WISP 2021 Proceedings

Pre-ICIS Workshop on Information Security and Privacy (SIGSEC)

12-12-2021

Expressing uncertainty in security analytics research: a demonstration of Bayesian analysis applied to binary classification problems

Douglas P. Twitchell
Boise State University

Christie M. Fuller
Boise State University, christiefuller@boisestate.edu

Expressing Uncertainty in Security Analytics Research: A Demonstration of Bayesian Analysis Applied to Binary Classification Problems

Douglas P. Twitchell
College of Business and Economics, Boise State University,
Boise, Idaho, USA

Christie M. Fuller[1]
College of Business and Economics, Boise State University,
Boise, Idaho, USA

ABSTRACT

A common application of security analytics is binary classification problems, which are typically assessed using measures derived from signal detection theory, such as accuracy, sensitivity, and specificity. However, these measures fail to incorporate the uncertainty inherent to many contexts into the results. We propose that the types of binary classification problems studied by security researchers can be described based on the level of uncertainty present in the data. We demonstrate the use of Bayes data analysis in security contexts with varying levels of uncertainty and conclude that Bayesian analysis is particularly relevant in applications characterized by high uncertainty. We discuss how to apply similar analyses to other information security research.

Keywords: Binary classification, security, Bayesian analysis, uncertainty, prevalence, positive predictive value

INTRODUCTION

The information security field contains many binary classification problems such as insurance fraud detection (Artís et al. 2002), finding security bugs (Jiang et al. 2020), intrusion detection (Li et al. 2020), and management fraud detection (Cecchini et al. 2010). Shaukat and colleagues (Shaukat et al. 2020) recently compared eighteen review papers related to the use of machine learning techniques in cybersecurity. The base rate in the samples, the size of the

[1] Corresponding author. christiefuller@boisestate.edu

sample, the prevalence of cases in the population, and the sensitivity and specificity of the test all contribute to varying levels of uncertainty in the contexts above. In this paper, we use Bayesian data analysis methods to show how this uncertainty impacts results and recommend when incorporating Bayesian data analysis is most useful to decision makers.

Evaluating Classification Problems

In most security research using binary classification, classification models are evaluated using signal detection measures such as accuracy, sensitivity, specificity, precision, F-score, and area under the ROC curve (AUC) as defined in Table 1 (Japkowicz and Shah 2011; Shaukat et al. 2020) derived from a confusion matrix (see Table 2). These measures are useful both to compare classification models and aid decision-makers (Green and Swets 1966). AUC, for example, provides a measurement that incorporates the accuracy of a model at every possible

Table 1: Definitions of common classification measures.

Measure	**Formula**	**Description**
Sensitivity, Recall, Hit Rate, True Positive Rate (TPR):	$\frac{TP}{TP + FN}$	Of all class A's, the fraction labeled class A
Specificity, True Negative Rate (TNR):	$\frac{TN}{TN + FP}$	Of all class B's, the fraction labeled class B
Precision, Positive Predictive Value (PPV):	$\frac{TP}{TP + FP}$	The posterior probability of a case labeled class A being class A
Negative Predictive Value (NPV):	$\frac{TN}{TN + FN}$	The posterior probability of a case labeled class B being class B
Accuracy:	$\frac{TP + TN}{TP + FP + TN + FN}$	Of all cases, the fraction correctly labeled
F1:	$\frac{2TP}{(2TP + FP + FN)}$	Average of sensitivity and precision
AUC:	$\int_{\infty}^{-\infty} (TPR(T))(FPR'(T))dT$	Area under the ROC curve

threshold or output probability of that model. This, however, is best used for comparing models rather than determining the model's usefulness. For exploratory studies on new topics or techniques, the measures above provide an adequate baseline for future research. For evaluating how detection systems might work in the field, however, these measures may fall short.

Table 2: Confusion Matrix

	Classified as: A*	**Classified as: B**
True Class: A	True Positives (TP)	False Negative (FN)
True Class: B	False Positive (FP)	True Negative (TN)

Note: Here we assume class A is the class of most interest to the researcher.

