Proceedings of Machine Learning Research 319:232–243, 2026 IndabaX Nigeria 2026

Advanced Machine Learning Models for Network Traffic Prediction and Management

Ayodele E. Awokoya AWOKOYAAE@MCU.EDU.NG
Department of Computer Science,
McPherson University, Nigeria
Department of Computer Science,
University of Ibadan, Nigeria

Adeduro Oladapo ADEDUROOO@MCU.EDU.NG
Department of Computer Science,
McPherson University, Nigeria

Kayode A. Oladapo OLADAPOKA@MCU.EDU.NG
Department of Computer Science,
McPherson University, Nigeria

Kayode O. Olumurewa OLUMUREWAKO@MCU.EDU.NG
Department of Computer Science,
McPherson University, Nigeria

Rejoice I. Omoyeni REJOICE.I.OMOYENI@GMAIL.COM
Department of Computer Science,
McPherson University, Nigeria

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

Network traffic prediction helps institutions maximise performance while strengthening security and achieving sustainable development goals, better infrastructure stability, and lower energy costs. As internet usage expands, reliable traffic predictions are essential for quality service and reduced operational expenses. The university encounters critical network congestion during peak hours, degrading user satisfaction and increasing costs. This study employs Random Forest, Long Short Term Memory (LSTM), and Support Vector Regression (SVR) to forecast network traffic volume at the university. Wireshark was used to capture a full month of campus network traffic; data preprocessing included feature extraction and normalization. Three models were trained on real world traffic data and evaluated across short term (hourly), medium term (daily), and long term (weekly) predictions. Across all horizons, LSTM consistently achieved lower RMSE and MAE than Random Forest and SVR. LSTM's ability to capture temporal dependencies and long range patterns makes it well suited for network traffic prediction in resource planning. Implementing LSTM based forecasting can help the university improve network performance, user experience, and progress toward environmental sustainability goals.

Keywords: network traffic prediction, machine learning, LSTM, random forest, support vector regression

1. Introduction

Network traffic prediction is fundamental to modern network administration. It helps administrators allocate resources more effectively, strengthen security, and work toward the Sustainable Development Goals (SDGs), particularly those related to resilient infrastructure and lower energy consumption (Lohrasbinasab et al., 2022). As global internet usage continues to rise, understanding and forecasting traffic patterns becomes critical not only for maintaining quality of service, but also for meeting sustainability commitments (Alsheikh et al., 2014). SDG 9, which calls for building resilient infrastructure and fostering innovation, sits squarely at the intersection of these concerns. Machine learning techniques, in particular, offer substantial improvements in operational efficiency by enabling smarter network management and reducing maintenance costs (Singh et al., 2020). At the university, network congestion during peak hours has become a recurring challenge. When usage spikes, the user experience suffers, and operating costs climb. Many institutions around the world face similar difficulties: they struggle to predict and control traffic fluctuations, and end up managing resources reactively, which complicates both software planning and capacity provisioning (Aouedi et al., 2022). This research evaluates three machine learning approaches, Random Forest, Long Short Term Memory (LSTM), and Support Vector Regression (SVR), to forecast network traffic at the university across three different time horizons: hourly, daily, and weekly. Our aim is to provide solution oriented guidance that network administrators can use to improve system performance and long term resilience. The remainder of the paper is structured as follows. Section II reviews relevant literature, highlighting the shortcomings of existing approaches. Section III describes the data collection, preprocessing, and model implementation pipeline. Section IV presents and discusses the experimental results. Finally, Section V offers conclusions and directions for future work.

