OVERVIEW STUDY OF MOBILE NETWORK TRAFFIC FOR BTS STATIONS

Hoang Van Thuc*, Pham Van Ngoc, Doan Thi Thanh Thao, Vu Chien Thang, Pham Thanh Nam, Mac Thi Phuong

University of Information and Communication Technology, Thai Nguyen University, Vietnam

ARTICLE INFORMATION

Journal: *Vinh University Journal of Science Natural Science, Engineering and Technology*
p-ISSN*: 3030-4563*
e-ISSN: *3030-4180*

Volume: 53
Issue: 4A
Correspondence:
hvthuc@ictu.edu.vn

Received: *03 July 2024*
Accepted: *30 August 2024*
Published: *20 December 2024*

Citation:
Hoang Van Thuc, Pham Van Ngoc, Doan Thi Thanh Thao, Vu Chien Thang, Pham Thanh Nam, Mac Thi Phuong (2024). Overview study of mobile network traffic for BTS stations.
Vinh Uni. J. Sci.
Vol. 53 (4A), pp. 5-14
doi: 10.56824/vujs.2024a076a

ABSTRACT

In recent years, Machine Learning (ML) has become a crucial and promising tool for forecasting and solving a wide range of complex problems. The rapid development of machine learning is closely linked to technological advancements and has also driven the growth of the AI community and open-source tools (e.g., TensorFlow, Keras, PyTorch, fast.ai). This enables researchers to deploy and apply machine learning algorithms more effectively. This paper provides an overview of mobile network traffic at BTS stations, conducted from a data-driven perspective, focusing on extracting and transforming data into information that serves production and business purposes within mobile networks, as well as describing the characteristics of user traffic. The authors used the Google Colab environment to analyze network time statistics to determine traffic in each area. Leveraging large volumes of information helps improve mobile network performance and address various issues (e.g., anomaly detection) that may impact network infrastructure. The study's findings contribute to addressing certain practical challenges in deployment, optimization, resource allocation, and energy savings for mobile networks.

Keywords: 5g traffic; base transceiver station; 5G BTS; 5G Traffic; 5G/BTS Traffic.

1. Introduction

Technology and data engineering have been and will continue to develop rapidly, driven by people's thirst for knowledge and attracting the attention of researchers across various fields such as machine learning, expert systems, and computer science. Numerous classification techniques have been proposed; however, no single classification approach has proven to be consistently optimal or more accurate than others [1], [2], [3].

This article provides an overview of mobile network traffic for transceiver stations, exploring their operating mechanisms as well as factors affecting network traffic. From this analysis, we evaluate and categorize traffic from NodeB/eNodeB stations using Vinaphone mobile network data from Thai Nguyen Telecommunications

[4], [5], [6]. The objective is to research machine learning models and algorithms that support the classification of base stations based on traffic patterns derived from mobile network data.

Research tools and languages that support data mining (such as Google Colab and Python) are installed and utilized for this project. Within the scope of this study, machine learning algorithms are applied to classify transceiver stations by traffic volume [7], [8], [9]. Relevant data for classifying broadcast stations includes total traffic and call setup success rates. The sample dataset comprises traffic information for mobile communication network transceiver stations across various regions [10], [11], [12].

Figure 1 shows the structural model of a base transceiver station.

