Deep Learning for Network Traffic Classification

Weston Jackson[1]; Niloofar Bayat[2]; Derrick Liu[3]
Columbia University in the City of New York

COLUMBIA ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Introduction

- Over HTTPS services, client and server first communicate through a TLS handshake.
- **Server Name Identification** (SNI) is an extension to the TLS handshake where the destination hostname can be extracted from the Client-Hello message.

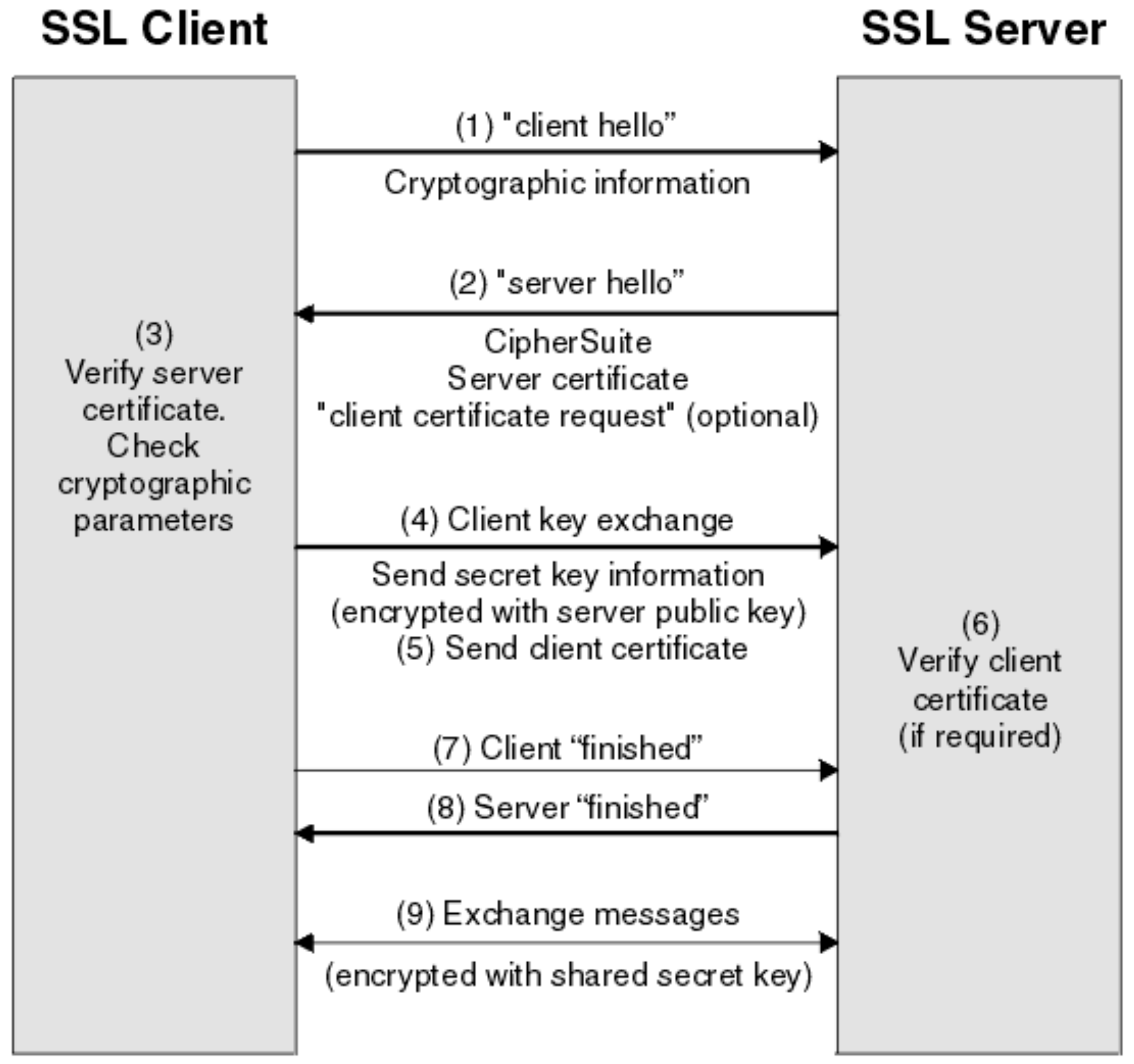

SSL Client
SSL Server
(1) "client hello"
Cryptographic information
(2) "server hello"
CipherSuite
Server certificate
"client certificate request" (optional)
(3)
Verify server certificate. Check cryptographic parameters
(4) Client key exchange
Send secret key information (encrypted with server public key)
(5) Send client certificate
(6)
Verify client certificate (if required)
(7) Client "finished"
(8) Server "finished"
(9) Exchange messages
(encrypted with shared secret key)

Figure 1. TLS handshake

Problem Formulation

- Firewalls inspect SNI to check if an SNI is allowed.
  - ➢ SNI can be faked to bypass such firewalls
- Since SNI is not encrypted, it does not preserve users' privacy and an adversary can detect it.
  - ➢ ESNI has been proposed to address this issue

