UNDERSTANDING POISONING ATTACKS WITH VISUALIZATION TECHNIQUES

(Technical Paper)

ALGORITHMIC BIAS AND DISCRIMINATION: AN INTERDISCIPLINARY PERSPECTIVE

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree

Bachelor of Science, School of Engineering

Evan Rose

Fall, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Advisors

Catherine Baritaud, Department of Engineering and Society

David Evans, Department of Computer Science

Developments in machine learning and artificial intelligence (AI) over the past several years have enabled breakthroughs in many different areas, including medicine, autonomous vehicles, computational biology, image and text generation, and many more. As the successes of AI and machine learning become even more abundant, the technologies will move into increasingly specialized applications, including critical applications.

If AI and machine learning are to continue to serve critical roles in everyday life, then they must be implemented as fair, secure systems. In particular, they must respect the social values of the social context into which they are embedded, and furthermore they must be robust to attacks from adversaries who wish to compromise them.

This technical research and its associated loosely-coupled STS research will study ways to make machine learning and AI more reliable. The technical research lies at the intersection of computer security and machine learning, called adversarial machine learning, and examines the behavior of machine learning systems in the presence of sophisticated adversaries who seek to undermine the reliability of the machine learning system (Biggio & Roli, 2018). More specifically, the technical research will study the properties of so-called poisoning attacks, a type of adversarial attack against a machine learning system. The STS research, loosely coupled with the technical topic, examines more broadly the ways in which AI technologies are influenced by, and in turn influence, the social context in which they operate. More specifically, the STS research studies the issues of fairness, bias, and discrimination in AI technologies, and how the social and technical considerations related to these issues interact with each other to result in a cohesive network of sociotechnical relationships through which the development of such technologies can be understood.

The technical work is in collaboration with Fnu Suya, Ph.D Candidate in the Department of Computer Science at the University of the Virginia, and Professor David Evans, Professor of Computer Science in the Department of Computer Science at the University of Virginia. The technical project is complete, with the only remaining deliverable being the technical report, to be completed in February 2023. The STS work is in collaboration with Professor Catherine Baritaud, Professor of STS in the Department of Engineering and Society. The writing of the STS report will take place from December 2022 to April 2023. As a milestone, at least 50% of the STS report will be completed by the end of February 2023.

UNDERSTANDING POISONING ATTACKS WITH VISUALIZATION TECHNIQUES

Machine learning is a process by which an automated system can use preexisting experience, in the form of data, to learn patterns about a complicated system. For example, a machine learning model might attempt to predict stock prices based on past market behavior, or distinguish between images of cats and dogs based on previously labeled examples.

Machine learning is a powerful and flexible tool capable of being adapted to a large number of computational problems. In general, the success of a machine learning system is dependent on a critical set of assumptions regarding the environment in which the system is developed and deployed. For example, it is important that the training data used to train the model is representative of the inputs the model will see in deployment, and moreover that the information reflected in the training data is accurate.

In small-scale applications, these assumptions may hold, but as machine learning systems are developed for increasingly broad and sensitive applications, the invalidity of these assumptions in general is becoming readily apparent. The response is the emergence of an area of research known as adversarial machine learning, which attempts to study machine learning in the presence of adversaries who seek to compromise the performance of the machine learning model (Biggio & Roli, 2018).

One kind of adversarial attack against machine learning takes place at the data collection step. Machine learning pipelines typically collect large amounts of training data from several untrusted sources, as illustrated in Figure 1. In a poisoning attack, an adversary controlling some small fraction of the training data chooses that data in order to induce some specific behavior in the trained model (Nelson et al., 2008, Biggio et al., 2012).

