IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 March 2023, accepted 21 March 2023, date of publication 29 March 2023, date of current version 7 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3263053

RESEARCH ARTICLE

Carrier Frequency Offset Estimation in 5G NR: Introducing Gradient Boosting Machines

MOSTAFA HUSSIEN[1,2], AHMED ABDELMOATY[1,3], (Member, IEEE), MAHMOUD ELSAADANY[1], (Senior Member, IEEE), MOHAMMED F. A. AHMED[2,4], GHYSLAIN GAGNON[1], (Senior Member, IEEE), KIM KHOA NGUYEN[1], (Member, IEEE), AND MOHAMED CHERIET[1], (Senior Member, IEEE)

[1]École de Technologie Supérieure (ÉTS), Univeristy of Québec, Montréal, QC H3C 1K3, Canada
[2]Department of Information Technology, Assiut University, Assiut 71515, Egypt
[3]Resilient Machine-Learning Institute (ReMI) of École de Technologie Supérieure (ÉTS), Univeristy of Québec, Montréal, QC H3C 1K3, Canada
[4]Canadian National Railway (CN), Montréal, QC H3B 2M9, Canada

Corresponding author: Mostafa Hussien (m.korashy@ieee.org)

ABSTRACT The Beyond fifth Generation (B5G) communication systems imposed several challenges on radio designers. For example, a machine is required to set up a call at a low Signal-to-Noise Ratio (SNR), as low as −10 dB, in the extended coverage mode. Moreover, only one receive antenna will be available, and virtually no frequency diversity. Such requirements present major challenges to maintaining timing and frequency synchronization. Carrier Frequency Offset (CFO) estimation is at the heart of these challenges. Different approaches have been proposed for CFO estimation such as maximum likelihood based on a cyclic prefix. Nevertheless, these methods remain limited in various ways. At the same time, Machine Learning (ML) techniques showed outstanding performance in several wireless communication problems. In this work, we propose an ML-based approach for CFO estimation in OFDM systems. Specifically, we propose a Gradient-Boosting Machine (GBM)-based solution to predict the CFO given the received Primary Synchronization Signal (PSS) and Secondary Synchronization Signal (SSS). Furthermore, we make our dataset available for public access to encourage other researchers to pursue this promising direction. We compare our results with different baseline models (i.e., artificial neural networks and support vector machines). The experimental results show that our model outperforms other baseline models due to its ensemble nature which enables ensemble models to obtain a better generalization behavior.

INDEX TERMS 5G, carrier frequency offset, gradient-boosting, machine learning, new radio.

I. INTRODUCTION

Massive integration of connected devices with emerging services provisions a successive increase in traffic demand and higher data rates. It is expected that data rates to be exploded by deploying 5G New Radio (NR). Globally, mobile data traffic is projected to margin 226 Exabytes (EB) per month in 2026. To cope with these requirements and to provide better user experience and services, it is anticipated for future communication networks (e.g., 6G) to integrate multiple advanced technologies, such as edge computing and Machine Learning (ML). This integration is stimulated by the increase in traffic and data requirements. Indeed, it is anticipated that the connected devices will margin 80 million by 2025 [1].

Orthogonal Frequency Division Multiplexing (OFDM) is one of the adopted technologies in 4G Long-Term Evolution (LTE), and it is expected to continue supporting the 5G NR. OFDM is proven to have the ability to work in harsh fading environments due to multipath. Furthermore, relying on 5G NR and IEEE 802.11ax is provisioned to exploit Quadrature-Amplitude Modulation (QAM) with higher orders of up to 1024. Additionally, different modulation schemes with higher orders (e.g., 64-QAM and 256-APSK) are expected to be supported by Millimeter-wave (mm-Wave) technologies and satellite TV standards, respectively.

Higher-order modulation schemes are susceptible to phase errors due to their highly dense constellation mapping.

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Huo.

TABLE 1. A list of the abbreviations (alphabetically sorted).

Abbreviation	Defintion
ANN	Artificial Neural Networks
APSK	Amplitude Phase Shift Keying
AWGN	Additive White Gaussian Noise
BEM	Basis Expansion Modelling
CFO	Carrier Frequency Offset
CIR	Channel Impulse Response
CP	Cyclic Prefix
CNN	Convolutional Neural Networks
CRC	Cyclic Redundancy Block
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
GBM	Gradient Boosting Machine
ICI	Inter-carrier Interference
JMLE	Joint Maximum Likelihood Estimation
LLR	Log Likelihood Ratio
NR	New Radio
OFDM	Orthogonal Frequency Division Multiplexing
PCA	Principal Component Analysis
PSS	Primary Synchronization Signal
PBCH	Physical Broadcast Channel
QAM	Quadrature Amplitude Modulation
RNN	Recurrent Neural Networks
RB	Resource Block
SSS	Secondary Synchronization Signal
SNR	Signal to Noise Ratio
SCS	SubCarrier Spacing
SVM	Support Vector Machines
SSB	Signal Synchronization Block
UE	User Equipment

These errors arise from residual Carrier Frequency Offset (CFO) that results from an intrinsic mismatch between the transmitter and the receiver oscillators. Interestingly, it is linearly increasing during frame reception and eventually translated to a large phase offset that causes a considerable declination in spectral efficiency and increases Bit-Error-Rate (BER). CFO resulting from Doppler shift may reach up to 2 KHz at 4.2 GHz band, this is equivalent to 13% of the sub-carrier spacing [2].

The CFO destroys the orthogonality between subcarriers and induces Inter-Carrier Interference (ICI). Consequently, there are degradations in the OFDM system performance. Therefore, estimating the CFO is crucial for future communication networks. In general, estimating the CFO can be classified into two categories: data-driven estimation and blind estimation. Blind estimation of the CFO can be performed with algorithms such as the Cyclic Prefix (CP)-based maximum likelihood. On the other hand, Zadoff-Chu (ZC)-based cross-correlation or auto-correlation algorithm is an example of the data-driven CFO estimation [3].

