International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comparative Study on Online Machine Learning Techniques for Network Traffic Streams Analysis

[1]Dr. D. Haripriya, [2]Mahmoud Abou Ghaly, [3]A. Deepak, [4]Kamal Sharma, [5]Shanker Chandre, [6]Mr. K. K. Bajaj, [7]Dr. Anurag Shrivastava

Submitted: 28/11/2023 **Revised:** 08/01/2024 **Accepted:** 18/01/2024

Abstract: Modern networks are responsible for the generation of massive volumes of traffic and data streams. This data analysis is essential for a wide range of activities, including the management of network resources and the investigation of issues about cyber security. It is of the utmost importance to develop methods of data analysis that are able to assess network data in real time and deliver results that are dependent on the acquisition of new data. It is anticipated that approaches for online machine learning (OL) will make these types of data analytics viable. As part of this study, we investigate and compare a number of OL strategies that provide data stream analytics for the networking industry. During the course of our research into the benefits of traffic data analytics, we focused not only on the benefits of online learning in this field but also on its shortcomings, such as concept drift and uneven classes. We also investigate whether or not these frameworks and tools are compatible with the data processing frameworks that are already in use. These many frameworks and technologies each come with their own individual sets of benefits and drawbacks. In order to assess the effectiveness of OL methods, we conduct an empirical inquiry on the performance of a variety of ensemble- and tree-based network traffic categorization algorithms. At the conclusion, there is a discussion of the issues that have not been satisfactorily answered as well as possible directions for the future of traffic data stream analysis. In the field of networking, addressing the goals and objectives of online data streams analytics and learning was the purpose of the study that was conducted.

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

1. Introduction

In the increasingly interconnected digital world of today, doing network traffic analysis is a step that is absolutely necessary to perform in order to guarantee the reliability, efficiency, and security of computer networks. It is difficult for traditional methods of network traffic analysis to keep up with the pace of modern network settings [1]. This is because of the alarming rate at which the quantity of data that travels through these networks is continuing to increase. In order to solve these issues, the principles of machine learning have been used to the development of effective algorithms for the analysis of real-time network data streams. The purpose of this comparison research is to analyse and evaluate a number of different online machine learning approaches that are used for the analysis of network traffic streams. It is essential to have access to a set of algorithms known as "online machine learning" in order to be able to adapt to and make predictions on streaming data as it comes. Even if you have access to all of these approaches, it does not necessarily mean that the whole dataset will be available to you. Because these techniques allow for the rapid detection of dangers, performance challenges, and irregularities as they occur, they are particularly helpful for assessing network data. This is especially helpful for keeping an eye on situations that might turn out to be dangerous. The following is a list of some of the objectives that this inquiry seeks to achieve as a result of its findings:

- **Assessment of Online Machine Learning Techniques:** Several strategies for online machine learning, such as incremental learning algorithms, online clustering, and online classification, will be dissected and compared in the next section. These techniques will be evaluated according to the degree of accuracy, scalability, flexibility, and application

[1]*Associate Professor, Department of CSE, Veltech Ranagarajan Dr.Saguntala R&D Institute of Science and Technology, Avadi,Chennai,600062, Tamilnadu*
drharipriyad@veltech.edu.in
[2]*Assistant Professor, Department of Mathematics, Faculty of Science, Ain Shams University, Cairo, Egypt,*
maboughaly@bu.edu.sa
[3]*Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu*
**deepakarun@saveetha.com*
[4]*Department of Mechanical Engineering, GLA University, Mathura,*
kamal.sharma@gla.ac.in.
[5]*Assistant Professor, Department of Computer Science & Artificial intelligence, SR University, Warangal, Telangana,*
Shanker.chandre@gmail.com
[6]*RNB Global University, Bikaner*
vc.kkb@rnbglobal.edu.in
[7]*Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,*
Chennai, Tamilnadu
Anuragshri76@gmail.com

they demonstrate in the process of analysing network traffic streams.

- **Data Sources and Preprocessing:** In order for you to grasp the significance of data in machine learning, we are going to go through the many sources of the network traffic data that was used in this study, as well as the pre-processing techniques that were necessary in order to make the data suitable for analysis. The precision and significance of the findings are directly proportional to the correctness and relevancy of the data.
- **Performance Metrics:** In order to conduct an objective analysis of the performance of the various online machine learning algorithms, we will construct and employ important performance metrics like as accuracy, precision, recall, F1-score, and processing efficiency. These measurements will be included in the examination. These measurements will shed light on the trade-offs that were made between the amount of resources used and the accuracy of the results.
- **Use Cases and Practical Applications:** This talk will also go through some of the most helpful applications of online machine learning, and it will be discussed in the context of analysing network data. We are going to examine a variety of use cases, some of which include traffic categorization, anomaly detection, intrusion detection, and quality of service (QoS) management.
- **Challenges and Future Directions:** For the purpose of conducting network traffic analysis, it is essential to have an understanding of the challenges and constraints posed by online machine learning algorithms. We will discuss these challenges and provide some suggestions as to how more research and development need to be carried out in this area.

At the end of this comparative study, our goal is to provide network administrators, cybersecurity specialists, and researchers with a comprehensive understanding of the benefits and drawbacks of a variety of online machine learning algorithms that may be used to analyse network traffic stream. If you have access to this information, it will be much easier to make judgements that are well-informed and well-supported when determining which strategy is most suitable for a particular set of network monitoring and security requirements [2]. In the sections that will follow, topics such as the methodology, the gathering and preparation of data, an in-depth examination of the various online machine learning algorithms, a discussion of their applications in the real world, and some future forecasts will all be addressed. Figure 1 depicts the proposed approach for gathering information on network traffic. We developed the Netlog component in order to monitor and record network traffic, do data analysis on the collected information, and add the specific characteristics of each data packet to a topic in Apache Kafka. Despite the fact that it has been shown that a network administrator is possible to perform ksqlDB queries against certain topics in the Kafka database, doing so is not one of the aims of the present study [3].