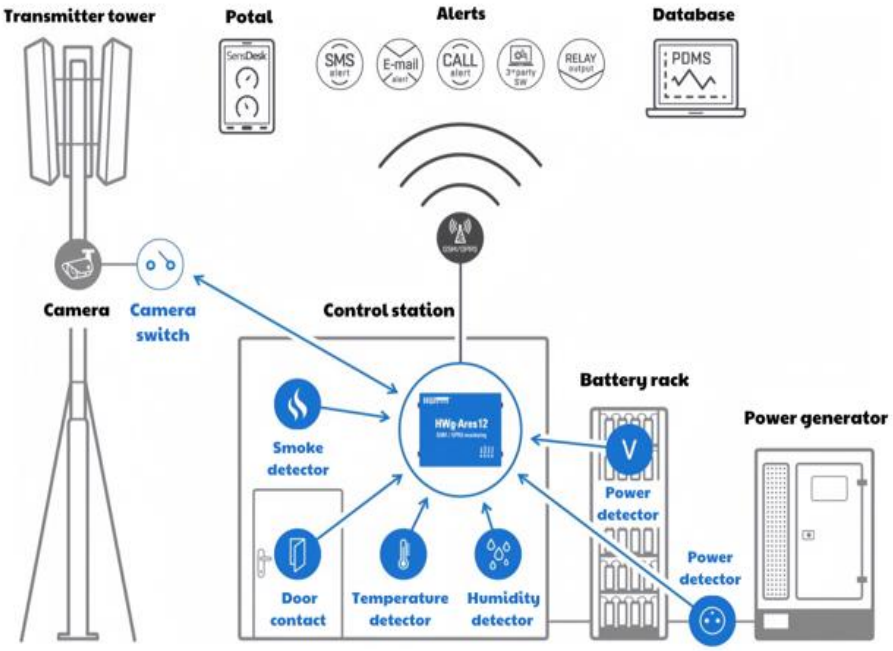

Figure 1: *BTS station structure model*

A base transceiver station (BTS) is a device that enables wireless communication between user equipment (UE) and the network. UE includes devices such as mobile phones (handheld devices), WLL phones, and computers with wireless Internet connections. The network can be based on any wireless communication technology, including GSM, CDMA, wireless local loop (WLL), Wi-Fi, WiMAX, or other wide-area network (WAN) technologies. BTS is also known as Node B in 3G networks or simply as a base station (BS). In LTE networks, the abbreviation eNB (evolved Node B) is commonly used, while in 5G networks, it is referred to as gNodeB [1].

A basic BTS includes:

 Transceiver (TRX): Responsible for transmitting and receiving signals to and from higher network elements.

 Combiner: Combines feeds from several base stations to be sent through a single antenna, thereby reducing the number of antennas that need to be installed.

 Power Amplifier: Amplifies the signal from the base station to enable efficient transmission through the antenna.

2. Operational mechanism and network traffic

2.1. Network operating mechanism

When users' mobile devices access the Internet, they send requests to mobile base stations. These requests are then relayed by transceiver stations to the radio network controller (RNC), which routes them into the core network and out to the intranet. Through the intranet, managers can gather statistics on the traffic of BTS stations to estimate the daily traffic at each base transceiver station [2].

Mobile network traffic refers to internal communication within a network where links to and from end nodes are wireless. The network is distributed across areas of land known as cells, with each area served by at least one fixed-location transceiver (usually three mobile sites or mechanical transceiver stations). These base stations provide the cell with network coverage that enables the transmission of voice, data, and other types of content. A typical cell uses a distinct set of frequencies from neighboring cells to avoid interference, ensuring guaranteed quality of service for each traffic flow [3].

2.2. Factors affecting network traffic

Several factors can affect network traffic during usage. While some of these factors are unavoidable, steps can be taken to minimize their negative effects on network performance. However, certain other factors can be completely addressed through essential equipment upgrades or effective network planning [4].

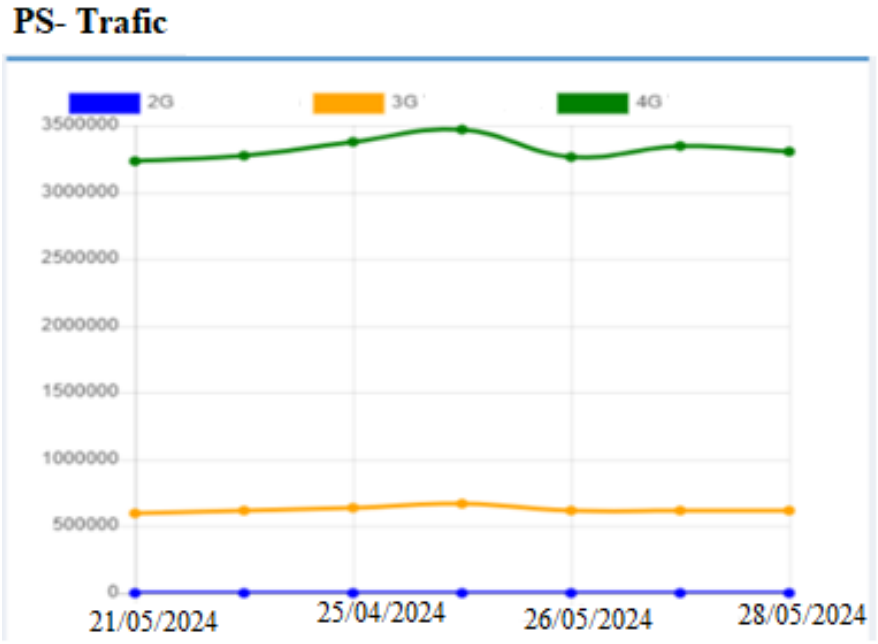

Figure 2: *Example model of mobile network traffic statistics*

Designing and constructing a database of station cells, known as the traffic database, was initially intended for statistical purposes; however, it did not effectively assist managers in categorizing stations based on traffic patterns.