Recently, state-of-the-art ML algorithms go all the way from data mining techniques, and resource allocation problems to tackle most of the issues in cellular networks. Interestingly, there is a significant trend for implementing powerful ML-based solutions for many complex problems in wireless communications such as link adaptation [4], resource allocation [5], Channel State Information (CSI) compression [6], beamforming [7], among others. ML algorithms can be categorized into three main categories namely supervised, unsupervised, and Reinforcement Learning (RL). Basically, supervised learning requires labeled data in order to train the system. Hence, the system learns from these labels to predict the target output. Conversely, unsupervised learning does not have the luxury of accessing labeled data. The expected output is not known priorly and the system needs to learn in a blind fashion. Finally, in the RL regime, an agent learns the best actions by itself, but with enforced guidance by a reward mechanism. The actions in RL are made by the agent toward the environment, which, in turn, replies by changing its state and sending back a reward value to the agent depending on how good is the agent's actions.

A. RELATED WORK

The importance of CFO synchronization attracted the attention of many researchers. Some studies investigated the Primary Synchronization Signal (PSS) in the time domain [8], while other works studied the frequency domain. The authors in [8] proposed a method to detect the PSS by finding the maximum cross-correlation. However, the Integer Carrier Frequency Offset (ICFO) degrades the accuracy of PSS detection. Hence, differential correlation-based PSS detection schemes have been proposed using Discrete Fourier Transform (DFT), or Fast Fourier Transforms (FFT) [9], [10].

To improve the overall performance of the OFDM systems, joint optimization of CFO with other parameters was proposed in [11] and [12]. In [11], a framework of pilot-aided joint estimation of CFO along with the Channel Impulse Response (CIR) for linear periodic channels was studied. A Joint Maximum Likelihood Estimator (JMLE) that guarantees higher spectral efficiency and lower computational complexity was proposed. The estimator exploits the periodicity and the sparsity of the channel to improve the estimation performance. The timing and channel estimation were accompanied by CFO in a joint approach for OFDM with high mobility systems in [12]. To overcome the problem of high complexity in joint estimation, a computationally efficient algorithm using Basis Expansion Modeling (BEM) was introduced. Basically, BEM tracks channel variations to reduce the number of unknown channel parameters. The proposed algorithm is proven to outperform other benchmark algorithms.

In the last few years, the challenges of traditional methods trigger researchers to deploy state-of-the-art data-driven ML techniques for CFO estimation [13], [14], [15]. The authors of [13] compared three different architectures of Artificial, Convolutional, and Recurrent Neural Networks

(ANNs, CNNs, RNNs) to three conventional techniques of CFO estimation namely: original periodogram, Welch's periodogram, and multiple signal classification estimator. They concluded that the ML-based estimators outperform the conventional techniques by 15 dB with one-bit resolution. In [14], RNNs have been used to jointly estimate CFO with packet detection in IEEE 802.11 systems. The performance of RNNs has been proven to surpass the conventional techniques in low-to-medium Signal-to-Noise-Ratios (SNRs). However, with high SNRs, the performance is degraded. Instead of estimating the residual CFO, the authors in [15] proposed an algorithm to estimate the variance of the CFO from SNR to optimize the Log-Likelihood Ratio (LLR). Large throughput gains have been proven even with the cases of high SNRs and high residual CFO. Multiple practical impairments at the receivers such as CFO, timing synchronization, and channel estimation in Multiple-Inputs Multiple-Outputs (MIMO)-OFDM systems have been covered in [16]. The authors proposed a Deep Convolutional Neural Network (DCNN) scheme to learn the adaptive coding and modulation given the CSI. The BER and the throughput of the proposed scheme outperform other ML techniques such as Support Vector Machine (SVM) and K-Nearest Neighbors (KNN).

The aforementioned works suffer from either higher complexity compared to the conventional techniques, or performance degradation in higher SNRs. To fill this gap, in this work, we propose a novel ML-based approach leveraging the power of GBM. Recently, GBM has outperformed various ML methods and acted as a seminal high-efficient approach in various challenging tasks. GBMs and deep learning are both powerful techniques for solving regression problems, but there are situations where GBM may be preferred over deep learning for various reasons and facts such as:

- **Data Efficiency**: GBMs require fewer data to train and can often achieve comparable performance with fewer samples. This is because GBMs can learn complex non-linear relationships between input features and the output variable using a small number of trees, whereas deep learning models typically require a large number of parameters to learn complex representations of the input data. In the CFO estimation problem, this is considered a main advantage compared with deep learning, especially, because it is expected to trigger several retrainings after the model deployment due to concept drift in the CFO estimation problem.
- **Robustness**: GBMs are less sensitive to outliers in the data than deep learning models. This is because GBMs use an ensemble of trees, which are less affected by individual data points than deep neural networks. This gives the GMB an important advantage in the context of the CFO estimation problem due to the impulse noise that is common in a wireless communication system.
- **Faster Training**: GBMs are typically faster to train than deep learning models, especially for small to medium-sized datasets. This is because GBMs are based on decision trees, which are simple and fast to train, whereas deep learning models require a large amount of computational resources and may take days or weeks to train on large datasets. This characteristic can be an advantage when we retrain the model after detecting a performance degradation.
- **Explainability**: GBMs provide an interpretable model, which can be important to understand the reasoning behind the model's decisions. In contrast, deep learning models are often considered to be "black boxes" that are difficult to interpret.
- **Tunability**: GBMs have several hyperparameters that can be tuned to optimize the model's performance, such as the number of trees, learning rate, and the maximum depth of trees. This provides greater control over the model's performance than deep learning models, which have a huge number of hyperparameters that can be difficult to tune.

B. CONTRIBUTION

In this work, we provide the first attempt to leverage the power of GBM for CFO estimation. We use the real and imaginary parts of the PSS and SSS to directly regress the CFO values, which eliminates the need for complicated feature engineering.

The contribution of this work can be summarized as follows:

- We propose the first work that leverages the power of ML techniques in the problem of CFO estimation. We proposed a novel GBM-based solution for CFO estimation in B5G communications.
- We present the first open well-annotated dataset for the problem of CFO estimation. This dataset encourages further work on the problem of CFO estimation and facilitates the comparison between different models and solutions.
- We highlight and discuss several promising research directions in this problem, exploiting the existence of the public dataset.

The rest of this article is organized as follows: Section II presents the system model and NR signal synchronization procedure. Section III presents our proposed GBM-based algorithm for CFO estimation. The evaluation scenario and results are detailed in section IV. Section V lists some promising research directions for future investigation and Section VI concludes our work. The mentioned abbreviations are listed in Table. 1 with an alphabetical order.

