Received: 29 April 2020 | Revised: 12 August 2020 | Accepted: 24 August 2020

DOI: 10.4218/etrij.2020-0188

SPECIAL ISSUE

ETRI Journal WILEY

A supervised-learning-based spatial performance prediction framework for heterogeneous communication networks

Shubhabrata Mukherjee[1] | **Taesang Choi**[2] | **Md Tajul Islam**[1] | **Baek-Young Choi**[1] | **Cory Beard**[1] | **Seuck Ho Won**[2] | **Sejun Song**[1]

[1]Department of Computer Science and Electrical Engineering, University of Missouri, Kansas City, MI, USA

[2]Electronics Telecommunications Research Institute, Daejeon, Rep. of Korea

Correspondence
Shubhabrata Mukherjee, Department of Computer Science and Electrical Engineering, University of Missouri, Kansas City, MI, USA.
Email: smpw5@mail.umkc.edu

Taesang Choi, Electronics Telecommunications Research Institute, Daejeon, Rep. of Korea.
Email: choits@etri.re.kr

In this paper, we propose a supervised-learning-based spatial performance prediction (SLPP) framework for next-generation heterogeneous communication networks (HCNs). Adaptive asset placement, dynamic resource allocation, and load balancing are critical network functions in an HCN to ensure seamless network management and enhance service quality. Although many existing systems use measurement data to react to network performance changes, it is highly beneficial to perform accurate performance prediction for different systems to support various network functions. Recent advancements in complex statistical algorithms and computational efficiency have made machine-learning ubiquitous for accurate data-based prediction. A robust network performance prediction framework for optimizing performance and resource utilization through a linear discriminant analysis-based prediction approach has been proposed in this paper. Comparison results with different machine-learning techniques on real-world data demonstrate that SLPP provides superior accuracy and computational efficiency for both stationary and mobile user conditions.

KEYWORDS

5G, heterogeneous communication network, machine learning, performance prediction, supervised learning

1 | INTRODUCTION

Future wireless communication networks are expected to support rapidly growing mobile data with huge numbers of complex Internet of things (IoT) applications, which require high data rates, complete coverage, and a thousand-fold capacity increase. For example, it is predicted that the total number of Internet users will be more than 5 billion, with almost 30 billion devices connected worldwide, by 2023 [1]. More than 70% of the global population will be connected and mobile and Wi-Fi network speeds will triple by 2023 [1]. These incremental networks will generate a massive amount of network traffic; Cisco predicted the generation of 3.3 ZB of traffic per year by 2021 [2]. Almost 53% of total network traffic was generated by mobile users in 2019. This ratio is likely to increase in the future [3]. Therefore, it is necessary to ensure the continuity of reliable and secure network services for heterogeneous mobile IoT devices.

Although traditional terrestrial networks can provide high-speed data services, non-terrestrial systems, such as satellites and the Loon network, may support extended coverage to otherwise unreachable areas. However, any individual system cannot achieve ubiquitous communication coverage and service continuity [4]. Therefore, as shown in Figure 1, we believe that heterogeneous communication networks (HCNs), including the emerging sixth-generation (6G) and

Loon wireless systems, terrestrial mobile and wireless networks, satellites, IoT, Wi-Fi, and Bluetooth, are crucial for meeting the challenges facing future mobile communications [5–7]. Using various terrestrial and non-terrestrial systems in parallel requires more than the simple combination of systems because different systems employ different protocol standards, data formats, and access technologies. Different network resources are expected to cooperate to support more efficient data transmission and provide more reliable services. However, current network architectures cannot efficiently exploit all the benefits of HCNs

In this paper, we propose a novel supervised-learning-based spatial performance prediction (SLPP) framework for next-generation HCNs that improves network performance, ensures seamless network management, and enhances service quality. To support the critical network functions of HCNs, including adaptive asset placement, dynamic resource allocation, and load balancing, it is highly beneficial to harness accurate performance prediction methods for different systems. Unlike many existing systems that use measurement data to react to network performance changes, SLPP proactively improves network performance, automates network optimization and management, and reduces operational costs and energy consumption [9,10]. SLPP can be used to determine the type of wireless network (cellular, Wi-Fi, satellite) in an appropriate location that is suitable for a specific application at a particular time. For example, if there are mission-critical communications occurring, which must satisfy ultra-reliable low-latency communication (URLLC) requirements, the SLPP model can anticipate which network can meet these requirements [11]. Additionally, in a secure virtual private network application, the proposed prediction framework can accurately predict which system can provide the maximum data security [12,13].

SLPP harnesses a robust network performance prediction framework for optimizing performance and enhancing the resource utilization of various HCN systems by taking advantage of recent improvements in both machine-learning and statistical learning methods. Specifically, SLPP defines a novel machine-learning-based performance prediction model for HCNs for both terrestrial and non-terrestrial networks, using previous data as a training dataset to predict the performance of a network.

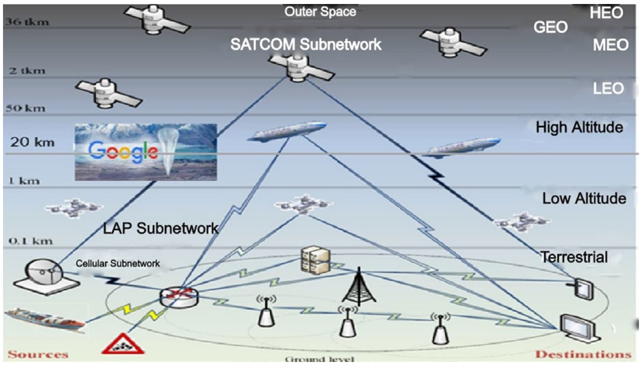


FIGURE 1 Co-existence of a terrestrial and non-terrestrial heterogeneous network [8]

As illustrated in Figure 2, the SLPP framework is added as an artificial intelligence (AI) layer on top of a traditional communication network structure, which contains the infrastructure and management layers. The infrastructure layer includes user equipment (UE), radio access nodes, core nodes, and network clouds, which interface to process control signals and user data. The management layer provides various essential network operations and management functions, including fault management, configuration management, and performance administration. These two functional representations are valid for any legacy communication network, such as 4G, 5G, or Wi-Fi, as well as any advanced and sophisticated network scenarios, such as HCNs or 6G.

The proposed SLPP framework represents an attempt to define machine-learning-based automation's role in the context of next-generation communication networks, where machine-learning-based automation is the key to orchestrating all network management functionalities efficiently according to [14,15]. As shown in Figure 2, the AI layer above the management layer supports various intelligent network functions by automating network management processes. The primary AI layer functionality includes event-based classification, such as automated root cause analysis (RCA), management of network faults, key performance indicators (KPIs), degradation prediction for preventive maintenance, and proactive fault recovery. The spatial performance prediction service function advocates service- and capacity-based network planning. Additionally, the SLPP supports reliability prediction for the performance- and capacity-based load balancing of HCNs. The AI layer maintains bi-directional information flows with the management layer. Therefore, the management layer periodically receives accurately predicted information regarding various types of network performance information, such as latency and signal level reliability, from the AI layer, allowing it to maintain optimal network performance in terms of faults, configurations, and performance management services. Using the SLPP framework, it is possible to determine which networks provide the best performance. According to these performance results, the SLPP can ensure optimal network resource (bandwidth and power) embedding. By employing the proposed model, network operators can perform service reliability prediction for efficient resource provisioning, preventive network maintenance, and the prompt deployment of alternative networks (including fast RCA) during natural disasters. SLPP can eliminate the requirements of traditional drive testing, which requires a substantial workforce and valuable network resources. SLPP can also be useful as a critical design algorithm for performance- and capacity-based load balancing for heterogeneous network environments. A

