IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# A Comparative Study On Online Machine Learning Techniques For Network Traffic Streams Analysis

**Dr.A.Mekala**

**Asst.Professor, PG Dept of Computer Science**

**Sacred Heart College,Tirupattur**

**Abstract:**

State-of-the-art networks generate a huge amount of company data channels. Labeling these dates is essential for people of color, plus support for network tapes and cyber security scanning. There is an urgent need for data logic styles, which can perform online network data processing according to the emergence of new models dates. Online Machine Learning (OL) promises to support a similar type of data analysis. In this document we examine and compare the LO methods that facilitate the analysis of data blocks in the web domain. We also examine the importance of business data analysis and highlight the benefits of internet literacy in this context and the challenges in analyzing the business block of a network based on OL technology, e.g. drifts and unbalanced classes. Let's look at the data flow processing tools and frameworks that can be used for online or on-the-fly processing of this data with its advantages and disadvantages and its inerrability into the de facto system data processing framework. To test the performance of OL techniques, we conduct an empirical evaluation on the performance of various tree-based and ensemble-based network traffic classification algorithms. Finally, presented the open questions and future directions of traffic data flow analysis. This is a technical study provides valuable information and insight to the network research community when it comes to meeting requirements and the goals of online data flow analysis and network domain learning.

**Keywords:** Machine learning, Online learning, Network traffic streams, Network traffic classification Internet of Things Deep Learning.

**Introduction:**

Growing interest in launching new networks paradigms, e.g. Internet of Things (IoT), Internet of Vehicles (IoV) and the new generations of cellular networks, namely 5G and 6G [1], led to a significant increase in the network and thus the amount of data Communication in the digital world [2].amount of mobile data The traffic generated in 2022 is estimated at one zettabyte reported in the Cisco Annual Internet Report [3]. Also due to remarkable advances in sensor technology and computing power miniaturized devices and communication protocols, it's very interesting for deploying smart devices in various applications such as smart factories, Health and Smart Cities. Data is generated for these applications and/or massively consumed and transferred over time different communication protocols, from device to device (D2D) to satellite communication types. New network and computing paradigms such as 5G and 6G [1] and edge computing [6] motivate researchers to equip networks with self-organizing network technologies (SON) and self-sustaining network technologies (SSN) [7]. Network performance monitoring is the basis of these organizational techniques. However, various platforms for monitoring network performance are being introduced by analyzing the network traffic flows and processing this amount the date is a big challenge [8]. The analysis of network traffic flows is considered to be a major challenge as the data can be large and complex in many cases. IoT applications such as augmented reality (AR), automotive networks, interactive games and event tracking [5]. The need for analysis of these data led to the development of Network Traffic Monitoring i Analytical Techniques (NTMA) that can be used for various purposes , including detection of hidden traffic flow patterns [9] and decision making (e.g., routing and resource allocation) and forecasting possible events and problems, such as B. network traffic predictions and error handling or security [8]. Time sensitivity e.g nature of the high-speed data flow of these applications and networks requires a real-time (or near real-time) and online data analysis approach. Although there are many traditional NTMA techniques [10], there are also new ones techniques should be offered or existing NTMA methods may need to be used can be customized to analyze data streams in such applications.

Various NTMA techniques have been introduced, such as heuristics models, statistics-based techniques and machine learning (ML) techniques mainly focused on pattern and anomaly extraction streams, but below that mostly ML techniques are promising best performance in terms of accuracy and speed [11]. Different ML techniques were introduced for processing streams and time series records including, Recursive and Long-Term Neural Networks (RNN) Memory (LSTM). Although there is an extensive literature on the subject there are significant challenges to consider by applying these techniques to real NTMA applications as shown below:

Conceptual drift [12]: The statistical analysis of data flows is a difficult task due to training variations and actual rehearsals Tempo - conceptual drift. Long-term learning under conceptual drift can affect the accuracy of ML models.

• **Network Theory**: refers to the challenge posed by ML models trained on reference data sets may not be as accurate as they were then used in real networks; This is what every ML model should look like trained separately for each specific network [13].