II. SYSTEM MODEL AND NR SIGNAL SYNCHRONIZATION

A. SYSTEM MODEL

For a typical NR OFDM system, the received signal $r(n)$ can be represented by:

$$r(n) = y(n) + w(n) = \Big[x(n) * h(n)\Big]e^{-j2\pi\epsilon n} + w(n), \quad (1)$$

where:

$x(n)$ is the transmitted signal.

$h(n)$ is the impulse response of the channel.

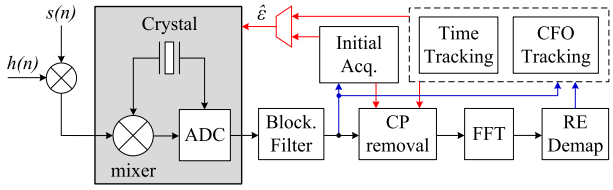

FIGURE 1. CFO tracking in NR UE receiver.

$w(n)$ is the Additive White Gaussian Noise (AWGN).
ϵ is the normalized CFO.
$*$ represents the convolution operator.

Fig. 1 illustrates the block diagram of the processing chain of the NR User Equipment (UE) receiver. We are assuming an OFDM system where a unified crystal is exploited to lock the carrier frequency to the sampling clock through the RF processing. Then, the initial acquisition phase starts by estimating the frequency/timing offsets experienced by both the RF crystal and the channel. However, a residual CFO caused by temperature changes and Doppler effects always exists and needs to be estimated. Hence, it is normal to consider the tracked CFO a fractional part of the Sub-Carrier Spacing (SCS). Lastly, the received symbols are transferred to the Frequency Domain (FD) using FFT as given by:

$$R_l^{(\delta)}(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} r_l^{(\delta)}(n)\, e^{-j2\pi kn/N}, \quad 0 \le k < N, \quad (2)$$

where:
$r_l^{(\delta)}(n)$ is the l^{th} OFDM symbol after removing the CP. The length of the OFDM symbol is denoted by N, and δ represents the drift samples.

B. NR SIGNAL SYNCHRONIZATION

Based on the 3GPP specifications release 15 [17], the NR system is defined by multiple SCS and a CP overhead. CP can be either normal or extended. The basic SCS 15 KHz is used as a base for obtaining any other SCS by the scale of 2^{μ}, where $\mu \in \{0, 1, 2\}$. The frame structure in the time domain consists of 10 subframes with a fixed duration of 1 ms. Regardless of the CP overhead, each SCS is aligning on symbol boundaries in every subframe. The period of each time slot is equal to $1/2^{\mu}$. For each slot, there are 14 and 12 OFDM symbols for normal and extended CP, respectively. While in the frequency domain, similarly to LTE; a resource block (RB) is defined by 12 consecutive subcarriers. An RB grid in the NR system is shown in Fig. 2.

The procedure of signal synchronization includes cell search, frame boundaries detection, and signal quality measurements. In NR, downlink synchronization signals are classified into two types:

a) Primary Synchronization Signal **(PSS)**:
PSS sequences can be denoted by $P_{\mu}[m]$ and composed of 127 samples of n-sequences that given by [17]:

$$\begin{aligned} P_{\mu}(m) &= 1 - 2n(q), \\ q &= (m + 43\mu) \mod 127, \\ 0 &\le m \le 127, \end{aligned} \quad (3)$$

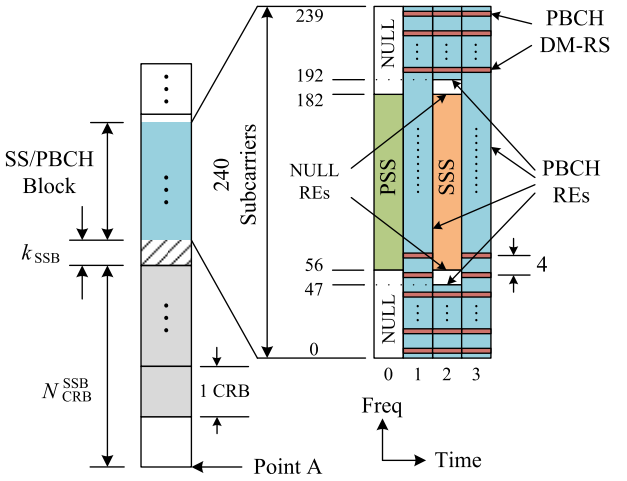

FIGURE 2. Frame structure of SS block [18].

where $\mu \in \{0, 1, 2\}$ stands for cell sector ID and

$$n(m + 7) = (n(m + 4) + n(m)) \mod 2, \quad (4)$$

where the first 7 samples of n(m) can be given by $\{0, 1, 1, 0, 1, 1, 1\}$. In the frequency domain, PSS channel consists of 240 subcarriers, and using Eq. (3) it can be given as:

$$D_{\mu}[f] = \begin{cases} P_{\mu}[f - 56] & 56 \le f \le 126, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

PSS is located in the first OFDM symbol of the synchronization block and occupies subcarriers with indexes from 57 to 183.

b) Secondary Synchronization Signal **(SSS)**:
SSS is a result of the combination of two n-sequences with a duration of 127 samples. SSS is generated depending on group ID $\in$ $[0, 355]$. SSS occupies subcarriers with the same indices as PSS but it is located in the third OFDM symbol of the Signal Synchronization Block (SSB).

Additionally, another type of signal is mapped to the SSB, namely the Physical Broadcast Channel (PBCH). 56 information bits representing four information fields are transmitted by the PBCH in each SSB. The first 24 bits are used for cell configuration, while the last 24 bits are reserved for Cyclic Redundancy Check (CRC). The remaining bits are used by the UE to detect the radio frame's beginning; accordingly, it starts the procedure of synchronization.

III. PROPOSED MODEL

In this section, we introduce the GBM-based framework in subsection III-A. The description of the dataset is presented in subsection III-B.