2. Literature Review

2.1. Overview of Network Traffic Prediction

Network traffic prediction has evolved sharply over the past decade, driven by the need for smarter load management and more robust cybersecurity. According to Cisco's global forecast, IP traffic volumes have been climbing steeply, making accurate predictive models essential for capacity planning (Cisco Systems, Inc., 2019). At the same time, threats like Distributed Denial of Service attacks demand that models cope with highly irregular traffic signatures (Buczak and Guven, 2015). Traditional Poisson based assumptions rarely hold; real network traffic often exhibits self similarity and long range dependence, as noted by (Alsheikh et al., 2014) in their survey on machine learning in wireless sensor networks (Alsheikh et al., 2014). These properties were first seriously examined in the late 1980s and early 1990s, spurring the development of fractal models. While those models could capture self similar behaviour reasonably well, they proved computationally expensive and struggled to adapt to rapidly changing conditions (Joshi and Hadi, 2015). Over time, researchers blended classical time series techniques with soft computing and neural methods to handle the erratic spikes that older models missed. (Fadlullah et al., 2017) surveyed how deep learning can be harnessed for intelligent network control and underlined its capacity to learn from raw traffic data without hand crafted features (Fadlullah et al., 2017). Still, deterministic tools like network calculus, which offers theoretical performance bounds, tend to falter when traffic dynamics shift quickly, an issue (Jiang, 2012) acknowledged in his overview of stochastic network calculus.

2.2. Causes of Network Congestion

Network congestion arises when the volume of data packets exceeds the network's forwarding capacity, disrupting smooth communication. The primary culprits are bandwidth limitations, network complexity, latency, and jitter (El Naqa et al., 2019). Narrow bandwidth and convoluted topologies make resource allocation unpredictable, while latency, the delay between transmission and reception, hurts real time applications such as VoIP and video conferencing. Jitter, or variation in packet delay, further degrades quality by causing packets to arrive out of order. In the context of a university campus, these factors often combine during peak usage windows, producing the very slowdowns reported by users.

2.3. Evolution of Network Traffic Model

Early modelling efforts leaned on M/M/1 queueing theory, which assumes Poisson arrivals and exponential service times. The mathematics was tractable, but the assumptions rarely matched the bursty nature of modern traffic (Bolch et al., 2006). Markov chain models offered a finer grained view by describing state transitions, yet they soon encountered the "state explosion" problem as network sizes grew (Bolch et al., 2006). Other methods, such as stop and go models, worked acceptably for highly periodic flows but could not generalise to the diverse traffic mixes seen today (Joshi and Hadi, 2015). As a result, the research community turned toward machine learning, which shows more flexibility in teasing out patterns from complex, noisy data.

2.4. Machine Learning Techniques for Network Traffic Prediction

Machine learning (ML) has become a compelling alternative to conventional analytical models. ML algorithms can discover intricate historical patterns without being explicitly programmed for each task. Supervised methods learn from labelled data pairs, unsupervised methods mine unlabelled data for hidden structure, and reinforcement learning refines behaviour through trial and error (El Naqa et al., 2019). For a broader overview of machine learning paradigms, including Bayesian methods, (Vanneschi and Silva, 2023) provide foundational definitions, though their treatment is not specific to time series prediction. Several surveys have catalogued how these strategies apply to network traffic. Singh et al. (2020) for example, proposed a machine learning based sub slicing framework in a sustainable 5G context, showing that accurate predictions can directly support resource efficiency goals. Others have validated Gradient Boosting or simpler regressors, though many prior studies lacked extensive real world validation, leaving questions about practical deployment unanswered (Ponmalar et al., 2022). Klaine et al. (2017) surveyed machine learning applications in self organizing cellular networks and highlighted the persistent challenge of data scarcity for model training.

2.5. Selected Machine Learning Models

This research focuses on three models that offer distinct strengths for traffic forecasting.

1. 1. **Random Forest** is an ensemble of decision trees that reduces overfitting and reliably captures non linear relationships (Liu et al., 2012). A practical benefit is its built in feature importance ranking, which helps identify which network characteristics most influence traffic volume.

2. 2. **Long Short Term Memory (LSTM)** is a recurrent neural network designed to learn long range temporal dependencies (Van Houdt et al., 2020). Through its gated memory cells, it can retain information across many time steps, an advantage when forecasting time series like network traffic, where what happened hours ago shapes the current load.

3. 3. **Support Vector Regression (SVR)** uses the kernel trick to handle non linear regression; it finds a function that deviates from the true values by at most a small margin while remaining as flat as possible (El Naqa et al., 2019). SVR can model non linear traffic relationships, but its performance heavily depends on careful kernel and parameter selection.