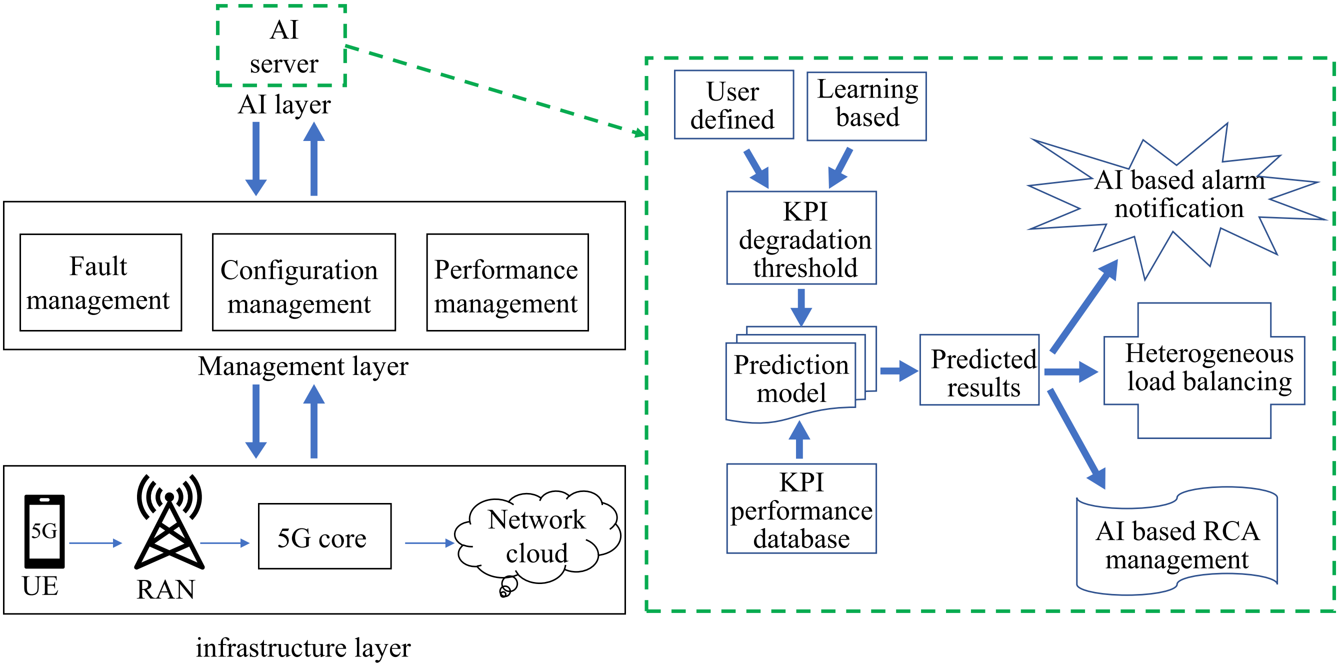


FIGURE 2 A supervised-learning-based spatial performance prediction (SLPP) framework

performance prediction framework can also be realized by constructing an event-based classification system for automated RCA management and KPI degradation prediction. We present the detailed procedure for a performance prediction scenario, where we predict the signal level of a long-term evolution (LTE) network based on historical performance data. However, we attempt to emphasize the detailed descriptions of algorithm- and real-data collection-based approaches, similar to [16,17], in contrast to the comprehensive system-level illustrations presented in [18] because of the unavailability of a complete industrial network architectural environment. We plan to overcome this limitation in future work by constructing an open-source HCN simulation platform.

As a proof of concept for our methodology, we collected real-world LTE network data as training data for both stationary and mobile users. We leverage a linear discriminant analysis (LDA)-based classification algorithm for predicting the signal levels of an LTE network and present performance comparisons with other frequently used supervised and unsupervised approaches, such as linear regression and K-means clustering. The main contributions of this paper can be summarized as follows:

- We propose a comprehensive framework for multi-network multi-parameter prediction with process flows.
- We use real data from networks and other relevant performance parameters that would facilitate faster product implementation based on a proof of concept of our model.
- We present detailed descriptions of a methodical approach for solving issues related to working with practical raw data.
- We present performance comparisons with other popular machine-learning approaches in terms of accuracy and computational efficiency.

The remainder of this paper is organized as follows. Section 2 discusses the background of our research and related works. Section 3 discusses the detailed process flow for constructing a robust performance prediction system. Section 4 describes our data collection methodology. Section 5 describes the challenges faced and the solutions used to construct the proposed model. Section 6 describes the proposed LDA-based prediction model and Section 7 presents obtained experimental results. Finally, Section 8 summarizes the conclusions derived from our study on heterogeneous network performance prediction.

2 | BACKGROUND AND RELATED WORK

One of the most ambitious projects created by Alphabet Inc.'s (Google's parent company) Google X laboratory is "Project Loon," which aims to connect people everywhere and expand Internet access to the billions who currently lack access [19]. Loon uses a temporospatial software-defined network (SDN) that applies supervised and unsupervised machine learning to model time-dynamic wireless signal propagation [20]. However, the Loon SDN is only designed for homogeneous network conditions, primarily LTE, meaning a more generalized approach is required to model current heterogeneous networks. Additionally, some recent studies have described prediction approaches to consider performance under both stationary and mobile conditions, which mainly pertain to cellular networks [9,21,22]. In a recent paper, the authors described a comprehensive framework for cellular performance prediction. However, this work can be extended to construct a generic prediction model for heterogeneous networks [10]. In another paper, the authors discussed various deep-learning techniques for space—air–ground integrated network optimization [23]. Some relevant studies have used neural-network-based handover for multi-radio-access-technology (multi-RAT) networks, deep belief networks for traffic flow prediction, and K-means clustering to enhance the results of network planning tools, but have not considered the relatively simple and efficient techniques of supervised machine learning [24–26]. Additionally, some researchers have used regression and other non-AI-based techniques to detect performance anomalies in cellular or WiMAX networks. However, the effects of performance degradation under stationary and mobile conditions have not been explicitly described [27,28]. Some researchers have also used supervised-learning techniques such as k-nearest neighbors for machine-learning-based handover in vehicular networks, which demonstrates the versatility of performance prediction for automating numerous network functions [29].

The authors of [30] proposed a supervised deep-learning-based system for appropriate input and output characterizations of heterogeneous network traffic with multiple hidden layers to compute non-linear transformations of previous layers. They used a greedy layer-wise training method to initialize their deep-learning system and the back-propagation algorithm to fine tune the deep-learning process. Tang and others [31] proposed a novel deep-learning-based traffic load prediction algorithm to forecast future congestion in SDN--IoT networks in combination with a partial channel assignment algorithm to allocate channels to each link intelligently. The authors of [32] proposed a deep-reinforcement-learning approach to minimize prediction uncertainty for dynamic resource allocation. In [33], the authors used deep learning to optimize traffic and enable low-latency and reliable content caching for transmitting virtual reality content from unmanned aerial vehicles. The authors of [34] followed a specific reinforcement-learning-based approach called the “multi-armed bandit” approach to solve the challenge of handover between macrocells and picocells. However, their work can be extended to optimize multi-RAT handover. Yu and others [35] also used a deep-reinforcement-learning-based medium access control protocol for heterogeneous wireless networking.

