An Anti-Sherif Artificial Intelligence-Driven Cybersecurity Audit Model: Beyond Artificial Intelligence Adoption in Cybersecurity Auditing

Ndaedzo Rananga * and H.S. Venter

Posted Date: 22 January 2026

doi: 10.20944/preprints202601.1703.v1

Keywords: artificial intelligence (AI); cybersecurity auditing; AI governance; human-in-the-loop; risk-based auditing; audit automation; cybersecurity controls; adaptive assurance

 doi:10.20944/preprints202601.1703.v1

Article

An Anti-Sherif Artificial Intelligence-Driven Cybersecurity Audit Model: Beyond Artificial Intelligence Adoption in Cybersecurity Auditing

Ndaedzo Rananga [1,*] and H.S. Venter [2]

[1] University of Pretoria, Pretoria 0002, RSA

[2] University of Pretoria, Pretoria, 0002, RSA

* Correspondence: u11329892@tuks.co.za

Abstract

The increasing adoption of artificial intelligence (AI) in cybersecurity has introduced new opportunities to enhance detection, response, and automation capabilities; however, applying AI within cybersecurity auditing remains constrained by traditional compliance-oriented approaches that rely profoundly on binary, checklist-based evaluations. Such approaches often reinforce a policing or "sheriff-style" perception of auditing, emphasizing enforcement rather than enablement, risk insight, and organizational improvement. This study proposes an Anti-Sherif AI-driven cybersecurity audit model that integrates AI-based analytics with human expert judgment to support a more adaptive, risk-informed auditing process. Grounded in design science research, the model combines conventional binary compliance checks with AI-derived intelligence and governance-based maturity assessments to evaluate cybersecurity controls across technical, operational, and organizational dimensions. The approach aligns with established standards and frameworks, including International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC) 27001, the National Institute of Standards and Technology (NIST), and the Center for Internet Security (CIS) benchmarks, while extending their application beyond static compliance. A fictional case study is used to demonstrate the model's applicability and to illustrate how hybrid scoring can reveal residual risk not captured by conventional audits. The results indicate that combining AI-driven insights with structured human judgment enhances audit depth, interpretability, and business relevance. The proposed model provides a foundation for evolving cybersecurity auditing from periodic compliance assessments toward continuous, intelligence-supported assurance.

Keywords: artificial intelligence (AI); cybersecurity auditing; AI governance; human-in-the-loop; risk-based auditing; audit automation; cybersecurity controls; adaptive assurance

1. Introduction

As articulated by Ndaedzo and Venter [1], among several factors directly influencing the exponential adoption of modern technologies, mobility remains one of the most significant. The emergence of high-speed technologies, such as 5G, and the proliferation of Internet of Things (IoT) devices have inevitably intensified society's reliance on and appreciation for seamless digital communication [2]. Network configurations are now designed to enable users to access computing resources anytime, anywhere, and from almost any device. This ease of access to networking resources has contributed directly to the accelerated adoption of modern technological advancements by end users and organizations.

Concurrently, the concept of artificial intelligence (AI) has moved beyond theoretical exploration and is becoming a central focus of the modern information and communication technology (ICT) landscape [3]. As a result of technologies such as the IoT, mobile cloud computing, and ubiquitous

 doi:10.20944/preprints202601.1703.v1

computing, the contemporary ICT environment is increasingly saturated with interconnected devices, influencing activities that range from everyday routines to professional work practices. While these advancements warrant recognition, they are inevitably associated with increasingly complex and sophisticated cyber threats. Growing concern persists regarding both the nature of modern cyber threats and their frequency. Notably, no single cybersecurity capability can be considered a universal solution to the increasingly complex and evolving threat landscape. As demonstrated by Yaker et al. [4], certain modern technologies, such as 5G, require novel cybersecurity approaches to mitigate the threats associated with their adoption.

While organizations and researchers continue to enhance cybersecurity controls through a range of innovations and applications, the nature of cybersecurity necessitates that such controls undergo continuous improvement and evaluation to ensure their effectiveness and adequacy. To facilitate ongoing evaluation of cybersecurity controls, organizations commonly adopt the practice of cybersecurity auditing. Cybersecurity auditing is closely related to traditional forms of auditing, including information systems (IS) auditing and financial auditing. It is intended to test the resilience of cybersecurity controls and to assess whether they comply with organizationally defined policies or internationally recognized standards, such as ISO, CIS, and NIST [5]; however, with the rise of increasingly sophisticated and evolving cyber threats, conventional cybersecurity audit approaches are becoming less effective in identifying and mitigating the actual risks organizations encounter. Traditionally, cybersecurity audits rely profoundly on binary compliance and noncompliance evaluations. To ensure that cybersecurity auditing remains relevant and effective, there is a need to rejuvenate conventional auditing methodologies so they respond more effectively to modern security threats, as traditional audit approaches are increasingly insufficient.

Of particular concern is the prevailing perception that cybersecurity auditing should be approached as policing, referred to herein as the "sheriff-style" approach. This methodology places more significant emphasis on enforcing compliance with predefined standards or policies through checklists and periodic assessments. In practice, these methods tend to prioritize control evaluation, with limited emphasis on human judgment and contextual insight that could support organizations in continuously improving their cybersecurity posture. As a result, such approaches often fall short, and stakeholders may intentionally mislead audit processes by withholding critical information to avoid unfavorable findings.

In the current AI-driven landscape, with concepts, such as augmented AI [6], there is an increasing need to embed AI within the cybersecurity auditing domain; however, organizations may remain hesitant to fully engage in AI-driven audit processes owing to mistrust in AI technologies, which are not guaranteed to produce correct outcomes or to account for human judgment adequately. The problem motivating the present study is, therefore, twofold. First, cybersecurity auditing is frequently perceived as a "sheriff-style" activity focused primarily on enforcement rather than enablement. Second, adopting AI in cybersecurity auditing often overlooks the critical role of human judgment and expert insight. This oversight risks undermining the effectiveness and integrity of the audit process and may cause more harm than benefit within the cybersecurity auditing landscape.

To approach this problem, the remainder of the paper is organized as follows: Section 2 presents background concepts relevant to the study. Section 3 outlines the research methodology adopted. Section 4 introduces the study's main contribution by outlining the proposed model. Section 5 discusses study limitations and future research directions. Section 6 concludes the paper, followed by a disclaimer regarding using certain resources in Section 7.

2. Background

As emphasized by Craigen et al. [1], much of the ICT terminology established in the literature is frequently subjective and occasionally lacks sufficient precision; therefore, it is necessary to provide explicit definitions for key terms used throughout this study, as outlined below.

 doi:10.20944/preprints202601.1703.v1

2.1. Cybersecurity

Diverse authors define cybersecurity concepts in varying ways; nevertheless, these descriptions generally align with established principles. Corporations likewise define and classify assets in distinct ways, depending on criteria specific to their business needs. Cybersecurity focuses on protecting digital assets, including data, information, systems, and services, from potential harm arising from successful cyber incidents. As cybercriminals become more innovative and introduce increasingly complex and sophisticated threats, such as advanced persistent threats (APTs) difficult to detect, there is an increasing need to strengthen cybersecurity controls [8]. Cybersecurity, therefore, involves systematic assessment and testing of an organization's internal controls, including policies, configurations, and procedures, to determine the resilience of these controls against potential cyber threats [3].

Cooke [9] emphasizes that cybersecurity broadly encompasses policies, technologies, and personnel that protect corporate network infrastructure from potential cyber threats, detect such threats, and remediate them within a reasonable recovery time objective (RTO) [10]. Diverse authors define cybersecurity in different ways, depending on their areas of work and study objectives; nevertheless, these definitions consistently focus on ensuring the confidentiality, integrity, and availability (CIA) of valuable digital assets. According to the NIST, cybersecurity can be defined as the:

> '*Prevention of damage to, protection of, and restoration of computers, electronic communications systems, electronic communications services, wire communication, and electronic communication, including information contained therein, to ensure its availability, integrity, authentication, confidentiality, and nonrepudiation*' [2].