Uncertainty refers to a lack of complete information that prevents perfect decision making (Twitchell and Fuller 2018). The lack of information, or uncertainty, may lead to overconfidence in results both by expressing unsupported precision and failing to recognize the full range of possibilities. For example, neither sensitivity nor specificity (as defined in Table 2) incorporate any information about the prevalence, or proportion of the population represented by the class of most interest. (Note that we refer to "prevalence" to mean the percentage of a population with the phenomenon being detected. We will use "base rate" to refer to the number of positive cases in the sample used in the study).

Researchers frequently build classification models on balanced samples to overcome the inherent problems of training a model on a low base rate sample. While this is a well-justified strategy, results should also be evaluated with respect to the prevalence. When results are evaluated in relation to the prevalence, sensitivity may be unacceptably low or the false positive rate too high for the model to have practical value (Twitchell and Fuller 2018; National Research Council, 2003). ROC curves, AUC and F-score are frequently reported, but are less useful for analyzing performance in specific scenarios or contexts than they are for comparing algorithms. Precision, or the proportion of positive identifications that are correct, does incorporate the sample's base rate, but does not help the user understand the uncertainty about the base rate and is often not representative of the true prevalence (Twitchell and Fuller 2018). These signal detection measures are point estimates that tell us how a classifier performs on the sample data. They do not tell us how well a classifier might perform on future data, nor do they incorporate any uncertainty in the data. To achieve its full potential, analytics research should more fully evaluate the results in the intended context.

Bayesian Data Analysis

Bayesian data analysis methods are well-suited for understanding the type of uncertainty we describe above. Unlike frequentist methods, they do not assume a sampling distribution. Instead, they require building a full probability model on all parameters and allow sampling from the posterior distribution of this model. The posterior distribution is not assumed to have any shape and includes uncertainty about all parameters in the model including any prior uncertainty we have about any of the parameters.

Bayesian data analysis goes beyond a simple application of Bayes' theorem. A well-known application of Bayes' theorem is calculating the probability of having a disease given a positive diagnostic test. In this application, a positive result from a diagnostic test with 99% accuracy when testing for a disease with a 1% prevalence in the general population yields a probability of having the disease of approximately 50% (Horgan 2016). Missing from this analysis is sources of uncertainty related to the accuracy of the test or related to the prevalence of the disease. The specificity and sensitivity of the test and the prevalence of the disease are all subject to measurement error and bias from the contexts in which those measurements were taken. Finally, the more uncertain the test, the more influence the prevalence has on the test precision. A very uncertain test combined with a very low prevalence results in a precision not much different than the prevalence.

Modern Bayesian data analysis—as described in extensive detail in (Gelman et al. 2013)—encourages researchers to include these errors as uncertainty in the analysis by working with probability distributions rather than point estimates. At its essence, Bayesian data analysis requires describing a *prior* distribution of all applicable parameters before incorporating the data and a *likelihood* distribution of the data. Multiplying them together and dividing by the data distribution (i.e., applying Bayes' theorem to the probability distributions) results in a joint posterior distribution of all the parameters. In many applications, this procedure can't be done by exact computation (Geyer 2011), so current methods employ Markov Chain Monte Carlo (MCMC), an approximate, probability-based, numerical method to sample directly from the posterior distribution. Descriptive statistics can then be derived from the sample to summarize the posterior distribution and make inferences. In the next section, we illustrate the use of Bayesian data analysis with examples from information security.

METHODOLOGY

While Bayesian data analysis is becoming more commonly used in the medical field (Gelman and Carpenter 2020), it has only recently appeared in IS research and behavioral business research (Cecchini et al. 2010; Dutta et al. 2018; Twitchell and Fuller 2018). To demonstrate the utility of incorporating Bayesian data analysis into security analytics research, we identified several studies, shown in Table 3, that use binary classification models to find phenomena of interest such as cases of fraud, bugs, and deception. These studies, unlike many others, provided a full confusion matrix and prevalence estimates. Many studies don't provide prevalence, likely because it is difficult to obtain or the study authors decided it wasn't important. Prevalence (Prev. in Table 3) is the estimated prevalence of the phenomenon being detected in a real-world setting, not the prevalence of the sample used to train or test the model. Prevalence n (Prev. n in Table 3) is the size of the real-world population used to determine prevalence. Sample size is the size of the testing set used to test the detection method in the study.