A. GRADIENT BOOSTING MACHINES (GBM)

GBM is a widely-adopted efficient classification and regression model. Given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ that consists of N observation-label pairs (x_i, y_i). GBM

iteratively constructs M weak learners (usually decision trees), $h(x, a_1), h(x, a_2), \ldots, h(x, a_M)$. Assume the labels are generated from an underlying function such that $y_i = f(x_i)$ is the true function to be approximated. A GBM model approximates $f(x)$ by a prediction function, $\hat{y} = g(x)$. The prediction function could be expressed as an additive expansion of a basis function $h(x, a_m)$ such that:

$$\begin{cases} \hat{y} = \sum_{m=1}^{M} \beta_m h(x; a_m), \\ h(x; a_m) = \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}), \end{cases} \tag{6}$$

where $I = 1$ if $x \in R_{jm}$ and zero otherwise. The input space is divided into J different non-overlapping regions $R_{1m}, R_{2m}, \ldots, R_{Jm}$. Each decision tree predicts a constant-value γ_{jm} for a region R_{jm}. In a certain decision tree, the mean values of each splitting variable are given by the parameter a_m. The hyperparameter β_m controls the contribution of each node to the final prediction [19]. The values of these hyperparameters are selected to minimize a specified loss function. Specifically, the mean square error function is a typical choice for regression problems. For a better approximation with small chances of overfitting, a regularization parameter is added to the loss function.

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{m}^{M} \Omega(h_m). \tag{7}$$

The second term is a regularization term that counts for the complexity of the model to prevent overfitting. The regularization term, Ω, can be given by:

$$\Omega(h_m) = \gamma T + \frac{1}{2}\lambda ||\mathbf{w}||^2, \tag{8}$$

where T is the number of leaves and $\mathbf{w}$ is the vector of leaf weights. The hyperparameters γ and λ control the hardness of the regularization, and accordingly, the complexity of the model. It is worth noting that there are different techniques that can be adopted to prevent overfitting during the training phase. Column subsampling and shrinkage are two examples of such techniques [20].

Following the principle of *empirical risk minimization*, we train a GBM model, h_M, to minimize the empirical risk given by:

$$\mathcal{R}(h_M) = \mathbb{E}[\mathcal{L}(y, h_M(x))], \tag{9}$$

where $\mathcal{R}(\cdot)$ is the empirical risk of the input model and $\mathcal{L}(\cdot)$ is the adopted loss function. Given an N-points dataset, we can compute the empirical risk given in (9) by:

$$\mathcal{R}(h_M) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y_i, h_M(x_i)). \tag{10}$$

Since we adopt the additive training approach to train the GBM model, the prediction of the i^{th} data point at a time step, t, is given by:

$$\hat{y}_i^{(t)} = \sum_{i=1}^{M} h(x_i; a_m) = \hat{y}_i^{(t-1)} + h^t(x_i), \tag{11}$$

Algorithm 1 : GMB Training Procedure

Input: A training dataset, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
Initialize a model with a constant value,

$$h_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{N} \mathcal{L}(y_i, \gamma)$$

for m=1 to M **do**

- Compute residuals, $r_{m,i}$ such that for $i = 1, \ldots, N$, compute:

$$r_{m,i} = -[\frac{\delta \mathcal{L}(y_i, h_{(x_i)}))}{\delta h(x_i)}]_{h(x_i)=h_{m-1}(x_i)}.$$

- Train a regression tree with features x and labels r and create terminal regions R_jm for $j = 1, \ldots, J_m$.
- Compute $\gamma_{jm} = \underset{\gamma}{argmin} \sum_{x_i \in R_{jm}} \mathcal{L}(y_i, h_{m-1}(x_i) + \gamma)$, for $j = 1, \ldots, J_m$.
- Update the model:

$$h_m(x) = h_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} 1(x \in R_{jm})$$

end for
Return the trained ensemble model, h_M.

which means that we add the t^{th} learner to minimize the objective in (7) in a greedily fashion. Therefore, the objective in (7) at the t^{th} time step can be expressed as:

$$\mathcal{L}^t = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t-1)} + h^t(x_i)) + \Omega(h^t). \tag{12}$$

The second-order Taylor expansion can be used to get a fast optimization for the objective function.

$$\mathcal{L}^t = \sum_{i=1}^{N} [l(y_i, \hat{y}_i^{(t-1)}) + g_i h^t(x_i) + \frac{1}{2} q_i h^{t^2}(x_i)] + \Omega(h^t), \tag{13}$$

where $g_i = \delta_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $q_i = \delta^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ represent the first and second order gradient statistics on the objective function. To further simplify the objective in (13), we can remove the constants to reduce the objective to the following form:

$$\mathcal{L}^t = \sum_{i=1}^{N} [g_i h^t(x_i) + \frac{1}{2} q_i h^{t^2}(x_i)] + \Omega(h^t). \tag{14}$$

The regularization term can be further expanded as follows:

$$\begin{aligned} \mathcal{L}^t &= \sum_{i=1}^{N} [g_i h^t(x_i) + \frac{1}{2} q_i h^{t^2}(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \\ &= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} q_i + \lambda) w_j^2] + \gamma T, \end{aligned} \tag{15}$$

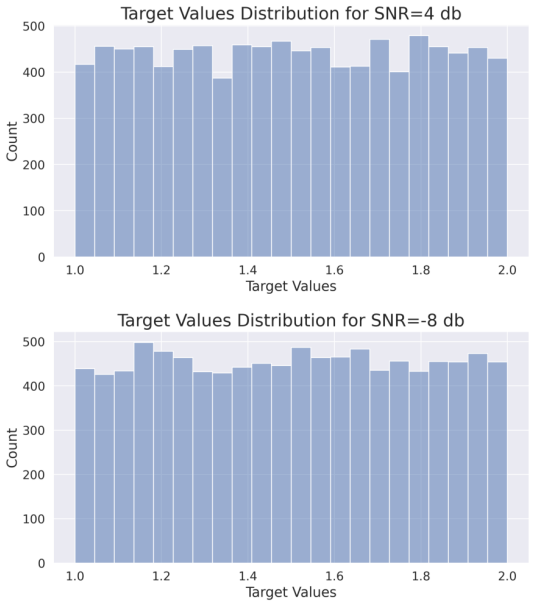

FIGURE 3. The target distribution for SNR=4 and SNR=-8. We can see that the target follows a uniform distribution.

where $I_j = \{i|z(x_i) = j\}$ is the instance set of all leaf nodes. For a certain structure $z(x)$, the optimal value for a leaf j is denoted by w_j^* and given by:

$$w_j^* = -\frac{G_j}{Q_j + \lambda}, \tag{16}$$

where $G_j = \sum_{i \in I_j} g_i$ and $Q_j = \sum_{i \in I_j} q_i$. We can calculate the corresponding optimal objective by:

$$\tilde{\mathcal{L}}^t(z) = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{Q_j + \lambda} + \gamma T, \tag{17}$$

where the equation in (17) can be used to measure the quality of a tree structure z. Algorithm 1 describes the training process of a GBM model.