Craigen et al. [1] reviewed cybersecurity definitions to formulate a definition that summarizes those from a multidisciplinary group. They define cybersecurity as:

> '*The organization and collection of resources, processes, and structures used to protect cyberspace and cyberspace-enabled systems from occurrences that misalign de jure from de facto property rights*' [1].

Cybersecurity auditing shares a symbiotic relationship with other traditional forms of auditing, including performance audits, financial audits, and IS audits. Among these, the IS audit bears the closest resemblance and relevance to cybersecurity auditing. For this reason, before examining the background and context of cybersecurity auditing, it is necessary to establish a clear understanding of IS auditing, as presented in the subsequent section.

2.2. Information System Audit

According to the Information Systems Audit and Control Association (ISACA), the term audit refers to the formal inspection and verification used to determine whether a standard or set of guidelines is followed, whether records are accurate, and whether efficiency and effectiveness targets are met [12]. Auditors conduct auditing to assess predefined controls or to evaluate controls against defined risks [13]. Auditing relies on several critical components: it must remain process-oriented, independent, and objective to determine the extent to which internal controls counter environmental risks [14].

In financial auditing, auditors aim to ensure the integrity of financial transactions while preventing fraudulent activities or human error that can result in material irregularities (MI). Conversely, the IS audit assesses the integrity of systems, storage, processing, and transmission of valuable information to ensure that stored information remains intact and trustworthy. In its simplest form, IS auditing can be defined as a formal, independent, and objective examination of an organization's IT infrastructure to determine whether the activities (for example, procedures and controls) involved in collecting, processing, storing, distributing, and using information comply with guidelines, safeguard assets, maintain data integrity, and operate effectively and efficiently to achieve the organization's business objectives [12].

 doi:10.20944/preprints202601.1703.v1

Cybersecurity vulnerability assessment and cybersecurity auditing remain closely related disciplines, yet they differ fundamentally in scope, methodology, and purpose. Both contribute to an organization's overall cybersecurity posture; however, their objectives serve distinct functions: one focuses on technical assessment, while the other focuses on governance. For the purposes of this study, this comparison remains of limited relevance and is not expanded further. The subsequent section, therefore, presents the concept of cybersecurity auditing, which supports the proposed model.

2.3. Cybersecurity Auditing

The ISACA introduced the concept of cybersecurity auditing at an elevated level through the official Certified Information Systems Auditor (CISA) manual published in 2015, which serves as a central authority in IS auditing. The CISA Review Manual, 26th Edition (2022), subsequently integrated the concept of cybersecurity auditing more comprehensively into auditing practice. This progression indicates that cybersecurity auditing, in its entirety, remains a relatively new concept that has only recently acquired traction within the auditing landscape.

Cybersecurity auditing constitutes a specialized process for assessing IT infrastructure against an organization's security policies, controls, governance arrangements, and compliance with defined standards, whether internal or international. The distinction between conventional vulnerability assessment and cybersecurity auditing remains subtle, with differences primarily evident in methodology, objectives, scope, and the anticipated outcomes of the process. According to Al-Matari [15], contemporary approaches to cybersecurity auditing require extensive examination of available technologies, methodologies, and processes. One notable development in efforts to strengthen cybersecurity capabilities is the adoption of AI. The subsection presents selected applications of AI in cybersecurity.

2.4. Artificial Intelligence in Cybersecurity

Cybersecurity researchers increasingly apply AI to strengthen cybersecurity controls and counter the evolving, complex, and sophisticated nature of cyber threats. Researchers also recognize AI as a potentially powerful tool for confronting selected cybersecurity challenges [16]. Common applications of AI in cybersecurity include real-time detection, automated incident response, predictive analytics, and vulnerability management [16]. Despite the increasing adoption of AI within the cybersecurity fraternity, several drawbacks remain unresolved. These drawbacks include heterogeneous data sources, challenges associated with explainable AI, the lack of a threat intelligence platform, and limited real-time data, as emphasized by Kaur et al. [3].

The preceding sections introduced the background and key terminologies used throughout the study. Before presenting the proposed Anti-Sheriff model, the next section presents the research methodology adopted in the present study. This ensures that the model development follows a systematic approach and can be validated through scientific and empirical assessments, thereby supporting credible and replicable outcomes.

3. Methodology

The study employed the design science research (DSR) methodology. The DSR methodology represents a widely used scientific approach supporting the development and evaluation of innovative artifacts, including models, methods, and frameworks. As succinctly explained by Vom Brocke et al. [17], DSR enhances technology, improves processes, and advances scientific knowledge through the creation of innovative artifacts that solve problems and improve day-to-day operations.

As illustrated in Figure 1, the DSR methodology comprises six research steps: problem identification and motivation, solution objectives definition, design and development, demonstration, evaluation, and communication. As emphasized by Venable et al. [18], these steps support research paradigms that produce innovative solutions to practical problems through

 doi:10.20944/preprints202601.1703.v1

developing scientifically sound artifacts, such as models, methods, and tools. In the present study, the authors apply the DSR methodology to establish the foundation for the proposed Anti-Sherif Model. As presented in the introductory section, the study examines the persistent perception of cybersecurity auditing as a tick-box or binary compliance exercise, particularly in the modern technological era, including AI.

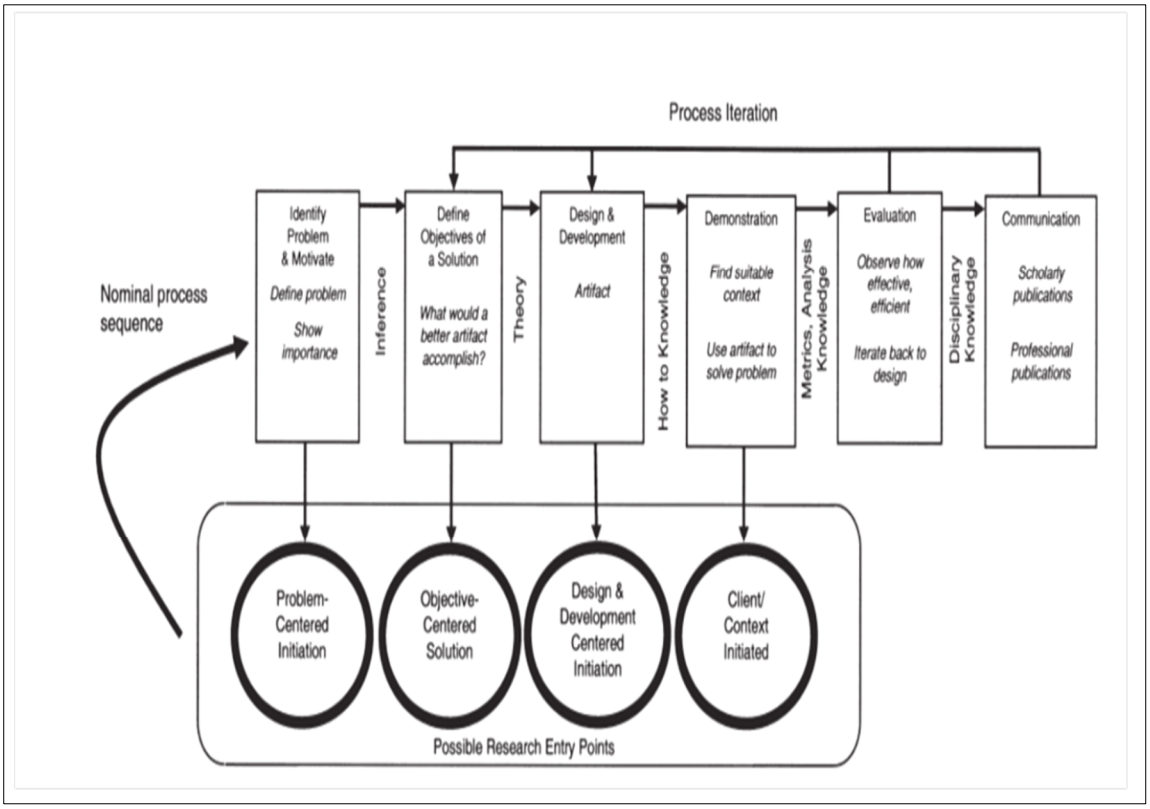


Figure 1. The design science research (DSR) methodology process model applied in the present study. (Adapted from Vom Brocke et al [3]).

