ANALYSIS OF STATISTICAL DISTRIBUTION CHANGES OF INPUT FEATURES IN NETWORK TRAFFIC CLASSIFICATION DOMAIN

Authors:
Lukáš Jančička, CTU in Prague
Dominik Soukup, CTU in Prague
***Josef Koumar**, CTU in Prague*
Doc. Tomáš Čejka Ph.D, CESNET a.l.e.

QoDaNeT, NOMS 2024

MOTIVATION

What I want to do?

I want automatic detection/classification of ....

Examples:

- Detection of cat in the image
- Detection of botnet
- Classification of TLS services

MOTIVATION

How to do it?

MOTIVATION

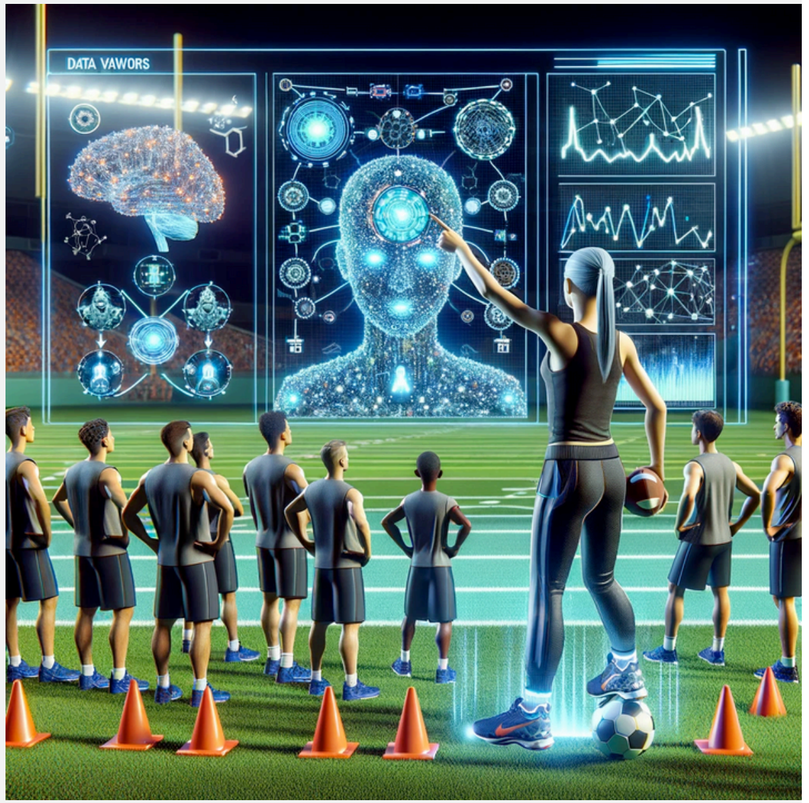
How to do it?

1. Collect the data or use existing dataset

MOTIVATION

How to do it?

1. Collect the data or use existing dataset
2. Train model



MOTIVATION

How to do it?

1. Collect the data or use existing dataset
2. Train model
3. Deploy the model and maintain the model

MOTIVATION

How to do it?

1. Collect the data or use existing dataset
2. Train model
3. Deploy the model and maintain the model

Examples:

