World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Understanding Overfitting in AI and its impact on Cybersecurity

Brad Russell *

College of Business and Technology, Columbia Southern University, United States.

World Journal of Advanced Research and Reviews, 2025, 27(02), 361-372

Publication history: Received on 27 June 2025; revised on 01 August; accepted on 04 August 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.2.2859

Abstract

Artificial intelligence is becoming an essential part of cybersecurity, but its advantages are not reaching everyone equally. This paper investigates the problem of overfitting in AI security systems, which happens when models learn too much from past data and fail to recognize new or evolving threats. Using a comparative case study approach, we analyze events like the 2017 WannaCry ransomware outbreak and examine how both large technology firms and public sector organizations respond to these challenges. Our findings show that overfitting is not just a technical flaw but is shaped by decisions about resources, maintenance, and access to expertise. Organizations with more funding and technical capacity are able to keep their AI models current and effective, while smaller and less resourced groups often rely on outdated systems that leave them exposed to attacks. This pattern raises concerns about growing inequality in digital security. The study concludes that addressing overfitting requires not only better technical solutions but also policy changes and industry standards that support fairness, transparency, and adaptability. By making advanced cybersecurity tools more accessible and focusing on ongoing improvement, we can help ensure that digital protection is available to all organizations, not just the privileged few.

Keywords: Artificial Intelligence; Cybersecurity; Digital Divide; Machine Learning; Overfitting; Security Vulnerabilities

1. Introduction

Artificial intelligence and machine learning have become essential tools in the fight against cyber threats. As attacks grow more complex and frequent, AI-driven systems offer the ability to analyze vast amounts of data quickly and identify subtle patterns that might escape human detection. These systems power critical security functions such as intrusion detection, malware classification, and automated threat response. They promise faster and more accurate defenses that scale with the growing demands of modern networks [1,2].

Yet, despite these advances, AI in cybersecurity faces a hidden but serious challenge: overfitting. Overfitting happens when a model learns the details of its training data too well, including noise and irrelevant patterns. This means the model performs excellently on known data but struggles with new, unseen threats. In cybersecurity, this is a major problem because attackers constantly change their tactics. A model that is too narrowly focused on past data may fail to detect new attack methods or generate excessive false alarms, which can overwhelm security teams and reduce trust in automated defenses [3,4].

Overfitting also creates privacy risks. Research shows that models that memorize their training data become vulnerable to attacks that can reveal sensitive information about individuals or organizations. These privacy attacks, such as membership inference, allow adversaries to extract confidential details from the model itself. This is especially concerning in cybersecurity, where models often process highly sensitive information from critical infrastructure and private networks [5,6].

[*] Corresponding author: Brad Russell

What makes the problem even more complex is the adversarial nature of cybersecurity. Unlike many other fields, attackers actively probe AI models for weaknesses. Overfitting not only reduces a model's ability to generalize but also makes it easier for attackers to fool or manipulate the system. This creates a pressing need for AI models that are both accurate and resilient against adversarial tactics [7,8].

Despite the growing use of AI in security, the specific risks and consequences of overfitting have not been fully explored. Much of the existing research focuses on algorithmic fairness, explainability, or adversarial robustness, but the unique challenges posed by overfitting in cybersecurity remain underexamined. There is a clear need for comprehensive studies that address these gaps, combining theoretical insights with practical solutions [9,10].

This paper seeks to fill that gap. We start by explaining what causes overfitting and how it shows up in machine learning models, especially those used in security. Then, we examine the operational, privacy, and adversarial risks linked to overfitting, supported by recent studies and real-world examples. After that, we review current strategies to reduce overfitting, including traditional methods like regularization and data augmentation, as well as newer privacy-preserving techniques. Finally, we outline important research directions to help build AI systems that are robust, adaptable, and trustworthy in the face of evolving cyber threats. By placing overfitting within the broader cybersecurity context, this work aims to guide both researchers and practitioners. Our goal is to provide actionable insights that will help improve AI security tools and protect critical digital infrastructure.

2. Literature Review: Overfitting in AI for Cybersecurity

2.1. Theoretical Foundations of Overfitting

The concept of overfitting has long been recognized as a core challenge in machine learning. Early foundational work by Geman, Bienenstock, and Doursat (1992) described overfitting as the tendency of models to capture noise and idiosyncrasies in training data, rather than learning generalizable patterns [1]. In the context of AI, overfitting is often seen when a model achieves high accuracy on training data but performs poorly on new, unseen data [2]. This gap between training and real-world performance is particularly problematic in cybersecurity, where models must adapt to constantly evolving threats.