3. Research model

This article introduces the Decision Forest (DF) model, an open-source machine learning framework built from the ground up using TensorFlow to construct predictive models. DF incorporates advanced machine learning algorithms designed to address supervised classification, regression, and ranking problems. The most commonly used algorithms in DF are Random Forests (RF) and Gradient Boosted Decision Trees (GBDT). Both of these are ensemble algorithms consisting of multiple "decision trees" though each algorithm uses unique techniques for implementation [5].

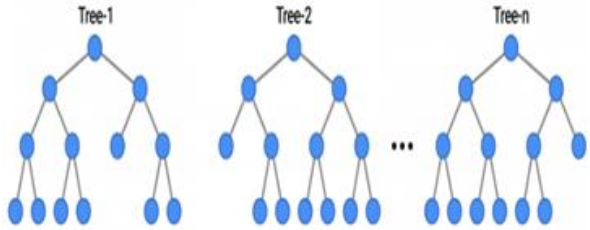

Figure 3*: Decision Forest (DF) model [3]*

3.1. Random Forest algorithm

The Random Forest (RF) algorithm, illustrated in Figure 4, is a supervised learning algorithm commonly used for classification and regression problems, as well as for model prediction. RF combines multiple decision trees, making it an ensemble method based on bootstrap bagging. In bagging, multiple decision trees are created, each from a different bootstrap sample of the training dataset. A bootstrap sample is a sample of the training dataset where each data point can appear multiple times, a method known as sampling with replacement [6], [7].

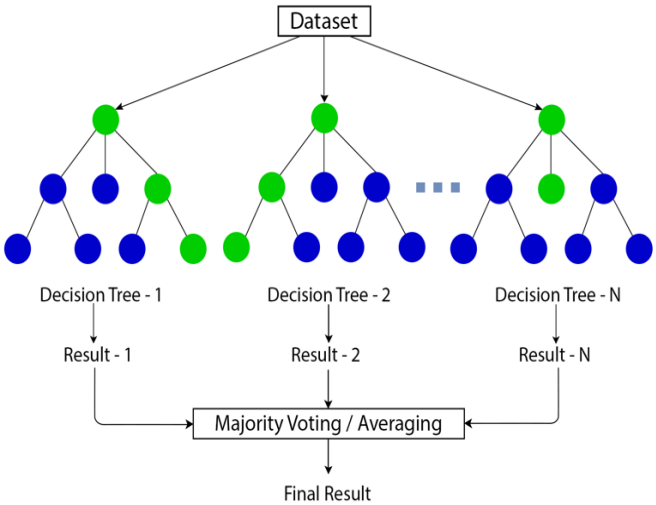

Figure 4: *Random Forest algorithm model [3]*

3.2. Gradient Boosted Decision Trees algorithm

Gradient Boosted Decision Trees (GBDT) is a machine learning technique that optimizes the predictive value of a model through successive steps in the learning process.

Each iteration of GBDT involves adjusting the coefficients, weights, or biases applied to each input variable to predict the target value, aiming to minimize the loss function - a measure of the difference between the predicted and actual target values. "Gradient" refers to the incremental adjustments made at each step, while "boost" accelerates the model's prediction accuracy toward an optimal level [8], [9], [10].

