Predicting the actual location of faults in underground optical networks using linear regression

Owusu Nyarko-Boateng | Adebayo Felix Adekoya | Benjamin Asubam Weyori

Abstract
Optical cables are enormous transmission media that carry high-speed data across transatlantic, intercontinental, international boundaries, and cities. The optical cable is essential in data communication. The cable has become an indispensable component in optical communications infrastructure; hence, conscious efforts are always adopted to prevent or minimize faults in the optical network infrastructure. Typically, tracing fault in the underground optical network has been difficult even though the optical time-domain reflectometer (OTDR) has been used to measure the distance of faults in the underground fiber cable. The methodologies deployed in the reviewed literature indicate a vast gap between the fault distance measured by the OTDR and the actual distance of fault. This paper observed the difficulties involved in tracing the actual spot of fault in the underground optical networks. The difficulty of tracing these underground faults mostly result in an undue delay and loss of revenue. This research presents a machine learning approach to predict the actual location of a fiber cable fault in an underground optical transmission link. Linear regression in the python sci-kit learn library was used to predict the actual location of a fault in an underground optical network. The mean square error and MAE evaluation matrix used provided good accuracy results of 0.061291 and 0.080143, respectively. The result obtained in this paper indicates that faults in underground optical networks can be found quickly to avoid the delays in the fault tracing process, which leads to an excessive revenue loss.

KEYWORDS
fiber optics cable, linear regression, machine learning, OTDR and fault tracing, underground optical networks

1 | INTRODUCTION

Machine learning (ML) applications in optical communications and networking are gaining more attention, particularly in the areas of nonlinear transmission systems, optical transmission amplification, passive optical performance monitoring, and cross-layer network optimizations for software-defined networks, fault detection, identification of bit error rate...
(BER), quality of transmission (QoT) and signal amplification. Several ML techniques have been deployed to solve problems relating to optical communications infrastructure. The process of using the ML technique to predict optical network performance has reached a stage where complex network management, fault tracing, configurations, and coding schemes have high accuracy. Not much has been done with ML in tracing fault in underground optical networks. Hard failures in underground optical networks are complicated to trace as the optical time-domain reflectometer (OTDR) measurements provide the distance of the fiber cable buried in the earth. The major problem the telecommunications companies experience in the underground fault tracing is how to find the exact spot of the cable cut on earth. Long-haul underground optical networks have fiber optics cables (FOCs) laid in trenches of 1.2 m deep inside the earth. This mode of transmission has encountered several challenges, which include most predominantly fiber cable cuts, which result in disruption of communication services.

In underground optical transmission, it is usually strenuous to trace hard failures in the fiber cable buried in the earth when the cable is not visible on the earth's surface. Tracing the exact spot of fault in the fiber cable requires some form of digging along the optical cable transmission path to identify the exact cable cut point. This situation causes enormous revenue loss due to delays in tracing and repairing the fault.

The OTDR measurements alone have not been useful in tracing fault in the underground optical infrastructure. This research provides a solution to the problem by using ML modeling to predict the actual fault location when there is a fiber cable cut in underground optical infrastructure. The identified drawbacks in the present approach of tracing faults in underground optical networks include the inconsistencies between the measured value of OTDR and the actual distance of fault, excessive delay in pinpointing the exact spot of fault on earth’s surface, waste of resources, and the economic losses. Many solutions, especially the use of OTDR for determining the precise length of fiber cable cuts in underground optical transmission systems, have been developed and the application of ML to trace faults in soft and hard failures of optical transmission infrastructures. Thus, telecommunication companies have, over the years, relied on OTDR due to its preciseness in estimating the distance of fault in fiber cables. The solution made it possible to measure the length of fault in the fiber cable using the OTDR device. The imprecise and inaccurate nature of OTDR measurement of tracing faults in the underground optical transmission system has made it impossible to identify the exact locations of the fiber cuts on earth, resulting in extra costs and workforce hours for the telecommunication companies. These extra costs include the cost implication of delay in resolving the faults, negative impact on the company’s brand, and the additional cost of digging and covering unaffected areas, among others. OTDR provides an estimated distance of fiber cable at which faults occur, but the actual distance of fault on earth and the estimated fiber cable distance varies considerably. Various research works have been carried out in this domain to resolve these problems, but all the reviewed works provided an estimated fault distance of the underground fiber cable. The previous research works include fault tracing using OTDR, photon probe fault locator, Raman-based fiber sensors, Tunable optical time-domain reflectometer (T-OTDR), Correlation technique utilizing traffic signal, step frequency method, published their work on Fault Detection Technique in an underground fiber network, emphasizing mainly on the limitations and drawbacks on the practical approach of measurement of OTDR, which focused not only on the distance of the FOC alone but the Euclidean distance on the earth surface from the optical transmitter to the point of the fiber cable cut. The solution provided by Reference 16 also introduced several cut points along the fiber cable transmission path, which eventually increased the losses in the FOC network.