B. DATASET DESCRIPTION

It is crucial to have a large well-annotated dataset to build any predictive model. Although being at the heart of any communication system, this standardized publicly available dataset is not available for the problem of CFO estimation. This dataset is presented here in order to stimulate further studies in this area. We built a dataset to cover a wide range of SNRs, namely, from SNR= -10 db to SNR=10 db with a step of 2 db. For each SNR $\in \{-10, -8, \ldots, 10\}$, we generated uniformly distributed CFO values. To validate the uniformity of the CFO values in the dataset, Fig. 3 shows the histogram of the CFO values in two different SNR values, namely SNR=4 db and SNR $= -8$ db. We generated a file for each SNR separately. This facilitates designing a model for each SNR value. A larger collection of randomly generated CFOs with different SNR values has been also generated that could be used for training a global model that predicts the CFO for any SNR value. The data has been formatted as Comma-Separated Values (CSV) files. Any data point in each file consists of 509 columns representing the real and imaginary parts of the PSS (127×2), the real and imaginary parts of the SSS (127×2), and the final column represents the target CFO value to be predicted.

TABLE 2. The description of dataset files.

File Name	File Size (MByte)	SNR Value (db)
(1)_SNR=-10	90.4	-10
(2)_SNR=-8	90.8	-8
(3)_SNR=-6	91.2	-6
(4)_SNR=-4	91.6	-4
(5)_SNR=-2	92	-2
(6)_SNR=0	92.3	0
(7)_SNR=2	92.6	2
(8)_SNR=4	90.2	4
(9)_SNR=6	92.9	6
(10)_SNR=8	93	8
(11)_SNR=10	11.5	10

To motivate the community to further explore this interesting research direction, we released the dataset as an open-source at the *GitHub* repository of this work.[1] The dataset consists of 12 CSV files. Each SNR value has one file as well as one file for the aggregated data from different SNRs. Providing the data of each SNR separately helps in building ensemble models with the help of a multiplexer to select the model corresponding to the estimated SNR. Table 2 shows the details of each file.

IV. RESULTS AND DISCUSSION

A. DIMENSIONALITY REDUCTION

The curse of dimensionality is one of the most common problems that raise from dealing with high-dimensional data. Sampling from a high-dimensional space makes the data sparse. Consequently, deriving important conclusions from such a sparse sample becomes more challenging. Unsurprisingly, the curse of dimensionality is presented in the problem under investigation [21]. Different dimensionality reduction techniques can be adopted in this problem such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), autoencoders, etc. Among those techniques, autoencoder is a nonlinear neural network-based technique that achieved outstanding performance in the prior art. However, we employed a PCA-based dimensionality reduction technique in our study for the sake of simplicity and explainability.

PCA is a statistical procedure introduced by *Karl Pearson* in his pioneering paper [22], that uses an orthogonal transformation to convert a group of correlated variables into a group of uncorrelated variables [23]. It has been widely used

[1] https://github.com/Mostafa-Korashy/ML-based-Frequency-Offset-Estimation-in-NR

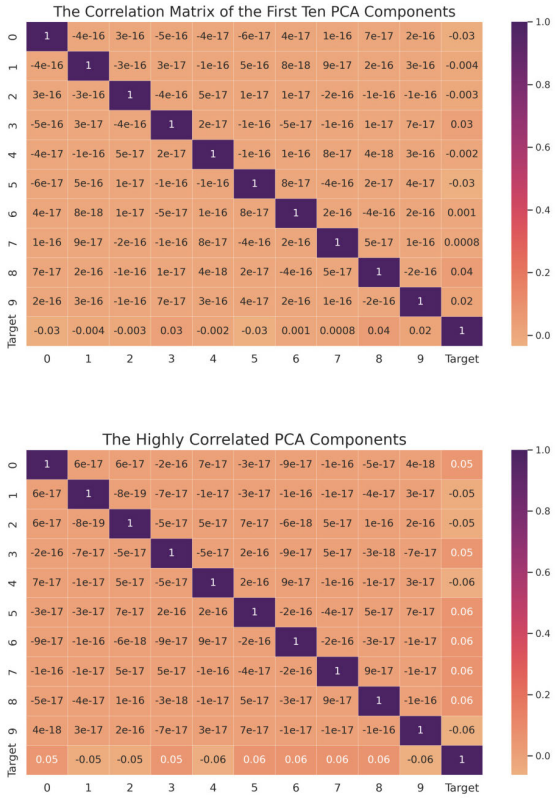

FIGURE 4. The correlation matrix of the first ten PCA components.

for many applications such as visualizing high-dimensional data, and dimensionality reduction for downstream tasks (e.g. regression or classification). Algorithm 2 summarizes the steps of PCA.

Algorithm 2 : Dimensionality Reduction Using PCA for Carrier Frequency Estimation

Input: A training dataset, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$. Each point, X_i is an n-dimensional vector such that $x_i = [x_i^{(1)}, x_i = [x_i^{(2)}, \ldots, x_i = [x_i^{(n)}]$

Standardize the raw data:

$$x_i^{(j)} = \frac{x_i^{(j)} - \bar{x}^{(j)}}{\sigma^j} \forall j$$

Calculate the covariance matrix for the standardized data:

$$\sum = \frac{1}{N} \sum_{1}^{N} (x_i)(x_i)^T,$$

where $\sum \in \mathbb{R}^{n \times n}$.

Compute the eigenvectors and eigenvalues of the covariance matrix, $\sum$.

A typical way of adopting PCA for dimensionality reduction is to use the first k components for the downstream tasks. When we analyzed the correlation of the first k components and the target value, we figured out that some later components maintain higher Pearson factors than the early components. This implies that these later components could be more relevant for the downstream task than the early ones. Therefore, we perform the PCA analysis using the highest possible dimension (i.e., the same dimension as the input). Then, we analyze the correlation between the target and all PCA components using the Pearson matrix. Finally, we consider the PCA components that show the highest correlation with the target label. Fig. 4 shows the correlation matrix for both the first ten PCA components and the ten PCA components with the highest correlation factors. The top figure shows the correlation matrix of the first ten PCA components where we can see the 8^{th} component has a much higher correlation factor than the 2^{nd} component. This motivated us to consider the 15 PCA components with the highest Pearson factors among the 508 components resulting from the PCA analysis.