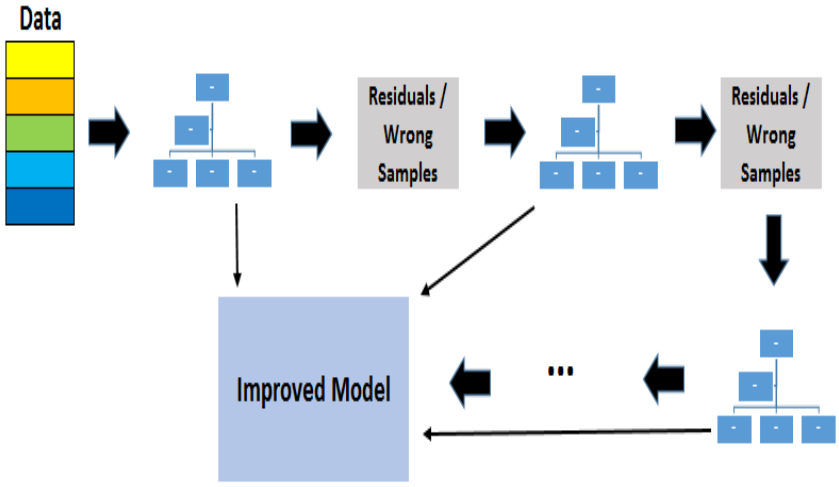

Figure 5: *Gradient boosting decision tree model [13]*

4. Experiment in Google Colab environment

The experiments were conducted in the Google Colab environment, utilizing TensorFlow and various machine learning libraries. Additional Python libraries used for computations include Pandas and Numpy [11], [12].

4.1. Experimental data

The network traffic dataset used for the experiments consists of 24 fields and 1000 records, aimed at evaluating the effectiveness of the model using the Random Forest algorithm. Key fields related to traffic, such as Traffic_Volume_UL_GB and Traffic_Volume_DL_GB, were weighted and labeled to enhance model performance. A reduced and detailed summary of the dataset fields is presented in Table 1.

Table 1: *Network traffic data set*

	Mean	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
IRATHO_SR	87.92	30.85	0.00	97.96	99.82	100.00	100.00
HSRate_via_Per	97.55	8.65	0.00	99.36	99.79	99.93	100.00
UDAT_Kbps	30721.87	8696.75	0.00	25535.90	31320.61	36454.97	64943.29
TraVol_UL_GB	2.24	2.90	0.00	0.71	1.42	2.68	38.01
TraVol_DL_GB	26.08	26.90	0.00	9.86	18.75	31.85	246.75
CMax_Throughput	31922.03	18243.68	0.00	16298.50	31392.50	46583.75	69771
EUTRAN	99.20	8.91	0.00	100.00	100.00	100.00	100.00
CDown_Avg_ Throughput	20.43	4.61	0.00	17.56	20.47	23.21	39.62

	Mean	Std	Min	25%	50%	75%	Max
IRHPS_Ratio	88.61	30.77	0.00	99.34	100.00	100.00	100.00
IRTHS	87.15	30.82	0.00	96.46	99.35	100.00	100.00
IeHS_total	96.68	17.35	0.00	99.94	100.00	100.00	100.00
UUAT_Kbps	2414.8 4	945.76	00.0	1718 .90	2392.82	3067.18	10840
CUp_Avg_Throughput	2.05	0.88	0.0 0	1.41	2.00	2.62	9.43
IRHL_toWPer	87.92	30.85	0.00	97.96	99.85	100.00	100.00
TDTV_GB	28.32	29.58	0.00	10.57	20.17	34.61	284.76
Downlink_Latency	21.18	12.00	0.00	15.77	18.61	23.01	169.26
CPDMax_Throughput	97.55	27.50	0.0 0	81.28	97.52	113.84	195.32
IFHPer	99.20	4.81	0.00	99.49	99.81	99.94	100.00
SD_all_Service	0.18	0.43	0.00	0.07	0.12	0.19	10.16
eSSRas_Per	99.80	3.19	0.00	99.91	99.96	99.98	100.00
RCESR_All_Service	99.83	3.16	0.00	99.92	99.98	100. 00	100.02
CSSRC_Per	99.73	3.19	0.00	99.83	99.93	99.97	100.00
INTRA_HOSR_ATT	497.96	731.75	0.00	112.00	286.50	571.25	9784.0
RBURD_Per	6.75	8.80	0.00	2.47	4.30	7.55	79.61

4.2. Model building and evaluation

The problem of classifying BTS stations based on traffic is described as follows: the input data for the model's training dataset comprises 70% of the original dataset (701 data points) with 24 different features. None of the 24 feature types are specified as input features, so all columns will be used as input features except for the labels. The features utilized by the model are displayed in the training logs and summarized in the model summary (model.summary).