- Detection of cat in the image --> cat photoshoot

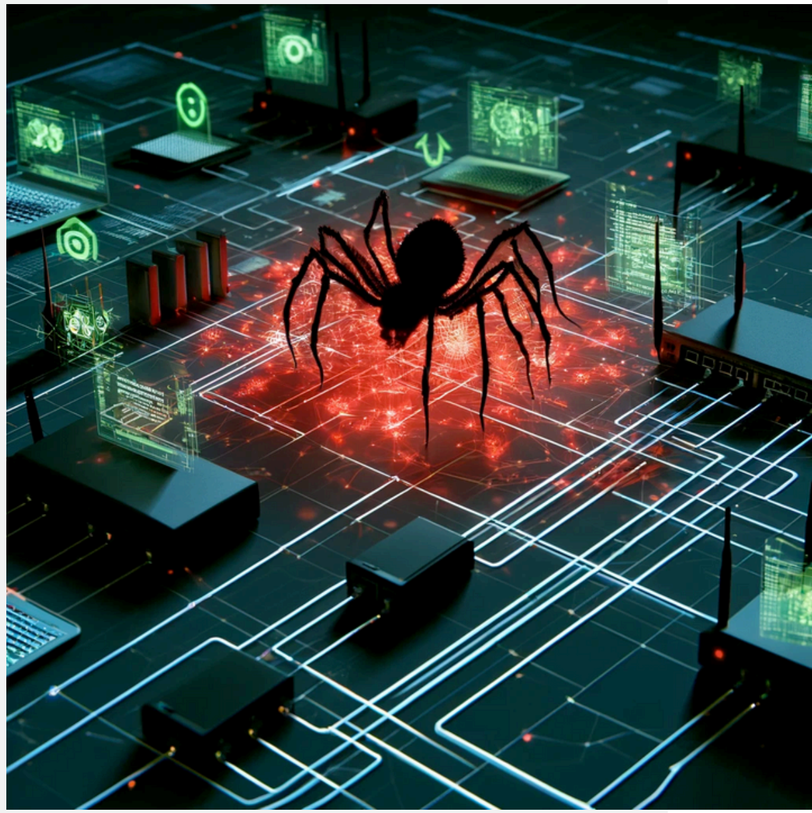
MOTIVATION

How to do it?

1. Collect the data or use existing dataset
2. Train model
3. Deploy the model and maintain the model

Examples:

- Detection of cat in the image --> cat photoshoot
- Detection of botnet --> run botnet in my network?!



MOTIVATION

How to do it?

1. Collect the data or use existing dataset
2. Train model
3. Deploy the model and maintain the model

Examples:

- Detection of cat in the image --> cat photoshoot (easy)
- Detection of botnet --> run botnet in my network?!
- Classification of TLS services --> collect data of users?!

MOTIVATION

What about data/concept drift?

MOTIVATION

What about data/concept drift?

MOTIVATION

What about data/concept drift?

MOTIVATION

What about data/concept drift?

- Will network traffic of botnet / TLS service change in time?

MOTIVATION

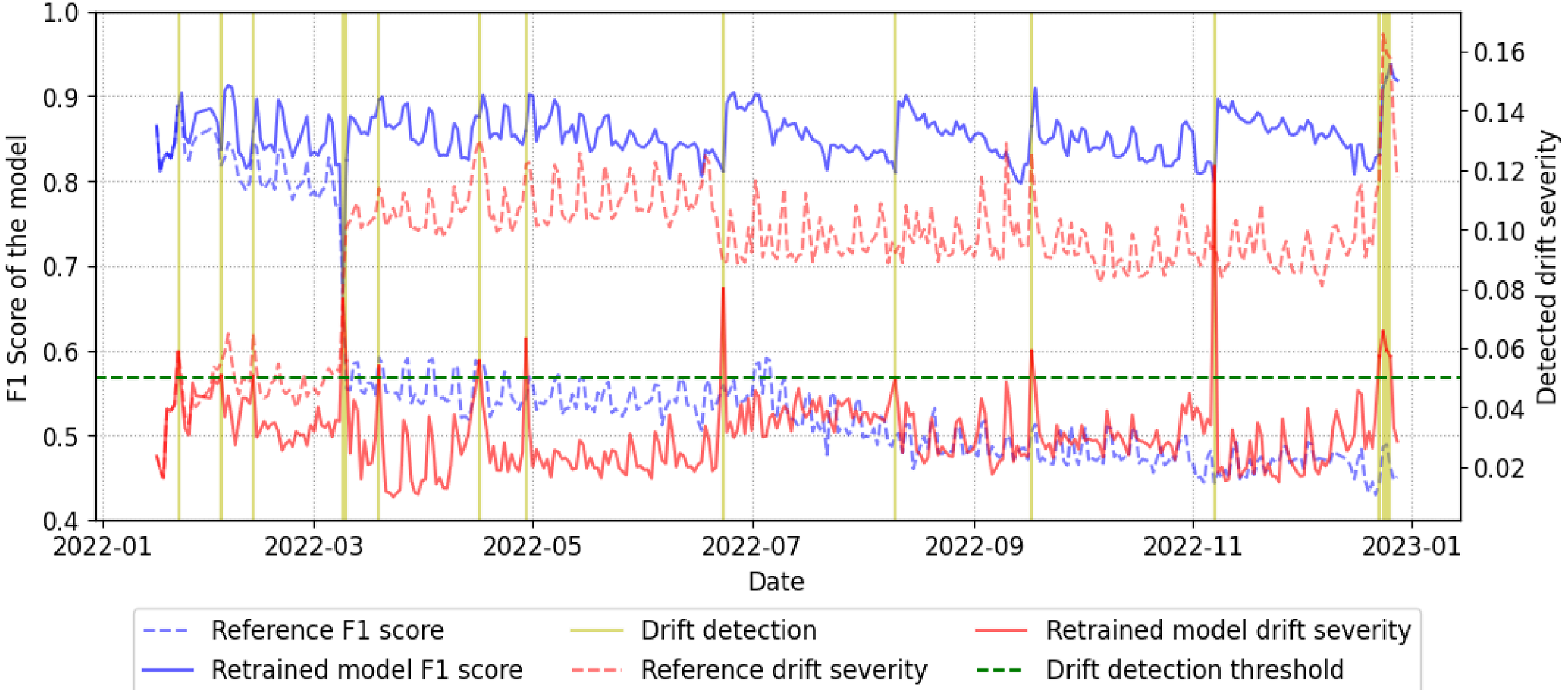
What about data/concept drift?

