A lightweight Machine-learning based Wireless Link Estimation for IoT devices

Khanh-Hoi Le-Minh[1,2], Kim-Hung Le[1,2], Quan Le-Trung[1,2]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
Email: {hoilmk,hunglk,quanlt}@uit.edu.vn
Corresponding author: hunglk@uit.edu.vn

Abstract—**Robust wireless communication between devices is crucial to ensuring the reliability of IoT systems. However, it is strictly relied on estimating the link quality of these devices, which is usually interfered with by environmental factors. In such a scenario, intelligent algorithms based on machine learning to select resistant communication links are promising solutions, but they demand sophisticated computation, limiting their deployment on resource-constrained IoT devices. Therefore, this paper introduces a lightweight link quality estimation algorithm, namely LLQE, built from the gradient boosting decision tree. The superiority of our proposal not only precisely assesses several levels of link quality but also is lightweight enough for resource-constraint devices. The evaluation results on publicly available datasets show that LLQE accurately estimates various link quality indicators with 97% accuracy.**

Index Terms—**Wireless link estimation, machine learning, Internet of Things devices.**

I. Introduction

We have been witnessing the transformation in several aspects of the industrial sector because of the rapid proliferation of the Internet of Things, which significantly increases the number of wireless devices deployed worldwide. Following the report in [1], over 170 billion IoT devices are expected to connect to the Internet by 2050. However, most IoT devices are resource-constrained, with limited processing, memory, and communication capabilities. In addition, these devices are deployed in a wide area with many obstacles, restricting radio signal propagation. Therefore, research on enhancing connectivity in IoT devices recently gained much attention from the research and industrial community [2].

Deploying IoT devices in a wide geographical area often causes various communication issues because of considerable fluctuation in the wireless signal generated by shadow fading, attenuation, and noises from the weather [3]. This negatively affects the reliability of communication links and may cause errors in transmitted data or event packet loss, damaging the fault-tolerance and dependability of the whole IoT system. Thus, recognizing the quality of wireless links in advance is crucial for IoT devices to select appropriate routing algorithms or alternative links. It also lowers the energy consumption to re-transmit network packets, extending the IoT device life cycle [4].

Most existing link quality estimation methods based on machine learning leverages classification models for modeling physical and link-layer information collected by observing network packets. For example, the authors in [5]–[7] indicate the link quality by modeling the relationship between received signal strength indicator (RSSI), signal-to-noise ratio (SNR), and packet reception rate (PRR). The other works apply statistical data analysis methods to this information [8], [9]. However, these works still own several limitations, including:

- Using the very limited level of quality information (only "bad" and "good"). This significantly reduces the benefit of indicating link quality.
- Using statistical information of network within a specific time-window, leading a delay in estimating link quality.
- Using high computational models that are insufficient for resource-constraint devices.

Motivated by these issues, in this paper, we first analyze the wireless network metric (e.g., RSSI, SNR, PRR) and its relationship with link quality indicators. Then, we introduce a link quality estimation method based on a gradient boosting decision tree that accurately estimates four levels of link quality and is lightweight enough to deploy to IoT devices with limited computation capability and being compatible with edge frameworks [10], [11]. The main contributions of this paper are summarized below:

- A method of data preprocessing is proposed during the link quality estimation process to remove noise in sample data and deal with an imbalanced sample of link quality. This paper uses interpolation to replace noise in the data sample. SMOTE is applied to deal with imbalanced sample data so that each link quality level sample is balanced.
- A generated synthetic feature is applied to affect model accuracy. Polynomial regression generates synthetic features by adding powers to the original features.
- A link quality estimation model for IoT devices based on the Gradient Boosting decision tree is proposed, which

978-1-6654-9927-9/22/$31.00 ©2022 IEEE

classifies the link quality into four groups: good, immediate, bad, and very bad.

The rest of the paper is organized as follows. Related research is presented in the section II. The proposed estimator is explained in section III. The systems are evaluated under the section IV. Finally, section V concludes and discusses future research opportunities.

