly. However, the specific locations and proportions of these outliers are unknown, so in the evaluation process of this algorithm, a suitable contamination rate parameter will be experimentally selected for the models by means of estimation.

3.4.3 The CPE-End RCA System Design

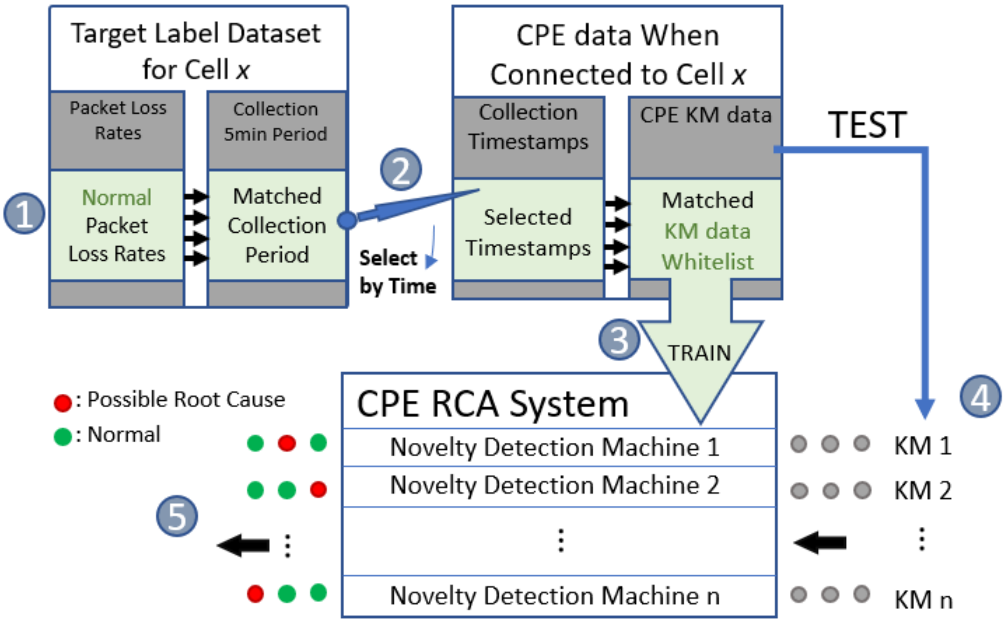

Target Label Dataset for Cell *x*
Packet Loss Rates
Collection 5min Period
1
Normal Packet Loss Rates
Matched Collection Period
2
Select by Time
CPE data When Connected to Cell *x*
Collection Timestamps
CPE KM data
Selected Timestamps
Matched KM data Whitelist
TEST
3
TRAIN
4
●: Possible Root Cause
●: Normal
CPE RCA System
Novelty Detection Machine 1
Novelty Detection Machine 2
Novelty Detection Machine n
5
KM 1
KM 2
KM n

Figure 8: The CPE-End RCA System Workflow

Figure 8 illustrates the design of the RCA system based on the whitelist concept. The root cause investigation on the CPE end is conducted with the cell as the basic unit. At the first step, after being extracted from a certain cell's database, both uplink and downlink PLR data is manually labelled together by a criterion that is suitable for the novelty detection. If the PLR of both uplink and downlink data are normal under that criterion, they will be selected together with the time interval in which they were recorded.

In the second step, from the pre-selected CPE data when the device is connected to the cell specified in the first step, all collection timestamps that are included in the normal PLR data's time intervals are located. The KM data corresponding to these timestamps are then added to the CPE whitelist, these KMs are explained in section 3.2.2.

In the third stage, the novelty detection models are engaged. Each KM to be studied is assigned to a dedicated novelty detection machine. After the model set is arranged, the KM data in the whitelist is fed into the corresponding detection machine for training.

After the novelty detection system is trained, the fourth step is using the system to test the abnormality of the CPE-end KM data that has not yet been labelled. When the input data deviate considerably from the training data, the system highlights the detection results on the location of these data, indicating that this is an abnormal value under the PLR standard.

Finally in stage 5, an RCA report is generated at the system output. In this report, each individual datum has a corresponding detection result.

Such a system enables the advantage of high-resolution sample in the CPE database to be utilised, and the problem can be accurately pinpointed when any packet loss anomaly occurs. The reason why the output report can be used as the result of RCA is because the novelty detection is performed under the criterion of packet loss. Different from the anomaly or novelty detection on the KMs themselves, this system is trained on the KM data corresponding to normal PLRs, so the detection results are also referring to the packet loss standard. In other words, the deviated performance of data detected by this system is likely to cause anomalies in the PLR.

3.5 Algorithm 2: Cell-Level Automatic Root Cause Analysis

Along with the PLR data, there are other KPIs recorded in the cell databases, which mostly records the KPIs for the PDCP layer of the network. These metrics are recorded every five minutes, which is much lower than the sampling frequency on the CPE side. Therefore, instead of locating temporary events that lead to packet loss, investigations at the cell level will zoom out to explore the relevant factors that is affecting the PLR for a relatively lengthy period.

In this section, the choice of model will be discussed first. Next, methods for localising fault will be introduced. Finally, the originally designed RCA system for cell level based on the selected model is presented and explained step by step.

3.5.1 Model Selection

At the cell level, the sampling frequency inconsistency problem described in section 3.4 no longer exists. Each PLR label in the cleaned cell level database corresponds to a set of KPIs. After such limitations are removed, a more comprehensive and efficient supervised analysis system can be pursued. In the RCA system designed for the cell level, machine learning algorithms are still used, but the purpose of these models is no longer to classify data. In this scenario, they are used as analysers to learn patterns in the data, and to collect information from it that would help advance the process of root cause analysis.

Since the main purpose of the model is to analyse the data, the selected model should have high interpretability, which means the importance of input features can be quantified. In addition, the quality of the model and the process of decision making need to be quantified or visualised as results. A classic eligible model is the decision tree. The quality of the decision tree model can be quantified by the number of leaf nodes or layers, the embedded feature importance scoring algorithm can explain the model well and be used as the result of RCA, and more importantly, the model itself can be visualised. These properties greatly satisfy the requirements of this RCA system.

Some algorithms developed from decision trees are also interpretable in terms of feature importance, including Random Forest and Gradient Booster. Both algorithms are based on a method of bagging multiple trees and outputting them in the form of a single tree through voting or merging. Since the original intention of these design is to mitigate limitations of the decision tree itself, such as the bias for high cardinality, they perform better than the decision tree in prediction accuracy wise. On the other hand, as the concept of greedy algorithm in decision tree is abandoned, these models also take less resources at runtime [39]. However, the nature of ensemble trees makes the decision-making process of these models more difficult to be visualised, and even if the final model can be visualised, the meaning of the results is less clear than that of decision trees.

In summary, although the prediction accuracy of the latter two models is proved to outperform decision trees, from the perspective of data analysis, decision trees are still the best choice since it can provide users with the most understandable model explanation. Eventually, the CART Decision Tree classifier algorithm is adopted for this system. The limitations can be minimised by other means in the system.

3.5.2 Tracing Root Causes by Feature Selection

The most important part of the output information is to quantify the impact of each interested KPI on packet loss, which can be achieved through a variety of model explanation methods. The decision tree itself has a scoring system for feature importance. It scores a feature's importance based on its contribution to the information gain and weighted by sample size. The CART Decision Tree's data impurity criterion is Gini, when a parent node is split to two child nodes by the threshold set on the selected feature (KPI), the information gain for this split is calculated by [40]:

$$\frac{N_P}{N} \times \left(Gini_P - \frac{N_L}{N_P} \times Gini_L - \frac{N_R}{N_P} \times Gini_R \right) \qquad (12)$$

- N is the total number of samples of the training data.
- N_p is the number of samples contained in the parent node.
- $Gini_p$ is the data impurity in current node (See Section 2.2).
- N_L and N_R is the number of samples contained in the new left and right-side child nodes after split.
- $Gini_L$ and $Gini_R$ is the data impurity in the new left and right-side child nodes after split (See Section 2.2).

The weighted information gain obtained from such a split is counted in the total contribution of the feature on the parent node that are used to formulate the split rule.

However, the built-in feature importance ranking system has certain flaws, the most serious one being the multicollinearity between the input features. On the parent node, the separation rule can be formulated by any one of the strongly correlated features, causing the features that can be used to create the same rule in the model to be not unique. Even though this has no effect on the predicted outcome, since a feature's importance is calculated by the reduction in data impurity when the feature is used as the separation rule, the importance for all mutually correlated features will be generally decreased [41]. This problem exists in the model itself and is difficult to solve even by means of interpretation outside the model. Therefore, before putting into training and testing, the entire dataset will be tested by Pearson's correlation first. If it is found that some features' data are strongly correlated, only one of them will be reserved as the feature to be fed into the system.

Another limitation is that when the cardinality of the variables in the training dataset is different, the impurity decreases based scoring system is biased towards high cardinal data [42]. Cardinality is a characteristic of judging the number of different categories in a piece of data. The cardinality of numerical data is high because each different value can be regarded as a category. However, a special case exists where the cardinality of the numerical data decreases when the same value appears repeatedly in the data.

To address such a limitation, the permutation feature importance method is introduced. This method is a black-box interpretation technique, it is used when testing the model. The scoring system determine the importance of a variable by randomly shuffles the test data of it and keeps the data of other input variables unchanged, then observes the effect of this behaviour on the prediction results. If shuffling the values of an input variable has a minor effect on the model's predictions, then it proves that the variable's impact on PLR prediction is insignificant. If the effect is opposite, it proves that this variable plays a key role in the system's decision-making process and is of great importance [31]. This measurement is done with the "*permutation_importance*" library provided by scikit learn [43], a more detailed introduction to this approach can be found in the documentation of this library.

At the output end, the cardinality test results of the input KPIs will be output together with the feature importance rank calculated by the two methods, and the user decides which interpretation method to take according to the cardinality report.

3.5.3 Cell Level RCA System Design

Since the cell station is in the middle of the source and the destination in the network, each KPI in the cell database records the performance of uplink and downlink separately. Therefore, the packet loss problems will also be studied separately to achieve targeted analysis. Even so, they can still be automatically analysed through the same system. The only requirement is to assign the RCA object argument as "Uplink" or "Downlink" in the system's setting argument, and the system will select the corresponding target and features.