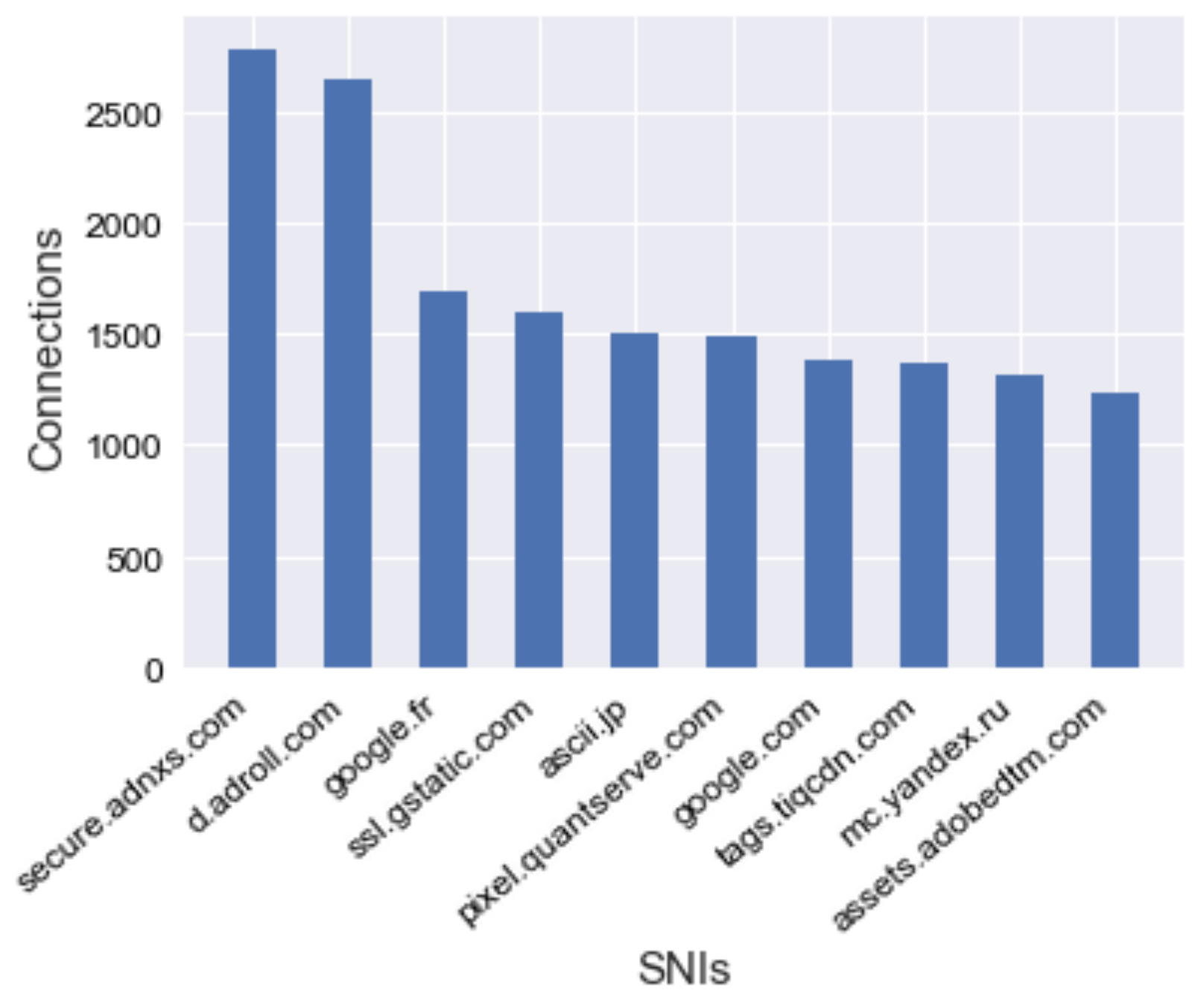