The effectiveness of the model is evaluated based on accuracy and loss. A higher accuracy value, closer to 1, indicates better model performance; conversely, a value closer to 0 suggests poor predictive ability. Similarly, the model's loss represents the prediction accuracy; a lower loss value, closer to 0, indicates more accurate predictions. With the number of trees varying as $K = \{1, 51, 151, 201, 251, 300\}$, the accuracy and loss were averaged over 7 runs. The results are listed in Table 2 as follows:

Table 2: *Results of running the model with the RF algorithm*

No.	The number of trees	Accuracy	Loss
1	1	0.95	0.320205
2	51	0.974215	0.146839
3	101	0.977.77	0.105082
4	151	0.97855	0.101885
5	201	0.97855	0.10257
6	251	0.984245	0.101035
7	300	0.979947	0.09969

Based on Table 2, we observe that after each change in the number of trees, the RF model yields high accuracy in the initial experiment, reaching 94% with the first tree and increasing by 3% to 97% at the 300th tree. Similarly, the model's loss improved significantly, decreasing by 2.2% from 3.2% with the first decision tree to 0.9% with the last tree.

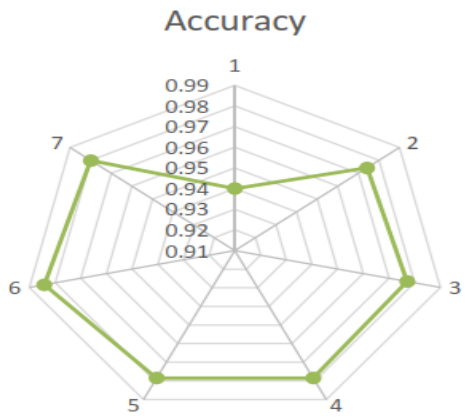

Figure 7: *Accuracy of the RF model in the first experiment*

Figures 7 and 8 illustrate the accuracy and loss of the model. As shown in the figures, the measurements gradually increase with each layer; the closer to the center, the lower the measurement value, and vice versa. In Figure 7, the model's accuracy reached approximately 94% in the first experiment and gradually increased in subsequent trials, ultimately reaching nearly 98% (97.99%) in the final evaluation.

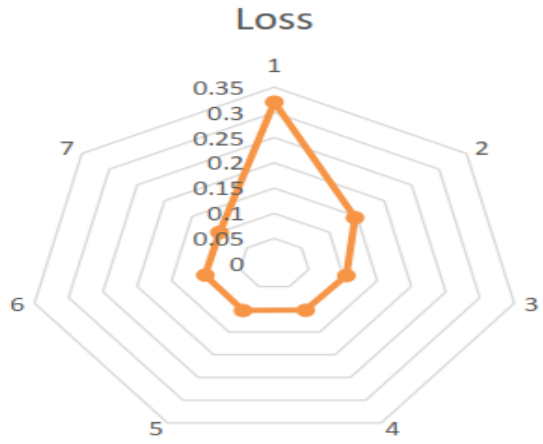

Figure 8: *Loss of the RF model in the first experiment*

Table 3 presents the station classification results of the two models with fixed input parameters. For the model utilizing the GBDT algorithm, the features that were re-selected as more important than the others during model construction include CellUpMax, TVU, and TDVT.

In Table 3 and Figure 9, we observe that the RF model provides more accurate prediction results than the GBDT model in most variations of the number of trees. In nearly all experiments, the accuracy of the RF model remained stable at an average level of about

94%, gradually increasing in subsequent decision trees. In the fourth experiment, the GBDT model achieved a significant accuracy of approximately 88%; however, this accuracy tended to decrease afterward, with the final decision tree yielding only about 72%.

Table 3: *Comparison of the accuracy of two algorithms RF and GBDT*

No.	Algorithm	K						
		1	51	101	151	201	251	300
1	RF	0.9855	0.9841	0.9826	0.9841	0.9855	0.9768	0.9918
	GBDT	0.8335	0.8182	0.7879	0.8030	0.8182	0.8030	0.8030
2	RF	0.9262	0.9826	0.9897	0.9916	0.9922	0.993	0.9943
	GBDT	0.7932	0.8135	0.8145	0.8208	0.8265	0.8417	0.8548
3	RF	0.9456	0.9521	0.955	0.9555	0.9731	0.9815	0.9852
	GBDT	0.8521	0.8337	0.8285	0.819	0.8081	0.8057	0.803
4	RF	0.9927	0.9844	0.9717	0.9711	0.9709	0.9567	0.9514
	GBDT	0.8849	0.8647	0.8061	0.8056	0.7589	0.7533	0.7246
5	RF	0.9059	0.9061	0.9275	0.9651	0.9755	0.9807	0.9888
	GBDT	0.8339	0.8304	0.8289	0.8254	0.8163	0.812	0.8007
6	RF	0.9208	0.923	0.951	0.9573	0.9698	0.9753	0.9895
	GBDT	0.839	0.8238	0.823	0.8198	0.7918	0.7786	0.7711
7	RF	0.9156	0.941	0.9473	0.955	0.9711	0.9837	0.9916
	GBDT	0.8655	0.8375	0.8267	0.7969	0.7952	0.7837	0.7761