Other researchers proposed embedded OTDR software on the small factor pluggable (SFP) module to be deployed at the end-to-end nodes in the FOC network infrastructure to monitor and report the distance of fault in the transmission link. The embedded OTDR in the SFP module was to reduce fault down-time and the delays in fault tracing in the underground optical networks.

In their study, investigated the application of ML-based techniques for soft-failure detection, identification, and causes of failures based on continuous monitoring of BER. The authors further explored the trade-off between the accuracy and complexity provided by different ML algorithms using several model parameters, such as BER sampling time and the amount of BER data needed to train the models. Predicting imminent transmission failures that could disrupt the optical network operations by continuous monitoring of the active optical links is essential. Monitoring optical network infrastructure to identify and localize failures during active operation intelligently has been achieved by the use of ML-based techniques.

An optical network generates a large number of different data streams that must be fetched, processed, and analyzed promptly by ML techniques to ensure optical network QoT. In order to enable data-driven ML analysis, it is important to explore several aspects of network data, including its extent, monitoring, query mechanisms, storage, and representation attributes. Generally, as long the underground fiber cable cut remains unsolved, the telecommunications company loses...
huge revenue, and users also suffer unnecessarily as a result of the devastating impact of the fault. This paper uses the ML technique to predict the actual location of faults in underground optical networks by applying a simple linear regression predictive model. The input dataset of the single-layer perceptron (SLP) neural network (NN) structure applied a sigmoid activation function to obtain useful output value with reasonable accuracy.

2  |  FAULT DETECTION IN UNDERGROUND OPTICAL NETWORKS

Failures in optical networks are mostly characterized by losses that hugely affect the QoT and quality of service. The primary classification of these failures is hard failures, which is a sudden event such as fiber cuts, power outages, etc.1 Soft failures are gradual transmission degradation due to equipment malfunctioning or filter misalignment. Failures in optical networks are caused by different sources such as filters misalignment, amplifier malfunctioning, fiber bends, etc. During network operations, several kinds of soft failures affect the signal quality and induce anomalies in the BER at the receiver, ultimately leading to packet losses or even service disruption. Hence, the techniques for soft failure detection, localization, and identification are crucial, as it is used to perform traffic re-routing and rapid failure recovery. Hard failure in underground optical networks such as fiber cable cut and bends has been traced and identified using OTDR device.22

2.1  |  OTDR

The OTDR device is used to trace faults in optical cables.14,23 The scientific principle employed by the OTDR device to measure the distance of faults in optical cables is Rayleigh scattering and Fresnel reflection techniques. The device is also used to verify splice loss, measures the length, and finds faults in optical cables. OTDRs are also commonly used to establish the integrity of fiber cables when it is newly installed. The OTDR’s operating principle is based on the measurement of the Raleigh backscattering signal, which is generated by sending a high power optical pulse from the OTDR through the optical fiber. When the light reaches the faulty spot in the fiber cable, it gets reflected back to the OTDR. This backscattered light measured by the sensitive optic receiver is converted to digital form and averaged to improve the signal to noise ratio. The resultant signal is displayed as a graph called a Trace. The trace is a visual representation of the backscattering coefficient created by the OTDR to determine the activities of the backscattered light. The trace shows the activities such as cuts, splice loss, bends, attenuation, and distance of fault in the optical networks.24

The OTDR measurements deliver much valuable information such as optical power loss (dB), attenuation (dB/km), bending, stressing, and breaks. Fresnel reflection is a discrete reflection which uses the activity of the backscattered light to determine the distance of the light signal, which travels back from the fault point to the optical transmitter. These fault locations are caused by a change in reverse coefficient elements such as air gap or severe particles obstructing the free flow of the light signal. At these points, there is always a strong light signal reflected back.25 However, by using the information of Fresnel reflection, OTDR can predict the soft and hard failures in the optical network infrastructure. The device measures the distance of the light source (optical transmitter) to the faulty zone.

Besides Rayleigh scattering, other famous scientific principles used in tracing faults in optical networks include Raman scattering, Mie scattering, and Brillouin scattering.26 All these principles, when deployed in OTDR, measures the distance of the underground fiber cable.

2.2  |  Application of machine learning in optical-network failures

The adoption of the ML approach in the field of optical networks has been inspired by the exceptional growth and the complexity faced by the continuous expansion of optical network infrastructure. The increase in such complexity is as a result of the introduction of a massive number of adaptable and interdependent system parameters such as routing, configurations, modulation, symbol rate, coding schemes, etc. These parameters are managed by the usage of coherent transmission and reception technologies, advanced digital signal processing, and compensation of nonlinear effects in optical fiber propagation. Optical signals are usually affected by fiber nonlinearity.

ML applications in optical networks provide several advantages such as traffic prediction and virtual topology design, failure detection, fault localization, and flow classification. Binary support vector-machine (SVM), random forest, multiclass SVM, NNs are some ML techniques used in optical network fault detection. At the same time, NNs, logistics, and
linear regressions are predominantly deployed in optical network fault identification. ML technique has been deployed in this paper to help predict the actual distance of fault when the value of OTDR measurement is known. Using the ML technique to trace faults in the underground optical cables is expected to be done without much delay, resulting in reduced loss of revenue.

