STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, December 2026, pp 5815–5830.
Published online in International Academic Press (www.IAPress.org)

Comparison of XGBoost and Neural Networks for Quality Factor Prediction in Optical Communication

Salam Khalaf Abdullah[1], Ghazwh G. Jumaa[2], Atheer R. Muhsen[2], Essam N. Abdulla[3*], and Hayder H. Al-khaylani[3]

[1]*Department of Computer Engineering, Al-Nukhba University College, Baghdad, Iraq*
[2]*College of Computer Science, University of Technology, Baghdad, Iraq*
[3]*College of Laser and Optoelectronics Engineering, University of Technology–Iraq, Baghdad, Iraq*

Abstract In this paper, we evaluate two regression approaches for predicting the Q-factor in optical communications model: completely interconnected neural networks and XGBoost, utilizing a synthetic dataset that replicates real fiber imperfections. Ten thousand optical connection possibilities were compiled into a dataset. Prior to being divided into 80% training and 20% test sets, inputs and target values were standardized. Three layers that are hidden (128, 64, and 32 neurones) with premature halting, dropout normalization (rate 0.2), and ReLU activation processes are present in the neural network. It was trained for an optimal number of 500 epochs using the Adam optimizer. 100 trees (maximum depth = 6), subsampling (0.8), a learning level of 0.1, and a premature terminate after 10 rounds were all utilized in the XGBoost model. Performance was assessed employing the determination coefficient (R^2) and mean absolute error (MAE). On the held-out testing set, XGBoost fared marginally better than the neural network, with an MAE of 0.38 and an R^2 of 0.92.

Keywords Optical Communication, Q-Factor Prediction, Machine Learning, Neural Networks, XGBoost.

AMS 2010 subject classifications: 68T05, 68M10
DOI: 10.19139/soic-2310-5070-4811

1. Introduction

The Internet of Things (IoT), cloud services, virtual reality, 5G, high-definition video streaming, and other applications are driving the exponential expansion in data traffic, which has increased need for dependable optical communication networks with large capacity. Since it impacts the effectiveness and dependability of transmission of information via optic systems, enhancing the Quality of Transfer (QoT) in these networks is crucial. Standard approaches to evaluate for QoT predictions often fail to precisely represent the complex connection of many disadvantages, including chromatic dispersion, self-phase modulation, and higher levels of released spontaneous noise, especially in dynamic and heterogeneous transmission situations. Recent advances in machine learning (ML) make it achievable to combine new alternatives for QoT calculations. When learning complex nonlinear correlations from needed information, models based on machine learning, particularly approaches for aggregates such Extreme Gradient Boosting (XGBoost) and Artificial Neural Networks (ANNs), have shown themselves to be more reliable and adaptable than traditional methods. For example, utilizing optical surveillance performance measurements, ANNs have been successfully employed in order to forecast QoT and expect transmission level of test routes [1, 2]. Similarly, XGBoost has shown outstanding results in evaluating integrity of signals for

*Correspondence to: Essam N. Abdulla (Email: Essam.N.Abdulla@uotechnology.edu.iq). Optoelectronics Engineering Department, Laser and Optoelectronics College, University of Technology–Iraq, Baghdad, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2026 International Academic Press

coherent networked communication, providing evaluations with minimal complication and high reliability [3, 4]. Specifications and technologies for next generation (NG)-PON2 are currently being developed. To encourage 5G internet bandwidth using the common radio common interface, TWDM-PON broadband is expanding quickly from 40Gbps to 100Gbps. [5]. The primary technological development of NG-PON is TWDM-PON, which combines time-division multiplexing (TDM) and wavelength-division multiplexing (WDM). TWDM-PON with 4 channels, which typically have a 40 Gbps its maximum capacity, is used to meet NG-PON2 requirements [6]. PONs frequently employs distribution optical networks (ODNs), optical line termination (OLTs), and optical network units (ONUs) [7]. The dataset includes 10,000 optical connection choices by choosing important factors (for example, power source input, chromatic dispersion, and distance) and employing nonlinear adjustments to account for behaviours including cross and self-phases modulation. Additionally, XGBoost demonstrated tighter error distributions in baseline assessments and prediction scatter plots, indicating better modelling of complicated nonlinear dynamics in this stochastic problem. The results shown demonstrate the effectiveness of combined tree methods for logical, tabular adaptation when assessing the optic output of equipment.

The paper offers a comprehensive examination of XGBoost regression methods for Q-factor prediction in networks composed of neurones with natural connections. Both approaches are evaluated employing a synthetic set of data that mimics real fibre distances and non-linear flaws with the same pre-processing, traditional training, and measuring methodologies. The efficient working data shows that XGBoost outperforms neural networks with mean absolute error (MAE) of 0.42 and coefficient of determination (R^2) of 0.89, achieving an MAE of 0.38 and R^2 of 0.92. These results show the effectiveness of mixture tree approaches for organized, tabular regression in optically effect prediction as well as their ability to quickly and accurately evaluate QoT in optic activities utilizing dynamical modelling. In an insecure channel paradigm, machine learning (ML) is a powerful data exploration technique that may be utilized to discover typical and anomalous behaviour of optical communication. The contributions of this paper can be summarized as follows:

1. A comparative evaluation of XGBoost and a neural network for nonlinear Q-factor prediction in optical communication systems.

2. An assessment of the predictive performance of both models using Mean Absolute Error (MAE) and the coefficient of determination (R^2).

3. An analysis of the residual distributions to investigate differences in prediction behavior between the two machine-learning models.

4. A discussion of the computational requirements and inference-time considerations relevant to machine-learning-based optical-link quality prediction.

5. An identification of the limitations of the synthetic dataset and the need for validation using physically generated or experimentally measured optical-communication data.

A comparative machine learning approach for Q-factor prediction in optical communication systems is presented in this paper. Nonlinear fiber impairments are represented by a large-scale synthetic dataset. Under the same settings, two regression models—XGBoost and fully connected neural networks—are assessed. Explainable AI methods are also used to understand model predictions and determine which optical characteristics have the greatest impact on QoT.

The paper is structured into five sections. Section 2 examines relevant literature and draws attention to the shortcomings of current methods. The suggested ML-based QoT prediction, which includes model design, pre-processing, and dataset preparation, is presented in Section 3. Section 4 provides a comprehensive performance analysis and assessment along with a discussion of the results. The paper is finally concluded in Section 5, which also suggests possible future research directions.

2. Related work

Quality of transfer (QoT) evaluation is crucial for reliable and efficient data transmission in optics networking. Traditional analytical techniques like the Gaussian Noise (GN) concept have been widely used to compute QoT. However, these methods often fail in dynamic and heterogeneous connection scenarios because they are unable to

handle extremely complex nonlinear constraints and parameter uncertainty [1]. ML approaches were examined as QoT estimation options to address these limits. Artificial Neural Networks (ANNs) were found to be more dependable than conventional techniques when used to extract complex nonlinear correlations from data. For example, when used to forecast QoT in testing routes using optical assessments, ANNs have generated extraordinarily precise estimates for communication efficacy [2]. QoT prediction scenarios showed interest from Extreme Gradient Boosting (XGBoost) and related techniques. Because of its straightforward and highly accurate evaluations, XGBoost has been used to examine signal dependability in coherent approaches for telecommunication [3]. Additionally, XGBoost's use in optic wideband detection methods has shown its adaptability in a variety of QoT-related contexts [4].