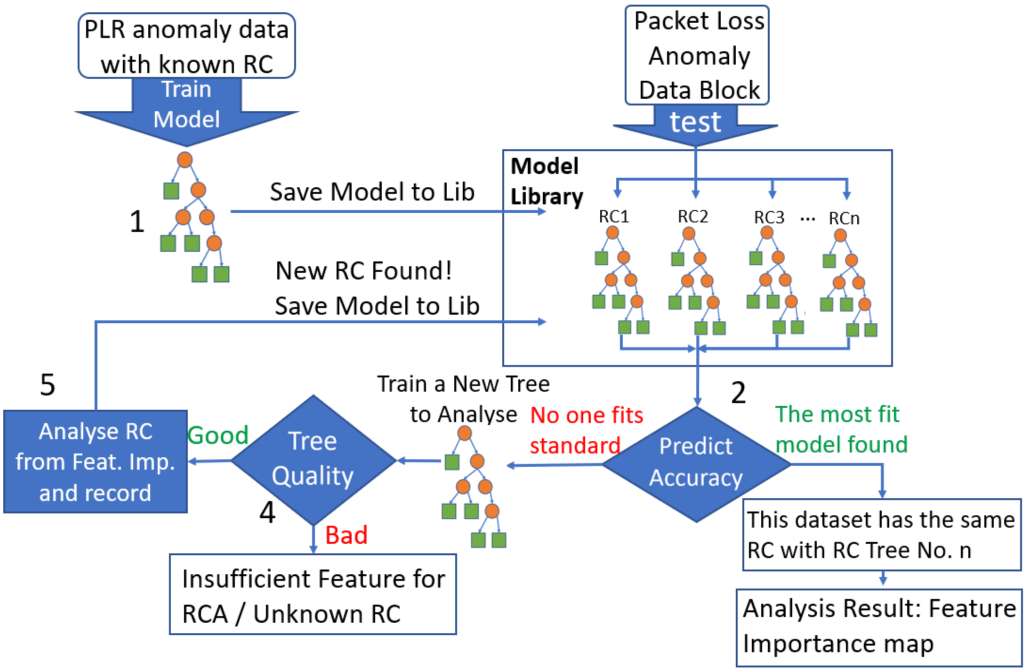

PLR anomaly data with known RC
Train Model
1
Save Model to Lib
New RC Found! Save Model to Lib
Packet Loss Anomaly Data Block
test
Model Library
RC1
RC2
RC3
...
RCn
5
Analyse RC from Feat. Imp. and record
Good
Tree Quality
4
Bad
Insufficient Feature for RCA / Unknown RC
Train a New Tree to Analyse
No one fits standard
2
Predict Accuracy
The most fit model found
This dataset has the same RC with RC Tree No. n
Analysis Result: Feature Importance map
3

Figure 9: Workflow of the RCA System at Cell Level

Figure 9 shows the system design for cell level automatic RCA. In the first stage, *N* decision tree models are trained on data with known root causes, and these trained models are stored in a model library. At this stage, decision tree models are stored as carriers of known abnormal patterns, makes the library $L = \{RC1, RC2, \ldots RCN\}$.

In the second stage, a block of labelled data with unknown root cause of the anomaly is fed into the system as test data. Each model in the model library performs binary classification on this data to predict anomalies, and the predicted results are compared with the actual labels of the test data to calculate the accuracy of the predictions for each model in Equation (13), this test is performed by the "accuracy_score" function provided by Scikit Learn [44].

$$\alpha_{model_x} = accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1\ (\hat{y}_i = y_i) \qquad (13)$$

Through the prediction accuracy of these models $\{\alpha_{model_1}, \dots, \alpha_{model_N}\}$, Equation (14) distinguishes that which model carries the most consistent data pattern with such a block of unseen data. If the prediction accuracy of the winner model reaches the pre-set standard σ, *e.g.*, 90%, then it means that the RC of this anomaly is highly likely to be matched with the RC of the winning model (15), and the information of this model will be directly output as the result of RCA.

$$\alpha_{model_j} = \max(\{\alpha_{model_1}, \dots, \alpha_{model_N}\}) \qquad (14)$$

$$\alpha_{model_j} \geq \sigma \rightarrow Match_{Model} = model_j \qquad (15)$$

At this stage, the decision tree acts as a classifier.

After the matched model is found, the third step is to produce a more comprehensive analysis at the output. The matching model is visualised, and the built-in feature importance scoring mechanism outputs the quantified feature importance. Outside the model, the permutation importance also explains the model for reference. In this step, the decision tree plays the role of the evaluator.

If the system is unable to find a matched RC model for the input data at the third step, a more in-depth analysis of this data is required. In the fourth step, a new decision tree model is trained on the test data, and the quality of the model priorly determines whether the system can find the root cause from this data. If the model is too complex, it will prove that the system is unable to find the root cause of the packet loss issue through the existing data or features. In this step, the decision tree model assesses the possibility of tracing root cause through this system.

Finally, if this abnormal data can train a high-quality model, it means that the information analysed by the system is accurate and significant. The decision tree feature importance and permutation importance scores for input KPIs are then visualised along with the cardinality test for reference. After being reviewed by the user, if the system output is decided to successfully locate the source of the problem or representative to certain specific RC, the model will be stored in the model library, and the system can quickly output the RCA result when an anomaly caused by the same factors occurs in the future. In this step, the decision tree model plays the role of the analyser.

Chapter 4 Evaluation for RCA Algorithm 1

The evaluation of the RCA Algorithm designed for CPE-end KMs is presented in this chapter. The data used for this evaluation are first introduced. Next, a set of experiments for system validation and model parameter selection is designed. The optimal parameters obtained from this experiment set is used to output the final analysis results. Finally, these RCA results and room for improvement of the system are discussed.

4.1 Data

The KM data used in the training and testing of the RCA algorithm 1 are included in a dataset provided by Vilicom in the form of a *.csv* file, which contains the data in the CPE device recoded from January 21, 2022, 01:19:38 (UTC) to January 24, 2022, 09:41:00 (UTC). On the other hand, the data recorded in the databases of the nine eNB cells are also provided, both uplink and downlink PLR data are extracted from these databases as the targets. The provided cell databases include data collected from January 21, 2022, 16:15 (UTC) to January 31, 2022, 23:50 (UTC).

In the CPE dataset, the name of the cell to which the CPE modem is connected is recorded for each set of KM samples. Since this system is designed to be trained and tested with data from a single cell, the CPE data is primarily sorted by the cell's name that the device is connected. Due to the equipment is in various operating areas for different durations, the distribution of available samples corresponding to each cell is not uniform. The number of samples obtained by connecting different cells within the time range covered by the experiment is shown in Table 1.

Cell Name	Number of Samples in CPE dataset
VME23	7860
VME32	5802
VME11	5504
VME33	3644
VME13	2985
VME12	474
VME22	369
VME21	272
VME31	156
Unknown/Out of Range	50
Grand Total	**27116**

Table 1: Sample Distribution from the Cell Aspect

For domain-based novelty detection such as OCSVM, the sample size used for model training needs to be as large as possible to cover a variety of situations, so as to ensure the model's actual performance [38]. Although the data put into the training dataset accumulate over time, we can only obtain a limited number of samples by the time of the evaluation is performed. Therefore, in this experiment, only the data from the cell connection with the largest number of samples are used to evaluate the model, the name of the eNB cell is VME23.

4.2 Experiments

4.2.1 Experiment Aim

A complete validation of the semi-supervised system is difficult as no reference set of known RC data is available. However, the effectiveness of the system can be demonstrated by the performance when predicting data with normal PLRs. Therefore, the performance evaluation will be conducted based on the concept of cross-validation by splitting the whitelist into training and testing data.

On the other hand, as mentioned in Section 3.4.2, the normal standard of PLR should be manually formulated to control the uncertainty risk. A threshold slightly higher than 0% is preferred, in order to allow a portion of anomalous PLRs to be wrongly classified as normal, the KM data outliers causing these abnormal PLRs are the contamination in the training data. The proportion of outliers in the training dataset under each KM defines the "contamination parameter (*c*)" of their novelty detection model. Due to the specific number and location of outliers under each KM is unknown, a reasonable range of *c* values is estimated first, the *c* value in this range that can make the system perform best is selected experimentally. In consideration of reducing complexity, this experiment assumes that the data under each KM contains a same share of outliers (*c*), so as to conduct performance analysis from a system-wide perspective.

4.2.2 Evaluation Method

If a whitelist containing only normal data is separated into training and test datasets, and the training dataset contains all normal data patterns, the trained novelty detection model should recognise all test data as normal. When the training data are declared to contain a proportion of contamination *c*, the model's evaluation criteria for the training data will become stricter [45]. If the proportion and distribution of outliers contained in the test data are consistent with the training data, then the model should also find outliers with the same proportion of *c* in the test data. This condition is approximated by a statistical method in the actual cross-validation process. At the output side, there will be an error in the proportion of anomalies found by the models compared with the pre-set *c*. We use this error as the evaluation standard and select the *c* that leads to the smallest error as the final parameter for results demonstration.

In the CPE whitelist generated when the device is connected cell VME23, for each KM, 80% of the data are used as training data, and 20% of the data are used as test data. This whitelist separation process is randomised, which means that samples in the training and testing datasets are randomly drawn from the whitelist [46]. The positive rate (α) of each novelty detector (for each KM) is expressed by the proportion of predicted normal samples in the test data:

$$\alpha = \frac{N_p}{N_s} \times 100\% \tag{16}$$

- N_p is the number of samples predicted as normal (positive)
- N_s is the total test samples for the corresponding KM

The whitelist contains a proportion of outliers, but the amount of these outliers in the whitelist is unknown. For each KM, we desire that the pre-set contamination proportion (c) is included in both training and testing data with the same distribution, but the training and testing datasets obtained by a single random separation obviously cannot meet this requirement. To approximate this condition, the experiments are performed by the Monte Carlo (MC) method. This uncertainty mitigation technique is more specifically discussed in the work of Hicks *et al.* [47]. In a set of MC experiments, the process of random whitelist splitting, model training and testing, and positive rate measurement is performed by $\boldsymbol{R}$ repetitions. The mean positive rate of the $\boldsymbol{R}$ tests for each KM's novelty detector α_{KM} is used as the result of this set of experiments:

$$\alpha_{KM} = \frac{\sum_{i=0}^{\boldsymbol{R}-1} \alpha_i}{\boldsymbol{R}} = \frac{\overline{N_p}_{KM}}{N_s} \tag{17}$$

The positive rate of the entire system (α_{sys}) is calculated by the mean accuracy of all KM models in the system (18). Since one packet loss rate corresponds to one set of KM data, the training data sample size for each KM is the same. Therefore, α_{sys} can also represent the proportion of the MC mean number of normal samples observed in the total training samples involved in the system.

$$\alpha_{sys} = \frac{\sum_{i=0}^{N-1} \alpha_{KM_i}}{N} = \frac{\sum_{i=0}^{N-1} \overline{N_p}_{KM_i}}{N_s \times N} \tag{18}$$

- N is the total number of KPI.

Every model in a well-performing system should be able to accurately predict that there are c proportion of outliers contained in corresponding KM's test data. Therefore, the criterion for system performance is defined as the error between the pre-set contamination proportion and the actual detected mean contamination proportion:

$$\delta = \left| c - \left(1 - \alpha_{sys}\right) \right| \tag{19}$$

4.2.3 Experiment Design

The specific experiment design is as follows:

1. Set the constant PLR threshold to 0.05 (5%) and select normal PLRs from the cell dataset.
2. Select the available KM data by their timestamps contained within the five-minute collection window of the normal PLRs, then use these data to form the CPE whitelist.
3. Arrange 10 sets of experiments with different contamination proportion estimation $c \in [0.001, 0.02)$, with a step of 0.002. For a single set of experiment, the contamination parameter of each novelty detection model is assigned to the corresponding estimation.
4. In each set of experiments, the system will be trained and tested with Monte-Carlo method on the CPE whitelist dataset corresponding to cell VME23, the MC repetition is set to 2000. The result positive rate (α_{sys}) vs. contamination parameter (c) curve will be plotted.
5. From the positive rate-contamination pairs recorded in step 4, the system's performance criterion δ can also be calculated. Hence, a δ vs. contamination rate parameter (c) plot can be made. The contamination parameter that leads to the smallest δ will be adopted.
6. The system output and the results statistics under this parameter will be demonstrated and discussed.

