Universal Data Anomaly Detection via Inverse Generative Adversary Network

Kursat Rasim Mestav, *Student Member, IEEE,* and Lang Tong, *Fellow, IEEE*

***Abstract*—The problem of detecting data anomaly under unknown probability distributions is considered. Whereas the probability distribution of the anomaly-free data is unknown, anomaly-free training samples are assumed to be available. For anomaly data, neither the underlying probability distribution is known nor anomaly data samples are available. A deep learning approach coupled with a statistical test based on coincidence is proposed where an inverse generative adversary network is trained to transform data to the classical uniform vs. nonuniform hypothesis testing problem. The proposed approach is particularly effective to detect persistent anomalies whose distributions have an overlapping domain with that of the anomaly-free distribution.**

***Index Terms*—Detection and estimation, Deep learning, Anomaly detection, Novelty detection, Semi-supervised learning, Coincidence test.**

I. Introduction

We consider the problem of universal data anomaly detection. By universal detection, we mean that both the anomaly-free and anomaly distributions are unknown and nonparametric. Specifically, under the null hypothesis $\mathcal{H}_0$ that models the anomaly-free data, measurements are from some unknown distribution f_0. Under the alternative $\mathcal{H}_1$ that models anomaly, measurements are from an unknown distribution that is at least ϵ distance away from f_0.

More precisely, given conditionally independent and identically distributed observations $Z_i, i = 1, \cdots, N$, we consider the following hypothesis testing problem:

$$\mathcal{H}_0 : Z_i \sim f_0 \quad \text{vs.} \quad \mathcal{H}_1 : Z_i \sim f_1 \in \mathcal{F}, \tag{1}$$

where $\mathcal{F} = \{f : ||f - f_0|| > \epsilon\}$ and $\|\cdot\|$ can be arbitrary distance measure such as the total variation or the KL divergence. In a paradigm of data-driven solutions to anomaly detection, we assume that only a set of training samples $\mathscr{Z}_0 = \{z_1, \cdots, z_T\}$ under $\mathcal{H}_0$ is available.

The assumption that the alternative distribution is unknown reflects the fact that data anomaly can happen in many ways, including the possibility that an adversary may have tampered the data in a man-in-the-middle attack [1]. Often in these cases, well-calibrated anomaly data are not available, or they are insufficient for learning.

The assumption that the distribution under the null hypothesis is unknown, but with some training data is also reasonable. For instance, data may be measured under a quasi-stationary environment that some samples can be authenticated but not enough to estimate the distribution accurately. A data-driven approach to anomaly detection may prefer using training samples directly to construct a test rather than estimating f_0 first from the training data and using the estimated distribution to construct a test.

The above hypothesis testing problem is general and has a wide range of applications such as healthcare informatics, image processing, and video surveillance [2]. In such applications, it is rare to have adequate training data for anomaly. A universal bad data detection method which only uses anomaly-free samples in training is needed in this setting.

Another application of the universal anomaly data detection is bad data detection in power system state estimation [3]. The bad data in power systems occur because of measuring equipment failures, communication channel failures, and cyber attacks. It is also hard to collect samples of such cases as they occur rarely. In addition, there may be more challenging to detect adversarial attacks such as unobservable data attacks [4].

A. Related Work

There are limited results in the classical statistics and the statistical signal processing literature that treats the hypothesis testing problem above. Indeed, pathological examples exist that consistent detection may not even be possible [2]. The problem is nonparametric and lacks a specific structure to place the problem in a well-studied class. The presence of training data under one hypothesis and the complete lack of training data in the other makes the problem a special machine learning problem. Here we review some recent approaches.

In the machine learning literature, the above problem is considered as semi-supervised anomaly detection [5]. Algorithms in this category can be classified into three groups: (i) density-based and nearest neighborhood-based techniques, (ii) one-class support vector machine algorithm and (iii) the auto-encoder based neural-network approaches.

Density-based methods estimate the probability density of the samples and detect anomalies according to this estimate.

Kursat Rasim Mestav (krm264@cornell.edu) and Lang Tong (lt35@cornell.edu) are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA.

Part of the work was presented at the 57th Allerton Conference on Communication, Control, and Computing.

This work is supported in part by the National Science Foundation under Award 1809830, 1932501, and, Power Systems and Engineering Research Center (PSERC) Research Project M-39

A nearest neighborhood-based technique is proposed in [6], where the anomaly-free data are assumed to be clustered with small nearest neighbor distances. A more recent density-based method using an energy-based model is proposed in [7]. Deep structured energy-based model is used to estimate the probability densities assuming the anomalies have a smaller density. Such assumptions may not be appropriate for anomalies that arise from data attacks where the attacker can design attacks to manipulate the data population.