II. Related works

Machine learning and deep learning algorithms are becoming more popular and used in various fields. Many scientists have implemented these methods to resolve the LQE problem, and they have proven to be quite effective.

Wei Sun et al. [12] proposed a WNN-LQE system to improve routers based on the SNR value. WNN-LQE can also determine the trust interval of the PRR value and assess whether the system meets the practical implementation criteria. The Link Quality Indicator (LQI) value was examined by the authors in [13] to enhance routers and create more efficient connection selections. The algorithm divides a connection into three levels based on LQI and other metrics, such as RSSI: good, connectable, and bad. The approach combines handling with fewer characteristics and classes, resulting in faster and more accurate prediction results and a higher level of trust. Miguel L. Bote-Lorenzo et al. [14] combine machine learning techniques and online perception algorithms to create a link quality estimation system that can be updated in real time to match the adaptability of the connection. The system may assess the quality of the linkages using this model by integrating algorithms by anticipating a particular trust or value matrix.

Jian Shu et al. [15] present an LQE model that classifies link quality into five categories: very good, good, medium, bad, and very bad, based on an SVM algorithm paired with a decision tree method. The suggested approach delivers greater outcomes and lower processing costs without reducing the accuracy of the prediction findings by transforming them into a classification issue rather than directly predicting the PRR value, as in prior research. The authors in [16] offer a systematic quantification of the impact of the design steps on the final performance of a wireless link quality classifier. The proposed used a decision tree to classify the link quality into three levels: good, intermediate, and bad. In the preprocessing stage, the approach analyzes the impact of re-sampling on wireless quality classification and evaluates an imbalanced dataset. The authors used the Rutgers [17] dataset to evaluate the effectiveness of several machine learning and deep learning models, including logistic regression, SVM, decision tree, random forest, and multilayer perception. The highest results have a 94 percent to 95 percent accuracy rate. Data preparation strategies can address issues such as noisy data and data imbalance.

The works in [18] proposed a lightweight, weighted Euclidean distance-based multi-parameter fusion link quality estimator. The Fused parameter combines Signal-to-Noise Ratio (SNR), Link Quality Indicator (LQI), and weighted Euclidean distance. The link quality estimator is constructed by logistic regression that estimates the mapping relation between the fused parameter and packet reception ratio. The paper in [19] proposes an effective link quality estimator method, namely RNN-LQI, which adopts a Recurrent Neural Network (RNN) to predict the Link Quality Indicator (LQI) series. RNN-LQI estimates the link quality according to the fitting model of LQI and PRR. The method is proper with low-power wireless links with more fluctuations. IoT nodes are often deployed in harsh environments and subjected to environmental noise during communication, lowering network quality. A small proportion of the imbalance is because of excellent and low connection quality samples. As a result, the sample must be processed before training. The preprocessing step in our suggested system was employed to eliminate sample imbalance and noise. Therefore, this paper proposed a link quality estimation model based on the Gradient Boosting Decision Tree (GBDT) algorithm to estimate the link quality.

III. Link Quality Estimation

A. Problem formation

Let $\{X \in \mathbb{R}^{NxM}, y \in \mathbb{R}^{N}\}$ be the wireless link quality data, where N and M denote the number of collected data and physical layer parameters (e.g., RSSI, SNR, PRR), respectively. The array $x_i = X[i,:]$ with $i \in (1, n)$ represents a connection stage at time i with the labeled link quality grade $y_n = y[n]$. In this work, we use four grades to evaluate the link denoted by values from one to four. Then, our goal is to build a link quality estimation:

$$f(x) : \mathbb{R}^{M} \rightarrow \{1, 2, 3, 4\} \tag{1}$$

which is lightweight enough to deploy to IoT devices. The overall link quality estimation is illustrated in Figure 1.