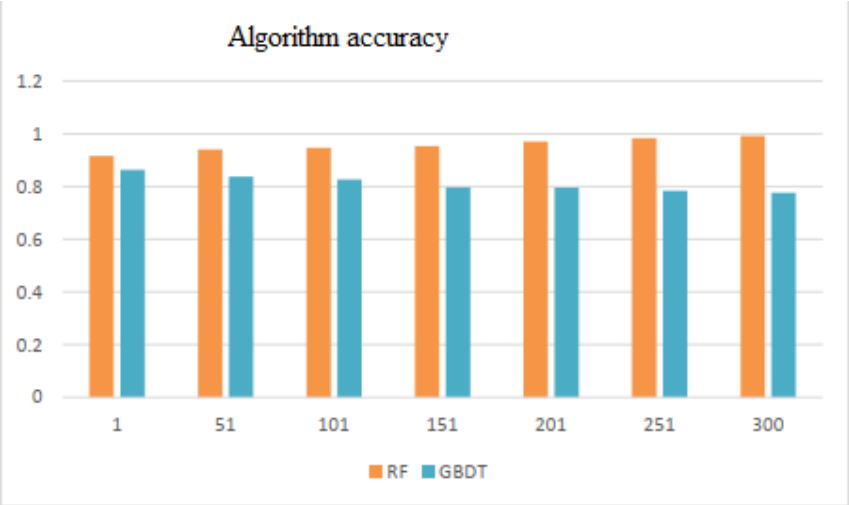

Figure 9: *Compare the accuracy of two algorithms RF and GBDT in the 7th run*

5. Conclusion

This article presents research findings on mobile network traffic analysis for BTS stations using the TensorFlow-Decision Forest (TF-DF) experimental model. The research process involved processing a dataset and applying it to the model, followed by an evaluation of the model's performance. Experimental results demonstrate high

prediction accuracy, with significant improvements in loss measures due to the integration of decision trees within the model. The article explores the theoretical foundations of mobile network traffic at BTS stations and reviews various algorithms addressing attribute selection challenges, with a focus on the TensorFlow-Decision Forest model and the Random Forest algorithm.

Based on these insights, the author proposes methods to improve feature labeling to optimize classification algorithms, especially for handling high-dimensional data. To validate the enhanced model's effectiveness, experiments were conducted using a network traffic dataset. The results indicate that the Decision Forest model utilizing the Random Forest algorithm achieves superior accuracy, suggesting it could be a viable option for application developers seeking reliable data classification solutions. Through these contributions, this research aims to address specific challenges in data mining and classification, potentially providing valuable insights for both the broader field of data analysis and particular applications.

Acknowledgments: This article is the result of a university-level project with code ĐH2024-TN07-03, funded by the University of Information and Communications, Thai Nguyen University.

REFERENCES

[1] P. Yu, Q. Yang, F. Fu, and K. S. Kwak, "Inter-cell cooperation aided dynamic base station switching for energy-efficient cellular networks," In *Proc. IEEE Asia-Pacific Conf. on Communications*, pp. 159-163, 2012. DOI: 10.1109/APCC.2012.6388122

[2] T. Adhikary, A. K. Das, M. A. Razzaque, M. O. Rahman, and C. S. Hong, "A distributed wake-up scheduling algorithm for base stations in green cellular networks," In *Proc. ACM Conf. on Ubiquitous Information Management and Communication*, p. 120, 2012. DOI: 10.1145/2184751.2184888

[3] J. Wu, Y. Zhang, M. Zukerman, and E. K. N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 803-826, 2015. DOI: 10.1109/COMST.2015.2403395

[4] J. Huang, Z. Zhong, and H. Huo, "A dynamic energy-saving strategy for green cellular railway communication network," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, pp. 1-13, 2015. DOI: 10.1186/s13638-015-0317-2