Table 3. Studies included along with reported results

Study	TP	FP	TN	FN	Sample Size	Prev.	Prev. n
Insurance Fraud (Artís et al. 2002)	768	290	708	229	1995	0.333	3
Security bugs (Jiang et al. 2020)	107	8	16,917	3,938	20970	0.038	138,982
Criminal Statements (Fuller et al. 2009)	63	93	194	16	366	0.216	366
Intrusion Detection (Li et al. 2020)	137,884	53	85,070	4,299	227306	0.001	2,830,540
Credit Card Fraud (Arya and G 2020)	99	40	56,821	2	56962	0.002	284,807
Data Loss Prevention (Faiz et al. 2020)	135	28	2,159	114	2436	0.105	8,117
Management Fraud (Cecchini et al. 2010)	20	92	890	5	1007	0.032	6,427
Crowdfunding Fraud (Siering et al. 2016)	288	38	232	94	652	0.007	44,054

The studies we found in the security literature did not attempt to estimate prevalence from a field test. Instead, if they reported any prevalence, they reported it as a feature of their sample or the population from which the sample was drawn. For example, in the criminal statements data, the sample the study used had a base rate of 79 deceptive statements out of 366 total statements. Since the base rate is given, we include the prevalence estimate as a prior instead of estimating it as part of the likelihood. Another example is the intrusion detection study (Li et al. 2020) which only provided the prevalence of its training set. The prevalence did not reflect the real-world ratio of true positives to overall number of cases. So, we used the closer-to-real-world prevalence found in the CICIDS2017 data set (Panigrahi and Borah 2018). We used Model 1 in the appendix to analyze the studies. We translated this model into the Stan probabilistic programming language (Carpenter et al. 2017) for fitting via MCMC. The code for the model is linked in the appendix.

RESULTS

The sensitivity, specificity and prevalence of these studies are summarized in Table 4, along with the credible intervals (CrI) for each. Because the results of the model are posterior distributions that may not be normally distributed, we follow (Edwards et al. 1963) and use *credible interval* to indicate the interval which contains 95% of the probability density starting from the 2.5% quantile and ending at the 97.5% quantile. We abbreviate this CrI to distinguish it from confidence intervals used in frequentist statistics. While insurance fraud and management fraud have similar sensitivity levels, management fraud has a much wider credible interval and thus less certain results than those for insurance fraud. We further investigate the impact of prevalence through its influence on positive predictive value (PPV), as shown in Table 5.

Table 4. **Summary statistics for the model posterior distribution**

Study	Model Sens. μ (95% CrI)	Model Spec. μ (95% CrI)	Model Prev. μ (95% CrI)
Insurance Fraud	0.770 (0.746, 0.793)	0.709 (0.684, 0.733)	0.394 (0.070, 0.801)
Security Bugs	0.026 (0.023, 0.030)	0.999 (0.999, 1.000)	0.038 (0.037, 0.039)
Criminal Statements	0.791 (0.704, 0.870)	0.675 (0.631, 0.717)	0.217 (0.176, 0.261)
Intrusion Detection	0.970 (0.969, 0.971)	0.999 (0.999, 1.000)	0.001 (0.001, 0.001)
Credit Card Fraud	0.971 (0.931, 0.994)	0.999 (0.999, 0.999)	0.002 (0.002, 0.002)
Data Loss Prevention	0.542 (0.489, 0.593)	0.987 (0.982, 0.991)	0.105 (0.098, 0.112)
Management Fraud	0.779 (0.611, 0.910)	0.906 (0.892, 0.919)	0.032 (0.028, 0.037)
Crowdfunding Fraud	0.753 (0.714, 0.789)	0.857 (0.816, 0.894)	0.007 (0.007, 0.008)

Table 5 shows the PPV of each study. We focus on PPV as the model output of interest since detection is focused on finding "positive" cases and PPV provides the probability that a case labeled as positive is actually positive. Additionally, we highlight PPV uncertainty to show how accuracy, prevalence, and sample size affect result in differing levels of uncertainty. Though we view uncertainty as a spectrum, to facilitate description, we describe three broad categories of uncertainty: high, medium, and low, and have labeled our results accordingly in We use these categories to frame the discussion below.