The adoption of DSR grounds the Anti-Sherif Model in real-world auditing challenges, including the need for empirical, AI-driven solutions that extend beyond traditional compliance checklists. This approach also safeguards the scientific rigor of the model by integrating established auditing standards, cybersecurity frameworks, and empirical validation techniques to support the proposed solution. Researchers consistently emphasize the importance of aligning proposed solutions with widely recognized standards, as Sulistyowati et al. [5] also emphasize.

The previous section presents the iterative process of identifying problems through environmental analysis. Before examining the remainder of the study, this section provides an overview of how the DSR processes are applied in the present study. Table 1 summarizes the methodology for the proposed model, using the design science research methodology (DSRM), specifically the Anti-Sherif Cybersecurity Audit Model.

 doi:10.20944/preprints202601.1703.v1

Table 1. Methodology summary using the design science research methodology (DSRM) for the Anti-Sherif cybersecurity audit model.

DRS model steps (Adopted from [3])	Description (What the step means) [3,4]	Application in the present study	Output description
1. Problem Identification & Motivation	This step provides the foundational motivation for the study by identifying the problem. This step presents the gaps and the reason a need exists to resolve those gaps through a proposed solution.	In the present study, the authors have observed an over-reliance on the process of binary checks in the cybersecurity audit landscape. More often, this exercise i.e., binary control check, falls short as it does not address risk at a business objective level. As a result, the effectiveness and relevancy of the cybersecurity audit become questionable to the stakeholders and can be perceived as a policing exercise rather than a risk-oriented process.	Clearly defined cybersecurity audit problem
2. Define Objectives of a Solution	This step narrows the study objective toward specific end goals and the new artifacts the study aims to present to the body of knowledge.	The proposed model introduces a multi-layered cybersecurity audit approach that surpasses a traditional means of binary check. These layers include components, such as AI-driven insights, human judgment, government maturity scoring, risk modeling, and audit outcomes aligned with business objectives.	Model functionalities and components

 doi:10.20944/preprints202601.1703.v1

3.	Design & Development	This is the first main contribution of the study, in which the proposed model is presented through a conceptual model and/or a systematic framework.	The proposed model is designed through components meant to strike a balance between binary check, AI integration, and human or expert judgment.	Functional Anti-Sherif Model
4.	Demonstration	The practical applicability of a solution should be applied through a scientific approach, such as experiments. This step demonstrates the applicability of the solution in practice.	Apply the model to the fictional Nany FinTrust MicroBank to generate governance and business insights.	Demonstrated model solving real cybersecurity audit challenges using fictional scenarios
5.	Evaluation	This step ensures that gaps in the proposed solution are identified and resolved. The step also guards against limitations of the proposed solution.	Quantitative evaluation (AI model accuracy, anomaly detection performance). Qualitative evaluation by cybersecurity experts and auditors. Refinements made based on feedback.	Model critical evaluation outcome aimed at confirming effectiveness and areas of improvement
6.	Communication	Finally, the output of the study is made available to the intended	Dissertation chapters and conference submissions are published.	Publishable research outputs and communicated findings

 doi:10.20944/preprints202601.1703.v1

society.

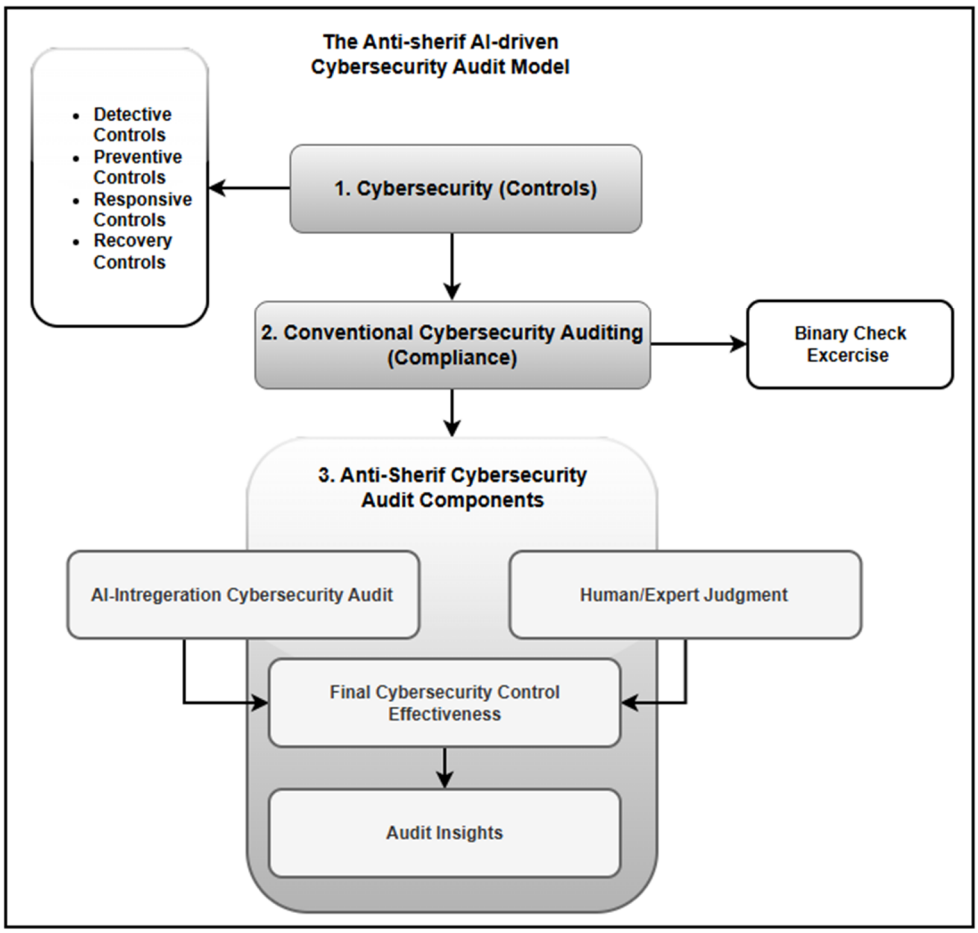
The preceding sections present the background and the details of the methodology employed in this study. The next section introduces the proposed model, beginning with a high-level representation and subsequently expanding it through a detailed representation.

4. Model Presentation

The proposed model serves as a foundational baseline for integrating AI into cybersecurity auditing, ensuring equal consideration of three key pillars: operational, compliance, and technical. As the name suggests, the model attempts to transition the prevailing perception of cybersecurity auditing from a policing or enforcement-driven approach to a more balanced, risk-informed, and adaptive methodology. The model also incorporates human societal dimensions, including ethics, human judgment, explainability, transparency, and public trust.

4.1. High-Level Anti-Sherif AI-Driven Cybersecurity Model

In the present study, the authors present three core components of the model: cybersecurity (technical security controls), conventional cybersecurity auditing (compliance), and the Anti-Sherif cybersecurity audit approach, which encompasses integrating AI and human judgment in cybersecurity audits, as graphically represented in Figure 2. As depicted in the figure, the proposed model represents a transition from conventional approaches to conducting cybersecurity binary compliance checks.



 doi:10.20944/preprints202601.1703.v1

Figure 2. Conceptual high-level model representation (Anti-Sherif AI-driven cybersecurity model).