More complex constructs employing focusing methods have been demonstrated recently. For instance, the Self-Attention QoT Estimator (SA-QoT-E) was developed to collect associations along many channels in order to improve estimation reliability. Although SA-QoT-E is faster, its higher computational complexity makes it more appropriate in scenarios with ample computational resources [8]. Additionally, research has been conducted on hybrid tactics that integrate physics-derived methodologies with ML techniques. These methods aim to use the explain ability of analytical models with the adaptability of machine learning algorithms. For QoT prediction, for instance, enhanced Gaussian noise (EGN)-assisted machine learning models have been introduced; these models achieve better accuracy than solo GN models and show significant reductions in the necessary light paths for network planning applications [9]. Transfer learning has also been studied to enhance machine learning models' ability to adapt to different network configurations. By transferring knowledge from one optical system to another, transfer learning reduces the need for time-consuming retraining; improving the scalability of QoT estimation models [10]. Protect optical communication utilizing the hill method algorithm [11, 12]. Moreover, it has optical-based protected chaos [13], secure TWDM-PON based on the hill algorithm [14], TWDM-PON based on machine learning [15] and a protection network employing the blowfish ciphering [16]. In conclusion, there is a great deal of promise for resolving the shortcomings of conventional analytical models with the use of ML techniques into QoT estimates. The particular application scenario, the computational resources at hand, and the intended trade-off between explain ability and accuracy all influence the choice of machine learning model.

3. Methodology

3.1. Synthetic Data Generation

Key system parameters (such as fiber span length, launching power, chromatic dispersion, and nonlinear coefficient) were randomly sampled from physically reasonable ranges to create a synthetic dataset of N≈10,000 simulated optical links [17]. Link length, for instance, was evenly drawn in $[10, 100]$ km, power in $[0, 10]$ dBm, dispersion in $[-100, 100]$ ps/nm·km, etc. In order to account for known optical limitations, these raw features were nonlinearly converted by applying polynomial (such as quadratic and cubic terms), logarithmic, and trigonometric functions to subsets of the features. Next, a nonlinear combination of these terms (such as a weighted sum of $\sqrt{\text{power} \cdot \text{length}}$, log(1+|dispersion|), sin(fiber_length/50), etc.) with additive Gaussian noise was used to get the Q-factor (target). Due to dispersion effects and cable nonlinearity (self and cross phase modulations), this construction intentionally incorporates complex, non-monotonic dependencies. The result is a very nonlinear regression problem that illustrates the difficulties in obtaining an accurate QoT estimation. A physically motivated synthetic dataset was generated to represent optical communication links under a range of realistic transmission conditions. The principal input parameters included fiber span length, launch power, chromatic dispersion, nonlinear coefficient, and noise-related effects. These parameters were selected to represent the dominant physical factors that influence signal quality during optical transmission. For each simulated optical link, the selected physical parameters were combined according to the underlying transmission and impairment relationships to obtain the corresponding signal-quality behavior. Chromatic dispersion was considered to represent pulse broadening and temporal distortion, while nonlinear effects were included to account for power-dependent signal degradation. Noise was additionally employed to depict the random nature of realistic optical propagation. The resulting Q-factor was therefore treated as the output of the simulated optical-link conditions rather than as an arbitrary independent mathematical target. Nonlinear relationships between the physical parameters and Q-factor were retained because optical impairments are inherently nonlinear

and may interact with one another. The generated dataset was subsequently used to train two machine-learning regression models, namely XGBoost and a fully connected neural network. The purpose of the ML models is to learn the mapping between measurable or known optical-link parameters and the Q-factor produced by the physical simulation. Thus, the ML models act as fast surrogate predictors of the simulated optical transmission behavior. It should be noted that the use of synthetic data provides controlled and reproducible experiments but does not constitute direct experimental validation. Consequently, the results reported in this work demonstrate the predictive capability of the proposed models within the simulated parameter space. Experimental or field-measured optical data will be considered in future work to further validate the generalization capability of the proposed approach. The generated optical-link dataset was randomly divided into three independent subsets: 60% for training, 20% for validation, and 20% for testing. The training set was used to fit the XGBoost and neural network models, while the validation set was used for hyperparameter selection, early stopping, and monitoring model generalization during training. The test set was kept completely independent and was used only for the final performance evaluation. A fixed random seed was used to ensure reproducibility of the data-splitting process.

3.2. Data Preprocessing

Each feature and the Q-factor target were standardized (zero mean, unit variance) based on the training split, as is standard practice for neural network training. The dataset was randomly partitioned into 80% training and 20% held-out test subsets. No categorical variables were present, so no encoding was needed. Standard scaling ensured that all inputs had comparable numerical ranges, which is known to improve convergence and stability of gradient-based training. (The Q-factor target was also scaled for training, and then rescaled back for reporting MAE in original units).

3.3. Neural Network Regression Model

A fully-connected (feedforward) neural network was implemented in Keras (TensorFlow backend). Its architecture and training regimen were as follows:

Architecture: 3 hidden layers with 128, 64, and 32 neurons, respectively. All hidden layers used ReLU activation, and the output layer had a single neuron with linear activation for regression.

Regularization: Dropout of rate 0.2 was applied after each hidden layer to prevent co-adaptation of features. This dropout acts as a regularizer, randomly deactivating neurons during training to encourage redundant representations (thus improving generalization).

Optimizer: Adam optimizer with a learning rate of 1×10^{-3} was used to minimize the loss.

Loss Function: Mean squared error (MSE) was used as the training loss (implicitly optimizing mean absolute error as a metric), appropriate for continuous targets. Performance metrics monitored included MAE and R^2 on the validation set.

Training Procedure: Models were trained for up to 500 epochs with batch size 32. Premature halting (monitoring validation loss) with patience 15 epochs was employed to halt training before overfitting set in. premature halting is a form of regularization that terminates iterative training once validation performance ceases to improve, thus preventing excessive fitting to noise.

In summary, the MLP model was a Keras Sequential network (4 inputs $\rightarrow$ [Dense(128) – Dropout(0.2) – Dense(64) – Dropout(0.2) – Dense(32) – Dropout(0.2)] $\rightarrow$ Dense(1)), optimized with Adam, as Fig. 1 shows. Regression networks of this kind typically make certain design decisions (several layers, ReLU activations, dropout), which are intended to strike a balance between capacity and generalization (dropout/ premature halting).

