December 2020

MULTIVARIATE TIME SERIES UNSUPERVISED ANOMALY DETECTION AND DIAGNOSIS IN 5G NETWORKS

Sreyan Ghosh

Vijay Kataria

MULTIVARIATE TIME SERIES UNSUPERVISED ANOMALY DETECTION AND DIAGNOSIS IN 5G NETWORKS

AUTHORS:
Sreyan Ghosh
Sakshi
Vijay Kataria

ABSTRACT

For a variety of reasons (including, for example, increasing cyber security threats, increased network heterogeneity, the increased use of virtualization technologies, etc.) maintaining the fifth generation (5G) networks of tomorrow will be challenging. To address such challenges techniques are presented herein that support a multivariate time series unsupervised method for anomaly detection using Key Performance Indicators (KPIs) that are derived from various infrastructure level metrics collected from all kinds of networking nodes deployed in 5G networks. Multivariate time series unsupervised anomaly detection is the future of anomaly detection with systems generating real-time time series data but this is an area that has not yet been explored with 5G. This invention introduces anomaly detection in 5G networks at the node level or deployment level instead of simply monitoring anomalous behavior in a particular KPI on a particular node. Additionally, aspects of the techniques presented herein provide a semi-automated method for anomaly diagnosis and deep-dive anomaly analytics in support of better systems of the future. The main focus herein lies in anomaly detection in modern 5G network deployments at the complete node or deployment level using infrastructure level data from 5G nodes to detect and prevent network failures. We present a multivariate time-series unsupervised AI algorithm to solve this problem which helps detect anomalous behavior and in turn facilitates better network troubleshooting and capacity forecasting.

DETAILED DESCRIPTION

Maintaining the 5G networks of tomorrow is set to be a challenging domain due to increasing cyber security threats and widening attack surfaces created by the Internet of Things (IoT), increased network heterogeneity, increased use of virtualization technologies,

and distributed architectures with many 5G nodes consuming significant amounts of memory and central processing unit (CPU) capacity.

Techniques are presented herein that support a software defined anomaly detection system based on neural networks, a subdomain of artificial intelligence (AI), which focuses on acquiring infrastructure level metrics from nodes used in 5G deployments at the infrastructure level (through, for example, bulk statistics from UPF nodes or Structure of Management Information (SMI) statistics generated from Virtual Network Functions (VNFs) or Simple Network Management Protocol (SNMP) polled metrics from transport nodes), calculating KPIs based on the acquired metrics, and feeding the time series KPIs into complex neural network models for anomaly detection and diagnosis to facilitate the detection of early symptoms and thus network failure prevention.

Aspects of the techniques that are presented herein introduce an unsupervised multivariate time series anomaly detection mechanism in support of maintaining and preventing failure in nodes used in 5G network deployments. This technique introduces monitoring anomalous behavior in 5G networks on the complete node level or deployment level by monitoring/using multiple KPIs from a single node or a combination of various KPIs from multiple nodes deployed in the network. Learning complex dependencies between multiple time series in high frequency data is a difficult task demanding complex AI algorithms. While such an approach is the future of anomaly detection in systems generating time series data, it is an area which has not yet been explored in 5G network anomaly detection and analysis. Aspects of the techniques that are presented herein pave the way for improved anomaly diagnosis with the help of better root cause analysis, anomaly severity identification through more accurate anomaly scoring, which when combined with smart rules pave the way for network failure prevention using an semi-automated approach. We can achieve this for example by introducing rules to trigger network orchestration services for node configuration management Further, this approach paves the way for deep dive anomaly analytics which facilitates future capacity forecasting and better systems of tomorrow.

Aspects of the techniques presented herein encompass a number of elements that are of particular interest and note. Various of those elements will be described and discussed below. At multiple places the words nodes, system and deployment are used

interchangeably. A node would mean a particular node in the 5G network deployed. A deployment would mean the entire deployment comprising of various nodes. Our approach can also be used for a combination of KPIs from multiple nodes. Thus, collectively all 3 are referred to as "system" at places in the document.

A first element of interest and note concerns the algorithmic approach to the particular use case.