• **Retraining problem**: Although this may be the case for a trained ML model good initial performance, thanks to the dynamics, its accuracy can be greatly reduced. ML models must be formed, but the network must not be left unattended when washing.

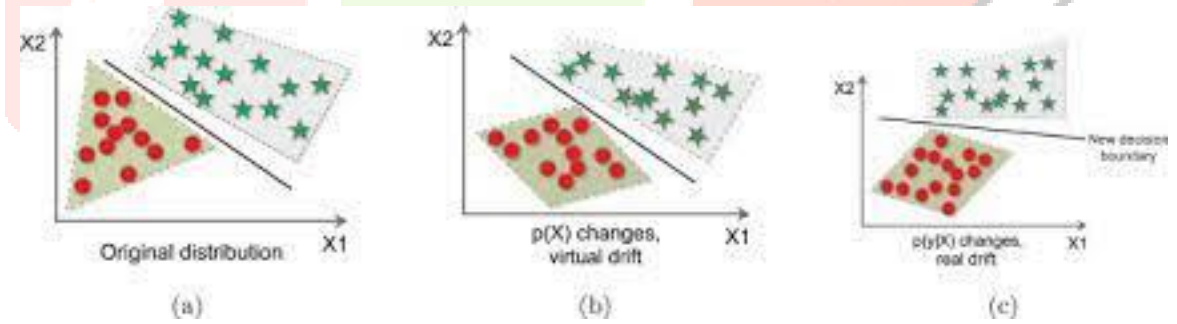
• **Lack of representative datasets**: Lacks of public datasets, which may represent new network paradigms, e.g.IoT networks are widespread. Even in the case of NTMA most Existing records are based on statistics collected from the network, such as network traffic matrix, which may not reflect real networks.

• **Network Interaction**: Most network paradigms are the, for example, is very dynamic in terms of mobility and generates different traffic flows with different characteristics. Then the goal ML should be able to interact with the network and update itself the model was gradually developed to support new situations.

**Related Work**

The OL and its challenges are well researched available literature [18]. However, its use in communication Systems and Networks is fairly new. As far as we know, is the first study to examine the special relationship between Analysis of OL data and traffic flows and challenges when using OL methods in communication systems. It should be noted that limited number of articles discuss LO and other learning-based techniques for specific types of network traffic analysis applications. ex., intrusion detection and security.

The article in [19] contains a detailed study of OL. The Article covers OL applications, the taxonomy of OL techniques, theoretical aspects and future directions of OL. Author at [20] reviewed and analyzed a variety of convex optimizations online algorithms. These algorithms include double averaging, mirror drop, FTRL, FTRL-Proxima and many earlier items. Authors at [18] reviewed the incremental OL paradigm and analyzed the state-of-the-art algorithms associated with it. The authors analyzed its main characteristics eight incremental techniques popular in your work because of these techniques represents several classes of algorithms. The comparison was based on online misclassification and the behavior of techniques in Limit . There is also an issue with hyperparameter optimization is discussed for selected techniques. Shalev-Swartz [21]. online bump optimization and the key role of bump in achieving it effective orienteering techniques.