- Will network traffic of botnet / TLS service change in time?

--> Yes, it will. (updates of code, new versions, ...)

cesnet

MOTIVATION



MOTIVATION

What about data/concept drift?

- Will network traffic of botnet / TLS service change in time?
  --> Yes, it will. (updates of code, new versions, ...)
- Will it be a problem?

MOTIVATION

What about data/concept drift?

- Will network traffic of botnet / TLS service change in time?
  --> Yes, it will. (updates of code, new versions, ...)
- Will it be a problem?
  --> Unfortunately yes, it will. (obsolescence of models leading to accuracy drop)

PROBLEM

Motivation:

- Statistical features commonly used for ML are highly dependent on the statistical distribution of these features for each class inside the dataset. This means that even for the initial high-quality dataset, we can get a nonrelevant dataset over time due to changes in the distribution of features

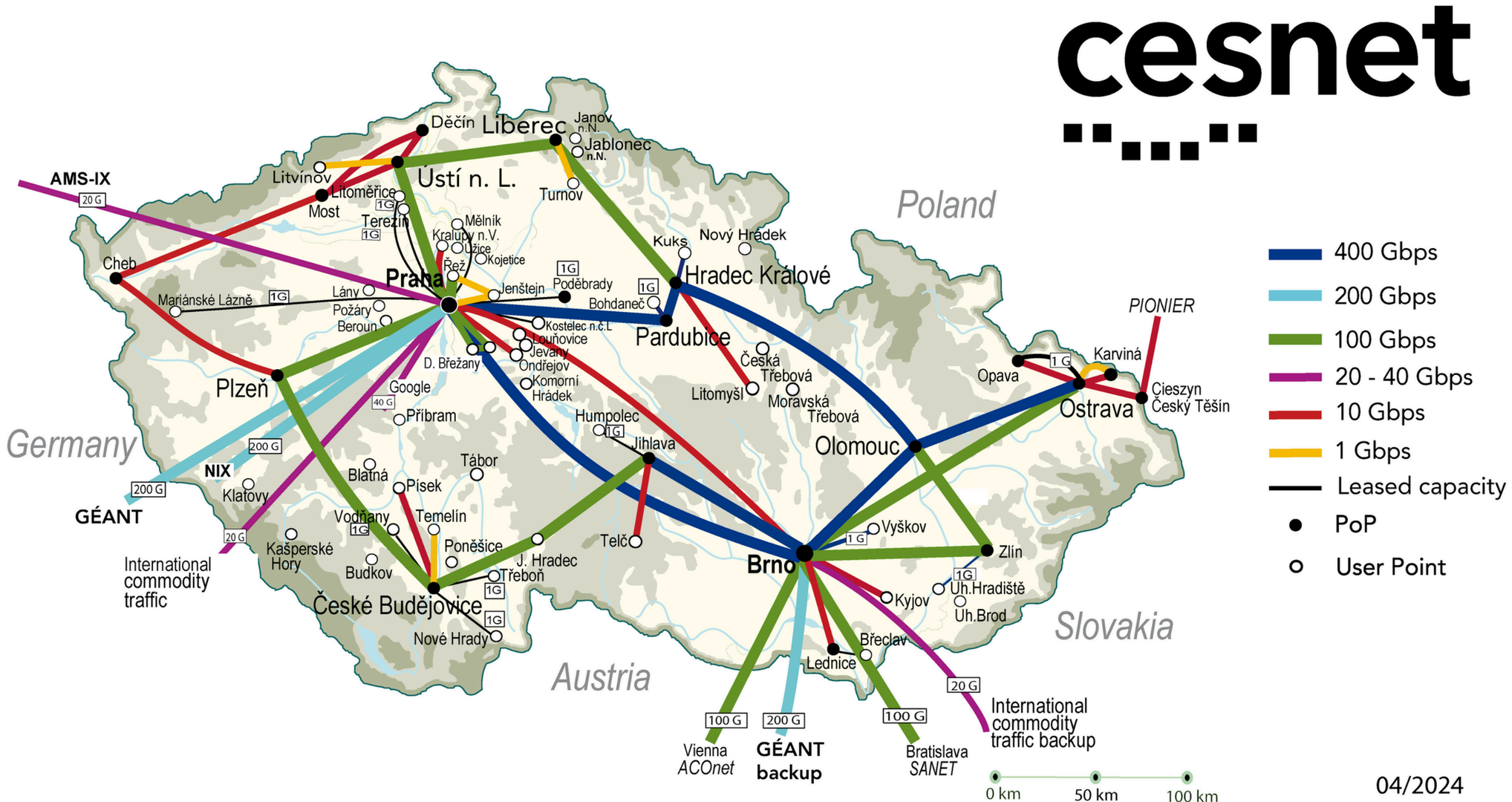
  ---> This phenomenon is called concept/data drift

