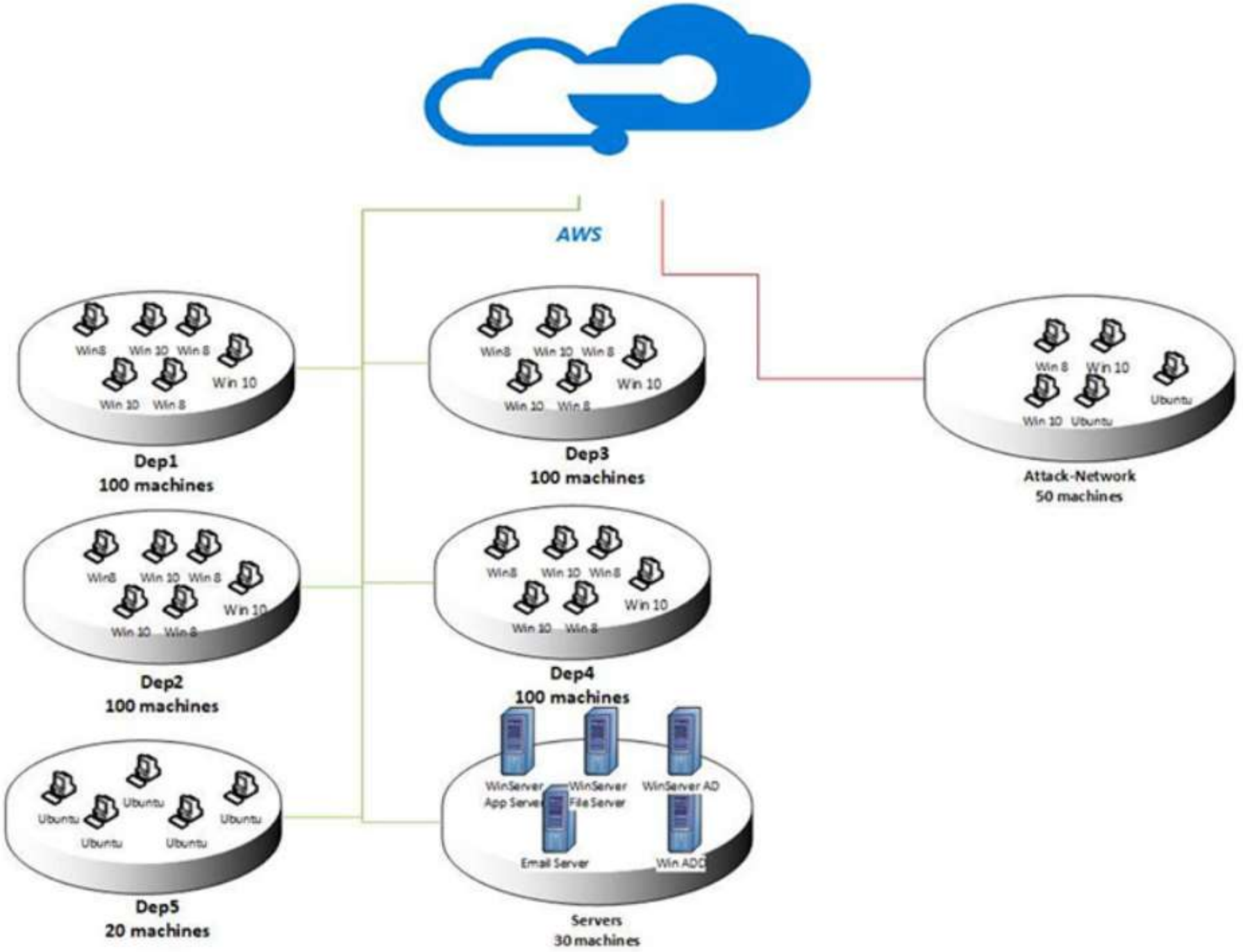
BOTNET DETECTION

DATASET

- A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018)
- The dataset includes seven different attack scenarios, namely Brute-force, Heartbleed, **Botnet**, DoS, DDoS, Web attacks, and infiltration of the network from inside.
- The attacking infrastructure includes 50 machines and the victim organization has 5 departments includes 420 PCs and 30 servers.
- This dataset includes the network traffic and log files of each machine from the victim side, along with 80 network traffic features extracted from captured traffic using CICFlowMeter-V3.
- The data set size is 6.41 GB of 10 csv files each one represents a specific day queries, this project was made on one file containing more than one million record.

Link for dataset: https://registry.opendata.aws/cse-cic-ids2018/


AWS
Win8
Win 10
Win 8
Win 10
Win 10
Win 8
Dep1
100 machines
Win8
Win 10
Win 8
Win 10
Win 10
Win 8
Dep3
100 machines
Win 8
Win 10
Ubuntu
Win 10
Ubuntu
Attack-Network
50 machines
Win8
Win 10
Win 8
Win 10
Win 10
Win 8
Dep2
100 machines
Win8
Win 10
Win 8
Win 10
Win 10
Win 8
Dep4
100 machines
Ubuntu
Ubuntu
Ubuntu
Ubuntu
Ubuntu
Dep5
20 machines
WinServer
App Server
WinServer
File Server
WinServer AD
Email Server
Win ADC
Servers
30 machines

DATA PREPROCESSING

A 5-step Process

01 Checking for wrong null or duplicated values

02 Assigning the correct type for each feature

03 Removing redundunt features and perform PCA

04 Checking data balance

05 Splitiing Data into Training and Testing sets

Training Models

Choosing the best one based on results

01	02	03	04	05
Decision Tree	K-Nearest Neighbour	Support Vector Machines	Naïve Bayesian	Neural Networks

RESULTS

Model	DT	KNN	SVM	NB	ANN
Accuracy	99.99%	99.92%	75.52%	52.10%	99.31%
Precision	99.99%	99.76%	90%	36.11%	98.80%
Recall	99.99%	99.93%	5.6%	98.38%	98.31%

RELIABILITY OF RESULTS

Referring to **Spiringer** journal of big data paper: A survey and analysis of intrusion detection models based on **CSE-CIC-IDS2018** Big Data.

- " With regard to CICIDS2018, the RF and **DT** learners scored an accuracy99.99%. Tied to this accuracy, the precision was 100% and the recall was 99.99% for both learners. The of RF and DT learners also had the **highest accuracy** for ISOT HTTP (99.94% for RF and 99.90% for DT)."
- Discussion of surveyed works:

In general, the best performance scores are unusually high for studies where scores are provided. This finding is notable. Accuracy scores are between 96 (D'hooge et al., 2020) and 100 (Atefinia & Ahmadi, 2020; Kanimozhi & Jacob, 2019a). Several papers show recall scores of 100 (Atefinia & Ahmadi, 2020; Kanimozhi & Jacob, 2019a; Kanimozhi & Jacob, 2019b; Li et al., 2020; Filho et al., 2019) and also precision scores of 100 (Atefinia & Ahmadi, 2020; Kanimozhi & Jacob, 2019a; Huancayo Ramos et al., 2020; Filho et al., 2019). In addition, three studies show a perfect AUC score (Kanimozhi & Jacob, 2019a; Kanimozhi & Jacob, 2019b ;Li et al., 2020). These noticeably high scores for the various metrics **may be due to overfitting**.