Aspects of the techniques presented herein apply multivariate time series anomaly detection techniques to metrics that are accumulated from each 5G node that is deployed in a network operator's broad network. This departs from the traditional univariate anomaly detection system which demands separately monitoring every individual component of a 5G node in the deployment(e.g., separately monitoring, among other things, each CPU, random access memory (RAM) usage, etc.), which in turn adds significant confusion to whether just a single component showing anomalous behavior would require any action to be taken on the system. With multivariate anomaly detection the system as a whole is checked to determine if it is showing anomalous behavior which would demand quick actions to be taken before the node, and in turn the network, breaks down. By incorporating a time series approach, anomalies may be detected based on the patterns found in previous time steps and not in, for example, a single time step. This becomes useful in systems producing data which is time-dependent (e.g., CPU usage over the last time steps helps to detect whether the CPU usage at the current time step is anomalous). See, e.g., Figure 1, below.

	KPI 1	KPI 2	KPI 3	KPI 4	KPI 5
Timestep 1	1	6	11	16	21
Timestep 2	2	7	12	17	22
Timestep 3	3	8	13	18	23
Timestep 4	4	9	14	19	24
Timestep 5	5	10	15	20	25

Figure 1: Illustrative Example of Multivariate Time series data

Further, aspects of the techniques presented herein employ an unsupervised approach to the learning algorithm. Modern networks have a very low positive anomaly rate. However, a single failure can cost the carrier a significant amount of money. But the low number of positive anomaly cases makes it difficult for supervised AI algorithms to learn the patterns behind an anomaly. Most previous anomaly detection systems in cellular networks are supervised and require the manual labelling of anomalies by experts and a significant number of positive anomaly cases so that the network can learn what anomaly and non-anomaly data points look like. This is due to the absence of complex and advanced algorithms that are capable of learning. The unsupervised nature of the algorithm that is employed under aspects of the techniques presented herein mitigates the need for manual labelling of the data and the need for a high number of positive anomaly cases in the data. The nature and the complexity of the algorithm helps find complex dependencies and correlations between multiple variables in time series data which, in turn, helps to create a strong representation of the input data with anomalies scored based on the reconstruction error. The model is trained on just non-anomalous data points and the main task here for the model is to recreate the same. Whenever a new data point comes in, the algorithm weights recreate the data point and if it is not well re-created it is deemed to be an anomaly.

Further still, aspects of the techniques presented herein address robustness to noise, anomaly scoring, and root cause identification. From the manner in which the AI algorithm and neural network are framed, the raw data is first converted into signatures using pair-wise correlations and then feed into the network. This specific methodology makes the algorithm robust to all kinds of input noise that are very prevalent in real-time data in large-scale systems, also facilitates better anomaly detection/scoring and root cause identification.

Well-established multivariate anomaly detection algorithms like the Auto Regressive Integrated Moving Average (ARIMA) perform well when the correlation among multiple features is linear and the input process is strictly ergodic. Multivariate time series anomaly detection with unsupervised AI algorithms, although a rapidly emerging field, has not been explored in cellular networks, especially in 5G networks where high volumes of time series data are generated by various components of a node or by various actions of end users. The instant algorithmic approach is mostly a compilation

of prior research made in the AI-based anomaly detection space but its application to 5G networks with infrastructure level node KPIs and it's application to complete node/deployment level anomaly detection is something that has yet been unexplored. Also, the neural network architecture from the number of layers and components used, and the type of layers used at various points, which is mostly dependent on the type of input data in most cases, is novel and fits well to the current use case as described herein.

A second element of interest and note concerns the detection of early symptoms and semi-automated troubleshooting. Aspects of the techniques presented herein detect early symptoms of network failure or outage and inform a network engineer of the root cause of the anomaly and anomaly severity so that he/she can take the appropriate actions. This procedure also facilitates semi-automated network failure prevention when combined with the algorithms root cause identification and anomaly severity identification techniques trigger an action to be taken (e.g., a configuration to be pushed) by network orchestration services so that most problems can be taken care of automatically.

A third element of interest and note concerns anomaly analytics. Aspects of the techniques presented herein provide anomaly analytics which, among other things, aid in the building the better networks of tomorrow and better capacity forecasting.