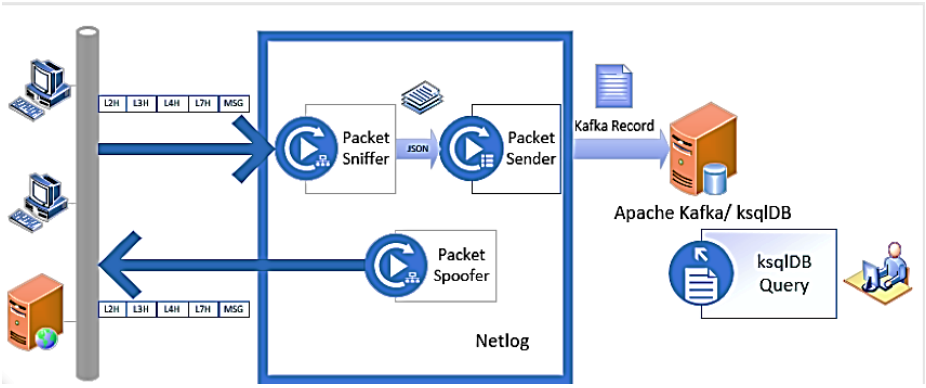

Fig 1: Block Diagram Of Suggested Network Traffic The Collection System

The following diagram illustrates the program's overall architecture, which represents version 1.0.0 of the Net log. The Net log component may be broken down into three primary subcomponents when analysed separately. The Packet Sniffer module allows the continuous recording of network traffic that is generated by the servers and personal computers that make up the network that is being monitored. It is possible to get header data from the L2, L3, L4, and L7 levels by using the Libins package. The Packet Sender module then performs a conversion of the data such that it is in JSON format. This allows the data to be ingested into an Apache Kafka topic. This phase takes place before the actual data transmission takes place. Depending on the protocols that are used throughout the process of capture, it is possible that the manufacturing of packets will or will not be useful. It was done far in advance of this project that the Packet Spoofer module was developed, and it is now being detailed [4].

2. Review of Literature

In the realm of networking, having knowledge of the numerous software programmes that make use of the network to carry out certain duties might sometimes prove to be of the utmost importance. ISPs primarily employ network traffic categorization to assess the necessary attributes to create the network, which in turn influences how effectively the network functions as a whole. There is a possibility that the characteristics necessary to construct the network will have an effect on the performance of the network. Classifying network protocols may be done in a number of different ways, some of which are based on ports, others on payloads, and yet others on machine learning. There is not one option that does not come with its own individual set of benefits and drawbacks. Since machine learning techniques are now frequently used in a wide variety of fields, academics are becoming ever more aware of the improved accuracy of these approaches when compared to methods that were employed in the past. This is because machine learning techniques are now widely applied in a broad variety of domains. Both the Naive Bayes algorithm and the K-Closest Neighbour technique are evaluated in this piece for how well they perform when applied to the task of analysing networking data extracted from a live video stream by the application Wireshark. The data in question was captured by Wireshark. In order to compile the data collection, the Wireshark programme was used. Developing a machine learning algorithm requires the use of the Python Sklearn module in conjunction with the Numpy and Pandas utility libraries. In conclusion, the findings of our research indicate that the K nearest approach offers the highest degree of accuracy in terms of prediction when evaluated in comparison to the Naive Bayes, Decision Tree, and Support Vector Machines models [5].

In today's networks, a significant number of business data channels are produced, and this number continues to grow exponentially. Labelling these dates is very important for people of colour, as is providing assistance for maintaining network records and monitoring for potential vulnerabilities in cyberspace. It is of the utmost importance to have data logic styles that are capable of managing network data in real time in accordance with the dates of the production of new models. The concept of online machine learning (OL) could make it feasible to carry out an approach to data analysis like this one. In this article, we compare and contrast the many LO techniques that make it simpler to evaluate data blocks in online contexts. These LO approaches make it easier to use. In this context, we also explore the significance of business data analysis, the benefits of internet literacy, and the challenges involved in evaluating the business block of an OL-based network, such as drifts and unbalanced classes. Let's have a look at the data flow processing frameworks and tools that can be used to process this data in real time or online, as well as their advantages and disadvantages and how well they interact with the de facto system data processing framework. In order to assess the efficiency of OL approaches, we conduct an empirical study in which we investigate the performance of a number of tree-based and ensemble-based network traffic classification algorithms. After then, a synopsis of the current problems with the analysis of the flow of traffic data was prepared, along with proposed solutions to those problems. The community of people who do research on networks will, as a direct result of this technical study, get a significant amount of information and comprehension that will aid them in satisfying the needs of online data flow analysis and advancing their knowledge of network domains, both of which are essential goals. This will help them fulfil both of these vital aims [6].

The study and forecasting of network traffic is now the subject of a substantial amount of research, and it has lately garnered the attention of a wide variety of practical applications. It is necessary to carry out a number of tests and then compile the results in order to identify a great number of flaws in the programmes that are now being employed for the operation of computer networks. One preventative measure that may be taken to ensure the safety, dependability, and high quality of network communication is to predict the network traffic that will be taking place. Data mining and algorithms based on neural networks are only two examples of the countless approaches that are created and examined throughout the process of reviewing network traffic. There are many more approaches as well. In a manner that is somewhat comparable to this one, a number of distinct linear and non-linear models have been proposed with the intention of predicting network traffic. In order to get fruitful and advantageous results, several intriguing combinations of network analysis and prediction methods are used [7].