Figure 2 presents a high-level representation of the model. The subsequent section presents each component of the model, beginning with cybersecurity controls.

4.1.1. Cybersecurity (Controls)

As depicted in Figure 2, the proposed model adopts cybersecurity technical controls designed to satisfy the core complementary cybersecurity components, commonly referred to as the security triad [19]: Confidentiality, Integrity, and Availability (CIA). Organizations implement these components through a range of mechanisms that protect the ICT environment.

Within the contemporary ICT landscape, characterized by increasingly complex and sophisticated cyber threats, organizations invest in advanced detection mechanisms that identify malicious activities in near real time [20], including AI-assisted intrusion detection systems (IDS), security information and event management (SIEM), user and entity behavior analytics (UEBA), and endpoint detection and response (EDR). In parallel, organizations strengthen preventive capabilities through mechanisms, such as operating system and database security hardening, network segmentation, multi-factor authentication (MFA), and AI-assisted controls that predict and block malicious activities in real time. Organizations also deploy security orchestration, automation, and response (SOAR) solutions to automate responses in near real time [21]. Advances in disaster recovery and backup planning further strengthen recovery capabilities.

Collectively, these mechanisms protect digital assets and align with the NIST Cybersecurity Framework (CSF) functions—protect, detect, respond, and recover—as depicted in Figure 2 [5]. These functions derive from the NIST CSF [22] and serve as essential building blocks for fluctuating from point-in-time defenses toward continuous, risk-based, and intelligence-driven cybersecurity auditing using modern approaches, including AI, within the proposed model.

To ensure that cybersecurity controls remain effective in supporting the CIA of information, organizations must perform regular cybersecurity auditing. Conventional approaches test the existence of these controls using binary compliance checks, as further discussed in the subsequent sections.

4.1.2. Conventional Cybersecurity Auditing (Binary Check)

As presented in Figure 2, the second component of the proposed model is cybersecurity auditing and compliance. As emphasized in the introductory section, the evolving nature of cybersecurity requires continuous scrutiny of the technical controls described in the preceding subsection to ensure that they remain effective and relevant in defending against modern cyber threats. Organizations also conduct cybersecurity audits to demonstrate compliance with requirements set by governing and regulatory bodies.

The proposed model evaluates the same components discussed earlier—detect, prevent, respond, and recover—through complementary assessments. Conventionally, organizations perform these assessments as binary compliance checks before applying a risk-based approach that ensures technical controls align with business needs. Compliance assessments against defined baselines and international standards verify whether implemented technical controls operate as intended in practice. In traditional cybersecurity auditing, binary checks place limited emphasis on control effectiveness and typically focus on confirming the existence of controls, such as SIEM and EDR, within the ICT environment.

Auditors often conduct this form of assessment as a policing-oriented exercise, referred to herein as the “Sheriff approach,” which reduces the audit to a tick-box activity. Common frameworks and standards, including the CIS Benchmarks, ISO/IEC 27001, and the NIST CSF, require auditors to verify the presence of controls but do not mandate in-depth measurement of how effectively those controls operate under dynamic threat conditions.

 doi:10.20944/preprints202601.1703.v1

Building on this binary approach, the proposed model, as expanded in the subsequent section, introduces more significant depth and accuracy to cybersecurity auditing. This approach not only verifies the existence of cybersecurity controls but also extends the audit to consider additional influential factors affecting control performance and organizational risk.

4.1.3. Anti-Sherif Cybersecurity Audit Components

As illustrated in Figure 2, the third component of the proposed model represents the Anti-Sherif cybersecurity audit components, which extend traditional binary assessment into a second analytical layer supported by AI-driven intelligence and human judgment. The first component, the binary check, provides a foundational assessment during the initial phase of the audit. Its objective is to determine whether control exists. The proposed Anti-Sherif Model strengthens this evaluation by introducing a more sophisticated, multidimensional assessment through integrating AI and human judgment.

Within this component, the model evaluates the performance of controls, such as SIEM and EDR systems, by analyzing telemetry completeness, anomaly patterns, log quality, and threat intelligence indicators. The human judgment component further incorporates governance context and maturity considerations, enabling auditors to assess whether processes are formally documented, consistently applied, and aligned with organizational and regulatory expectations.

Through this process, the model moves beyond confirming the existence of controls, such as SIEM or EDR, and evaluates their resilience by assessing detection capability, alert fidelity, and response readiness. A typical example involves the assessment of disaster recovery (DR). The model verifies not only the presence of backups but also whether the RTO and recovery point objective (RPO) meet business-defined requirements. RTO denotes the maximum tolerable downtime following a system failure or cyber incident, such as ransomware, whereas RPO represents the maximum acceptable amount of data loss measured from the last valid backup.

By integrating binary checks, AI-driven analytics, and human governance evaluation, the model converts raw technical control outputs into measurable, intelligence-based audit indicators. This approach enables more accurate, risk-informed, and business-aligned cybersecurity audit insights.

The preceding discussion provides a high-level representation of the proposed model. The subsequent section presents a detailed representation of the model and expands on its underlying logic, which follows the sequence binary check → AI layer → human judgment layer.

4.2 Detailed Model Representation

As emphasized by Kaur et al. [3], AI applications in cybersecurity aim to strengthen technical capabilities, including IDSs and intrusion prevention systems (IPSs). The concept of defense-in-depth plays a critical role in mitigating risks presented by complex cyber threats, such as APTs, ransomware, and zero-day exploits. Although technical cybersecurity controls remain essential, a layered defense strategy is increasingly necessary to ensure resilience against sophisticated attacks in the modern threat landscape.

4.2.1. Cybersecurity Controls (a Detailed Description)

In the present study, the proposed model uses multiple technical controls and combines them into a unified set, as indicated in Equation 1.

Let $T = \{t_1, t_2, \dots \dots, t_n\}$, $t_i \epsilon [0,1]$ (1)

Where T: a set of technical security controls and $t_i \epsilon [0,1]$: is the effectiveness score of each control (e.g., firewall, access control, IDS, IPS, antivirus, data loss prevention (DLP), and EDR, etc.). Drawing from the Matryoshka mathematical approach [5], it is further assumed that the presence of more technical controls decreases the likelihood of a successful cyberattack. Let $P_i \epsilon [0,1]$ denote the probability that control i fails to prevent or interrupt the attack (so, the $P_i = 1 - t_i$). Assuming independence and that each technical control enhances the model's capability and reduces

 doi:10.20944/preprints202601.1703.v1

the overall defined cybersecurity risk. The overall probability of attack success across the implemented controls can be defined as follows (Equation 2):

$$P_{success} = \prod_{i=1}^{n} P_i = \sum_{i=1}^{n} (1 - t_i) \quad (2)$$

Based on this assumption, the aggregated (baseline) technical control score T is defined as follows (Equation 3):

$$T_{avg}^{*} = \frac{1}{n} \sum_{i=1}^{n} t_i^{*} \quad (3)$$

The mathematical formulation of the technical controls simplifies the aggregation of multiple security controls into a unified framework aimed at reducing potential cybersecurity threats. These equations provide a foundational representation supporting the conceptualization of how individual controls interact and collectively contribute to risk reduction. The equations also support subsequent sections of the study, particularly the introduction of the AI-based scoring mechanism and the integration of human judgment into the evaluation process. This combined approach emphasizes the need to move beyond binary checks toward a dynamic, hybrid auditing model that balances automated analysis with expert interpretation.

4.2.2. Cybersecurity Auditing (Binary Check for Compliance)

Once the study defines cybersecurity controls, each control must be tested to determine its effectiveness in supporting organizational objectives and compliance requirements. Before introducing the integration of AI and human judgment, the authors present Table 2, which lists the technical-layer controls and specifies how each control is evaluated using a binary check (1 = met, 0 = not met). Table 2 also indicates the standards and frameworks informing each audit check.