2.6. Comparative Studies

Several comparative efforts have shaped the design of this study. Lu and Yang implemented an LSTM based predictor for network traffic and showed that it outperformed conventional neural networks, but their approach required careful data preprocessing to handle noisy measurements (Lu and Yang, 2018). In contrast, Mohammed presented a case study of campus network analysis using traditional statistical tools, providing a baseline that our work extends with ML (Mohammed et al., 2013). Mo (2015) explored particle swarm optimization for traffic prediction and reported modest accuracy improvements, though without the temporal depth that recurrent architectures bring. Other surveys have underscored the need for interpretable and scalable systems that function across different network environments and evolving traffic mixes (Ponmalar et al., 2022),(Afolabi et al., 2018). Similarly, Foukas et al. (2017) provided an early survey of 5G network slicing challenges, emphasizing the need for dynamic resource allocation. Wang et al. (2014) proposed an optimal slicing strategy for SDN based smart home networks, but highlighted scalability challenges that remain when extending such approaches to campus wide environments (Wang et al., 2014). Our study directly addresses this gap by evaluating three models on a real university network and reporting performance across multiple prediction horizons.

3. Methodology

This section details how the data were collected, how the predictive models were built, and what metrics were used to assess them.

3.1. Data Collection

We began by mapping the university's network infrastructure. The campus serves roughly 1,550 students and 300 staff through a combination of wired and wireless connections, with

Starlink providing primary internet access at 100–150 Mbps. To capture representative traffic, we connected Wireshark, a network protocol analyser Mohammed et al. (2013), to a mirror port on the core router, ensuring that normal operations were not disturbed. The capture recorded timestamps, communicating device identifiers, protocol types (TCP, UDP, HTTP, DNS), packet sizes, and TCP header fields. The monitoring period lasted a full month, from 7 June 2024 to 28 June 2024, covering both high and low activity periods.

3.2. Ethical Consideration and Data Anonymization

Because network traffic can contain sensitive information, strict ethical protocols were followed. All IP addresses were anonymized using a one way hash function, and no encryption breaking was attempted. Application layer payloads were stripped of personally identifiable information in accordance with established privacy practices.

3.3. Data Preprocessing

The raw capture files were cleaned by removing malformed or incomplete packets and filtering out non IP traffic such as ARP and ICMP. We then extracted features relevant to volume forecasting: hourly byte and packet counts, average packet size, protocol shares (TCP, UDP, HTTP), source and destination port frequencies, and temporal markers (hour of day, day of week). All numerical features were normalized to the range [0, 1] using min max scaling; categorical variables were one hot encoded.

3.4. Machine Learning Model Implementation

The forecasting problem was framed as a multi step regression task: using historical traffic features, predict the total byte volume for the next hour, day, or week. This consistent target allowed a direct comparison of the three models.

i. **Features and target:** The input features included hour of day, day of week, a weekend flag, lagged byte volumes at 1 , 2 , and 24 hour offsets, protocol share percentages, and average packet size. The target was the total byte count in the upcoming prediction window.

ii. **Random Forest:** A Random Forest regressor was implemented using scikit learn with 100 trees, a maximum depth of 20, and the sqrt feature per split policy (Liu et al., 2012). This ensemble method handles non linear interactions well and offers built in feature importance scores.

iii. **Support Vector Regression (SVR):** An SVR with a radial basis function kernel, C=1.0, gamma=‘scale’, and epsilon=0.1 was built, also in scikit learn. SVR can model non-linear patterns, though training time increases noticeably on larger datasets (El Naqa et al., 2019).

iv. **Long Short Term Memory (LSTM):** The LSTM was constructed with TensorFlow/Keras. The input consisted of the previous 24 time steps of features; a single LSTM layer with 100 units, a dropout of 0.2, and recurrent dropout of 0.2 was followed by a dense layer with a linear activation. The model was compiled with mean squared error loss and the Adam optimizer (Van Houdt et al., 2020), (Lu and Yang, 2018). Training stopped early if validation loss did not improve for 10 consecutive epochs.

v. **Baseline model:** A naïve persistence model that simply repeats the last observed volume was included as a sanity check. Comparing against this trivial forecast shows whether the more complex models are extracting meaningful patterns.