The risk of overfitting is not limited to any one type of machine learning. It appears in supervised learning, where models can memorize training labels, and in unsupervised learning, where clustering algorithms may identify patterns that exist only in the training set [3]. Reinforcement learning agents, too, can overfit by exploiting quirks of simulated environments that do not hold in real-world scenarios [4]. The underlying causes are often insufficient data diversity, excessive model complexity, or a lack of appropriate regularization [5,6].

Theoretical frameworks for understanding overfitting have evolved alongside advances in AI. Modern research emphasizes the importance of balancing model capacity with data availability and diversity. The bias-variance tradeoff remains a central concept, highlighting the tension between underfitting (too simple) and overfitting (too complex) models [7]. In cybersecurity, this balance is especially delicate, as the cost of missed detections or false positives can be high [8].

2.2. Overfitting and Security System Vulnerabilities

The introduction of AI into cybersecurity has brought both promise and new vulnerabilities. Intrusion detection systems (IDS), for example, rely on machine learning models to distinguish between benign and malicious activity. When these models overfit, they may only recognize attacks that closely resemble those in the training data, leaving organizations exposed to novel or slightly altered threats [9,10]. This limitation is exacerbated by the adversarial nature of cybersecurity, where attackers actively probe for weaknesses.

Malware detection is another area where overfitting can undermine security. Some models rely on features that are specific to the training dataset, such as certain byte sequences or metadata, rather than behavioral indicators that generalize across malware families [11]. As a result, these models may fail to detect new malware variants or generate excessive false positives, which can overwhelm security analysts and reduce trust in automated systems [12].

Empirical studies have documented these challenges. For instance, Ring et al. (2019) reviewed network-based IDS datasets and found that many models trained on static datasets failed to generalize to real-world traffic, often due to overfitting on outdated or unrepresentative data [13]. Similarly, Saxe and Berlin (2015) demonstrated that deep learning models for malware detection could achieve high accuracy on benchmark datasets but struggled with real-world samples that differed from the training distribution [11]

2.3. Overfitting, Privacy and Adversarial Attacks

The consequences of overfitting extend beyond operational failures. Recent research has shown that overfitted models are more susceptible to privacy attacks, such as membership inference and attribute inference. Shokri et al. (2017) demonstrated that adversaries could determine whether specific data points were used in training a model, exploiting the model's memorization of its training set [14]. Yeom et al. (2018) further explored the connection between overfitting and privacy risk, showing that models with high generalization gaps are more vulnerable to information leakage [15].

Adversarial attacks also exploit overfitting. Biggio and Roli (2018) reviewed how attackers can craft inputs that deceive overfitted models, causing them to misclassify malicious activity as benign or vice versa [16]. The dynamic nature of cyber threats means that models must not only avoid overfitting but also remain robust against adversarial manipulation. This dual challenge is unique to cybersecurity and underscores the need for continuous monitoring and adaptation [17].

2.4. Mitigation Strategies and Open Challenges

A range of strategies has been developed to combat overfitting, including regularization, data augmentation, cross-validation, and ensemble methods [18,19]. In cybersecurity, these approaches must be tailored to address the specific challenges of evolving threats and adversarial environments. For example, data augmentation techniques can help increase training diversity, but generating realistic synthetic attack data remains difficult [20]. Ensemble methods can improve robustness but may increase computational demands and complexity [21].

Despite these advances, significant gaps remain. Many studies focus on improving model accuracy without systematically addressing generalization or robustness. There is also a lack of standardized benchmarks for evaluating overfitting and resilience in real-world cybersecurity settings [22]. As AI adoption in security operations accelerates, there is a growing need for research that bridges the gap between theoretical advances and practical deployment.

This literature review highlights the complexity of overfitting in AI for cybersecurity. It is not simply a technical nuisance but a multidimensional challenge that affects operational effectiveness, privacy, and resilience. Addressing these issues requires ongoing collaboration between researchers, practitioners, and policymakers.

3. Methodology: Multi-Case Technical Analysis of Overfitting in Cybersecurity

3.1. Research Design

This study adopts a multi-case technical analysis to investigate how overfitting emerges and impacts AI-driven cybersecurity systems in real-world settings. The research design focuses on understanding both the technical and operational factors that contribute to overfitting, as well as the consequences for system performance, privacy, and adversarial robustness. Rather than treating overfitting as a generic statistical artifact, this approach centers the unique context of cybersecurity, where threat landscapes are dynamic and adversarial actors actively seek to exploit model weaknesses [1,2].