2.2.1 Simple linear regression model

The simple linear regression (SLR) model is the most intuitive and ubiquitous ML algorithm, which uses one or more independent variables to predict a dependent variable. SLR models a target value based on independent variables. The model performs tasks to predict a dependent variable value \( y_i \) based on a given independent variable \( x_i \). So, the regression technique finds out a linear relationship between the OTDR measured value as \( x_i \) (input) and the actual location of a fiber cable cut, \( y_i \) (output). Hence, SLR is given by:

\[
y_i = \alpha_0 + \alpha_1 x_i,
\]

where \( \alpha_0 \) is the intercept and \( \alpha_1 \) is the coefficient of \( x_i \).

In the simple linear regression model applied in this paper, the training data were defined as \( x_i \), which was the input value obtained from the OTDR measurement, and \( y_i \) is the actual location (predicted value). To achieve the best fit of the predicted value, \( y_i \), for the regression model in (1); given \( x_i \) as the dependent variable, \( \alpha_0 \) and \( \alpha_1 \) values must reach its best in the specified epoch. Ones the best of \( \alpha_0 \) and \( \alpha_1 \) values were found, the best fit for the simple linear regression was achieved.

2.2.2 Single-layer perceptron neural network

SLP NN is an ML technique which has an input layer and an output layer of processing units with no feedback connections. The SLP NN structure for the linear regression prediction consists of two input units. The input units to the SLP are the actual distance of the underground fiber cable between the optical transmitter and the point of the fault \( (x_1) \) with weighted value \( (w_1) \) and the measure of OTDR \( (x_2) \) with its weighted value \( (w_2) \). The single hidden layer computes the input value by applying the sigmoid activation function to the expected predictive distance of the underground fault to achieve \( y_{pred} \) as indicated in Figure 1.

\[
y_{pred} = \sum_{i=1}^{n} (x_1 w_1 + x_2 w_2),
\]

\( w_1 x_1 \) represents the actual distance between the optical transmitter and the point of fiber cable cut.
\( w_2 x_2 \) represents the measurement of OTDR and its weighted value.

3 PROPOSED MODEL

The identified drawbacks in computing the actual location on earth of fault tracing in optical networks have severe effects on the efficiency of the fault location process. It is as a result of these inefficiencies that this paper adopted an ML predictive model that can compute the actual location of the fault on earth based on the measured distance of OTDR.
Figure 2 represents an underground optical transmission link between node A and node B with three chambers, which contains a coiled FOC of length \( q \). The distance between any two chambers in the link has an equidistance of length \( p \). The depth of the underground FOC from the surface of the earth is 1.2 m (this value has no significance in the modeling of the ML predictive technique in this paper). The link also has some spliced closures fixed along the transmission path due to the splicing of other forms of hard failures. In case the link experience a cut in the earth at point \( x \), and the underground cable is not visible to the surface of the earth; location the spot of the FOC cut underground is much complicated. The paper develops an intelligent system to solve this problem. The parameters extracted from the link AB were essential is the prediction of the actual distance.

Parameter of Figure 2:

a. the stock of coiled FOC in each duct well (chamber) is \( q = 15 \) m.
b. The stock of fiber cable between the two chambers is \( p = 1000 \) m.
c. The OTDR measurement between the optical transmitter, \( A \), and the fiber cable cut point is \( x \)
d. Earth distance of the transmission link is the Euclidean distance between the optical transmitter \( A \) and the optical receiver \( B \).
e. The stock of fiber cable in each splice closure box is 1 m
f. The actual fault distance on earth is \( E_d = y_{pred} \)

3.1 | The stock of coiled fiber cables

The difficulty in tracing the precise location of faults in the underground FOC has been attributed to the number of coiled fiber cables placed in the various chambers (\( C \)) along the optical transmission path AB and the stock of cable in the spliced enclosure box. The coiled fiber cables accounted for the inconsistencies in the OTDR measurement and the actual fault location. The proposed predictive model considered the number of chambers in the optical transmission link, the coiled fiber cable, the number of chambers between the optical transmitter, the stock of cable in the spliced enclosure box and the spot of fault in the underground optical network, as shown in Figure 2.

3.2 | Number of chambers in the optical transmission link

In the single-mode underground transmission system, chambers are placed at equal intervals, \( p \), along the optical transmission path AB. The chambers contain coiled optical cable of length, \( q \). This indicates that each chamber contains an equal length of coiled optical cable at the time of installation. This phenomenon is done to allow slag in the underground optical cable.
The number of chambers in the transmission link between A and x has been represented mathematically as \( C_x = C_i \). Where \( C_x \) is the summation of the chambers between A and x; \( i = 0, 1, 2, \ldots, n \).

\[
C_x = \sum_{i=0}^{n-1} C_i
\]  

(3)

In Figure 2, the number of chambers between A and x plays a critical role in computing the exact location of the fault on earth. Having \( C_i, q \) and \( p \) between A and x optimize the process of predictability of the precise spot of the fault on earth.

