Analysis of Video Traffic over Internet using Machine Learning Algorithms

Project report submitted for

Vth Semester Minor Project-II

in

Department of Computer Science Engineering

By,

Atilli Sanjeet-18100011

Abhishek Pragada-18101031

Sayyed Bashar Ali-18100051

Dr. Shyama Prasad Mukherjee

International Institute of Information Technology, Naya Raipur

(A Joint Initiative of Govt. of Chhattisgarh and NTPC)

Email: iiitnr@iiitnr.ac.in, Tel: (0771) 2474040, Web: www.iiitnr.ac.in

CERTIFICATE

This is to certify that the project titled "**Analysis of Video Traffic over Internet using Machine Learning Algorithms**" by "Atilli Sanjeet, Abhishek Pragada, Sayyed Bashar Ali" has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree/diploma.

(Signature of Guide)

Dr. Venkanna U

Assistant Professor, CSE

Department of Computer Science Engineering

Dr. SPM IIIT-NR

December, 2020

DECLARATION

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature of Author)

Sayyed Bashar Ali

(18100051)

(Signature of Author)

Atilli Sanjeet

(18100011)

(Signature of Author)

Abhishek Pragada

(18101031)

Date : 02-December-2020

ABSTRACT

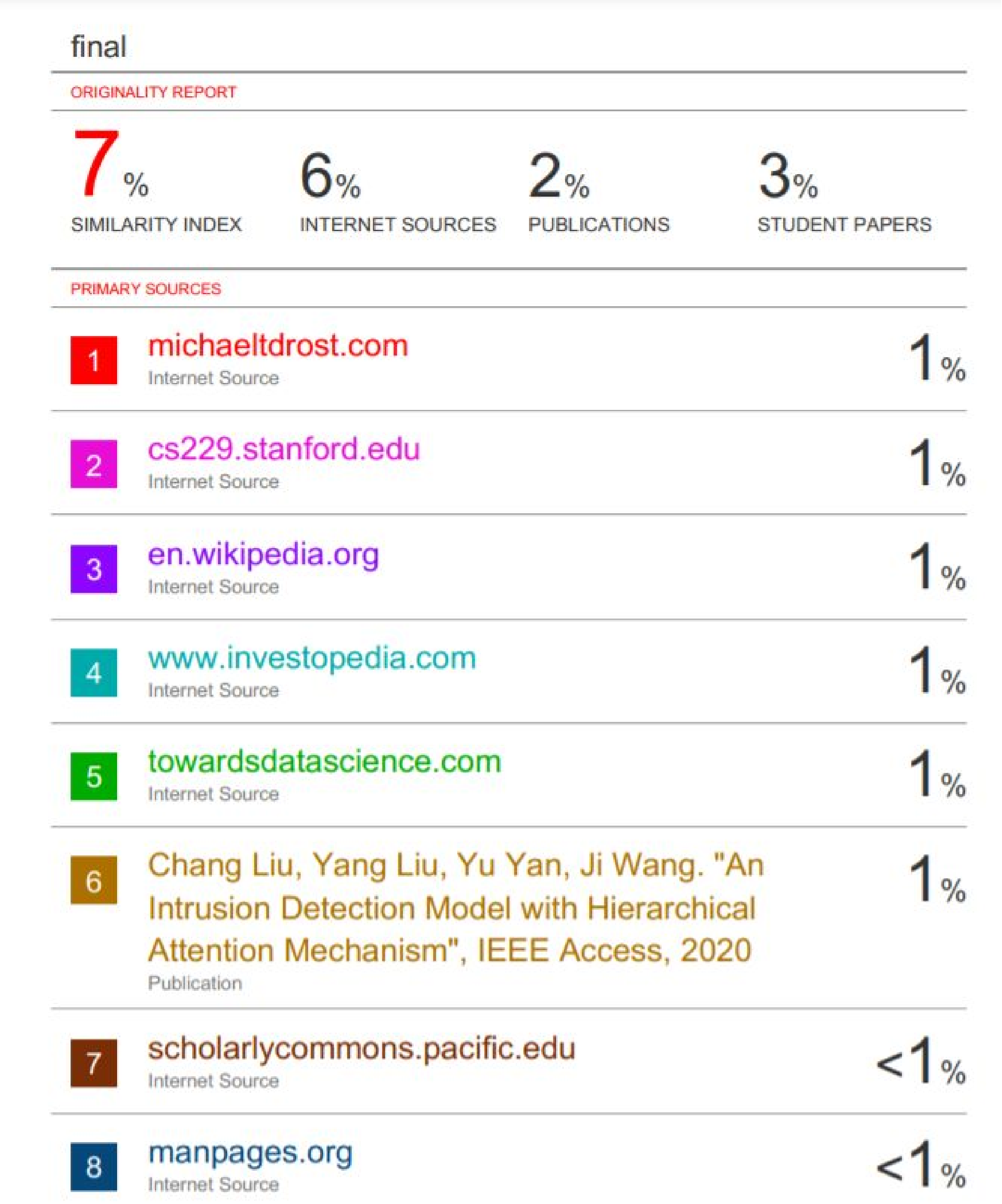
Classification of internet traffic has become quite a necessity these days for internet service providers. The obsolete method of classification to gain advantage and to capture insights of network used methods like port-based classification and deep packet inspection which requires manual work. This method can’t identify encrypted traffic and applications are using dynamic ports nowadays. This paper proposes the classification of video traffic using state of the art technique and technology using Machine Learning. As the video traffic consists of 77% of total traffic and may reach 82% by the end of 2023, this traffic needs to be channeled properly to save bandwidth. This will increase QoS (Quality of Service) for ISPs. The classification is based on three classes is Subscription-based traffic which consists of payment based services like Netflix taking 15% of video traffic and Amazon Prime which takes 3.7 % of video traffic, Non-subscription based video traffic which consists of free streaming services like YouTube and Dailymotion which takes 11.4 % of video traffic and Video Conferencing that is Google Meet and Cisco WebEx. The classification of networks is done with the help of 4 machine learning algorithms that are K-NN, SVM, Gradient Boosted Decision Tree, and XGBoost. To collect the data set 3 PCs running on both Linux and Windows have been used with internet connection speed varying from 0 Mbps to 100 Mbps. The classification of video traffic again uses stacking techniques of machine learning to further improve the accuracy of the model. Through stacking we were able to achieve 98% accuracy for a new packet.