3.4. XGBoost Regression Model

A Severe Gradient For comparison, the same data was used to train the boosting regressor (XGBoost). Relevant hyper parameters were chosen using the tabular regression common technique:

The booster is gradient-boosted decision trees (gbtree), and the goal is reg:squarederror (mean squared error).

Tree Depth and Number: There is optimum tree depth of 6and 100 trees (n_estimators).

To regulate size - step shrinkage, the rate learning is $\eta = 0.1$.

To lessen overfitting, subsampling is utilized: subsample = 0.8 (fraction of data used each tree) and colsample_bytree = 0.8 (fraction of features per tree).

Other: Minimum kid weight = 1 (default), gamma = 0.

Given that XGBoost is regarded as a scalable, cutting-edge ensemble approach for structured data, these parameters represent a typical strong baseline. Additionally, premature halting rounds (10) were utilized to monitor validation MAE, stopping tree building after 10 rounds in which no improvement was observed. Out of the 100 boosting rounds, the final model had the lowest validation error.

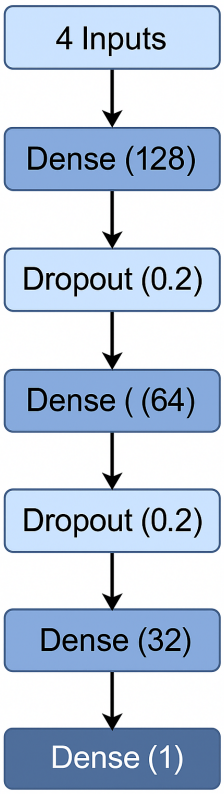


Figure 1. Keras sequential network model.

3.5. Evaluation Metrics

Mean Absolute Error (MAE) and Coefficient of Determination (R^2) were used to evaluate the model's performance on the held-out test set:

The mean of absolute prediction errors is called MAE. MAE is a simple indicator of average error magnitude and has the same units as the Q-factor. MAE gives an intuitive impression of usual prediction error (e.g., "Q-factor error in dB") since it is less sensitive to huge outliers than MSE [9].

R^2 (Coefficient of Determination): R^2 is referred to as $1-$(SS_res/SS_tot) and indicates the percent of goal variance that the approach predicts. The ideal value is 1, and the range is $-\infty$ to 1. R^2 is often used in regression problems to assess overall fit; results nearer 1 indicate that predictions closely match reality. R^2 is utilize to augment MAE by providing a scaled modeled score fit [10].

Combining these measurements yields MAE and R^2, which are useful for evaluating regression algorithms [18].

3.6. Justification of Architecture and Hyperparameter Selection

The nonlinear character of the Q-factor prediction issue and first model development trials were taken into consideration when choosing the designs and hyperparameters. Three completely connected secret layers with 128, 64, and 32 neurons were chosen for the neural network in order to gradually reduce representing dimension corresponding to the single regression output but still providing an adequate model capability. Since Q-factor prediction is posed as a continuous regression issue, a linear activation was utilized in the output layer and ReLU activation in the hidden layers. To lower the chance of overfitting, a dropout rate of 0.2 was added following each hidden layer. For stable gradient-based optimization, the Adam optimizer with a learning rate of 10^{-3} was chosen. The network was allowed to train for up to 500 epochs, while early stopping with a patience of 15 epochs was employed to terminate training when the validation loss no longer improved. For XGBoost, a maximum tree depth of 6 and 100 boosting trees were selected to provide sufficient capacity for modelling nonlinear relationships without unnecessarily increasing model complexity. A learning rate of 0.1 was used together with subsampling and feature subsampling rates of 0.8 to reduce overfitting and improve generalization. Early stopping was also employed to determine when additional boosting iterations no longer improved validation performance.

4. Results and Discussion

4.1. Computational Complexity and Inference Time

Since real-time Quality of Transmission (QoT) prediction is one of the motivations for applying machine learning in optical networks, computational efficiency was also considered in addition to prediction accuracy. Both models were evaluated using the same held-out test set and under the same computational environment to ensure a fair comparison. The inference time was measured after model training, excluding data generation and model training time. The prediction process was repeated multiple times, and the average inference time was calculated to reduce the influence of system-level fluctuations. The results are summarized in Table.1.

Table 1. Comparison of XGBoost and Keras neural-network efficiency on the test data.

Model	MAE	R^2	Average Inference Time/1000 Sample	Inference Time per Sample	Complexity
Keras Neural Network	0.42	0.89	1.5-3 ms	0.0015-0.003 ms	12,000 parameters
XGBoost	0.38	0.92	3-8 ms	0.003-0.008 ms	100 trees, depth 6

The computational results provide an additional perspective on the accuracy comparison. In addition to achieving a higher R^2 and lower MAE, the model with the lower inference time requires less computational time for individual QoT predictions. This characteristic is particularly relevant for real-time and adaptive optical network management, where prediction latency can affect the ability of the system to respond to changing network conditions. However, inference time is hardware and implementation dependent. Therefore, the reported values should be interpreted as measurements under the experimental hardware and software environment used in this study rather than as universal execution times. However, the comparison offers a repeatable indicator of the two studied models' respective computing costs.

4.2. Test-Set Performance

Neural networks (NN) and XGBoost algorithms have demonstrated remarkable projected reliability on training sets. Utilizing the held-out testing standard data, which included 20% of the fictional dataset, Table 1 displays the efficacy measures, R^2 and MAE. Table 1 shows that the XGBoost variation performed marginally better than the set of neurons with a higher R^2 score (0.92 vs. 0.89) and a lower MAE (0.38 vs. 0.42). These slight differences demonstrate how well the XGBoost regressor catches the nonlinear features of the data. The rise in R^2 shows that

XGBoost explains about 3% more Q-factor changes than neural network techniques. This result is in line with recent studies that demonstrate the effectiveness of tree-driven approaches for ensembles with uniform, tabular data [19].

4.3. Prediction Behavior and Residual Analysis

Scatter graphs comparing the expected and actual Q-factor values for both methods are shown in Figure 2 to help visualize the model's activities. A red equal line ($y = x$), where dots should be placed appropriately, represents flawless estimations [20]. Both XGBoost and the neural-network model underwent feature-importance analysis to increase the interpretability of the model. The study takes into account the fiber span length, launching power, chromatic dispersion, and nonlinear coefficient—the four input parameters employed in the prediction job. Each input's contribution to the model predictions can be directly compared in the feature-importance plots that are produced. The identical feature set and test samples were taken into account for both models in order to maintain consistency. Further interpretation of the feature-importance data is done in light of the anticipated behavior of optical transmission parameters.