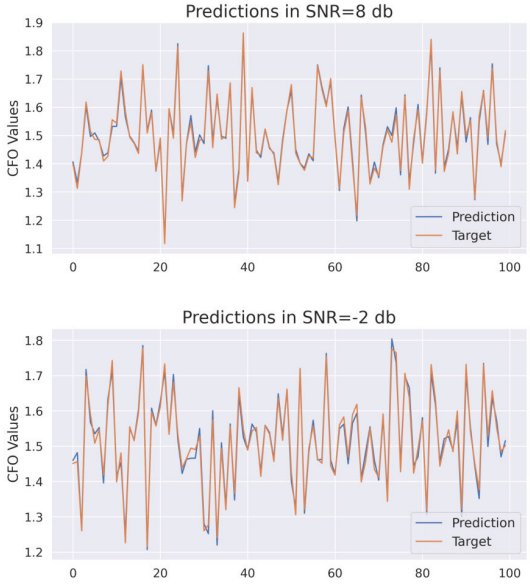

FIGURE 5. The prediction of the CFO using GBM.

B. PREDICTION ACCURACY

We trained a GPM to predict the CFO. For each SNR value, we trained a GPM model using 80% of the data. The remaining 20% has been used for testing. The training set is further divided into train and validation sets. Fig. 5 shows the prediction accuracy for a 100-point sample from the test set. We can see that the GPM model is capable of predicting the target CFO with a considerable level of accuracy. Again, this can be attributed to the power of the ensemble model and the benefit of the adopted boosting technique.

To illustrate the distribution of prediction errors, we plot the histogram of the error bins for different SNR values. For example, we plot the histogram of the prediction error as shown in Fig. 6. We can see that most of the prediction errors lie in the early bins that correspond to the smaller error bins. Note that we normalize our target to be in the range of [2, 3] which means the early error bins have a small percentage compared with the target values.

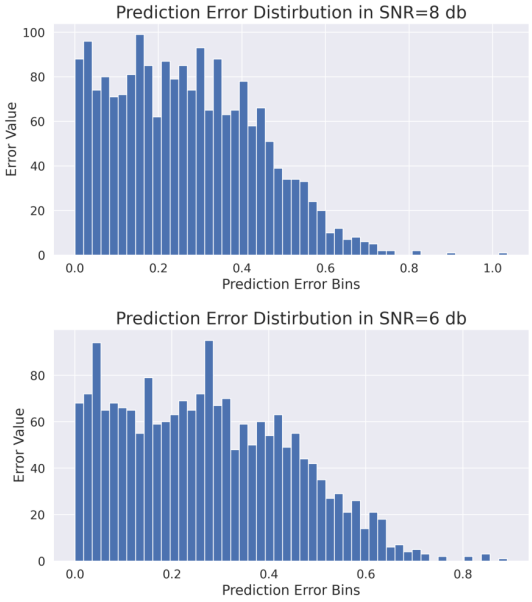

FIGURE 6. The histogram of the prediction error for SNR=8 and SNR=6, as examples.

C. BASELINE COMPARISONS

In this section, we compare our proposed GBM model with two widely adopted models for regression, namely Artificial Neural Networks and support vector machines.

1) ARTIFICIAL NEURAL NETWORK (ANN)

ANN models have shown outstanding performance in many fields and problems such as computer vision, link adaptation, CSI compression, etc [24]. Multilayer Perceptron (MLP) is one of the widely adopted models, especially for processing tabular data such as the problem under consideration. In this work, we used an architecture that consists of one input layer, one hidden layer with *ReLU* activation, and one output layer with *linear* activation. The model is trained to minimize the mean square error (18) using the *Adam* optimizer [25] with a learning rate, $lr = 0.1$. We set the batch size to 128. An $l2$ regularization has been adopted to prevent the high-bias/low-variance behavior (overfitting). The model has been trained for 500 epochs.

$$\mathcal{L} = \frac{1}{N} \left\| y - \hat{y} \right\|_2 . \tag{18}$$

2) SUPPORT VECTOR MACHINES (SVM)

For several decades, SVM dominated most predictive tasks due to their powerful modeling capabilities and inherited simplicity and explainability. In the literature, the term SVM has been used to refer to a classification model, while Support Vector Regressor (SVR) referred to regression models. In this work, we use the term *SVM* to denote the adopted regression model. SVM is a margin maximization technique that assumes the data is linearly separable. We adopt a kernelized version of SVM since the linear separability assumption is not guaranteed in practice, especially in our dataset that exhibits a high degree of nonlinearity. Several kernels can be adopted

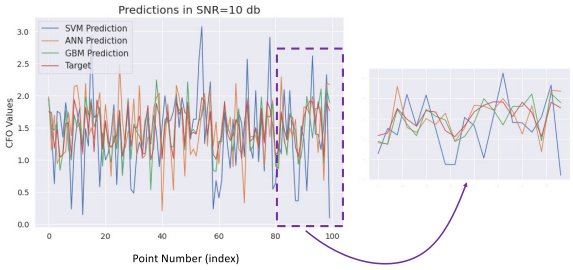

FIGURE 7. The prediction accuracy of our proposed GBM with two baseline models (i.e., ANN and SVR).

such as Polynomial Kernel, Gaussian Kernel, or Radial Basis Function (RBF) [26]. In this work, we adopted an RBF kernel (19).

$$\mathcal{K}(x, y) = \exp[\frac{\|x-y\|^2}{2\sigma^2}]. \tag{19}$$

The prediction results of the GBM and the two baselines (i.e., ANN, SVM) are shown in Fig. 7. We can see that the prediction accuracy of our proposed GBM outperforms the predictions of the other two models. We can see the predictions of the ANN model are closer to the target compared with the predictions of the SVM model. However, the predictions of the GBM are the closest to the target values. This can be attributed to the ensemble nature of the GBM model. Many weak learners can perform better than a single stronger learner. Ensemble models exploit multiple weak learners to produce weak predictions based on features extracted through various data projections. The produced results are then fused with any voting mechanisms to achieve better performances than that obtained by any standalone learner. This gives the GBM an extra advantage over the other two baseline models [27].