Therefore, in this work, we analyse statistical distribution of long-term dataset from real-world environment.

DATASET

- **CESNET-TLS-Year22** - one-year-long dataset of TLS network traffic from CESNET3
  - For work with the dataset, we used our open-source tool **CESNET-DataZoo**
  - 507,739,073 IP flows
  - 180 classes
  - 54 features

CESNET3 NETWORK

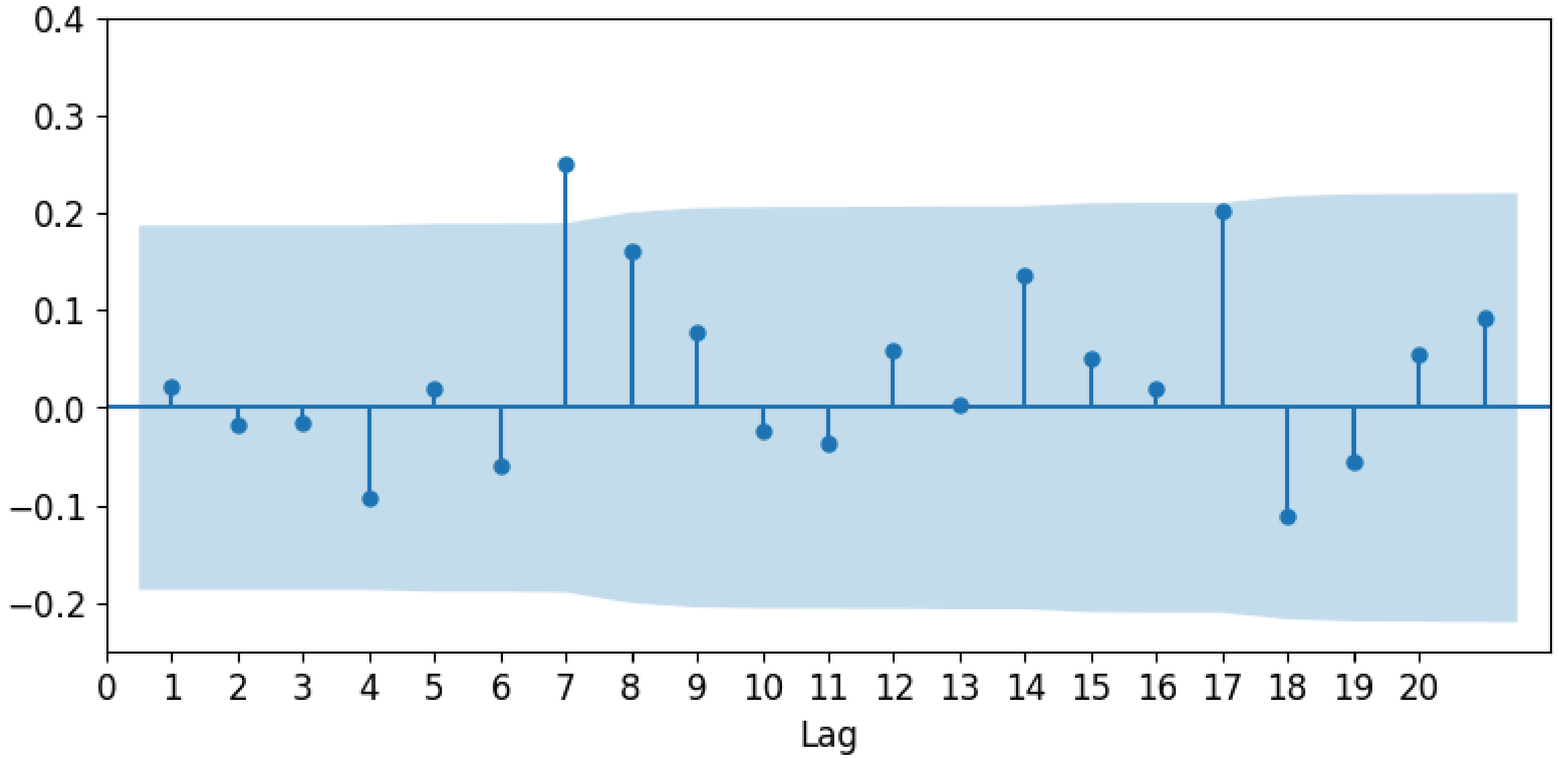