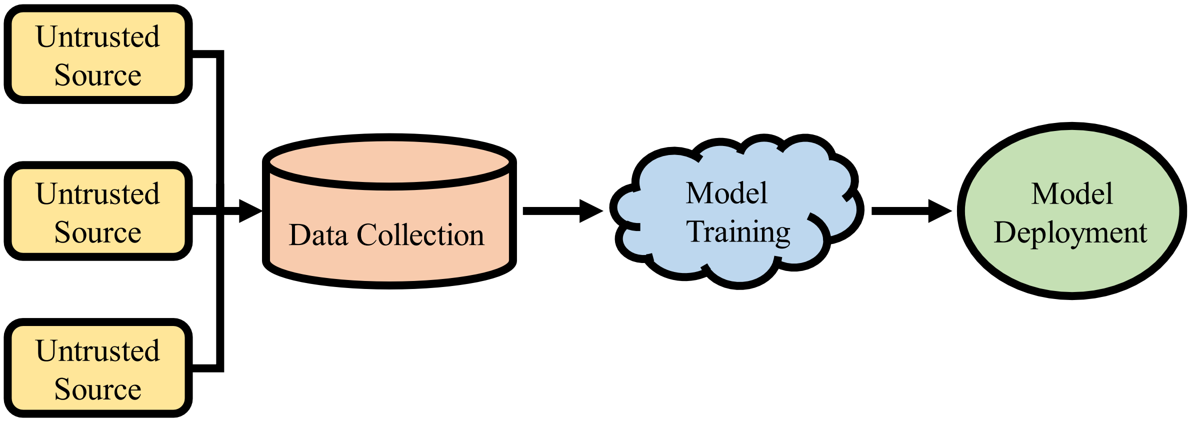

Figure 1. Prototypical Machine Learning Pipeline. Notice how the data collection step collects data from several untrusted sources, which creates an attack surface for a poisoning attack (Rose, 2022).

Previous works on poisoning attacks consider two extreme attacker objectives: indiscriminate attacks, in which the attacker tries to reduce overall model accuracy (Xiao et al., 2012), and instance-targeted attacks, in which the attacker tries to reduce model accuracy on a specific instance (Shafahi et al., 2018). Recently, Jagielski et al. (2021) introduced the subpopulation attack, a more realistic attacker objective which focuses on compromising the model's behavior on a preselected subpopulation while not affecting the model behavior on points outside the subpopulation.

The broad goal of this technical research is to study and understand subpopulation poisoning attacks against machine learning. More specifically, the technical work studies a question enabled by the recent introduction of subpopulation poisoning attacks: *which subpopulations are must susceptible to a subpopulation poisoning attack, and what affects their susceptibility?*

METHODS OF DEVELOPMENT

To achieve the stated research goal, we have designed and conducted a number of subpopulation poisoning attack experiments. First, poisoning attacks were studied in a synthetic dataset setting, where direct control over the dataset parameters and choice of subpopulation could be exploited to produce direct visualizations of poisoning attacks over a variety of simplified scenarios. Then, poisoning attacks against a real-world dataset were conducted, and dimensionality reduction techniques were employed to visualize the result.

The anticipated outcome of this work is an improved understanding of subpopulation poisoning attacks. The results of the experiments, as well as interactive visualizations and an accompanying poisoning attack demo, have been developed and are available as an explorable conference paper at uvasrg.github.io/poisoning (Rose et al., 2022).

ALGORITHMIC BIAS AND DISCRIMINATION: AN INTERDISCIPLINARY PERSPECTIVE

Machine learning and AI technologies are being developed and deployed at such an unprecedented pace that it is difficult to imagine a future in which such technologies do not dominate the technological landscape. With the widespread success and exploding popularity of these technologies, it is more important now than ever to consider deeply the full set of ramifications of using such technologies. Critical limitations in these technologies must be rapidly identified and accounted for so that they do not limit the opportunities of those affected by them.

Algorithmic risk assessment is a process which takes as input information about a scenario and produces as output a score reflecting the risk associated with that scenario. For example, a loan application screening pipeline could use a risk assessment tool to predict the risk of a borrower defaulting on a loan, providing valuable information for the lender when evaluating loan proposals from applicants (Beshr, 2020; Quinn, 2021). In the criminal justice system, risk assessment instruments are used to predict risks related to defendants and criminals, such as the risk that a defendant fails to appear in court or the risk that a convicted criminal will commit a violent crime in the future (Chohlas-Wood, 2020, para. 2).