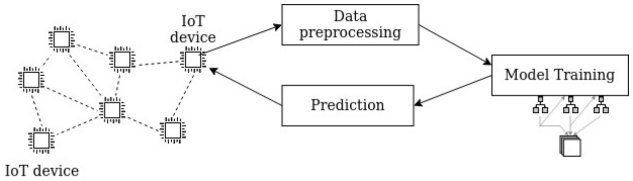

Fig. 1. System workflow

B. Link Quality Estimation Model

Our proposed approach is based on a Gradient Boosting Decision Tree (GBDT), which receives RSSI and network features as input parameters and output the quality of the wireless link. The main idea to build an additive classification model by successively fitting a weak classifier to current residuals. In other words, the model is combined from all weak classifiers to create an ensemble series classifier. As a result, freshly learned weak classifiers could adaptively repair earlier weak classifiers' errors, significantly increasing prediction performance. We also employs Gradient Boosting technique that uses decision trees as weak classifiers and a loss function to discover residuals. A decision tree, comprising a collection of nodes and edge arrangements in a hierarchical structure, may break down complicated issues into a more understandable hierarchy. In detail, given the network information at time i denoted by x_i, the estimated link quality $\hat{y}_i$ with N estimators is:

$$\hat{y}_i = E_N(x_i) = \sum_{n=1}^{N} f_n(x_1) \tag{2}$$

where the E_N and f_n are the assembled and single decision tree, respectively. In other words, the assemble E_N is built up from cumulative tree f_n. Let denote E_{n-1} is the previous stage of E_n:

$$E_n(x_i) = E_{n-1}(x_i) + f_n(x_i) \tag{3}$$

Where $f_n(x_i)$ is the new tree added to reduce the loss $l_n(y_i, \hat{y}_i)$. Combining Eq. (10) and Eq. (3), $f_n(x_i)$ could be redefined as:

$$f_n = \arg\min_{f} \sum_{i=1}^{n} l(y_i, E_{n-1}(x_i) + f(x_i)) \tag{4}$$

Applying the first-order Taylor approximation method to approximately calculate the loss function l

$$l(y_i, z) \approx l(y_i, a) + (z - a)\frac{\partial l(y_i, a)}{\partial a} \tag{5}$$

with z and an equal $E_{n-1}(x_i) + f(x_i)$ and $E_{n-1}(x_i)$, respectively. Note that $\frac{\partial l(y_i,a)}{\partial a}$ is the derivative of the loss function (negative gradient) denoted by d_i. The single decision tree f_n mentioned in Eq. (4) can be approximately defined as:

$$f_n \approx \arg\min_{f} \sum_{i=1}^{n} l(y_i, E_{n-1}(x_i)) + f(x_i)d_i \tag{6}$$

After removing constant equations, the short form of f_n is:

$$f_n \approx \arg\min_{f} \sum_{i=1}^{n} f(x_i)d_i \tag{7}$$

For each iteration in the training phase, the network information x is fitted to minimize value of f_n by updating the negative gradients d_i. The overall process is illustrated in Figure 2.

C. Selection link quality parameters

To accurately determine the quality of network links, we have to analyze and understand the influence of observed physical layer parameters on them. In fact, the physical layer parameters (e.g., RSSI, LQI, and SNR) may quickly identify connection changes, but they are significantly affected by several environmental factors, such as attenuation, multipath type of channel distortions, and background noise's temporal. In addition, public trace sets limit the valuable parameters for LQE. For example, the Rutgers trace sets only publish raw RSSI values and sequence numbers. Therefore, eliminating redundant features and generating synthetic features are vital to improving estimation quality.

Synthetic feature generation: To increase the number of valuable features for model training, we analytically examine the impact of the synthetic features generated from raw RSSI values model performance. In detail, we create a feature matrix consisting of degree-2 polynomial features generated from raw RSSI values and their interaction, such as the average ($\overline{RSSI}$), standard deviation (σ_{RSSI}), first order derivative of RSSI ($f'(RSSI)$) over k packets. The degree-2 polynomial features (denoted by θ_{RSSI}) of raw RSSI and its $\overline{RSSI}$ are defined as

$$\theta_{RSSI} = [1, \overline{RSSI}, RSSI, \overline{RSSI}^2, \overline{RSSI} * RSSI, RSSI^2] \tag{8}$$