Most of the methods discussed above require significant data processing and are inefficient in terms of training and testing because they also require considerable amount of time for model deployment and learning. Additionally, rather than centralized model deployment, these methods use nodal analysis or implementation, which increases complexity and scalability issues. Accuracy and effectiveness are also concerns for such models because heterogeneous environments and networks exhibit data variation and feature dissimilarity. Unlike the studies discussed above, we propose a framework for the optimal combination of supervised and unsupervised learning to maximize accuracy with minimal complexity. We consider both terrestrial and non-terrestrial networks for modeling our prediction algorithm instead of using only limited standards such as reliability and throughput. Our model considers network security and availability as crucial parameters for predicting network performance and performing decision making. As a proof of concept, we consider real-time raw network data for multiple user conditions and locations, which increases our model's generalizability and effectiveness.

3 | PROCESS FLOW

We propose a performance-based prediction approach called “œspatial performance prediction” as the basis for an algorithm to predict the performance of heterogeneous networks effectively. The detailed process flow for this approach is presented in Figure 3. In this approach, the system is initially trained using various network performance metrics, namely throughput, reliability, latency, current user statistics, and the deployed security schemes of multiple access networks, such as LTE, 5G, Wi-Fi, and satellite. Based on this initial training data, the system will generate a robust spatial performance map of the heterogeneous environment. Finally, with help from the predicted network model, the system can determine the load distribution paradigm for the overall heterogeneous network. Our system assumptions are based on the desire to handle end-to-end load balancing for any combination of terrestrial and non-terrestrial networks because we incorporate both types of networks. Because it is spatial in nature, our model leverages the advantages of prototyping a multi-access network for any combination of space and network types.

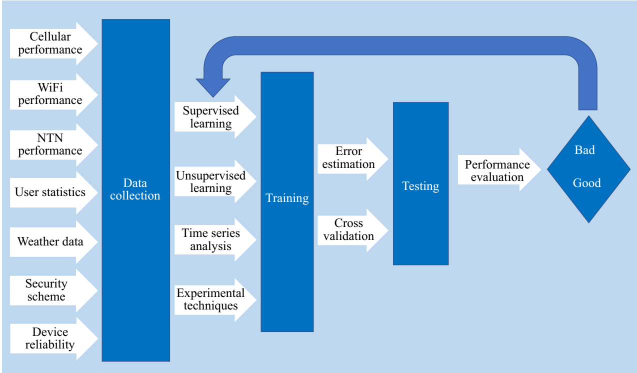


FIGURE 3 Network performance prediction process flow

Additionally, instead of simply using traditional performance metrics such as latency or throughput, our model includes security performance as a significant criterion for categorizing network preferences. To the best of our knowledge, ours is the first model that accounts for security performance when ranking heterogeneous terrestrial and non-terrestrial networks.

4 | DATA COLLECTION METHODOLOGY

We incorporated an Android application called "Network Cell Info" [36] to record network performance data in real time. The collected data include various heterogeneous network environments (primarily UMTS, LTE, GSM, and Wi-Fi). For our current prediction model, we focus on LTE data, where "T-Mobile" is considered as a cellular operator and the "One Plus 6" and "Samsung S5" devices are considered as user equipment. We accumulated performance data from various locations, such as residences, streets, universities, and airports, under both stationary and mobile conditions. Our mobile condition data contain variations such as walking, travel via public transport, and travel via private vehicles, which represent the effects of numerous mobility conditions. Finally, we collected historical weather data for the target coverage locations from [37]. We recorded and accumulated a total of 34 network performance parameters and nine weather data parameters, namely temperature, dew point, humidity, wind direction, wind speed, wind gust, atmospheric pressure, precipitation, and weather conditions (cloudy, rainy, etc.). Table 1 lists some examples of these parameters. These 43 independent predictors serve as the primary source of data for predicting the performance of cellular (LTE, UMTS, and GSM) networks.

TABLE 1 Collected network data types and notations

Parameter	Descriptions
Clid	ECI for LTE, UCID for UMTS, CID for GSM
Acc	Device location accuracy at the time of measurement
Bearing	Device direction of travel at the time of measurement
Alt	Device altitude at the time of measurement
rsrq	LTE reference signal received quality (RSRQ)
rssnr	LTE RSSNR (reference signal signal-to-noise ratio)
ta	LTE TA (timing advance)
Ws	Speed and direction of wind
Wg	Sudden, brief increase in wind speed
Ppt	Hourly precipitation measure

5 | COLLINEARITY ISSUES OF PERFORMANCE DATA

We encountered a few challenges related to our raw collected data and had to perform multi-stage preprocessing prior to using them as inputs for the main prediction model. The primary issue is that the raw performance data exhibit the "multicollinearity" property. Multicollinearity or collinearity refers to a situation in which two or more predictor variables are closely related [38]. Collinearity reduces the accuracy of the estimates of fitting coefficients $\widehat{\beta}_j$, which leads to the miscalculation of standard error. The t-statistic for each predictor is calculated by dividing $\widehat{\beta}_j$ by its standard error. Collinearity results in a decline in the t-statistic, which leads to the failure to reject the null hypothesis $H_0\colon \widehat{\beta}_j = 0$. The best way to identify multicollinearity is to compute the variance inflation factors (VIF) of data. The VIF is the ratio of the variance of $\widehat{\beta}_j$ when fitting the full model divided by the variance of $\widehat{\beta}_j$ when fitting that variable alone. The smallest possible value for the VIF is one, which indicates the complete absence of collinearity. It has been observed that in practical cases, there is almost always a small amount of collinearity among predictors. As a rule of thumb, a VIF value that exceeds 10 indicates a problematic amount of collinearity. The VIF of each variable can be computed using (1), where $R^2_{(X_j|X_{-j})}$ is a coefficient of determination that is derived from the regression of X_j over all other predictors. If $R^2_{(X_j|X_{-j})}$ is close to one, we can conclude that collinearity is present and the VIF will be large.

$$\mathrm{VIF}(\widehat{\beta}_j) = \frac{1}{1 - R^2_{(X_j|X_{-j})}}. \tag{1}$$

In the first step of data preprocessing, we found that the data exhibit "perfect multicollinearity," which refers to the condition when there is an exact linear relationship between two or more variables [38]. We can conclude that a set of variables is perfectly multicollinear if there are one or more exact linear relationships among any variables, which is similar to (2), where $X_{1i}, X_{2i}, \ldots, X_{ki}$ are the predictor variables and $\lambda_0, \lambda_1, \ldots, \lambda_k$ are the constants.

$$\lambda_0 + \lambda_1 X_{1i} + \lambda_2 X_{2i} + \cdots + \lambda_k X_{ki} = 0. \tag{2}$$

Perfect multicollinearity may be encountered frequently when handling raw datasets that contain redundant information. To rectify the issue of perfect multicollinearity in our data, we performed linear regression between the predicted variables and all other predictors. We adopted an alias function to identify the variables responsible for perfect multicollinearity, as shown in (3).

$$\text{myfit}=\text{lm}(\text{sigl} \sim .,\ \text{data}=\text{lte});\ \text{alias}(\text{myfit}). \tag{3}$$

We excluded the identified predictors from our dataset and rechecked to confirm that the perfect multicollinearity was eliminated from our data. We then performed Eigensystem analysis to identify any other predictors responsible for multicollinearity in our data by leveraging (4) and (5).

$$\frac{\text{max(eigen(cor(lte))\$values)}}{\text{min(eigen(cor(lte))\$values)}}, \tag{4}$$

$$\text{kappa(cor(lte), exact}=\text{TRUE)}. \tag{5}$$

We derived the "condition number" and "Kappa" from (4) and (5). Both values were in the range of ~2 × 10^6, which indicates the presence of strong collinearity. If the condition number is above 30, it indicates that the regression contains severe multicollinearity [38]. We can also confirm multicollinearity if two or more of the variables are related to the high condition number with high levels of variance [10]. The condition number is computed by calculating the square root of the maximum eigenvalue divided by the minimum eigenvalue in the design matrix. This method identifies which variables are responsible for the multicollinearity problem. In the next step, we used the VIF to identify the variables responsible for this collinearity, as shown in (6). We calculated the mean VIF using (6) and found a large value (~2 × 10^4 range). We acquired the VIFs for individual predictors to identify the predictors responsible for collinearity. Next, we recalculated the mean VIF and individual VIFs and removed the predictors with VIF > 5. After removing the predictors responsible for collinearity, we determined that the condition number and Kappa value were less than 15 and that the mean VIF was reduced to the range of ~1.4. Therefore, we confirmed that our dataset was free from the collinearity problem and ready to use for our prediction model.