### 3.3 Distance on earth

Computing the exact fault location on the earth’s distance (\( E_d \)) is critical in the fault tracing process of underground optical networks. As indicated in Figure 2, \( E_d \) in the underground optical cable transmission is the point between the optical transmitter to the point of the cable cut. \( E_d \) is the predicted distance, \( y_{pred} \). The essential parameters of determining the fault distance are the length of the coiled cable in the chambers and the measurement of the OTDR device. In the instances where there are no chambers between the cable cut point and the optical transmitter, the actual location is given by,

\[
E_d \sim x
\]

(4)

where \( x \) is the distance between the optical transmitter and the point of the underground optical cable cut.

### 4 ML MODEL FOR FAULT TRACING OF UNDERGROUND OPTICAL NETWORKS

In order to resolve the problem of delays and loss of revenue in the process of tracing fault in the underground optical network infrastructure, there is the need to adopt the technique of an intelligent system to aid the process. The intelligent system uses three main algorithms to learn the behavior of the input data. These main algorithms include supervised, unsupervised, and re-enforcement learning techniques. \(^4\) \(^3\) \(^2\) \(^3\) \(^5\) SLR supervised predictive model was adopted by the paper to predict the actual location of the fault based on the SLP NNs. \(^3\) \(^6\) \(^3\) \(^7\) Several ML algorithms were considered appropriate for this study. However, the SLR algorithm was adopted due to its ease of use, efficiency, faster training and testing time, which results in low computational cost.

The datasets collected from the field were fed into the predictive model; datasets were pre-processed and transformed by normalization. Data patterns were extracted based on the SLR analytics by dividing the data into training and testing datasets. The partitioned dataset was taken through several epochs to produce the expected output. Python Scikit-learn library was used for SLR modeling. The proposed ML framework to predict the actual location of the fault in the underground optical network is summarized in Figure 3.

### 4.1 Data collection

The input data used for the experiment were collected from the field. The data collection was done on a real-time basis as faults occurred at various locations of the optical network infrastructure, which belong to one of the telecommunication companies in Ghana. The dataset was collected over 8 weeks in three regions of Ghana. Some of the optical links had multiple mobile network operators, which causes the impact of faults to be severe. The datasets we collected over the period were 1250. The attributes of the datasets were five (5); thus, the region, OTDR measurements, actual distance, number of splice closures, and the stock of cable in each chamber. The regional distribution of the datasets in the eight regions has been presented in Table 1. The actual fault tracing process in the underground optical network has been very complicated to determine due to the nonlinearity of the underground optical cable. The use of the ML technique in this paper aims at efficiency in fault tracing processes in the underground optical network, which addresses our main objective to reduce the delays associated with the fault tracing procedure.
In this work, the value of $y$ is the computed distance of the fault on the earth. The $x$ and $y$ values ranged from 1 m to 50,000 m. The SLP NN applied in the linear regression predictive model had an input variable $x_i = [x_1, x_2]$ and the weight $w_i = [w_1, w_2]$ as indicated in Equation (2). SLR performs well when the dataset has a linearly separable nature of the relationship among the variables. We adopted SLR because we found it more appropriate for our model due to its ability to implement, interpret, and efficiently train the dataset we used.

### 4.2 Data transformation

The input dataset was taken through selection, data cleansing, normalization function was also used in each of the input variables, which transformed them into a value between 0 and 1. The data pre-processing was performed to remove outliers.

### 4.3 Pattern extraction

In ML modeling, data partitioning is essential. The overall dataset was partitioned into training and test dataset; the more training dataset a model has, the better the model becomes; hence 80% (1000) and 20% (250) of the original dataset, respectively, was implemented in the SLR model. By using the SLP learning technique, SLR predictive model was applied to train the partitioned data. The practical approach to the ML modeling, applied a regression analysis in which the data was integrated into a sigmoid activation function, as shown in Equation (5).

$$Sigmoid.f(x) = \frac{1}{1 + e^{-x}}.$$  

By achieving the best-fit regression values, the model aimed to predict $y$ value such that the error difference between the predicted value and actual value is minimum. It is crucial to update $\alpha_0$ and $\alpha_1$ values, to reach the
best value that can minimize the error between predicted \( y \) value (\( \text{pred} \)) and actual \( y \) value (\( y_i \)) computed by Equation 6 to 14.

\[
\text{minimize} \quad \frac{1}{n} \sum_{i=1}^{n} (\text{pred}_i - y_i)^2, \quad (6)
\]

\[
J = \frac{1}{n} \sum_{i=1}^{n} (\text{pred}_i - y_i)^2, \quad (7)
\]

\[
J = \frac{1}{n} \sum_{i=1}^{n} (a_0 + a_1 \cdot x_i - y_i)^2, \quad (8)
\]

\[
a_0 = a_0 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (\text{pred}_i - y_i), \quad (9)
\]

\[
a_1 = a_1 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (\text{pred}_i - y_i) \cdot x_i, \quad (10)
\]

A partial derivative was also applied in the regression analysis to improve the accuracy of the result.