Keywords: ISPs, Network classification, QoS

Acknowledgments

Primarily, we would like to pay a deep sense of gratitude to our supervisors Venkanna U. for providing us the opportunity to prepare this project. His valuable guidance and kind supervision towards us has served as the major contribution towards the completion of this project. His great guidance throughout the project has played a crucial part in building this project. With so much enthusiasm He has provided an extraordinary amount of feedback on my work and given me many enlightening explanations related to the theory of this thesis. In addition he has also provided the computational and hardware resources to make this possible, which I am very grateful for. Finally, I would also like to thank him for the opportunity to work on this topic for the first time from which I have learned a lot. At last thanks to M.Tech students at IIIT Naya Raipur for their guidance, support and kind supervision in the completion of this project. So with due regards, we express our gratitude to all of them.

Plagiarism Report



final

ORIGINALITY REPORT

7%	6%	2%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	michaeltdrost.com Internet Source	1%
2	cs229.stanford.edu Internet Source	1%
3	en.wikipedia.org Internet Source	1%
4	www.investopedia.com Internet Source	1%
5	towardsdatascience.com Internet Source	1%
6	Chang Liu, Yang Liu, Yu Yan, Ji Wang. "An Intrusion Detection Model with Hierarchical Attention Mechanism", IEEE Access, 2020 Publication	1%
7	scholarlycommons.pacific.edu Internet Source	<1%
8	manpages.org Internet Source	<1%

Approval Sheet

This project report entitled "**Analysis of Video Traffic over Internet using Machine Learning Algorithms**" by "Atilli Sanjeet, Abhishek Pragada, Sayyed Bashar Ali" is approved for Vth Semester Minor Project-II.

(Signature of Examiner - I)

____

Name of Examiner -I

(Signature of Examiner - II)

____

Name of Examiner -II

(Signature of Chair)

____

Name of Chair

Date: 02-December-2020 Place: IIIT Naya Raipur

Table of Contents

List of All Tables

List of All Figures

CHAPTER 1

INTRODUCTION

Due to the increase in users of the internet over years, many new challenges have evolved on data management. Videos are considered as bandwidth hogs that take and waste a lot of bandwidth. The cost of the internet is rapidly increasing in most countries so if we try to reduce the network traffic it might be able to reduce the wastage of data. Here the wastage of bandwidth refers to Bad Video Clients, Congestion in-network, improper design of buffer, unwanted updates. When during the day in office hours or in any prime time the user faces a slow internet issue which is very unsatisfactory. Implementing the technology proposed ISPs can profit a lot where they can manage different bandwidth with respect to the application which the user is accessing.[1] Through proper channeling and classification network engineers and administrators can also manage internet traffic and resolve congestion.

Videos on the internet is the field which gives out with transmission of digital video over the internet. There are many formats for the videos available on the internet with the most being MPEG-4, AVC and MP4. There are several online video hosting services, including YouTube, as well as Vimeo, Twitch, and Dailymotion out of which we have selected 6 service providers which are YouTube, Netflix, Amazon Prime, Dailymotion and Video conferencing services like Google Meet and Cisco WebEx. Each of these services uses different rules and regulations where one needs to maintain quality of service and some needs to maintain quality of product. The installed switches and routers in our home and small scale offices are not smart devices where only network algorithms are present. If some algorithm which runs in either smart devices , servers and modern switches and routers then wastage of data can be minimized which will save millions of dollars and the internet would become cheaper and accessible to all.[2] Also, existing visually analysing methods need to be improvised for making it more efficient. As they are mostly based upon the experience and local level knowledge.

What we are trying to achieve is to prioritize the video traffic. With the increase in users of OTT platforms and cheap subscription based streaming the broadband at home doesn't have a proper mechanism to categorize it. There are more than 159 million subscribers on Netflix and have the greatest market share.[3] Watching TV shows or movies on Netflix uses about 1 GB of data per hour for each stream of standard definition video, and up to 3 GB per hour for each stream of HD video . If some important conference needs to be done ,or some important system updates need to take place and connected to some low to mid speed connection and someone is streaming Netflix , Prime Video connected to router or modem the important tasks of video streaming or internet is compromised by these services and face internet connection issues or if some prioritization in quality of service like lag free streaming , fast buffering of video is needed it can be done in that order too. The paper described the machine learning algorithms which will help classifying the video traffic to the network packet level.

Major Contributions of this paper are as follows:

- We have used the Argus package to increase the number of meaningful features required for video traffic classification.

- We have made 3 classes for classification namely video conferencing, subscription-based and non-subscription based.

- The Proposed Solution, i.e., Stacking of SVM, KNN and Gradient Boosting with meta learner as XG Boost gives high accuracy without overfitting. These models are able to work in new networks.The stacked model was able to classify 1,33,551 test packets.

- Rate of false positives is all below 1 %. And average accuracy of 98% for classification of each class using stacking.

- Machine Learning doesn't take additional GPUs. Thus, they can predict the classes using CPUs and implement them over any device from low specs server to high specs servers or local applications for PCs, Mobile Phones etc .

CHAPTER 2

LITERATURE SURVEY

2.1 EXISTING WORKS

Traffic Classification is a focus point in the present scenario as reducing the traffic saves a lot of bandwidth and also one can be able to achieve high throughput demands simultaneously providing QoS (quality of service), pricing, and anomaly detection which are very essential for advanced network management techniques. In the path of attaining high accuracies, many research works are done using different methodologies. In this section we will give an overview of some of the works which include their advantages and disadvantages.

Klenilmar Lopes Dias[4] in 2019, made his work on Real-time network traffic classification using ML. This uses Naive Bayes algorithm for video classification and also it uses the help of 14 features (both derived and calculated) which are more in number, than that used in other related works. But it also had it's negatives which include lack of clarity on protocol that are used by the services like Youtube that keep shifting between UDP and TCP for video data packets and also this methodology was completely based on QoS which restricts this work's exposure.

In 2017, Ricky Andersson [5] made his work accessible, where he used two algorithms Random Forests and Gradient Boosted Trees which are compared based on their accuracies while analysing the video traffic but the output was of a mixed option where Random Forest was performing better when it came to classification time and Gradient Boosted Trees was performing better when the classification time was ignored. So he summarizes that the random forest classifier managed to achieve accuracies over 93% and declared it as the better of the two the only flaw in his work was that he was able to produce only few classes which later he mentions that would increase in his future work.