$$v=\text{vif(myfit)};\ \text{sort}(v);\ \text{mean}(v). \tag{6}$$

6 | SPATIAL PERFORMANCE PREDICTION MODEL

We prepared our final prediction model based on 10 predictors, which are listed in Table 1. We designed a classification model for the prediction of signal levels under stationary conditions and another similar classification model for mobile conditions. An LDA approach was used for signal level prediction. LDA is a supervised-learning approach for estimating the conditional distribution of the response Y to given predictors X [39]. LDA represents the distributions of the predictors X separately in each of the response classes and then uses Bayes' theorem according to (7) to estimate a conditional probability $P_r(Y=k|X=x)$. If we wish to classify an observation into one of K classes, where $K \geq 2$, we denote $f_k(x) \equiv P_r(X=x|Y=k)$ as the density function of X for an observation that comes from the k-th class. We assume that in (7), X is a discrete random variable. If it is assumed that $\sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2$, then there is a shared variance term across all K classes, which can be denoted as σ^2.

$$P_r\,(Y=k|X=x)=\frac{\pi_k f_k\,(x)}{\sum_{l=1}^{k} \pi_l f_l\,(x)}, \tag{7}$$

$$p_k(x)=\frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{(x-\mu_k)^2}{2\sigma^2}}{\sum_{l=1}^{k} \pi_1 \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{(x-\mu_l)^2}{2\sigma^2}}. \tag{8}$$

π_k denotes the prior probability that an observation belongs to the k-th class, and μ_k and σ_k^2 are the mean and variance of the k-th class, respectively. The LDA classifier assumes that the observations within each class come from a normal distribution with a class-specific mean vector and common variance σ_k^2. When plugging estimates for these parameters into the Bayes' classifier in (8), it is assumed that there is only one predictor ($p=1$). However, for the LDA analysis of more than one predictor ($p>1$), it is assumed that $X=(X_1, X_2, \ldots, X_p)$ is drawn from a multivariate Gaussian or multivariate normal distribution with a class-specific multivariate mean vector and common covariance matrix [39]. To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu, \sum)$. Here, $E(X)=\mu$ is the mean of X (a vector with p components) and $\text{Cov}(X)=\sum$ is the $p \times p$ covariance matrix of X. Finally, the multivariate Gaussian density and discriminant function are defined as shown in (9).

$$f(x)=\frac{1}{2\pi^{\frac{p}{2}}|\sum|^{\frac{1}{2}}} \exp - \frac{1}{2}(x-\mu)^T 1 - \widehat{\sum}(x-\mu), \tag{9}$$

$$\delta_k(x)=x^T 1 - \widehat{\sum}\mu_k^T - \frac{1}{2}\mu_k^T 1 - \widehat{\sum}\mu_k + \log\pi_k. \tag{10}$$

The Bayes' classifier assigns an observation $X=x$ to the class for which (8) is maximized. $f(x)$ in (9) is the multivariate Gaussian density function and $\delta_k(x)$ in (10) is the discriminant function. Although (10) has a complex form, the discriminant function is linear, as indicated in (11). Therefore, this technique is called LDA.

$$\delta_k(x)=C_{k0}+C_{K1}x_1+C_{K2}x_2+\ \ldots\ +C_{Kp}x_p. \tag{11}$$

Initially, based on the raw data, we performed data preprocessing and rectified the multicollinearity issue as described

		Predicted class		
		Class 2	Class 3	Class 4
Actual class	Class 2	**63**	0	11
	Class 3	12	**30**	28
	Class 4	33	20	**3404**

FIGURE 4 Confusion matrix

above. After applying LDA to the training data, we obtained a confusion matrix. This confusion matrix compares the LDA predictions to the true classes for the training observations in the dataset. Elements on the diagonal of the matrix represent samples whose signal levels are correctly predicted, while off-diagonal elements represent samples that are misclassified, as shown in Figure 4. In the next step, we conducted a cross-validation procedure to calculate the actual test error. We used K-fold cross-validation for this purpose. This approach involves randomly dividing the set of observations into k groups or folds of approximately equal size. The first fold is treated as a validation set and the proposed method is fitted to the remaining k1 folds. Subsequently, the mean squared error MSE_1 is computed for the observations in the held-out fold. This procedure is repeated k times. Each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error $\mathrm{MSE}_1, \mathrm{MSE}_2, \cdots . \mathrm{MSE}_k$. The K-fold cross-validation estimate is computed by averaging these values according to (12). We set $K = 10$ for optimal validation.

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{MSE}_i. \tag{12}$$

Signal levels are denoted by the numbers 0, 1, 2, 3, and 4. In our experimental results, zero indicates that LTE coverage is absent, one is the weakest LTE signal level, and four is the strongest signal level. We obtained three signal levels of 2, 3, and 4 for the stationary data and denoted them as Class 2, Class 3, and Class 4, respectively. For the mobile data, we observed signal levels of 1, 2, 3, and 4, and denoted them as Class 1, Class 2, Class 3, and Class 4, respectively.

7 | PERFORMANCE EVALUATION

In this section, we evaluate SLPP in terms of accuracy, performance stability, variation, and computational complexity for predicting the signal levels in under stationary and mobile conditions. We assess the most pertinent related approaches, including a Friis equation-based method, linear regression-based approach, and K-means-clustering-based procedure. Such methods are commonly used for network performance prediction. Finally, we present comparative analysis results for all of the tested methods in Table 2. As shown in Table 2, SLPP performs prediction tasks with the highest accuracy, least computational complexity, best performance stability, and lowest mean variance inflation factor for both stationary and mobile conditions.

After solving the multicollinearity problem present in the raw data, we applied LDA to our data for performance prediction. This is a supervised-learning-based approach. Based on the performance results, one can see that LDA is a stable and accurate classification model. The results in Figure 5 reveal that we can achieve an overall accuracy of 95% for predicting signal levels under stationary conditions and 90% under mobile conditions by using LDA. Furthermore, it is observed that the overall condition number, or Kappa, is 0.5 for the stationary condition and 0.48 for the mobile condition. Additionally, the overall VIF is only 1.58 for stationary conditions and 1.33 for mobile conditions. Overall, small values of both Kappa (0.5 for stationary and 0.48 for mobile) and the VIF (1.58 for stationary and 1.33 for mobile) indicate that we were successful in eliminating the collinearity issue in the raw data.

We achieved this performance using only 10 predictors. As shown in Figure 6, after solving the multicollinearity issue, our mean VIF remains within the ranges of 1.4 to 1.8 for stationary conditions and 1.2 to 1.6 for mobile conditions. Our results remain steady with an increase in the sample size. Additionally, the condition number (Kappa) remains below one for both the stationary and mobile conditions, regardless of the sample size. The consistently insignificant values of kappa and the VIF indicate that all of our data are free from collinearity for any number of samples and can be used to construct a stable prediction model. In Figure 7, one can see that the prediction accuracy of our method consistently remains between 94% and 96% for stationary conditions with varying sample sizes. Additionally, the prediction accuracy of our method consistently remains between 89% and 91% for mobile conditions. The consistent performance of user signal level prediction for both the stationary and mobile conditions demonstrates the stability of the classification accuracy of our model for any sample size (Figure 8).