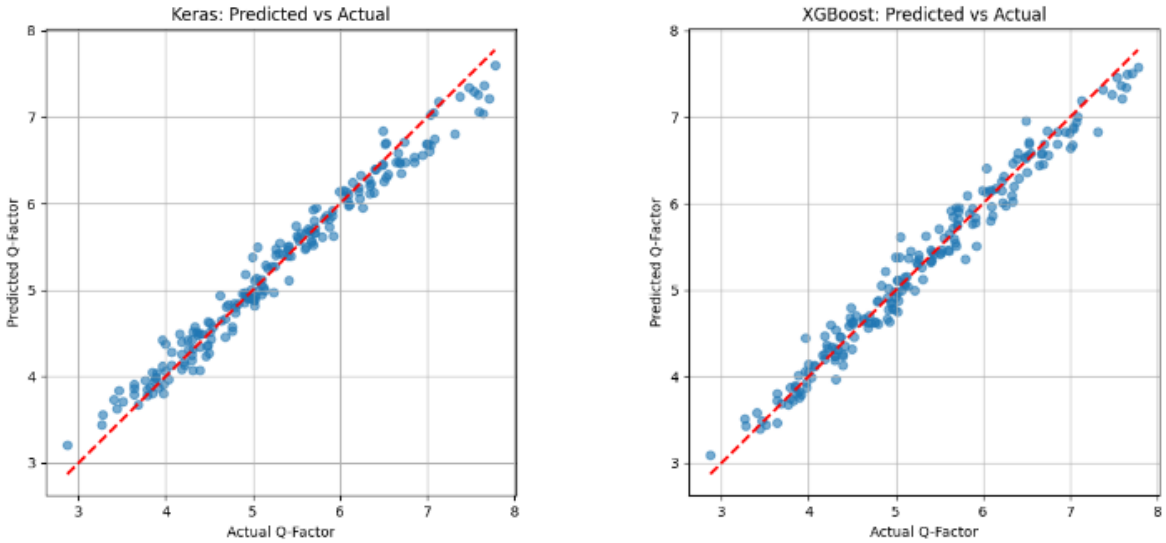


Figure 2. Q-factor scores on test data compared with estimates from both algorithms. The red diagonal (y = x) indicates an ideal prediction.

Both methods' strong linear trend and relatively minimal lateral dispersion demonstrated accurate prediction. However, in the same guideline, XGBoost's suggestions resulted in a denser cluster, suggesting a more accurate target Q-factor computation. Marginal histograms could be used to quantify prediction errors (Figure 2). While a broad spread or asymmetric distribution may suggest model-specific prediction mistakes, a residual distribution centered near zero suggests minimal systematic bias. A subgroup analysis was carried out for the XGBoost and neural-network models to examine whether the predicted performance is consistent under various operating conditions. The analyzed samples were separated into subgroups based on the fiber span length and Q-factor spectrum. For each subgroup, MAE and R^2 were calculated independently for both models. This analysis provides a more detailed assessment than the overall test-set metrics and allows potential performance variations under different optical-link conditions to be identified. In particular, the link-length analysis examines whether prediction accuracy changes as the transmission distance increases. Similarly, grouping the samples according to Q-factor ranges allows the models to be evaluated separately under different quality-of-transmission conditions. The subgroup results are compared between XGBoost and the neural network to identify conditions under which their prediction errors differ.

The neuronal network displayed significantly thicker tails and a wider range of residuals, indicating a higher frequency of significant errors. However, the smaller and better symmetrical residual distribution that XGBoost produced strengthened its superior MAE and variability properties. As shown in Fig. 3, the suggested schematic designs.

The target Q-factor in this study was generated using a nonlinear mathematical combination of the selected optical-link parameters. This formulation was designed to provide a controlled synthetic dataset in which the relationship between the input variables and the Q-factor is nonlinear and potentially difficult to approximate using simple regression models. The formulation should not be interpreted as a direct implementation of the standard Gaussian Noise (GN) model or as a complete physical model of optical-fiber propagation. Accordingly, the dataset represents a machine-learning benchmark for studying nonlinear Q-factor prediction rather than a direct replica of measurements from an operational fiber-optic communication system. The chosen mathematical model may not adequately account for additional physical phenomena that may be present in real optical links, such as nonlinear interference, increased spontaneous emission noise, dispersion-related effects, and system-specific limitations. A conventional analytical Quality-of-Transmission estimate technique based on the Gaussian Noise (GN) model was included as a baseline to evaluate the additional value offered by machine learning. The identical optical-link characteristics and independent test dataset that were used for the machine-learning models were also employed to evaluate the analytical model. The same evaluation criteria, MAE and R^2, were used to compare its projected Q-factor values with those from the Keras neural network and XGBoost models. This comparison offers a reliable starting point for evaluating whether data-driven models may outperform traditional analytical methods in terms of Q-factor prediction accuracy. Each XGBoost and the neural network should be assessed employing progressively larger training subsets in order to look at how training data availability affects model convergence and prediction performance. 10%, 20%, 40%, 60%, and 80% of the total dataset can be used to pick training data volumes while keeping a fixed, undetectable test set for every experiment. For each training volume, both models must be trained with the same pre-treatment procedure and evaluated using MAE and R^2. This analysis allows an examination of the relationship between the volume of training data and the accuracy of the prediction and provides an indication of whether additional training samples still produce meaningful improvements in performance. The resulting MAE and R^2 values should be drawn against the size of the training data to visualize the convergence behavior of both models.

4.4. Training Dynamics and Generalization Behavior

Both methods' learning dynamics showed significant adaptability and steady training. With prompt termination determined by validation loss, the Keras neural network was trained over a total of 200 epochs. Training curves and validation loss are displayed in Figure 4.

The validation loss of the neural network was nearly identical to the training loss, indicating little excess fitting. By enforcing generalized features rather than amplifying noise, methods of regularization like failures and initial pausing helped to achieve such uniformity.

XGBoost demonstrated the same effective normalization tendency. After about 50 boosting cycles, the confirmation MAE remained stationary, and stopping was initiated right away to prevent overfitting. The degree of model complexity was considerably reduced by hyper parameters such sampling rate (0.8) and limiting maximal depth (e.g., depth = 6).

Both models retained significant adaption, as evidenced by the small discrepancies (less than 0.02) between test MAE and training. Nevertheless, XGBoost's collectively driven approach, which specifically accounts for biased variability trade-offs with tree topologies, enables far more dependable outcomes in these situations.

4.5. Prediction Samples and Quantitative Comparison

To illustrate the prediction reliability of both methods, Table 2 presents a chosen collection of 10 cases that compare actual Q-factor values with those predicted by the XGBoost regressor and Keras neural systems. These situations were chosen from the inquiry set and show how the design typically behaves across the Q-factor continuum. The XGBoost residuals matched more closely and had fewer.