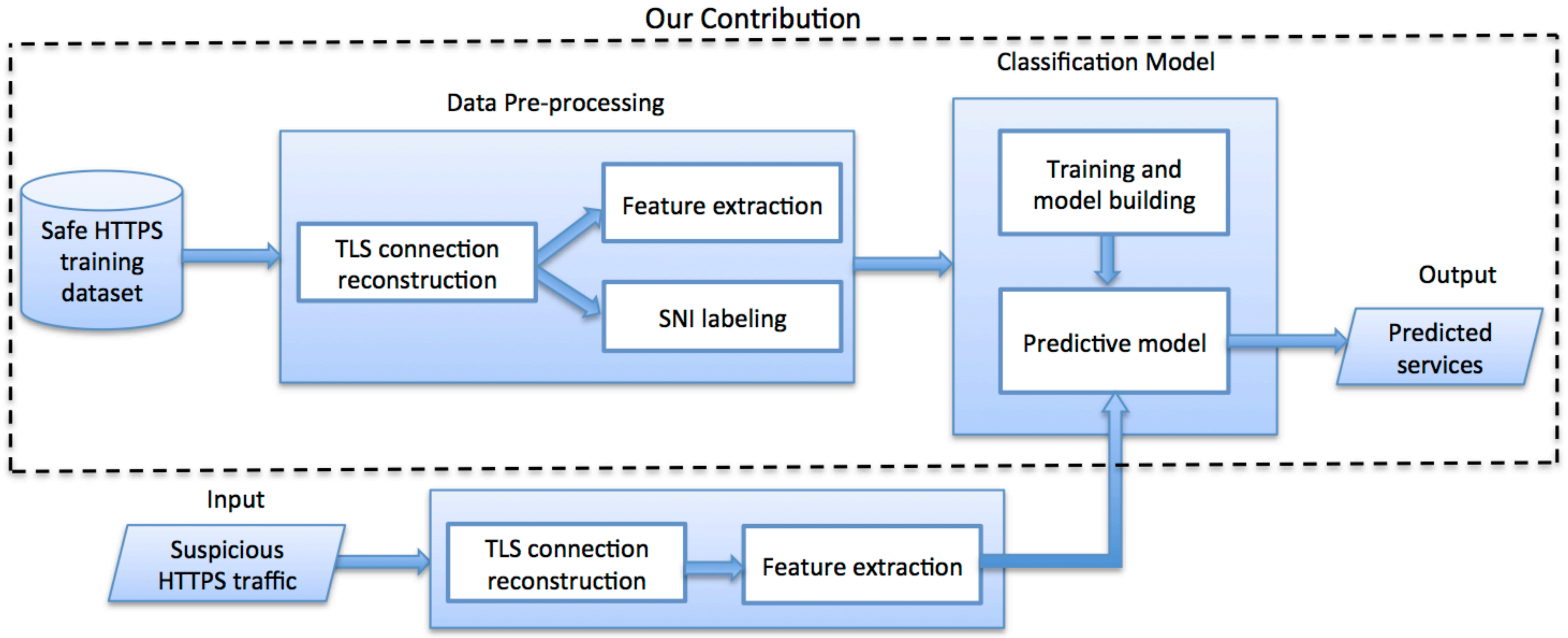