\[
\frac{\partial J}{\partial a_0} = \frac{2}{n} \sum_{i=1}^{n} (a_0 + a_1 \cdot x_i - y_i), \quad (11)
\]

\[
\frac{\partial J}{\partial a_0} = \frac{2}{n} \sum_{i=1}^{n} (\text{pred}_i - y_i), \quad (12)
\]

\[
\frac{\partial J}{\partial a_1} = \frac{2}{n} \sum_{i=1}^{n} (a_0 + a_1 \cdot x_i - y_i) \cdot x_i, \quad (13)
\]

\[
\frac{\partial J}{\partial a_1} = \frac{2}{n} \sum_{i=1}^{n} (\text{pred}_i - y_i) \cdot x_i, \quad (14)
\]

where \( J \) is the cost function.

The partial derivatives are the gradient descent, and it was used to update the values of \( a_0 \) and \( a_1 \). \( \alpha \) is the specified learning rate. A lower learning rate was chosen to get the cost function closer to the minima, but it took more time to reach the minima. A tremendous learning rate was not applied because it converged quickly, and it tends to overshoot the minima.

The dataset was pre-processed, thus, data-cleaning and data-transformation. At the data pre-procession stage, we replaced missing values within a column (set of attributes) with the mean value of the said column and removed outliers where needed. In order for our ML model to perform efficiently to reduce computational time, the datasets were scaled to the range of \([0, 1]\) using the max-min function as defined in (15).

\[
m'_x = \frac{m_x - m_{x}\text{min}}{m_{x}\text{max} - m_{x}\text{min}}, \quad (15)
\]

where \( m'_x \) is the normalized value, \( m_x \) is the value normalized, \( m_{x}\text{min} \) and \( m_{x}\text{max} \) are the minimal and maximal values of the datasets.

5 | RESULTS AND DISCUSSION

The regression Equations (1) was used to predict the actual location of the fault in the underground optical network. The predicted value and actual value are shown in Table 2 and Figure 4. As indicated in the figure, the simple regression model better fits than the actual value. The values in Table 2 are normalized data measured in meters. The goal of adopting normalization was to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. In ML, it is common practice to scale the dataset to a range that ensures better modeling. Normalization is the process we adopted to transform the dataset into the range of \([0, 1]\).
TABLE 2  The actual and predicted values of fault distance (m)

| Actual values ($y_i$) | Predicted values ($y_{pred}$) |
|-----------------------|-------------------------------|
| 0.039656982           | 0.030709101                   |
| 0.089107197           | 0.077275265                   |
| 0.037703717           | 0.028475712                   |
| 0.086880799           | 0.070130088                   |
| −0.024037119          | 0.015633724                   |
| 0.069267752           | 0.024539363                   |
| 0.066537393           | 0.042177778                   |
| 3.69546E-05           | 2.15985E-05                   |
| 0.030504379           | 0.012639084                   |
| 0.081982683           | 0.061063093                   |
| 0.031220966           | 0.024215522                   |
| 0.07001415            | 0.064332775                   |
| 0.084927534           | 0.075608599                   |
| 0.039656982           | 0.038586711                   |
| −0.001637077          | 0.001002021                   |
| −0.007129979          | 0.001830821                   |
| 0.0779012             | 0.052456728                   |
| 0.066537393           | 0.082137521                   |
| 0.0298648             | 0.010608599                   |
| 0.0298648             | 0.010887772                   |
| 0.000316188           | 0.000238414                   |
| 0.122042494           | 0.221173311                   |

FIGURE 4  The actual and predicted values of underground optical networks

At the same time, the actual values randomly fluctuate irregularly and are independent of each other with no abnormalities, as indicated in the horizontal and vertical axis of Figure 4. Hence, it is feasible and effective to use the simple linear regression model to predict the actual distance of an underground optical network. There were few observed outliers in the actual distance prediction. In Figure 4, the $y$-axis represents the distance of the cable cut, and the $x$-axis represents the number of faults. These outliers had an insignificant effect on the error margin of the entire process. In SLR data analysis,38,39 argued that it is a common practice to identify outliers in data analyses. That notwithstanding, outliers are of interest because they could significantly affect the results of research. However, an outlier, rather than being due to a measurement error, could be due to some interesting changes or behavior in the data generating process, and it is often
of interest to investigate such changes. The publication of Reference 40, suggested a method called massive unsupervised outlier detection (MUOD) for detecting outlying curves in functional data sets. MUOD is an unsupervised algorithm that automatically detects the magnitude, shape, and amplitude of outliers.

The outliers identified in the predicted values were as a result of the input dataset of the SLR model. Some of the underground transmission links have had several cuts and splicing joints. The underground FOC repairing process requires extra cable to be inserted to allow efficient splicing process or some cables drawn from the coiled cable in the nearby chambers to enable flexibility in cable splicing. These error measurement in the data collection process brought about some outliers and may be corrected when the model is fully implemented on the field. This is the reason why we adopted ML, which is essential because, as the predictive model is exposed to new data, the model would be able to adapt independently. The model learns from the previous computations to produce reliable, repeatable decisions and results of the exact distance of fault on earth.