In 2019, Nyashadzashe Tamuka[6] has used K-means and Silhouette analysis on unsupervised data for classifying video traffic in his work. It Uses fields of IPv4 header like length and TCP/UDP fields arrival time between packets as features for classification. It had it's limitations like it was unable to produce high accuracies while providing an accuracy of only 86.5%. and also it was only able to classify only 2 classes which are Youtube and Netflix.

Moving on to the work of Zhang and Jun which was produced in 2014[7], shows that it has better accuracy than the above mentioned works this is because it was using multiple algorithms to get the best result like Naive Bayes, Logistic Regression, SVM and K-means and in addition to this they used only the features they require with the help of PCA which works with dimensional reduction of the dataset. As every other work, it also has some disadvantages, it was unable to classify more than 2 types of videos which makes it lack in the diversity aspect. Also they used only a small data set.

Now that we have some overview of the previous works, we will try to reduce the amount of limitations along with the aim of improving diversity and accuracy in classification which will be discussed in the next section where we wrote a detailed description of our proposed model.

2.2 SUMMARY OF EXISTING WORK

Table 2.2 summarizes all the existing works with their methodologies to solve the problem and the limitations of the methodologies.

Table 1 . Summary of Related Work

S.No	Name of the solution	Methodology	Limitations
1	Real-time network traffic classification using ML[4].(2019)	Uses Naive Bayes algorithm for video traffic classification. Uses 14 features for classification both derived and calculated.Ex: Length, Arrival time can be derived from it's Ip header and mean and variance of	Based completely on QoS. Not much clearance on protocol used by services like Youtube that keep shifting b/w UDP and TCP for video data packets.

		TTL, arrival time comes in calculated. K-means for cross validation	
2	Video Traffic Classification using Random Forests and Gradient Boosted Trees[5]	Uses different machine learning tree methodologies i.e Random Forest and Gradient Boosting Decision tree	Lower accuracy No detailed explanation of performance metrics Classes not mentioned
3	Classification of Video traffic streaming using Machine Learning[6].(2019)	Uses K-Means and Silhouette analysis on unsupervised data. Uses fields of IPv4 header like length and TCP/UDP fields arrival time b/w packets as features for classification.	Uses very few features. Accuracy of 86.5% only using k-means algorithm. Classifies only two streaming platforms i.e Youtube and Netflix. No cross validation.
4	Robust Streaming Video Traffic Classification[7].(2014)	Machine Learning approach. Used different models to get best result Algos(Naïve bayes, Logistic regression, SVMs, K-means) PCA used for dim. reduction	PCA is not very useful in feature importance. Binary classification(Video or non-video stream classes) They have captured the traffic for very less time(60 secs).

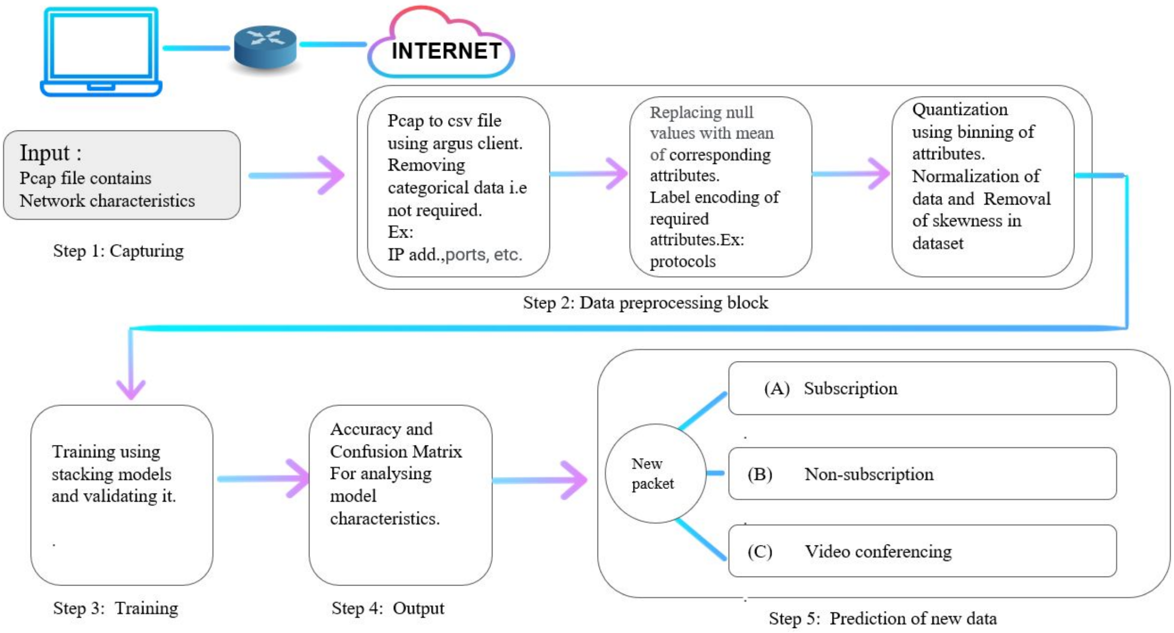
CHAPTER 3

PROPOSED MODEL

As our model deals with classification of video traffic at packet level the proposed model is distributed into three parts. Most important part is the collection of a proper dataset. Dataset will help to recognize the data properly. The following steps were followed to implement the proper classification model:-

1. Dataset construction and refinement of dataset
2. Construction of Machine Learning Model using different classification algorithms.
3. Analysis of classification and results

Below is the conceptual diagram to implement our model.