Table 5. Summary statistics for the model posterior distribution of NPV and PPV

Study	NPV μ (95% CrI)	PPV μ (95% CrI)	PPV Uncertainty Level
Insurance Fraud	0.800 (0.436, 0.976)	0.594 (0.167, 0.916)	high
Security Bugs	0.963 (0.962, 0.964)	0.672 (0.524, 0.816)	high
Criminal Statements	0.921 (0.881, 0.954)	0.402 (0.330, 0.477)	medium
Intrusion Detection	1.000 (1.000, 1.000)	0.517 (0.453, 0.586)	medium
Credit Card Fraud	1.000 (1.000, 1.000)	0.701 (0.642, 0.761)	medium
Data Loss Prevention	0.948 (0.942, 0.955)	0.829 (0.772, 0.879)	low
Management Fraud	0.992 (0.986, 0.997)	0.215 (0.165, 0.264)	low
Crowdfunding Fraud	0.998 (0.997, 0.998)	0.039 (0.029, 0.052)	low

High Uncertainty

The Insurance Fraud study (Artís et al. 2002) illustrates a PPV with high uncertainty. Even though the sensitivity and specificity seem good at 0.77, PPV is very uncertain with a mean of 0.594 and a 95% CrI that spans from 0.167 to 0.914. This high uncertainty stems from the prevalence. To determine the prior prevalence of fraudulent claims, the researchers asked the company's claim inspectors for an estimate, which they gave as 1/3 of all claims. We entered this into the model as $prior_{pos} = 1$ and $prior_{neg} = 2$. The resulting beta distribution is very wide, which results in the PPV also being very wide. The PPV is much lower than the sensitivity or specificity.

The Security Bugs (Jiang et al. 2020) study's PPV also has high uncertainty. This is despite its very certain sensitivity, in terms of credible interval (95% CrI: 0.023–0.030), specificity (95% CrI: 0.999–1.000), and prevalence (95% CrI: 0.37–0.39), which are a result of the high sample size. Despite the low uncertainty in these measures, the PPV has a high uncertainty, which results from the low accuracy due to low sensitivity (μ: 0.026). Calculating the raw PPV from the sample confusion matrix results in 107/(107 + 8) = 0.930, but incorporating the prevalence reported in the paper of 0.038 rather than the sample base rate of 0.193 and using a probabilistic model shows that this point estimate is misleading. Instead, Model 1 predicts a PPV with a mean of 0.672 whose 95% CrI stretches from 0.524 to 0.816. In situations of high uncertainty, results must be applied with caution.

Medium Uncertainty

The Criminal Statements, Intrusion Detection, and Credit Card Fraud studies all have PPVs of medium uncertainty. All three have more narrow intervals than the high uncertainty studies, with Criminal Statements having a 95% CrI that spans from 0.330 to 0.477, Intrusion Detection having a 95 % CrI between 0.453 and 0.586, and Credit Card Fraud having a 95% CrI that ranges from 0.642 to 0.762. Even in the medium uncertainty category, the prevalence can have a large effect on the PPV. Criminal Statements and Credit Card Statements are heavily influenced by the prevalence. Criminal Statements has a sensitivity of 0.797 but has a much lower mean PPV (0.402). Similarly, Credit Card Fraud has a high sample sensitivity of 99/101, but, because of its very low prior prevalence of 492/284807 (0.002), a much lower model mean PPV of 0.632 than the sensitivity alone would imply.

Low Uncertainty

The studies with the most certain PPVs include Data Loss Prevention, Management Fraud, and Crowdfunding Fraud. Of these, Crowdfunding fraud is the most certain with its certainty resulting from a very low and certain model prevalence of 0.007 (95% CrI: 0.029–0.052). This prevalence results in a very low PPV of 0.039. The low PPV itself also drives its certainty. The model assumes a binomial distribution for the PPV; therefore, the PPV distribution can't be lower than zero. As the distribution gets closer to zero or one it gets narrower since there isn't room for it to spread out without crossing zero or one. In low uncertainty contexts, the usefulness of signal detection measures is not greatly enhanced by incorporating results from Bayesian analysis.