3.5. Model Training and Validation

The dataset was divided chronologically: the first 70% of days were used for training and the remaining 30% for testing; no shuffling was applied to preserve temporal order. All models were evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R squared (R^2), and Mean Absolute Percentage Error (MAPE). To account for randomness in weight initialization and data ordering, each experiment was repeated five times, and results are reported as mean±standard deviation.

3.6. Additional Data Collection Via survey

Complementing the captured traffic data, a Google Forms survey collected feedback from students and staff on their internet experience. Questions covered perceived slow speed periods, online activities, and productivity impacts. In total, 108 responses were received; 68% of respondents reported encountering slow internet at least occasionally, most frequently between midday and 5:00 PM, with weekends also highlighted.

3.7. Implementation Environment

All experiments were conducted on a workstation featuring an Intel Core i7 12700H processor, 32 GB of RAM, and an NVIDIA RTX 3060 GPU (used only for LSTM training). The software stack included Python 3.10, TensorFlow 2.12, scikit learn 1.3.0, pandas 2.0.3, and Wireshark 4.0.7. The captured pcap files and the processed feature sets can be made available to researchers upon request, subject to the university's data sharing policy.

4. Results and Discussion

4.1. Interpretation of Survey Results

The Google Forms survey gathered 108 responses from students and staff at the university. Slow internet speeds were reported by 68% of respondents at least occasionally; 31.4% pointed specifically to the midday to 5:00 PM window as the worst period, and 29.6% reported similar problems during weekends. Over half of the respondents (57.9%) stated that sluggish connectivity directly hindered their academic or work productivity. These self reported patterns map neatly onto the measured traffic peaks discussed in the next section.

4.2. Data Analysis and Visualization

A detailed look at the captured traffic reveals the rhythms of campus usage. Figure 1 shows a clear double peak distribution in packet sizes: a tight cluster of small control packets

Table 1: RMSE, MAE, R^2, and MAPE for Random Forest, LSTM, SVR, and the persistence baseline across hourly, daily, and weekly forecasts.

Model	Metric	Hourly	Daily	Weekly
Random Forest	RMSE	2.58	3.12	3.61
	MAE	1.84	2.25	2.58
	R^2	0.84	0.79	0.73
	MAPE	10.20%	12.50%	15.10%
LSTM	RMSE	2.07	2.84	3.27
	MAE	1.52	2.04	2.36
	R^2	0.89	0.84	0.78
	MAPE	8.40%	10.60%	13.20%
SVR	RMSE	2.76	3.18	3.73
	MAE	1.95	2.34	2.73
	R^2	0.81	0.76	0.70
	MAPE	11.50%	13.70%	16.80%
Baseline	RMSE	3.62	3.93	4.26
	MAE	2.85	3.12	3.43
	R^2	0.68	0.62	0.55
	MAPE	15.30%	17.60%	20.90%

(40–60 bytes) and a larger cluster near the Ethernet MTU of 1500 bytes. This profile is typical of educational networks, where chatty protocols and large data transfers coexist.

TCP dominates overall traffic volume, accounting for about 74% of total bytes during weekdays, while UDP—mainly voice and real time conferencing—carries the remainder. The hourly breakdown is even more telling. Weekday traffic begins climbing around 08:00, plateaus between 10:00 and 12:00 as lectures and online labs intensify, dips slightly during the lunch hour, and then reaches its absolute daily peak from 14:00 to 17:00. This afternoon surge, which coincides exactly with the survey respondents' complaints, likely reflects the overlap of academic activities and entertainment use. Weekend traffic is lighter overall but exhibits a broad hump from roughly 12:00 to 20:00, driven by video streaming and social media. Packet length distributions further differentiate the two main transport protocols. TCP shows a wide tail, reaching 1500 bytes, while UDP packets cluster tightly around 100–200 bytes, as illustrated in Figure 3. These patterns are consistent with those reported in prior campus network studies (Mohammed et al., 2013).