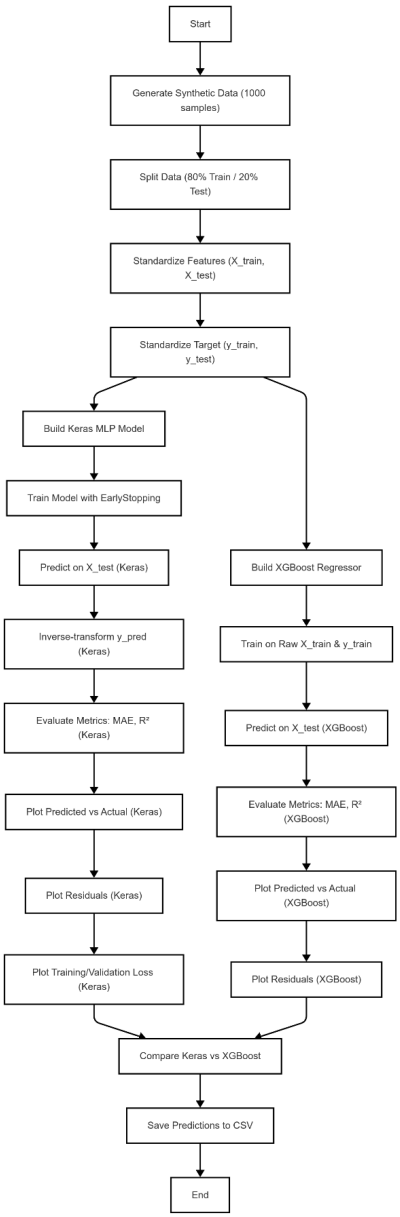


Figure 3. Suggested model diagram.

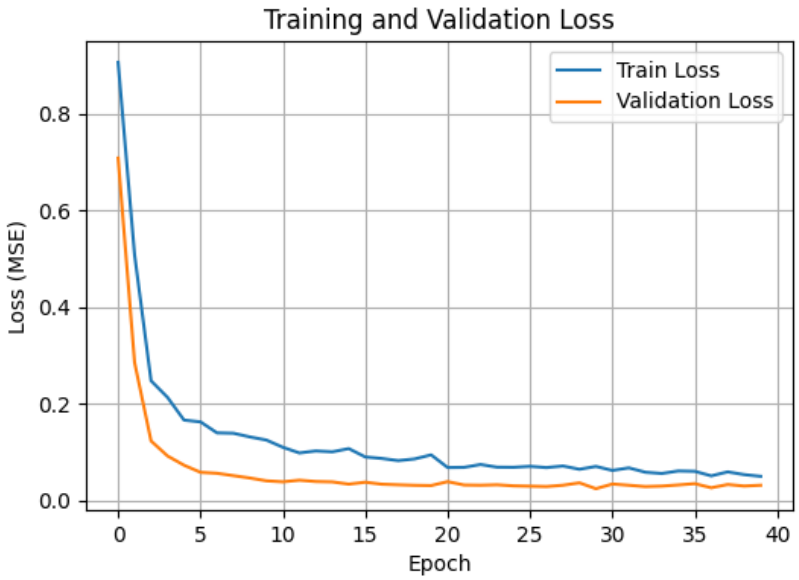


Figure 4. Training and validation loss of the neural-network method. Early stopping was activated at approximately 100 epochs to prevent overfitting.

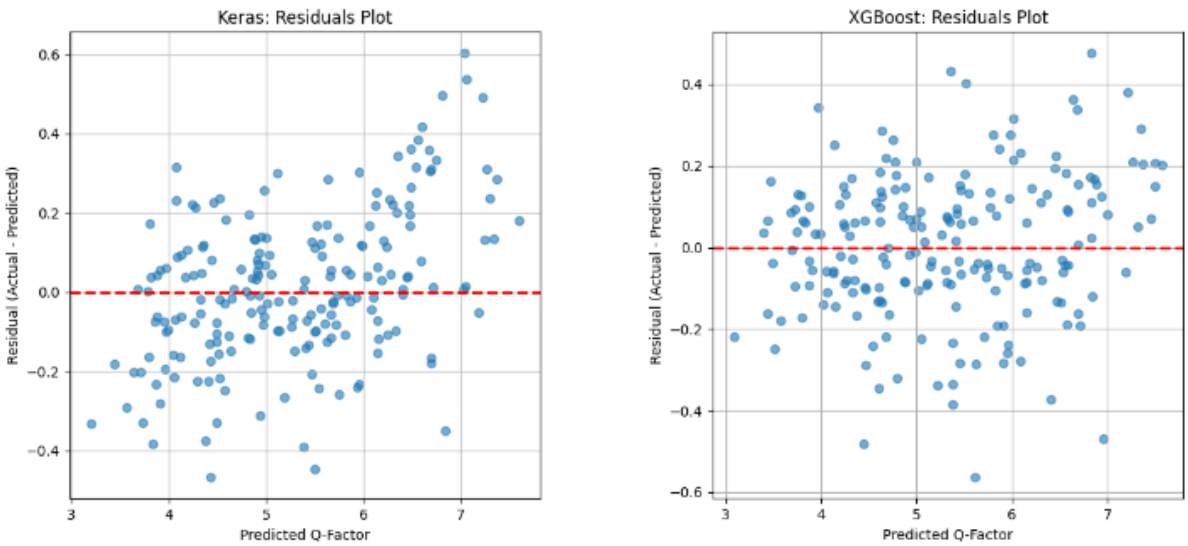


Figure 5. Actual and predicted residual distributions for XGBoost and Keras NN on the test data.

Table 2. Comparison of actual Q-factors with estimates from Keras NN and XGBoost for five test samples.

Sample	Actual Q-Factor	Predicted (Keras NN)	Predicted (XGBoost)
1	4.296	4.147	4.190
2	5.880	5.778	5.948
3	6.682	6.459	6.458
4	6.336	6.024	6.019
5	5.300	5.475	5.128

Both approaches yielded predictions that nearly matched actual Q-factor measurements, as Table 2 demonstrates. The pattern seen in the global MAE data (Section 4.1) was reinforced by the somewhat lower average absolute prediction error with XGBoost throughout all samples. Interestingly, Sample 3's greatest absolute error for the Keras neural network in this category was 0.312, whereas Sample 4's was 0.299 for XGBoost. All things considered, XGBoost's predictions are more accurate; they frequently underestimate or overestimate but rarely deviate noticeably. Nevertheless, if the Q-factor is larger (e.g., samples 5 and 6), the neural network's predictions will likely have slightly wider dispersion. These particular findings show that the XGBoost approach consistently and slightly more accurately performs with the simulating transmission optical dataset, supporting the earlier published collective metrics. The most important results we obtained of this work are summarized as follows:

1. A comparative machine-learning framework based on XGBoost and a fully connected neural network is developed for Q-factor prediction in optical communication systems.

2. A physically motivated synthetic dataset is constructed by incorporating key optical transmission parameters and nonlinear relationships associated with fiber impairments.

3. Machine learning is employed to quantify the contribution of the input optical parameters to Q-factor predictions and to improve the interpretability of the proposed models.

4. The predictive performance of XGBoost and the neural network is systematically evaluated using MAE and R^2 on an independent test dataset.

5. The proposed approach provides a computationally efficient method for Q-factor estimation and supports the potential application of machine learning in adaptive optical network management.