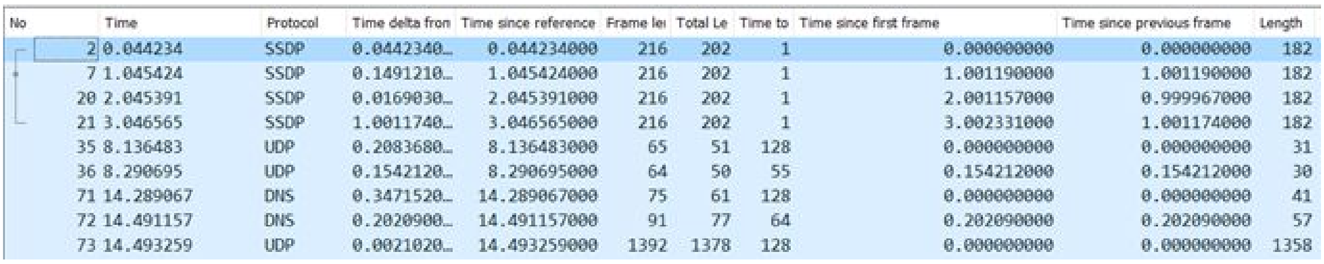
INTERNET
Input :
Pcap file contains
Network characteristics
Step 1: Capturing
Pcap to csv file using argus client. Removing categorical data i.e not required. Ex: IP add.,ports, etc.
Replacing null values with mean of corresponding attributes. Label encoding of required attributes.Ex: protocols
Quantization using binning of attributes. Normalization of data and Removal of skewness in dataset
Step 2: Data preprocessing block
Training using stacking models and validating it.
Step 3: Training
Accuracy and Confusion Matrix For analysing model characteristics.
Step 4: Output
New packet
(A) Subscription
(B) Non-subscription
(C) Video conferencing
Step 5: Prediction of new data

Fig. 1 Conceptual diagram

All the process is performed using a personal computer with 8GB of RAM and network is captured with the help of wifi and lan cable with network card installed following IEEE 802.11ac protocol for wireless transmission. The model is constructed in Windows 10 and Linux based distro Ubuntu.

3.1 DATASET CONSTRUCTION AND PREPROCESSING OF DATASET

The construction of the dataset is done using wireshark software which is installed both in windows and linux. The capturing of network packets is done in the machines which take information when different streaming services are run in the browser ,application provided by services like Netflix App , Prime Video app. The wireshark tool from default captures number, time, source and destination IPs protocol , length and info and generates a pcap or pcapng file which contains all details of the packets as shown in Fig. 2. We generated pcapng files for all the services when the video was running. The main reason to use pcapng files instead of pcap is because as pcapng contains captures from different interfaces , improved timestamp solution, additional metadata is also contained in the pcapng file and it is very extensible.



No	Time	Protocol	Time delta fron	Time since reference	Frame le	Total Le	Time to	Time since first frame	Time since previous frame	Length
2	0.044234	SSDP	0.0442340…	0.044234000	216	202	1	0.000000000	0.000000000	182
7	1.045424	SSDP	0.1491210…	1.045424000	216	202	1	1.001190000	1.001190000	182
20	2.045391	SSDP	0.0169030…	2.045391000	216	202	1	2.001157000	0.999967000	182
21	3.046565	SSDP	1.0011740…	3.046565000	216	202	1	3.002331000	1.001174000	182
35	8.136483	UDP	0.2083680…	8.136483000	65	51	128	0.000000000	0.000000000	31
36	8.290695	UDP	0.1542120…	8.290695000	64	50	55	0.154212000	0.154212000	30
71	14.289067	DNS	0.3471520…	14.289067000	75	61	128	0.000000000	0.000000000	41
72	14.491157	DNS	0.2020900…	14.491157000	91	77	64	0.202090000	0.202090000	57
73	14.493259	UDP	0.0021020…	14.493259000	1392	1378	128	0.000000000	0.000000000	1358

Fig. 2 Capturing of data

To capture and generate the dataset some constraints are set in relegation , some are used in top priority.

While capturing the dataset following constraints were set:-

1) Allowing protocols of UDP , TCP and HTTP only. These are the protocols followed by the network for video transmission. When using this protocol other protocols which are using the machine are left out like ARP requests , SMTP protocols for mails are excluded while capturing.
2) The videos were captured for 1 hour unrestricted on different machines.

3) The source and destination IP address were dropped off.
4) Length of packet size is fixed to a certain threshold.

After the packets are captured and converted to pcapng file, the file is converted to argus format and then using terminal commands to csv file using the help of argus client. Argus is a package in ubuntu which is a network utilization audit system. Through which various attributes and features of the network can be accessed. We selected 30 such features which are shown in Table (2).

Table 2. List of attributes

Features	***Description***
Rank	Ordinal value of this output flow record i.e. sequence number
Dur	Record total duration
RunTime	total active flow run time. This value is generated through aggregation, and is the sum of the records duration
IdleTime	time since the last packet activity. This value is useful in real-time processing, and is the current time - last time
Proto	transaction protocol
sTos	source TOS byte value
dTos	destination TOS byte value
sDSb	source diff serve byte value
dDSb	destination diff serve byte value
sTtl	source -> destination TTL value
dTtl	destination -> source TTL value
dMpls	destination MPLS identifier
sMpls	source MPLS identifier
NStrok	Number of observed keystrokes
TotPkts	total transaction packet count
TotBytes	total transaction bytes
TotAppByte	total application bytes
PCRatio	producer consumer ratio

Load	bits per second
Loss	packets retransmitted or dropped
Retrans	packets retransmitted
SrcGap	source bytes missing in the data stream
DstGap	destination bytes missing in the data stream
Rate	pkts per second
SIntPkt	source inter packet arrival time (mSec)
DIntPkt	destination inter packet arrival time (mSec)
SrcJitter	source jitter (mSec)
DstJitter	destination jitter (mSec)
srcUdata	source user data buffer
dstUdata	destination user data buffer
SrcWin	source TCP window advertisement
DstWin	destination TCP window advertisement.
TcpRtt	TCP connection setup round-trip time, the sum of 'synack' and 'ackdat'
Offset	record byte offset in file or stream

Using the argus client the pcapng file is converted to an argus file then to a csv file which will have the appropriate values according to its feature values. The csv files are then imported into the model. Argus can be implemented using Linux based terminal commands. After the conversion of pcapng to csv file using argus the csv files for each of the services are sent to imported in the model. The pre processing part for the dataset is done using following steps:-

1) Removal of data from each column having other value than numerical and categorical data. Some of the features were giving irrelevant values and having ambiguous values were dropped.
2) Secondarily, we check for null values. After getting the null values in respective columns we fill it with the mean values.

3) Label encoding the values of categorical type i.e protocol and removing unwanted protocols like ARPs misclassified by wireshark or skipped the check.

4) To increase the effective features binning is done.Binning is a method to generate a group of a number of more or less continuous values into a smaller number of bins or groups.The process of binning is done .for 9 attributes i.e TotBytes,Load,Loss,Ttl,Tos, win ,Gap,Offset,TcpRtt. The binning is done by bin by means , bin by boundary and bin by variance method.

5) As we are following supervised machine learning we have to insert a column which contains the output value for video conferencing subscription , non-subscription as 0 , 1 , 2 respectively based on priority order.