However, algorithmic assessment tools have caused controversy in the past. In a recent audit of the popular COMPAS criminal risk assessment tool, Angwin et al. (2016) argued that the tool, whose purpose was to predict recidivism, exhibited strong bias against black defendants. The main basis for this claim was rooted in a statistical analysis which demonstrated a higher false positive rate for black defendants, as well as several examples examining the tool’s behavior across different defendants whose recidivism outcome was known. While others, including Northpointe, the company responsible for creating COMPAS, have pointed out flaws (Dieterich, 2016; Gong, 2016) in the original analysis by Angwin et al., the story still prompted lots of discussion surrounding algorithmic fairness. Abe Gong (2016), data scientist, puts it strikingly after performing an analysis which undermines the claims made by Angwin et al.:

“Powerful algorithms can be harmful and unfair, even when they’re unbiased in a strictly technical sense” (para. 27).

Such challenges are not just limited to risk assessment tools, either. In the medical field, a prime opportunity for the application of machine learning presents itself: image-based cancer diagnosis. Using techniques from computer vision, which uses computational methods to process images and other visual data, it is possible to train a machine learning classifier to identify skin cancer just by taking a picture of the patient’s skin (“AI Skin Cancer”, 2021). However, the model developer’s ability to create a sufficiently fair model is limited by the current skin cancer datasets. As David Wen et al. (2022) showed in their review of publicly available skin cancer datasets, there is “substantial under-representation of darker skin types” (p. e64). In the few datasets that did report data on ethnicity, “no images were from individuals with an African, Afro-Caribbean, or South Asian background” (p. e71). These limitations could result in cancer detection models which perform worse on underrepresented groups, since some skin cancers manifest differently depending on skin color (p. e71).

BIAS AND DISCRIMINATION AS A SOCIAL CONSTRUCTION

Ferrer et al. (2021) analyzed the current state of bias and discrimination in AI through a cross-disciplinary lens, identifying several key sociotechnical forces contributing to the development of bias and discrimination considerations in AI technologies. Ferrer et al. defined bias for the purposes of their discussion as “deviation from the standard,” and remarked that bias defined this way is necessary for statistical analysis of data (p. 72). Discrimination, on the other hand, is defined legally as “the unfair treatment of an individual (or group) based on certain protected characteristics” (p.72).

Ferrer et al. have identified four perspectives necessary for understanding algorithmic discrimination through a cross-disciplinary lens: technical, social, legal, and ethical (p. 72). In the technical domain, bias is often conflated with discrimination, resulting in a research focus on measuring the presence of bias in AI without addressing the question of determining discrimination from bias (p. 72). In the legal domain, legislation codifies social values against discrimination, but suffers from an inability to respond quickly to dynamic technical and social changes, resulting in requirements which are often unactionable (p. 76). The social domain plays an important role in determining what values are important to preserve in AI systems, and additionally introduces historical considerations (p. 77). The ethical domain relieves the mentioned shortcomings of the legal domain in that the ethical domain is more reactive, but also may be difficult to reconcile with the technical (pp. 77-78).

Ferrer et al. have argued that the very notion of algorithmic discrimination couples technology and society in an inseparable way. This idea is reflected in the discussion on algorithmic bias and discrimination in the examples of COMPAS and skin cancer datasets: in both cases, a strictly technical consideration of the algorithmic task would have obscured the presence of issues related to algorithmic discrimination.