The algorithm implemented for conducting such experiments is illustrated in Algorithm 1:

***Algorithm 1**: Experiment for Contamination Parameter Selection*

```
Initialize Dataset: CPE_whitelist, Contamination_List, Accuracy_list
Initialize variable:  Repetition
for Contamination in Contamination_List do
    for Repetition do
        Randomly split whitelist to train (80% samples) and test (20% samples) dataset
        Train novelty detectors with KM train data
        Novelty detectors predict KM test data
        Count discovered anomalies
        Calculate $\alpha$ for each KM
    End for
    Calculate ach KPI's Monte Carlo mean $\alpha_{KM} = \frac{\sum_{i=0}^{R-1} \alpha_i}{R}$
    Calculate system $\alpha$ for "Contamination"  $\alpha_{sys} = \frac{\sum_{i=0}^{N-1} \alpha_{KM_i}}{N}$

    Add to Accuracy_list

End for
Plot Accuracy_list vs. Contamination_List curve
Calculate $\delta$
Plot $\delta$ vs. Contamination_List
```

4.3 Results and Discussion

4.3.1 Cell Database Packet Loss Rate Selection

Figure 10 shows the results of the manual labelling of PLRs in the VME 23 cell database. The figure on the left contains all the uplink and downlink PLR samples provided by this cell, totalling 1286 samples. Among them, there are 1011 samples marked as normal, which are shown in green dots. The remaining 275 samples are anomalies and are shown as red dots in the figure. On the right graph, the area containing the normal samples is zoomed in, and the decision boundary of the normal samples is marked, which means that the normal threshold of the uplink and downlink packet loss rates is set to 0.05 (5%).

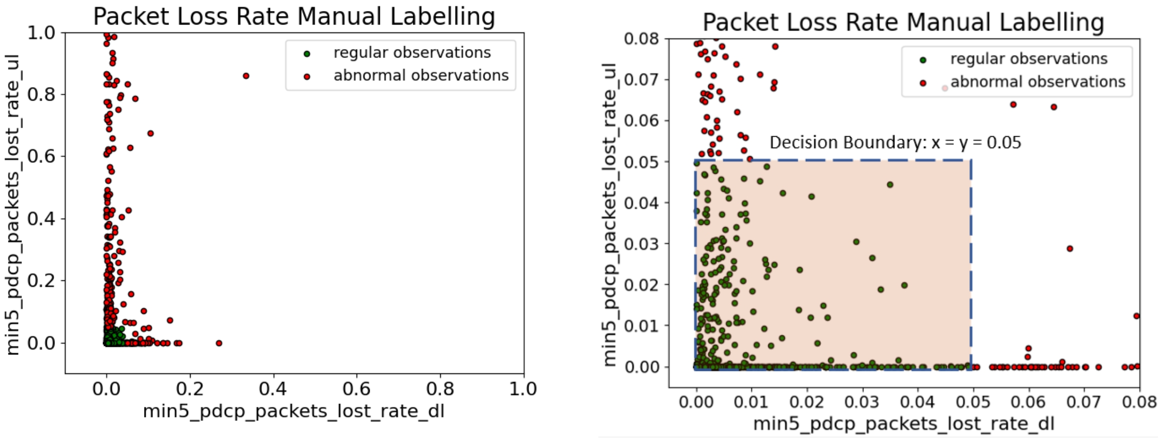


Figure 10: Uplink and Downlink Packet Loss Rate Labelling for Cell VME 23 Dataset, All Packet Loss Rate Samples range from 0% to 100% (Left) and Zoom-in (0% to 8%) Demonstration of the Decision Boundary and Normal Samples (Right)

4.3.2 Results of Experiment for Contamination Parameter

All available CPE data referencing to VME23 cell contained in the collection window of the selected normal PLR are extracted. A CPE whitelist with a sample size of 4547 for each KM is constructed. The experiment for contamination rate parameter (*c*) selection is conducted with this whitelist dataset.

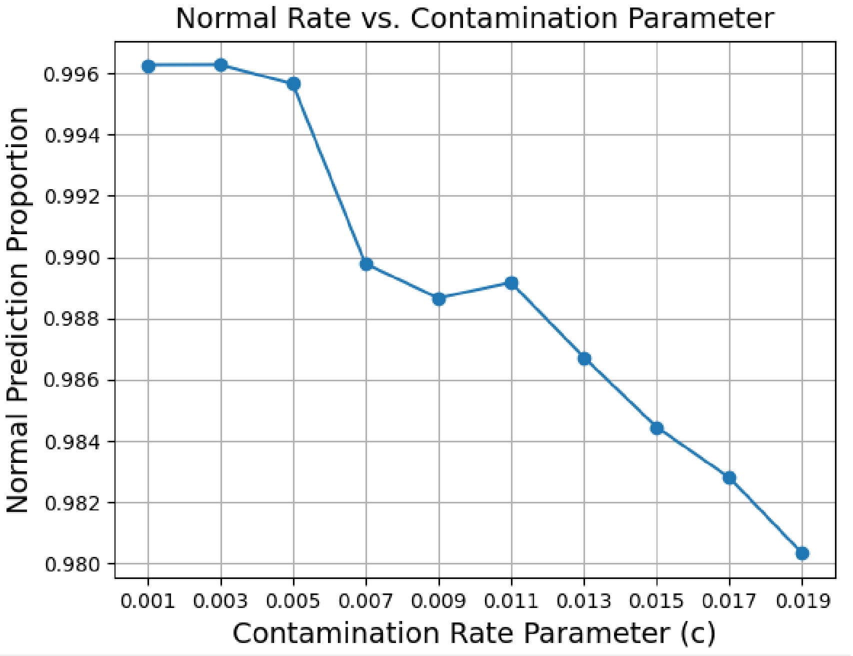


Figure 11: The Plot of Predictions' Positive Rate against Pre-set Contamination Parameter

Figure 11 demonstrates the positive rate vs. contamination parameter *c* plot obtained by the Monte Carlo method. Each point in the figure is the average result of 2000 repetitions of randomly splitting the whitelist and system predictions under the corresponding *c*. Theoretically, the greater the proportion of data defined as contaminated in the training data, the stricter the normality criteria set by the novelty detector is, resulting in more anomalies being found in the test data where the anomalies are distributed in the same way as the training data. In this experiment, such a phenomenon is manifested as a drop in the normal rate of the test results, which can be observed in Figure 11. In the results, as *c* increases, *i.e.*, the proportion of anomalies defined in the training data increases, the predicted normal rate in the test results decreases accordingly.

However, such a relationship is not linear, which means that the system behaves differently under different *c*-parameters. The system is not sensitive to the changes in *c* when *c* is less than 0.05. When *c* is greater than 0.05, the normality of the detection results firstly declines dramatically, then rebounds. It reaches the peak when *c* is set to 0.011, and then declines at a relatively stable rate. To investigate further, the error between the predicted anomaly rate and the pre-set contamination rate for each point is calculated using the method mentioned in Section 4.2.2.

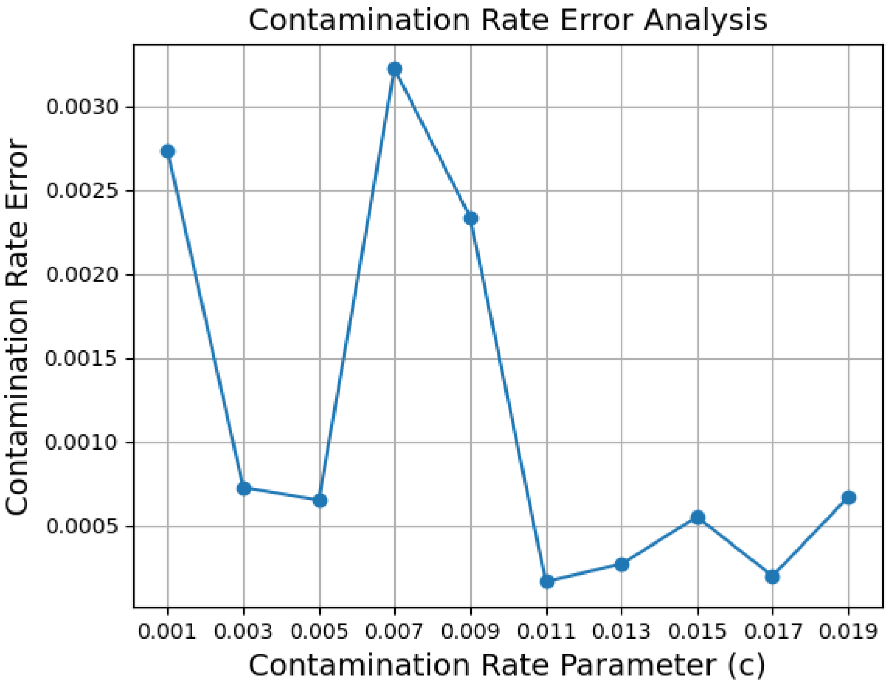


Figure 12: Result of Contamination Error Analysis

From the results demonstrated in Figure 12, it can be observed that the contamination rate error δ is very unstable at low *c* values. When c is set between 0.011 to 0.019, the resulting contamination error no longer fluctuates drastically, and the average error decreases. The contamination rate error δ is the smallest when the *c* value is 0.011, which is only 0.000167.

This also means that from the sample size wise, when the *c* value is 0.011, the average prediction accuracy of each novelty detection model in the system for this CPE whitelist reaches 99.98%, and under any other *c* value considered in this experiment, the mean accuracy is not lower than 99.68%. This proves that the system excels in the whitelist cross-validation experiments. From the normality patterns defined in the training data, it can accurately classify the test data.

4.3.3 System Output

The system output under the optimal contamination parameter *c* obtained in section 4.3.2 is illustrated in this section. The testing data sample size is 3125 for each KM.

GPS Coordinates

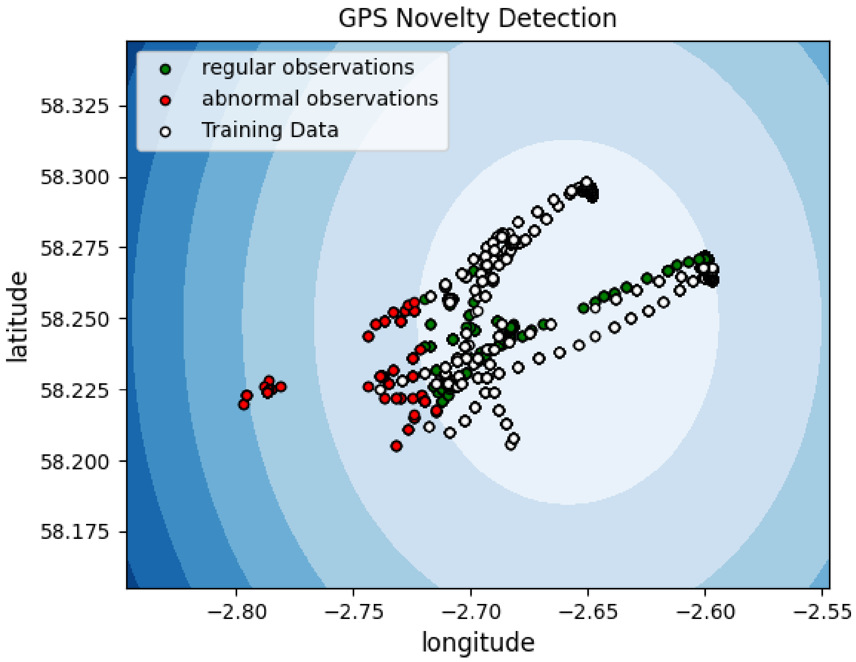


Figure 13: Result of GPS coordinates Novelty detection

Figure 13 shows the training data and the hard-decision results of the GPS coordinate novelty detection. The decision boundary generated by the OCSVM based on the RBF kernel is also shown in the figure. Several concentric ellipse-shaped decision domains are divided. The light blue ellipse in the centre represents the normal area, the points distributed in this area will be labelled as normal. The colour of the decision region outside the normal area gradually deepens as the number of margin layers increases. Each layer represents a different decision score range. This score can be used when making soft decisions, representing the abnormality degree of the points scattered in the corresponding decision area.