DISCUSSION

In this research, we have shown that including prevalence and uncertainty in addition to standard performance measures may better provide researchers and practitioners the information they need to evaluate the utility of binary classification models. We see credible intervals indicating high uncertainty in the Insurance Fraud (Artís et al. 2002) and Security Bugs (Jiang et al. 2020) studies with 95% credible intervals of 75 and 28 percentage points, respectively. In both studies, what might look like reasonable accuracies when looking at point estimates become much less useful when uncertainty is included.

These high uncertainty estimates are dramatic, but estimating uncertainty is also useful when it is medium and low. First, without estimating uncertainty we can't know whether it is high, medium, or low. Second, in some contexts where the impact of misclassification is high. even what we labeled medium or low uncertainty in this paper may be too uncertain to be useful. Security researchers can estimate prevalence and then use that prevalence to demonstrate the usefulness of the detection method using PPV and NPV. Even when prevalence is only crudely estimated as demonstrated by the "one-third" estimate in the Insurance Fraud (Artís et al. 2002) study, researchers can still use that information to estimate include prevalence.

Despite prior relevant studies not incorporating uncertainty and prevalence, they still add value to the field. We see these previous studies and the current paper's contributions as points along a line of maturation. In the past, reporting accuracy, sensitivity, specificity, and precision from sample data were considered sufficient to understand a binary classification problem. Many studies have gone beyond this by reporting ROC curves and AUC. This paper furthers this process of maturation by including uncertainty to enable assessment of their practical application.

Recommendations for researchers

We recommend researchers estimate the amount of uncertainty in their models when performing security analytics research. They can then know whether their research has low, medium, or high uncertainty. When uncertainty is low, reporting point estimates and indicating that uncertainty was tested and is low should be sufficient. When uncertainty is high, however, we recommend researchers include measures and visualizations of that uncertainty. We also recommend researchers temper their conclusions based on the measures of uncertainty. If, for example, the PPV of a detection mechanism has a wide credible interval, basing conclusions on the mean PPV would be inappropriate if the low or high end of the credible interval would lead to a different conclusion. When uncertainty falls in the medium category, researchers should report it, but how much it affects the conclusions or how much emphasis they should put on the uncertainty depends on the context. In contexts like credit card fraud (where banks assume the cost of a certain amount of fraud) some uncertainty may not have much consequence: the bank would do the same thing whether the result was at the high or low end of the credible interval. In contexts with higher stakes such as security intrusions in critical infrastructure, results at different ends of the credible interval may change which security measure are put in place.

To fully estimate uncertainty researchers must estimate the prevalence the phenomenon of interest for their detection mechanism. In our extensive search for relevant studies that included prevalence we only found 13 (8 of which we included as exemplars) out of several hundred candidates. As the studies above show, without an estimate of prevalence it is impossible to know the level of uncertainty in PPV. We recommend researchers include some estimate of prevalence, even if that estimate is crude.

Limitations and Future Directions

Future studies should incorporate a larger variety of data sets to obtain a more comprehensive view of uncertainty in detection studies across the security literature. Using more samples will allow researchers to explore the issue of context dependency and incorporate a wider variety of prevalences more fully. Though we have demonstrated the impact of Bayesian data analysis across multiple studies and contexts (and provide additional results through the link in the appendix), we were restricted by the number of studies that provided sufficient information to conduct the analysis. The description of the samples used to train and test models is also inconsistent, preventing reproducibility of their results and more extended performance evaluation. We recommend that researchers include a confusion matrix and as mentioned above prevalence (known or estimated) for their data in the context of interest, if the prevalence is different than the sample base rate.

The Bayesian data analysis presented in this paper may not be appropriate for every study. For example, studies that are simply comparing the performance of various classification algorithms (Lessmann et al. 2015) or determining important classification features (Ho et al.

2016), prevalence and PPV may be less of a priority. It may also be beyond the scope of exploratory research. However, we advocate for including measures that could enable this analysis as research in a context accumulates. Reviewers and editors should also encourage publishing data and a consistent set of metrics. We have also shown that Bayesian data analysis may be less useful where the researcher can establish that their context and sample is one of low uncertainty.