Three cases were selected to reflect a range of cybersecurity applications: intrusion detection, malware classification, and adaptive threat response. By comparing cases across different domains and data environments, this methodology reveals patterns and vulnerabilities that might be overlooked in single-case studies. The comparative approach provides insights into how overfitting manifests under different operational constraints and adversarial pressures, while also highlighting the limitations of current mitigation strategies [3].

The research combines quantitative analysis of model performance with qualitative review of operational incidents and published reports. Performance metrics such as training and validation accuracy, generalization gap, and false positive/negative rates are used to assess the extent of overfitting. At the same time, case documentation - including academic literature, industry white papers, and incident reports - provides context for understanding how overfitting affects real-world deployments [4,5].

3.2. Case Selection and Data Sources

3.2.1. Case Selection Criteria

Cases were chosen to maximize technical diversity and relevance to current cybersecurity challenges. The selection includes:

- A widely used open-source intrusion detection system trained on a benchmark dataset.
- A commercial malware detection platform evaluated on proprietary and public datasets.
- An adaptive defense system employing reinforcement learning in both simulated and live network environments.

This selection strategy prioritizes breadth of application and diversity of data sources, aiming to uncover both common and unique patterns of overfitting across the cybersecurity landscape [6].

3.2.2. Data Sources

- Technical Benchmarks: Publicly available datasets such as NSL-KDD, CICIDS, and VirusShare provide standardized testbeds for evaluating overfitting in intrusion and malware detection [7,8].
- Operational Reports: Incident analyses and post-mortems from industry sources offer insights into how overfitting has contributed to security failures or privacy breaches in practice [9].
- Academic Literature: Peer-reviewed studies supply detailed accounts of experimental setups, model architectures, and mitigation techniques, supporting cross-case comparison [10].

3.3. Analytical Methods and Limitations

3.3.1. Quantitative Analysis

Model performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve. Overfitting is assessed by comparing training and validation results, analyzing the generalization gap, and tracking model behavior on adversarial perturbed data [11,12].

3.3.2. Qualitative Analysis

Operational incidents are examined through document review, including published case studies, technical blogs, and security advisories. This mixed-methods approach acknowledges that numbers alone may not capture the full operational impact of overfitting, while qualitative evidence provides context for technical findings [13].

3.3.3. Limitations

The study is limited by the availability of high-quality, up-to-date datasets and the lack of access to proprietary data from commercial deployments. While the selected cases represent a range of applications, they cannot capture every possible scenario in the fast-evolving field of cybersecurity. Additionally, public reports may underrepresent failed deployments or security incidents due to confidentiality concerns.

Despite these constraints, the combination of quantitative and qualitative analysis provides a robust foundation for identifying patterns, vulnerabilities, and research gaps in the management of overfitting in cybersecurity AI.

3.3.4. Ethical Approval

This research relies exclusively on publicly available datasets, published literature, and documented incident reports. No experiments were conducted on human or animal subjects, and no proprietary or confidential data were accessed.

4. Case Study Analysis

4.1. The 2017 WannaCry Attack: Overfitting and AI-Driven Detection Failures

4.1.1. Technical and Operational Context

In May 2017, the WannaCry ransomware attack swept across the globe, crippling hospitals, businesses, and government agencies in more than 150 countries [1]. The attack exploited a vulnerability in Microsoft Windows, encrypting victims' files and demanding ransom payments in Bitcoin. While the rapid spread and impact of WannaCry have been widely studied, less attention has been paid to how overfitting in AI-driven security systems contributed to the scale of the disaster.

At the time, many organizations had deployed machine learning-based intrusion detection systems (IDS) and endpoint protection platforms. These systems were trained primarily on historical malware samples and known attack vectors. In practice, this meant that the models often learned to recognize specific patterns, signatures, and behaviors present in their training datasets, which were heavily weighted toward previously observed threats [2,3].

The demographic context of organizations affected by WannaCry was diverse. The attack did not discriminate by geography or sector, but its impact was particularly severe in public health systems, such as the United Kingdom's National Health Service (NHS), where legacy infrastructure and limited cybersecurity budgets left systems vulnerable. Many of these organizations relied on off-the-shelf security solutions that promised AI-powered protection but were often optimized for compliance benchmarks rather than real-world adaptability [4].

4.1.2. Overfitting Analysis

The technical failure of many AI-driven security tools during WannaCry can be traced to overfitting. Models trained on historical malware and attack patterns failed to generalize to the novel propagation techniques used by WannaCry. Specifically, these models often relied on static features - such as file hashes, network ports, or known command-and-control domains - that did not capture the behavioral anomalies introduced by the worm's rapid lateral movement and use of the EternalBlue exploit [5].