The white dots in the graph represent the training data, which was used to create the decision margins. A small portion of the training data is located outside the central normal area. They are identified by the algorithm as the contamination in the training data. The green and red dots represent the coordinates in the training data determined as normal and abnormal, respectively. Among them, the coordinates in the vicinity of the major samples in the training data are accurately identified as normal locations. The coordinates that are remote to the training data are marked as anomalies, which means that they may lead to a surge in the packet loss rate.

Although the CPE dataset with the largest amount of data has been selected in this experiment, these data are still not enough for the GPS coordinates RCA based on empirical research. The model needs more accumulated relevant data for training in future runs to achieve more accurate judgement. In the current experiments with limited data, soft decisions are the preferred method. For example, in Figure 13, although the points in the second-layer decision region are marked as abnormal in the hard-decision case, they are still not far from most normal data, and packet loss anomalies due to weak coverage are unlikely to occur at these coordinates. The small clusters scattered in the dark blue area to the left of them are farther from all normal data, they have higher decision scores, and therefore are more likely to cause the occurrence of PLR anomalies.

Other KMs

Before the whitelisted data corresponding to the scalar KMs are fed into models for training, they are first standardised, which means that they are recorded in the form of z-scores:

$$Z = \frac{x - \mu}{\sigma} \tag{20}$$

- x is the KPI value to be standarised.
- μ is the mean of all whitelist data under the same KM.
- σ is the standard deviation of all whitelist data under the same KM.

In the subsequent testing process, the test data is also standardised, except that the μ and σ that appear in the z-score formula of the training data are still calculated from the training data, which is the same as that used in the training data z-score formula. Therefore, the test data are standardised with respect to the training data. This process is done with the “Standard Scaler” library provided by Scikit-Learn [48].

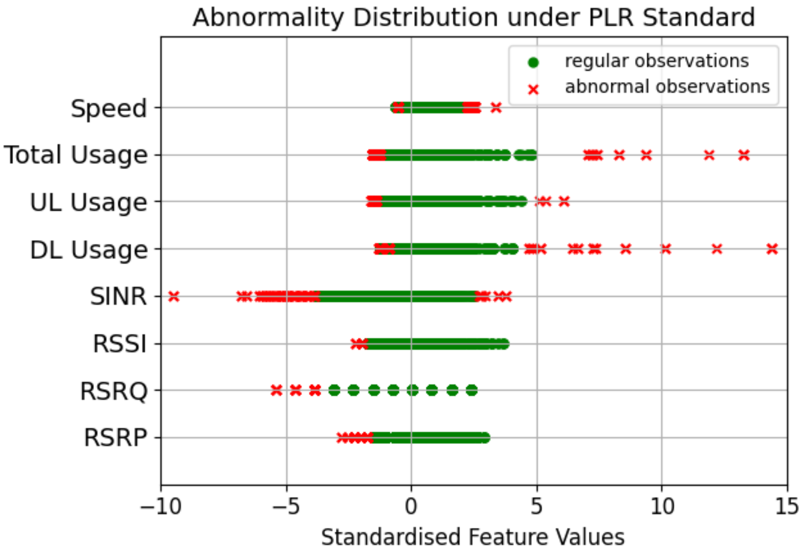


Figure 14: KMs’ abnormality distributions

Figure 14 shows the results' distrubution of RCA performed on all other scalar KMs. Since each KM is assigned a dedicated novelty detector, and their values are all scalars, their judgment results can be displayed in the form of a set of one-dimensional data. However, the values of different KM vary widely and are in different units, to make it easier to display their results in one graph, the data shown in the figure have not been reverse transformed, which means they are still shown with z-scores with regards to the training data.

The standardised data can show the distribution more intuitively. The 0 scale in the figure represents the mean value of the training data corresponding to each KM, the negative scales (*x*) represent that the data point is lower than that mean value by *x* times of standard deviation of training data, and the meaning of positive scales are vice versa.

As can be seen from the graph, these training data are marked as anomalous when their values are deviated from the mean of training data to a certain extend. However, some KMs are case sensitive. As mentioned in the data description in Section 3.2.2, when the device's movement speed is exceptionally low or static, it will not cause packet loss problems even if it is determined as abnormal by the system. The usage data is temporarily ignored during late night. In addition, some exceptionally high SINR values are also labelled as anomalies, which are invalid results because a high signal-to-noise ratio indicates that the signal strength is much higher than the interference strength, which is what makes the network more stable. Therefore, values that are labelled by the system to be outliers in these cases are removed from the final report.

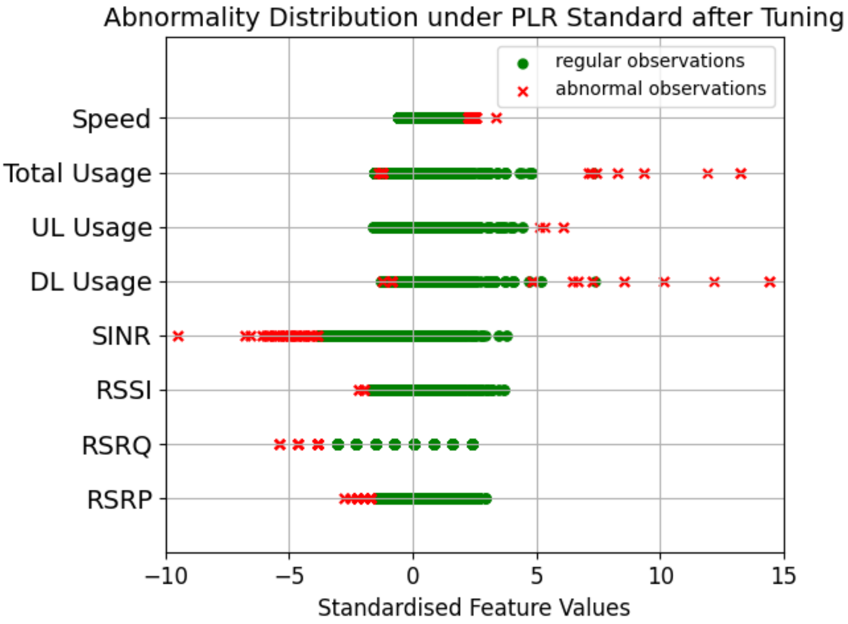


Figure 15: Adjusted KMs' abnormality distributions

The points marked as abnormal in Figure 15 will be output as the final result. In terms of speed, When the test data for movement speed is higher than the mean of the training data by around 0.4 times the standard deviation of the training data, it will be regarded as abnormal by the system.

The anomalies for the usage metrics are found in the data with high values, this type of anomaly indicates that network resources are heavily used at the time when it was collected. If such a situation persists for a long time, it may cause congestion and delay in the network, which is also one of the possible causes of the increased PLR.

Among the four KMs related to the electromagnetic environment, the data marked as abnormal are all extremely low values. An irregularly low SINR means that the interference and noise in the channel is too large relative to the signal power. When this KPI is marked as anomaly by the system, it means that there is electromagnetic interference in the RF environment where the device is located, which is likely to disrupt the information transmission chain. If the RSSI is too low, it means that the power of all co-channel signals or noise on the in the spectrum, including referenced signals and interference, are low, which is also a manifestation of weak links. The abnormally low RSRP proves that the referenced signal power is weak. Finally, RSRQ is a measurement of the referenced signal quality, an abnormally low value indicates that the strength of the current signal link is less than the strength of other co-channel links. Table 2 records the outlier range of each RF condition indicator in the detection results.

Key Metrics	**Anomaly Range in RCA Results**
SINR	2.2 to 11.1 dB
RSSI	-65 to -64 dBm
RSRP	-98 to -94 dBm
RSRQ	-18 to -16 dB

Table 2: Range of RCA Numerical Results for RF Conditions

Result Statistics and Discussion

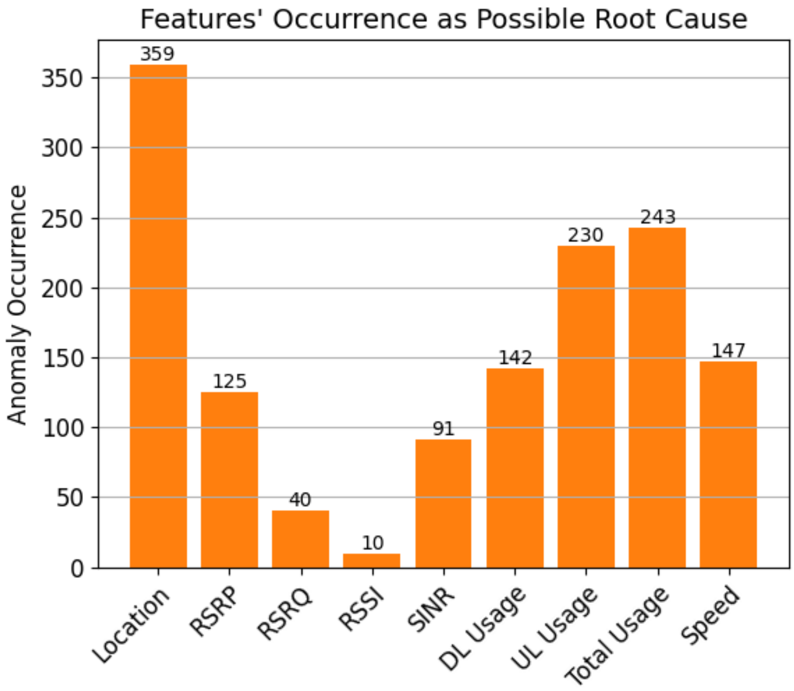


Figure 16: The Statistics of The RCA Results

The system generates an investigation report at the output end, which includes the location of all anomalies and the KM data corresponding to them. The statistics of the report are shown in Figure 16, the amount of data labelled as abnormal under the PLR standard in each KM is calculated and demonstrated. Among the 3125 sets of testing KM data, a total of 608 sets of data contain at least one KM that has been identified as the possible RC of PLR abnormality. The most frequently identified indicator that may lead to an increase in the packet loss rate is the abnormal geographic location of the device, with a total of 359 abnormal samples.

In the evaluation of the RCA system on the CPE side, a model parameter selection method based on the system performance on the whitelist cross-validation was proposed and used. With the models established by the selected parameters, the system is trained on the whitelist CPE data that correspond to the VME23 cell, and the factors that may cause the high packet loss rate are found as anomalies from the remaining non-whitelisted data.