V. FUTURE WORK

As shown previously, the adoption of ML techniques for learning the CFO is a promising direction that can bring plenty of advantages, especially for B5G communication systems. However, challenges such as dimensionality reduction, hyperparameter optimization, and building universal predictive models require further investigation. In this section, we propose various research directions for future exploration such as:

- In this work, we proposed and validated the use of PCA as a dimensionality reduction technique. Other dimensionality reduction techniques are worth exploring. Specifically, neural network architectures such as autoencoders have been widely used for dimensionality reduction and they showed an outstanding performance in this regard. Additionally, they are capable of capturing the nonlinearities inherent in communication systems through nonlinear dimensionality reduction.
- Towards the goal of zero-touch network management, Fine-tuning the different hyper-parameter values to obtain the best model becomes a challenging task, especially when we need to retrain the model (e.g., after

data drift). We proposed Bayesian optimization as a solution for our work. The authors believe that more contributions in this area will be appreciated.

- A dedicated model for each SNR value has been proposed in this work. Another direction that looks more appealing is the use of a universal predictive model that can be applied to all SNR values. This solution is of more interest, and more challenging as well.
- In order to avoid wasting resources training models from scratch for each new device being installed in a new environment, transfer learning techniques can be utilized and optimized to reduce the required resources for training CFO predictive models. The impact of such techniques on the performance of the CFO predictive models should be evaluated.
- The problem can be extended to span different channel models (slow versus fast-fading, etc.) and different numerology settings.

We believe that releasing our dataset for public access can encourage other researchers to investigate more in these directions, among others.

VI. CONCLUSION

In this work, we proposed a machine-learning approach for carrier frequency offset (CFO) estimation using gradient-boosting machines (GBM). Compared with various baseline models, our proposed model achieved a competitive performance in terms of prediction accuracy. Moreover, we released our dataset as open source to motivate other researchers to continue investigating data-driven solutions for CFO. We also proposed several promising research directions for further investigating the feasibility of data-driven CFO for new radio.

REFERENCES

[1] A. Abdelmoaty, D. Naboulsi, G. Dahman, G. Poitau, and F. Gagnon, "Resilient topology design for wireless backhaul: A deep reinforcement learning approach," *IEEE Wireless Commun. Lett.*, vol. 11, no. 12, pp. 2532–2536, Dec. 2022.

[2] P. Siyari, H. Rahbari, and M. Krunz, "Lightweight machine learning for efficient frequency-offset-aware demodulation," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2544–2558, Nov. 2019.

[3] K. Ramadan, K. F. Ramadan, A. S. Fiky, H. Alam, M. I. Dessouky, and F. E. AbdEl-Samie, "Joint low-complexity equalization and CFO estimation and compensation for UWA-OFDM communication systems," *Int. J. Commun. Syst.*, vol. 33, no. 3, p. e3972, 2020.

[4] M. Hussien, M. F. Ahmed, G. Dahman, K. K. Nguyen, M. Cheriet, and G. Poitau, "Towards more reliable deep learning-based link adaptation for WiFi 6," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[5] H. Lee, S. H. Lee, and T. Q. Quek, "Deep learning for distributed optimization: Applications to wireless resource management," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2251–2266, Aug. 2019.

[6] M. Hussien, K. K. Nguyen, and M. Cheriet, "PRVNet: A novel partially-regularized variational autoencoders for massive MIMO CSI feedback," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 2286–2291.

[7] J. Chen, W. Feng, J. Xing, P. Yang, G. E. Sobelman, D. Lin, and S. Li, "Hybrid beamforming/combining for millimeter wave MIMO: A machine learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11353–11368, Dec. 2020.

[8] M. H. Nassralla, M. M. Mansour, and L. M. A. Jalloul, "A low-complexity detection algorithm for the primary synchronization signal in LTE," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8751–8757, Oct. 2016.

[9] M. Morelli and M. Moretti, "A robust maximum likelihood scheme for PSS detection and integer frequency offset recovery in LTE systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1353–1363, Feb. 2016.

[10] J.-C. Lin, Y.-T. Sun, and H. V. Poor, "Initial synchronization exploiting inherent diversity for the LTE sector search process," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1114–1128, Feb. 2016.

[11] R. Shaked, N. Shlezinger, and R. Dabora, "Joint estimation of carrier frequency offset and channel impulse response for linear periodic channels," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 302–319, Jan. 2018.

[12] H. Abdzadeh-Ziabari, W.-P. Zhu, and M. N. S. Swamy, "Joint maximum likelihood timing, frequency offset, and doubly selective channel estimation for OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2787–2791, Mar. 2018.

[13] R. M. Dreifuerst, R. W. Heath Jr., M. N. Kulkarni, and J. Charlie, "Deep learning-based carrier frequency offset estimation with one-bit ADCs," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.

[14] V. Ninkovic, A. Valka, D. Dumic, and D. Vukobratovic, "Deep learning-based packet detection and carrier frequency offset estimation in IEEE 802.11ah," *IEEE Access*, vol. 9, pp. 99853–99865, 2021.

[15] J. Zhang and H. Wang, "Learning the optimal LLR under carrier frequency offset," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 441–446.

[16] M. Elwekeil, S. Jiang, T. Wang, and S. Zhang, "Deep convolutional neural networks for link adaptations in MIMO-OFDM wireless systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 665–668, Jun. 2019.

[17] *NR; Physical Channels and Modulation*, 3GPP, document TS 138 211, 2020.

[18] A. Ali, M. Elsaadany, and G. Gagnon, "Performance of time and frequency approaches for synchronization tracking in 5G NR systems," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2021, pp. 1–6.

[19] I. B. Mustapha and F. Saeed, "Bioactive molecule prediction using extreme gradient boosting," *Molecules*, vol. 21, no. 8, p. 983, 2016.

[20] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[21] C. de Bodt, D. Mulders, M. Verleysen, and J. A. Lee, "Nonlinear dimensionality reduction with missing data using parametric multiple imputations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1166–1179, Apr. 2019.

[22] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Phil. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901.

