Tree-Based Algorithms for Weakly Supervised Anomaly Detection

Thorben Finke
Institut für Theoretische Teilchenphysik und Kosmologie
RWTH Aachen University
52074 Aachen, Germany
finke@physik.rwth-aachen.de

Marie Hein
Institut für Theoretische Teilchenphysik und Kosmologie
RWTH Aachen University
52074 Aachen, Germany
marie.hein@rwth-aachen.de

Gregor Kasieczka
Institut für Experimentalphysik
Universität Hamburg
22761 Hamburg, Germany
gregor.kasieczka@uni-hamburg.de

Michael Krämer
Institut für Theoretische Teilchenphysik und Kosmologie
RWTH Aachen University
52074 Aachen, Germany
mkraemer@physik.rwth-aachen.de

Alexander Mück
Institut für Theoretische Teilchenphysik und Kosmologie
RWTH Aachen University
52074 Aachen, Germany
mueck@physik.rwth-aachen.de

Parada Prangchaikul
Institut für Experimentalphysik
Universität Hamburg
22761 Hamburg, Germany
parada.prangchaikul@uni-hamburg.de

Tobias Quadfasel
Institut für Experimentalphysik
Universität Hamburg
22761 Hamburg, Germany
tobias.quadfasel@uni-hamburg.de

David Shih
NHETC, Dept. of Physics and Astronomy
Rutgers University
Piscataway, NJ 08854, USA
shih@physics.rutgers.edu

Manuel Sommerhalder
Institut für Experimentalphysik
Universität Hamburg
22761 Hamburg, Germany
manuel.sommerhalder@uni-hamburg.de

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

Abstract

Particle physics searches that rely on a specific signal model have so far failed to find evidence for physics beyond the Standard Model. Model-agnostic methods provide an important alternative approach, as they can analyze large amounts of data for a wide range of potential anomalies. Many state-of-the-art anomaly detection algorithms are based on a weakly supervised classification task, where the data samples are distinguished from samples of a background template. A key challenge for such algorithms is their performance degradation in the presence of uninformative features, which introduces model dependence by requiring feature selection. In this work, we propose the use of tree-based algorithms in weakly supervised anomaly detection with tabular data, as they are not only significantly faster to train and evaluate than deep learning–based methods, but are also robust to uninformative features and achieve better performance.

1 Introduction

Despite rigorous efforts by the particle physics research program at the Large Hadron Collider (LHC) and other facilities, no conclusive evidence for physics beyond the Standard Model (BSM) has yet been found. However, most searches for BSM physics are driven by specific signal models, such as supersymmetric models, models based on extra-dimensional theories, etc. Since it is not feasible to search for every possible signal model, and the regions of LHC phase space yet to be searched are vast, model-agnostic machine learning methods have recently attracted considerable interest [1–11].

In particular, these methods have been shown to be sensitive to signal models that produce localized resonances, such as a particle decaying into two jets (so-called di-jet resonances). In this case, model-agnostic methods must detect the anomalous signal resonance, which is immersed in an overwhelming amount of background consisting of quantum chromodynamics (QCD) multi-jet events. Within this *resonant anomaly detection* regime, many state-of-the-art methods rely on a weakly supervised classification task: First, events are separated into a signal region (SR), where the majority of signal events are assumed to reside, and the sidebands (SB), which consist almost entirely of background events. A classifier is then trained to distinguish samples of the (potentially anomalous) data in the SR from samples of a background template typically learned from events in the SB. Since model-agnostic methods must be unsupervised, the only labels used in classification are whether an event is from the SR or the background template, making the task significantly more challenging than direct signal/background discrimination.

In this difficult weakly supervised regime, the discriminative power of the input features — i.e. the separation of signal and background in a feature — is crucial, since the inclusion of uninformative features adds noise to an already noisy classification task. For model-agnostic anomaly detection, this is a significant problem, as it requires the selection of features specific to a particular signal model in order to achieve sufficient sensitivity. In particular, deep neural network (NN) classifiers — which are used in most state-of-the-art resonant anomaly detection methods — show a significant drop in performance once uninformative features are included. However, a fully model-independent search would use a high-dimensional input space containing all available features, the majority of which are expected to be uninformative for a given signal model.