Table 2. Cybersecurity audit checks mapped to cybersecurity technical controls (with binary check column).

Function(s)	**Cybersecurity audit check**	**Control example**	**Standard reference**	**Binary criteria (per control)** T_{avg}
Detect	Compliance	SIEM	**NIST CSF**: DE.AE-1 (Anomalous activity detected); DE.CM-1 (Monitoring for events) **CIS**: Control 8 – Audit Log Management **ISO/IEC 27001**: A.12.4 (Logging and monitoring)	Is logging/alerting enabled and retained as per compliance requirements? (1 = Yes, 0 = No)
Prevent	Effectiveness	Firewall	**NIST CSF**: PR.AC-5 (Network integrity maintained); PR.IP-1 (Baseline configuration) **CIS**: Control 4 –	Do firewall rules effectively reduce unauthorized access risks? (1 = Yes, 0 = No)

 doi:10.20944/preprints202601.1703.v1

			Secure Configuration of Enterprise Assets and Software **ISO/IEC 27001**: A.13.1 (Network security management)	
Respond	Resilience	Incident response (IR) plan	**NIST CSF**: RS.RP-1 (Incident response plan executed); RS.CO-1 (Response coordination) **CIS**: Control 17 – IR Management **ISO/IEC 27001**: A.16.1 (Information security incident management)	Can the IR Plan be executed and remain effective during outages? (1 = Yes, 0 = No)
Recover	Predictive	Backups/ DR Strategy	**NIST CSF**: RC.IM-1 (Recovery planning); RC.RP-1 (Recovery plan tested) **CIS**: Control 11 – Data Recovery **ISO/IEC 27001**: A.17.1 (Information security continuity); A.17.2 (Redundancies)	Is the RTO accurately forecasted and tested? (1 = Yes, 0 = No)

Almuhammadi and Alsaleh [24] indicate that combining multiple standards strengthens cybersecurity assessments, as individual standards focus on distinct risk areas and provide broader coverage when applied collectively. Table 1 further illustrates how the integration of AI and human judgment mitigates the limitations and potential bias inherent in binary checks and enables risk-based cybersecurity auditing. The audit applies defined criteria, including compliance, effectiveness, resilience, and predictive assessments.

Table 2 demonstrates that cybersecurity auditing can be conducted using a binary check approach. Under this approach, the auditor applies a checklist and records whether each control is present or absent, without applying expert judgment or aligning audit procedures with the particular

 doi:10.20944/preprints202601.1703.v1

business needs of the auditee (the audited entity). Although this method offers a structured means of assessing compliance, it provides limited contextual depth.

There is broad agreement within the cybersecurity fraternity that modern cyber threats increasingly challenge the relevance and adequacy of existing cybersecurity control evaluation practices, as also emphasized by Yang et al. [2]. Despite this consensus, a substantial number of organizations continue to place limited emphasis on enhanced or intelligence-based cybersecurity assessments. Instead, they rely on traditional checkbox-style exercises, referred to in this study as binary checks, which confirm only the existence of a control without evaluating its effectiveness, resilience, or alignment with evolving cyber risks. An additional influencing factor, particularly for small and medium-sized enterprises (SMEs), is the complexity associated with implementing publicly available standards and frameworks [24].

Recent literature repeatedly emphasizes the need to reconsider prevailing cybersecurity audit approaches. The primary challenge lies in integrating advanced capabilities, such as AI, into cybersecurity audits while preserving essential societal, organizational, and human considerations. Maintaining this balance ensures that audits remain technologically relevant and contextually grounded. The subsequent section demonstrates how the proposed model responds to this challenge by examining the limitations of binary assessment practices and explaining why auditees may remain vulnerable, particularly in the presence of emerging technologies, such as AI. These technologies introduce complex and dynamic risks that simple binary evaluations cannot adequately assess, necessitating more adaptive, risk-based, and intelligence-driven auditing approaches.

4.2.3. AI Integration and Expert Judgment in Cybersecurity Auditing

After presenting the binary check approach used in cybersecurity auditing, the proposed model expands through integrating AI and human judgment, enabling cybersecurity audits to adopt a risk-based approach rather than a binary evaluation. Although AI aims to simplify and automate processes, developing sound and explainable AI models require expert input in the form of human judgment, which remains essential [25].

Building on the foundational logic presented in Table 1, the proposed model incorporates AI and human judgment, as illustrated in Table 2. Within this extended framework, the model evaluates cybersecurity controls across four dimensions: compliance, effectiveness, resilience, and predictability. For example, the audit of a SIEM capability commences with a binary compliance check to confirm the existence and configuration of the control. The AI integration layer then assesses the SIEM's ability to detect and analyze potential cyber threats using intelligent mechanisms, such as anomaly or behavioral analysis, including capabilities enabled through SOAR. As illustrated in Table 2, the auditor subsequently applies human judgment to contextualize these outcomes and determine whether the controls mitigate technical risks and align organizational objectives and operational priorities. This layered auditing approach is demonstrated later through an experimental evaluation that balances automation-driven intelligence with expert human oversight.

$$\boldsymbol{T}^{*}_{\boldsymbol{avg}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{t}^{*}_{\boldsymbol{i}} \quad (4)$$

Equation 4 derives the average baseline score from the binary assessment of control availability, without accounting for external factors that influence the effectiveness of technical control. The model now introduces the AI-based score and human expert judgment, as follows:

$A_i \in [0,1]$ *and* $H_i \in [0,1]$

Where A_i Is the AI-based predicted effectiveness of the defined cybersecurity controls represented by i, derived from models trained on historical cyberthreats and previous audit findings, and/or from other data sources, such as the dark web and threat intelligence. Conversely, H_i Represents the expert auditor's judgment score for the defined cybersecurity control i, factoring in business needs and human experience. The final blended control can be defined as follows:

$$\boldsymbol{t}_{\boldsymbol{i}}^{(\boldsymbol{AI})} = \alpha\boldsymbol{t}_{\boldsymbol{i}}^{(\boldsymbol{0})} + (1-\alpha)A_i \quad (5)$$

 doi:10.20944/preprints202601.1703.v1

Where $\alpha \in [0,1]$ is a weight that balances a mere cybersecurity technical control with AI prediction. The next step for the proposed model is to introduce the human aspect, and this can be defined as follows:

$$t_i^* = \beta t_i^{AI} + (1 - \beta) H_i \quad (6)$$

Where $\beta \in [0,1]$, which balances cybersecurity controls, auditing, and human oversight.

All parameters of the proposed model are now defined. The human judgment (H_i) component requires further elaboration because its values are inherently subjective and cannot be quantified using a fixed base score, unlike binary control checks. To mitigate this limitation, the proposed model adopts the Capability Maturity Model (CMM) as the basis for establishing a standardized scoring baseline for human judgment. As demonstrated by Regulwar et al. [26], CMMs play a critical role in gauging the maturity of organizational processes in cybersecurity and engineering contexts. In the present study, the CMM framework enables the assignment of base scores reflecting the maturity level of each evaluated cybersecurity control by translating qualitative expert assessments into measurable values. Table 3 illustrates the distribution of these maturity-based scores by mapping CMM levels to their corresponding quantitative values within the proposed model. This approach supports the model's objective of balancing AI-driven analysis with human-centric cybersecurity auditing capabilities.

Table 3. Cybersecurity audit (AI integration and human judgment).