LIMITATIONS OF CURRENT WORK AND OPPORTUNITIES FOR FUTIRE RESEARCH

In addition to listing several sociotechnical factors related to bias and discrimination in AI, Ferrer et al. identify several open challenges to be addressed moving forward. One of these challenges is in transforming social values regarding what constitutes discrimination into an actionable, technical specification (p. 78). As it stands, determining and rectifying discrimination in AI depends too heavily on a social and ethical context and does not admit a natural transformation into technical systems, especially systems without a human in the loop. Another key issue is improving AI literacy among the public, a task which in its definition requires bridging between the technical and the social (pp. 78-79). A final challenge identified by Ferrer et al. is related to discrimination-aware AI, which is concerned with using technical approaches to identify and inform users and developers about social issues like discrimination in an automated way (p. 79).

The existence of the above open problems is of special importance, as it demonstrates that more work is necessary to understand and improve the current methods for approaching discrimination in AI. Critically, the authors assert that the coupling of discrimination and AI is intrinsically sociotechnical in nature, and thus benefits from being examined from a sociotechnical perspective (p. 72). A main priority of this research is thus elaborate on the relationships from the viewpoint of a specific STS framework. For this purpose, we will use the Social Construction of Technology (SCOT) framework (Pinch & Bijker, 1984) to explicitly map out the relationships between social and technical forces. In this sociotechnical framework, interactions between the engineer and social groups are modeled in order to understand the effect each social group has on the development of a technological artifact and how the development of the artifact influences the social groups.

One of the advantages of modeling the social construction of AI technologies using SCOT lies in the clear identification of key social groups. In the cases of risk assessment and algorithmic cancer diagnosis, the main problem with the technological artifact was directly related to the social groups the artifact affected, and more precisely how the artifact behaved differently for different social groups. In general, the important interactions between a technological artifact and its surrounding social context can be extracted by first examining the relevant social groups the artifact affects. The full range of social groups affected by AI technologies is extensive, and more examples of such key social groups are illustrated in Figure 2.

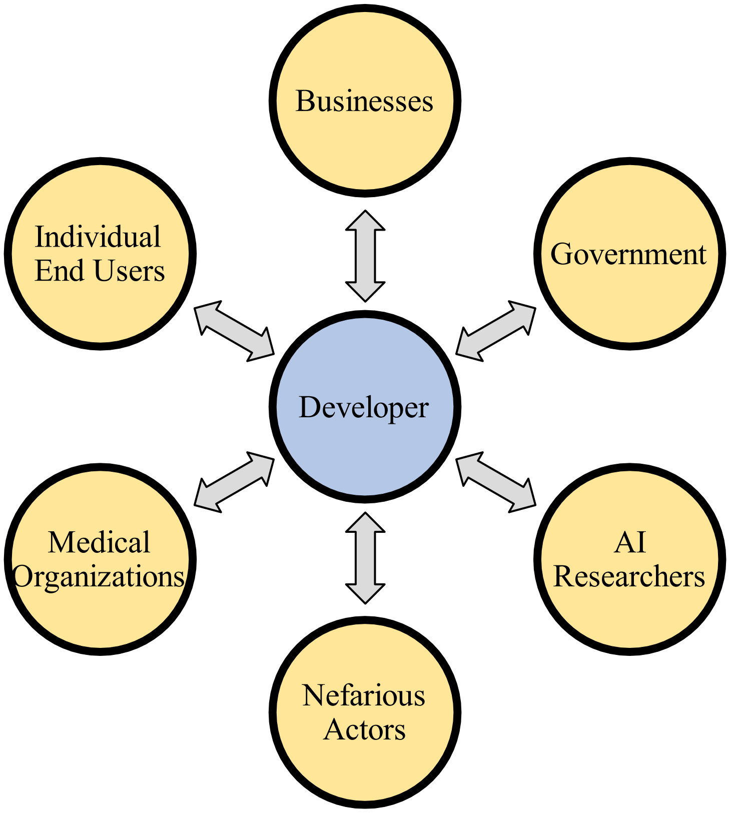

Figure 2. Relevant Social Groups for AI SCOT Model. Bidirectional arrows indicate two-way interactions between relevant social groups and the developer of an AI technological artifact. (Adapted by Rose (2022) from Carlson, 2009)

The SCOT framework will be applied in order to understand how social and technical forces influence each other in the development in AI systems. More specifically, this sociotechnical research will study the development of technical strategies for addressing bias and discrimination in machine learning systems, as well as how society's perception of bias and discrimination, as exhibited by machine learning systems, is affected by their deployment.