The selection of the contamination parameter is based on the assumption that the number of outliers under each KM is equal. However, the experimental results for the testing dataset show that the amount of data identified as abnormal for each KM is quite different. Instead of setting the same contamination parameter for all models in the system, if the parameter settings of the novelty detection machine corresponding to each KM can be individually researched and optimally selected with a similar experimental style, the system's prediction results for unseen data could be more accurate.

On the other hand, the system only makes judgments on new data based on the currently available whitelist data, which may not be enough to cover the complete range of normal KM values. With the accumulation of data, a whitelist with a complete range and uniform distribution can be generated by statistical methods. The models trained through this whitelist can be stored, so that can be directly used for detection on new data.

Chapter 5 Evaluation for RCA Algorithm 2

This chapter will demonstrate the performance of the system designed for RCA on the cell level KPI data from two aspects. First, in the absence of a known or matching RC model, the system will analyse a fragment of data, and the decision tree model trained on these data will be visualised, together with the data analysis report. The second aspect is to investigate whether the system can accurately identify the matching RC of the anomalous data when the RC is pre-stored in the RC model library, and distinguish it from other RC models.

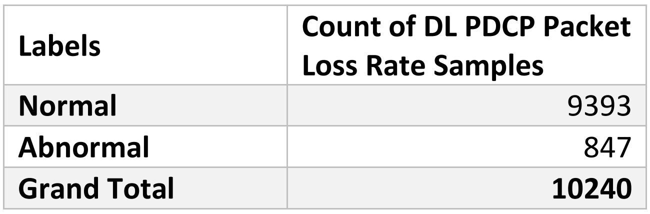
5.1 Data and Feature Selection

The KPIs in the eNB cell databases for uplink and downlink are recorded separately, and the system is also designed to investigate them individually. In this chapter, the system will only be evaluated with downlink KPI data. There are two reasons for doing this. Firstly, in addition to the metrics related to data usage, the PDCP layer SDU delay and drop rate measurements are only provided for downlink data. These two KPIs evaluate the performance of data packet delivery. Among the limited KPIs offered, such metrics are valuable since they may be more relevant to packet loss rate. On the other hand, the system is highly inclusive to the types and quantities of features involved. Even if the number and types of input KPIs are different, the system will still function and achieve the same analysis effect. Due to limited space, the KPI names used in the presentation of results are simplified. The correspondence between these simplified names and the actual KPI names recorded in Appendix 2 is shown in Table 3.

Simplified KPI Names	Actual KPI Names
DL Throughput	Effective Cell Throughput: Downlink
Total Payload	PDCP Total User Data Payload
Data Volume	PDCP User Data Volume All QCI: Downlink
SDU Delay	PDCP SDU Delay: Downlink
SDU Drop Rate	PDCP SDU Drop Rate: Downlink

Table 3: KPI Name Correspondences

The data from all cell's CPU databases are provided from February 1st, 2022, 00:00:00 (UTC) to March 1st, 2022, 00:00:00 (UTC). After missing and null data are removed, a total of 10,240 available samples are generated. Since the system is still in the validation stage, the labelling of target data (Downlink packet loss rate) is still done manually, a PLR threshold is set to 0.03. The labelled data sample size for each class is shown in Table 4.



Labels	Count of DL PDCP Packet Loss Rate Samples
Normal	9393
Abnormal	847
Grand Total	**10240**

Table 4: Labelling Result Statistics

In order to investigate the multi-collinearity issue that exists between input features, the correlations between KPIs are tested using the Pearson correlation coefficient.

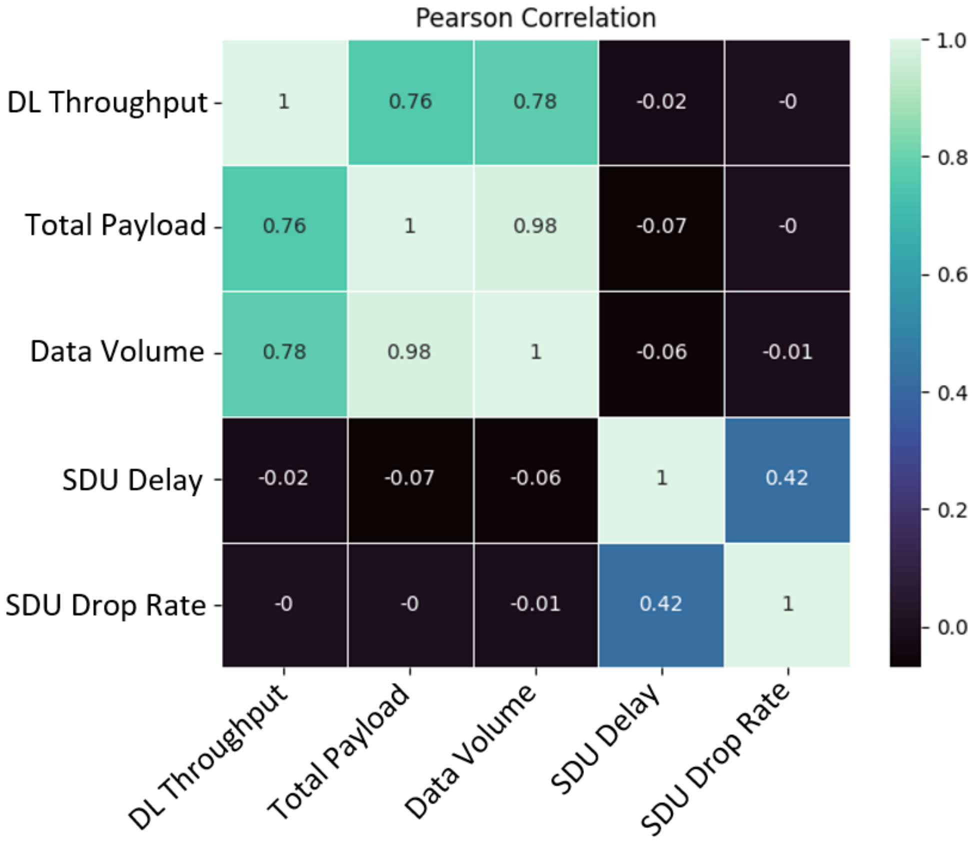


Figure 17: Results of Correlation Test for the Input KPIs

The correlation measurement results presented in Figure 17 indicate that the data of usage related KPIs share a strong correlation. These three KPIs are downlink throughput, downlink data volume, and uplink and downlink total data payload. The correlation coefficients between these features are above 0.75, which means they have strong positive collinearity. In order to avoid a misleading feature importance analysis, only one KPI should be remained. The total data payload is eventually selected as the KPI characterising the data usage. The reason is that the data payload represents the amount of content information that is intended to be transmitted in data packets, it intuitively reflects the usage of the network by users. In addition, it measures the total amount of uplink and downlink data payload, which is helpful in studying the impact of total network resource usage on the packet loss rate.

5.2 Scenario 1: No Matching RC Model Found

The first situation to be evaluated is when the system cannot find an RC model in the model library that matches the input abnormal data segment. In this case, the system uses these data to train a new decision tree model, then visualises and interprets the model to gain information that helps RCA.

In the cleaned KPI dataset, the packet loss rate labelled as abnormal is not evenly distributed. In most cases, these samples are clustered in periods where high packet loss occur intensively. Since the system is designed to analyse such anomalous patterns, the anomalous data segment used in this evaluation are selected by intercepting the data in one of such periods. Eventually, the downlink KPI data recorded in the VME33 site station on 19 February 2022 from 12:35:00 to 23:50:00 (UTC) are selected for this evaluation. In these data, there are 67 samples labelled as normal PLR and 68 abnormal samples, which is a set of data with a balanced sample size for each target category. This dataset is then used to train a new decision tree model.

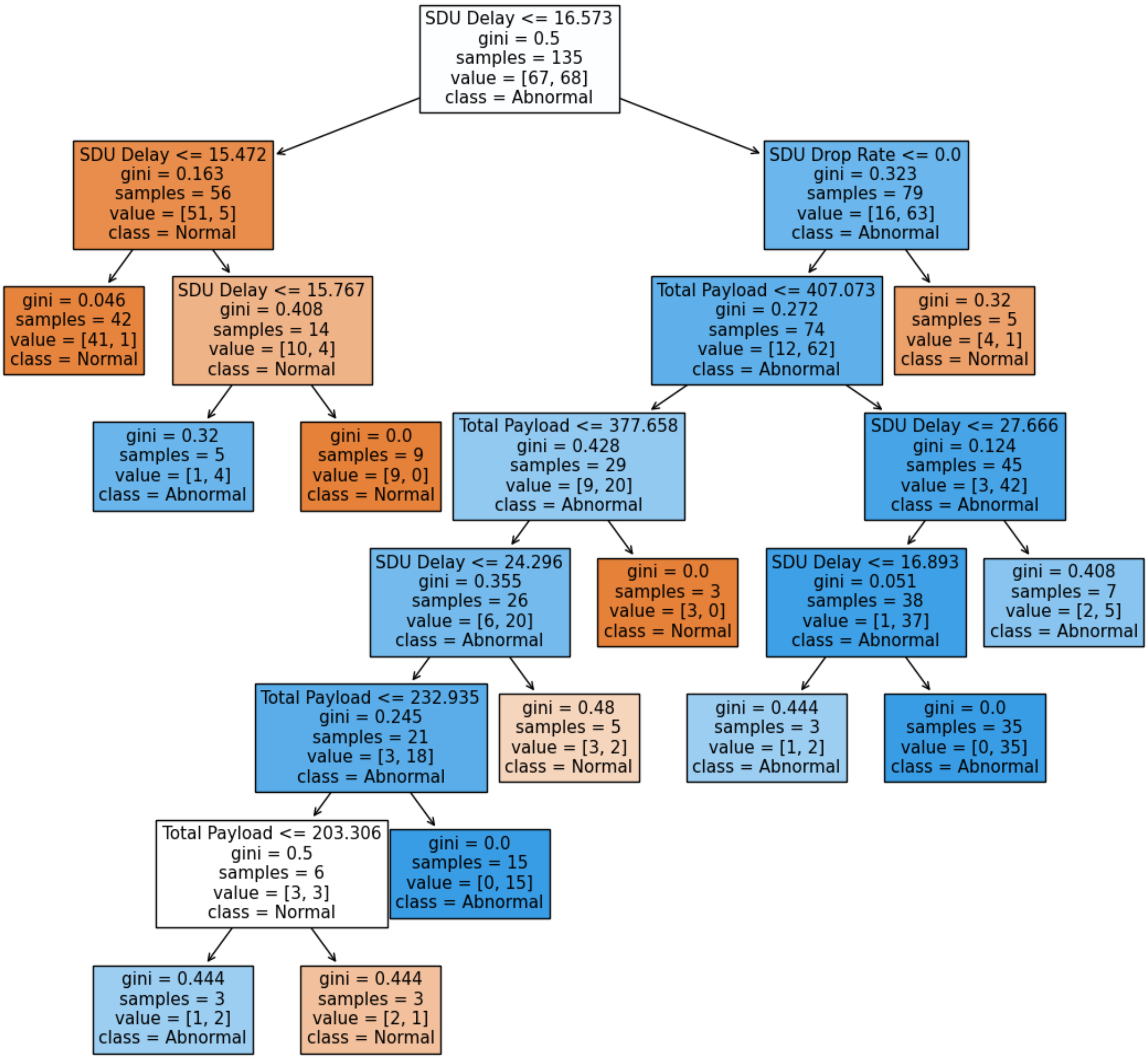


Figure 18: The Decision Tree Model Trained by Selected Data

Figure 18 shows the visualised decision tree model. In this model, each node contains a set of data sorted by classes. The sample size under each class is shown by the value tag on each node, and the class with the largest sample size is noted in the class tag. It can be observed that at the root node, the number of normal and abnormal samples are almost equal, and the data impurity is the highest in this state. Next, based on the Gini impurity decrease, the decision tree searches for features and thresholds that can further minimise data impurity at each separation node.