Security researchers later found that many IDS and endpoint detection models flagged only a small fraction of WannaCry's activities, missing the early signs of infection and allowing the ransomware to spread unchecked. In some cases, the models generated excessive false positives for unrelated activity, further overwhelming analysts and delaying effective response [6].

This case highlights how overfitting can create a false sense of security. Organizations believed their AI tools would detect new threats, but the models' narrow focus on past data left them blind to innovative attack methods. The generalization gap - where models perform well on test data similar to training samples but poorly on truly novel threats - was starkly exposed by WannaCry's success [7].

4.1.3. Organizational Impact and Response

The impact on organizations was immediate and severe. Hospitals in the NHS had to cancel surgeries and divert ambulances, while businesses lost access to critical data and systems. The financial costs ran into billions of dollars, but the human cost - in delayed medical care and disrupted services - was equally significant [1,4].

In the aftermath, organizations and vendors undertook major reviews of their AI-driven security tools. Many shifted toward incorporating behavioral analytics, anomaly detection, and continual retraining with fresh threat intelligence to reduce overfitting risks [8]. Some vendors adopted ensemble methods and adversarial training to make models more robust against novel attacks. There was also a renewed emphasis on human-in-the-loop systems, where expert analysts could override or supplement automated decisions in real time [9].

The WannaCry case serves as a cautionary tale. It demonstrates that overfitting is not just a theoretical concern but a practical vulnerability with real-world consequences. The lessons learned have shaped the development of more resilient, adaptive AI systems in cybersecurity, but the risk of overfitting remains an ongoing challenge as attackers continue to innovate.

4.2. Google's Adaptive Security in Silicon Valley: A model for Robust AI

4.2.1. Technical and Organizational Context

Silicon Valley is home to some of the world's most resource-rich technology companies, including Google. The region is known for its high median income, advanced infrastructure, and a corporate culture that prioritizes innovation and technical excellence [1]. Google's cybersecurity division operates with access to vast data resources, a highly skilled workforce, and the financial flexibility to invest in continual improvement of its AI-driven security systems [2].

Unlike many organizations that rely on static datasets and periodic model retraining, Google's security teams have implemented a continuous learning framework. This approach integrates live data streams from billions of endpoints, enabling models to adapt rapidly to new threats. The company's commitment to transparency and technical rigor is reflected in its regular publication of security research, open-source tools, and detailed incident reports [3].

4.2.2. Proactive Mitigation

Google's approach to AI security stands in stark contrast to the reactive, compliance-driven models seen in less-resourced organizations. The company employs a suite of advanced techniques to prevent overfitting, including:

- Continuous Data Refresh: Models are retrained on up-to-date, diverse datasets that include recent attack patterns, reducing the risk of memorizing outdated or irrelevant features [4].
- Adversarial Testing: Security teams routinely simulate new attack scenarios, using adversarial examples to probe model weaknesses and ensure robustness against novel threats [5].
- Ensemble Methods: Google combines multiple models with different architectures and training regimens to balance bias and variance, achieving both high accuracy and generalizability [6].
- Human-in-the-Loop Oversight: Automated systems are supplemented by expert analysts who review model outputs, provide feedback, and intervene when anomalies are detected [7].

4.2.3. Organizational Impact and Community Engagement

The benefits of Google's proactive stance are evident in its track record. The company has successfully detected and neutralized several large-scale threats, including zero-day exploits and nation-state attacks, before they could cause widespread harm [9]. When vulnerabilities are discovered, Google's security team works closely with affected parties, providing timely disclosures and mitigation guidance.

Community engagement is also a hallmark of Google's approach. The company invests in public education initiatives, funds security research, and supports open-source projects that benefit the wider digital ecosystem. This collaborative model fosters trust and positions Google as a leader in responsible AI deployment [10].

The contrast with organizations that struggle with overfitting is clear. Where others have suffered costly breaches due to narrow, outdated models, Google's investment in adaptive, transparent, and community-oriented security has paid dividends in resilience and public trust.

4.2.4. Analysis: Technical Capacity and Security Equity

The Google case illustrates how technical capacity, organizational resources, and a culture of transparency can dramatically reduce the risks of overfitting in AI for cybersecurity. Unlike environments where limited data, budget constraints, or compliance pressures lead to brittle, overfitted models, Google's approach demonstrates that robust, generalizable AI is achievable when organizations prioritize continual learning and community engagement.