CONCLUSION

This paper shows how Bayesian analysis can be used to better understand studies that use machine learning for security tasks. It demonstrates that studies that use machine learning may that expressing uncertainty in these studies provides a better measure of their usefulness. More specifically, it shows that incorporating Bayes data analysis techniques, specifically prevalence and PPV, provides a clearer view of a detection method's certainty and therefore usefulness.

REFERENCES

Artís, M., Ayuso, M., and Guillén, M. 2002. "Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims," *Journal of Risk and Insurance* (69:3), pp. 325–340. (https://doi.org/10.1111/1539-6975.00022).

Arya, M., and G, H. S. 2020. "DEAL – 'Deep Ensemble ALgorithm' Framework for Credit Card Fraud Detection in Real-Time Data Stream with Google TensorFlow," *Smart Science* (8:2), Taylor & Francis, pp. 71–83. (https://doi.org/10.1080/23080477.2020.1783491).

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. 2017. "Stan: A Probabilistic Programming Language," *Journal of Statistical Software* (76:1), pp. 1–32. (https://doi.org/10.18637/jss.v076.i01).

Cecchini, M., Aytug, H., Koehler, G. J., and Pathak, P. 2010. "Detecting Management Fraud in Public Companies," *Management Science* (56:7), INFORMS, pp. 1146–1160. (https://doi.org/10.1287/mnsc.1100.1174).

Dutta, H., Kwon, K. H., and Rao, H. R. 2018. "A System for Intergroup Prejudice Detection: The Case of Microblogging under Terrorist Attacks," *Decision Support Systems* (113), pp. 11–21. (https://doi.org/10.1016/j.dss.2018.06.003).

Edwards, W., Lindman, H., and Savage, L. J. 1963. "Bayesian Statistical Inference for Psychological Research," *Psychological Review* (70:3), US: American Psychological Association, pp. 193–242. (https://doi.org/10.1037/h0044139).

Faiz, M. F., Arshad, J., Alazab, M., and Shalaginov, A. 2020. "Predicting Likelihood of Legitimate Data Loss in Email DLP," *Future Generation Computer Systems* (110), pp. 744–757. (https://doi.org/10.1016/j.future.2019.11.004).

Fuller, C. M., Biros, D. P., and Wilson, R. L. 2009. "Decision Support for Determining Veracity via Linguistic-Based Cues," *Decision Support Systems* (46:3), pp. 695–703. (https://doi.org/10.1016/j.dss.2008.11.001).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. 2013. *Bayesian Data Analysis, Third Edition*, CRC Press.

Gelman, A., and Carpenter, B. 2020. "Bayesian Analysis of Tests with Unknown Specificity and Sensitivity," *MedRxiv*, Cold Spring Harbor Laboratory Press, 2020.05.22.20108944. (https://doi.org/10.1101/2020.05.22.20108944).

Geyer, C. J. 2011. "Introduction to Markov Chain Monte Carlo," in *Handbook of Markov Chain Monte Carlo*, CRC Press, pp. 3–47. (http://www.mcmchandbook.net/HandbookChapter1.pdf).

Green, D. M., and Swets, J. A. 1966. *Signal Detection Theory and Psychophysics*, Signal Detection Theory and Psychophysics, Oxford, England: John Wiley, pp. xi, 455.

Ho, S. M., Hancock, J. T., Booth, C., and Liu, X. 2016. "Computer-Mediated Deception: Strategies Revealed by Language-Action Cues in Spontaneous Communication," *Journal of Management Information Systems* (33:2), pp. 393–420. (https://doi.org/10.1080/07421222.2016.1205924).

Horgan, J. 2016. "Bayes's Theorem: What's the Big Deal?," *Scientific American Cross Check Blog*, , January 4. (https://blogs.scientificamerican.com/cross-check/bayes-s-theorem-what-s-the-big-deal/, accessed July 14, 2020).

Japkowicz, N., and Shah, M. 2011. *Evaluating Learning Algorithms: A Classification Perspective*, Cambridge University Press. (https://doi.org/10.1017/CBO9780511921803).