The SCOT framework additionally provides a way to directly model relationships between social groups, problems associated with the technological artifact, and potential solutions respecting the goals of the social group (Pinch & Bijker, 1984, pp. 415-419). This line of analysis stresses the uncertainty of technological development by indicating multiple possible technical resolutions to a single problem (p. 416). A preliminary application of this concept is illustrated in Figure 3, which shows one way in which a relevant social group affected by AI technologies may be associated with different problems and solutions.

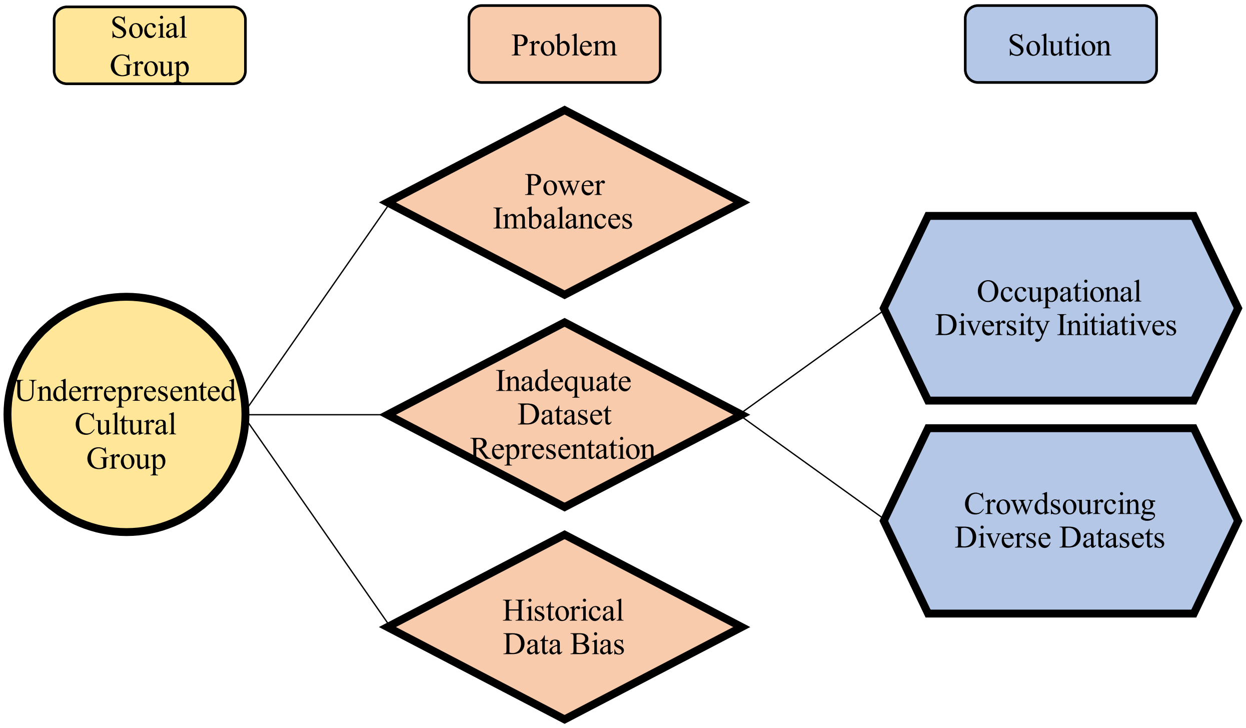
Social Group
Problem
Solution
Power Imbalances
Occupational Diversity Initiatives
Underrepresented Cultural Group
Inadequate Dataset Representation
Crowdsourcing Diverse Datasets
Historical Data Bias