We reviewed some other relevant algorithms for performance prediction for comparative analysis. We considered three types of methods: (i) Classical approaches (Friis equation-based methods), (ii) supervised-learning-based approaches (linear regression methods), and (iii) unsupervised-learning-based approaches (K-means clustering methods).

7.1 | Friis equation-based classical methods

Received signal power can be calculated using the Friis formula shown in (13), where P_r is the received signal power

TABLE 2 Performance comparison table

Approach used	Friis' equation-based	Linear regression-based	*K*-means clustering-based	SLPP: Our approach
Accuracy	Low: Dependant on specific environment and user mobility condition	Limited: Accuracy is not stable due to variance issue	Very Low: Only Up-to 40% in stationary and 60% in moving condition	High: Up-to 95% in stationary and 90% in moving condition
Computational complexity	Low: Mostly linear equation-based static method.	Low: Linear equation-based $O(n)$ computation	High: Increasingly higher order computation, based on number of cluster (k) and initial centroid (l)	Low: Linear equation-based $O(n)$ dynamic computation
Performance Stability	Good: Single static equation-based calculation.	Poor: Highly unstable performance due to variance issue.	Limited: Performance stability highly dependant on number of cluster and initial centroid definition	Excellent: 94%-96% throughout for static and 89%-91% throughout for moving condition
Variance Issue	Low: Variance issue is low due to static nature.	High: High variance and heteroscedasticity issue.	Low: Comparatively stable variation in prediction for specific number of cluster and initial centroid	Very Low: Mean variance inflation factor ≪2 for both static and moving condition performance

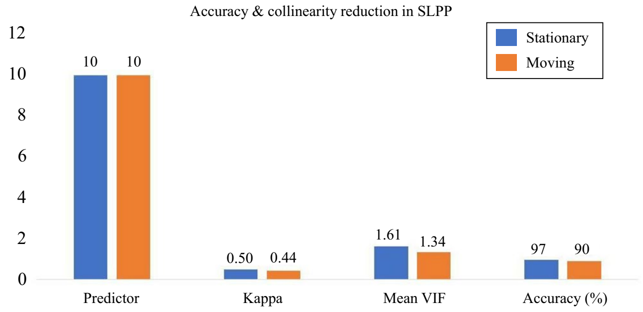


FIGURE 5 SLPP overall performance

at a distance R, P_t is the transmitter antenna power, and G_t and G_r are the transmitter and receiver antenna gains, respectively. The Friis' path loss in (13) is only valid for free space and it assumes that the receiving and transmitting antennas are isotropic. Therefore, the accuracy of this formula is extremely low for real-world scenarios.

$$P_r = \frac{P_r G_t G_r \lambda^2}{(4\pi R)^2}, \tag{13}$$

$$P_r(d) = P_r(d_0) - 10n \log\left(\frac{d}{d_0}\right) + X;\ d > d_0. \tag{14}$$

A generalized form of the free-space Friis equation is presented in (14), where the received power (in dB) decreases at a rate of $(1/d)^n$ and d is the distance between the transmitter and receiver [40]. Here, n is a path loss exponent that can be used for a generalized environment without free space. X is a Gaussian random variable used to capture shadowing probabilities defined in dB. Our model was constructed based on spatially distributed data, meaning (14) is not an appropriate application method for our model because it assumes a Gaussian distribution.

We have considered signal level data in different environments, including indoor and outdoor environments, with various mobility scenarios, such as stationary, slow-moving, and fast-moving vehicles. We also consolidated other environmental values, including weather conditions (ie, Ws, Wg, and Ppt in Table 1). We also integrated fixed wireless channel models such as Gaussian distribution-based, pedestrian-only, and vehicle-only data models in heterogeneous network scenarios to predict perceived signal levels.

7.2 | Linear regression methods

We applied a linear regression method to our dataset and observed the resulting performance. Although the error performance (route mean square error or RMSE, the relative measure of the percentage of the dependent variable variance or R^2, and Mean absolute error or MAE) is

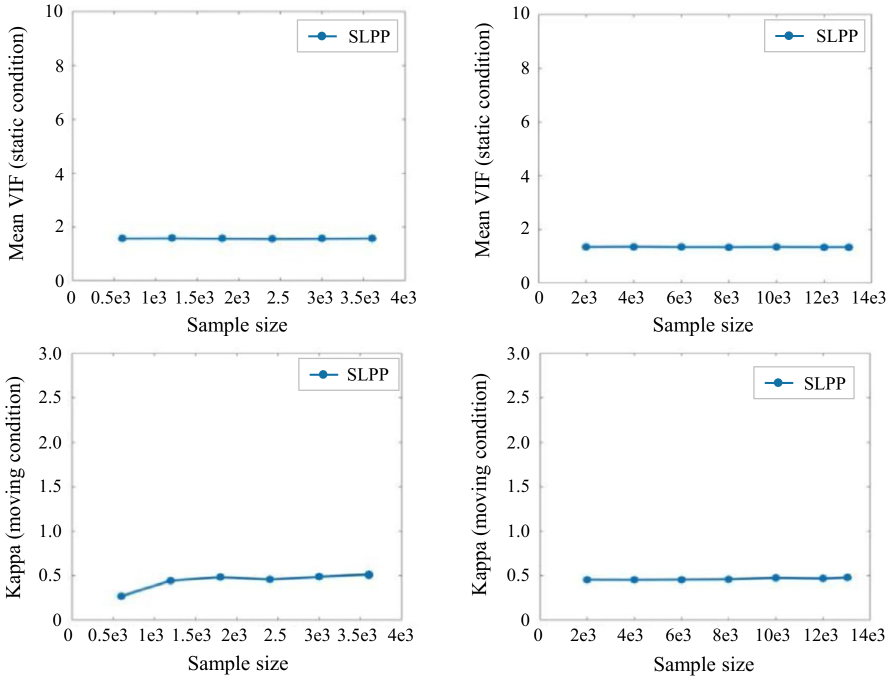


FIGURE 6 SLPP multicolinearity performance

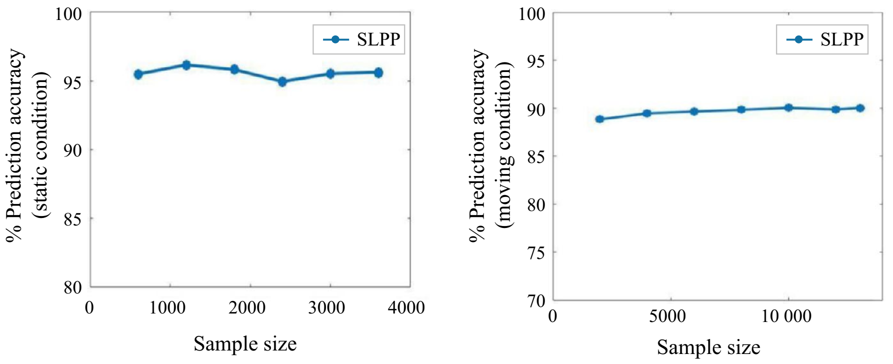


FIGURE 7 SLPP accuracy stability

stable for the mobile conditions when using regression, it varies significantly for stationary signal strength prediction (particularly RMSE), as shown in Figure 9, which results in unreliable error performance for the prediction model. The main issue in the regression-based approach is non-constant variance. An important assumption of linear regression is that error terms have constant variance.