Under two conditions, the node can be used as leaf nodes and no longer need to be separated further. The first one is that the data contained in the node is exceptionally pure, such as the leftmost node in the second layer and the rightmost node in the fifth layer (with the root node being the 0th layer). These nodes are the most weighted leaf nodes. They contain a large amount of data with the same target class, and extremely few or no samples from other classes. Such a leaf node represents that the path that leads to it can effectively distinguish different classes of data, so these nodes are high-quality nodes.

Another situation is when the data samples under all classes contained on a node are few relative to the total amount of data, and the data impurity is still high. In this case, it is pointless to conduct further separation. Such nodes are inferior nodes, if they appear frequently in the model and the occurrence is much higher than that of high-quality leaf nodes, it means that the overall quality of the model is poor. This situation is often accompanied by numerous layers, which also indicates that the algorithm is unable to organise the data by their class through the existing input features.

This model completes the classification of data through a seven-layer decision tree structure. There are five leaf nodes with Gini sample impurity lower than 0.1, accounting for 41.7% of the total number of leaf nodes. On the two most weighted leaf nodes, the sample size of a single class accounts for more than 50% of the corresponding class’s total sample size. The evidence proves that the model can distinguish data by existing features. After the feature importance analysis, the model is stored in the RC model library.

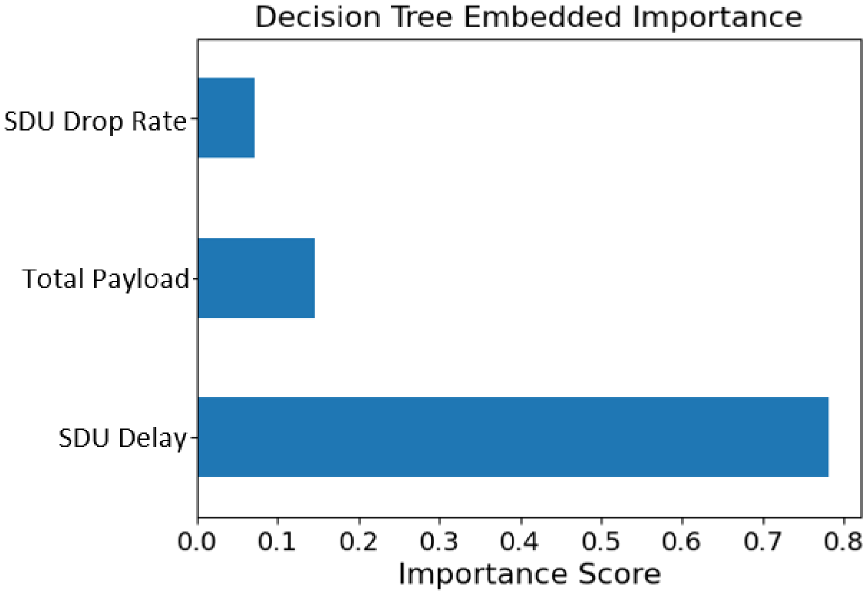


Figure 19: KPI's Importance Ranking by the Embedded Method

The importance scores calculated by the features' contribution to the data impurity decrease are shown in Figure 19. Such analysis result shows that the downlink PDCP SDU delay has the greatest impact on the packet loss rate, much larger than the other two KPIs. The total data payload of uplink and downlink is the second, and the effect of downlink PDCP SDU drop rate on PLR is the slightest. However, as mentioned in chapter 3.5.2, the scoring system is biased towards highly cardinal data. Although this problem is not common when all features are continuous numerical data *i.e.*, all feature's data have high cardinality, there may still be the data of a certain KPI repeatedly appears the same value, which reduces the cardinality of the data, causing the scoring system to discriminate against this KPI. In order to solve this problem, cardinality test and permutation importance test are included in the system output for users' reference.

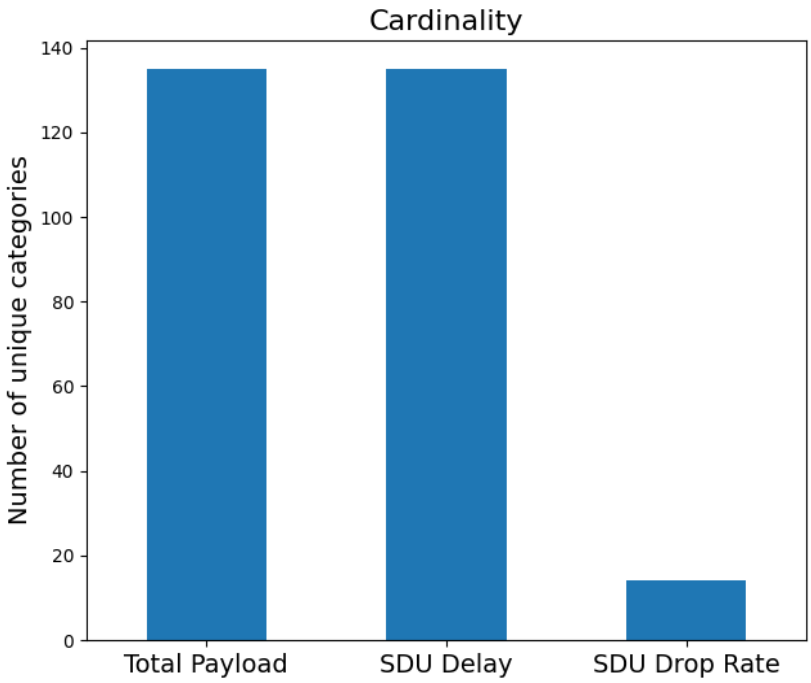


Figure 20: Results of Cardinality Test on Input KPIs

Figure 20 shows the results of the cardinality test, the cardinality of the SDU drop rate data is significantly lower than the other two KPIs. This is due to the infrequent occurrence of SDU drop, which leads to more samples with a value of 0, hence reduce the diversity of the data. In this case, the embedded scoring system is likely to evaluate the KPI unfairly. Therefore, feature importance needs to be measured using permutation importance.

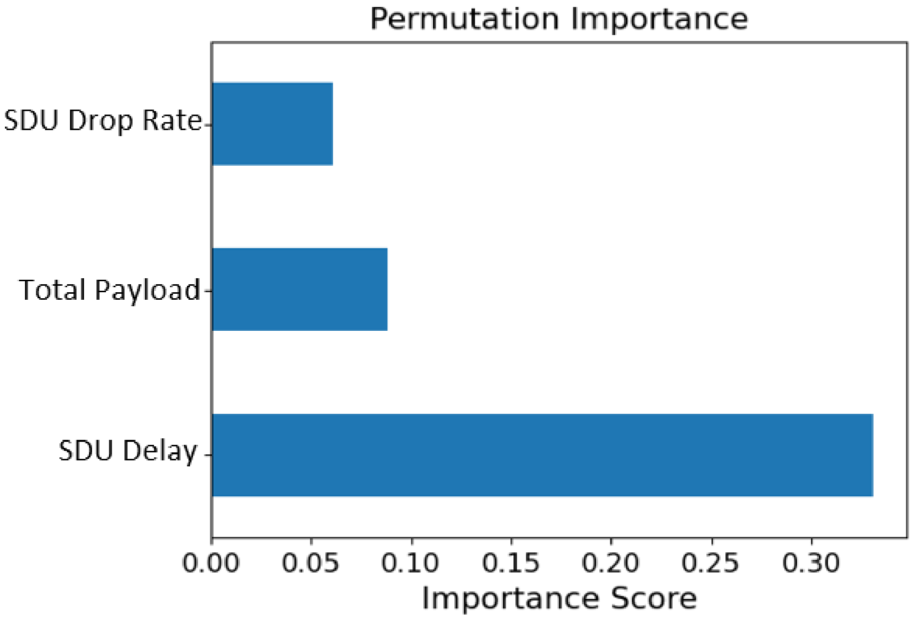


Figure 21: KPI's Permutation Importance

Figure 21 shows the feature importance calculated by the permutation importance score for each KPI. The permutation importance of input KPIs are measured from the test dataset, in this case, the test data is the training data. When the model is tested with training data, the KPI's importance is calculated by the strength of the impact of shuffling its data on the model's predictions. Compared with the embedded importance calculation result in Figure 20, although the SDU drop rate measured by this method is still the least important feature, the importance of this KPI is closer to that of the data payload, and the importance gap between it and the SDU delay is also narrowed.

Although the scoring mechanisms of these two methods are different and the obtained scores are not to scale, as feature importance ranking systems nevertheless, their original intention is to intuitively show the degree of influence of a KPI on the PLR and contrast it with that of other metrics entered the system. Therefore, the score is insignificant in the results of this type of analysis, the information of one KPI’s level of importance among all KPIs can be acquired through visual judgment.

5.3 Scenario 2: Repeated Anomaly for the Same RC

An advantage of this RCA system is the ability to quickly identify recurring root causes from past observations. When the RC of the abnormal data segment is the same as a certain RC model stored in the model library, the prediction accuracy of the model for this segment of data should be exceptionally high, and significantly different from that of other models. This section verifies the system from this aspect.

Since no anomalous PLR dataset with known RC is currently available, the validation will be done with a dummy dataset. We assume that in a piece of data, when the SDU delay is higher than 89 milliseconds, it will cause an abnormal PLR. In order to form a dataset that can simulate this situation, data with SDU delay below 89ms and normal PLR labels are randomly selected from the entire dataset. Next, the data with SDU delay higher than 89ms and anomalous PLR labels are also selected and mixed into the obtained normal samples. The dataset constructed by this method contains 96 samples in the normal target class and 55 samples in the abnormal target class for each KPI. Such a dataset is then manually divided into two parts, in which 48 sets of normal class KPI samples and 32 sets of abnormal class KPI samples are selected as training data. The remaining data are used as test data.

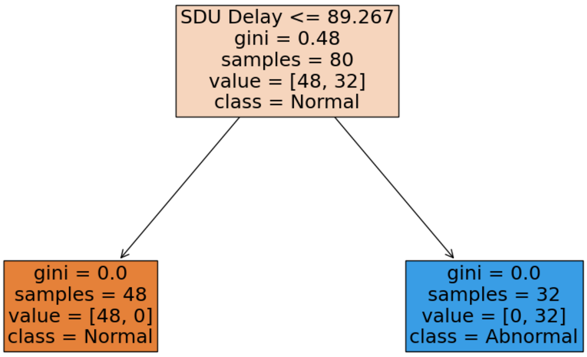


Figure 22: The Decision Tree Model Trained by Scenario 2 Training Data

Figure 22 shows the decision tree model generated from the training data. The algorithm sets the threshold of SDU delay to 89.267ms. The threshold is not exactly 89ms because the lowest SDU delay value labelled as anomaly in the training data is 89.267ms. The model completely separates all data by class using only one split, and the Gini impurity of both leaf nodes is 0. Such a model indicates that when the value of SDU delay in a set of KPI data is higher than 89.267ms, regardless of other participating KPIs' values, the model will define the PLR class corresponding to this set of data as abnormal. Such a model meets the criteria we defined when selecting training data, so it can be stored in the RC model library.