6) Now for feature selection we have used correlation as the basis. It reduces the number of features in our model feature size to smaller size. The correlation values greater than 75% are not statistically significant so it also is deleted.

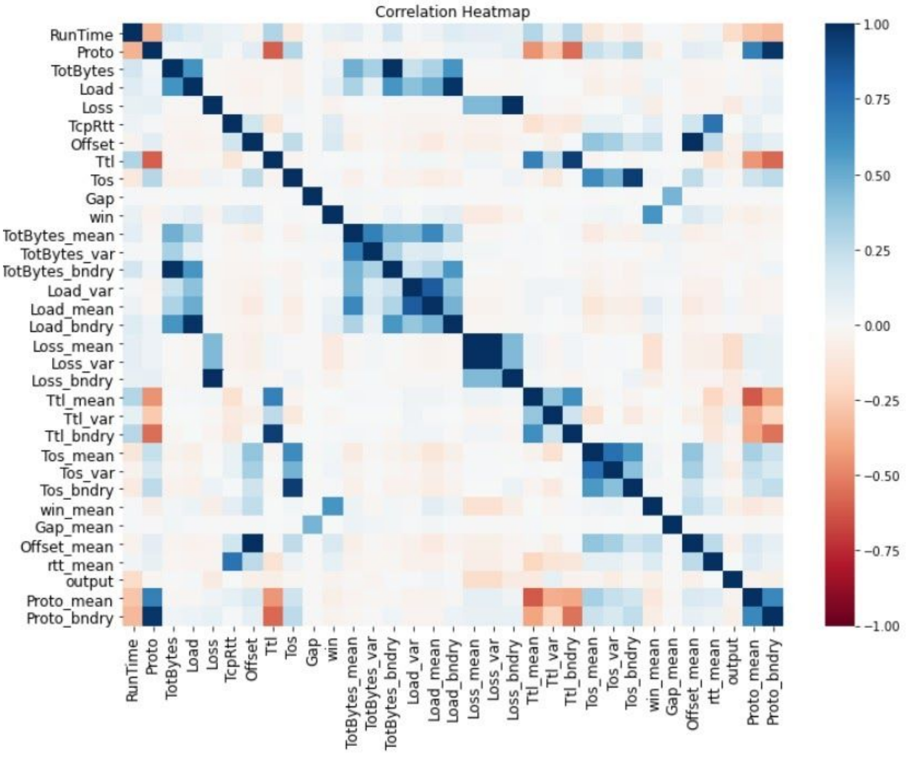


Fig. 3 Correlation between features

7) Last step in data preprocessing is skewness removal.Skewness means the asymmetry in a symmetrical normal distribution, in a set of data. If the curve is shifted to the left or to the

right, it is said to be skewed. Using predefined skew() of python library it can directly be implemented.

3.2 MACHINE LEARNING MODEL CONSTRUCTION

From Section 2 we came to know about the disadvantages from previous works so we tried to turn them into advantages in our work. We did it so by increasing the accuracy, increasing the cardinality of features and also by defining them into classes which are really helpful in saving the bandwidth.

So, as to achieve these goals we used four algorithms K-NN, XGBoost, SVM and Gradient Boosted Decision Trees (GBDT) which are mainly used for classification purposes to produce high accuracies.

3.2.1 XGBoost

XG Boost also known as eXtreme Gradient Boosting is an application of Gradient Boosted Decision Trees. In this model, a continuous array of decision trees are created. Weights are like the heart of this model as most of the model’s concept involves weights where all the independent variables are allocated weights and then fed into decision trees which further produce the results.

So on running the XG Boosting algorithm on the dataset produced, was giving us the results shown in the below Fig. 4

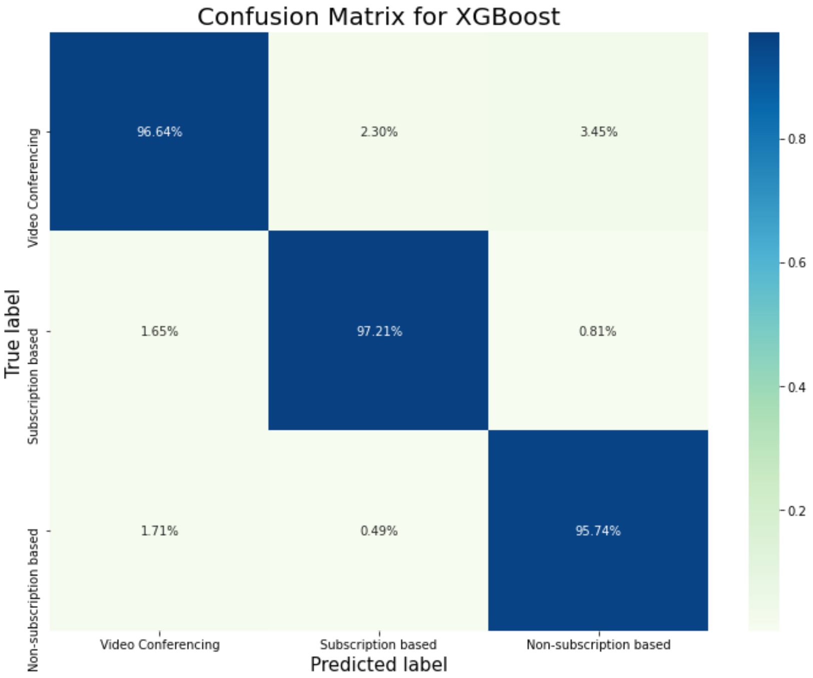


Fig. 4 Confusion matrix of XG Boost

Also it was giving the following overview on accuracies :

```
Training Accuracy ->0.965

Validation Accuracy ->0.964

Final Testset Accuracy ->0.965

Precision score ->0.965

Recall score ->0.965
```

3.2.2 SVM

Support Vector Machine is comparatively a straightforward algorithm used for classification and regression. It is mostly used for classification and prefers regression for only a few cases. Predominantly SVM realizes a hyper-plane which produces a boundary between different kinds of data. In 2D space, this hyper-plane is nothing but a line.In SVM, every data attribute present in the dataset is plotted in an N-dimensional space, here N represents the cardinality of attributes in the dataset. The final step would be finding the optimal hyperplane to separate the data.