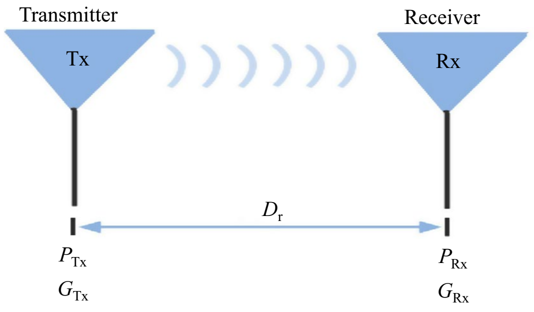


FIGURE 8 Friis' path-loss model

Non-constant variation in data is known as heteroscedasticity. The presence of heteroscedasticity in a model results in the erroneous estimation of standard error, R^2 score, and hypothesis testing results. We performed two different tests to determine the variance in the data: the standardized Breusch Pegan test and the non-constant variance test or Chi test, as shown in Figure 10. Both the standardized Breusch Pegan and non-constant variance test data reveal monotonic increases in variance for both the stationary and mobile conditions, resulting in strong heteroscedasticity that causes erratic overall prediction performance.

7.3 | *K*-means clustering

We selected *K*-means clustering as an unsupervised-learning approach and applied it to our data. This method provides a good overall cluster separation performance, as shown in Figure 11. Typically, the performance of a *K*-means model

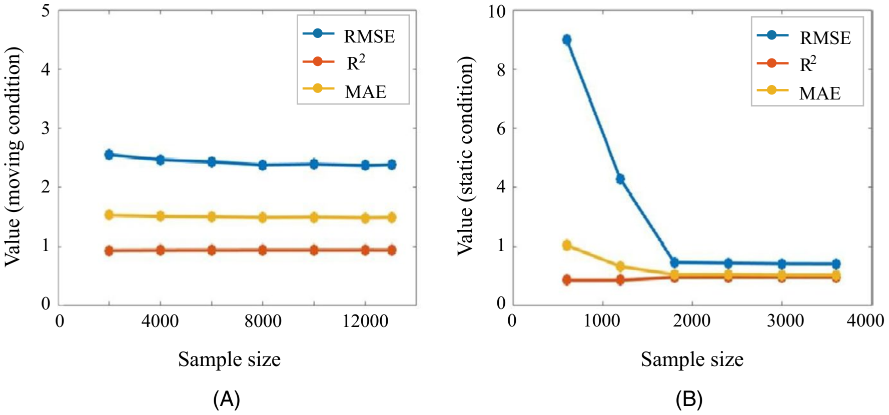


FIGURE 9 Linear regression accuracy performance

improves when additional initial centroids are considered, which allows a greater number of clustering models to run in parallel and consumes additional computation resources. From a computational perspective, K-means clustering is a function of $n*k*l$, where n is the sample size, k is the number of clusters, and l is the number of initial centroids. One can see that for a greater number of clusters ($k \geq 2$), achieving a better silhouette distance requires a greater number of initial centroids ($l \geq 1$). As shown in Figure 11, the best cluster separation performance is achieved when the number of clusters is set to four ($k=4$) with five or more ($l=5$) initial centroids. Therefore, the complexity is $20(n)$.

K-means clustering performs poorly in terms of identifying accurate signal levels. As shown in Figure 12, despite reasonable cluster separation performance, the accuracy of signal level prediction reaches a maximum of <40% for the stationary condition and <60% for the mobile condition when using the K-means clustering algorithm. Further, it is observed that improving accuracy costs significantly more computational resources based on increasing the sample size (n), number of clusters (k), and initial cluster centroids (l).

7.4 | Comparison of overall results

In this section, we discuss the overall results of the methods that were tested in this study. According to the comparisons in Table 2, the classical Friis equation-based path loss model and unsupervised-learning-based models yield very low accuracy rates. Additionally, the unsupervised-learning-based model has a very high computational cost compared to the other three approaches. Linear regression suffers from variance issues, which yields very inconsistent prediction performance. Although K-means clustering achieves good cluster separation, it incurs huge computational costs. Overall, our conventional LDA-based machine-learning SLPP method

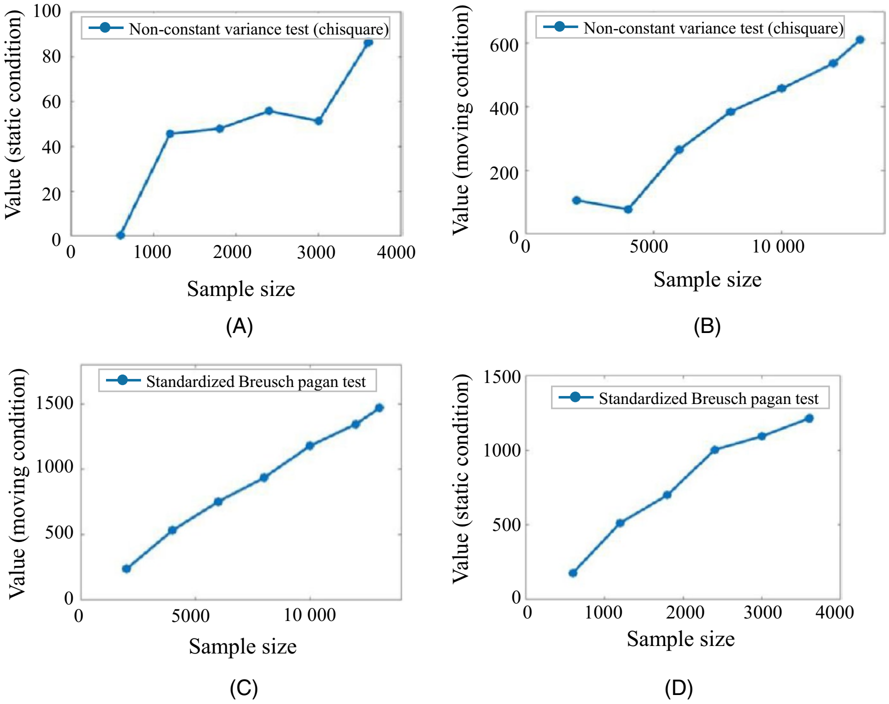
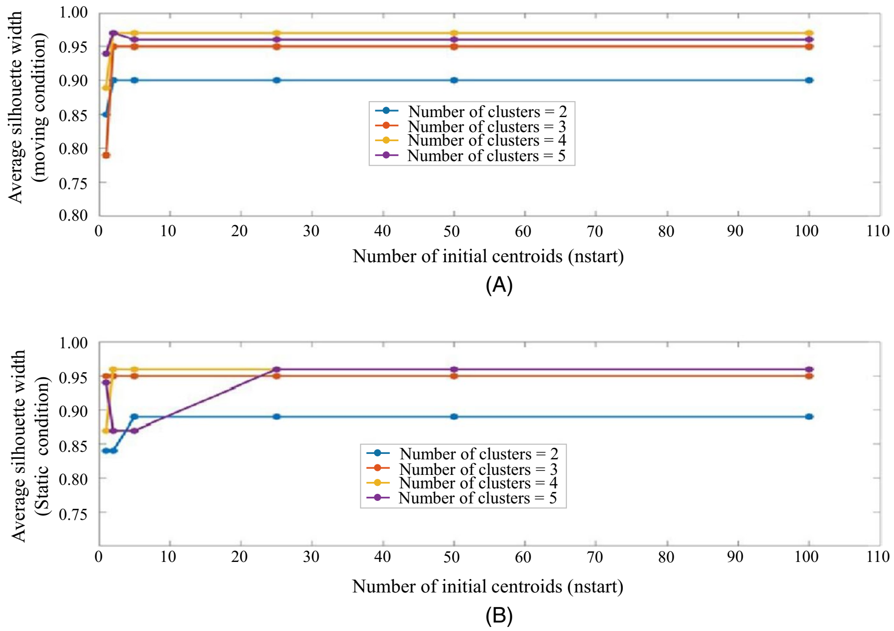