Oversampling minority class: Most public trace sets for LQE are unbalanced and contains high-quality links, ranging from 50% to 80% in total [20]. In such a scenario, the classification model may exhibit a specific deviation, increasing the susceptibility to majority class samples and, consequently, lowering the model performance. To solve this issue, we employ the synthetic minority oversampling technique (SMOTE) [21] to create synthetic samples of minority classes instead of duplication. Let x_n denote a sample (feature vector) belonging to a minority class and x'_n be its nearest neighbor; the synthetic sample s_n is calculated by:

$$s_n = x_n + r_n * (|x_n - x'_n|) \tag{9}$$

with r_n is a random number between 0 and 1.

IV. Evaluation Results

In this section, we present the evaluation results of our proposal in terms of its estimation performance. We first open by describing the evaluated datasets, followed by the evaluation metrics, and finally, presenting our results.

A. Evaluated Datasets

Evaluated dataset overview: The Rutger dataset contains 4,060 link traces extracted from 812 separate connections at five different noise levels of 0, -5, -10, -15, and -20 dB. It includes several attributes, such as raw RSSI, sequence numbers, source node ID, destination node ID, and fake noise levels. After deeply analyzing, we figure out that the valid value of RSSI is from 0 to 127, so the RSSI values higher than 128 are removed.

Division of link quality grade: This proposal classifies links into four categories based on their PRR values: good link, intermediate link, bad link, and very bad link. We classify links with PRR values between 90% and 100% as good links, values between 50% and 90% as intermediate links, values between

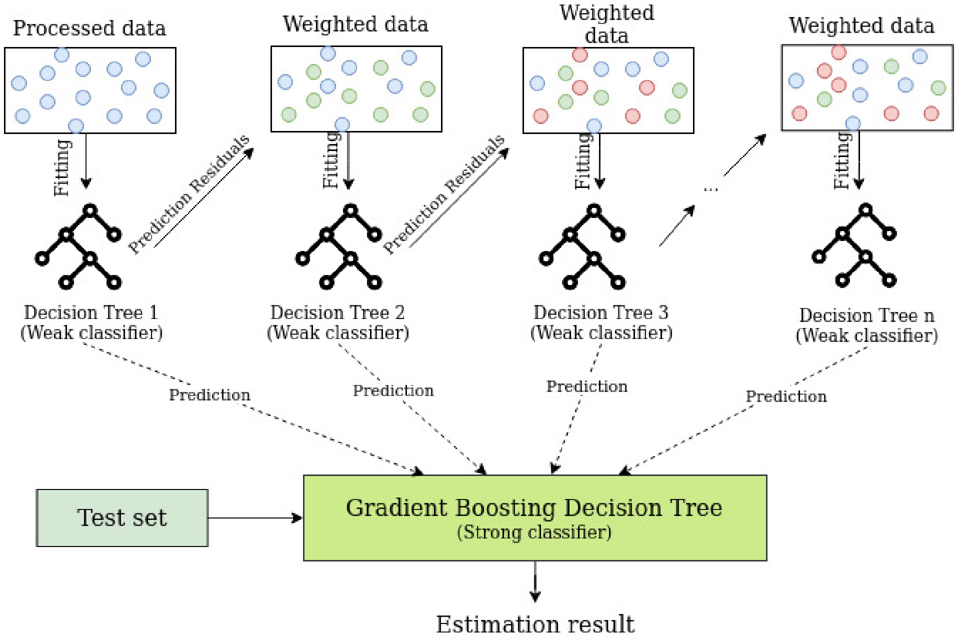

Fig. 2. Link quality estimation model based on gradient boosting decision tree

10% and 50% as bad links, and values between 0% and 10% as very bad links. The PRR value clearly distinguishes the relationship. The link quality grade as the link quality estimate metric is defined in Table I.