One possible approach to mitigate this problem is to use classification algorithms that are more robust to uninformative features. One family of such algorithms are boosted decision trees (BDTs), which have been shown to outperform deep learning–based classifiers on tabular datasets while being less affected by features that carry little or no information [12, 13]. In this contribution, we investigate the use of BDTs in the weakly supervised regime, specifically testing their robustness to uninformative features as well as their overall performance compared to NN classifiers. In addition, we test the use of ensembling multiple BDT classifiers to further improve performance, which is possible due to the significantly faster training time of these algorithms.

2 Methods

In a traditional model-specific search for new physics, the goal is to find the best classifier that distinguishes the signal of interest from the background. In a model-agnostic search, the signal is unknown, which means that the classification must be entirely data driven. In such a classification task, the goal is to distinguish potentially anomalous events in the data from background events. Given the Neyman-Pearson lemma [14], the most powerful classifier in the model-agnostic regime is based on the likelihood ratio between the data and background probability densities in the input features $\boldsymbol{x}$:

$$R_{\mathrm{optimal}}(\boldsymbol{x}) = \frac{p_{\mathrm{data}}(\boldsymbol{x})}{p_{\mathrm{bg}}(\boldsymbol{x})} \,. \tag{1}$$

The likelihood ratio R_{optimal}, which we will refer to as the *optimal anomaly score*, is only an idealization that cannot be achieved in practice: The true background density $p_{\mathrm{bg}}(\boldsymbol{x})$ is unknown and can only be approximated by imperfect physics and detector simulations. The densities are also intractable and must be approximated from samples. Existing resonant anomaly detection methods attempt to approximate this value by training binary classifiers that discriminate between samples drawn from the SR and samples drawn from a data-driven background template. It should be noted, however, that while most methods have focused on the case of resonant anomaly detection, the idea of approximating the optimal anomaly score with a classifier is not limited to this use case, see e.g. [15, 16].

Since the main point of interest of this work is the weakly supervised classification task, we use an *idealized anomaly detector* (IAD) where the events of the background template and the background events within the SR data come from the identical distribution. This ensures that the quality of the data-driven background template generation is fully factored out, and any improvement in classifier performance is based solely on the improvement of the classification algorithm itself.

Algorithms based on histogrammed gradient boosting have been shown not only to achieve state-of-the-art performance on tabular data [17], but also to allow for fast training on large datasets. Additionally, due to their simple tree structure and easily computable feature importance scores, they provide a better understanding of the model's decision-making compared to NNs, which is a critical requirement in high energy physics use-cases. Therefore, we also use histogrammed gradient boosting in this work, using the `HistGradientBoostingClassifier` implementation of the `scikit-learn` package [18]. The hyperparameters used are set to the `scikit-learn` defaults.

To further improve performance and increase stability, we use an ensemble of N independent training runs, each using a randomized training and validation sample. We have chosen the training and validation sets to be equal in terms of sample size. After all models in the ensemble have been trained, their N predictions on the separate test set are averaged to assess the final performance. Unless otherwise stated, we use $N = 50$ models per ensemble.

For comparison with an NN classifier, we implemented a simple fully connected neural network using the `Keras` package [19] with a `TensorFlow` [20] backend. The NN consisted of three layers, each with 64 nodes. The activation function used for the hidden layers was Rectified Linear Unit (ReLU), while the softmax activation function was used for the output layer. The binary cross-entropy loss function was used to measure the performance of the model and to guide the training process. The optimization algorithm chosen was Adam [21], with a learning rate of 0.001. During training, the data was processed in batches of 128 samples and the training process was repeated for a total of 100 epochs. In terms of ensembling, we use the same scheme as described above for the BDT classifiers. Also, for both classifier models, early stopping is enabled with a patience of 10 iterations/epoch.