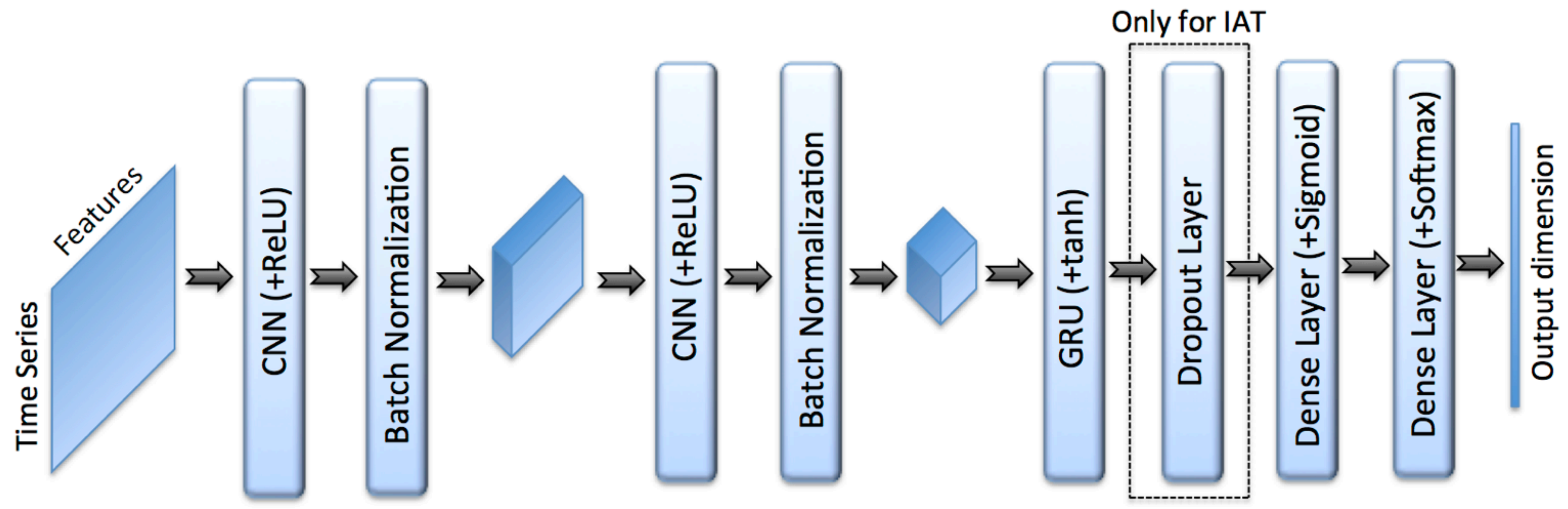