TABLE I
DEFINITION OF LINK QUALITY GRADE

Link Quality Grade(LQG)	Description	The range of PRR
1	Good link	90%$\leq PRR \leq$ 100%
2	Intermediate link	50% <PRR <90%
3	Bad link	10% <PRR $\leq$ 50%
4	Very bad link	0% $\leq PRR \leq$ 10%

B. Evaluation metrics

The accuracy metric is widely-used to evaluate the performance of link quality estimators. However, it is insufficient for analyzing unbalanced trace-sets, which are mostly constituted of high-quality links. If the model correctly classifies all samples as positive, the model's accuracy is 90%, a high performance, but this is inappropriate for negative samples. Therefore, in our work, we employ several metrics, such as the accuracy, precision, recall, F1-score, and confusion matrix. The details of these metrics are summarized below: Let TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively. The confusion matrix is used to assess classifier precision and recall.

$$Precision = \frac{TP}{TP + FP}$$

where TP is the number of true positives and FP the number of false positives.

$$Recall = \frac{TP}{TP + FN}$$

where TP is the number of true positives and FN the number of false negatives.

$$F1 = \frac{2 * (precision * recall)}{precision + recall} \tag{10}$$

C. Results

Our proposal is implemented using Scikit-learn and Keras with TensorFlow libraries and intensively evaluated on the Rutger dataset to demonstrate its effectiveness.

Estimation quality: Figure 3 and 4 show the detailed estimation performance and confusion matrix of our proposal on the Rutger dataset, respectively. We note that the estimated accuracy of high and low link qualities (good and very bad labels) is reported about 99%, meaning that LLEQ accurately selects effective network links and ignores low-quality ones. With a deeper investigation of these results, the area under the receiver operating characteristics (AUC-ROC curve) illustrated in Figure 5 proves again the effectiveness of our proposal with AUC values of all classes higher than 99%.

Resource consumption and baseline comparison: To demonstrate the lightweight of our proposal, we monitor resource consumption (memory and CPU footprints) when training and running LLQE on a Raspberry Pi 4 Model B (Cortex-A72, 4GB Ram) for 8 minutes and report the results in Figure 6. The training phase starts from the beginning to time index

TABLE II
THE COMPARISION OF ML ALGORITHMS ON THE RUTGER DATASET.

	Accuracy	**Levels of Quality**
Decision Trees [16], 2020	95.2%	3
Random Forest [16], 2020	95.3 %	3
Multilayer perception [16], 2020	95.1 %	3
Logistic Regression [22], 2021	96.8 %	3
LLQE	97%	4

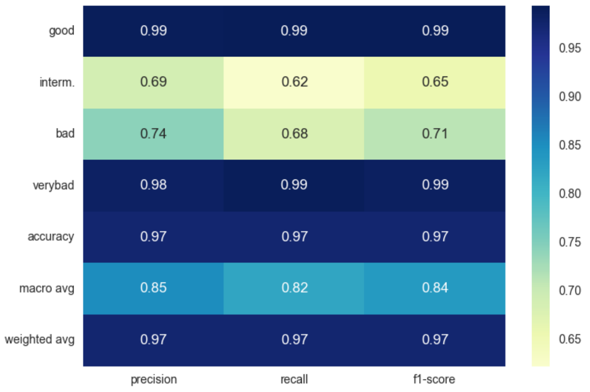

Fig. 3. The detailed link quality estimation results of our proposal.

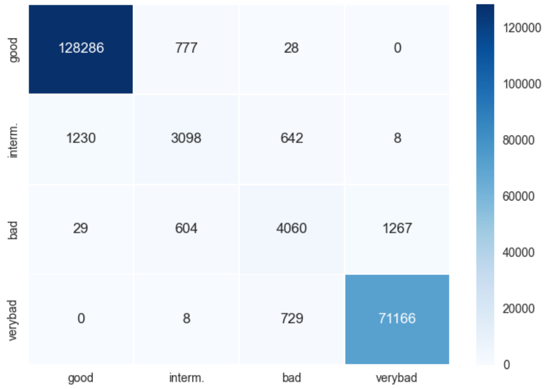

Fig. 4. The confusion matrix of our proposal over the Rutger dataset.

number 23, and the running phase is the remaining time. We can see from the results that LLQE only demands about 0.75 GB of memory for both training and running phases. Its computational footprint is about 53% and 25% for training and running phases, respectively. We compare LLQE with its competitors and report the results in Table II. It is interesting to see that LLQE not only estimates more level of quality than its competitors but also achieves higher accuracy.