3 Dataset and metrics

The studies are performed on the LHC Olympics 2020 R&D dataset [22, 23]. The dataset is based on simulated di-jet events. It contains $1\,000\,000$ QCD dijet events for the background and $100\,000$ signal events. The signal model used is a hypothetical Z' particle decaying into hadrons: $Z' \to X(\to qq)Y(\to qq)$, leading to two jets in the final state, each containing two sub-jets of quarks. The masses of the particles are $m_{Z'} = 3.5\,\mathrm{TeV}$, $m_X = 500\,\mathrm{GeV}$ and $m_Y = 100\,\mathrm{GeV}$.

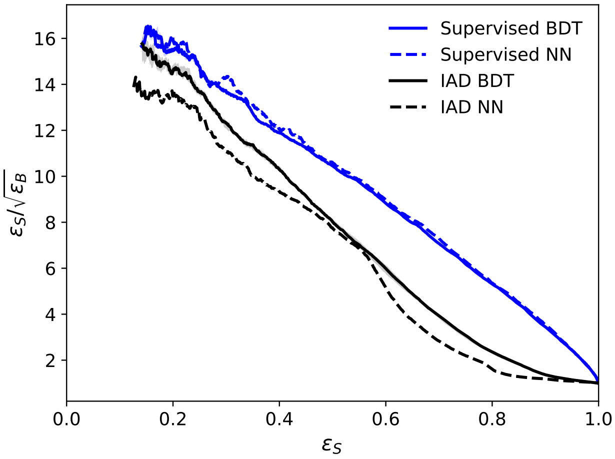

Figure 1: SIC curve comparison of BDT and NN classifier models for fully supervised and IAD classification tasks.

The SR for the weakly supervised classification is defined between $3.3\,\mathrm{TeV}$ and $3.7\,\mathrm{TeV}$, which contains about $120\,000$ background events. Unless otherwise stated, we inject $1\,000$ of the signal events into the dataset, resulting in 722 signal events in the SR, corresponding to $S/B = 6 \times 10^{-3}$. For the background template in the SR, $612\,858$ additional QCD events from the same simulation are used [24]. For the training and validation sets combined, a total of about 272k background template events and 120k SR data events are used. The training/validation split is always set to $50\,\%$. For the separate test set we use 340k background and 20k signal events in the SR. For the comparison study with a fully supervised classifier, the mentioned 120k SR data events are replaced by 54k SR signal events.

The features used for the training are the invariant mass of the lighter of the two jets m_{J_1}, the difference in jet mass between the two jets, $\Delta m_J = m_{J_2} - m_{J_1}$ as well as the subjettiness ratios τ_{21}^{J1} and τ_{21}^{J2} of the jets [25, 26].

Throughout this paper, the main performance metric of interest is the improvement in significance using the weakly supervised classifier relative to the inclusive significance. Therefore, the significance improvement characteristic (SIC) is used, which is defined as

$$\mathrm{SIC} = \frac{\epsilon_S}{\sqrt{\epsilon_B}}\,, \qquad (2)$$

where ϵ_S is the fraction of correctly identified signal events (also referred to as signal efficiency or true positive rate) and ϵ_B is the fraction of background events misidentified as signal (also referred to as background efficiency or false positive rate).

4 Results

Fig. 1 shows a performance comparison of a BDT and an NN classifier model for both the weakly supervised IAD and the fully supervised tasks. It can be seen that BDT classifiers are at least as powerful as NNs and that in the case of IAD, the BDT even outperforms the NN classifier for low signal efficiencies.

To investigate the robustness of the classifiers in a more realistic anomaly detection setting, we introduce features into the dataset that are completely uninformative: feature values are drawn from a standard Gaussian distribution and added to the original training features for both the SR data and the background template samples. This procedure can be repeated several times, so that the influence of adding an increasing number of uninformative features can be assessed. The result of this study can be seen in Figure 2. Looking at the results of the NN classifier, the rapid drop in performance is