Jiang, Y., Lu, P., Su, X., and Wang, T. 2020. "LTRWES: A New Framework for Security Bug Report Detection," *Information and Software Technology* (124), p. 106314. (https://doi.org/10.1016/j.infsof.2020.106314).

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. 2015. "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," *European Journal of Operational Research* (247:1), pp. 124–136. (https://doi.org/10.1016/j.ejor.2015.05.030).

Li, X., Chen, W., Zhang, Q., and Wu, L. 2020. "Building Auto-Encoder Intrusion Detection System Based on Random Forest Feature Selection," *Computers & Security* (95), p. 101851. (https://doi.org/10.1016/j.cose.2020.101851).

National Research Council. 2003. *The Polygraph and Lie Detection*, Washington, DC: The National Academies Press. (https://doi.org/10.17226/10420).

Panigrahi, R., and Borah, S. 2018. "A Detailed Analysis of CICIDS2017 Dataset for Designing Intrusion Detection Systems," *International Journal of Engineering & Technology* (7:3.24), pp. 479–482. (https://doi.org/10.14419/ijet.v7i3.24.22797).

Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., Chen, S., Liu, D., and Li, J. 2020. "Performance Comparison and Current Challenges of Using Machine Learning Techniques in Cybersecurity," *Energies* (13:10), Multidisciplinary Digital Publishing Institute, p. 2509. (https://doi.org/10.3390/en13102509).

Siering, M., Koch, J.-A., and Deokar, A. V. 2016. "Detecting Fraudulent Behavior on Crowdfunding Platforms: The Role of Linguistic and Content-Based Cues in Static and Dynamic Contexts," *Journal of Management Information Systems* (33:2), pp. 421–455. (https://doi.org/10.1080/07421222.2016.1205930).

Twitchell, D. P., and Fuller, C. M. 2018. "Advancing the Assessment of Automated Deception Detection Systems: Incorporating Base Rate and Cost into System Evaluation," *Information Systems Journal* (1–24). (https://doi.org/10.1111/isj.12231).

APPENDIX

Model 1.

$$y \sim \text{Binomial}(n, p)$$

$$p = (sens)(prev_{test}) + (fpr)(1 - prev_{test})$$

$$n_{tp} \sim \text{Binomial}(n_{pos}, sens)$$

$$n_{fp} \sim \text{Binomial}(n_{neg}, fpr)$$

$$sens \sim \text{Beta}(1,1)$$

$$fpr \sim \text{Beta}(1,1)$$

$$prev \sim \text{Beta}(prior_{pos} + 1, prior_{neg} + 1)$$

Where *y* is the number people who tested positive, which is distributed binomially according to *n*, the number of tests, and *p* the rate of positives for the test. *p* is the true-positive rate (i.e., the sensitivity, *sens*) plus the false-positive rate (i.e., *fpr* or 1 minus the specificity). These are multiplied by the prevalence, *prev*. The model jointly estimates the parameters *sens, fpr*, and *prev*. *prev* is estimated from *y*, the number of positives and *n*, the number of tests, and it is constrained by *sens* and *fpr*. *sens* is estimated as a probability for a binomial distribution that generates the number of true positives, n_{tp}, from the number of positive cases, n_{pos}. *fpr* is similarly estimated from the number of known false positives n_{fp} and the number of negative cases, n_{neg}. $prev_{test}$ is the prevalence of the test set used to establish the sensitivity and sensitivity of the detection algorithm. $prior_{pos}$ and $\text{prior}_{\text{neg}}$ are the prevalence given by the study rather than estimated by the model. Finally, we assume we have no prior information about the three parameters, *sens* and *fpr*, which are given flat priors from the Beta distribution.

Once the model estimates *sens*, *fpr*, and *prev* we can calculate the positive predictive value (*ppv*), or the probability of a positive test given that a subject is positive, as follows:

$$ppv = \frac{(sens)(prev)}{p}$$

Similarly, we calculate the negative predictive value (*npv*, or the probability of a negative test given the subject is negative).

$$npv = \frac{(spec)(1 - prev)}{1 - p}$$

Code and Supplementary Analysis

The R and Stan code along with the data used for the analysis and creating the tables and figures are here: https://osf.io/rbe8p/?view_only=a1aa9b6553a4464ebe3761c56f3a0832.