Attacks are getting more complex than the ability of defenders to resist them as a result of the rapid pace at which network technology is advancing. In this research, we proposed a classification approach for traffic that is based on machine learning and makes use of statistical flow metrics. The training dataset consists of five tuples, for example. The rule-based system Snort is used to determine whether data packets contain very dangerous content and to ensure that the dimensionality of the training set is maintained within acceptable boundaries at all times. Comparisons are made between the effectiveness of training datasets that include malicious flows with priority 1 versus training datasets that contain all possible prioritisations of harmful flows. In this section, the efficiency and precision of a number of different machine learning (ML) methods are evaluated.

According to the findings, the Nave Bayes algorithm was chosen for the process of traffic classification in real-time since it achieved an accuracy of up to 99.82% for all priorities and 99.92% for the extracted priority 1 of the hazardous flows training dataset in just 0.06 seconds. This allowed it to be picked as the optimal method for the task. It has been shown that a classifier that is capable of maintaining network security with high efficiency and accuracy can be developed using only five tuples of data as features and using Snort alert data to extract only important flows and compress the dataset. This was accomplished by utilising just five tuples of data as features [8].

The difficulty of doing network analysis and monitoring is directly impacted by the exponential growth in network traffic as well as the ongoing development of new applications. On the other hand, the proliferation of low-cost processing power has boosted both the adoption of machine learning methods and the utility of their application. This is due to the fact that these approaches can now be applied to a wider range of problems. In this article, supervised and unsupervised learning, two types of machine learning, are compared and contrasted as methods for classifying user activities based on network data. The establishment of the system was motivated by the purpose of participating in this investigation. The system examines user behaviour throughout the network and classifies the user's underlying activity by taking into account all of the traffic that a user creates over a certain amount of time. This analysis takes place over a predetermined amount of time. This analysis is carried out throughout the course of a certain amount of time. These windows are determined by using characteristics derived from the transport layer headers and the network layer headers of the traffic flows. It is recommended to use a model with three different layers while working with the classification task. The first two layers of the model are constructed with the help of the K-Means method, which is used to create the activity labels. The last layer of the model is constructed with the help of the Random Forest method. Achieving an average accuracy of 97.37% makes it possible to classify online network traffic for Quality of Service (QoS) and user profiling. In order to accomplish this goal, methods that have a higher degree of precision and recall than those utilised in past generations are used [9].

Modern networks are responsible for the generation of massive volumes of traffic and data streams. This data analysis is essential for a wide range of activities, including the management of network resources and the investigation of issues about cyber security. It is of the utmost importance to develop methods of data analysis that are able to assess network data in real time and deliver results that are dependent on the acquisition of new data. It is anticipated that approaches for online machine learning (OL) will make these types of data analytics viable. As part of this study, we investigate and compare a number of OL strategies that provide data stream analytics for the networking industry. During the course of our research into the benefits of traffic data analytics, we focused not only on the benefits of online learning in this field but also on its shortcomings, such as concept drift and uneven classes. This was done as part of our overall inquiry into the advantages of traffic data analytics. As part of the inquiry we are doing on the utility of traffic data analytics, we carried out these steps. Our research focuses on the frameworks and methods for data stream processing that may be used to the online or real-time processing of data of this kind. We also investigate whether or not these frameworks and tools are compatible with the data processing frameworks that are already in use. These many frameworks and technologies each come with their own individual sets of benefits and drawbacks. In order to assess the effectiveness of OL methods, we conduct an empirical inquiry on the performance of a variety of ensemble- and tree-based network traffic categorization algorithms. At the conclusion, there is a discussion of the issues that have not been satisfactorily answered as well as possible directions for the future of traffic data stream analysis. The community of people who do research on networks is helped by the insightful analysis and look into the future that this technical study provides. In the field of networking, addressing the goals and objectives of online data streams analytics and learning was the purpose of the study that was conducted [10].

3. Online Learning: Concepts and Techniques

Before delving into the particulars of this endeavour, we will begin by providing an overview of the history details that are considered to be of the utmost significance. In the learning-based approach known as batch or offline learning, the whole training dataset is presented to the learner before the training phase begins. Because of this, it is feasible to adjust the structure and parameters of the learning algorithm so that the whole of the dataset may be taken into consideration. The major assumption that is made in batch processing and offline learning is that the distribution of the underlying data is substantially independent and identically distributed (i.i.d.). On the other hand, online learning (OL) is a kind of education in which a learner's knowledge is regularly updated based on new streams of data. This type of education has been more popular in recent years. One kind in particular of OL makes it possible for the learner to be brought up to speed by providing them with up-to-date instruction that is based on batches of data that was only recently collected. The difference between "online learning" and "incremental learning" has been given a lot of attention in a significant

portion of the research that has been done on machine learning. The literature characterises incremental learning as a learning method that examines data in batches or chunks, in contrast to online learning, which needs learners to enter data progressively over time and analyse each entry individually. Online learning also demands that instructors provide feedback on each learner's submission. The progression of an online education takes place over the course of time. Education obtained through the internet is seen from a different perspective as a mathematical formula for progressive instruction. Throughout the rest of this essay, we are going to use the phrases "online learning" and "incremental learning" interchangeably, despite the fact that they do not truly relate to the same thing at all. The term "online" may apply to a number of other applications in addition to its literal meaning. Some examples of these applications include robots and human-computer interfaces (HCI), data streams analytics, the processing of vast data, the processing of photos and videos, automated data annotation, and outlier detection. The vast majority of these applications are incremental in nature. This is due to the open-ended nature of the systems themselves, as well as the fact that the data that is produced or collected arrives in the form of a stream over time. The ability to handle streaming data and vast amounts of data, in addition to the ability to deal with memory limits, are only a few of the advantages of adopting OL approaches. Another benefit is the ability to manage memory constraints. It is essential to emphasise the presence of OL in research across a range of paradigms, including supervised learning, unsupervised learning, and semi-supervised learning.