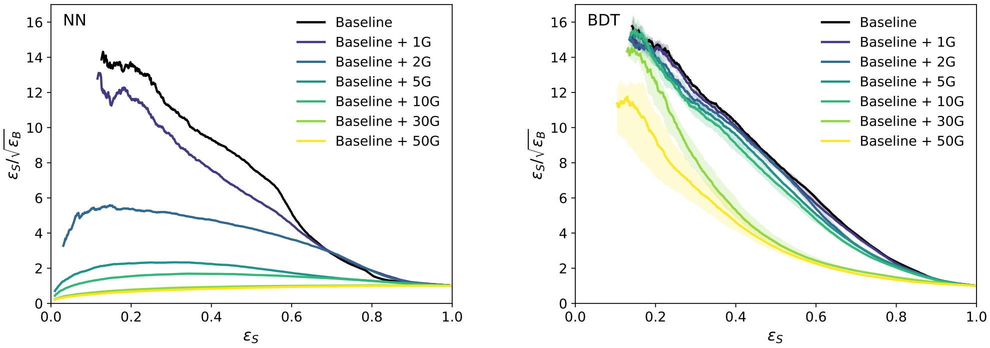

Figure 2: Impact of uninformative features on the IAD classification task. Comparison is done showing SIC curves for NN classifier models (left panel) and BDT classifier models (right panel). The legend label *Baseline* refers to the original feature set and the numbers before "G" indicate the number of Gaussian noise features added. For the BDT, solid lines correspond to the median performance of ten independent training runs and areas of reduced opacity describe the respective inner 68 % confidence intervals. Also, for 30 and 50 Gaussian noise features, N was increased to 100. For the NN classifiers, only a single ensemble was trained to reduce computational time.

clearly visible even when only a few uninformative features are added: Already with two Gaussian features, the NN performance is halved, and it deteriorates rapidly when five and more Gaussian features are added.

For the BDT classifiers, however, the situation is entirely different: up to ten Gaussian features, the performance is hardly affected. At 30 Gaussian features, the performance starts to degrade, but still achieves a significant improvement in significance, especially at low signal efficiencies. When 50 Gaussian features are added, the performance decreases across the entire signal-to-noise ratio range and the variance of the results increases significantly. However, a large proportion of the baseline significance can still be retained even in this challenging regime. This shows that the use of histogrammed boosted decision trees and ensembling allows truly model-agnostic searches, where classifiers are trained on a large number of different input features and only a small fraction are expected to contain information for any given signal.

5 Conclusion

We have investigated the use of (ensembled) histogrammed gradient boosted decision trees for weakly supervised anomaly detection. Our studies show that BDT-based classifiers not only perform at least as well as NN classifiers when no uninformative features are present, but also perform well in a scenario where the vast majority of features are completely uninformative and where the performance of NN classifiers breaks down. It can therefore be concluded that the robustness of BDTs to uninformative features, which has been extensively studied in the literature for supervised classification [12, 13, 27], also extends to the weakly supervised regime. Since this work considered an idealized scenario, further research should focus on possible limitations and extensions of the method. In particular, the performance when using low-level features (e.g. jet constituent features) as well as the behavior for non-idealized background estimation remain to be studied.

In summary, the use of BDT-based classifiers represents a significant improvement in anomaly detection methods based on weak supervision, allowing truly model-agnostic searches without the need to fine-tune features to a specific family of signal models.

Table 1: Subjettiness feature sets considered for training. Full training feature sets always include m_{J_1} and Δm_J as well. Details of the observables are given in the text.

Name	# features	Features
Baseline	4	$\{m_{J_1},\, \Delta m_J,\, \tau_{21}^{\beta=1,J_1},\, \tau_{21}^{\beta=1,J_2}\}$
Extended 1	10	$\{m_{J_1},\, \Delta m_J,\, \tau_{N,N-1}^{\beta=1,J_1},\, \tau_{N,N-1}^{\beta=1,J_2}\}$ for $2 \leq N \leq 5$
Extended 2	12	$\{m_{J_1},\, \Delta m_J,\, \tau_{N}^{\beta=1,J_1},\, \tau_{N}^{\beta=1,J_2}\}$ for $N \leq 5$
Extended 3	56	$\{m_{J_1},\, \Delta m_J,\, \tau_{N}^{\beta,J_1},\, \tau_{N}^{\beta,J_2}\}$ for $N \leq 9$ and $\beta \in \{0.5, 1, 2\}$

Appendix

Expanding the pool of features

We have established that algorithms based on histogrammed gradient boosted trees perform well when a significant number of uninformative features is added. In particular, we studied the performance in the worst-case scenario by adding features that consist of pure Gaussian noise. In a more realistic scenario, additional features based on actual physics variables would be used instead.

