Although we found the random forest ensemble classifier to be the most robust for this problem domain, we demonstrated that the choice of features had a much greater effect on performance. The enhanced feature set was created by augmenting a standard feature set used in this domain with features identified by a domain expert and specifically tailored for encrypted network sessions. By not relying solely on features that were convenient to gather and engaging with domain experts to iterate on how the data would be best represented, all machine learning algorithms had significant improvements in performance. Combining diverse views of the data, such as features pertaining to how the application is transmitting data with features that are representative of the application, was the key innovation. As an example of the magnitude of the improvement, linear regression using the enhanced feature set easily outperformed the random forest ensemble using a standard network connection representation on all criteria considered.

11 ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their helpful comments. We would also like to thank the Cisco Technology Fund for funding this work, and Greg Akers for his support. This work was in part made possible by the help of Cisco CSIRT and ThreatGrid with respect to obtaining data, and Philip Perricone and Bill Hudson by their contributions to Joy.

REFERENCES

[1] Blake Anderson and David McGrew. 2016. Identifying Encrypted Malware Traffic with Contextual Flow Data. In *ACM Workshop on Artificial Intelligence and Security (AISec)*. 35–46.
[2] Blake Anderson, Subharthi Paul, and David McGrew. 2016. Deciphering Malware's Use of TLS (without Decryption). In *ArXiv e-prints*.
[3] Mike Belshe, Roberto Peon, and Martin Thomson. 2015. Hypertext Transfer Protocol Version 2 (HTTP/2). RFC 7540 (Proposed Standard). (2015). http://www.ietf.org/rfc/rfc7540.txt
[4] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2011. Support Vector Machines Under Adversarial Label Noise. In *Asian Conference on Machine Learning*. 97–112.
[5] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning Attacks against Support Vector Machines. In *International Conference on Machine Learning (ICML)*. 1807–1814.
[6] Leyla Bilge, Davide Balzarotti, William Robertson, Engin Kirda, and Christopher Kruegel. 2012. Disclosure: Detecting Botnet Command and Control Servers through Large-Scale NetFlow Analysis. In *ACM Annual Computer Security Applications Conference (ACSAC)*. 129–138.
[7] Christopher Bishop. 2006. Pattern Recognition. *Machine Learning* 128 (2006), 1–58.
[8] Simon Blake-Wilson, Nelson Bolyard, Vipul Gupta, Chris Hawk, and Bodo Moeller. 2006. Elliptic Curve Cryptography (ECC) Cipher Suites for Transport Layer Security (TLS). RFC 4492 (Informational). (2006). http://www.ietf.org/rfc/rfc4492.txt
[9] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.
[10] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and Regression Trees*. CRC press.
[11] J Michael Butler. 2013. Finding Hidden Threats by Decrypting SSL. *SANS Institute* (2013).
[12] François Chollet. 2017. Keras. (2017). https://github.com/fchollet/keras Accessed: 2017-04-19.
[13] Cisco Talos. 2017. IP Blacklist Feed. (2017). http://www.talosintel.com/feeds/ip-filter.blf Accessed: 2017-04-19.
[14] Benoit Claise. 2004. Cisco Systems NetFlow Services Export Version 9. RFC 3954 (Informational). (2004). http://www.ietf.org/rfc/rfc3954.txt
[15] Benoit Claise, Brian Trammell, and Paul Aitken. 2013. Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information. RFC 7011 (Proposed Standard). (2013). http://www.ietf.org/rfc/rfc7011.txt
[16] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 3 (1995), 273–297.
[17] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, and others. 2004. Adversarial Classification. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 99–108.
[18] Tim Dierks and Eric Rescorla. 2008. The Transport Layer Security (TLS) Protocol Version 1.2. RFC 5246 (Proposed Standard). (2008). http://www.ietf.org/rfc/rfc5246.txt
[19] Pedro Domingos. 2012. A Few Useful Things to Know about Machine Learning. *Communications of the ACM* 55, 10 (2012), 78–87.
[20] Floriana Esposito, Donato Malerba, Giovanni Semeraro, and J Kay. 1997. A Comparative Analysis of Methods for Pruning Decision Trees. *Transactions on Pattern Analysis and Machine Intelligence* 19, 5 (1997), 476–491.
[21] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9, Aug (2008), 1871–1874.
[22] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Springer.
[23] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33, 1 (2010).
[24] Guofei Gu, Roberto Perdisci, Junjie Zhang, and Wenke Lee. 2008. BotMiner: Clustering Analysis of Network Traffic for Protocol-and Structure-Independent Botnet Detection. In *USENIX Security Symposium*. 139–154.
[25] Matt Harrigan. 2016. Machine Learning is not the Answer to Better Network Security. (2016). https://techcrunch.com/2016/02/29/machine-learning-is-not-the-answer-to-better-network-security/ Accessed: 2017-04-19.
[26] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. In *ArXiv e-prints*.
[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* 521, 7553 (2015), 436–444.
[28] Daniel Lowd and Christopher Meek. 2005. Adversarial Learning. In *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD)*. 641–647.
[29] David McGrew and Blake Anderson. 2016. Enhanced Telemetry for Encrypted Threat Analytics. In *IEEE ICNP Workshop on Machine Learning in Computer Networks (NetworkML)*. 1–6.
[30] David McGrew, Blake Anderson, Bill Hudson, and Philip Perricone. 2017. Joy. https://github.com/cisco/joy. (2017).
[31] Andrew W Moore and Denis Zuev. 2005. Internet Traffic Classification Using Bayesian Analysis Techniques. *SIGMETRICS Performance Evaluation Review* 33 (2005), 50–60.
[32] Vern Paxson. 1999. Bro: a System for Detecting Network Intruders in Real-Time. *Computer Networks* 31, 23-24 (1999), 2435–2463.
[33] Fabian Pedregosa, Ga el Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[34] Eric Rescorla. 2017. The Transport Layer Security (TLS) Protocol Version 1.3 (draft 20). Intended Status: Standards Track. (2017). https://tools.ietf.org/html/draft-ietf-tls-tls13-20
[35] Martin Roesch. 1999. Snort - Lightweight Intrusion Detection for Networks. In *USENIX Large Installation System Administration Conference (LISA)*. 229–238.
[36] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 1528–1540.
[37] Robin Sommer and Vern Paxson. 2010. Outside the Closed World: On using Machine Learning for Network Intrusion Detection. In *IEEE Symposium on Security and Privacy (S&P)*. 305–316.
[38] Florian Tegeler, Xiaoming Fu, Giovanni Vigna, and Christopher Kruegel. 2012. Botfinder: Finding Bots in Network Traffic without Deep Packet Inspection. In *ACM International Conference on Emerging Networking Experiments and Technologies (Co-NEXT)*. 349–360.
[39] Nigel Williams, Sebastian Zander, and Grenville Armitage. 2006. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. *SIGCOMM Computer Communication Review* 36, 5 (2006), 5–16.
[40] David H Wolpert and William G Macready. 1997. No Free Lunch Theorems for Optimization. *Transactions on Evolutionary Computation* 1, 1 (1997), 67–82.