Figure 3. Expansion of AI SCOT Model. A SCOT model can be expanded to cover social groups, problems faced by those social groups, and technical solutions to those problems. (Adapted by Rose (2022) from Pinch & Bijker, 1984)

The intended outcome of this research is to obtain a more precise description of the sociotechnical relationships driving the development of algorithmic discrimination. The resulting analysis should be able to offer insight into the open problems discussed above. This STS research project will culminate in a scholarly article examining the sociotechnical interactions which have and continue to shape the development of AI and machine learning technologies, especially with respect to bias and discrimination detection and mitigation.

ARTIFICIAL INTELLIGENCE AND SOCIETY

Recent growth in the popularity of AI and machine learning places a distinct pressure on those developing AI technologies. Not only must developers conscientiously craft their algorithms to be mindful of the social contexts in which they operate, but they must also prepare to reinforce their algorithms against nefarious actors who may wish to override critical systems. On the other hand, society should realize the ways in which it can affect and is affected by the development of AI technologies. As ethical, social, and security concerns continue to surround AI, it will be the social forces which ultimately determine the future of AI technology.

REFERENCES

AI skin cancer assessment tool granted award by NHSX. (2021, June 18). Med-Tech Innovation. https://www.med-technews.com/api/content/5bcd2cea-d00c-11eb-a08a-1244d5f7c7c6/

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Beshr, S. (2020, November 2). *A machine learning approach to credit risk assessment*. Medium. https://towardsdatascience.com/a-machine-learning-approach-to-credit-risk-assessment-ba8eda1cd11f

Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 1467–1474.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, *84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Chohlas-Wood, A. (2020, June 19). Understanding risk assessment instruments in criminal justice. *Brookings*. https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Northpointe Inc. Research Department. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Ferrer, X., Nuenen, T. van, Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine*, *40*(2), 72–80. https://doi.org/10.1109/MTS.2021.3056293

Gong, A. (2016, August 3). Ethics for powerful algorithms (1 of 4). *Medium*. https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84

Jagielski, M., Severi, G., Harger, N., & Oprea, A. (2021). Subpopulation data poisoning attacks. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 3104–3122.

Nelson, B., Barreno, M., Jack Chi, F., Joseph, A., Rubinstein, B., Saini, U., Sutton, C., & Xia, K. (2008). Exploiting machine learning to subvert your spam filter. *First USENIX Workshop on Large Scale Exploits and Emergent Threats*, *7*, 1–9.

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Quinn, M. (2021, November 11). *Machine learning in loan risk analysis*. https://www.bluegranite.com/blog/machine-learning-in-loan-risk-analysis

Rose, E. (2022). *Prototypical machine learning pipeline.* [Figure 1]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Rose, E. (2022). *Relevant Social Groups for AI SCOT Model.* [Figure 2]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Rose, E. (2022). *Expansion of AI SCOT Model.* [Figure 3]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Rose, E., Suya, F., & Evans, D. (2022, September 21). *Poisoning attacks and subpopulation susceptibility* [Paper presentation]. 5th Workshop on Visualization for AI Explainability. https://uvasrg.github.io/poisoning/

Shafahi, A., Ronny Huang, W., Najibi, M., Suciu, O., Struder, C., Dumitras, T., & Goldstein, T. (2018, December). Poison frogs! targeted clean-label poisoning attacks on neural networks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

Wen, D., Khan, S. M., Xu, A. J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., Perez, C. de B., Denniston, A. K., Liu, X., & Matin, R. N. (2022). Characteristics of publicly available skin cancer image datasets: A systematic review. *The Lancet Digital Health*, *4*(1), e64–e74. https://doi.org/10.1016/S2589-7500(21)00252-1

Xiao, H., Xiao, H., & Eckert, C. (2012, August). Adversarial label flips attack on support vector machines. *ECAI*.