Therefore, we also studied this case by including extended feature sets containing jet substructure information. A summary of the different feature sets is provided in Table 1. Three extended feature sets exist: Extended set 1 includes additional subjettiness *ratios*, extended set 2 also includes the *individual* subjettiness features up to τ_5 and extended set 3 containing 54 subjettiness features that were computed using different angular weighting parameters β. The performance comparison was again made for NN-based classifiers and BDT-based classifiers and the result can be seen in Figure 3.

As shown in previous studies, the NN classifier is highly sensitive to the selection of input features: There is a significant performance drop when using extended set 1 compared to the baseline performance, while the BDT classifier achieves a higher SIC with respect to the baseline for the same

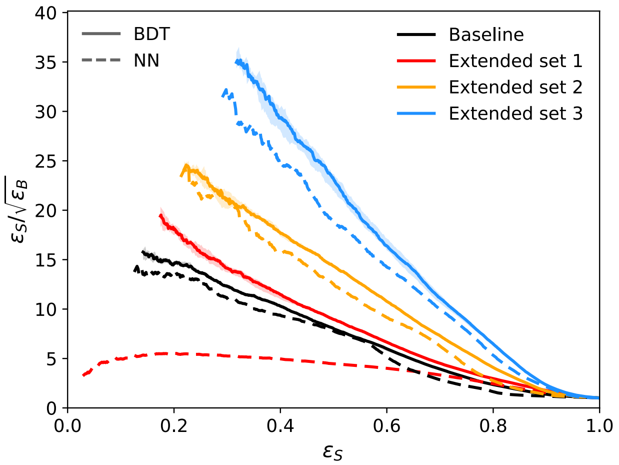

Figure 3: Impact of including additional physics features in the IAD classification task. Comparison is done showing SIC curves for NN classifier models (dashed lines) and BDT classifier models (solid lines). The different feature sets are summarized in Table 1. For the BDT, the solid lines correspond to the median performance of ten independent training runs and areas of reduced opacity describe the respective inner 68 % confidence intervals. For the NN classifiers, only a single ensemble was trained to reduce computational time.

feature set. For extended feature sets 2 and 3, the performance of both classifiers increases due to the information content from the additional subjettiness features. This behavior is in line with the previous findings: While for the BDT-based classifier the performance is at least as good as the baseline when more features are added, the NN performance can increase or decrease significantly depending on which particular set of features is used.

References

[1] E. M. Metodiev, B. Nachman, and J. Thaler, "Classification without labels: Learning from mixed samples in high energy physics," *JHEP*, vol. 10, p. 174, 2017. DOI: 10.1007/JHEP10(2017)174. arXiv: 1708.02949 [hep-ph].

[2] J. H. Collins, K. Howe, and B. Nachman, "Anomaly Detection for Resonant New Physics with Machine Learning," *Phys. Rev. Lett.*, vol. 121, no. 24, p. 241 803, 2018. DOI: 10.1103/PhysRevLett.121.241803. arXiv: 1805.02664 [hep-ph].

[3] J. H. Collins, K. Howe, and B. Nachman, "Extending the search for new resonances with machine learning," *Phys. Rev.*, vol. D99, no. 1, p. 014 038, 2019. DOI: 10.1103/PhysRevD.99.014038. arXiv: 1902.02634 [hep-ph].

[4] B. Nachman and D. Shih, "Anomaly Detection with Density Estimation," *Phys. Rev. D*, vol. 101, p. 075 042, 2020. DOI: 10.1103/PhysRevD.101.075042. arXiv: 2001.04990 [hep-ph].

[5] A. Andreassen, B. Nachman, and D. Shih, "Simulation Assisted Likelihood-free Anomaly Detection," *Phys. Rev. D*, vol. 101, no. 9, p. 095 004, 2020. DOI: 10.1103/PhysRevD.101.095004. arXiv: 2001.05001 [hep-ph].