So running the SVM algorithm on the dataset produced, was giving us the results shown in the below Fig. 4

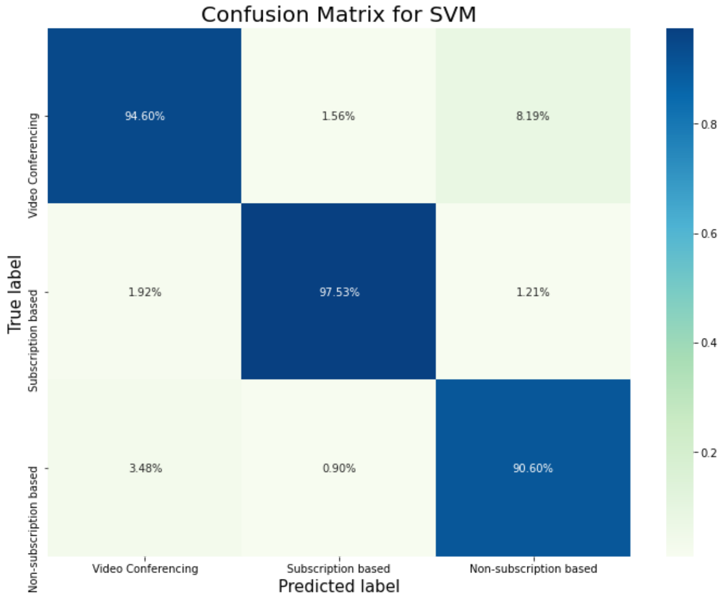


Fig. 5 Confusion matrix of SVM

Also it was giving the following overview on accuracies :

```
Training Accuracy ->0.945

Validation Accuracy ->0.944

Final Testset Accuracy ->0.940

Precision score ->0.942

Recall score ->0.940
```

3.2.3 GBDT

Gradient Boosting Decision Trees algorithm is a kind of Gradient Boosting algorithm that is used for both classification and regression. GBDT helps in the prediction of weak ensemble prediction models, which are mainly decision trees. Generally in boosting we combine a series of learning algorithms in order to achieve a strong learner from the sequentially attached weak learners, here the weak learners are the decision trees.

So on running the Gradient Boosted Decision Tree algorithm on the dataset produced, was giving us the results shown in the below Fig. 6

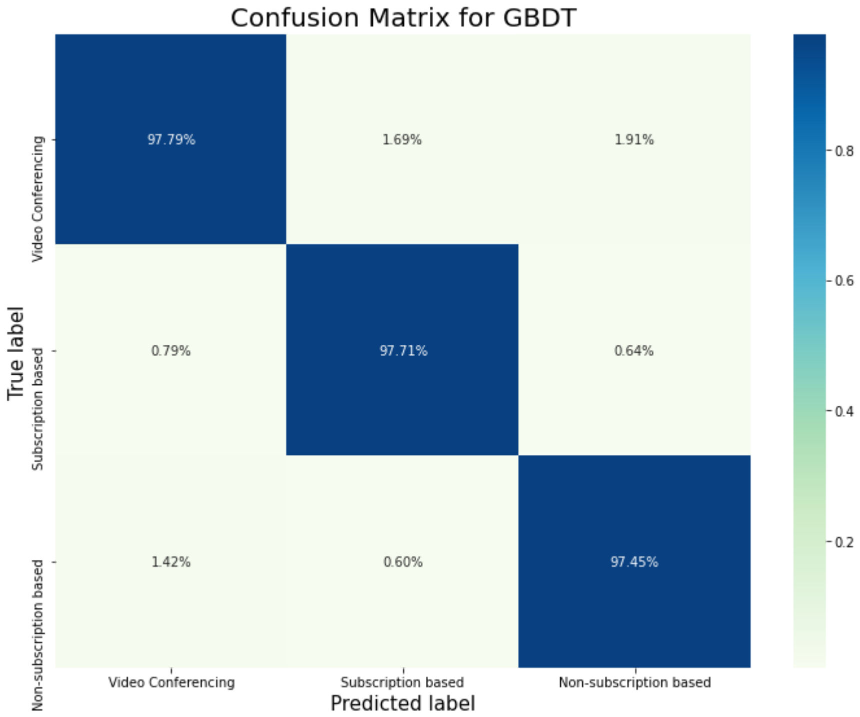


Fig. 6 Confusion matrix of Gradient Boosted Decision Trees

Also it was giving the following overview on accuracies :

```
Training Accuracy ->0.977

Validation Accuracy ->0.976

Final Testset Accuracy ->0.976

Precision score ->0.976

Recall score ->0.976
```

3.2.4 KNN

K-Nearest Neighbours (KNN) is also a prominent algorithm which is used for many classification problems like intrusion detection, pattern recognition, data mining etc.,. It fits in the supervised learning realm. It can be used in most of the real life situations as it is doesn't make any false presumptions about the data it mainly uses attributes to classify the data

So running the KNN algorithm on the dataset produced, was giving us the results shown in the below Fig. 7

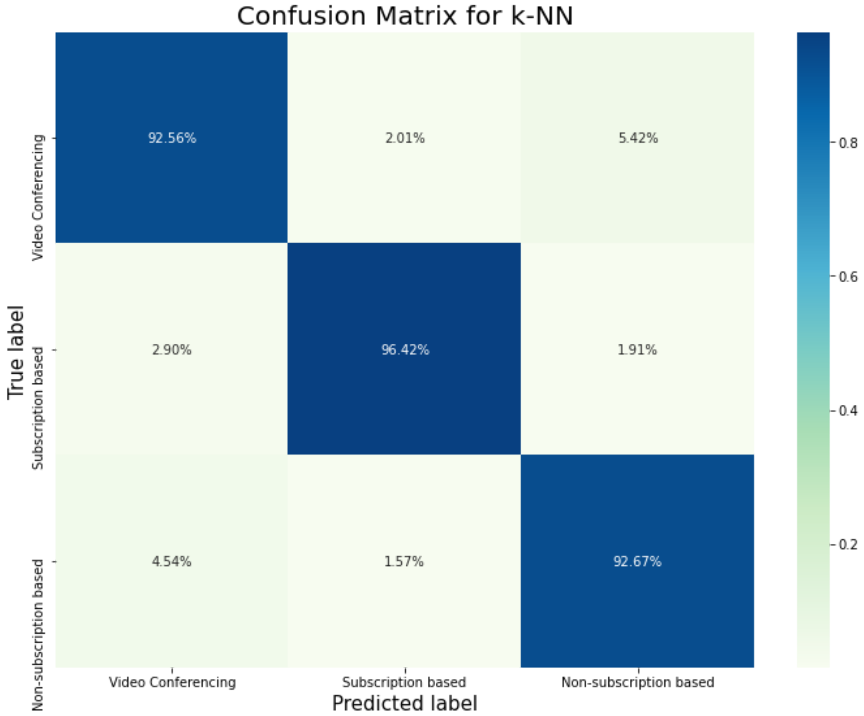


Fig. 7 Confusion matrix of KNN

Also it was giving the following overview on accuracies :

```
Training Accuracy ->0.944

Validation Accuracy ->0.940

Final Testset Accuracy ->0.938

Precision score ->0.939

Recall score ->0.938
```

On implementation of all the above models on the datasets produced, we are getting a reasonable accuracy of 90~93% but to further improve it we have a method called stacking which we have implemented in our work and we will be discussing it below.