[5] J. Zheng, Y. Cai, X. Chen, R. Li, and H. Zhang, "Optimal base station sleeping in green cellular networks: A distributed cooperative framework based on game theory," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4391–4406, 2015. DOI: 10.1109/TWC.2015.2420233

[6] M. Kulin, T. Kazaz, E. De Poorter, and I. Moerman, "A survey on machine learning-based performance improvement of wireless networks: PHY, MAC and network layer," *Electronics*, 2021. DOI: 10.3390/electronics10030318

[7] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Trans. Netw*, 2017. DOI: 10.1109/TNET.2016.2623950

[8] H. D. Trinh, A. F. Gambiny, L. Giupponi, M. Rossiy, and P. Dini, "Mobile traffic classification through physical control channel fingerprinting: A deep learning approach," *IEEE Access,* vol. 7, pp. 152187-152201, 2019. DOI: 10.1109/ACCESS.2019.2947742

[9] S. Dujardin, D. Jacques, J. Steele, and C. Linard, "Mobile phone data for urban climate change adaptation: Reviewing applications, opportunities and key challenges," *Sustainability,* 2020. DOI: 10.3390/su12041501

[10] P. Muñoz, R. Barco, E. Cruz, A. Gómez-Andrades, E. J. Khatib, and N. Faour, "A method for identifying faulty cells using a classification tree-based UE diagnosis in LTE," *EURASIP Journal on Wireless Communications and Networking*, 2017. DOI: 10.1186/s13638-017-0914-3

[11] D. Wu, Z. Zhang, S. Wu, J. Yang, and R. Wang, "Biologically inspired resource allocation for network slices in 5G-enabled Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9266–9279, 2019. DOI: 10.1109/JIOT.2018.2888543

[12] B. Han, V. Sciancalepore, D. Feng, X. Costa-Perez, and H. D. Schotten, "A utility-driven multi-queue admission control solution for network slicing," In *Proc. IEEE INFOCOM,* pp. 55–63, Apr. 2019. DOI: 10.1109/INFOCOM.2019.8737517

TÓM TẮT

NGHIÊN CỨU TỔNG QUAN VỀ LƯU LƯỢNG MẠNG DI ĐỘNG CHO TRẠM BTS

Hoàng Văn Thực, Phạm Văn Ngọc, Đoàn Thị Thanh Thảo, Vũ Chiến Thắng, Phạm Thành Nam, Mạc Thị Phượng

Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên, Việt Nam

Ngày nhận bài 03/7/2024, ngày nhận đăng 30/8/2024

Trong những năm gần đây, Học máy (Machine Learning - ML) đã trở thành một công cụ quan trọng và đầy hứa hẹn trong việc dự báo và giải quyết nhiều vấn đề phức tạp. Sự phát triển nhanh chóng của học máy gắn liền với sự tiến bộ của công nghệ và cũng thúc đẩy sự phát triển của cộng đồng AI cùng các công cụ mã nguồn mở (ví dụ: TensorFlow, Keras, PyTorch, fast.ai). Điều này giúp các nhà nghiên cứu triển khai và áp dụng các thuật toán học máy một cách hiệu quả hơn. Bài báo này tổng quan về lưu lượng mạng di động cho các trạm BTS, được thực hiện theo hướng dữ liệu, tập trung vào việc khai thác và chuyển đổi dữ liệu thành thông tin phục vụ sản xuất kinh doanh trong mạng di động, cũng như mô tả đặc điểm lưu lượng truy cập của người dùng. Nhóm tác giả đã sử dụng môi trường Google Colab để phân tích các số liệu thống kê về thời gian của mạng, nhằm xác định lưu lượng tại từng khu vực. Việc khai thác một lượng lớn thông tin giúp cải thiện hiệu suất mạng di động và giải quyết nhiều vấn đề (ví dụ: phát hiện bất thường) có thể ảnh hưởng đến cơ sở hạ tầng mạng. Kết quả nghiên cứu trong bài báo đã góp phần nhỏ vào việc giải quyết các vấn đề liên quan đến triển khai, tối ưu hoá, phân bổ tài nguyên và tiết kiệm năng lượng cho mạng di động trong thực tế.

Từ khóa: Lưu lượng 5G; trạm thu phát sóng gốc; BTS cho mạng 5G; lưu lượng 5G/ BTS; lưu lượng truy cập 5G.