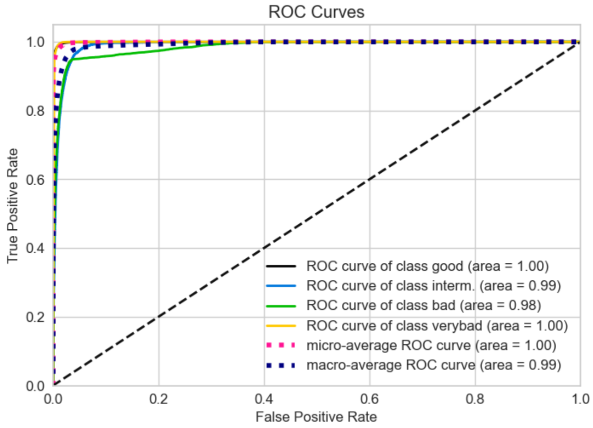

Fig. 5. The ROC-AUC curve of each link quality evaluated on the Rutger dataset.

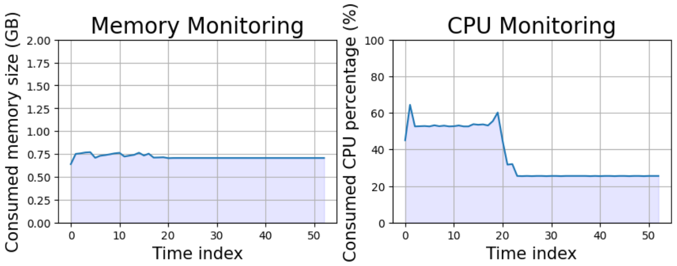

Fig. 6. The resource consumption of Raspberry Pi 4 when training and running LLQE.

V. CONCLUSION

In this paper, we propose a link quality estimation method named LLQE based on the gradient boosting decision tree algorithm for IoT devices. We first analyze the wireless network metric (e.g., RSSI, SNR, PRR) and its relationship with link quality indicators. Then, we introduce a method of data preprocessing during the link quality evaluation process. The proposed approach uses decision trees as weak classifiers and a loss function to discover residuals. We also employ the synthetic minority oversampling technique to create synthetic samples of the minority classes instead of duplication. Our proposed approach is implemented using Scikit-learn and Keras with TensorFlow libraries and is intensively evaluated on the Rutger dataset to demonstrate its effectiveness. The evaluation result reveals that LLQE accurately estimates four levels of link quality. In future work, we will focus on optimizing the energy consumption of LLQE-enabled IoT devices.

Acknowledgment

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D1-2022-26.

References

[1] M. Ayaz, M. Ammad-Uddin, Z. Sharif, A. Mansour, and E.-H. M. Aggoune, “Internet-of-things (iot)-based smart agriculture: Toward making the fields talk,” *IEEE access*, vol. 7, pp. 129551–129583, 2019.

[2] G. Cerar, H. Yetgin, M. Mohorčič, and C. Fortuna, “Machine learning for wireless link quality estimation: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 696–728, 2021.

[3] Y. A. Qadri, A. Nauman, Y. B. Zikria, A. V. Vasilakos, and S. W. Kim, “The future of healthcare internet of things: a survey of emerging technologies,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1121–1167, 2020.

[4] L. Kim-Hung and Q. Le-Trung, “User-driven adaptive sampling for massive internet of things,” *IEEE Access*, vol. 8, pp. 135798–135810, 2020.