Figure 2. SNIs with most connections in our dataset

The Pipeline



CNN-RNN Architecture



Data Collection and Feature Selection

- Data Collection and preprocessing:
  - ➢ Used publicly available Internet traffic data (pcap format)
  - ➢ Applied SSL filter to obtain HTTPS traffic
  - ➢ Unified two directions of communication over TCP channel
  - ➢ Removed unnecessary characters and unknown SNIs
- Statistical Features
  - ➢ Remote -> Local; Local -> Remote; Combined
    - Packet size: {size, 25th, 50th, 75th, max, avg, var}
    - Payload size: {25th, 50th, 75th, max, avg, var}
    - Inter-arrival time: {25th, 50th, 75th}
- Sequential Features
  - ➢ Combined
    - Packet size; payload size; inter-arrival time (log)
  - ➢ First 25 packets per TCP connection; ordered by arrival time
  - ➢ Shorter sequences padded with zero

Architectures

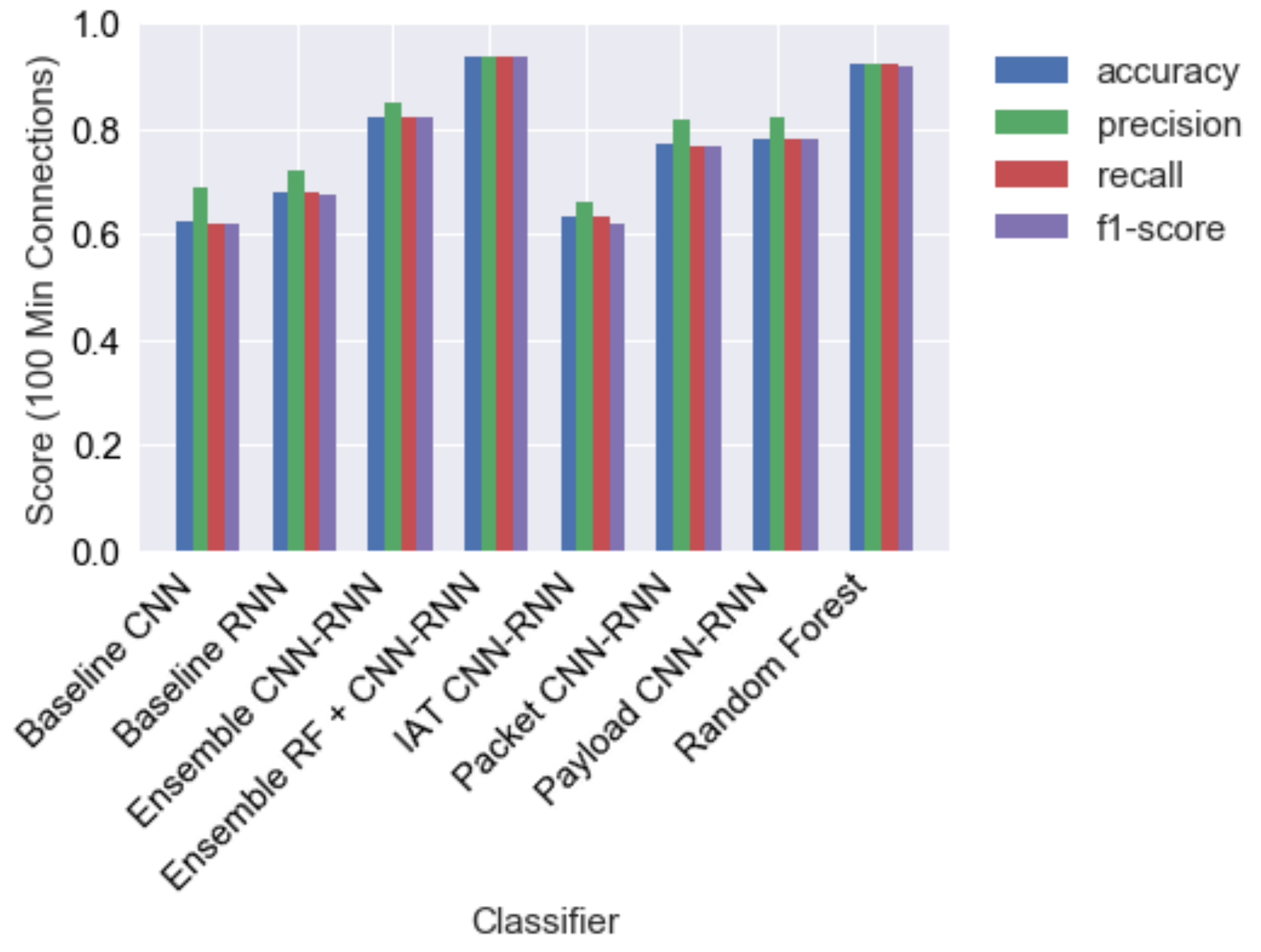


Figure 3. Accuracy, precision, recall and f1-score for different classifiers. Ensemble of random forest with CNN-RNN leads to best results.