The best cost function was obtained by improving the values of $\alpha_0$ and $\alpha_1$ such that the mean square error (MSE) value settled at minima, as indicated in the error values in Table 3.

| MAE  | MSE    | $R^2$ |
|------|--------|-------|
| 0.080143 | 0.061291 | 0.000357 |

**MAE**: is the absolute error, which is translated as the absolute percentage error; that is, MAE$x100$.

**MSE**: Mean square error is the cost function.

**$R^2$**: The ratio between predicted variation and actual variations,

$$R^2 = \frac{SS_e}{SS_t},$$

where,

$$SS_e = \sum_i (y_i - \hat{y}_i)^2,$$

and,

$$SS_t = \sum_i (y_i - \bar{y})^2,$$

The accuracy obtained in the test indicates how close the predicted values are to the real values. The result obtained in the simple linear regression prediction has the value that can be used for fault tracing in the underground optical networks. Even though the predictive model had outliers at the output, the evaluation matrix deployed suppressed the residual effect of the outliers on the final results. The values obtained for the evaluation matrix MSE and MAE were 0.061291 and 0.080143, respectively, indicating the predictive model is good for the prediction of the exact distance of faults in underground fiber cable cut. Identifying and repairing the fault in the underground optical networks significantly reduce the cost of repairing the faults and then overcome the challenges identified by Reference 42.

That notwithstanding, the ML model deployed produced smaller error margins that have the ability to precisely predict the exact distance of fault in underground optical networks. When the variables of the fault are available, our empirical ML model has the tendency to provide the result of the exact location quickly than to deploy any conventional mathematical computations. The modeling would be achieved quickly when the input variables are updated according to the observed variables of the new faults. The use of ML modeling in this research was significant because the model overcomes the delays in tracing faults in underground fault using only the measurements of the OTDR device. As the delays in the faults tracing process have been overcome by the model used, there will be a significant reduction in the loss of revenue, and the detrimental effect on customer experience will be heavily minimized. The attainment of this research result is a great achievement that supports the novelty of our aim to deploy the ML model to determine the exact distance of faults in underground optical networks.

## 6 | CONCLUSION

An intelligent model based on a SLP NN technique was applied in a simple linear regression predictive model to predict the precise distance of fault in an underground fiber cable cut. The evaluation of the test data shows that the simple linear
regression model is capable of predicting the location of an underground fiber cable cut correctly. In this case, OTDR measurement must be made available to predict the actual distance of fault on earth whenever a failure occurs in the underground optical networks. The solution provided in this paper will help the telecommunication industry players to reduce significantly the time used in tracing fault; this means the loss of revenue, which arises as a result of delays in the fault tracing and repairing process, will also reduce significantly.

At the same time, the simple linear regression prediction method achieved in this paper does not only apply to the prediction of faults in underground optical networks, but also other underground transmission infrastructures. The primary evaluation matrixes employed in the predictive model were MSE and MAE, and it provided good accuracy results of 0.061291 and 0.080143, respectively. The result obtained in the paper has provided an excellent research opening to apply other ML techniques to predict faults in the underground optical networks. Using the ML model to determine the exact distance of fault in underground optical networks has been a great achievement which has a high tendency to overcome the difficulty in tracing fault in underground optical networks. The correlation between some of the predicted values and the actual value was higher than then expected values, as the outliers were few and had no significant influence on the overall predicted values.

However, the attainment of this research result is a great achievement that supports the novelty of our aim to deploy the ML model to determine the exact distance of faults in underground optical networks in an effort to overcome the delays in the fault tracing process. The relevance of this result is essential for the MNOs whose long-haul network infrastructure is underground fiber cables.

ACKNOWLEDGEMENT
We are most grateful to the telecommunication company that allowed us to collect data from its infrastructure for this work. The company has decided to remain anonymous in this work.

PEER REVIEW INFORMATION
Engineering Reports thanks the anonymous reviewers for their contribution to the peer review of this work.

PEER REVIEW
The peer review history for this article is available at https://publons.com/publon/10.1002/eng2.12304.

CONFLICT OF INTEREST
The authors declare no potential conflict of interest.

AUTHOR CONTRIBUTIONS
Owusu Nyarko-Boateng: Conceptualization; data curation; formal analysis; investigation; methodology; and writing original draft. Adebayo Felix Adekoya: Supervision; writing-review and editing. Benjamin Asubam Weyori: Supervision; writing-review and editing.

DATA AVAILABILITY STATEMENT
The data that support the findings have been made available by the telecommunication company mentioned above. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors upon reasonable request and with the permission of the third party.