The technique of one-class SVM [8] learns a hyperplane to separate an anomaly-free region from the rest of the space. A kernel function can be used to generalize the technique for nonlinearly separable hypotheses. For the universal data anomaly detection, choosing the right kernel function is highly nontrivial.

The auto-encoder based approaches [9]–[11] train an auto-encoder on anomaly-free samples. The reconstruction errors of new samples are used as test statistics for anomaly detection. The work in [10] is particularly relevant to the approach in this paper for its use of GAN in the training of auto-encoders. The authors of [11] introduce Gaussian mixture distribution for latent variables in training an auto-encoder, which has the potential to capture the underlying distribution of the anomaly-free data.

Although algorithms based on auto-encoder have shown promising performance in popular image processing data sets, the premise that anomaly data tend to generate greater mismatch in an auto-encoder trained with anomaly-free data is a suspect. When the anomaly and anomaly-free distributions have substantial overlapping domains, auto-encoder based techniques tend to have low false positive rates but also low true positive rate.

The key idea that allows us to distinguish the null hypothesis under f_0 from the alternative distributions in $\mathcal{F}$ is rooted in the classical birthday problem [12]: given M people, what is the coincidence probability P_c that there are at least two people having the same birthday?

It turns out that this probability is the lowest when the underlying birthday distribution is uniform [12]. This suggests that a test on some measure of coincidence can serve as a way to distinguish the uniform distribution from all other distribution. Such a test was proposed earlier by David in [13] and more recently by Paniski [14]. By thresholding the number of unique people who do not share a birthday with others, the Paninski's test is shown to have both false alarm and miss-detection approach to zero in the asymptotic regime.

B. Summary of contributions

The main contribution of this work is twofold. First, we propose a novel solution architecture consisting of an inverse generative adversarial network (IGAN), a quantizer, and a non-parametric coincidence test, illustrated in Fig. 1 The design of the three functional blocks is detailed in Sec. II.

Second, the algorithmic contribution of this work includes the use of IGAN as a nonparametric preprocessing step that transforms data modeled by an unknown distribution under the null hypothesis to data with a uniform distribution. This transformation allows us to apply the idea of coincidence test for anomaly detection. Whereas Paniski's test is formulated to achieve diminishing probabilities of two types of errors in the asymptotic regime, we provide a specific test threshold when the data samples are finite. We should also note that other uniformity tests [15], [16] can also be used within the framework developed here.

Fig. 1: A schematic of universal data anomaly detection (UAD).

Comparing with existing solutions, the proposed approach achieves diminishing detection error probabilities asymptotically assuming that IGAN is trained successfully. In the finite data sample regime, on the other hand, the proposed approach has low sample complexity in the sense that the number of testing samples is considerably smaller than the size of the quantization alphabet.

We show through numerical examples that the proposed universal data anomaly detection algorithm is effective for some of the very challenging anomaly data scenarios.

II. Universal Anomaly Detection

A. A Schematic for Universal Anomaly Data Detection

The idea of the proposed universal anomaly detection (UAD) is captured in the schematic in Fig. 1. Observation samples $\{Z_i\}$ are passed through an *inverse generator* H that maps $Z_i \sim f_0$ to uniformly distributed samples $Y_i \sim \mathcal{U}[0,1]$ in interval $[0,1]$. The existence of such a mapping is guaranteed by the fact that the cumulative distribution function $F_Z(\cdot)$ of Z_i is one (but not the only one) such mapping. Because f_0 is unknown, the mapping is to be learned from the available historical data as shown in Sec. II-B.

Upon successful training of the inverse generator H, under H_0, analog samples Y_i are approximately i.i.d. uniformly distributed. They are then quantized uniformly with M levels, which results in M-alphabet discrete uniformly distributed samples $X_i \sim \mathcal{U}(M)$.

A coincidence test using 1-coincidence statistic $K_1(x)$ produces the test outcome. The threshold is set depending on the level of acceptable false-alarm (the size) of the detector, the quantization level M, the number of test samples N, and the detection resolution ϵ (see Sec. II-D).