04/2024

ANALYSIS OF DATA DISTRIBUTION

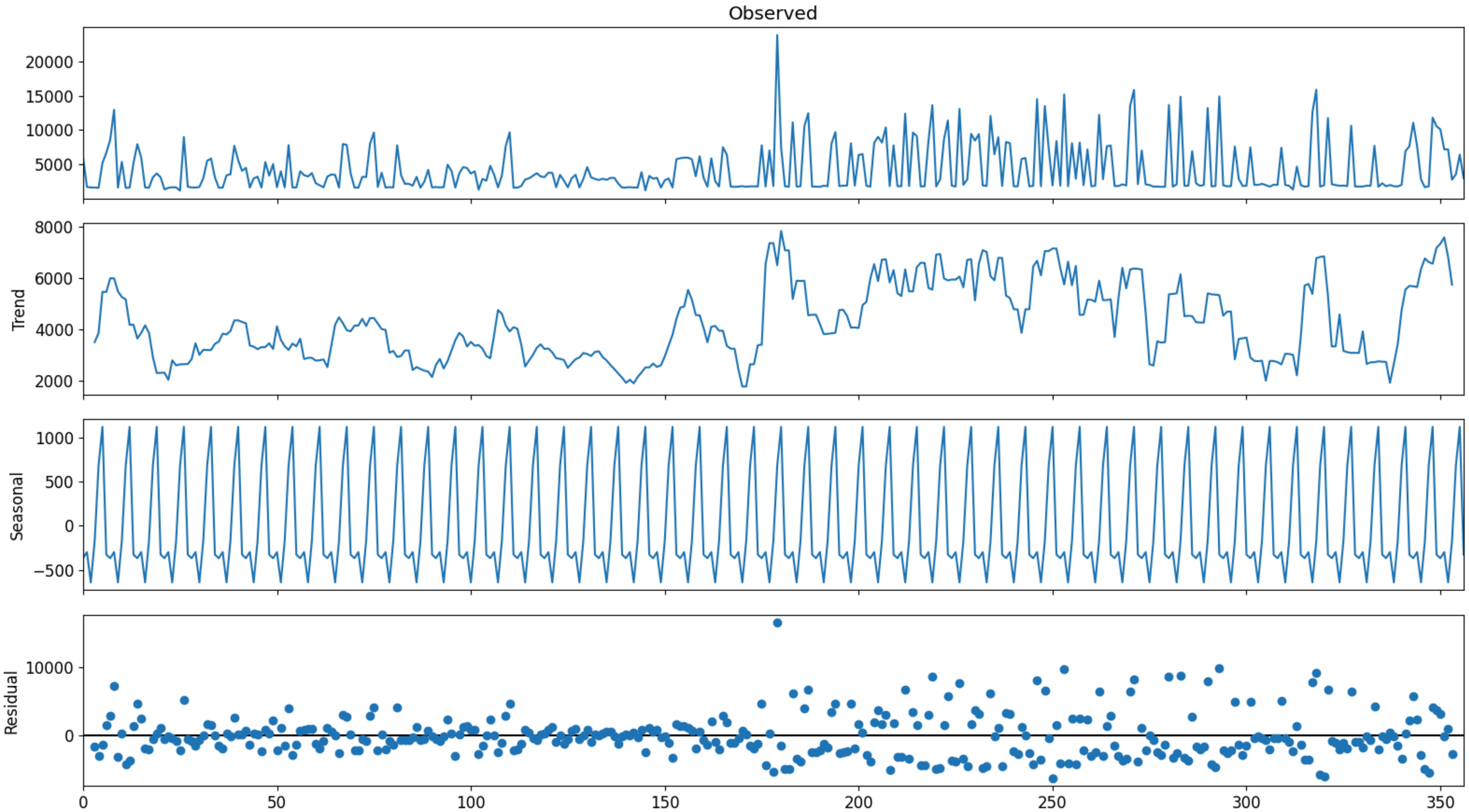
In the network traffic classification domain models rapidly lose accuracy and dataset relevance in time. Thus, we try to find out if concept drifts exist in the network traffic domain and also find out the type of concept drifts. Moreover, we try to find out if concept drifts repeat in time and how often.