ORCID
Owusu Nyarko-Boateng https://orcid.org/0000-0003-0300-2469
Adebayo Felix Adekoya https://orcid.org/0000-0002-5029-2393
Benjamin Asubam Weyori https://orcid.org/0000-0001-5422-4251

REFERENCES
1. Barletta L, Giusti A, Rottondi C, Tornatore M (2017, March). QoT estimation for unestablished lighpaths using machine learning. In Optical Fiber Communication Conference (pp. Th1J-1). Optical Society of America.
2. Rottondi C, Barletta L, Giusti A, Tornatore M. A machine learning method for QoT estimation of Unestablished Lightpaths. IEEE/OSA J Opt Commun Networking. 2018;10(2):A286–A297.
3. Heng X, Gan J, Zhang Z, Qian Q, Yang Z. Amplification of orbital angular momentum modes in an erbium-doped solid-core photonic bandgap fiber. *Opt Commun.* 2019;433:132-136. https://doi.org/10.1016/j.optcom.2018.09.072.

4. Mata J, de Miguel I, Durán RJ, Merayo N, Singh SK, Jukan A, Chaminada M (2018); Artificial intelligence (AI) methods in optical networks: A comprehensive survey; 1573–4277. Published by Elsevier. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/). vol. 28, April 2018, pp. 43–57.

5. Salani M, Rottondi C, Tornatore M (2019). Routing and Spectrum assignment integrating machine-LearningBased QoT estimation in elastic optical networks. Paper presented at: Proceedings of INFOCOM 2019, Paris, April 2019.

6. Ghobadi M, Mahajan R (2016). Optical layer failures in a large backbone. Paper presented at: Proceedings of the 2016 Internet Measurement Conference. ACM, 2016.

7. Pointurier Y. Design of low-margin optical networks. *IEEE/OSA J Opt Commun Networking.* 2017;9(1):A9–A17. https://doi.org/10.1364/JOCN.9.0000A9.

8. Hayford-Acuah T, Asante B. Causes of fiber cut and the recommendation to solve the problem. *IOSR J Electron Commun Eng (IOSR-JECE).* 2012;17(12):46-64.

9. Satish A, Amiru IS, Yupapin P. Review of optical fibers-introduction and applications in optical fibers. *Results Phys.* 2018;10(2018):743-750. https://doi.org/10.1016/j.rinp.2018.07.028.

10. Ali TA, Ameen JH. Study of fault detection techniques for optical fibers. *ZANCO J Pure Appl Sci.* 2019;31(3):143-149.

11. Fernández MP, Rossini LAB, Pascual JP, Caso PAC. Enhanced fault characterization by using a conventional OTDR and DSP techniques. *Opt Express.* 2018;26(21):27127–27140.

12. Khan FN, Fan Q, Lu C, Lau APT. An optical communication’s perspective on machine learning and its applications. *J Lightwave Technol.* 2019;37(2):493-516.

13. Boutaba R, Salahuddin MA, Limam N, et al. A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *J Internet Serv Appl.* 2018;9(1):16. https://doi.org/10.1186/s13174-018-0087-2.

14. Shariati B, Ruiz M, Comellas J, Velasco L. Learning from the optical Spectrum: failure detection and identification. *J Lightwave Technol.* 2019;37:433-440.

15. Wang G, Mididoddi CK, Bai F, Gibson S, Su L, Liu J, Wang C (2018). UltraFast Optical Imaging using Multimode Fiber based Compressed Sensing and Photonic Time Stretch. arXiv preprint arXiv:1803.03061.

16. Kumar V, Rajouria D. Fault detection technique by using OTDR: limitations and drawbacks on practical approach of measurement. *Int J Emerging Technol Adv Eng.* 2016;2(6):www.ijetae.com.

17. Eriksson TA, Bülow H, Leven A. Applying neural networks in optical network communication systems: possible pitfalls. *IEEE Photonics Technol Lett.* 2017;29(23):2091-2094.

18. Schmuck H, Hehmann J, Straub M, Pfeiffer T (2006). Embedded OTDR techniques for cost-efficient fibre monitoring in optical access networks: Alcatel Research and Innovation, Alcatel SEL AG, Lorenzstr. 10, 70435 Stuttgart, harald.schmuck@alcatel.de

19. Shakarami S, Musumeci F, Cugini F, Tornatore M (2018, March). Machine-learning-based soft-failure detection and identification in optical networks. Paper presented at: 2018 Optical Fiber Communications Conference and Exposition (OFC) (pp. 1–3). IEEE.

20. Vela AP, Shariati B, Ruiz M, et al. Soft failure localization during commissioning testing and lightpath operation. *J Opt Commun Networking.* 2018;10(1):A27–A36.

21. Rafique D, Velasco L. Machine learning for network automation: overview, architecture, and applications [invited tutorial]. *J Opt Commun Networking.* 2018;10(10):D126–D143.

22. Nakamura A, Okamoto K, Koshikiya Y, Manabe T (2019, July). Potential Fault Detection in Optical Cables Using OTDR Operating in Two-Modes. Paper presented at: 2019 24th OptoElectronics and Communications Conference (OECC) and 2019 International Conference on Photonics in Switching and Computing (PSC) (pp. 1–3). IEEE.

23. Tamura Y, Sakuma H, Morita K, et al. The first 0.14-dB/km loss optical fiber and its impact on submarine transmission. *J Lightwave Technol.* 2018;36(1):44-49.