4.3. Overall Model Performance

LSTM consistently achieved the lowest RMSE and MAE, followed by Random Forest. The persistence baseline, which simply repeats the last observed value, posted the highest errors across all horizons, confirming that the other models are extracting non trivial temporal patterns. SVR performed better than the baseline but fell short of the tree based and

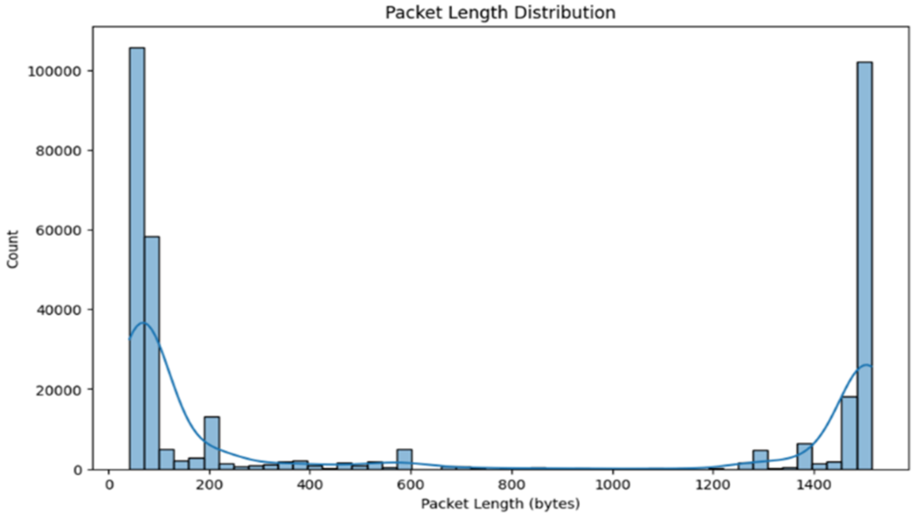

Figure 1: Distribution of packet sizes

recurrent models, particularly on the weekly horizon, where its R^2 dropped noticeably. The standard deviations indicate that LSTM's advantage is stable across runs, whereas SVR exhibited greater sensitivity to parameter choices (El Naqa et al., 2019). Paired t tests between LSTM and Random Forest showed that LSTM's improvements on hourly and daily forecasts were statistically significant ($p < 0.05$); for the weekly horizon the gap narrowed and was not consistently significant across all runs.

4.4. Feature Importance and Interpretability

Random Forest's built in feature importance scores offer insight into what drives the predictions. The most influential features were the 1 hour and 24 hour lagged traffic volumes, followed closely by the hour of day indicator. Protocol share features contributed modestly, while source port frequencies had negligible impact. This ranking aligns with intuition: recent behaviour and the time of day are strong predictors of near term load in a campus environment (Liu et al., 2012).

LSTM's predictions for a sample week are overlaid on the actual traffic in Figure 5. The model tracks the morning ramp up and the afternoon peak reasonably well, though it occasionally underestimates extreme spikes during unusual events. A shaded error band around the forecast shows that uncertainty widens during volatile periods, which is expected.

4.5. ANOVA Results

An analysis of variance (ANOVA) across the four models confirmed significant differences in prediction accuracy for all forecasting horizons ($p < 0.05$). Post hoc comparisons indicated that LSTM outperformed Random Forest and SVR on hourly and daily tasks, while the pairwise differences between Random Forest and SVR were not always significant for the weekly horizon. These statistical tests simply reinforce the pattern visible in Table 1