3.2.5 STACKING OF MODELS

As it can be inferred from the above section our main goal is to process a new algorithm of stacking in machine learning. The above single algorithms gave us sufficient accuracy though to have a precisive accuracy everytime and lower false positive percentage we used stacking algorithms.

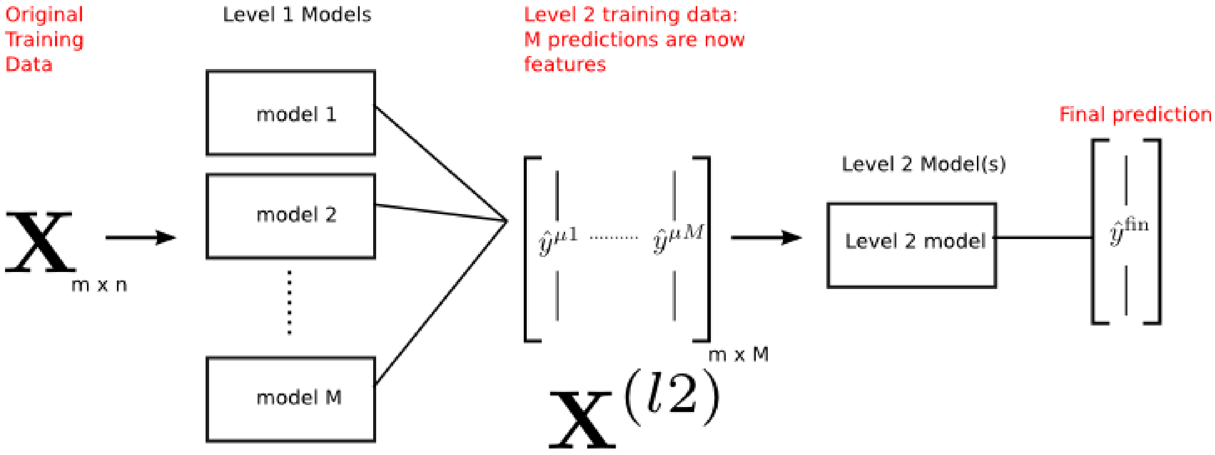


Fig. 8 Conceptual diagram of Stacking

Stacking learns to combine the base models using a meta-model. The algorithm behind stacking is to learn the various weak learning models by combining them by the help of a metamodel to generate predictions that are generated from individual weak learners. There are 2 things needed to fit this model first are the base models and secondly a meta-model. We have used this model with the help of 3 classifiers as a base model; those are Decision Tree SVM ,KNN and XGBoost to learn one neural network as a meta model.

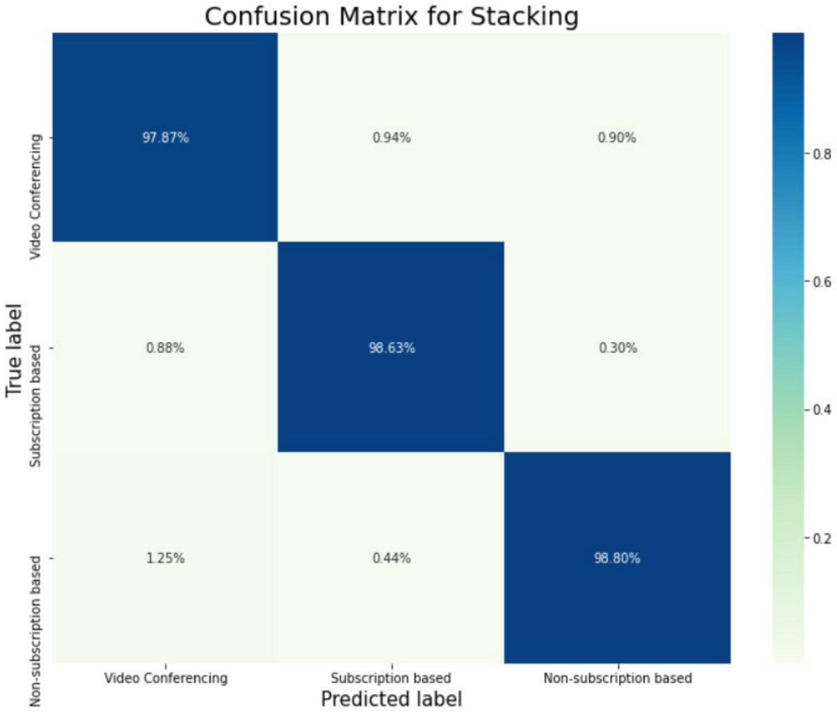


Fig. 9 Stacking result

```
Training Accuracy ->0.989

Validation Accuracy ->0.985

Final Testset Accuracy ->0.984

Precision score ->0.984

Recall score ->0.984
```

As we can observe from the above stacking model which was giving undesirable results we can see significant increase in accuracy and true label classification over validation data. Video conferencing was our priority so false positive rates needed to be minimized and stacking helped to reduce it to 0.90 % for Non-subscription class and 0.94% for subscription class the rest of the values can be inferred from Fig. 9

3.3 Comparison with Existing Solutions

After deployment of the model on our data set we got the factors like Accuracy, Precision and Recall. So, we tried to compare these factors with the previous models mentioned in Chapter 2 and visualize the progress we made, as shown in Table 3.