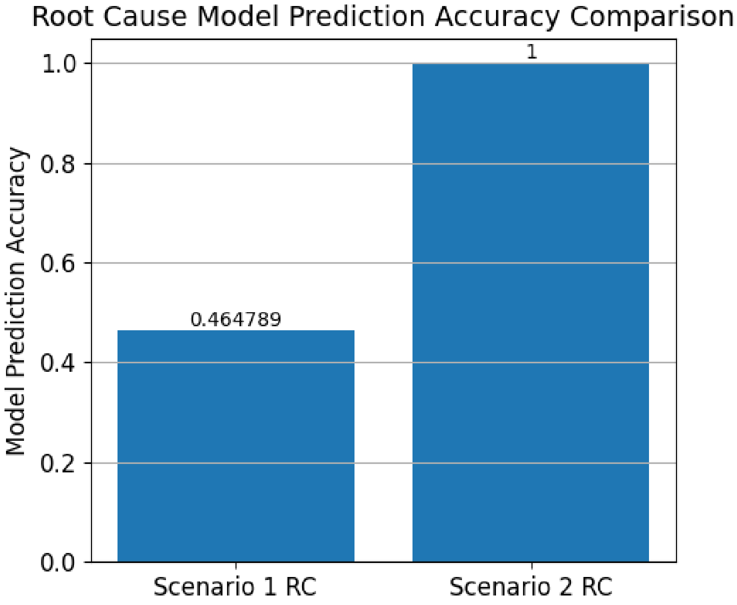


Figure 23: Prediction Accuracy of Each RC Model when Testing on the Test Data

When entering the test data into the system, the system evaluates this data on each RC model in the model library, and the test accuracy given by each model is shown in Figure 23. The results show that this dataset has a test accuracy of 46.48% when testing on the model derived from scenario 1. A prediction accuracy of 100% is reached when the data is tested on the model derived from scenario 2. This indicates that the system can accurately match the RC of this test data to the scenario 2 model. Although the information stored in scenario 1 also means that PLR anomalies are caused by high SDU delay, the prediction accuracy of the model is much lower than the model trained in scenario 2 due to the different thresholds set. This result confirms that the system can accurately match the simulated abnormal data to the correct RC model, which carries the identical root cause data pattern with the input data.

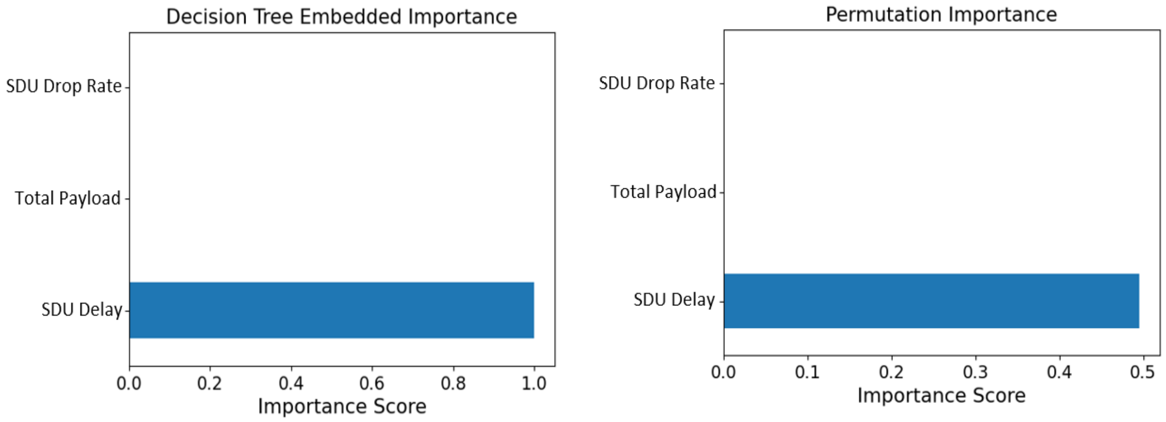


Figure 24: The KPIs’ Embedded Importance Scores (Left) and the Permutation Importance Scores (Right)

After match RC model is found, the embedded and permutation feature importance results are output by the system, which are shown in Figure 24. Since the model achieves complete purification of the data only by a threshold set on the SDU delay, the model-embedded importance scoring system determines it as the only feature with the ability to reduce impurity, hence it is the only feature that has an importance score. Under the permutation importance rule, changing the values of other input KPIs does not have any effect on the prediction results, so the SDU delay also becomes the only important feature.

Chapter 6 Conclusion

The design of the packet loss root cause analysis system for the Moray East offshore LTE network is demonstrated in this article. With the support of machine learning algorithms, the two subsystems search for the root cause of packet loss in the relevant metrics recorded in the CPE and eNB (Cell) databases, respectively. These two subsystems are verified and applied to the data captured over the network, which is provided by Vilicom Ltd. The results showed that the system is capable of locating faults that may cause packet loss. From the perspective of objectives listed in section 1.2, all goals except goal 4 have been achieved to some extent.

Goal 1 is not fully achieved since the automatic anomaly detection mentioned in Section 3.3.1 is not practically utilised to model evaluation. All PLR data labelling currently applied to the testing is done by manually setting PLR normality thresholds.

Goal 2 is fully accomplished. Each one of the network elements, namely CPE and eNB, is allocated to a RCA subsystem. The subsystem designed for the CPE data adopts novelty detection algorithms that operate based on data with normal PLR only, which solves the problem of inconsistent sampling frequencies of target and features. The trained CPE end subsystem can locate the abnormality in the unlabelled KM data under the PLR standard. The subsystem designed for the cell level is based on the decision tree algorithm, which can both analyse data for RCA and quickly locate known RC based on saved RC models.

Goal 3 is partly accomplished. Since the CPE database records the root-level metrics, the anomalies detected in these metrics can be directly used as root causes. Therefore, when new data is entered to a well-trained CPE end subsystem, it can automatically locate the RC of the packet loss. The cell level subsystem stores more complex KPI data patterns referring to certain predefined RCs, and when these RCs cause packet loss again, the system can automatically match the input anomalous data to its RC. However, in the case where no matched pre-defined RC is found, the KPIs involved at current stage cannot directly provide the RC. The analysis results need to be reviewed by experienced specialists as the results provide them with information that is useful for further RCA.

Due to the time limitation, the function of automatically capturing abnormal data segments in real-time stated in Goal 4 has not been developed in the system.

Goal 5 is completed. Verification and testing of the two subsystems are presented in Chapters 4 and 5. The subsystem on the CPE end was validated with the normal dataset. After a set of experiments is carried out through the Monte Carlo method, the results showed that the prediction accuracy on normal data is at least 99.68% under all tested contamination parameter, and under the optimal contamination parameter, the accuracy reached 99.98%, which means that the system is confirmed to be able to accurately predict the PLR state of normal data. The novelty detectors set by the optimal parameter are used to evaluate each input KM data collected from the offshore LTE network. At the output side, the anomalies that may lead to packet loss can be found.

The cell level RCA system was first used on KPI data analysis tests, and the results proved that the system could quantify and visualise the impact of input KPIs on PLR. The verification of the automatic detection of known RCs is currently remained in the simulation stage due to the lack of reference data. The verification results showed that the system can match the simulated abnormal data to the corresponding RC model in the model library and distinguish it from other models, with an accuracy of 100% on matched model and 46.5% on the other pre-saved model.

6.1 Future Works

On the CPE end subsystem, we are interested in the further research on novelty detection algorithms for different metrics. The algorithm currently used is selected based on limited data, but when longer-term data is acquired, we can use statistical methods to make the whitelist cover a more comprehensive normal range of data, and the contamination rate will also be more controllable. In this case, the detection algorithm suitable for each metric may be different from the ones we currently used. On the other hand, as mentioned in 4.3.3, the parameter optimisation experiments for each novelty detection model can be performed individually to obtain specialised best contamination parameter.

For the cell level subsystem, we hope to obtain the data with abnormal PLR corresponding to known RCs in the future and fill the RC model library through the models trained with these datasets, so that we can evaluate the RC matching performance of the system on the real data.

From a system point of view, the long-term goal of this project is that the system can be applied in the industry as a network analysis tool. In order to achieve this purpose, the algorithm should be packaged reasonably and have a user interaction system to facilitate parameter setting and feature selection outside the system. On the other hand, in order to achieve a higher level of automation, the attribute of automatic anomaly data interception stated in Goal 4 can also be developed in future system upgrades together with the untested automatic anomaly detection on PLR, so that the completed system can eventually be applied to real-time data, and accurately provide high-quality analysis results when packet loss problems occur.

References

[1] Moray Offshore Renewable Power, “Moray East Powering Scotland's Future through renewable energy,” Moray Offshore Renewable Power, 2020. [Online]. Available: https://www.morayeast.com/. [Accessed 29 Mar. 2022].

[2] Vilicom, “THE MORAY EAST OFFSHORE PROJECT: Private LTE Network Uses,” Vilicom, 2021. [Online]. Available: https://www.vilicom.com/wireless-network-case-studies/private-lte-network-for-moray-east-offshore/. [Accessed 30 Mar 2022].

[3] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *JMLR* , vol. 12, pp. 2825-2830, 2011.

[4] M. Nohrborg, “LTE,” 2010. [Online]. Available: https://www.3gpp.org/technologies/keywords-acronyms/98-lte. [Accessed 25 Jan. 2022].

[5] K. Budigere, N. Panchakarla, B. Chemmagate and S. Roy, “LTE: Long Term Evolution of 3GPP,” 2010.

[6] Vilicom, Interviewee, *Introduction to the Project.* [Interview]. 22 Oct. 2021.

[7] D.Rai and A. Dwivedi, “LTE Theory to Practice- KPI Optimization (A 4G Wireless Technology),” *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 8, no. 2, 2018.

[8] Freescale Semiconductor Inc., “Long Term Evolution Protocol Overview,” Freescale Semiconductor Inc., Oct. 2008. [Online]. Available: https://www.nxp.com/docs/en/white-paper/LTEPTCLOVWWP.pdf. [Accessed 15 Dec. 2021].

[9] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification And Regression Trees, New York: Routledge, 1984.

[10] J. R. Quinlan, “Induction of Decision Trees,” *Machine Learning 1,* pp. 81-106, 1986, [Online]. Available: http://dx.doi.org/10.1023/A:1022643204877.

[11] J. R. Quinlan, C4.5: programs for machine learning, San Francisco: Morgan Kaufmann Publishers Inc., 1993.

[12] M. Moon and S. K. Lee, “Applying of Decision Tree Analysis to Risk Factors Associated with Pressure Ulcers in Long-Term Care Facilities,” *Healthcare Informatics Research,* vol. 23, no. 1, pp. 43-52, 2017 [Online] Available: https://doi.org/10.4258/hir.2017.23.1.43.

[13] W. T. Tseng, W. F. Chiang, S. Y. Liu, J. s. Roan and C. N. Lin, “The Application of Data Mining Techniques to Oral Cancer Prognosis,” *Journal of Medical Systems,* vol. 39, no. 59, 2015. [Online] Available: https://doi.org/10.1007/s10916-015-0241-3.