Results

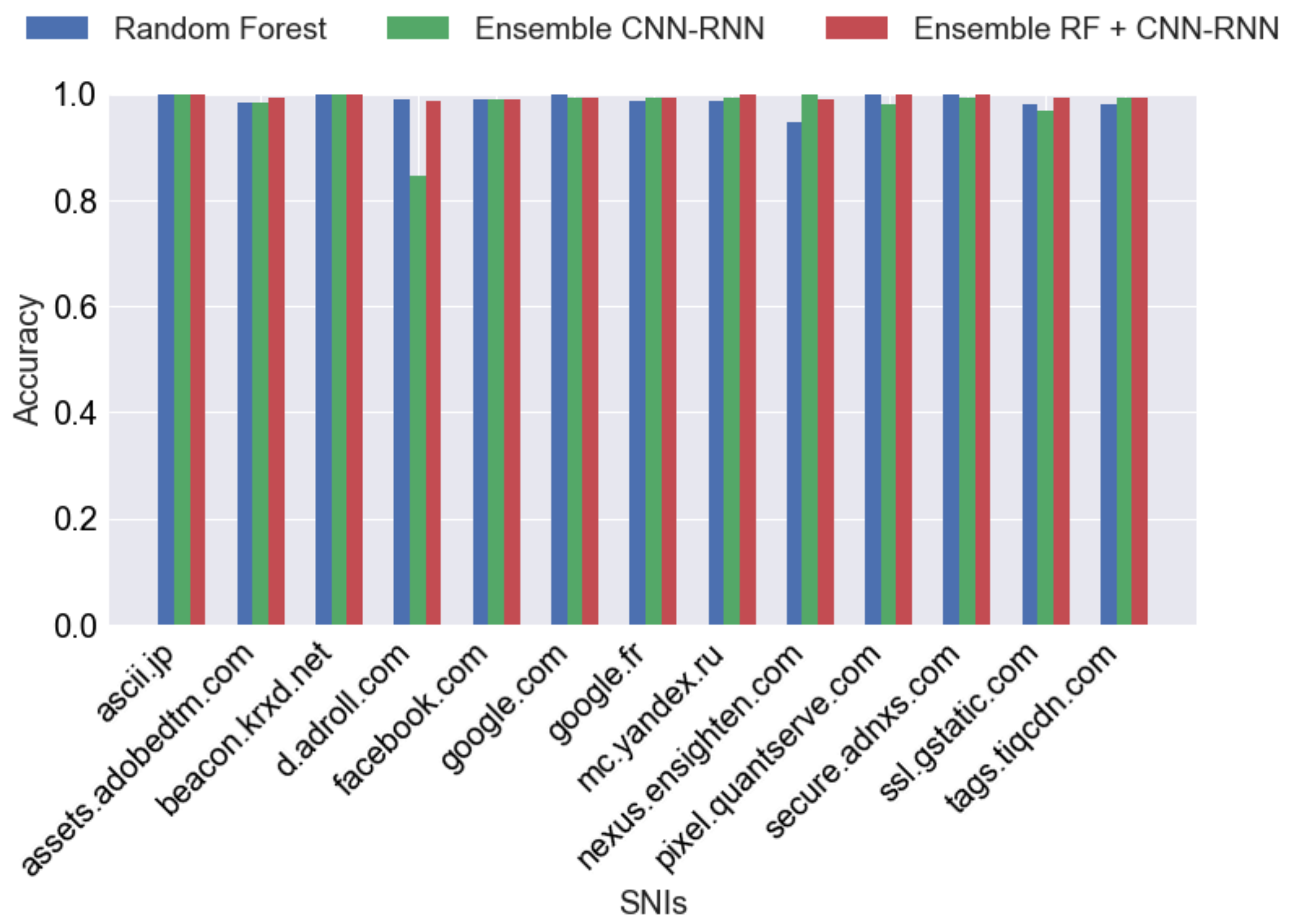


Figure 4. Prediction accuracy of the most used SNIs for different classifiers. Ensemble of CNN+RNN+RF outperforms the rest.