Table 3. Performance comparison of the proposed model with the existing works

Research Paper by	Model Used	Accuracy	Precision	Recall	Classes
Klenilmar Lopes Dias* , Mateus Almeida Pongelupe [4]	Naive Bayes	98.8%	72.01%	97.65%	3
Andersson R.[5]	Random Forest and Decision Tree	93%	94.6%	92.2%	2
Nyashadzashe Tamuka1 , Khulumani Sibanda [6]	K-NN	86.5%	Nil	Nil	2
Zhang, Jun & Chen, Xiao & Xiang[7]	SVM	95.1%	94.8%	94.7%	2
*Our Model	Stacking	98.4%	98.4%	98.4%	3

3.4 Practical Application of the Proposed Solution

Our proposed algorithm are beneficial to both ISPs and on client side where an application would be running in desktop which can classify the network packets but would take RAM from local machine but if implemented on server side of ISPs would help ISPs to classify the incoming packets and cut their loss which will save a lot of money.

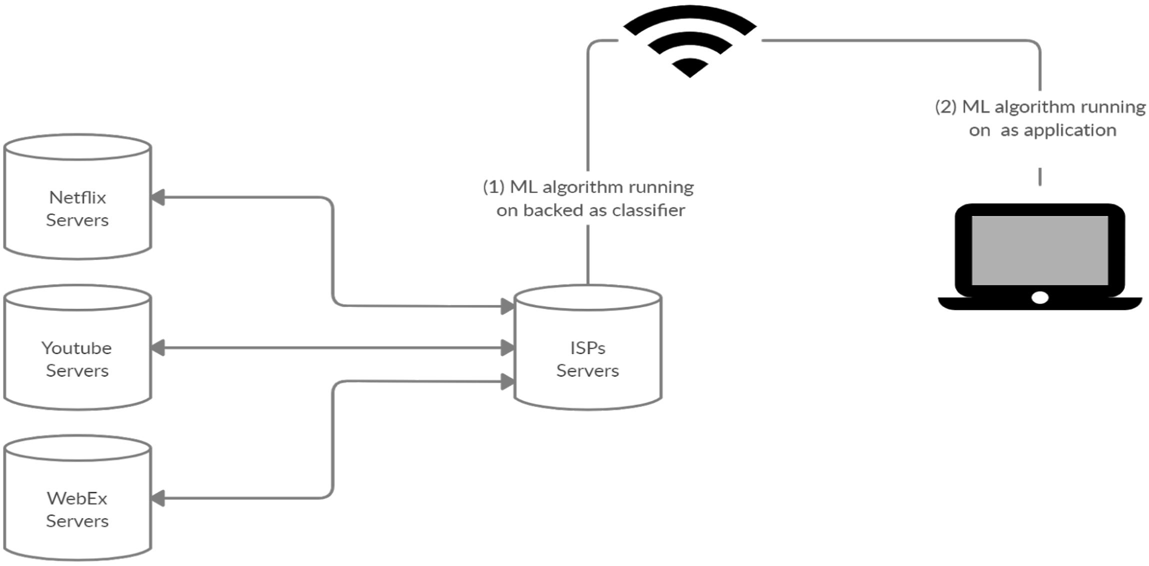

(2) ML algorithm running on as application
(1) ML algorithm running on backed as classifier
Netflix Servers
Youtube Servers
WebEx Servers
ISPs Servers

Fig. 10 Practical implementation

CHAPTER 4

RESULTS

This section explains the results of models which we have constructed for the video network traffic classification. We have classified the video traffic into three categories, i.e Video conference, Subscription based and non-subscription based services network. Now our proposed model is evaluated based on four performance metrics, i.e Accuracy, Precision, Recall, and F1-score. We also provide in-depth analysis of our proposed model with the help of a confusion matrix, which explains to us how much percent of data was predicted true and what were the percentages of false positives and false negatives in our model.

Table 4. Comparative analysis of different models

Model	Train accuracy	Test accuracy	Precision	Recall	F1-score
Random Forest	91.4	91.3	91.7	91.3	91.50
k-NN	94.4	93.8	93.9	93.8	93.85
SVM	94.5	94.0	94.2	94.0	94.09
XGBoost	96.5	96.5	96.5	96.5	96.5
GBDT	97.7	97.6	97.6	97.6	97.6
Stacking	98.9	98.4	98.4	98.4	98.4

From the above table we can see that we are achieving an accuracy of approximately ~99% with the help of our Stacked model.

CHAPTER 5

CONCLUSION

Traffic classification is a critical capacity that enables effective Internet network management. As it can be concluded from the above analysis our model outperforms the other model when implemented using a stacking algorithm instead of modifying a single algorithm. Stacking model is fast and can be implemented on low spec servers. Stacking uses a neural network approach for it's calculation but unlike deep learning uses less time and computations. This work has also detailed about the single models and how it fails to achieve desired results. The single model analysis of KNN, SVM , XGBoost has been provided in the above sections. The system fits the requirements of an effective, practical, and efficient modern Internet traffic classifier.

REFERENCES

[1] Singh, K. and Agrawal, S., 2011, April. Comparative analysis of five machine learning algorithms for IP traffic classification. In 2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC) (pp. 33-38). IEEE.

[2] Li, W. and Moore, A.W., 2007, October. A machine learning approach for efficient traffic classification. In 2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (pp. 310-317). IEEE.

[3] https://help.netflix.com/en/node

[4] Klenilmar Lopes Dias, Mateus Almeida Pongelupe, Walmir Matos Caminhas, Luciano de Errico, “An innovative approach for real-time network traffic classification”, Computer Networks, Volume 158, 2019, Pages 143-157, ISSN 1389-1286,

[5] Andersson R. “Classification of Video Traffic: An Evaluation of Video Traffic Classification using Random Forests and Gradient Boosted Trees”. Karlstad University, Sweden, 2017.

[6] Tamuka, Nyashadzashe & Sibanda, Khulumani. (2019). Modelling the Classification of Video Traffic Streaming Using Machine Learning.

[7] Zhang, Jun & Chen, Xiao & Xiang, Yang & Zhou, Wanlei & Wu, Jie. (2014). Robust Network Traffic Classification. IEEE/ACM Transactions on Networking. 23. 1-1. 10.1109/TNET.2014.2320577.

[8] Zander, S., Nguyen, T. and Armitage, G., 2005, November. Automated traffic classification and application identification using machine learning. In The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05) l (pp. 250- 257). IEEE.

[9] Robust Streaming Video Traffic Classification Jordan Ebel, jebel@stanford.edu

[10] https://manpages.debian.org/testing/argus-client/ra.1.en.html