Aspects of the techniques presented herein encompass a series of functional components supporting our methodology, various of which are illustrated in Figure 2, below, and which are described and discussed in the following narrative.

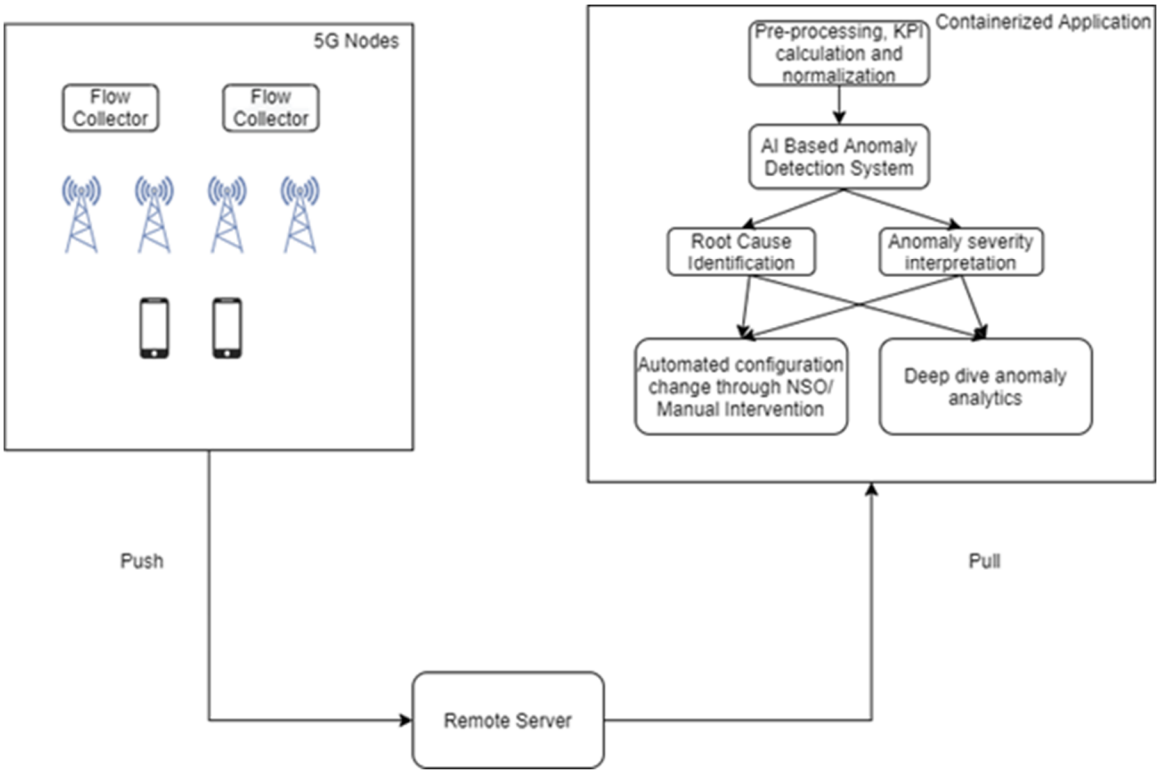

Figure 2: Illustrative System and Data Flow

A first functional component encompasses a data collection module. Such a module focuses mainly on four types of time series KPIs derived from infrastructure level data collected from 5G deployment nodes in support of the anomaly detection task, for example performance KPIs, telemetry KPIs, radio KPIs, microservice KPIs, SNMP polled KPIs, etc. All of these KPIs are calculated from raw metrics that are collected from 5G nodes (through, for example, bulk statistics from UPF nodes or Structure of Management Information (SMI) statistics generated from Virtual Network Functions (VNFs) or Simple Network Management Protocol (SNMP) polled transportation metrics) on an end application using a formula and a combination of metrics (e.g., system throughput) or a single metric (e.g., CPU utilization). Data points are collected from 5G nodes and stored in a centralized server from where an anomaly detection application may pull data at regular intervals (e.g., every five minutes). Depending upon the timeframe in which data is pulled the KPIs may also be aggregated using any aggregation function (e.g., minimum, maximum, average, etc.).