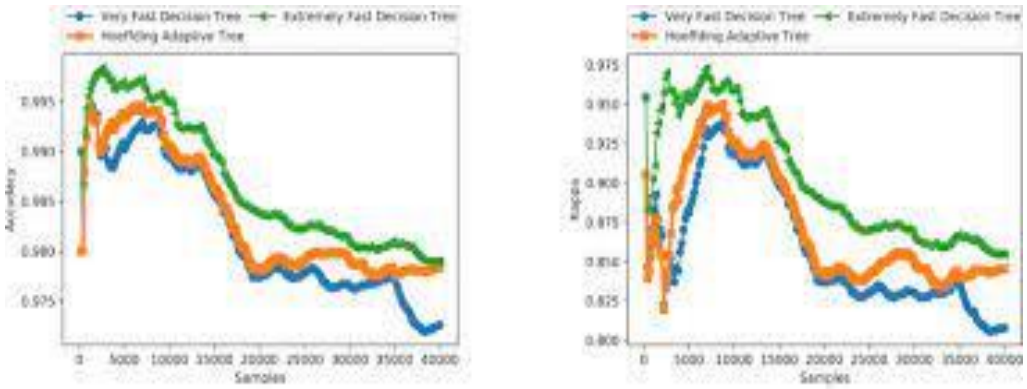



**Learning Procedure:**

learning method Supervised online learning assumes that the data $D$ = (($x1$ , $y1$ ), ($x2$ , $y2$ ), ($x3$ , $y3$ ), …, ($xm$, $ym$)) are available – one sample per times. We refer to $xi$ as the input instance and $yi$ as the target value or designation. For classfication problem, $yi$ takes on discrete values and in the regression problem, assumes continuous values. A couple ( , $yi$)) is an example of training.

Our goal is to create a predictive model $F \approx (y \mid X)$ from the formation copies. Algorithms are often trained in batch machine learning of all training samples because the data is available earlier. Anyway there are situations where $D$ data is not available in advance, e.g. B. fall detection system. In such applications, data arrives over time, either, set instance $xt$ in time step $t$ and $xt$ + 1 in time step $t$ + 1. Given the classification task, we strive to construct a working classifier $F$ after each time step using a training example

( , $yt$ ) AND Previous step model $Ft$-1. This is made possible by the introduction of OL techniques are given practical examples one by one as long as they are true tags are unknown at the time of receipt. Use OL algorithms training examples to optimize their loss/cost function and fit of their attitudes. For this purpose, OL can use stochastic optimization methods such as online backpropagation algorithms [31] and self-organization. cards (SOM) [32].Note that the online classifier will be $F$ gets a real $yt$ label after a period of time, so use it to rate classifier performance and further classifier improvement. The time intervals between different training examples (e.g. to $t$+1 and $t$+1 to $t$+2) are not essentially the same.




(a) Accuracy (b) Kappa

**Analyzing network traffic streams:**

In this section, we focus on the importance of data generation in communication systems and networks illustrate the greatest weaknesses traditional batch learning methods in network analysis traffic flows and discuss the main advantages of OL methods for communication systems. Currently, IoT is considered the ultimate an important emerging network paradigm. So, in this section, we emphasizes IoT and related paradigms, such as B. Industrial IoT (IIoT) one of the most important uses of network traffic in the real world flow analysis.

**Challenges of offline learning in network traffic streams analytic:**

**Dramatic data growth:** Cellular networks have seen tremendous growth over the past decade on mobile devices (e.g. smartphones and tablets). These devices generate (send or receive) huge amounts of data, e.g. Videos, images, text and geolocated information. Analyze such a large amount of data that you can capture the traffic different network levels (e.g.(e.g. application layer or data link) or from various sources (e.g. devices). In this case, store and organize this amount of data

**No resources available to process the stored data:**

As mentioned the speed of data generation in communication systems and networking is booming and growing. With offline machine algorithms, the collected data (saved data) is usually used for training and for evaluation purposes. However, it is fast and continuous data generation causes serious problems for our storage and computing resources. Also a big part traffic data are available as streams. door so huge amount of traffic data transferred to the main memory of the device, particularly low-cost and energy-efficient IoT devices with memory [44]. almost impossible and not feasible.

**No data is reported in real apps:**

monitored offline learning algorithms depend on label training samples. In concrete terms, most of these algorithms assume this properly labeled workout examples are available ahead of time. This assumption can affect the functionality of these algorithms in real applications; when the amount of available data is highlighted The training targets are small or the selected training instances are not representatives [46]. Then supervised offline learning algorithms needs a lot of high quality tagged data to deliver the goods efficiency.