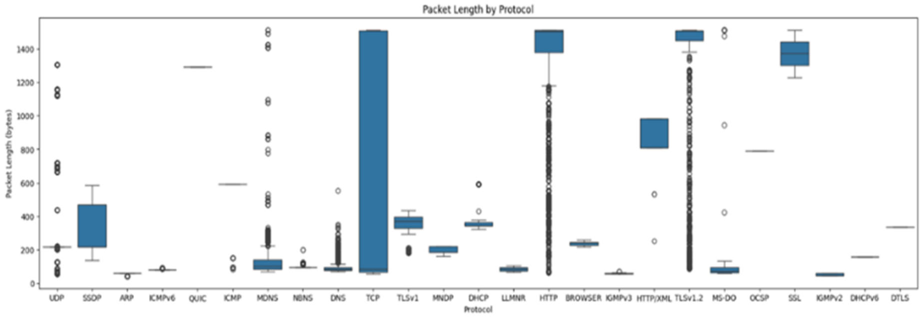

Figure 2: TCP vs UDP volume share by hour

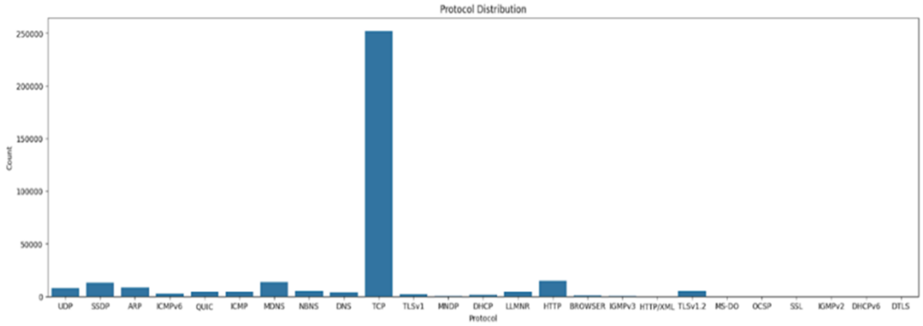

Figure 3: Packet length comparison: TCP vs UDP

Figure 4: Random Forest feature importance bar chart

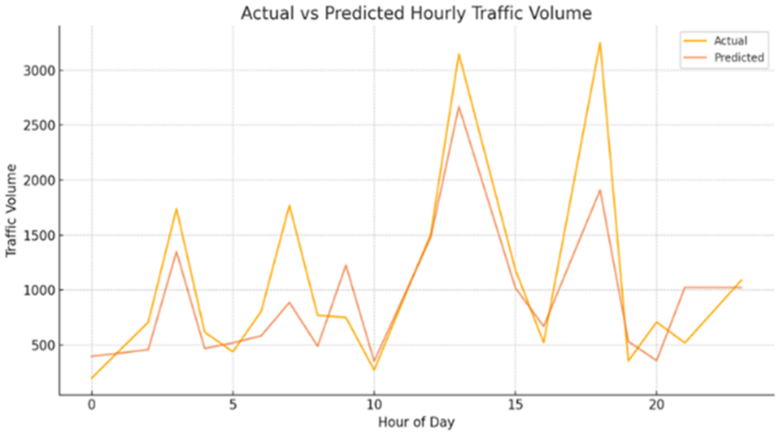

Figure 5: LSTM predicted vs actual hourly traffic (sample week)

5. Conclusion

This study set out to compare three machine learning models, Random Forest, Long Short Term Memory, and Support Vector Regression, for forecasting network traffic at the university. Across hourly, daily, and weekly horizons, LSTM consistently returned the lowest RMSE and MAE, while Random Forest offered a solid and interpretable alternative. SVR struggled with the non linear, bursty patterns typical of campus traffic. The persistence baseline served as a useful reminder that even modest temporal modelling can substantially improve on a naïve guess. What seems to give LSTM its edge is precisely the ability to track both short term fluctuations and the deeper daily rhythms that shape campus network demand. The survey feedback aligns neatly with the measured data: users feel the pinch most acutely in the afternoon window where the traffic loads peak. Taken together, the numbers and the user reports make a practical case for adopting recurrent forecasting as part of routine capacity planning. Still, several limitations temper these conclusions. The data came from a single university, so the findings may not transfer directly to other institutions with different network topologies or usage cultures. We did not incorporate external factors, weather, academic calendar shifts, or special events, that could sharpen predictions further. The dataset, while available on request under the university's data sharing policy, cannot be made fully public because of privacy constraints, which limits independent replication. The performance gap between LSTM and Random Forest narrowed at the weekly horizon, suggesting that for longer range forecasts the benefit of depth may plateau without richer input features. Future work should explore hybrid architectures that combine Random Forest's interpretability with LSTM's temporal sensitivity, test transfer learning across distinct campus networks, and systematically fold in exogenous variables. Even with these open questions, the current results indicate that a carefully tuned recurrent model can give network administrators a practical head start against congestion and a quieter afternoon for everyone on campus.

References

Ibrahim Afolabi, Tarik Taleb, Konstantinos Samdanis, Adlen Ksentini, and Hannu Flinck. Network slicing and softwarization: A survey on principles, enabling technologies, and solutions. *IEEE Communications Surveys & Tutorials*, 20(3):2429–2453, 2018.