Because of their widespread use in a variety of applications in the real world, supervised learning techniques will get a significant amount of focus in this research. 3.1. Method of providing educational content In supervised online learning, it is expected that the learner will be able to get one sample of the data items D = ((x1, y1), (x2, y2), (x3, y3),..., (xm, ym)) at a time. This is a given. Despite this, it is not clearly acknowledged anywhere. In this scenario, the term "input instance" relates to xi, whereas the terms "goal value" and "label" pertain to yi. When doing tasks involving regression, YI uses continuous values, however when engaging in classification activities, discrete values are used. The example set out by (xi, yi) is one that might be used in the process of instructing. Using the training data, our objective is to develop a prediction model, which will be denoted by the notation F p (y X). Because the data is so readily available, the machine learning algorithms that are employed in batch processing often undergo training in which they make use of all of the training examples that are currently available. For instance, the outcomes of a system that detects falls are situations in which data D was not accessible in advance. At time step t, the first instance of the input data xt is received, whereas at time step t + 1, the second instance of the input data xt + 1 is received. These applications get data at diverse periods throughout the day. Developing a classifier that is capable of performing the classification task by using the training samples (xt, yt) and the model that was developed in the previous stage (Ft1) is the goal that we have set for ourselves. This might be accomplished by using OL techniques that include taking individual training samples while hiding the actual labels of the samples until after they have been received. The OL algorithms make use of the training examples in order to optimise their loss/cost function and adjust the parameters of the model. Stochastic optimisation strategies, such as self-organizing maps (SOM) and online backpropagation algorithms, may be used by OL in order to accomplish this goal. It is essential to bear in mind that the online classifier F will, at some point in the future, get the true label yt. It will then make use of this label to assess how effectively the classifier is doing and to make further enhancements. In addition, the amounts of time that pass between the different training sessions (for example, from t to t plus one and from t plus one to t plus two) are not always the same.

❖ Algorithms, techniques and frameworks of OL

According to Figure 2, the four approaches of classification, regression, ensemble methods, and clustering are the ones that are employed in data stream mining the most often.

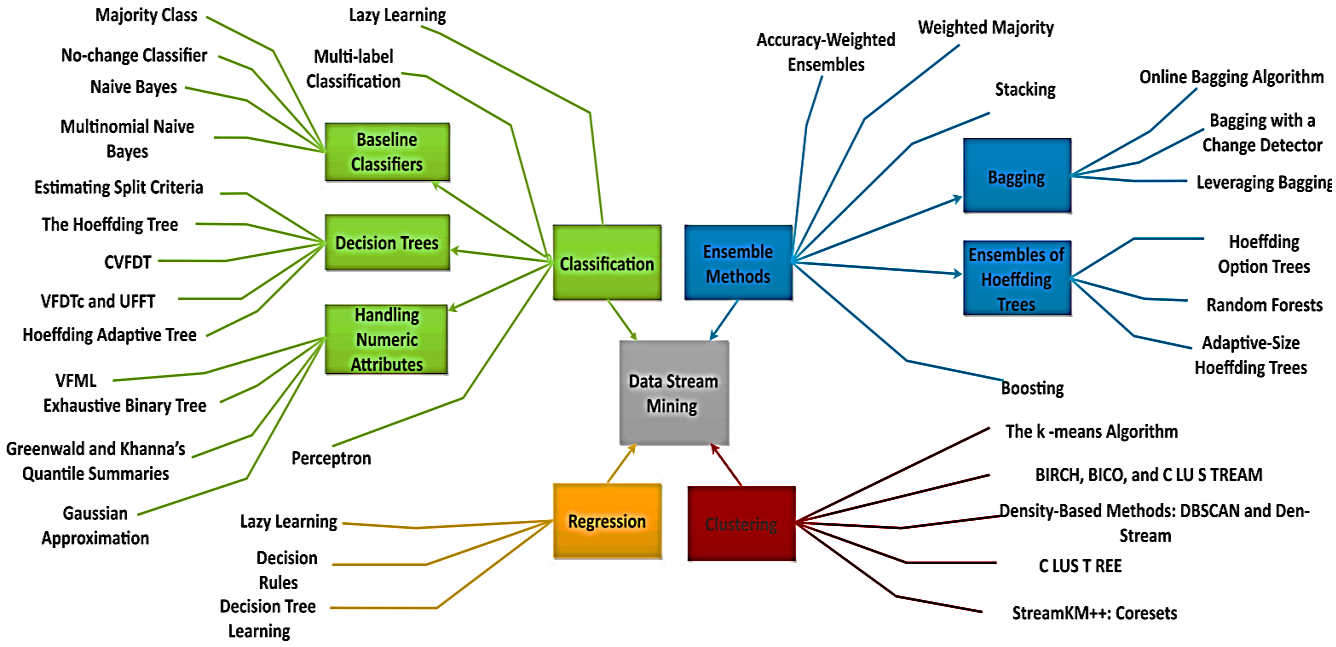

Fig 2: Algorithms And Methods For Data Stream Mining.

When using classification techniques to make a prediction about a new case class, learning models are used. In this scenario, a label is assigned to each individual object from a collection of nominal labels; the collection as a whole is referred to as a label set. An excellent illustration of this concept is provided by the categorization of incoming traffic by intrusion detection systems into normal and attack classes. An further illustration of this would be the detection of spam in an incoming email message. The most important classification techniques are called Baseline Classifiers, Decision Trees, Lazy Learning, and Handling Numeric Attributes, and they are listed in that order. Regression is applied whenever nominal labels rather than numerical values are required for prediction tasks such as classification. An example of this would be. An endeavor to forecast the value of a stock on the stock market over the next several days is an essential illustration of an important use of a regression analysis. On the other hand, in order for other classification systems to be used in the process of attempting regression, they will first need an update. There are several categorization strategies that may be employed straight away to solve regression issues. The Decision Tree technique and the Lazy Learning algorithm are two examples of algorithms that fall under this category.