FIGURE 10 Linear regression variance performance

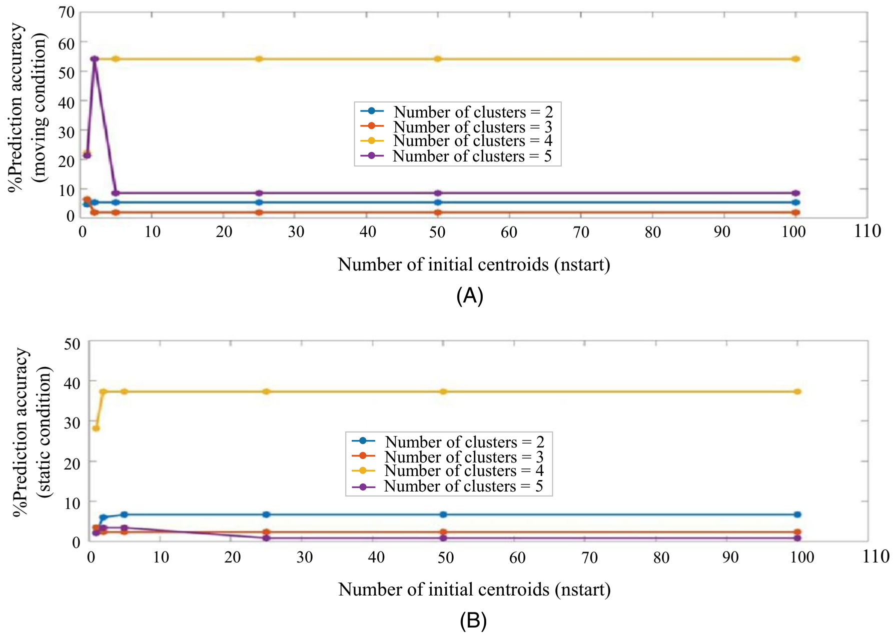
FIGURE 11 *K*-means clustering: cluster separation performance



(A)

(B)

FIGURE 12 *K*-means clustering: accuracy performance



(A)

(B)

outperforms all other approaches in terms of accuracy, performance stability, and computational complexity after solving the multicollinearity issue in the raw data.

8 | CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrated how a supervised classical machine-learning method can be applied to real-time network data collected from various network devices in multiple locations for both stationary and mobile users while considering location-based temporal weather data. We also demonstrated how these data can be used to predict the signal levels perceived by users based on the classical LDA approach, which was used to forecast the signal strength at a location according to multiple categories (ie, strong, good, moderate, or poor). Based on our analysis, it can be concluded that this prediction approach yields excellent accuracy, simplicity, speed, and resource efficiency. Furthermore, we discussed multicollinearity issue detection in raw data and its resolution. The experimental efficiency of the proposed model proves that in many cases, classical machine-learning approaches are not only effective, but also computationally inexpensive. In the future, a similar model will be used to predict the performance of various combinations of networks, such as 5G, Wi-Fi, satellites, and balloons. We will incorporate the data from various networks in the spatiotemporal plane to determine the effectiveness of our model for heterogeneous networks and identify machine-learning and deep-learning models that are efficient, scalable, and accurate.

ACKNOWLEDGEMENT

This research was funded in part by the Institute for Information & communications Technology Promotion (IITP) grant by the Korea government (MSIT No. 2018-0-00175) in the Horizon 2020 EU-Korea project 5G-ALLSTAR. This paper and the research behind it would not have been possible without exceptional support from Dr. Taesang Choi of ETRI. The first (Shubhabrata Mukherjee) and second (Taesang Choi) corresponding authors contributed equally to this paper.

ORCID

Shubhabrata Mukherjee https://orcid.org/0000-0002-9093-539X
Taesang Choi https://orcid.org/0000-0003-2831-811X

REFERENCES

1. Cisco, *Cisco annual internet report (2018–2023) white paper*, Mar. 2020, available at https://bit.ly/3dul4d2
2. H. Gebre-Amlak et al., *Protocol heterogeneity issues of incremental high-density WI-FI deployment*, in Proc. Int. Conf. Wired/Wireless Internet Commun. (Boston, MA, USA), June 2018, pp. 159–170.
3. Broadbandsearch, *Mobile vs. desktop usage (latest 2020 data)*, 2020, available at https://bit.ly/33Hc4Nd
4. N. Chakchouk, *A survey on opportunistic routing in wireless communication networks*, IEEE Commun. Surveys Tutorials **17** (2015), 2214–2241.
5. M. Zekri, B. Jouaber, and D. Zeghlache, *A review on mobility management and vertical handover solutions over heterogeneous wireless networks*, Comput. Commun. **35** (2012), 2055–2068.
6. D. Niyato and E. Hossain, *Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach*, IEEE Trans. Veh. Technol. **58** (2008), no. 4, 2008–2017.
7. Q. Zhiguo et al., *Multilevel pattern mining architecture for automatic network monitoring in heterogeneous wireless communication networks*, China Commun. **13** (2016), 108–116.
8. Anttonen, P. Ruuska and M. Kiviranta, *3GPP nonterrestrial networks: A concise review and look ahead*, 2019.
9. C. Qiu et al., *Spatio-temporal wireless traffic prediction with recurrent neural network*, IEEE Wireless Commun. Lett. **7** (2018), no. 4, 554–557.
10. J. Riihijarvi and P. Mahonen, *Machine learning for performance prediction in mobile cellular networks*, IEEE Comput. Intell. Mag. **13** (2018), no. 1, 51–60.
11. S. Mukherjee and C. Beard. *A framework for ultrareliable low latency mission-critical communication*, in Proc. Wireless Telecommun. Symp. (Chicago, IL, USA), 2017, pp. 1–5.
12. *Configuring location awareness rules for pulse secure client*, 2020, available at https://bit.ly/2wnnmtB
13. T. Tuglular, *Automatic enforcement of location aware user based network access control policies*, in Proc. WSEAS Int. Conf. Telecommun. Inform. (Istanbul, Turkey), May 27–30, 2008, pp. 49–54.
14. H. Viswanathan and P. E. Mogensen, *Communications in the 6G era*, IEEE Access **8** (2020), 57063–57074.
15. J. Crawshaw, *AI in Telecom Operations: Opportunities & Obstacles*, 2020, available at https://bit.ly/3hwyzJZ
16. Q. Huang et al., *Machine learning-based cognitive spectrum assignment for 5G URLLC applications*, IEEE Netw. **33** (2019), no. 4, 30–35.
17. S. M. Kala, M. P. K. Reddy and B. R. Tamma, *Predicting performance of channel assignments in wireless mesh networks through statistical interference estimation*, in Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (Bangalore, India), July 2015, pp. 1–6.
18. Y.-J. Cho et al., *AI-enabled wireless KPI monitoring and diagnosis system for 5G cellular networks*, in Proc. Int. Conf. Inf. Commun. Technol. Convergence (Jeju Island, Rep. of Korea), Oct. 2019, pp. 899–901.
19. *Expanding internet connectivity with stratospheric balloons*, available at https://x.company/projects/loon/
20. B. Barritt, V. Cerf and S. D. N. Loon. *Applicability to Nasa's next-generation space communications architecture*, in Proc. IEEE Aerospace Conf. (Big Sky, MT, USA), Mar. 2018, pp. 1–9.
21. C. Yue et al., *Link forecast: Cellular link bandwidth prediction in lTE networks*, IEEE Trans. Mob Comput. **17** (2017), no. 7, 1582–1594.
22. D. Liang et al., *Mobile traffic prediction based on densely connected CNN for cellular networks in highway scenarios*, in Proc. Int. Conf. Wireless Commun. Signal Process. (Xi'an, China), 2019, pp. 1–5.
23. N. Kato et al., *Optimizing space air-ground integrated networks by artificial intelligence*, IEEE Wirel. Commun. **26** (2019), no. 4, 140–147.
24. M. Mroue et al., *A neural network based handover for multirat heterogeneous networks with learning agent*, in Proc. Int. Symp. Reconfigurable Commun.-Centric Syst.-Chip (Lille, France), July 2018, pp. 1–6.
25. W. Huang et al., *Deep architecture for traffic flow prediction: deep belief networks with multitask learning*, IEEE Trans. Intell. Transp. Syst. **15** (2014), no. 5, 2191–2201.
26. Z. Nouir et al., *Supervised prediction for radio network planning tool using measurements*, in Proc. IEEE Int. Symp. Personal, Indoor Mobile Radio Commun. (Helsinki, Finland), Sept. 2006, pp. 1–5.
27. W. Jun et al., *CellPAD: Detecting performance anomalies in cellular networks via regression analysis*, in Proc. IFIP Netw. Conf. Workshops (Zurich, Switzerland), May 2018, pp. 1–9.
28. M. Porjazoski and B. Popovski. *Coverage predictions and performance analysis of metropolitan and cellular system based on IEEE 802.16*, in Proc. Int. Conf. Telecommun. Modern Satellite, Cable Broadcasting Services (Nis, Serbia), Sept. 2007, pp. 238–242.
29. L. Yan et al., *Machine learning based handovers for Sub-6 GHz and mmWave integrated vehicular networks*, IEEE Trans. Wireless Commun. **18** (2019), no. 10, 4873–4885.
30. Z. Md, F. NeiKato et al., *The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective*, IEEE Wirel. Commun. **24** (2016), no. 3, 146–153.
31. F. Tang et al., *An intelligent traffic load prediction-based adaptive channel assignment algorithm in SDN-IoT: A deep learning approach*, IEEE Internet Things J. **5** (2018), no. 6, 5141–5154.