A second functional component encompasses an application module. The application is a containerized application deployed on a cloud server. The application is used for anomaly detection and troubleshooting in real time so that network failures and outages may be prevented. The application has four main tasks, including:

- Data collection and KPI calculation. The time series data collected from the 5G nodes and stored in a remote server is pulled to the application and stored in its local container volumes. The data is then preprocessed and used to calculate various KPIs for which the application already has predefined formulas. The KPIs may be just the same plain preprocessed data or a mathematical formula comprising a collection of various fields from the preprocessed data. The KPIs are then min-max normalized before being input to the model.
- Anomaly detection. The groups of time series KPIs, denoting the performance of a particular node, are fed into the pre-trained neural network for anomaly detection. Additionally, the model may be periodically retrained (e.g., every week) on a separate server.
- Alerting and semi-automated troubleshooting. On the occurrence of an anomaly the system alerts the concerned parties to the anomaly severity and a root cause and checks its set of rules for any predefined rules regarding troubleshooting. If a rule is found the system triggers a network orchestration service and the anomaly is taken care of automatically before the system breaks down.
- Anomaly analytics. The system stores information about all previous anomaly occurrences in a report format including, for example, their root cause, severity, duration, and recovery and troubleshooting mechanism. This helps in building the better networks of tomorrow and may also help in capacity forecasting for the future.

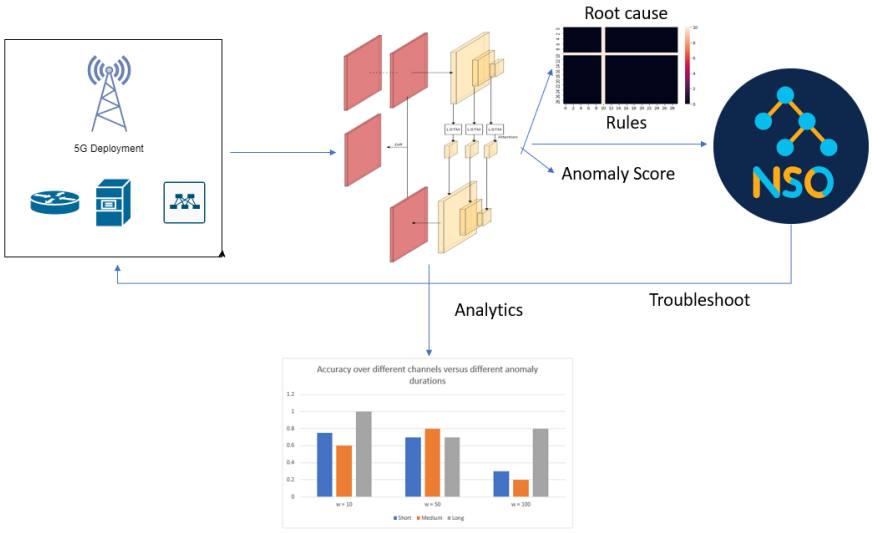

Figure 3: Illustrative Description of the proposed semi-automated troubleshooting procedure

A third functional component encompasses an anomaly detection algorithm. The anomaly detection algorithm is based on a deep learning algorithm and comprises a multi-layer neural network. It is a central part of the techniques presented herein. Learning complexities from high frequency multivariate time series data are challenging for native machine learning (ML) algorithms, necessitating a complex and state of the art algorithm. For purposes of exposition the algorithm explanation that is presented below has been divided into three areas, including input data transformation (supporting the signature calculation), neural network layers, and the residual matrix calculation and usage.

A first area of the anomaly detection algorithm encompasses input data transformation. The input data may be transformed into a sequence of concatenated, separate signature matrices. Consider a case where a total of 'n' KPIs, as part of a system being monitored, are provided as input. A signature matrix of shape n x n is formed based on the pairwise inner product of two time series for the multivariate time series segment (t-w) to t. For example, given two time series segments their correlation may calculated using the formula that is presented in Figure 4, below.

$$m_{ij}^{t} = \frac{\sum_{\delta=0}^{\omega} x_i^{t-\delta} x_j^{t-\delta}}{\omega}$$

Figure 4: Signature Creation Formula

In the formula that is presented in Figure 2, above, w is the rescale factor. The value of w may be taken as 10, 50 and 100 though this value may be fine-tuned for better performance during initial model training depending on the system KPIs being monitored.