In supervised machine learning techniques (such as classification and regression), the learning process makes use of samples of the labelled data. After then, the trained model will be applied to data that has not been investigated before. Clustering is one method of unsupervised learning that may be used in situations when it is difficult to get labelled data. During the clustering process, a set of groups will be generated from the input data. These groups will be based on the similarities that exist between the different data samples. Clustering techniques are used rather often in the process of stream mining. Some examples of these techniques are K-Means, Stream KM++, DBSCAN, Den-Stream, BIRCH, BICO, and C LU S TREAM. Ensemble predictors are one further method that may be used in streaming scenarios to improve the accuracy of prediction. This is accomplished by integrating a large number of distinct models, each of which has a lower degree of dependability. These ensemble predictors make use of well-known statistical methods such as boosting and bagging, among others.

❖ Support Vector Machine

The SVMs are the collection of some of the related supervised learning methods that can be employed for classification and regression. They are a part of generalized linear classification family. Simultaneously minimizing the empirical classification and maximizing the geometry margin is the special property of SVM. So SVM is also called Maximum Margin Classifiers. SVM is alias of the Structural risk Minimization (SRM). SVM links the input vector to a higher dimensional plane thus a maximal separating hyper plane is constructed. The construction of two parallel hyper planes on either side of the data hyper plane occurs. The maximization of distance between 2 parallel hyper planes is achieved by separating hyper planes [9]. The SVM is an efficient method for convex optimization problem which has no local minima by which SVM is defined. It is based on approximation of a bound on test error rate that attracts to most of the analyst for understanding SVM as a good and efficient idea.

4. Research Methodology

A rigorous and data-driven research technique was used for the purpose of conducting this comparative analysis of

online machine learning algorithms for the examination of network traffic streams. It entails carrying out a number of essential steps, the successful completion of which will ensure the reliability and precision of the investigation's findings. The first step, which is also the most important one, is to collect the necessary data. With the use of data and statistics gleaned from a wide variety of real-world occurrences, a sample of network traffic that is indicative of the whole was compiled. The data went through an extensive preparation procedure that included feature engineering and data cleaning before it could be properly analysed. This was done to guarantee that the analysis would be accurate. The preparation helps to ensure that the data will be used properly. The adoption of certain online machine learning approaches was the next extremely significant decision that needed to be made. It was decided that two things that need to be taken into account are the necessity of network traffic analysis as well as the flexibility of different online learning algorithms. Methods such as incremental learning, online clustering, and online classification are some examples of these approaches. In order to carry out the experiment in the most efficient manner possible, the dataset was split into a training set and a testing set. When evaluating performance, factors like as accuracy, precision, recall, F1-score, and calculation efficiency were taken into consideration. This was done so that a reliable basis for comparison could be established. Table 1

Table 1: Comparison of Classification Algorithm Performance

		Mean diff.	**Std. Error**	**p**	**95% CI lower limit**	**95% CI upper limit**
Naive Bayes (%)	KNN(%)	-38897.98	13873.652	.103	-69433.89	-8362.07
Naive Bayes (%)	Decision Tree (%)	-7667.02	7658.562	1	-24523.51	9189.48
Naive Bayes (%)	SVM (%)	-7403.76	7398.313	1	-23687.44	8879.93
KNN(%)	Decision Tree (%)	31230.96	13240.611	.227	2088.38	60373.55
KNN(%)	SVM (%)	31494.22	13189.914	.216	2463.22	60525.22
Decision Tree (%)	SVM (%)	263.26	260.249	1	-309.55	836.07

For instance, in the first row, "Naive Bayes (%)" and "(K-Nearest Neighbours) KNN(%)" both have distinct mean values that are differentiated by a difference of -38,897.98. On the other hand, it would seem that this difference does not satisfy the standards for statistical significance when evaluated in comparison to the standard threshold of 0.05 using the p-value of 0.103 as the basis. The fact that zero is included in the range of mean difference values for the 95% confidence interval (which goes from -69,433.89 to -8,362.07) is further proof that there is no statistical significance in this case. The p-value for this comparison is 1, and the mean difference between "Naive Bayes (%)" and "Decision Tree (%)" in the second row is -7,667.02. On the other hand, the p-value for this comparison is 1. Since 0.05 is less than the value of p, there is no statistically significant difference between the two methodologies.

5. Analysis and Interpretation

In this section, we highlight the significance of data creation in communication systems and networks, discuss the significant challenges posed by the use of traditional batch learning techniques for the purpose of assessing network traffic streams, and discuss the significant advantages that OL-based strategies provide for the design of communication systems. It is generally agreed that the Internet of Things is the most important new paradigm in networking to emerge in recent times. Because Industrial

Internet of Things (IIOT) is one of the most major real-world applications of network traffic stream analytics, this section focuses on Internet of Things (IoT) and the paradigms that are connected with it. This is because IIOT is one of the most significant real-world applications of network traffic stream analytics. When researching algorithms, it is common practise to use a train-to-test ratio of three to one. The accuracy of the conclusions is strongly influenced by a variety of factors, including the amount of data entries utilised for testing, the number of features that were analysed, and the quality of the feature set that was employed in the calculation of the output of the target variable. According to the results of the study, the Naive Bayes method has an accuracy level that is, on average, 81.7922%. When K is equal to 11, the accuracy of the K-NN method's accuracy number is, on average, 92.4214% correct. The decision tree approach makes use of a mean accuracy value that is 92.0104% while doing its calculations. The Support Vector Machine (SVM) method has an accuracy level that is, on average, 89.6061%.