Mohammad Abu Alsheikh, Shaowei Lin, Dusit Niyato, and Hwee-Pink Tan. Machine learning in wireless sensor networks: Algorithms, strategies, and applications. *IEEE Communications Surveys & Tutorials*, 16(4):1996–2018, 2014.

Ons Aouedi, Kandaraj Piamrat, Salima Hamma, and JK Menuka Perera. Network traffic analysis using machine learning: an unsupervised approach to understand and slice your network. *annals of telecommunications*, 77(5):297–309, 2022.

Gunter Bolch, Stefan Greiner, Hermann De Meer, and Kishor S Trivedi. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. John Wiley & Sons, 2006.

Anna L Buczak and Erhan Guven. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*, 18(2): 1153–1176, 2015.

Cisco Systems, Inc. Cisco visual networking index: Forecast and trends, 2017–2022 white paper. White paper, Cisco, February 2019. URL https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html.

Issam El Naqa, Ruijiang Li, and M Murphy. Machine learning in radiation oncology, 2019.

Zubair Md Fadlullah, Fengxiao Tang, Bomin Mao, Nei Kato, Osamu Akashi, Takeru Inoue, and Kimihiro Mizutani. State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems. *IEEE Communications Surveys & Tutorials*, 19(4):2432–2455, 2017.

Xenofon Foukas, Georgios Patounas, Ahmed Elmokashfi, and Mahesh K Marina. Network slicing in 5g: Survey and challenges. *IEEE communications magazine*, 55(5):94–100, 2017.

Yuming Jiang. Stochastic network calculus for performance analysis of internet networks—an overview and outlook. In *2012 International Conference on Computing, Networking and Communications (ICNC)*, pages 638–644. IEEE, 2012.

Manish Joshi and Theyazn Hassn Hadi. A review of network traffic analysis and prediction techniques. *arXiv preprint arXiv:1507.05722*, 2015.

Paulo Valente Klaine, Muhammad Ali Imran, Oluwakayode Onireti, and Richard Demo Souza. A survey of machine learning techniques applied to self-organizing cellular networks. *IEEE Communications Surveys & Tutorials*, 19(4):2392–2431, 2017.

Yanli Liu, Yourong Wang, and Jian Zhang. New machine learning algorithm: Random forest. In *International conference on information computing and applications*, pages 246–252. Springer, 2012.

Iraj Lohrasbinasab, Amin Shahraki, Amir Taherkordi, and Anca Delia Jurcut. From statistical-to machine learning-based network traffic prediction. *Transactions on Emerging Telecommunications Technologies*, 33(4):e4394, 2022.

Haipeng Lu and Fan Yang. Research on network traffic prediction based on long short-term memory neural network. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pages 1109–1113. IEEE, 2018.

Nian-Fa Mo. Network traffic prediction based on particle swarm optimization. In *2015 International Conference On Intelligent Transportation, Big Data And Smart City*, pages 531–534. IEEE, 2015.

SBA Mohammed, SM Sani, and DD Dajab. Network traffic analysis: A case study of abu network. *Network*, 4(4), 2013.

P Sridevi Ponmalar, Ayushi Kamboj, Vikram Shete, and R Harikrishnan. Machine learning based network traffic predictive analysis. *Rev. Comput. Eng. Res*, 9:96–108, 2022.

Sushil Kumar Singh, Mikail Mohammed Salim, Jeonghun Cha, Yi Pan, and Jong Hyuk Park. Machine learning-based network sub-slicing framework in a sustainable 5g environment. *Sustainability*, 12(15):6250, 2020.

Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. *Artificial intelligence review*, 53(8):5929–5955, 2020.

Leonardo Vanneschi and Sara Silva. Bayesian learning. In *Lectures on Intelligent Systems*, pages 259–270. Springer, 2023.

Shiwei Wang, Xiaoling Wu, Hainan Chen, Yanwen Wang, and Daiping Li. An optimal slicing strategy for sdn based smart home network. In *2014 International conference on smart computing*, pages 118–122. IEEE, 2014.