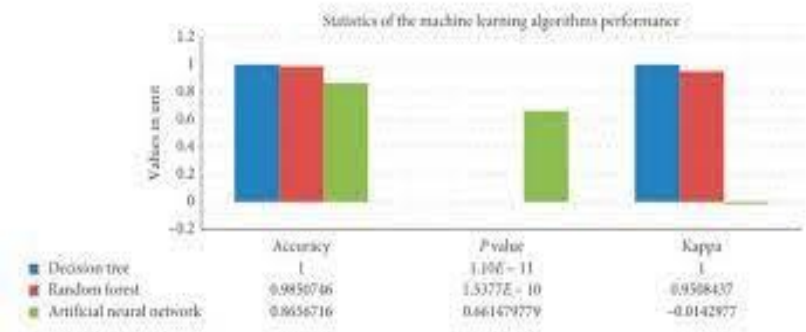
**E-learning challenges in analyzing network traffic flow:**

Although OL theoretically has many advantages in this case, there are none are some of the challenges that need to be addressed to apply OL to the field of network traffic flow analysis. Compared to traditional offline mode machine learning algorithms, design online learning algorithms The can learn from large streams of data, which is more difficult. It's controversial explains the fact that OL algorithms have to learn from a training instances at a time, so you need to design more training process [53]. In addition, the data were published from transient or dynamically changing systems, e.g., IoT and IIoT systems have an unsteady distribution (concept drift). ref The fact that only one training instance (or limited instance) is available in OL. at a time, identify those changes and take opportunities to address them These changes are challenging. Also in many cases supervised learning Tasks such as network traffic classification and class designations are not present too (unbalanced classes) because there are minority/majority classes. Unbalanced activities complicate the training process and increase the load complications data status [27,54]. Concept drift and imbalance The Class challenges are detailed below Top OL challenges in analyzing network traffic flows.

**Challenges and future directions:**

**IoT data streams:**

Despite recent advances in OL algorithms for large traffic data streams, there are still problems Consider to further develop this approach to learning network paradigms. For example IoT data properties, such as heterogeneity, highly dynamic environments, noisy data, and the spatio-temporal correlation can aggravate the situation. THE large IoT data streams can increase learning costs time, memory and calculations because a large number Data instances and associated attributes produce a very complex result Model OL, thus extending working hours and Performance drop. While present OL The does reasonably well with existing, current network data flows The shift in network paradigms is changing the essence network traffic flows, resulting in more complex data , so existing OL algorithms need to be extended to process new types of data in terms of speed, complexity and heterogeneity.



**Conclusion:**

In this Paper, we examined the OL paradigm from a network perspective. OL arouses great interest and has since become a hot topic in the scientific community different practical use cases in traffic data stream analysis. Communication systems and networks can use LO e.g. Chain consumer traffic data generator . Specifically in this chain devices or network users generate raw data that can be analyzed by OL algorithms. So OL models extract valuable knowledge Data important for decision making, QoS/QoE assurance and predictive modeling.However, the performance OL algorithms can be affected by the unbalanced class and transient nature of network environments (e.g. concept drift). IN of this article we discussed the data properties and network traffic of his challenges for OL algorithms. In particular, we emphasized key characteristics of IoT data.In addition, we analyzed the problems related to traditional offline and online learning methods benefits for contractors. As the research core, we provide comprehensive services In addition, dealt with algorithms and techniques for data stream mining on data stream processing tools and facilities. Especially us compared data processing tools and highlighted features and benefits and the disadvantages of all stream processing tools. As part of a comparative study, we evaluated the performance of some online team and tree-based algorithms to show how online machine learning techniques can work with network traffic flow analysis. We also did some research identified challenges and future research directions for the deployment of OAs for communication systems and networks.

**References:**

[1] A. Shahraki, M. Abbasi, M. Piran, M. Chen, S. Cui, et al., A comprehensive survey on 6g networks: Applications, core services, enabling technologies, and future challenges, 2021, arXiv preprint arXiv:2101.12475.

[2] L. Yang, D. Holtz, S. Jaffe, S. Suri, S. Sinha, J. Weston, C. Joyce, N. Shah, K. Sherman, B. Hecht, and J. Teevan, “The effects of remote work on collaboration among information workers,” Nature Human Behaviour, Sep. 2021.

[3] L. Stewart, G. Armitage, P. Branch, and S. Zander, "An Architecture for Automated Network Control of QoS over Consumer Broadband Links," TENCON 2005 - 2005 IEEE Region 10 Conference, pp. 1-6, November 2005.