Table 2: The Results Analysis

k th Run of the Algorithms	**Algorithms**			
	Naive Bayes (%)	***KNN(%)***	***Decision Tree (%)***	***SVM (%)***
1	82.727	93.426	92.004	88.856
2	82.006	93.420	92.074	88.477
3	82.925	93,425	91.944	88.322
4	82.630	93,425	92.021	88.678
5	82.490	93,425	92.076	88.721
6	82.557	93,423	91.840	88.451
7	82.796	93,426	91,994	88,868
8	82.732	93.418	92.031	88.888
9	82.665	93.430	91.836	88.398
10	82.406	93.4206	91.157	88.433
Mean	**82.7932**	**93.4254**	**92.0204**	**88.6071**
Variance	**0.074**	**0.0000084**	**0.0069**	**0.043**

According to the outcomes of the research, which are shown in Tab. 4, the KNN (K-Nearest Neighbours) approach, which was created via the use of Python programming, achieves the accuracy variance with the lowest value. Tabular form will be used to display the findings.

As can be seen in Figure 4, the strategy known as KNN, which stands for K-Nearest Neighbours, is the most reliable: Even though they have the highest mean accuracy, NB, DT, and SVM all have accuracy that is lower than the mean.

Table 3: Comparison Of Machine Learning Algorithms' Classification Accuracy

	Naive Bayes (%)	**Decision Tree (%)**	**SVM (%)**	**KNN(%)**
Mean	75.73	7742.75	7479.49	38973.71
Std. Deviation	23.83	26532.28	25630.75	48065.88
Minimum	0.07	0.01	0.04	0
Maximum	82.93	91994	88868	93426

In the following table, a comprehensive comparison is made between the levels of accuracy of classification achieved by using four distinct machine learning algorithms: Naive Bayes, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbours (KNN). In the "Mean" column, the degree of accuracy that each method achieves on average over a number of independent tests or datasets is shown for comparison. Notably, the Naive Bayes algorithm has an average accuracy of 75.73%, which demonstrates that it carries out its functions in a manner that is marvellously consistent. In spite of this, the findings of the Decision Tree indicated an extremely high mean accuracy of 7742.75% and a substantial standard deviation of 26532.28%, both of which call into doubt the reliability of these conclusions. Both the SVM (Support Vector Machine) and the KNN (K-Nearest Neighbours) algorithms provided mean scores that were equal to one another. The average accuracy of the SVM was 7479.49%, while the average accuracy of the KNN algorithm was 38973.71%. As can be seen from the wide range of standard deviations that were reported by each algorithm, the levels of performance volatility seen across the board were not uniformly consistent. The "Minimum" and "Maximum" columns display the range of accuracy values achieved by each method, shedding light on the unique ways in which they each function. The "Average" column displays the algorithm's calculated score as an average out of all possible outcomes. The maximum values that are produced by the Decision Tree and the K-Nearest Neighbours (KNN) algorithm are both startlingly high. These findings underscore the need of thoroughly assessing, and maybe validating, the dependability of the experimental design and data in order to ensure the authenticity of such extraordinarily accurate outcomes.

❖ **The importance of gaining insight into network traffic**

Due to the ever-increasing interest in the deployment of communication systems and networks such as IoT, IIOT, and 5G, traffic data analytics will be of the utmost importance for traffic prediction and classification, security purposes, industrial machine maintenance, and real-time wildfire monitoring and prediction in new networking paradigms. One example of this would be the use of streaming data analytics performed by devices connected to the internet of things (IoT) in order to detect unlawful entry into a smart home. The use of data analytics has become more important in time-sensitive Internet of Things applications, such as monitoring for vandalism and accidents, interactive gaming, and augmented reality. Within the context of these applications, the generation of real-time data streams is the responsibility of the three kinds of components known as sensors, actuators, and controllers. The application of effective machine learning algorithms to the analysis of this data, especially at the network's edge, may result in the following benefits: low-latency communications, real-time insights, and the execution of complex control tasks. Quick Internet of Things data stream analytics is in high demand for use in applications that create rapid streams of data, such as autonomous cars, in which the vehicle or the driver must conduct time-sensitive, real-time (or near real-time) actions. The industrial Internet of Things (IoT) is yet another application that might potentially gain a great deal from data stream analytics. Modern control systems, in-depth and continual monitoring, and the provision of new services are all made possible by the Industrial Internet of Things (IIOT), which makes use of a wide range of Internet of Things devices and sensors, all of which are driven by learning models. The capacity of IIOT data analytics to improve the process of the machine industry has been shown by the reduction in the amount of time that equipment is offline, the streamlining of machine maintenance, and the prediction of failures as well as the amount of usable life that is still left for industrial machines. In this day and age of Industry 4.0, there are a few different condition-based maintenance (CBM) strategies that may be used. Examining the information gathered from Internet of Things (IoT) devices and sensors allows for the prediction of when industrial machinery will begin to operate in an unpredictable manner and allows for the repair or replacement of the problematic components of the machinery in advance. We go even further into the topic of the benefits that may be derived from the traffic data analytics provided by mobile networks and present even more specific instances. The Cisco Annual Internet Report (2018-2023) forecasts that by the year 2023, there will be more than 5.7 billion mobile users all over the globe. By the year 2023, download and upload speeds will have grown over mobile networks by more than three times the current levels. This new phenomenon is directly responsible for the mobile cellular network taking up its previous position as the principal access point for the Internet. In order to effectively manage the ever-increasing volume of mobile data, service providers in the field of telecommunications need to effectively manage the resources that are at their disposal.