B. Inverse Generative Adversary Network

In contrast to GAN [17], IGAN aims to find the inverse of the generator of a data set. Fig. 2 shows a learning architecture of IGAN that consists of two simultaneously

trained neural networks: (i) an inverse generator and (ii) a discriminator. The training data passes through the inverse generator and the output is tested against synthetic uniformly distributed data by a discriminator. Ideally, the inverse generator converges to a function that transforms the distribution of the data to the uniform distribution.

We use the 1-Wasserstein distance to measure the similarity between probability distributions. As shown in [18], the Wasserstein distance measure tends to have improved stability of the training process.

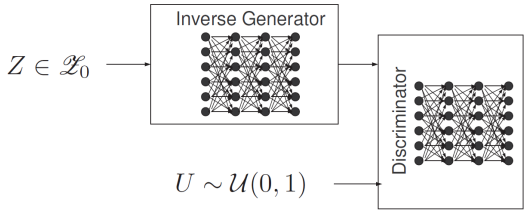

Fig. 2: An inverse generative adversary network (IGAN) learning of an inverse generator.

An implementation of IGAN is shown in Algorithm 1. In our approach, the weights in both networks θ, and ω are initialized randomly and updated with the learning rate of α. To enforce the Lipschitz constraint of the 1-Wasserstein distance, we used gradient penalty on the discriminator's loss function as it is proposed in [19]. The discriminator f_ω is updated more frequently than the inverse generator g_θ. We used Adam algorithm [20] for the weight updates.

Algorithm 1 IGAN. The experiments in the paper used the values $\alpha = 0.0001$, $\lambda = 10$, $m = 100$, $c = 5$.

Require: : α, the learning rate. λ, the gradient penalty coefficient. m, the batch size. c, the number of iterations of the discriminator per generator iteration.

for Number of training iterations **do**
  for $t = 0, 1, ..., c$ **do**
    for $i = 1, ..., m$ **do**
      Sample $U \sim \mathcal{U}(0,1)$ from uniform distribution.
      Sample $Z \sim f_0$ from real data.
      $\tilde{U} \leftarrow g_\theta(Z)$
      $\hat{U} \leftarrow \epsilon U + (1-\epsilon)\tilde{U}$
      $L_i \leftarrow f_\omega(\tilde{U}) - f_\omega(U) + \lambda(\|\nabla_{\hat{U}} f_{\omega(\hat{U})}\|_2 - 1)^2$
    Update the discriminator parameters ω by descending its stochastic gradient:
    $\nabla_\omega \big[\frac{1}{m}\sum_{i=1}^{m} L_i\big]$
  Sample $\{Z_i\}_{i=1}^m \sim f_0$ from real data .
  Update the inverse generator parameters θ by descending its stochastic gradient:
  $\nabla_\theta \big[\frac{1}{m}\sum_{i=1}^{m} -f_\omega(g_\theta(Z_i))\big]$

C. Quantization and Coincidence Test

Once the inverse generator is learned, we have the transformed data samples Y_i that are uniformly distributed under $\mathcal{H}_0$ and nonuniform under $\mathcal{H}_1$. Testing the uniformity of continuously distributed random samples without any assumptions on the density function is nontrivial [21]. Here we apply the M-level uniform quantization to Y_i, which gives us M-ary discrete random samples X_i that are uniformly distributed under $\mathcal{H}_0$. The distribution of X_i under $\mathcal{H}_1$ depends on the hyper-parameter M. Finding the optimal choice of M is beyond the scope of this paper. We assume that almost everywhere in $\mathcal{F}$, the inverse transformed and quantized samples X_i are ϵ distance away from being uniform. At the heart of the proposed approach is the coincidence test for uniformity proposed by Paninski [14] for the following binary hypotheses using conditionally IID samples $\{X_i, i = 1, \cdots, N\}$ from M-alphabet discrete distributions

$$\begin{aligned}\mathcal{H}_0' :&\quad X_i \sim P_0 = (\frac{1}{M}, \cdots, \frac{1}{M}),\\ \mathcal{H}_1' :&\quad X_i \sim P_1 \in \{p = (p_1, \cdots, p_M) |\ ||p - P_0|| > \epsilon\}.\end{aligned}$$

The intuition of uniformity test is that, when X_i are from the uniform distribution, the probability of coincidence is the lowest, and $K_1(x)$, the number of "unique" valued samples, is the highest. Thus, Paninski's test for uniformity is given by

$$K_1(x) \underset{\mathcal{H}_1'}{\overset{\mathcal{H}_0'}{\gtrless}} T_\alpha$$

where the threshold T_α is a function of false positive level α as well as the alphabet size (quantization level) M, the sample size N, and distance between two hypotheses ϵ.