Function(s)	**Cybersecurity audit check**	**Control example**	**Binary Criteria (per control)** $t_i^{(0)}$	**AI integration (A_i)**	**Human judgment (H_i)**
Detect	Compliance	SIEM	Is logging/alerting enabled and retained as per compliance requirements? (1 = Yes, 0 = No)	AI validates log coverage & anomaly detection patterns	Auditor checks alignment with business needs & regulatory context
Prevent	Effectiveness	Firewall	Do firewall rules effectively reduce unauthorized access risks? (1 = Yes, 0 = No)	AI tests rules against known & emerging threat signatures (threat intel, ML-based traffic analysis)	The auditor evaluates the adequacy of rules against entity-specific risks
Respond	Resilience	IR Plan	Can the IR Plan be executed and remain effective	AI simulates IR scenarios, MTTR	Auditor verifies practicality,

 doi:10.20944/preprints202601.1703.v1

			during outages? (1 = Yes, 0 = No)	forecasting	governance fit, stakeholder readiness
Recover	Predictive	Backups/ DR Strategy	Is the RTO accurately forecasted and tested? (1 = Yes, 0 = No)	AI models downtime costs & predicts the likelihood of successful restores	Auditor validates business continuity alignment & real-life test results

As depicted in Table 3, the CMM defines maturity levels ranging from 1 to 5, with each level reflecting the degree of process maturity. To quantify human judgment in the auditing of cybersecurity controls, the authors apply the CMM and assign scores based on expert assessment. The base score applied in this study ranges from 0 to 1, ensuring alignment with the binary checklist defined earlier.

Table 3. Cybersecurity control evaluation reference according to the Capability Maturity Model (CMM) framework.

CMM level →	**Description**	**Interpretation in the cybersecurity audit context**	**Assigned base score (H_i)**
Level 1 Initial (Ad-hoc)	At this level, organizational processes are chaotic and disorganized, and the initiatives are conducted haphazardly.	Cybersecurity controls are not defined, highly dependent on manual measures.	0.2
Level 2 Repeatable / managed	Basic processes are established, however, at an entry level.	Basic cybersecurity controls exist; however, no enforcement processes are implemented.	0.4
Level 3 Defined	At this level process is standardized, documented, and integrated throughout the organizational structure.	Cybersecurity controls are implemented and enforced in a reasonable manner across the ICT environment.	0.6
Level 4 Quantitatively managed	Metrics are key performance indicators (KPIs) defined against	Cybersecurity controls are defined, and the effectiveness of each	0.8

 doi:10.20944/preprints202601.1703.v1

	the processes for qualitative and quantitative measurements of success factors.	control can easily be measured using a defined matrix.	
Level 5 Optimizing	Processes are grounded, and the organization focuses on continuous improvements.	Most cybersecurity controls are automated and advanced, such as the implementation of continuous control monitoring (CCM) and CCA.	1

$H_i \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. These defined base scores will be crucial for the empirical research avenues.

The preceding sections present the components of the proposed model. Before introducing the experimental setup, this section provides a concise summary of the overall model logic, as illustrated in Figure 3. Figure 3 depicts the complete audit process, beginning with the initial binary compliance check, progressing through the AI integration assessment, and culminating in the incorporation of human judgment. The final output represents the overall effectiveness of the cybersecurity audit process by accounting for influential factors, including corporate business needs and organizational objectives.

 doi:10.20944/preprints202601.1703.v1

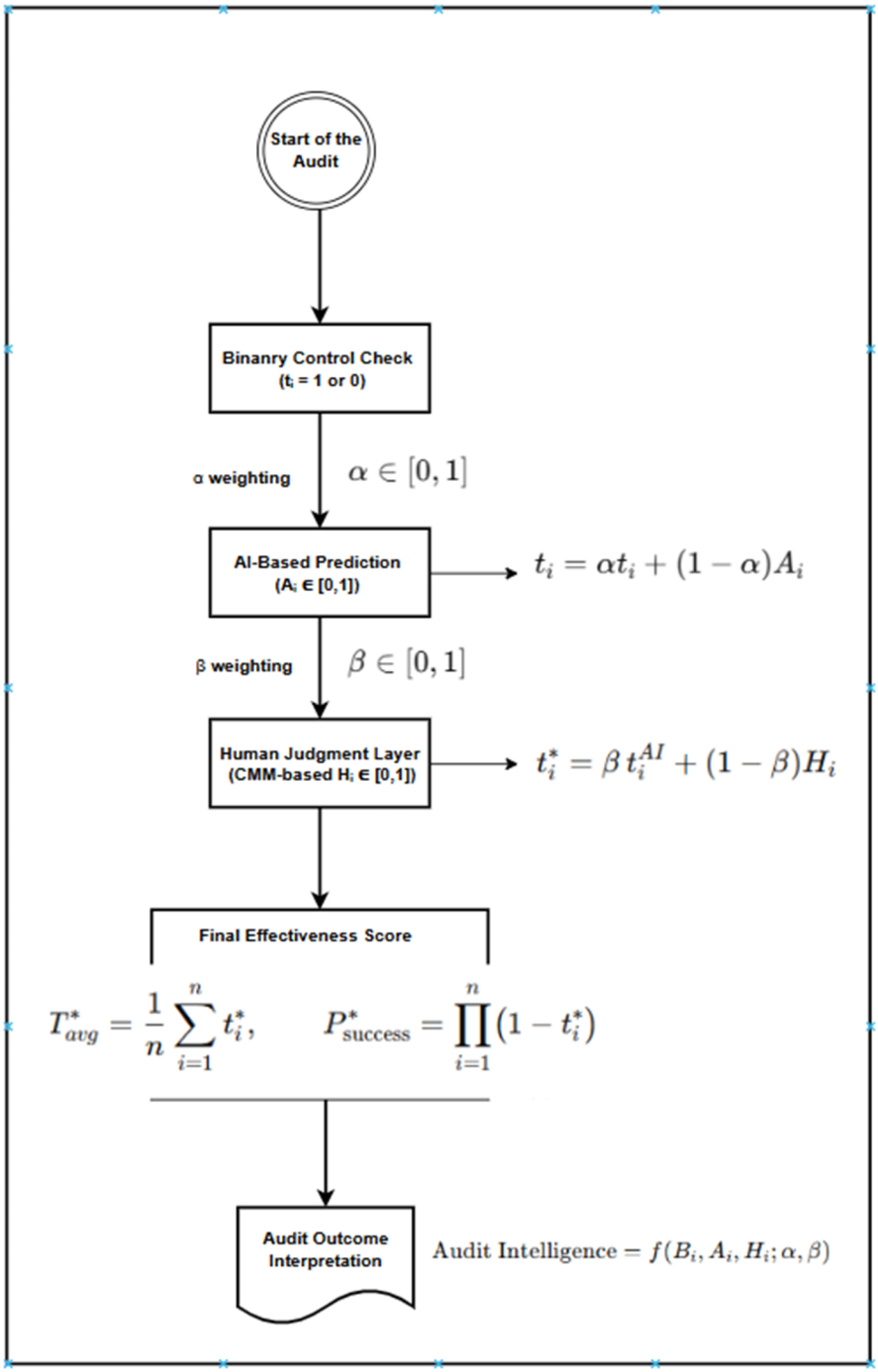


Figure 3. Model process flow representation.

The preceding sections present the detailed structure of the proposed model. As demonstrated by Rananga and Venter [1], experimental demonstrations play a critical role in establishing the applicability of a proposed solution through using fictional scenarios that closely reflect real-world conditions. The subsequent subsection applies the proposed model to such a scenario and provides an empirical analysis.

4.3. Empirical Research

To support clarity in the empirical demonstration of the proposed model, the study presents a fictional logical case scenario. In practice, such a scenario would correspond to a business case or an audit requirement defined within the audit scope, depending on the organization's risk appetite.

 doi:10.20944/preprints202601.1703.v1

4.3.1. Fictional Case Scenario

Nany FinTrust MicroBank (Pty) Ltd is a fictional SME specializing in micro-lending services. The institution employs approximately 200 staff members and operates under a hybrid working arrangement supporting remote and office-based work. Provided the nature of its services, the institution subscribes to several regulatory and data privacy standards, including the Protection of Personal Information Act (POPIA), the Payment Card Industry Data Security Standard (PCI DSS), the NIST CSF, and the CIS Benchmarks. These standards represent commonly adopted cybersecurity and data protection frameworks, as expressed by Sulistyowati et al. [5].