24. Thongdaeng R, Worasucheep D-r. Effect of bending radius and bending location on insertion loss in single mode fibers and polarization maintaining fibers. *Procedia Comput Sci.* 2016;86:15-18. https://doi.org/10.1016/j.procs.2016.05.004.

25. Kovacevc MS, Kuzmanovic L, Djordjevich A. Estimation of Rayleigh scattering loss in a double-clad photonic crystal fiber. *Opt Quantum Electron.* 2018;50(5):217.

26. Vamsi AVN, Rao AB. Fault detection in fiber optic communication cable by coherent anti-stokes Raman scattering using superconducting nanowire single-photon detector. *Indian J Sci Technol.* 2016;9(S(1)):1-6. https://doi.org/10.17485/ijst/2016/v9i8(1)/92819.

27. Musumeci F, Rottondi C, Nag A, et al. An overview on application of machine learning techniques in optical networks. *IEEE Commun Surv Tutorials.* 2018;21(2):1383-1408.

28. Pino-Mejías R, Pérez-Fargallo A, Rubio-Bellido C, Pulido-Arcas JA. Artificial neural networks and linear regression prediction models for social housing allocation: fuel poverty potential risk index. *Energy.* 2018;164:627-641. https://doi.org/10.1016/j.energy.2018.09.056.

29. Chen X, Guo J, Zhu Z, Priettti R, Castro A, Yoo SJB. (2018, March). Deep-RMSA: A deep-reinforcement-learning routing, modulation and spectrum assignment agent for elastic optical networks. In 2018 Optical Fiber Communications Conference and Exposition (OFC) (pp. 1–3). IEEE.

30. Junior UCP, Manito ARA, Rocha GVS, Monteiro FP, de Carvalho CCMM, Bezerra UH, de Tostes MEL (2018, September). Evaluation of Harmonic Contribution Impacts in the Electric Grid Through Linear Regression, Artificial Neural Networks and Regression tree. In 2018 IEEE PES Transmission & Distribution Conference and Exhibition-Latin America (T&D-LA) (pp. 1–5). IEEE.

31. Cohen E, Malka D, Shemer A, Shahmoon A, Zalevsky Z, London M. Neural networks within multi-core optic fibers. *Sci Rep.* 2016;6:29080.
32. Seve E, Pesic J, Delezoide C, Bigo S, Pointurier Y. Learning process for reducing uncertainties on network parameters and design margins. *J Opt Commun Networking*. 2018;10:A298–A306.

33. Azzimonti D, Rottondi C, Tornatore M (2019). Using active learning to decrease probes for QoT estimation in optical networks. Paper presented at: Proceedings of OFC 2019, San Diego, Feb 2019.

34. Sun M, Wang M, Guo Y, Mu H, Wu B, Jian S (2018, October). Optical modulation format recognition for polarization-multiplexed coherent system using Chinese restaurant process. In *Real-time Photonic Measurements, Data Management, and Processing III* (Vol. 10822, p. 108220W). International Society for Optics and Photonics.

35. Amari A, Lin X, Dobre OA, Venkatesan R, Alvarado A. A machine learning-based detection technique for optical fiber nonlinearity mitigation. *IEEE Photonics Technol Lett*. 2019;31(8):627-630.

36. Kashi AS, Zhuge Q, Cartledge JC, et al. Nonlinear signal-to-noise ratio estimation in coherent optical fiber transmission systems using artificial neural networks. *J Lightwave Technol*. 2018;36(23):5424-5431.

37. Floris I, Sales S, Calderón PA, Adam JM. Measurement uncertainty of multicore optical fiber sensors used to sense curvature and bending direction. *Measurement*. 2019;132:35-46. https://doi.org/10.1016/j.measurement.2018.09.033.

38. Ojo OT, Anta AF, Lillo RE, Sguera C (2019). Detecting and classifying outliers in big functional data. *arXiv preprint arXiv:1912.07287*.

39. Azcorra A, Chiroque LF, Cuevas R, et al. Unsupervised scalable statistical method for identifying influential users in online social networks. *Sci Rep*. 2018;8(1):6955.

40. López-Pintado S. Discussion of multivariate functional outlier detection by M. Hubert, P. Rousseeuw and P. Segaert. *Stat Methods Appl*. 2015;24(2):253-256.

41. Nyarko-Boateng O, Adekoya AF, Weyori BA. Using machine learning techniques to predict the cost of repairing hard failures in underground fiber optics networks. *J Big Data*. 2020;7(1):1-16. https://doi.org/10.1186/s40537-020-00343-4.

42. Nyarko-Boateng O, Xedagbui FEB, Adekoya AF, Weyori BA. Fiber optic deployment challenges and their management in a developing country: a tutorial and case study in Ghana. *Eng Rep*. 2020;2(2):e12121.

How to cite this article: Nyarko-Boateng O, Adekoya AF, Weyori BA. Predicting the actual location of faults in underground optical networks using linear regression. *Engineering Reports*. 2021;3:e12304. https://doi.org/10.1002/eng2.12304