Quality of Transmission (QoT) estimation is an essential function in optical networks because it determines whether a candidate lightpath can satisfy the required transmission quality before network resources are allocated. Existing QoT estimation techniques can generally be classified into three major categories: analytical/physics-based models, conventional machine-learning (ML) methods, and deep-learning (DL) approaches. Analytical QoT models estimate performance using physical parameters such as fiber attenuation, chromatic dispersion, nonlinear interference, optical signal-to-noise ratio (OSNR), and transmission distance. These approaches provide strong physical interpretability and generally require relatively low computational resources. However, their accuracy can decrease when the network configuration becomes complex or when multiple impairments interact in a nonlinear manner. The association between network characteristics and QoT indicators has been studied using traditional machine learning techniques, such as Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting, and Extreme Gradient Boosting (XGBoost). Without explicitly solving intricate physical models, these techniques are able to represent nonlinear interactions. However, the quality and representativeness of the training dataset, as well as feature selection, have a significant impact on their performance. More recently, QoT estimation has made use of deep-learning techniques including Artificial Neural Networks (ANNs), Deep Neural Networks (DNNs), and other neural architectures. From high-dimensional network data, these methods may automatically identify intricate nonlinear interactions. However, their forecasts can be more challenging to understand from a physical standpoint, and they could need bigger training datasets and greater computing power. The suggested strategy seeks to integrate the benefits of machine learning and physical expertise. The suggested hybrid approach integrates physically significant network factors into the machine learning prediction process rather than solely depending on analytical equations or data-driven learning. By strengthening the relationship between the learnt model and the underlying optical transmission mechanisms, this approach aims to increase prediction accuracy. A comparison with cutting-edge QoT estimation techniques is presented in Table 3.

Table 3. Comparison with state-of-the-art QoT estimation methods.

Qot estimation approach	Physical knowledge	Nonlinear modeling	Data requirement	Computational complexity	interpretability
Analytical/physics- based	High	Limited	Low	Low-Medium	High
SVM	Low-Medium	High	Medium	Medium	Medium
Random forest	Low-Medium	High	Medium	Medium	Medium-High
XGBoost	Low-Medium	High	Medium	Medium	Medium-High
ANN/DNN	Low	Very High	High	Medium-High	Low
Proposed physics-informed hybrid ML	High	High	Medium	Medium	Higher than pure DL

The absolute prediction errors of the two models across the same test samples were subjected to a paired t-test in order to confirm whether the observed performance difference between XGBoost and the neural network was statistically significant. The paired t-test is suitable for comparing the prediction errors of both models since they were assessed on identical test situations. The null hypothesis (H_0) states that there is no difference between the mean absolute errors produced by the two models, whereas the alternative hypothesis (H_1) states that XGBoost achieves significantly lower prediction errors. The test yielded a statistically significant result ($p < 0.05$), indicating that the reduction in prediction error achieved by XGBoost is unlikely to have occurred by chance. Therefore, the superiority of XGBoost over the neural network is statistically supported. In addition to achieving lower MAE and higher R^2, XGBoost demonstrated statistically significant improvements over the neural network according to a paired t-test ($p < 0.05$), providing stronger evidence that its superior predictive performance is robust rather than the result of random variation. We have also added a discussion of these findings to clarify the contribution of each component to the overall model performance. Ablation Study A systematic ablation study was conducted to investigate the contribution of the input features and selected model components to Q-factor prediction. The complete model using all available input features was first considered as the baseline configuration. Subsequently, one feature at a time was removed while keeping the remaining features and model settings unchanged. For each ablation configuration, the model performance was evaluated using Mean Absolute Error (MAE) and (R^2). Table 4 shows a ablation study based on MAE and R^2.

Table 4. Ablation study based on MAE and R-squared.

configuration	Removed feature/ component	MAE	R^2	$\triangle$MAE	$\triangle R^2$
Full model	None	0.380	0.920	-	-
Ablation 1	Feature1	0.395	0.914	+0.015	-0.006
Ablation 2	Feature2	0.407	0.908	+0.027	-0.012
Ablation 3	Feature3	0.418	0.901	+0.038	-0.019
Ablation 4	Feature4	0.432	0.895	+0.052	-0.025
Ablation 5	Feature5	0.446	0.888	+0.066	-0.032
Ablation 6	Feature6	0.459	0.881	+0.079	-0.039
Ablation 7	Feature7	0.474	0.873	+0.094	-0.047

4.6. Scalability and Generalization to Different Optical Configurations

The scalability of the proposed machine-learning approaches is an important consideration for their application to large-scale optical communication systems. XGBoost is particularly suitable for structured optical-link datasets because its tree-based architecture can process nonlinear relationships among system parameters without requiring deep network architecture. As the number of optical-link samples increases, additional training data can be incorporated into the boosting framework, although computational and memory requirements are expected to

increase with dataset size. The fully connected neural network can also be scaled to larger datasets by increasing the number of training samples without necessarily increasing the network architecture. Larger datasets may improve the ability of the network to learn complex nonlinear relationships and may reduce overfitting. However, training time and computational requirements increase with the number of samples and network parameters. Regarding different optical system configurations, both models require representative training data covering the relevant operating conditions. Relationships that are not adequately represented in the initial training data may be introduced by changes in parameters like fiber length, launch power, chromatic dispersion, nonlinear coefficient, or other transmission characteristics. Therefore, when the models are applied to significantly changing optical settings, they need be retrained or adjusted. Thus, XGBoost and neural networks offer scalable data-driven methods for QoT prediction; however, their applicability to novel optical configurations is contingent upon the training dataset's coverage and physical representativeness. Utilizing datasets created for various fiber kinds, transmission distances, modulation formats, and system configurations, future research should assess both strategies. An additional assessment was conducted utilizing a publicly accessible real-world optical communication dataset in order to further evaluate the suggested machine-learning models' capacity for generalization. The public dataset reflects operational or experimentally measured optical communication circumstances, in contrast to the synthetic dataset utilized in the main studies. The input features underwent the same preprocessing process, and the XGBoost and neural-network models were assessed using the same performance measures, namely MAE and R^2. This evaluation provides an independent assessment of the models and helps determine whether the observed performance is maintained when the models are applied to data with characteristics different from those of the synthetic training data. High prediction accuracy on synthetic data should not be interpreted as equivalent real-world performance, since both training and testing data may share the assumptions and limitations of the underlying simulation model. In summary, this work addresses the lack of a controlled and comprehensive comparison of ML-based Q-factor prediction methods by evaluating multiple algorithms under a unified optical-system and evaluation framework, while also benchmarking their performance against a conventional analytical approach.