[4] T. Karagiannis, A. Broido, M. Faloutsos, and K. claffy, “Transport layer identification of P2P traffic,” Proceeding of the 4 th ACM SIGCOMM conference on Internet measurement (IMC '04), New York, pp. 121–134, September 2004

[5] P. B. Park, Y. Won, J. Chung, M. Kim, and J. W.-K. Hong, “Fine-grained traffic classification based on functional separation,” International Journal of Network Management, vol. 23, no. 5, pp. 350–381, Aug. 2013.

[6] G. Aceto, A. Dainotti, W. de Donato and A. Pescape, "PortLoad: Taking the Best of Two Worlds in Traffic Classification," 2010 INFOCOM IEEE Conference on Computer Communications Workshops, pp. 1-5, March 2010.

[7] Z. Yuan and C. Wang, "An improved network traffic classification algorithm based on Hadoop decision tree," 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), pp. 53-56, May 2016.

[8] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn and F. Abdessamia, "Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms," 2016 2nd IEEE International Conference on Computer and Communications (ICCC), pp. 2451-2455, October 2016.

[9] Z. Fan and R. Liu, "Investigation of machine learning based network traffic classification," 2017 International Symposium on Wireless Communication Systems (ISWCS), pp. 1-6, August 2017.

[10] A. Pasyuk, E. Semenov and D. Tyuhtyaev, "Feature Selection in the Classification of Network Traffic Flows," 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), pp. 1-5, October 2019.

[11] Y. Wang, Y. Xiang and S. Yu, "Internet Traffic Classification Using Machine Learning: A Token-based Approach," 2011 14th IEEE International Conference on Computational Science and Engineering, pp. 285-289, August 2011.

[12] S. Dong and R. Jain, “Flow online identification method for the encrypted Skype,” in Journal of Network and Computer Applications, vol 132, pp. 75-85.

[13] M. Dixit, R. Sharma, S. Shaikh and K. Muley, "Internet Traffic Detection using Naïve Bayes and K-Nearest Neighbors (KNN) algorithm," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1153- 1157, May 2019.

[14] F. Zhang, Y. Wang and M. Ye, "Network Traffic Classification Method Based on Improved Capsule Neural Network," 2018 14th International Conference on Computational Intelligence and Security (CIS), pp. 174-178, November 2018.

[15] H. Lim, J. Kim, J. Heo, K. Kim, Y. Hong and Y. Han, "Packet-based Network Traffic Classification Using Deep Learning," 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 046-05, February 2019.

[16] J. Kwon, D. Jung and H. Park, "Traffic Data Classification using Machine Learning Algorithms in SDN Networks," 2020 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1031-1033, October 2020.

[17] Z. Li, Z. Qin, K. Huang, X. Yang, and S. Ye, “Intrusion Detection Using Convolutional Neural Networks for Representation Learning,” Lecture Notes in Computer Science, pp. 858–866, 2017.

[18] A. S. Iliyasu and H. Deng, "Semi-Supervised Encrypted Traffic Classification with Deep Convolutional Generative Adversarial Networks," in IEEE Access, vol. 8, pp. 118-126, 2020.

[19] G. D’Angelo and F. Palmieri, "Network traffic classification using deep convolutional recurrent autoencoder neural networks for spatial–temporal features extraction," Journal of Network and Computer Applications, vol. 173, pp. 102890, 2021.

[20] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, “Characterization of Encrypted and VPN Traffic using Time-related Features,” Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP2016), pp. 407-414, February 2016.

[21] H. A. H. Ibrahim, O. R. Aqeel Al Zuobi, M. A. Al-Namari, G. Mohamed Ali, and A. A. A. Abdalla, "Internet traffic classification using machine learning approach: Datasets validation issues," 2016 Conference of Basic Sciences and Engineering Studies (SGCAC), pp. 158-166, February 2016.

[22] A. Moldagulova and R. B. Sulaiman, "Using KNN algorithm for classification of textual documents," 2017 8th International Conference on Information Technology (ICIT), pp. 665-671, May 2017.