32. J. Li et al., *Deep reinforcement learning-based mobility-aware robust proactive resource allocation in heterogeneous networks*, IEEE Trans. Cognitive Commun. Netw. **6** (2020), no. 1, 408–421.
33. M. Chen, W. Saad and C. Yin, *Echo-liquid state deep learning for 360 content transmission and caching in wireless VR networks with cellular-connected UAVs*, IEEE Trans. Commun. **67** (2019), no. 9, 6386–6400.
34. S. C. Pakhrin and D. R. Pant. *Multi-armed bandit learning approach with entropy measures for effective heterogeneous networks handover scheme*, in Proc. Int. Conf. Adv. Comput., Commun. Contr. Netw. (Greater Noida (UP), India), Oct. 2018, pp. 451–455.
35. Y. Yu, T. Wang and S. C. Liew, *Deep-reinforcement learning multiple access for heterogeneous wireless networks*, IEEE J. Sel. Areas Commun. **37** (2019), no. 6, 1277–1290.
36. *Network cell info: the ultimate network cell signal information tool*, available at https://m2catalyst.com/apps/network-cell-info
37. Wunderground, *Local weather forecast, news and conditions|weather underground*, available at https://www.wunderground.com
38. Wikipedia, *Multicollinearity*, available at https://en.wikipedia.org/wiki/Multicollinearity
39. G. James, et al., *Statistical Learning, An Introduction to Statistical Learning*, Springer, New York, NY, 2013. pp. 15–57.
40. C. Beard and W. Stallings, *Wireless Communication Networks and Systems*, Pearson, New Jersey, NJ, 2015.

AUTHOR BIOGRAPHIES

Shubhabrata Mukherjee received his MS degree in Electrical Engineering from the University of Missouri-Kansas City (UMKC), US, in 2017 and his BS degree in Electronics and Communication Engineering from West Bengal University of Technology, India, in 2009. He is currently working as a Graduate Intern for Network AI for the Intel Corporation at Intel HQ, Santa Clara. He previously worked for Project Loon of the Google X Laboratory (2017 to 2018). He is pursuing his PhD in Computer Networking and Communication Systems at UMKC. His research interests include applications of machine learning in wireless communication, ultra-reliable low-latency communication, and autonomous network management.

Taesang Choi received his MS degree and PhD in Computer Science and Telecommunications from the University of Missouri-Kansas City, US, in 1990 and 1995, respectively. In 1996, he joined ETRI, where he is currently a Principal Research Staff Member. He has been actively involved in the research and development of traffic engineering, traffic measurement and analysis, SDN/NFV management, and 5G network slice management. He has also actively contributed to various SDOs and open-source activities, such as IETF, ITU-T, ONF, and ONOS. He is currently acting as an ITU-T SG13 Question 6 Rapporteur and International IT Standardization Expert representing South Korea.

Md Tajul Islam received his MS degree in Computer Science from the University of Missouri-Kansas City (UMKC), US, in 2019 and BS degree in Electronics and Communication Engineering from Khulna University of Engineering and Technology, Bangladesh, in 2011. He worked at the Samsung R&D Institute of Bangladesh as a Software Engineer until mid 2015. He is currently a PhD student of Computer Networking and Communication Systems at UMKC. His research interests include optimized management of wireless networks, software-defined networking, and machine-learning applications in wireless communications.

Baek-Young Choi received her PhD in Computer Science and Engineering from the University of Minnesota, Twin Cities. She is an Associate Professor with the Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City. She was a Post-Doctoral Researcher with Sprint Advanced Technology Labs and a 3M McKnight Distinguished Visiting Assistant Professor at the University of Minnesota, Duluth. She has been a fellow with the US Air Force Research Laboratories Visiting Faculty Research Program and Korea Telecom-Advance Institute of Technology. Her research interests include the broad area of algorithm and system development for diverse types of networks, particularly software-defined networking, cloud computing, and IoT.

Cory Beard is an Associate Professor at the Department of Computer Science and Electrical Engineering at the University of Missouri-Kansas City, US, where he focuses on ultra-reliable low-latency cellular communication and its application to emergencies, public safety, air traffic control, and electric utility smart grid communications, as well as scheduling and service prioritization for wireless systems, sensor networks for home energy savings, and preemptive and queuing policies for emergency traffic in wireless cellular networks. He received his BS and MS degrees in Electrical Engineering from the University of Missouri, Columbia, US, in 1990 and 1992, respectively. He received his PhD in Electrical Engineering from the University of Kansas in 1999.

Seuck Ho Won received BS degrees in Clinical Pathology and Electrical Engineering from Kwangwoon University, Seoul, Korea, in 1985 and 1990, respectively, and his PhD in Electrical Engineering from Chungnam National University, Daejeon, Korea, in 2002. Since 1985, he has been a Clinical Pathologist at Sin-Chon General Hospital, Gyeonggi-do, Korea. Since 1990, he has been a Principal Engineer at ETRI, Daejeon, Korea. He was a research faculty member at Virginia Tech, US, in 2005. His research interests include information theory, error correction coding, MIMO, and beamforming, with an emphasis on mobile communications and broadcasting.

Sejun Song received his MS degree and PhD degrees in Computer Science and Engineering from the University of Minnesota, Twin Cities, US, in 1999 and 2001, respectively. He is an Associate Professor with the Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City. He and his research team have conducted research in the areas of trustworthy information and computing systems, and software such as resilient network and system management, software-defined networks, cloud computing auditability, mobile cloud computing, security, high availability, data storage, and embedded systems.