[6] K. Benkendorfer, L. L. Pottier, and B. Nachman, "Simulation-assisted decorrelation for resonant anomaly detection," *Phys. Rev. D*, vol. 104, no. 3, p. 035 003, 2021. DOI: 10.1103/PhysRevD.104.035003. arXiv: 2009.02205 [hep-ph].

[7] A. Hallin *et al.*, "Classifying anomalies through outer density estimation," *Phys. Rev. D*, vol. 106, no. 5, p. 055 006, 2022. DOI: 10.1103/PhysRevD.106.055006. arXiv: 2109.00546 [hep-ph].

[8] J. A. Raine, S. Klein, D. Sengupta, and T. Golling, "CURTAINs for your sliding window: Constructing unobserved regions by transforming adjacent intervals," *Front. Big Data*, vol. 6, p. 899 345, 2023. DOI: 10.3389/fdata.2023.899345. arXiv: 2203.09470 [hep-ph].

[9] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih, and M. Sommerhalder, "Resonant anomaly detection without background sculpting," *Phys. Rev. D*, vol. 107, no. 11, p. 114 012, 2023. DOI: 10.1103/PhysRevD.107.114012. arXiv: 2210.14924 [hep-ph].

[10] T. Golling, S. Klein, R. Mastandrea, and B. Nachman, "Flow-enhanced transportation for anomaly detection," *Phys. Rev. D*, vol. 107, no. 9, p. 096 025, 2023. DOI: 10.1103/PhysRevD.107.096025. arXiv: 2212.11285 [hep-ph].

[11] T. Golling *et al.*, "The Interplay of Machine Learning–based Resonant Anomaly Detection Methods," arXiv: 2307.11157 [hep-ph].

[12] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2022. DOI: 10.1109/TNNLS.2022.3229161.

[13] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," arXiv: 2207.08815 [cs.LG].

[14] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Phil. Trans. Roy. Soc. Lond. A*, vol. 231, no. 694-706, pp. 289–337, 1933. DOI: 10.1098/rsta.1933.0009.

[15] T. Finke, M. Krämer, M. Lipp, and A. Mück, "Boosting mono-jet searches with model-agnostic machine learning," *JHEP*, vol. 08, p. 015, 2022. DOI: 10.1007/JHEP08(2022)015. arXiv: 2204.11889 [hep-ph].

[16] G. Bickendorf, M. Drees, G. Kasieczka, C. Krause, and D. Shih, "Combining Resonant and Tail-based Anomaly Detection," Sep. 2023. arXiv: 2309.12918 [hep-ph].

[17] H. Carlens, *State of competitive machine learning in 2022*, https://mlcontests.com/state-of-competitive-data-science-2022.

[18] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] F. Chollet *et al.*, *Keras*, `https://keras.io`, 2015.

[20] Martín Abadi *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from `https://www.tensorflow.org/`, 2015.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," eprint: `arXiv:1412.6980`.

[22] G. Kasieczka, B. Nachman, and D. Shih, *R&d dataset for lhc olympics 2020 anomaly detection challenge*, `https://zenodo.org/record/6466204`, 2019.

[23] G. Kasieczka *et al.*, "The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics," arXiv: `2101.08320 [hep-ph]`.

[24] D. Shih, *Additional qcd background events for lhco2020 r&d (signal region only)*, `https://zenodo.org/record/5759086`, 2021.

[25] J. Thaler and K. Van Tilburg, "Identifying Boosted Objects with N-subjettiness," *JHEP*, vol. 03, p. 015, 2011. DOI: `10.1007/JHEP03(2011)015`. arXiv: `1011.2268 [hep-ph]`.

[26] J. Thaler and K. Van Tilburg, "Maximizing Boosted Top Identification by Minimizing N-subjettiness," *JHEP*, vol. 02, p. 093, 2012. DOI: `10.1007/JHEP02(2012)093`. arXiv: `1108.2701 [hep-ph]`.

[27] A. Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," in *Proceedings of the Twenty-First International Conference on Machine Learning*, ser. ICML '04, Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 78, ISBN: 1581138385. DOI: `10.1145/1015330.1015435`. [Online]. Available: `https://doi.org/10.1145/1015330.1015435`.