Ongoing technological advancement and the evolving cybersecurity threat landscape place increasing pressure on the institution to implement and enforce robust cybersecurity mechanisms supporting continuous auditing and control improvement. This requirement reflects the need for regular scrutiny of cybersecurity controls to identify weaknesses that could result in security breaches or system downtime.

Management and institutional stakeholders express concern regarding the current cybersecurity audit practices, which they perceive primarily as tick-box or binary compliance exercises that emphasize regulatory adherence over risk-driven auditing aligned with business objectives. As a result, decision-makers remain hesitant to endorse AI-based initiatives to enhance cybersecurity auditing, viewing such initiatives as potentially disproportionate investments with limited direct business value.

To mitigate this gap, the Head of Audit adopts the Anti-Sherif Model, an AI-driven cybersecurity auditing framework that aims to:

- strengthen business alignment in cybersecurity audits;
- improve accuracy and consistency in audit scoring; and
- retain the application of human judgment.

Invoking this fictional case scenario, the study explores empirical evaluation through experiments conducted using simulated data, as presented in the subsequent sections.

4.3.2. Experimental Demonstration

The study defines the variables used throughout the experiment through variable operationalization before presenting the experimental setup for validating the proposed model. Variable operationalization translates abstract research concepts into measurable indicators supporting empirical and quantitative observation [27], as summarized in Table 4. All symbols listed in Table 4 derive from the model process flow illustrated earlier in Figure 3.

Table 4. Experimental variable operationalization.

Construct	Symbol	Measurement Approach	Data Type
Cybersecurity Technical Control Binary Check	$t_i^{(0)}$	0 = not implemented, 1 = implemented	Nominal
AI integration	A_i	ML-based risk probability (normalized [0,1])	Predefined
Human Judgment Integration (Using CMM levels)	H_i	Derived from Table 3 maturity scale (0.2–1.0)	Predefined
Controlled weights	α and β	$\alpha = \beta = 0.6$	Predefined

 doi:10.20944/preprints202601.1703.v1

AI-influenced	t_i^{AI}	$\beta t_i^{AI} + (1 - \beta)H_i$	Continuous
Probability of attack success	$P_{success}^*$	$(1 - t_i^*)$	Continuous

In practical applications, the weights presented in **Table 4** can be computed using methods, such as a Delphi-based expert process informed by historical cybersecurity incident data. For the purposes of this demonstration, the study defines the weights using proportional influence. The weighting parameters remain bounded within the interval $[0.1, 1.0]$to regulate the relative contribution of each scoring dimension. In the present study, the authors assign the weights as $\alpha = \beta = 0.6$to support a balanced integration of AI-driven analytical insights and human expert judgment. This configuration preserves methodological neutrality and robustness by avoiding undue bias toward algorithmic assessment or subjective evaluation in computing the final control score.

The cybersecurity audit focus is narrowed to reflect the defined case scenario. The auditor establishes the technical boundaries and identifies the in-scope systems for evaluation. Audit exercises are inherently scope-defined, meaning that an audit cannot assess all controls simultaneously. To ensure meaningful outcomes, auditors commonly adopt risk-based approaches [28].

In the present study, the authors selected cybersecurity controls for testing. These controls are deliberately chosen because they constitute the minimum viable security foundation required to establish a defense-in-depth, or layered defense, posture across the four core cybersecurity functions—detect, prevent, respond, and recover—for SMEs, such as Nany FinTrust MicroBank (Pty) Ltd. The selection also reflects the inclusion of these controls across multiple authoritative industry frameworks, including the NIST CSF, CIS Critical Security Controls, and ISO/IEC 27001 [29], which consistently identify them as high-priority safeguards with a strong risk-reduction effect for minimizing exploitable attack surfaces. The cybersecurity controls audited for the experimental evaluation are listed in Table 5.

Table 5. Cybersecurity controls included in the audit scope.

Control ID	Audit Function (NIST CRF)	Description
C1	Prevent	The implementation of the MFA
C2	Detect	Implemented and functional EDR/Antivirus capabilities
C3	Prevent	Regular firewall rule review
C4	Prevent	Patch management processes are implemented
C5	Prevent	Admin privilege hardening processes
C6	Detect	Audit logging is enabled on critical applications
C7	Prevent	Network segmentation is implemented

 doi:10.20944/preprints202601.1703.v1

C8	Detect	Regular vulnerability scanning
C9	Detect	Database logging is enabled for all critical applications
C10	Recover	Backup testing process implemented

The authors apply the proposed model to the defined case scenario step by step using the controls listed and described in Table 5. The first guiding question establishes whether the controls in Table 5 are in place at Nany FinTrust MicroBank (Pty) Ltd. Table 6 records the binary check results.

Table 6. Cybersecurity Control Binary Check Results.

Symbol	**Audit result (s)** **Binary score** $t_i^{(0)}$
C1	1
C2	1
C3	1
C4	1
C5	1
C6	1
C7	1
C8	1
C9	0
C10	1

As depicted in Table 6, a conventional cybersecurity audit compliance check computes the final binary control score as the means of all control indicators, as expressed below:

$$T^{(0)} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{t_i^{(0)}} \quad (7)$$

Where:

- $t_i^{(0)} \in \{0,1\}$
- N=10=number of total cybersecurity audits.

Following this logic, the final binary score for the audited cybersecurity controls, as depicted in Table 6, can be computed as:

$$1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0 + 1 = 9$$

$$T^{(0)} = \frac{9}{10} = 0.9$$

This result indicates that 90% of the audited controls are present. Under a conventional cybersecurity compliance audit, this outcome classifies Nany FinTrust MicroBank (Pty) Ltd as largely compliant, indicating the implementation of a reasonable baseline of cybersecurity controls.

 doi:10.20944/preprints202601.1703.v1

The proposed model mitigates the limitations of conventional binary cybersecurity assessments by transitioning toward a risk-based evaluation framework. It reduces reliance on pass–fail indicators by incorporating complementary inputs, specifically human expert judgment and AI-derived intelligence sourced from open-source threat data. This integrated approach enables a more nuanced and context-aware assessment of cybersecurity risk, as demonstrated in the following steps.

Following the binary check, the auditor applies AI-driven intelligence (A_i) to determine the AI-based score derived from network scan results. After conducting a network scan of the target environment to identify common vulnerabilities and exposures (CVEs) affecting the server infrastructure of Nany FinTrust MicroBank (Pty) Ltd, the model assesses the likelihood of exploiting these CVEs with limited effort using the exploit prediction scoring system (EPSS) [30]. EPSS derives its score from the output of the network scanning tool used in the experimental demonstration. For clarity, this tool is referred to hereafter as the network tool.

Step 1. Record the EPSS score from the network scan results:

From the Network toolscan results, assume that two Critical CVEs are found, such as CVE-2024-23692 and CVE-2014-6287.

Table 7. Base exploit prediction scoring system (EPSS) scores obtained from the National Vulnerability Database (NVD).

CVE ID	EPSS (Exploit probability)
CVE-2024-23692	0.9430
CVE-2014-6287	0.94316

The EPSS values associated with the identified CVEs indicate a high likelihood of exploitation with minimal effort, as indicated in Table 6. Both CVEs demonstrate a 94% probability of exploitation. While this result provides a robust initial indication of risk, the proposed model applies additional mechanisms to substantiate the probability rationale and derive a final AI-based score.