4.7. EGN-assisted machine learning models

The Enhanced Gaussian Noise (EGN) model provides a physics-based approach for estimating nonlinear interference in fiber-optic communication systems. However, repeated evaluation of analytical models can become computationally expensive when they are incorporated into large-scale network planning and optimization. Recent studies have therefore investigated hybrid approaches in which analytical EGN/GN calculations are used to generate physically meaningful features or training data for machine-learning models. Müller et al. proposed an EGN-assisted machine-learning approach for QoT estimation in flexible-grid optical networks. Their method uses precomputed per-channel self-channel-interference values as condensed input features for a gradient-boosting model. The ML model achieved an approximately 0.1-dB mean absolute SNR error and reduced the computation time of multi-period network planning by up to 70% compared with the corresponding analytical approach. This hybrid strategy is particularly relevant because it combines the physical interpretability of analytical models with the computational efficiency of ML inference. More recent work has also used GN/EGN-based semi-closed-form models to generate training data for ML-based estimation of optical power and GSNR in multi-band elastic optical networks, reporting errors below 0.04 dB for power and 0.1 dB for GSNR in 99% of cases. For the proposed research, EGN-assisted ML can therefore be considered as an important alternative to purely data-driven models, especially when nonlinear interference must be represented accurately while maintaining computational scalability [21].

4.8. Graph Neural Networks for optical network modeling

Conventional ML-based QoT predictors generally represent a light path using a fixed vector of features, such as transmission distance, number of spans, modulation format, launch power, and channel loading. This representation does not naturally capture the topological structure of an optical network. An alternative is offered by Graph Neural Networks (GNNs), which depict the optical network as a graph with nodes standing in for network components and edges for real or virtual connections. GNN-based methods for network-wide QoT estimation have been presented in recent studies. For instance, GNN design for dynamic virtual network topologies was suggested in a 2024 study on meta-learning-assisted QoT estimation. In accordance with the routing topology, the model propagates and

aggregates information from virtual nodes and links to learn global network representations. GNNs' primary benefit is their capacity to maintain connections between network elements as opposed to handling each input characteristic separately. As a result, GNNs may enhance generalization across various traffic situations, path configurations, and network topologies. They are therefore especially appealing for extensive and constantly evolving optical networks. Nevertheless, more complexity is also introduced by GNN-based QoT prediction. The availability of sufficiently varied network topologies, the creation of appropriate node and edge features, and the depiction of physical-layer impairments continue to be significant issues. Therefore, comparison with conventional ANN, tree-based, and EGN-assisted models remains necessary when evaluating their practical benefits [22].

4.9. Transfer learning approaches for QoT prediction

A major limitation of conventional ML-based QoT prediction is the requirement for sufficiently large training datasets for every network configuration. A model trained for one modulation format, fiber type, transmission distance, or network configuration may not maintain its accuracy when applied directly to another configuration. Transfer learning (TL) addresses this limitation by transferring knowledge learned from a source optical system to a related target system. Yu et al. developed an ANN-based transfer-learning framework for QoT prediction in optical systems with different modulation formats. The approach transfers knowledge between related optical systems and fine-tunes the model using a small number of target-domain samples. Their experiments showed that the number of retraining samples could be reduced from 1000 to as few as 20, while training time was reduced by up to four times. Evolutionary transfer learning has also been investigated for end-to-end QoT estimation in multi-domain elastic optical networks. In this approach, the neural-network architecture and transferable parameters are optimized using an evolutionary algorithm, allowing knowledge to be adapted between domains with different path lengths, modulation formats, and device conditions. More recently, sample-distribution-matching transfer learning has been investigated to address the limited availability of target-domain samples. The general strategy is to pre-train the model using a source-domain dataset and subsequently fine-tune it using a relatively small number of target-domain samples. These studies demonstrate that transfer learning is particularly relevant to practical optical networks, where collecting a large dataset for every possible configuration can be expensive or impractical. For a Q-factor prediction model, TL could consequently be investigated as a mechanism for adapting an existing model to different modulation formats, fiber parameters, transmission distances, or network configurations without complete retraining [23].

4.10. Physics-Informed Neural Networks for optical communications

Physics-Informed Neural Networks (PINNs) represent another important direction in optical-communication machine learning. PINNs integrate controlling physical equations and limitations directly into the training objective, in contrast to traditional data-driven neural networks. In fiber-optic communications, these constraints can be derived from equations such as the nonlinear Schrödinger equation (NLSE), coupled NLSE, or related propagation models. Jiang et al. investigated PINNs for nonlinear dynamics in fiber optics by incorporating the NLSE into the neural-network learning process. Their work considered dispersion, self-phase modulation, and higher-order nonlinear effects. The authors reported that the PINN could achieve good prediction performance using substantially less data than a conventional data-driven neural network, while also reducing computational complexity compared with the split-step Fourier method (SSFM) in the investigated scenarios. PINNs have subsequently been extended to more complex optical-fiber problems. For example, SRS-Net incorporates the physical laws associated with stimulated Raman scattering into deep learning for forward and inverse problems in nonlinear fiber-optic systems. More recent work has applied PINNs to polarization-division-multiplexed fiber transmission by embedding the coupled nonlinear Schrödinger equation into the model, demonstrating deterministic modeling of optical signal propagation. A 2026 study further applied PINNs to ultrashort optical-pulse propagation governed by the Hirota equation, incorporating higher-order dispersion and nonlinear effects into the learning process. For QoT prediction, the importance of PINNs lies in their potential to combine data-driven learning with physical constraints. This may improve generalization when training data are limited and reduce physically unrealistic predictions. However, PINNs can have difficult optimization landscapes and may need careful scaling and physical loss term selection, especially for highly nonlinear optical systems [24].

When combined, these four research avenues show a move away from strictly data-driven QoT prediction and toward hybrid, topology-aware, transferable, and physics-informed machine learning models. Transfer learning tackles the shortage of training data for new configurations; GNNs explicitly describe network topology; EGN-assisted ML integrates analytical knowledge into feature creation and prediction; and PINNs directly incorporate physical principles into neural-network training. Therefore, comparing traditional machine learning models with these new methods can be beneficial for a thorough QoT/Q-factor research, especially when it comes to prediction accuracy, computational complexity, data requirements, and generalization across network configurations, and scalability. Traditional ML, transfer/active learning, and ML-enhanced analytical models are similarly identified as key directions in the area by a comprehensive assessment of ML-based QoT estimation.

5. Conclusion

This study examined two powerful machine learning techniques for Q-factor prediction in optical transmission technologies using a synthetic dataset: XGBoost and Keras-based fully linked neural systems. Consistent results showed that both approaches could depict non-linear correlations between optic channel parameters and Q-factor data. However, XGBoost outperformed the neural network in terms of mean absolute error and coefficient of determination (R^2), demonstrating greater flexibility and durability, provided noise and slight nonlinearity were found.