This case also highlights a form of "security privilege." Organizations with ample resources and technical expertise can afford to implement best practices that remain out of reach for many others. The result is a digital landscape where some entities enjoy advanced protection and rapid response, while others remain vulnerable to the pitfalls of overfitting and evolving threats [11].

The evidence shows that effective overfitting mitigation is not just a technical challenge, but a question of organizational will and capacity. Google's model offers a blueprint for others, but also underscores the need for broader access to resources, data, and expertise across the cybersecurity field.

5. Findings: Systematic Patterns and Impacts of Overfitting in Cybersecurity

5.1. Patterns of Overfitting and Organizational Vulnerability

Analysis of real-world incidents and technical reports reveals clear, recurring patterns in how overfitting shapes the effectiveness of AI-driven cybersecurity systems. Organizations that rely on static datasets and infrequent model updates consistently experience higher rates of undetected attacks, false positives, and operational disruptions [1,2]. This pattern is not simply the result of technical oversight or resource constraints. Instead, it reflects a broader industry tendency to prioritize rapid deployment and compliance benchmarks over genuine model generalization and adaptability [3].

Just as in other sectors where infrastructure decisions reinforce social inequities, the distribution of advanced cybersecurity protections often follows organizational lines. Well-resourced companies with dedicated security teams and access to diverse, up-to-date data are able to deploy adaptive models that generalize well to new threats. In contrast, smaller organizations and public sector entities, such as hospitals and local governments, are more likely to use off-the-shelf solutions that overfit to outdated or narrow datasets [4,5]. This results in a digital divide where some entities enjoy robust protection, while others remain vulnerable to evolving cyber threats.

The speed and manner in which AI security solutions are deployed also reflect these disparities. Large technology firms invest in continual retraining, adversarial testing, and ensemble methods, while less-resourced organizations often lack the capacity for ongoing model maintenance. The result is a landscape where overfitting-related vulnerabilities persist longest in environments with the least ability to respond or recover from attacks [6].

5.2. Operational and Social Consequences

The operational impacts of overfitting in cybersecurity AI are substantial. Overfitted models frequently miss new attack techniques, such as zero-day exploits or novel malware variants, leading to costly breaches and service disruptions [7]. At the same time, these models may generate excessive false positives, overwhelming analysts and diverting attention from genuine threats [8]. The cumulative effect is a reduction in both the efficiency and trustworthiness of automated security systems.

These technical failures have broader social and economic implications. For example, public health organizations affected by ransomware attacks have faced not only financial losses but also disruptions to patient care and public services [9]. In the private sector, small businesses and community organizations may lack the expertise or resources to recover from breaches, resulting in lost income, reputational damage, and even closure.

Recent reporting has highlighted how these disparities in cybersecurity protection can mirror and reinforce existing social inequalities. Just as the placement of data centers and other digital infrastructure has been shown to disproportionately burden low-income and minority communities with environmental and economic costs [10,11], the uneven adoption of robust AI security measures can leave the most vulnerable organizations exposed to the greatest risks.

5.3. Economic and Workforce Impacts

The economic impact of overfitting in cybersecurity AI is twofold. First, organizations that suffer breaches due to undetected attacks face direct costs in the form of ransom payments, remediation expenses, and lost productivity [12]. Second, the broader economy absorbs indirect costs through increased insurance premiums, regulatory penalties, and diminished consumer trust in digital services.

Workforce impacts are also significant. As AI-driven security tools become more widespread, there is a growing demand for professionals who can monitor, retrain, and validate these systems. However, the technical expertise required is often concentrated in large firms and elite research institutions, leaving smaller organizations at a disadvantage. This concentration of expertise and resources can exacerbate existing disparities in cybersecurity outcomes, mirroring patterns seen in other technology-driven fields [13,14].

5.4. The Digital Security Divide

The cumulative effect of these patterns is the emergence of a digital security divide. Organizations with the resources and expertise to manage overfitting enjoy greater resilience and protection, while those without remain exposed to the evolving threat landscape. This divide is not accidental - it is the result of industry practices, resource allocation, and policy decisions that favor certain sectors and communities over others.

Recent surveys show that while most Americans support the development of AI-driven security and infrastructure, few are willing to accept the risks and costs in their own communities [2]. This “not in my backyard” attitude extends to digital security, where the benefits of AI are often concentrated in well-funded organizations, while the risks are offloaded onto those with the least capacity to manage them.

The findings underscore the need for more equitable access to robust, adaptive AI security solutions, as well as policies that address the root causes of overfitting and its unequal impacts across the digital landscape.