[5] R. D. Gomes, D. V. Queiroz, A. C. Lima Filho, I. E. Fonseca, and M. S. Alencar, “Real-time link quality estimation for industrial wireless sensor networks using dedicated nodes,” *Ad Hoc Networks*, vol. 59, pp. 116–133, 2017.

[6] M. Deb, S. Roy, B. Saha, P. Das, and M. Das, “Designing a new link quality estimator for sensor nodes by combining available estimators,” in *2017 IEEE 7th International Advance Computing Conference (IACC)*, pp. 179–183, IEEE, 2017.

[7] S. Rekik, N. Baccour, M. Jmaiel, and K. Drira, “Holistic link quality estimation-based routing metric for rpl networks in smart grids,” in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–6, IEEE, 2016.

[8] W. Sun, W. Lu, Q. Li, L. Chen, D. Mu, and X. Yuan, “Wnn-lqe: Wavelet-neural-network-based link quality estimation for smart grid wsns,” *IEEE Access*, vol. 5, pp. 12788–12797, 2017.

[9] W. Sun, X. Yuan, J. Wang, Q. Li, L. Chen, and D. Mu, “End-to-end data delivery reliability model for estimating and optimizing the link quality of industrial wsns,” *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 3, pp. 1127–1137, 2017.

[10] K.-H. L. Minh and K.-H. Le, “Odlie: On-demand deep learning framework for edge intelligence in industrial internet of things,” in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 458–463, 2021.

[11] K.-H. Le, K.-H. Le-Minh, and H.-T. Thai, “Brainyedge: An ai-enabled framework for iot edge computing,” *ICT Express*, 2021.

[12] W. Sun, W. Lu, Q. Li, L. Chen, D. Mu, and X. Yuan, “Wnn-lqe: Wavelet-neural-network-based link quality estimation for smart grid wsns,” *IEEE Access*, vol. 5, pp. 12788–12797, 2017.

[13] H.-J. Audéoud and M. Heusse, “Quick and efficient link quality estimation in wireless sensors networks,” in *2018 14th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*, pp. 87–90, 2018.

[14] M. L. Bote-Lorenzo, E. Gómez-Sánchez, C. Mediavilla-Pastor, and J. I. Asensio-Pérez, “Online machine learning algorithms to predict link quality in community wireless mesh networks,” *Computer Networks*, vol. 132, pp. 68–80, 2018.

[15] J. Shu, S. Liu, L. Liu, L. Zhan, and G. Hu, “Research on link quality estimation mechanism for wireless sensor networks based on support vector machine,” *Chinese Journal of Electronics*, vol. 26, no. 2, pp. 377–384, 2017.

[16] G. Cerar, H. Yetgin, M. Mohorčič, and C. Fortuna, “On designing a machine learning based wireless link quality classifier,” in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–7, Ieee, 2020.

[17] S. K. Kaul, I. Seskar, and M. Gruteser, “CRAWDAD dataset rutgers/noise (v. 2007-04-20).” Downloaded from https://crawdad.org/rutgers/noise/20070420/RSSI, Apr. 2007. traceset: RSSI.

[18] W. Liu, Y. Xia, S. Hu, and R. Luo, “Lightweight multi-parameter fusion link quality estimation based on weighted euclidean distance,” in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, IEEE, 2019.

[19] M. Xu, W. Liu, J. Xu, Y. Xia, J. Mao, C. Xu, S. Hu, and D. Huang, “Recurrent neural network based link quality prediction for fluctuating low power wireless links,” *Sensors*, vol. 22, no. 3, p. 1212, 2022.

[20] R. Jacob, R. D. Forno, R. Trüb, A. Biri, and L. Thiele, “Dataset: Wireless link quality estimation on flocklab-and beyond,” in *Proceedings of the 2nd Workshop on Data Acquisition To Analysis*, pp. 57–60, 2019.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[22] G. Cerar, H. Yetgin, M. Mohorčič, and C. Fortuna, “Learning to fairly classify the quality of wireless links,” in *2021 16th Annual Conference on Wireless On-demand Network Systems and Services Conference (WONS)*, pp. 1–8, Ieee, 2021.