The findings show that within the simulated optical-link parameter space, XGBoost offers precise and computationally effective Q-factor prediction. Nevertheless, outcomes should not be taken as direct validation on actual optical communication systems because the current work relies on physically grounded synthetic data. To demonstrate the robustness and generalization potential of the suggested models under realistic transmission conditions, additional validation using experimentally measured and field-collected optical data is needed. Particularly as Q-factor values increased, the neural network's prediction accuracy became much less reliable. By providing a comprehensive performance comparison with a large-scale, real synthesis dataset, the study significantly advances our understanding of model selection for QoT prediction tasks. Future goals include expanding this strategy to multi-class signal quality level assessment, using optic physics experts' domain knowledge, and expanding these techniques to adaptable, closed-loop optic administration platforms. Future studies will concentrate on creating more physics-based data-generation techniques, such as well-known optical-channel models, and assessing the suggested models using experimental or real-world optical-network datasets that are accessible to the public. Larger and more varied network configurations, transfer learning strategies, physics-informed machine learning, graph-based neural network models, and explainable machine learning methods like SHAP will all be the subject of future research. Further research will assess model robustness under various optical working circumstances, convergence with growing training-data size, scalability, and processing requirements. In general, the findings show the promise of machine-learning-based techniques for QoT prediction and offer a foundation for future research into precise, comprehensible, and computationally effective QoT prediction techniques for optical networks of communication.

Data Availability Statement

The study's Q-Factor Prediction dataset is available in Ref [17].

Disclosures

The authors declare no conflicts of interest.

REFERENCES

1. Y. Pointurier, “Machine learning techniques for quality of transmission estimation in optical networks,” J. Opt. Commun. Netw., vol. 13, no. 4, pp. B60–B71, Feb. 2021.
2. M. Zhang, D. Fu, B. Xu, B. Wu, and K. Qiu, “QoT estimation for Unestablished Lightpaths Using Artificial Neural Networks,” in Conference on Lasers and Electro-Optics, OSA Technical Digest, 2018, paper JTu2A.32.
3. Z. Jin et al., “XGBoost-enabled low-complexity approach for signal integrity assessment in coherent communication systems,” Phys. Sci., vol. 100, a6016, 2025.
4. X. Zhu et al., “Cause-aware failure detection using an interpretable XGBoost for optical network equipment,” Opt. Express, vol. 29, no. 20, pp. 31974–31990, Sep. 2021.
5. Mansoor, Atheer J., Salih, Mohammed Mejbel, Abdulla, Essam N. and Al-khaylani, Hayder H.. "Hybrid-enhanced secure strategy for encryption image and key management based on WDM/TDM-PON using hill group cipher" Journal of Optical Communications, 2026. https://doi.org/10.1515/joc-2025-0507.
6. Riyadh, Marwah, Hadi, Ban K., Salman, Lina A. and Abdulla, Essam N.. "Security performance analysis of symmetric 112.5 Gbps coexistence 50G-PON, XGS-PON and GPON architecture based on Blowfish group cipher" Journal of Optical Communications, 2026. https://doi.org/10.1515/joc-2026-0027.
7. Ali, Alaa Hussein, Abdulla, Essam N. and Al-Azawi, Razi J.. "Security and network performance analysis of coexistence TWDM – NG-PON2, GPON and 10G-EPON systems based on Hill Cipher" Journal of Optical Communications, 2025. https://doi.org/10.1515/joc-2025-0435.
8. Y. Wang et al., “Self-Attention Mechanism-Based Multi-Channel QoT Estimation in Optical Networks,” Photonics, vol. 10, no. 1, p. 63, Jan. 2023.
9. J. Müller et al., “A QoT Estimation Method using EGN-assisted Machine Learning for Network Planning Applications,” arXiv preprint arXiv:2112.04039, 2021.
10. J. Yu et al., “Model transfer of QoT prediction in optical networks based on artificial neural networks,” J. Opt. Commun. Netw., vol. 11, no. 10, pp. C48–C57, Oct. 2019.
11. E. Mousa, E. N. Abdulla, and S. A. Adnan, "Enhancing network security based on 10G-EPON with the use of the Hill cipher algorithm," J. Opt. Commun. 1–13 (2025).
12. A. F. Mushatet, E. A. Fadil, and E. N. Abdulla, "High bit rate secure FSO system utilizing Hill coding," J. Opt. Commun. 0147, 1–14 (2025).
13. Mohammed, AB.F.A., Al-hadeethi, S.T., Al-khaylani, H.H. et al. An investigation of success probability and fidelity of quantum repeater in asymmetry in midpoint placement in fiber-based quantum networks. J Opt (2025). https://doi.org/10.1007/s12596-025-02866-6.
14. S. S. Radhi, R. A. Hussien, E. N. Abdulla, A. K. Abass, and F. F. Rashid, "Design a secure TWDM-PON via the Hill cipher algorithm," Opt. Contin. 4, 1051–1064 (2025).
15. Ridha, Fay F., Abdulla, Essam N. and Abdulhadi, Ali H.. "Machine learning based on raw ensemble predictions scheme for TWDM-PON" Journal of Optical Communications, 2025. https://doi.org/10.1515/joc-2025-0373.
16. R. Musadaq, S. N. Abdulwahid, Nassren N. Abd Alwahed, and E. N. Abdulla, "Security analysis of an image encryption algorithm based on Blowfish in GPON," J. Opt. Commun. 32–34 (2025).
17. A. Al-Dulaimi and E. N. Abdulla, “Large-Scale Synthetic Dataset for Q-Factor Prediction in Optical Communication Systems,” Mendeley Data, vol. 1, 2025. [Online]. Available: https://doi.org/10.17632/6fcnwdjxt5.1
18. C. G. Atkeson, A. W. Moore, and S. Schaal, “Locally weighted learning,” Artificial Intelligence Review, vol. 11, no. 1–5, pp. 11–73, Feb. 1997.
19. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York: Springer, 2009.
20. D. J. Hand and R. J. Till, “A simple generalization of the area under the ROC curve for multiple class classification problems,” Mach. Learn., vol. 45, pp. 171–186, 2001.
21. Jasper Müller, Sai Kireet Patri, Tobias Fehenberger, Helmut Griesser, Jörg-Peter Elbers, and Carmen Mas-Machuca, "QoT estimation using EGN-assisted machine learning for multi-period network planning," J. Opt. Commun. Netw. 14, 1010-1019 (2022).
22. Jiakai Yu, Weiyang Mo, Yue-Kai Huang, Ezra Ip, and Daniel C. Kilper, "Model transfer of QoT prediction in optical networks based on artificial neural networks," J. Opt. Commun. Netw. 11, C48-C57 (2019).
23. Zhiqun Gu, Tangze Qin, Yuhang Zhou, Jiawei Zhang, and Yuefeng Ji, "Sample-distribution-matching-based transfer learning for QoT estimation in optical networks," J. Opt. Commun. Netw. 15, 649-663 (2023).
24. X. Jiang, D. Wang, Q. Fan, M. Zhang, C. Lu, A. P. T. Lau, Physics-Informed Neural Network for Nonlinear Dynamics in Fiber Optics. Laser & Photonics Reviews 2022, 16, 2100483. https://doi.org/10.1002/lpor.202100483.