6. Discussion: Breaking the Cycle of Overfitting in Cybersecurity

6.1. How Overfitting Reinforces Old Vulnerabilities Through New Technology

The persistent problem of overfitting in AI-driven cybersecurity systems shows how new technology can end up repeating old mistakes. While AI is often promoted as a solution to modern threats, its deployment sometimes deepens existing divides between organizations with resources and those without. When security vendors and large enterprises focus on compliance checklists and rapid deployment, they often overlook the need for continual learning and adaptation. This results in models that are technically advanced but practically brittle systems that work well in the lab yet fail to protect against the ever-changing tactics of real-world attackers [1,2].

Just as digital redlining exposes how infrastructure decisions follow historical patterns of exclusion, the uneven distribution of robust AI security mirrors broader patterns of technological privilege. Well-funded organizations in tech hubs enjoy the benefits of adaptive, continually retrained models, while public institutions and small businesses are left with outdated, overfitted solutions. The result is a digital landscape where the risks and costs of cyberattacks are disproportionately borne by those with the least ability to recover [3,4].

This cycle is perpetuated by the language of innovation. Vendors promise “next-generation” AI security, but rarely address the underlying challenges of data diversity, adversarial adaptation, and ongoing maintenance. When breaches occur, blame is often placed on user error or “sophisticated attackers,” rather than on the structural issues that leave some organizations perpetually exposed. This framing allows the industry to celebrate progress while sidestepping responsibility for persistent vulnerabilities [5].

6.2. Building Real Security

Breaking the cycle of overfitting and its unequal impacts requires more than technical fixes. It demands a shift in how organizations, vendors, and policymakers approach AI security. First, there must be a commitment to transparency and accountability in how models are trained, validated, and updated. Security tools should be evaluated not just on benchmark performance, but on their ability to adapt to new threats and protect the most vulnerable users [6].

Second, the field needs broader coalitions that bring together technical experts, policymakers, and affected communities. Just as environmental justice movements have united diverse groups to challenge harmful infrastructure, digital security justice will require collaboration across sectors. This means sharing threat intelligence, investing in workforce development, and supporting open-source tools that are accessible to organizations of all sizes [7].

Finally, policy interventions may be necessary to ensure that robust, adaptive AI security is not a privilege reserved for the few. This could include funding for public sector cybersecurity, requirements for regular model retraining, and incentives for vendors to serve under-resourced organizations. By centering equity and resilience in AI security policy, the field can move beyond patchwork solutions and build a digital environment where everyone is protected [8,9].

7. Policy Recommendations: Moving Towards Robust AI Security

7.1. Federal and Regulatory Requirements

To address the risks of overfitting and ensure that AI-driven cybersecurity benefits all organizations, federal agencies should require comprehensive risk and impact assessments before approving or funding major AI security deployments. These assessments must go beyond technical benchmarks, explicitly evaluating how proposed models will affect the most vulnerable organizations - including public institutions, small businesses, and critical infrastructure with limited cybersecurity resources.

Regulatory frameworks should mandate that vendors and developers demonstrate their models’ ability to generalize beyond narrow training data, including transparent reporting on validation procedures, adversarial testing, and ongoing model retraining. Just as recent executive orders have called for environmental reviews and public transparency in AI infrastructure projects, similar standards should apply to the deployment of high-impact AI security systems. This includes requiring detailed documentation of model performance across diverse environments and public disclosure of known limitations.

Federal procurement policies can play a powerful role. The government should prioritize contracts with vendors who commit to regular model updates, independent audits for overfitting and bias, and clear communication of risks in accessible language. Agencies like the Environmental Protection Agency have begun to require ongoing air-quality monitoring and public reporting for AI infrastructure; cybersecurity regulators should require analogous transparency for model performance and risk.

7.2. Community Oversight and Capacity Building

Communities and organizations most affected by cyber threats deserve a real voice in how AI security tools are developed and deployed. This means establishing oversight boards or advisory councils that include representatives from under-resourced sectors, public interest groups, and technical experts. These bodies should have the authority to review proposed deployments, set conditions for approval, and demand remediation if models fail to perform as promised.

Funding should be made available for independent technical assistance, enabling smaller organizations and local governments to evaluate AI security products and participate meaningfully in oversight. Federal and state programs can provide grants or low-interest loans to support workforce development, model validation, and ongoing system maintenance - helping close the gap between well-resourced tech hubs and vulnerable sectors.