AUTOCORRELATION



cesnet

SEASONAL DECOMPOSE



KOLMOGOROV–SMIRNOV TEST

The Kolmogorov-Smirnov (KS) test is used to test the goodness of fit of a given set of data to a theoretical distribution.

- We use KS test to compare the distribution of two set of data:
  - Weekdays data
  - Weekend data
- The KS test requires the data generated from one process

  ==> each class separately
- The KS test can be used only for one-dimensional data

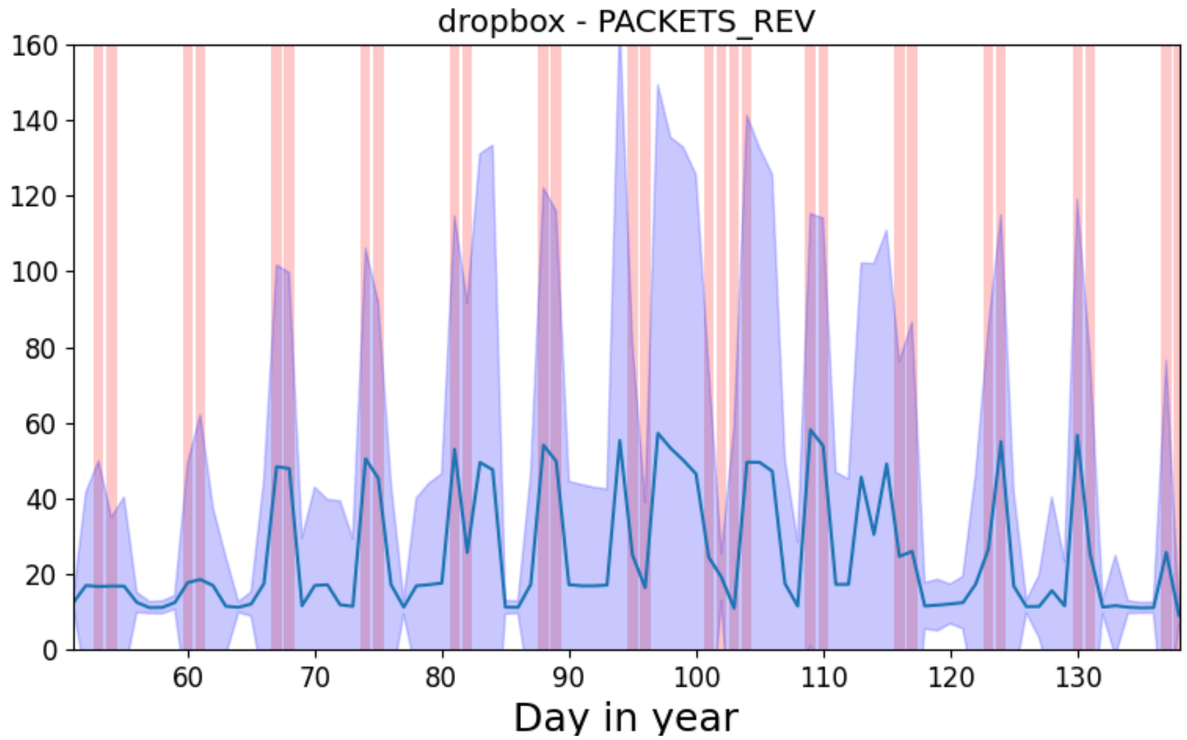
  ==> each feature separately

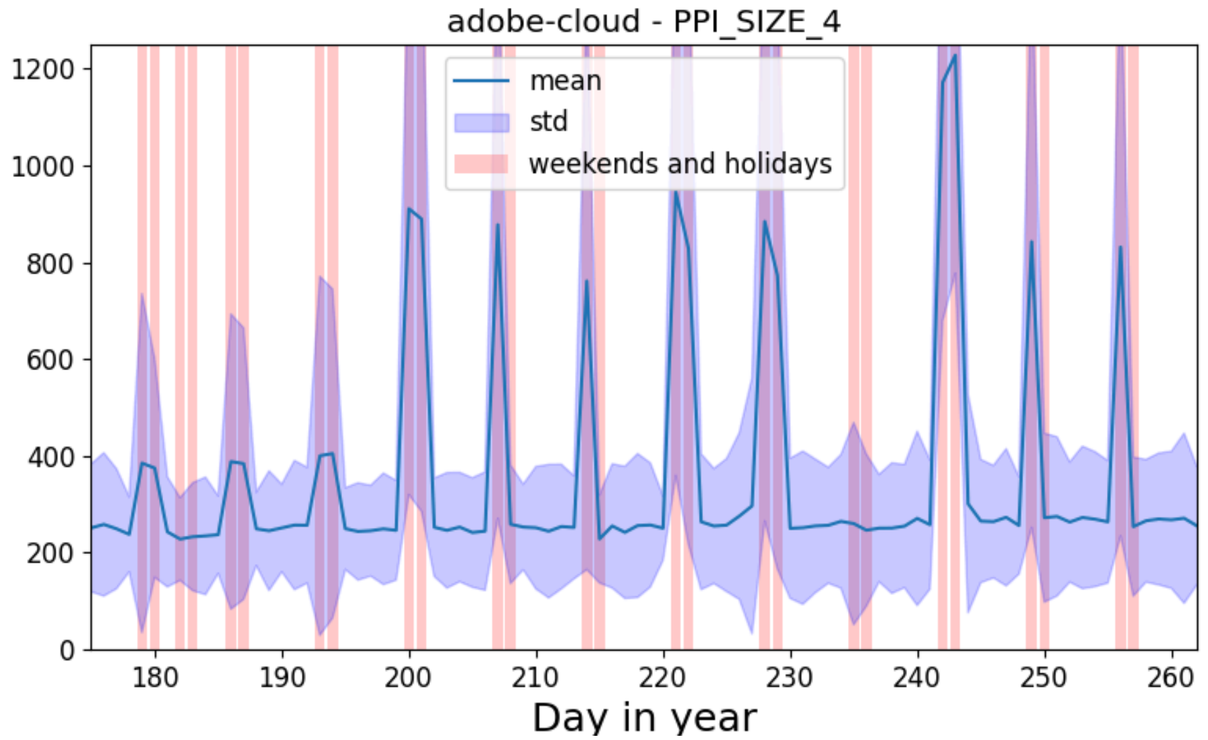
KOLMOGOROV–SMIRNOV TEST

The experiment results shows:

- 38% of all tests (features X classes) result in the rejection of the null hypothesis ==> different distribution in Weekdays and Weekend
- Some features have usually the same distribution
  - TCP flags, PPI Inter Packet Times, Flow End Reason Features, ...
- Some features have usually different distributions
  - Number of Bytes and Packets, Durations, PPI Sizes, Directions,...
- Some classes have the mostly same distribution
  - docker-registry, kaspersky, doh, chrome-remotedesktop, apple-updates, ...
- Some classes have different distributions for most of the features
  - avast, spotify, slack, dropbox, snapchat, google-drive, youtube, ...