Paninski showed that the coincidence test is consistent so long as N grows faster than $\sqrt{M}$ as $N = o(\frac{1}{\epsilon^4}\sqrt{M})$. Remarkably, the sample complexity can be significantly less than the size of the alphabet. A large-deviation bound is later established in [22].

D. Test threshold.

When the sample size N is finite, the threshold of the test statistics affects the true and false-positive probabilities of the detection.

Let $P_0(\mathcal{E})$ be the probability of event $\mathcal{E}$ under hypothesis $\mathcal{H}_0$. The threshold T_α of the K_1 coincidence test with the constraint on the false-positive probability to no greater than α is given by

$$T_\alpha = \min\{t : P_0(K_1 \le t) \le \alpha\}. \tag{2}$$

The computation of T_α amounts to evaluating $P_0(K = 1)$, which was given by Von Mises in [12]:

$$P_0(K_1 = k) = \sum_{j=k}^{M} (-1)^{j+k} \binom{j}{k} \binom{m}{j} \frac{N!}{(N-j)!} \frac{(M-j)^{N-j}}{M^N}.$$

III. Simulation

We tested the proposed methods on Gaussian and Gaussian Mixture models that are more commonly used in signal processing applications. We used a composite hypothesis for the alternative hypothesis to capture the variability of the alternative hypotheses. In general, when alternative (anomaly) distributions do not have overlapping domains, most existing techniques (the one proposed here) work well. Our experiments aimed to show the cases when the underlying distributions overlap. We evaluated the following 3 scenarios,

Case 1: $\mathcal{H}_0 : Z_i \sim \mathcal{N}(0,1)$ $v.s.$ $\mathcal{H}_1 : Z_i \sim \mathcal{N}(\mu,1)$ where $-1 < \mu < 1$.

Case 2: $\mathcal{H}_0 : Z_i \sim \mathcal{N}(0,1)$ $v.s.$ $\mathcal{H}_1 : Z_i \sim \mathcal{N}(0,\sigma)$ where $0.5 < \sigma < 0.8$.

Case 3: $\mathcal{H}_0 : Z_i \sim \mathcal{N}(0,1)$ $v.s.$ $\mathcal{H}_1 : Z_i = 0$.

We used 100000 anomaly-free training samples to train the iGAN. To test our algorithm, we generated 20000 batches of $N = 50$ samples from the distribution in $\mathcal{H}_0$ and another 20000 batches of $N = 50$ samples from the distribution in $\mathcal{H}_1$. For each batch, we varied the μ and σ. After using the IGAN, we simply used a fixed value of 200 for the quantization parameter M for all experiments. However, there is a space for improvement by choosing M more judiciously.

For each case, we compared the proposed approach with three major deep learning benchmarks: the Deep auto-encoding Gaussian mixture model (DAGMM) [11], the autoencoder approach based on the reconstruction error of f-AnoGAN [10] and the one-class SVM [8]. We used the scikit-learn library for the implementation of one-class SVM [23]. We implemented Universal Data Anomaly Detection and the algorithms in [11] and [10] using TensorFlow-GPU [24].

We trained the one-class SVM using Radial Basis Function (RBF) kernel. Using the RBF kernel one-class SVM tests samples according to closeness to the center of training samples.

Fig. 3 shows the receiver operating curve (ROC) of the tested algorithms where UAD showed significant improvement over the benchmarks. The two auto-encoder techniques showed peculiar characteristics that the TPR is below the trivial random choice detector. This behavior is further explored in test Cases 2 and 3.

In Case 2, we tested a case where the anomaly samples are much denser around the mean of the anomaly-free samples. It is the case totally opposite to what F-AnoGAN and one-class SVM rely on.

As shown in Fig. 4, F-AnoGAN and one-class SVM performed poorly with very low TPR, indicating that they classified almost all anomaly data as anomaly free. DAGMM showed better results due, perhaps, to the use of Gaussian mixture distribution in training.

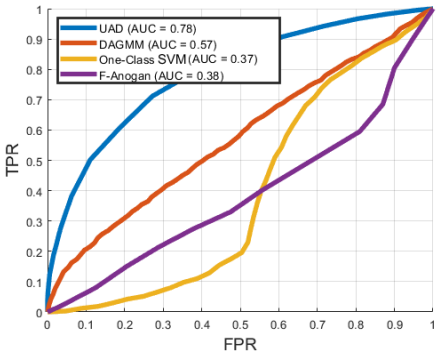

Fig. 3: ROC curve of the methods for Case 1.