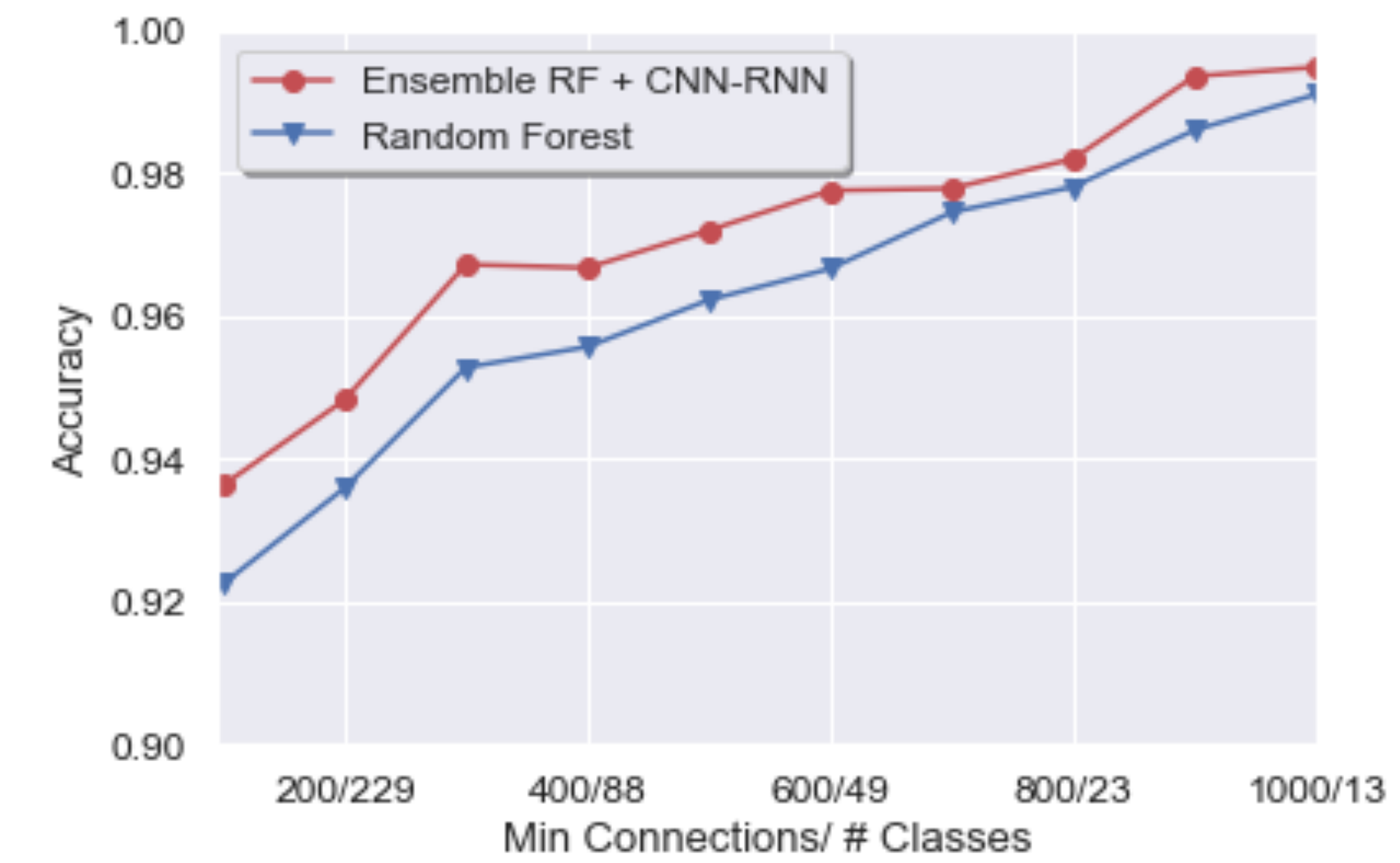


Figure 5. Comparison between RF accuracy (statistical features), and ensemble of RF (statistical features) and CNN-RNN (sequential features). CNN-RNN helps the state of art to perform better. Note that y-axis is restricted to [0.9, 1].