Thus, at every time step a set of three signature matrices are calculated with w = 10, 50 and 100. Consider an anomaly detection system that is started from the 100th second, where signature matrices of time segments 90-100 seconds for w = 10, 50-100 seconds for w = 50, and 0-100 seconds for w = 100 are used. Since there is a set of three signature matrices for one time step (i.e., the 100th second), 100 – 110th time steps may be used to calculate an anomaly score for the 111th time step. The number of time steps k to use for the calculation of anomalous scores for a k+1th time step may be decided upon during training based on what best fits the given data.

A second area of the anomaly detection algorithm encompasses the actual neural network architecture . The basic structure of the neural network is based on an autoencoder architecture comprising an encoder, a decoder, and hidden layers, where the output of the neural network is the reconstructed input data (i.e., input signature matrices in the instant case) and anomalous behavior is detected from the residual matrices calculated as the difference between the input and the output (i.e., the reconstructed input). In recent research this architecture has proven well in anomaly detection tasks in highly complex data. Thus, based on this architecture a tailored neural network may be built.

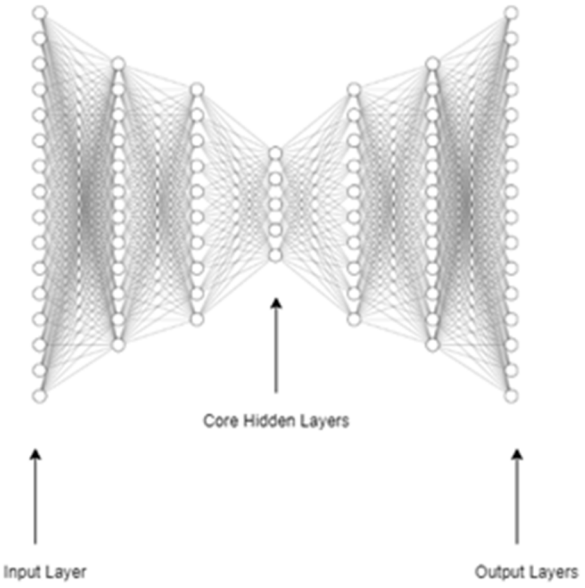

Figure 5: Autoencoder Architecture for Anomaly Detection

The autoencoder architecture is generally divided into three sections, which are depicted in Figure 5, above, and which are described and discussed below.

A first section of the autoencoder architecture encompasses an encoder. The encoder layer compresses the input for the hidden layer to learn useful information from same. For the encoder portion of the network a fully convolutional encoder, comprising of any number of convolution operations. The three layers may have 30x30x32, 20x20x64, and 10x10x128 outputs, respectively.

A second section of the autoencoder architecture encompasses core hidden layers. The core hidden layers in an autoencoder type of a network have a proven ability to learn in a way that extracts useful information from the compressed encoder layer. Long short-term memory (LSTM) layers with attention may be employed for the hidden layers. This helps in learning temporal information from time steps and multivariate time sequences and the attention layers help extract information from the time steps that provide important information. The set of three signature matrices from each time step is fed from each encoding layer to an LSTM layer with attention. This layer also compresses multiple input time steps into one, returning just a single set of three matrices, one for each value of w that may be passed to the decoder.

A third section of the autoencoder architecture encompasses a decoder. Each convolution operation in the decoder takes as input the concatenated output channels from both LSTM attention layers and the previous deconvolution layer and produces an output with a specified number of output channels. Thus, the final output of the model is a reconstructed signature matrix of the exact size and number of channels (i.e., the number of signature matrices at each time step – three in the instant case) as the input. The algorithm decoder part may contain any number of convolution functions, for example layers with outputs of 15 x 15 x 64, 30 x 30 x 64, and 3 x 30 x 3 (as final output), respectively. The final output layer should always have the number of outputs in the 3rd dimension as the number of different values of w's used.