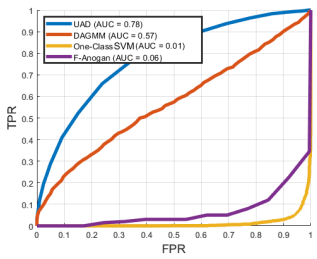

Fig. 4: ROC curve for Case 2.

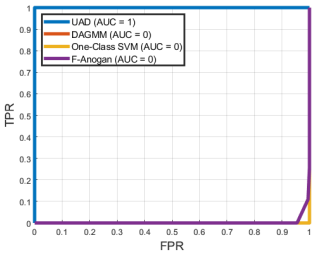

Fig. 5: ROC curve for Case 3.

Finally, we simulated an extreme scenario, where all anomalies are accumulated at the mean value of the anomaly-free samples. In practice, such a case may represent the physical "stuck-at" faults when sensors produced constant values. As shown in Fig. 5, UAD performed perfectly whereas competing techniques poorly. Granted that this is perhaps a special case that favors tests that derived based on distinguishing the underlying probability distributions, it does expose the potential pitfall of techniques such as auto-encoder methods based on direct matching data with anomaly-free samples.

IV. Conclusion

This paper presents a novel method for the problem of detecting data anomaly under semi-supervised settings based on distinguishing underlying probability distributions between the anomaly and anomaly-free data. Comparisons with some of the benchmark solutions showed the advantages of UAD when the distribution of the anomaly overlaps with that of the anomaly-free data. The developed technique is suitable for applications when an anomaly occurs persistently when multiple but limited samples are available for detection.

Acknowledgements

The authors wish to thank the anonymous reviewer for suggesting the use of WGAN-gp [19] in IGAN implementation.

REFERENCES

[1] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man in the middle attacks," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2027–2051, thirdquarter 2016.

[2] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215 – 249, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016516841300515X

[3] K. R. Mestav and L. Tong, "Learning the unobservable: High-resolution state estimation via deep learning," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2019, pp. 171–176.

[4] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 645–658, Dec 2011.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009. [Online]. Available: http://doi.acm.org/10.1145/1541880.1541882

[6] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, May 2000. [Online]. Available: http://doi.acm.org/10.1145/335191.335388

[7] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, p. 1100–1109.

[8] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pp. 582–588, 1999. [Online]. Available: http://dl.acm.org/citation.cfm?id=3009657.3009740

[9] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674, 2017. [Online]. Available: http://doi.acm.org/10.1145/3097983.3098052

[10] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, vol. 54, pp. 30 – 44, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841518302640

[11] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International Conference on Learning Representations*, 2018.

[12] R. Von Mises, "Über aufteilungs-und besetzungswahrscheinlichkeiten," *Revue de la Faculté des Sciences de l'Université d'Istanbul*, vol. 4, p. 145–163, 1939.

[13] F. N. David, "Two combinatorial test of whether a sample has come from a given population," *Biometrika*, vol. 37, no. 1/2, pp. 97–110, 1950. [Online]. Available: http://www.jstor.org/stable/2332152

[14] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4750–4755, Oct 2008.

[15] E. J. Dudewicz and E. C. V. D. Meulen, "Entropy-based tests of uniformity," *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 967–974, 1981.

[16] J. Acharya, A. Jafarpour, A. Orlitsky, and A. Suresh, "A competitive test for uniformity of monotone distributions," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, C. M. Carvalho and P. Ravikumar, Eds., vol. 31. Scottsdale, Arizona, USA: PMLR, 29 Apr–01 May 2013, pp. 57–65. [Online]. Available: http://proceedings.mlr.press/v31/acharya13a.html

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *ArXiv preprint: 1406.2661*, 2014.

[18] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *ArXiv preprint: 1701.07875*, 2017.

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5767–5777.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ArXiv preprint: 1412.6980*, 2014.

[21] M. Adamaszek, A. Czumaj, and C. Sohler, *Testing Monotone Continuous Distributions on High-Dimensional Real Cubes*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 228–233. [Online]. Available: https://doi.org/10.1007/978-3-642-16367-8_13

[22] D. Huang and S. Meyn, "Generalized error exponents for small sample universal hypothesis testing," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8157–8181, Dec 2013.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/