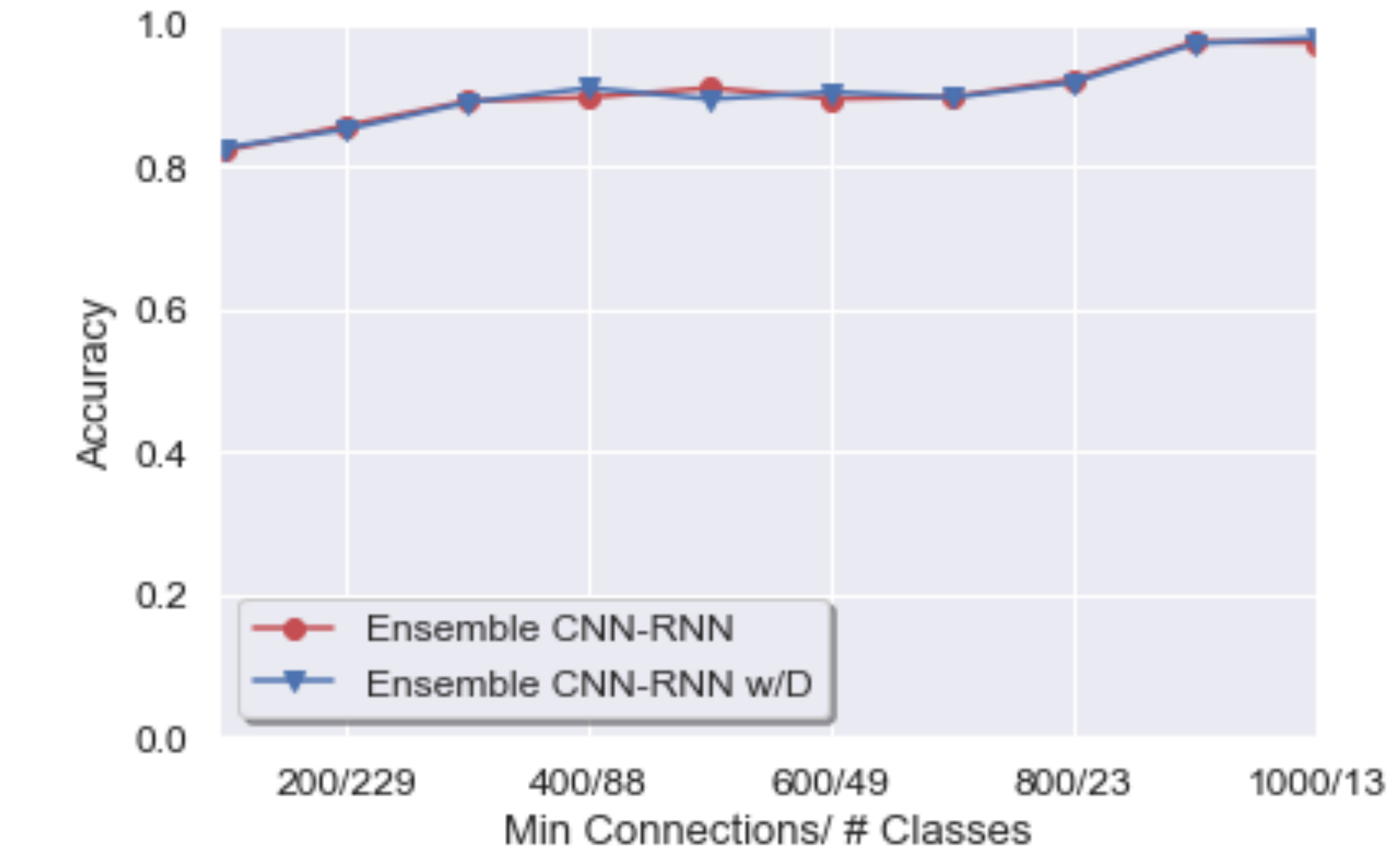


Figure 6. Accuracy of Ensemble CNN-RNN with/without directionality. No consistent improvement is observed using directional features.

Discussion and Conclusion

- Directionality can be studied more in the future.
- This system can be used to detect SNI for suspicious traffic.

Contact Information

[1]Weston.j.jackon@gmail.com
[2]Niloofar.bayat@columbia.edu
[3]dl3122@columbia.edu

Acknowledgements

Special thanks to professor Iddo Drori for his helpful comments.