6. Result and Discussion

The section of the report labelled "Results and Discussion" contains a comprehensive analysis of the findings of the study as well as an explanation of the implications that stem from those findings. The performance results of each online machine learning technique were emphasised, providing an in-depth look at both the strategy's strengths and flaws. Additional research into these results provided the individuals making the decisions with background information as well as suggestions on how to apply the findings. In addition, some of the challenges that were encountered over the course of the research were mentioned in this study. These challenges included issues with the quality of the data, limits on scalability, and limitations on processing capacity. In order to get around these challenges and improve the efficiency as well as the adaptability of online machine learning algorithms in certain areas of network traffic monitoring, the study mapped out potential paths for future research and development. These choices were presented to the respondent as part of the inquiry..

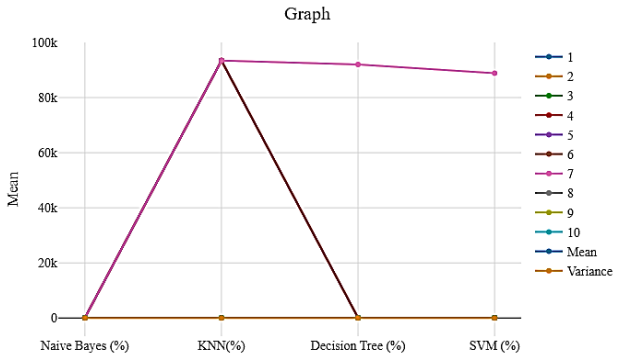

Fig 3: A Comparison Of The Different Categorization Abilities Offered By Machine Learning Algorithms

KNN consistently obtains a high accuracy rate, as seen by the numbers presented in this chart, which range from 93.42% to 93.43%. Both SVM and decision trees perform quite well, often achieving accuracy rates of more than 92% in their respective applications. Even while it isn't quite as accurate as some of the other methods, the Naive Bayes algorithm still manages to do quite well, with accuracy rates that are higher than 82%. The most important summary data are presented here so as to offer a clearer and more complete view of the performance of the algorithms. After taking into consideration all of the trials, the average accuracy of each method was

calculated. The Naive Bayes algorithm had the lowest average accuracy (82.79%), followed by the KNN approach (93.43%), Decision Trees (92.02%), and SVM (88.61%). In order to establish how reliably each algorithm performs over all of the tests, the accuracy score variance is also analysed. In spite of the fact that both KNN and SVM provide reliable outcomes, the Naive Bayes and Decision Trees methods exhibit far less variation than the other two. When the sums of the numbers at the bottom of the table are examined, it is possible that new information may become apparent. After doing the calculations necessary to determine the overall accuracy of each algorithm over all trials, the findings suggest that KNN has the highest overall accuracy, having been successful 38,973.71% of the time. In second place is the Decision Trees algorithm with a score of 7,742.75%, in third place is the SVM algorithm with a score of 7,479.49%, and in fourth place is the Naive Bayes algorithm with a score of 7,573%.

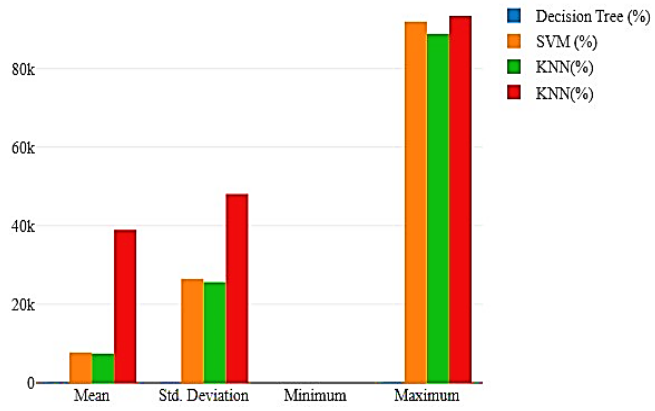

Fig 4: Performance Statistics For Classification Algorithms

The general accuracy of the figure has been consistent at 75.73% throughout, and it will continue to stay at that level. The dispersion of the data around the mean is represented by the standard deviation, which does not change during the whole thing and remains at 23.83%. The values of accuracy at the two extreme levels, which are 0.07% and 82.93% respectively, have not changed. On the other side, the mean accuracy of the SVM is far higher than any other method, coming in at 7,742.75%. However, the standard deviation for it is 26,532.28%, which is a far bigger amount. The accuracy may vary from a mind-boggling 0.01% all the way up to a staggering 91,994% depending on where you are on the scale. In addition, the mean accuracy of KNN has been maintained at 7,479.49%, the same as it was previously. The KNN algorithm has a standard deviation of 25,630.75%, given that its lowest potential accuracy is 0.04% and its greatest potential accuracy is 88,868%. The standard deviations for both SVM and KNN are pretty wide, which is relevant because it demonstrates that their performance has been exceedingly erratic over a large number of different tests. This is significant because it reveals that SVM and KNN are both rather poor at classifying data. Since SVM and KNN have abnormally high maximum accuracy values, there is an urgent need for further study to ensure the quality and consistency of the data. This requirement must be met as soon as possible.