7.3. Transparency, Accountability, and Environmental Integration

Transparency must become a core requirement. Vendors and developers should be required to publish plain-language summaries of their models' capabilities, limitations, and data requirements. Public education campaigns, modeled after those used to explain the energy and water demands of AI infrastructure, can help organizations and communities understand the trade-offs and risks of adopting AI security tools.

Environmental and social impact assessments should be integrated into the risk management frameworks for high-impact AI systems, as proposed in recent policy discussions in both the US and EU. This means not only tracking technical performance but also considering the broader implications for equity, resilience, and community well-being

7.4. Incentivizing Innovation and Fairness

Finally, policy should incentivize the development and adoption of AI security solutions that are robust, adaptable, and accessible. This could include tax incentives for vendors who open-source their models, share threat intelligence, or partner with public sector organizations to pilot new approaches. Regulatory agencies should set minimum standards for model generalization and adversarial robustness, and require regular third-party evaluations to ensure compliance. By centering equity, transparency, and continuous improvement in AI security policy, we can break the cycle of overfitting and build a digital landscape where all communities are protected - not just those with the most resources.

8. Conclusion

8.1. Key Contributions

This research documents how overfitting in AI-driven cybersecurity systems creates systematic patterns of vulnerability and exclusion across the digital landscape. By examining real-world incidents, technical literature, and organizational practices, we show that overfitting is not simply a technical flaw but a reflection of deeper industry tendencies - prioritizing speed, compliance, and cost over genuine resilience and equity. Our analysis reveals that organizations with limited resources, such as public institutions and small businesses, are consistently left with brittle, overfitted models, while well-funded technology firms enjoy adaptive, robust defenses.

The concept of a digital security divide provides a powerful framework for understanding how technological progress can reinforce existing inequalities rather than closing them. While much of the current AI ethics debate focuses on algorithmic bias within software, our approach exposes how the deployment and maintenance of AI security tools can become a mechanism for concentrating digital risk in already vulnerable sectors. This perspective broadens the conversation about AI fairness, highlighting the need for structural solutions that address both the technical and social dimensions of cybersecurity.

Through comparative case studies and analysis of industry practices, we provide clear evidence that disparities in cybersecurity protection cannot be explained by technical needs or market forces alone. Instead, these patterns reflect deliberate industry choices that maximize efficiency and profit while externalizing risk onto those with the least capacity to resist or recover. This insight gives policymakers, advocates, and affected organizations concrete evidence to challenge the status quo and demand more equitable solutions.

Future Research Directions

The next frontier for research lies in centering the knowledge and experiences of those most affected by cybersecurity failures. Organizations on the front lines - public hospitals, schools, local governments, and small businesses - have unique insights into the real-world impacts of overfitting and digital exclusion. Building partnerships that prioritize these voices can lead to more accurate assessments of AI security risks and more effective, community-driven solutions.

Longitudinal studies that track the operational, economic, and social outcomes of AI security deployments across different sectors are urgently needed. Current research often relies on benchmark datasets and technical metrics, but lacks investigation of how overfitting shapes actual incident response, recovery, and trust in digital systems. Community-based participatory research methods could help fill this gap, ensuring that future policy and technical interventions are grounded in lived experience.

It is also critical to examine how intersecting factors - such as organizational size, sector, workforce diversity, and geographic location - shape both exposure to cyber threats and the capacity to manage overfitting. Comparative studies across regions, industries, and international contexts could reveal how local policies and resources influence digital security outcomes, providing a roadmap for targeted interventions.

Compliance with ethical standards

Disclosure of conflict of interest

This author has no conflicts of interest to disclose

References

[1] Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutor. 2016;18(2):1153-76.

[2] Sommer R, Paxson V. Outside the closed world: On using machine learning for network intrusion detection. In: 2010 IEEE Symposium on Security and Privacy. IEEE; 2010. p. 305-16.

[3] Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Syst Appl. 2014;41(4):1690-700.

[4] Saxe J, Berlin K. Deep neural network based malware detection using two dimensional binary program features. In: 2015 10th International Conference on Malicious and Unwanted Software (MALWARE). IEEE; 2015. p. 11-20.

[5] Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press; 2016.

[6] Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.

[7] Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. In: 5th International Conference on Learning Representations (ICLR 2017). 2017. Available from: https://openreview.net/forum?id=Sy8gdB9xx

[8] Roshanaei V, Ghorbani AA, Amirkhani A. Machine learning for cyber threat intelligence: A review. Comput Secur. 2024;133:103464.

[9] Kim Y, Kim W, Kim H, Kim S. False positive reduction in intrusion detection using deep neural networks. IEEE Access. 2020;8:168335-45.