[14] B. Schölkopf, R. C. Williamson, A. Smola, J. S. Taylor and J. Platt, “Support vector method for novelty detection,” *Advances in neural information processing systems,* vol. 12, pp. 582-588, 1999.

[15] M. Hearst, S. Dumais, E. Osuna, J. Platt and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their Applications,* vol. 13, no. 4, pp. 18-28, July-Aug. 1998, [Online] Available: https://doi.org/10.1109/5254.708428.

[16] A. Bounsiar and M. G. Madden, “Kernels for One-Class Support Vector Machines,” in *International Conference on Information Science & Applications (ICISA),* 2014.

[17] R. Vlasveld, “Introduction to One-class Support Vector Machines,” 12 Jul. 2013. [Online]. Available: http://rvlasveld.github.io/blog/2013/07/12/introduction-to-one-class-support-vector-machines/. [Accessed 25 Jan. 2022].

[18] R. ZHANG, S. ZHANG, S. MUTHURAMAN and J. Jiang, “One Class Support Vector Machine for Anomaly Detection in the Communication Network Performance Data,” in *5th WSEAS Int. Conference on Applied Electromagnetics, Wireless and Optical Communications,* Tenerife, Spain, 2007.

[19] P. Winter, E. Hermann and M. Zeilinger, “Inductive Intrusion Detection in Flow-Based Network Data Using One-Class Support Vector Machines,” in *2011 4th IFIP International Conference on New Technologies, Mobility and Security,* 2011, pp.1-5.

[20] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, “ LOF: Identifying Density-based Local Outliers,” in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data,* Dallas, TX, U.S.A, 2000.

[21] J. Lee, B. Kang and S. H. Kang, “Integrating independent component analysis and local outlier factor for plant-wide process monitoring,” *Journal of Process*

Control, vol. 21, no. 7, pp. 1011-1021, 2011, [Online] Available: https://doi.org/10.1016/j.jprocont.2011.06.004.

[22] E. H. Budiarto, A. E. Permanasari and S. Fauziati, “Unsupervised Anomaly Detection Using K-Means, Local Outlier Factor and One Class SVM,” in *5th International Conference on Science and Technology (ICST),* Yogyakarta, Indonesia, 2019.

[23] A. M. Miyim and A. Wakili, “Performance Evaluation of LTE Networks,” in *2019 15th International Conference on Electronics, Computer and Computation (ICECCO),* 2019, [Online] Available: https://ieeexplore.ieee.org/document/9043271.

[24] D. Baltrunas, A. Elmokashfi, A. Kvalbein and Ö. Alay, “Investigating packet loss in mobile broadband networks under mobility,” in *2016 IFIP Networking Conference (IFIP Networking) and Workshops,* Vienna, Austria, 2016, [Online] Available: https://ieeexplore.ieee.org/document/7497225.

[25] E. Masum and J. Babu, “End-to-End Delay Performance Evaluation for VoIP in the LTE network,” M.Sc. Thesis, Dept. Telecomm. Eng. Blekinge Institute of Technology, Blekinge, Sweden, 2011.

[26] J. Wang, H. Zhou, Y. Li, Q. Sun, Y. Wu, S. Jin, T. Q. S. Quek and C. Xu, “Wireless Channel Models for Maritime Communications,” *IEEE Access,* vol. 6, pp. 68070-68088, 2018, [Online] Available: https://ieeexplore.ieee.org/document/8528349.

[27] Y. Huo, X. Dong and S. Beatty, “Cellular Communications in Ocean Waves for Maritime Internet of Things,” *IEEE Internet of Things Journal,* vol. 7, no. 10, pp. 9965-9979, 2020, [Online] Available: https://ieeexplore.ieee.org/document/9072093.

[28] V. Forsberg, “AUTOMATIC ANOMALY DETECTION AND ROOT CAUSE ANALYSIS FOR MICROSERVICE CLUSTERS,” M.Sc. Thesis, Dept. Computer Sci. UMEA University, Umea, Sweden, 2019.

[29] S. L. ÁLVAREZ, “Anomaly Detection and Root Cause Analysis for LTE Radio Base Stations,” M.Sc. Thesis, Dept. E.E.&C.S., KTH, STOCKHOLM, SWEDEN, 2018.

[30] C. M. Roelofs, M.-A. Lutz, S. Faulstich and S. Vogt, “Autoencoder-based Anomaly Root Cause Analysis for Wind Turbines,” *Energy and AI,* vol. 4, 2021 [Online] Available: https://doi.org/10.1016/j.egyai.2021.100065.

[31] D. Rengasamy, B. C. Rothwell and G. P. Figueredo, “Towards a More Reliable Interpretation of Machine Learning Outputs for Safety-Critical Systems Using Feature Importance Fusion,” *Appl. Sci.*, vol. 11, no. 24, 2021 [Online] Available: https://doi.org/10.3390/app112411854.

[32] T. Josefsson, “Root-cause analysis through machine learning in the cloud,” Dept. I.T., Uppsala Univ., Uppsala, 2017.

[33] T. Kohonen, “The self-organizing map,” *Proceedings of the IEEE,* vol. 78, no. 9, pp. 1464-1480, 1990 [Online] Available: https://ieeexplore.ieee.org/abstract/document/58325.

[34] D. J. Dean, H. Nguyen and X. Gu, “UBL: Unsupervised behavior learning for predicting performance anomalies in virtualized cloud systems,” 2012 [Online] Available: https://dl.acm.org/doi/10.1145/2371536.2371572.

[35] A. Gómez-Andrades, P. Muñoz, I. Serrano and R. Barco, “Automatic Root Cause Analysis for LTE Networks Based on Unsupervised Techniques,” *IEEE Transactions on Vehicular Technology,* vol. 65, no. 4, pp. 2369-2386, 2016 [Online] Available: https://ieeexplore.ieee.org/abstract/document/7105410.

[36] M. Demetgul, “Fault diagnosis on production systems with support vector machine and decision trees algorithms,” *The International Journal of Advanced Manufacturing Technology,* vol. 67, 2012 [Online] Available: https://link.springer.com/article/10.1007/s00170-012-4639-5.

[37] M. Solé, V. Muntés-Mulero, A. I. Rana and G. Estrada, “Survey on Models and Techniques for Root-Cause Analysis,” arXiv, 2017 [Online] Available: https://doi.org/10.48550/arxiv.1701.08546.

[38] F. Pimentel et al., “A review of Novelty Detection,” *Signal Processing,* vol. 9, pp. 215-249, 2014.

[39] J. Ali, R. Khan, N. Ahmad and I. Maqsood, “Random Forests and Decision Trees,” *International Journal of Computer Science Issues(IJCSI),* vol. 9, 2012 [Online] Available: https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees.

[40] Scikit Learn, “sklearn.tree.DecisionTreeClassifier,” Scikit Learn, Dec. 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn-tree-decisiontreeclassifier. [Accessed 12 Feb. 2022].

[41] S. Chowdhury, Y. Lin, B. Liaw and L. Kerby, “Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance,” 2021, [Online] Available: https://arxiv.org/abs/2111.02513.

[42] C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn, “Bias in random forest variable importance measures: Illustrations, sources and a solution,” *BMC Bioinformatics,* vol. 8, no. 25, 2007 [Online] Available: https://doi.org/10.1186/1471-2105-8-25.

[43] Scikit Learn, “4.2. Permutation feature importance,” Scikit Learn, Dec. 2021. [Online]. Available: https://scikit-learn.org/stable/modules/permutation_importance.html. [Accessed 10 Mar. 2022].

[44] Scikit Learn, “3.3. Metrics and scoring: quantifying the quality of predictions,” Scikit Learn, Dec. 2021. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score. [Accessed 21 Mar. 2022].

[45] Scikit Learn, “2.7. Novelty and Outlier Detection,” Scikit-Learn, Dec. 2021. [Online]. Available: https://scikit-learn.org/stable/modules/outlier_detection.html. [Accessed 15 Mar. 2022].

[46] Scikit Learn, “sklearn.model_selection.train_test_split,” Scikit Learn, Dec. 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html?highlight=train%20test%20split#sklearn.model_selection.train_test_split. [Accessed 22 Feb. 2022].

[47] W. Warren-Hicks, J. P. Carbone and P. L. Havens, “Using Monte Carlo techniques to judge model prediction accuracy: validation of the pesticide root zone model 3.12,” *Environ Toxicol Chem,* vol. 21, no. 8, p. 1570–1577, 2002 [Online] Available: https://pubmed.ncbi.nlm.nih.gov/12152756/.

[48] Scikit Learn, “sklearn.preprocessing.StandardScaler,” Scikit Learn, Dec. 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. [Accessed 18 Mar. 2022].

Appendix 1: Software Used

Filename/Algorithm/ Package	Supplier/Source/Author/ website	Use/Modifications made/ Student written
Pandas	Pandas, pandas.pydata.org	Data processing
NumPy	NUmpy, numpy.org	Data processing
Scikit Learn	Scikit Learn, https://scikit-learn.org/stable/	Machine learning algorithms provider, including OCSVM, LOF, Decision Tree. Data Pre-processing tool. Cited in thesis when used.
Matplotlib	matplotlib.org	Results visualisation
PyCharm	https://www.jetbrains.com/pycharm/	Programming IDE
joblib	https://joblib.readthedocs.io/en/latest/	Saving machine learning models
read.py	https://github.com/menglinyao/RCA-System-for-packet-loss	CPE data selection algorithm developed by student, see thesis 3.2.2, 3.3.2, 3.4
UE_RCA.py		CPE-end RCA algorithm developed by student, see thesis 3.4
PM_Preprocessing.py		Cell data selection algorithm developed by student, see thesis 3.2.1, 3.3.2
Feature_selection.py		Cell-level RCA algorithm developed by student, see thesis 3.5
MC_contam_research.py		Experiment algorithm see section 4.2.3

Appendix 2: Cell (eNodeB) KPI

KPI Names	**Units**	**Recording Methods (5 mins)**	**Measure-ments**	**Specification**
PDCP Packet Loss Rate	Percent	Averaging	Uplink, Downlink	3GPP TS 36.314 version 15.1.0
Effective Cell Throughput	Bit per second	Grand total	Uplink, Downlink	
PDCP Total User Data Payload	Bytes	Grand total	Uplink and Downlink Total	
PDCP User Data Volume All QCI	kbits	Grand total	Uplink, Downlink	
PDCP SDU Delay	ms	Averaging	Downlink	
PDCP SDU Drop Rate	Rate (0-1 scale)	Averaging	Downlink	3GPP TS 32.425 version 9.2.0

Appendix 3: CPE Key Metrics

KM Names	Units	Description	Specification
Longitude, Latitude	Coordinates	GPS Coordinates of the CPE location	-
Temperature	Celsius	Temperature of the CPE device	-
Speed	m/s	Movement speed of the CPE device	-
SINR	dB	Signal to interference and noise ratio	3GPP TS 36.214 version 14.3.0 Release 14
RSSI	dBm	Received signal strength indicator	
RSRP	dBm	Reference signal received power	
RSRQ	dB	Reference signal received quality	
Tx, Rx, Usage	Bit per Second	Upload, download data usage and total bandwidth usage	