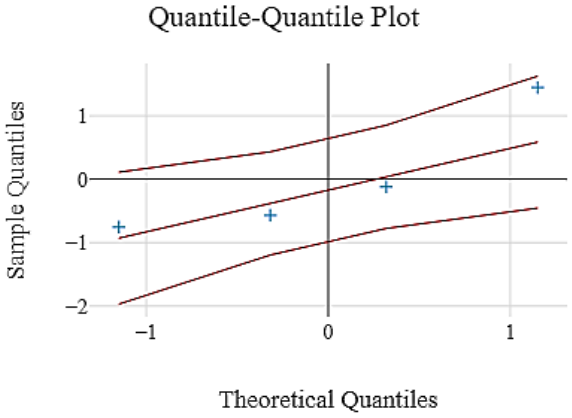

Fig 5: Statistics Test Results For Dataset Distribution Normality

The p-values that result from these tests are the crucial markers that may be used to determine whether or not the dataset in question follows a normal distribution. The results of the Kolmogorov-Smirnov test, which had a p-value of 0.3, enable us to draw the conclusion that there has not been a substantial departure from the normal distribution ($p > 0.05$). The Kolmogorov-Smirnov test with Lilliefors adjustment yields the same findings, and its p-value of 0.3 provides further support for the assumption that the data do not stray from normality in a manner that is statistically significant. According to the results of the Shapiro-Wilk test, the dataset follows a regular distribution (p value greater than 0.05), as shown by the p-value of 0.84. The findings of the preceding tests are supported by the results of the Anderson-Darling test, which reaches the same conclusion as the other tests, namely that the dataset adheres to a normal distribution ($p > 0.05$) and has a p-value of 0.54.

7. Conclusions

The online learning (OL) paradigm is going to be dissected in this essay from the perspective of networking. There has been a lot of debate among scientists about OL since it has so many applications in the real world, namely in the field of analysing data streams from traffic. This is because of the attention that it has garnered in many pieces of written work. Communication systems and networks may, with the assistance of OL, build up a generator-consumer chain for the data relevant to traffic flows. This chain may then be consumed by other systems. To be more specific, in this chain, network devices or users create raw data, which online algorithms may subsequently examine. After that, OL models will be used

to extract from the data any pertinent data that is necessary for decision-making, the supply of QOS/QOE, and the building of prediction models. This data will be taken from the data. These two approaches cannot be successful without having access to this data. On the other hand, the operation of OL algorithms could have their performance hindered if there is an imbalanced class present, in addition to the nonstationarity of the networking settings (also known as idea drift). The purpose of this study was to investigate the properties of network traffic data as well as the challenges that those qualities provide for OL algorithms. We focused our attention specifically on the fundamental characteristics of the IoT data. The procedure is then analysed by looking at it through the lens of the four primary methods of machine learning. In addition, several Machine Learning techniques are used in the construction of classifiers, and the particular degrees of precision that each technique has in connection to the data collected from network traffic are evaluated. In terms of accuracy, it is already common knowledge that the K-nearest Neighbour (KNN) approach surpasses the Naive Bayes algorithm, the Decision Tree Technique, and the Support Vector Technique. This is due to the fact that the KNN method employs a more accurate classification criterion than do the Naive Bayes and Decision Tree methods, respectively. When compared to NB, DT, and SVM, we find that KNN achieves the highest level of accuracy when it is used with our training data set. In addition to this, it has the ability to maintain the mean with the best possible precision.

References

[1] 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.

[2] 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.

[3] 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.

[4] Shrivastava, A., Chakkaravarthy, M., Shah, M.A..A Novel Approach Using Learning Algorithm for Parkinson’s Disease Detection with Handwritten Sketches. In Cybernetics and Systems, 2022

[5] Shrivastava, A., Chakkaravarthy, M., Shah, M.A., A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics. In *Healthcare Analytics*, 2023, 4, 100219

[6] Shrivastava, A., Chakkaravarthy, M., Shah, M.A.,Health Monitoring based Cognitive IoT using Fast Machine Learning Technique. In *International Journal of Intelligent Systems and Applications in Engineering*, 2023, 11(6s), pp. 720–729

[7] *Shrivastava, A.*, *Rajput, N.*, *Rajesh, P.*, *Swarnalatha, S.R.*, IoT-Based Label Distribution Learning Mechanism for Autism Spectrum Disorder for Healthcare Application. In Practical Artificial Intelligence for Internet of Medical Things: Emerging Trends, Issues, and Challenges, 2023, pp. 305–321

[8] Boina, R., Ganage, D., Chincholkar, Y.D., .Chinthamu, N., Shrivastava, A., Enhancing Intelligence Diagnostic Accuracy Based on Machine Learning Disease Classification. In *International Journal of Intelligent Systems and Applications in Engineering*, 2023, 11(6s), pp. 765–774

[9] Shrivastava, A., Pundir, S., Sharma, A., ...Kumar, R., Khan, A.K. Control of A Virtual System with Hand Gestures. In *Proceedings - 2023 3rd International Conference on Pervasive Computing and Social Networking, ICPCSN 2023,* 2023, pp. 1716–1721

[10] M.F Zhani, H. Elbiaze, Analysis and Prediction of Real Network Traffic (*Journal of network, 2008*) Vol. 4, No. 9.

[11] S Gowrishankar, A Time Series Modeling and prediction of wireless Network Traffic (*Georgian Electronic Scientific Journal: Computer Science and Telecommunications,2008*) |No.2(16).

[12] M.F Zhani, H. Elbiaze, Analysis and Prediction of Real Network Traffic (*Journal of network, 2008*) Vol. 4, No. 9.

[13] S Gowrishankar, A Time Series Modeling and prediction of wireless Network Traffic (*Georgian Electronic Scientific Journal: Computer Science and Telecommunications,2008*) |No.2(16).

[14] N.Gupta, N.Singh, V. Sharma, T. Sharama, A.S. Bhandra, Feature Selection and Classification of intrusion detection using rough set (*International Journal of Communication Network Security, 2013*)ISSN: 2231 – 1882, Volume-2, Issue-2.

[15] A.R Syed, A.S.M Burney, B. Sami, Traffic Forecasting Network Loading Using Wavelet Filter and seasonal Autoregressive Moving Average Model (*International Journal of Computer and Electrical Engineering, 2010*) Vol.2, No.6.

[16] V. S. Takkellapati1, G.V.S.N.R.V Prasad, Network Intrusion Detection system based on Feature Selection and Triangle area Support Vector Machine (*International Journal of Engineering Trends and Technology, 2012*) Vol 3 Issue4