[10] Ring M, Wunderlich S, Scheuring D, Landes D, Hotho A. A survey of network-based intrusion detection data sets. Comput Secur. 2019;86:147-67.

[11] Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy. IEEE; 2017. p. 3-18.

[12] Yeom S, Giacomelli I, Fredrikson M, Jha S. Privacy risk in machine learning: Analyzing the connection to overfitting. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF). IEEE; 2018. p. 268-82.

[13] Papernot N, McDaniel P, Sinha A, Wellman M. Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814. 2016.

[14] Biggio B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognit. 2018;84:317-31.

[15] Huang L, Joseph AD, Nelson B, Rubinstein BIP, Tygar JD. Adversarial machine learning. In: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence. ACM; 2011. p. 43-58.

[16] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138-60.

[17] Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. Found Trends Mach Learn. 2021;14(1–2):1-210.

[18] Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural Comput. 1992;4(1):1-58.

[19] Xu D, Tian Y. A comprehensive survey of clustering algorithms. Ann Data Sci. 2015;2(2):165-93.

[20] Zhang J, Springenberg JT, Boedecker J, Burgard W. Deep reinforcement learning with successor features for navigation across similar environments. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE; 2017. p. 2371-8.

[21] Hastie T, Tibshirani J, Friedman J. The Elements of Statistical Learning. 2nd ed. New York: Springer; 2009.

[22] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6:60.

[23] Sagi O, Rokach L. Ensemble learning: A survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8(4):e1249.

[24] Dietterich TG. Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems. Springer; 2000. p. 1-15.

[25] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE; 2009. p. 1-6.

[26] Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: ICISSP 2018. p. 108-16.

[27] Greenberg A. The WannaCry ransomware attack has a link to suspected North Korean hackers. Wired. 2017 May 15.

[28] Martin A, Ghafur S, Kinross J, Hankin C, Darzi A. WannaCry - a year on. BMJ. 2018;361:k2381.

[29] Google Security Blog. Our approach to security and transparency. Google; 2023. Available from: https://security.googleblog.com/

[30] Google AI Blog. Sharing research and best practices. Google; 2022. Available from: https://ai.googleblog.com/

[31] Thomas K, Bursztein E, et al. Data breaches, phishing, or malware? Understanding the risks of stolen credentials. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM; 2017. p. 1421-34.

[32] Google Open Source. Our commitment to open security. Google; 2023. Available from: https://opensource.googleblog.com/

[33] Saxenian AL. Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Cambridge: Harvard University Press; 1996.

[34] Ristenpart T, Yilek S. When good algorithms go bad: Security lessons from machine learning. Commun ACM. 2020;63(7):24-6.

[35] Survey: 93% of Americans support AI data centers, but not locally. HostingJournalist. 2025 Feb 24. Available from: https://hostingjournalist.com/news/survey-93-of-americans-support-ai-data-centers-but-not-locally

[36] Black communities face more pollution due to demand for AI. Capital B News. 2025 Mar 10. Available from: http://capitalbnews.org/ai-data-centers-south-carolina-black-communities/

[37] How the data center boom could harm Black communities. Canary Media. 2025 Feb 28. Available from: https://www.canarymedia.com/articles/fossil-fuels/how-the-data-center-boom-could-harm-black-communities

[38] Environmental and community impacts of large data centers. Gradient. 2025. Available from: https://gradientcorp.com/trend_articles/impacts-of-large-data-centers/

[39] Bashir S, Olivetti E. Water and energy use in data centers: Environmental impacts and policy responses. J Environ Manage. 2025;320:115860.

[40] Why the winners of the trillion-dollar data center gold rush are overwhelmingly white men. Bisnow. 2024 Nov 24. Available from: https://www.bisnow.com/national/news/data-center/data-center-diversity-126950

[41] Biden Administration releases executive order on AI infrastructure. Inside Energy & Environment. 2025 Jan 18.

[42] Data centers and local environmental considerations. National League of Cities. 2025 May 23.

[43] How AI infrastructure could help form a sustainable future. World Economic Forum. 2025 Jun 30.

[44] Summary of artificial intelligence 2025 legislation. National Conference of State Legislatures. 2025 Apr 24.

[45] AI, climate, and regulation: From data centers to the AI Act. arXiv preprint arXiv:2410.06681. 2024 Oct 8.

[46] Executive Order on Advancing United States Leadership in Artificial Intelligence Infrastructure. White House. 2025 Jan 14.

[47] The US must balance climate justice challenges in the era of artificial intelligence. Brookings. 2024 Feb 20.