WEEKEND PHENOMENON



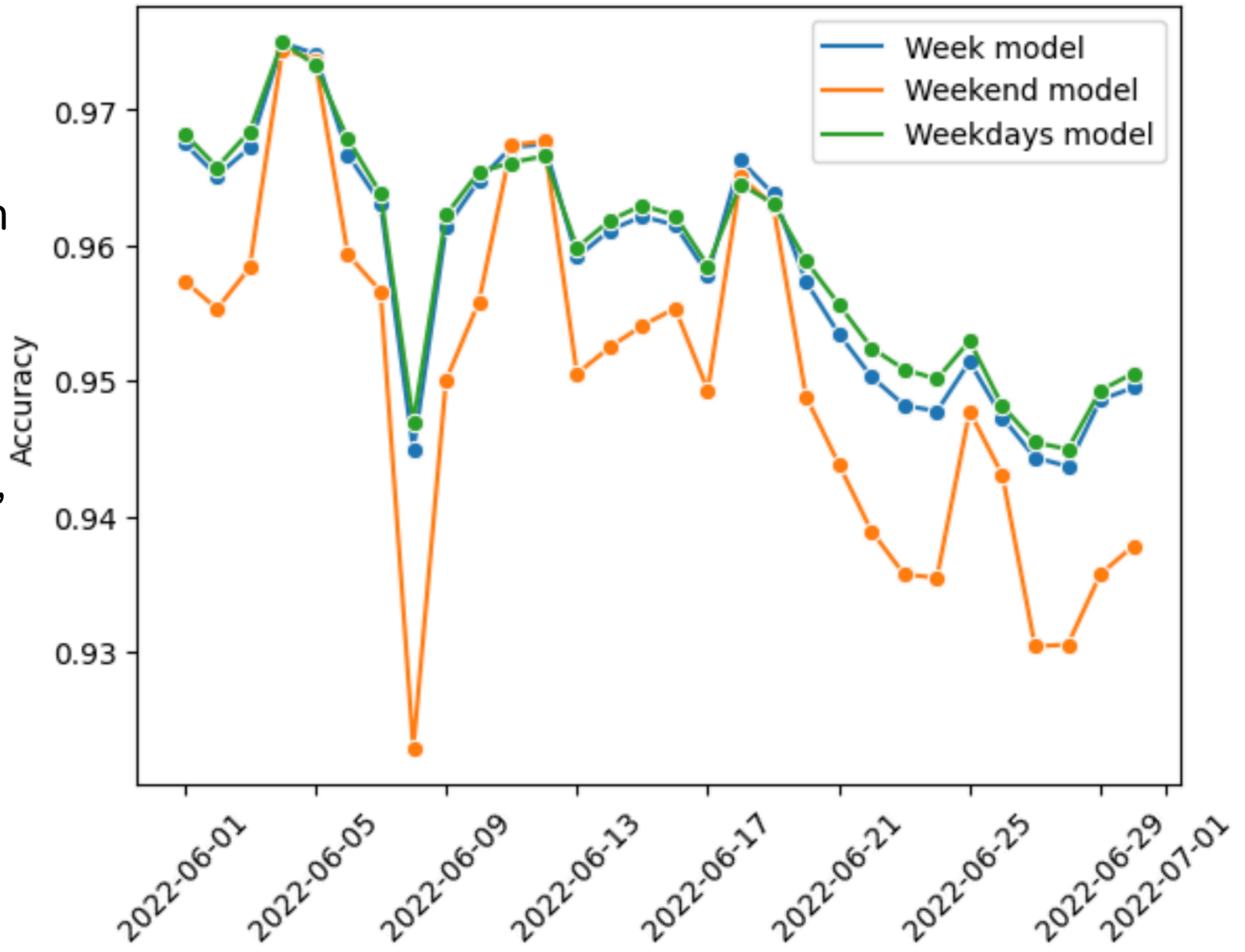


CONCLUSION AND FUTURE WORK

- We provide evidence of the existence of the Weekend phenomenon in the network traffic classification domain
- Future work:
  - Multivariate distribution tests (like MMD)
  - How much will distribution changes affect the model?
  - How much will distribution change of one single class aggeft the model?
  - On the CESNET-QUIC dataset, we also noticed that the traffic for each class differs between day hours and night hours. Exists also some Day/Night phenomenon?

HOW DOES AFFECTS MODEL?

- Trained on May 2022 using CESNET-DataZoo
- Neural Network model from CESNET-Models
- The weekend phenomenon affects all models, however, despite the fact that the weekend model does not work on weekdays, the weekday model performs comparatively well on the weekend



CESNET-DataZoo:

CESNET-MODELS:

CONTACTS

janciluk@fit.cvut.cz
koumajos@fit.cvut.cz
soukudom@fit.cvut.cz
cejkat@cesnet.cz