Figure 6, below, illustrates aspects of the neural network narrative that was presented above.

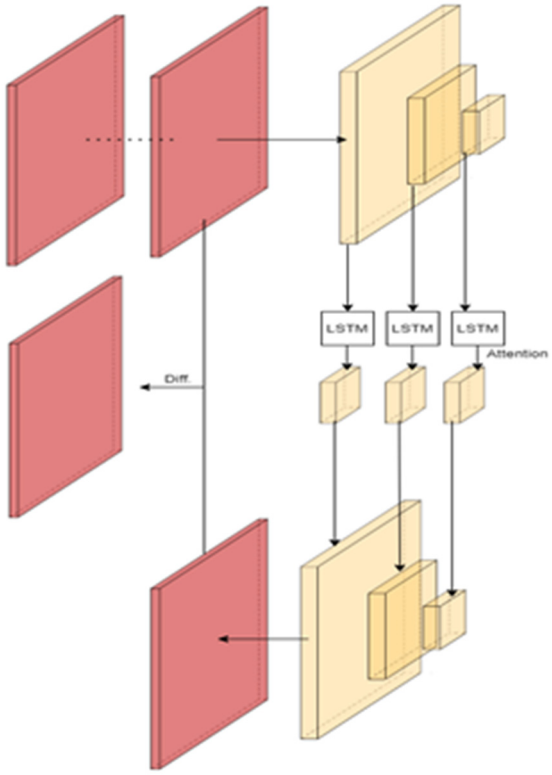

Figure 6: Exemplary Neural Network

A third area of the anomaly detection algorithm encompasses residual matrices.

The residual matrices are calculated by finding the difference between the set of original matrices and the final matrices. More precisely, the difference between the last time step of the input (of shape n x n x 3) and the (only) single output (also of shape n x n x 3) is calculated in support of forming the residual matrix (again of shape n x n x 3). Note that 'n' denotes the number of input variables.

The residual matrix is then used for all further actions, including, for example, anomaly root cause identification, anomaly scoring, etc. Various of these actions are described and discussed below.

Aspects of the techniques presented herein encompass anomaly scoring. Anomalies are scored based on the number of poorly reconstructed pairwise correlations in the residual matrices. This calculation is based on a threshold theta which may be is set by a domain expert and remains fixed once decided upon while training. In other words, the number of elements in the residual set of signature matrices that crosses the given threshold are flagged as an anomaly. For this purpose, any one of the signature matrices (representing a single w) may be taken or an average of all three of the matrices may be taken.

Aspects of the techniques presented herein encompass root cause identification. The residual matrices can be inferred to help detect root cause for identification. Since the three output residual matrices for different values of w are each of shape n x n, where n denotes the input features, the exact feature causing the anomaly may be tracked down. An illustrative example is presented in Figure 7, below. This can be used in support of, for example, semi-automated troubleshooting as discussed earlier.

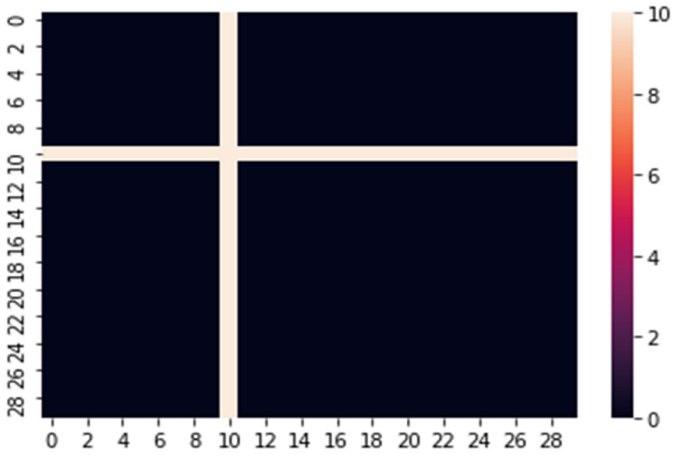

Figure 7: Example Output Signature Matrix Showing KPI 10 as Root Cause on Dummy Data

Aspects of the techniques presented herein also encompass anomaly severity identification. Anomaly scores are used for anomaly severity identification. This anomaly severity taken into consideration together with the anomaly root cause when combined with hand crafted rules by domain experts can trigger network services using network service orchestrator tools for network configuration management thus facilitating semi-automated network failure prevention.