[23] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[24] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2014.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[26] L. Nguyen, "Tutorial on support vector machine," *Appl. Comput. Math.*, vol. 6, nos. 1–4, pp. 1–15, Jul. 2017.

[27] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020.

MOSTAFA HUSSIEN received the B.Sc. (Hons.) and M.Sc. degrees from Assiut University, Assiut, Egypt. He is currently pursuing the Ph.D. degree with École de Technologie Supérieure (ÉTS), University of Québec, Montréal, Canada. He has completed various internships in industrial companies and research institutes, such as Samsung AI Center (SAIC), Nokia Bell-Laboratories, and Ciena. He has contributed to developing AI-based power-saving algorithms for Samsung 5G RAN. He holds many fellowships, such as MITACS and DAAD from the Canadian and German Governments, respectively. His research interests include artificial intelligence and next-generation wireless systems. He was a reviewer of different conferences (e.g., IEEE CVPR, ICCV, and WACV) and journals (e.g., IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, and IEEE COMMUNICATIONS LETTERS).

AHMED ABDELMOATY (Member, IEEE) received the B.S. degree in electronics and telecommunications engineering from Al-Azhar University, Cairo, Egypt, in 2004, the M.S. degree in electrical engineering from the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, in 2017, and the Ph.D. degree from École de Technologie Supérieure (ETS), University of Quebec, Canada, in 2022. Currently, he is a Postdoctoral Fellow with the Electrical Engineering Department, ETS. His research interests include wireless communications, network design, machine learning, and RF propagation.

MAHMOUD ELSAADANY (Senior Member, IEEE) received the B.Sc. (Hons.) and M.Sc. degrees in electrical engineering from Cairo University, Giza, Egypt, in 2006 and 2010, respectively, and the Ph.D. degree in electrical and computer engineering from Concordia University, Montreal, QC, Canada, in 2018. He is currently an Assistant Professor with the ECE Department, Concordia University, and also a Research Professional with École de Technologie Supérieure (ETS), Université du Québec, Montreal. He was a Researcher with Qatar University, Doha, Qatar, from 2008 to 2010. His current research interests include digital signal processing, optimization of microwave components, machine-type communication, and algorithm design for 5G cellular networks.

MOHAMMED F. A. AHMED received the B.Sc. and M.Sc. degrees in communication engineering from Assiut University, Assiut, Egypt, in 2001 and 2004, respectively, and the Ph.D. degree in communications engineering from the University of Alberta, Edmonton, AB, Canada, in 2011. He was a Postdoctoral Research Fellow with King Abdullah University of Science and Technology (KAUST), Saudi Arabia. From 2012 to 2014, he was with École de Technologie Supérieure (ÉTS), Montreal, QC, Canada. He joined NSERC/Ultra Electronics TCS Industrial Research Chair, ÉTS, in 2017, and the Resilient Machine Learning Institute (ReMI), in 2019. In both positions, his research and development focus was on ML models/algorithms for critical communications systems. Currently, he is with the Canadian National Railway (CN) as an Expert Data Developer.

GHYSLAIN GAGNON (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Carleton University, Ottawa, ON, Canada, in 2008. He is currently the Dean of Research and a Professor with École de Technologie Supérieure, Montreal, QC, Canada. He is also a Board Member of ReSMiQ. He was the Director of the Research Laboratory, LACIME, from 2013 to 2020, a group of 15 faculties and nearly 150 highly dedicated students and researchers in microelectronics, digital signal processing, and wireless communications. He is highly inclined toward research partnerships with the industry. His research interests include microelectronics, digital signal processing, and machine learning with various applications, including health care, media art, and building energy management.

KIM KHOA NGUYEN (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Concordia University. He is an Associate Professor with the Department of Electrical Engineering, Ecole de Technologie Supérieure (ETS), University of Quebec, Montreal, Canada. In the past, he served as the CTO at Inocybe Technologies (currently Kontron Canada), the world's leading company in software-defined networking (SDN) solutions. He was an architect of Canarie's GreenStar Network and was also involved in establishing CSA/IEEE standards for green ICT. He has led research and development in large-scale projects at Ericsson, Ciena, Telus, InterDigital, and Ultra Electronics. His research interests include cloud computing, network optimization, the IoT, big data, machine learning, AI, smart building, smart city, high-speed networks, and green ICT. He was a recipient of the Microsoft Azure Global IoT Contest Award, in 2017, and Ciena's Aspirational Prize, in 2018.

MOHAMED CHERIET (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the University of Pierre and Marie Curie (Paris VI), in 1985 and 1988, respectively. Since 1992, he has been a Professor with the Systems Engineering Department, École de Technologie Supérieure (ÉTS), University of Quebec, Montreal, where he was appointed as a Full Professor, in 1998. He is the Founder and Director of the Synchromedia Laboratory for Multimedia Communication in Telepresence Applications, since 1998. He has extensive research experience in cloud computing, network virtualization, and softwarization. In addition, he is an expert in computational intelligence, pattern recognition, machine learning, artificial intelligence, and perception. He has published more than 450 technical papers in the se fields. He is a holder of the Tier 1 Canada Research Chair on sustainable and smart eco-cloud, in 2013, Since then, he leads the establishment of the first smart-university campus in Montreal, Canada, created as a hub for innovation and productivity. He is the General Director of the FRQNT Strategic Cluster on the operationalization of sustainability development, CIRODD (2019–2026). He is a 2016 fellow of the International Association of Pattern Recognition (IAPR), a 2017 fellow of the Canadian Academy of Engineering (CAE), a 2018 fellow of the Engineering Institute of Canada (EIC), and a 2019 fellow of Engineers Canada (EC). He was a recipient of the 2016 IEEE J. M. Ham Outstanding Engineering Educator Award, the ÉTS Research Excellence Prize in 2013, for his outstanding contribution to green ICT, cloud computing, and big data analytics research areas, and the 2012 Queen Elizabeth II Diamond Jubilee Medal. He is the Founder and the Former Chair of the IEEE Montreal Chapter of Computational Intelligent Systems (CIS), a Steering Committee Member of the IEEE Sustainable ICT Initiative, and the Chair of the ICT Emissions Working Group. He published along with his working group, the first standard ever, IEEE 1922.2, on the real-time calculation of ICT emissions, in April 2020. He serves on the editorial boards for several renowned journals and international conferences.