Further we provide some analysis on dummy data on how our model fairs on dummy data and which also stands against the reason on why we should leverage multiple input matrices with different scales when only one would have been sufficient. Different output signature matrices for different scales or values of w may be used for better predicting anomalies for short, medium and long duration anomalies. See Figure 8, below, for an illustrative example. The flexibility that the system provides with multiple anomaly output matrices at different scales, taking into account different numbers of time steps, helps to improve anomaly detection and diagnosis.

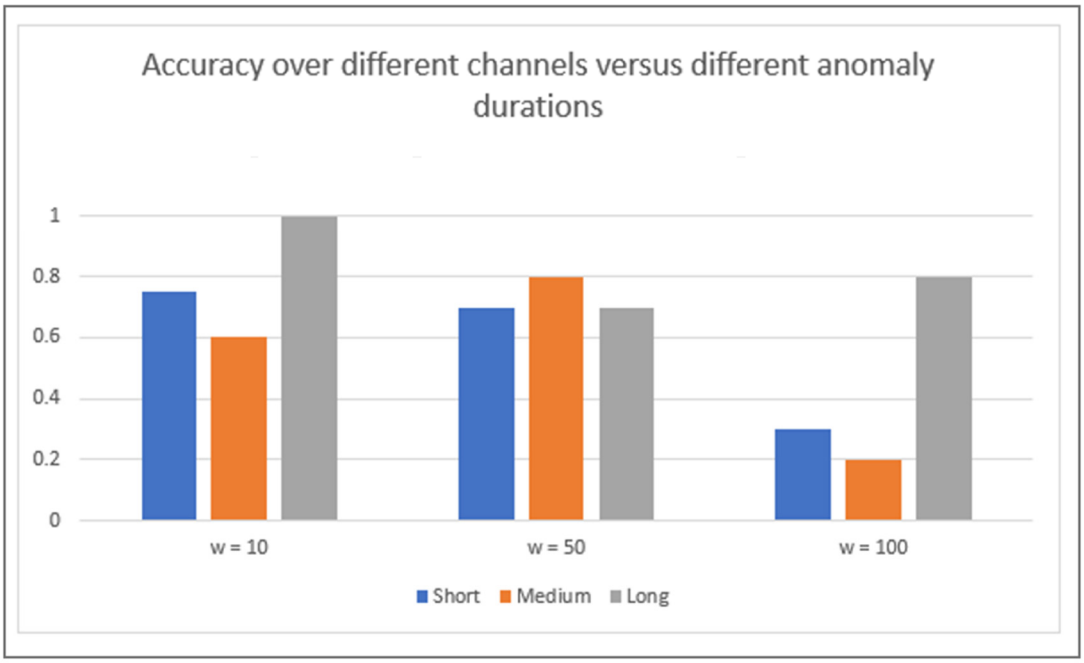

Figure 8: Illustrative Short, Medium, and Long Duration Anomalies on Dummy Data

For the techniques that are presented herein, of particular interest and note are, for example:

- Introduction of complete node/system/deployment level anomaly detection in 5G network deployments, instead of single KPI anomaly detection, using multiple KPIs from a particular node, a combination of KPIs from various nodes or all KPIs being monitored in the network deployment.
- The anomaly detection approach and algorithm supporting our innovation that was described above – which is the current state of the art and the future of anomaly detection in systems generating high frequency time series data with little or no positive anomaly cases for supervised models to train on –. Aspects of the techniques presented herein focus on infrastructure level data that is collected from a 5G nodes
- An anomaly diagnosis mechanism supporting our mechanism with the algorithm output and a semi-automated mechanism for timely fault detection and troubleshooting in networks to prevent network breakdowns. Adding to this,

our algorithm supporting the approach also supports anomaly analytics for better future capacity management and better future networks.

In summary, our work introduces node/system/deployment level anomaly detection in 5G networks and also a supporting algorithm for the same together with a methodology for semi-automated anomaly diagnosis.