Step 2 introduces an algorithm incorporating open-source intelligence (OSINT). Public repositories frequently provide guidance and proof-of-concept (PoC) material for exploiting known CVEs. In the present study, GitHub serves as the primary OSINT source because of its widespread adoption, accessibility, and support for application programming interface (API) integration, which are required for the search algorithms presented in Algorithm 1. Prior studies, including Cosentino et al. [31] and Wang et al. [32], also identify GitHub as a dominant platform for publicly shared vulnerability-related artifacts.

As illustrated in Algorithm 1, the algorithm automates the assessment of online visibility and exploitation maturity for a provided CVE by querying GitHub for publicly accessible references linked to that vulnerability. The process constructs a search query using the CVE identifier and submits it to the GitHub Search API. The API returns repositories, commits, issues, and code fragments containing the CVE identifier. The algorithm then extracts the total number of references and applies a predefined bucket-based normalization scheme to transform the raw count into a standardized score within the interval $[0, 1]$. Lower reference counts correspond to lower scores, indicating limited public exploitation activity, whereas higher counts map to values approaching 1.0, indicating extensive discussion, available PoC material, or operational weaponization. The resulting normalized OSINT-derived metric feeds into the broader AI scoring computation used to evaluate cybersecurity controls within the proposed model.

 doi:10.20944/preprints202601.1703.v1

Algorithm 1 Compute INTELIGENCE From GitHub CVE Mentions $M \in [0, 1]$

procedure COMPUTE_GITHUB_MENTION_SCORE(CVE_ID)
  Input: CVE_ID
  Output: Normalized Mention Score M
  Step 1: Build Search Query for the CVEs found during Network Scan Results
  $query \leftarrow$ "https://api.github.com/search/code?q=" + CVE_ID
  Step 2: Send REST API Request and get response with total mentions about the CVEs found from the Network Scan results
  $response \leftarrow HTTP.GET(query)$
  $total\_mentions \leftarrow response.items.count$
  Step 3: Apply Normalization (Predefined Logic)
  if $total\_mentions = 0$ **then**
    $M \leftarrow 0.0$
  else if $1 \leq total\_mentions < 5$ **then**
    $M \leftarrow 0.2$
  else if $5 \leq total\_mentions < 20$ **then**
    $M \leftarrow 0.4$
  else if $20 \leq total\_mentions < 50$ **then**
    $M \leftarrow 0.6$
  else if $50 \leq total\_mentions < 100$ **then**
    $M \leftarrow 0.8$
  else
    $M \leftarrow 1.0$ ▷ Highly discussed / widely weaponised
  end if
  Step 4: Return Normalized Score
  return M
end procedure

Algorithm 1. Computing OSINT using GitHub CVEs mentions and popularity.

This section presents the logic underlying the integration of OSINT sources to enhance insight into the identified CVEs. A Python script then validates the applicability of the proposed algorithm by retrieving results through the GitHub search functionality. For demonstration purposes, the script uses the two identified CVEs as input and records the corresponding outputs, as illustrated in Figure 4 and summarized in Table 8.

Table 8. Public exploit mentions for target CVEs with weights.

CVE ID	Exploit source	EPSS	Code search count	Issues/ PRs count	Total GitHub mentions	Normalized score	M
CVE-2024-23692	GitHub	0.9430	338	30	368	0.80	
CVE-2014-6287	GitHub	0.94316	780	8	788	1.00	

```
[redacted]VirtualBox:~/Documents/Anti-Sherif Algorithm$ python3 cve_github_men
tions.py CVE-2024-23692 --token [redacted]24us
============================================================
CVE: CVE-2024-23692
  Code search count    : 338
  Issues/PRs count     : 30
  Total GitHub mentions: 368
  Normalized M score   : 0.80
```

 doi:10.20944/preprints202601.1703.v1

Figure 4. Script results.

As indicated in Figure 4, the methodology used in the proposed model to compute the final normalized value incorporates three GitHub-derived intelligence signals: code search counts, issues and pull requests (PRs), and the total number of references to a vulnerability across GitHub repositories. The algorithm aggregates these signals and applies the normalization thresholds defined earlier in Algorithm 1 to derive the threat intelligence score M.

For example, the vulnerability CVE-2024-23692, identified during the ICT vulnerability scan of the fictional organization Nany FinTrust MicroBank (Pty) Ltd, is processed using the Python script. The results, presented in Table 8, indicate 338 code references and 30 issue or PR references, yielding a combined total of 368 GitHub references associated with this CVE. Applying the normalization rules defined in Algorithm 1, the final normalized score for CVE-2024-23692 is 0.80, reflecting a high level of community attention, exploit maturity, and threat relevance. The same computational logic applies consistently to the second CVE included in the demonstration.

After computing the individual normalized scores for each CVE, the model computes the final score as follows:

$$t_i^{AI} = \alpha . EPSS_i + (1 - \alpha) . M_i \quad (8)$$

Where:

$$EPSS_i = exploit\ probability\ score\ as\ depicted\ in\ table\ 8$$

M_i= GitHub normalized popularity/mentions score as indicated in Table 8.

$$\alpha = Pre - defined\ weights$$

The model now defines all values required to compute the final AI score for each CVE. For CVE-2024-23692, the final AI score is computed as follows:

$$t_1^{AI} = 0.6(0.9430) + 0.4(0.80) = 0.5658 + 0.32 = 0.8858$$

Final AI score (CVE-2024-23692): 0.886

For CVE-2014-6287, the final AI score is computed as follows:

$$t_2^{AI} = 0.6(0.94316) + 0.4(1) = 0.565896 + 0.4 = 0.965896$$

Final AI score (CVE-2014-6287): 0.966

After defining the AI scores for each CVE, the model combines these values to compute the final AI contribution score.

$$t_{AI}^{final} = \frac{t_1^{AI} + t_2^{AI}}{2} = \frac{0.8858 + 0.9659}{2} = 0.92585$$

Final combined AI score = 0.926

After computing the AI-related results, the model defines the human judgment component. As presented earlier in Table 3, the cybersecurity control evaluation reference adopts values derived from the CMM framework. Within the defined scenario, the assigned auditor or penetration tester assigns a value of 0.8 based on expert judgment. This score corresponds to Level 2 (managed/repeatable) on the adopted maturity scale, indicating that the control exists and operates as intended but remains insufficiently optimized or resilient against advanced threat scenarios.

$$H_i = 0.8$$

With the AI-based and human judgment values defined, the model computes the final risk score, referred to as the Anti-Sherif hybrid score, as follows:

$$S_i = \beta t_i^{AI} + (1 - \beta) . H_i \quad (9)$$

$$S_i = 0.6 \times 0.92 + 0.4 \times 0.8 = 0.552 + 0.32 = 0.872$$

The final binary compliance assessment yields a score of 90%, indicating that most controls are present and formally satisfy audit requirements. This result alone, however, can present a potentially misleading representation of the organization's security posture, as a binary evaluation classifies a control as passing solely on its presence. Conversely, the hybrid Anti-Sherif score demonstrates that, despite the high binary compliance rate, active exploitation intelligence combined with moderate control maturity indicates a substantial level of residual risk. The model computes 87% residual risk, emphasizing a clear divergence between binary compliance outcomes and risk-informed assessment.

 doi:10.20944/preprints202601.1703.v1

This divergence illustrates that pass–fail evaluations are insufficient for accurately capturing real-world cybersecurity risk.

The demonstration of the proposed Anti-Sherif cybersecurity audit approach adopts a risk-based perspective, integrating AI-driven intelligence with expert human judgment. This finding aligns with the work of Muthukrishnan et al. [33], who emphasize the increasing importance of human-in-the-loop AI models within modern augmented AI approaches.

The study concludes with a comparative analysis of a conventional binary check exercise and the hybrid control assessment, summarized in Table 9.

Table 9. Comparison of binary and hybrid control assessment.

Assessment dimension	**Binary compliance model**	**Hybrid Anti-Sheriff Model**
Scoring logic	Pass / Fail (0 or 1)	Overall risk inheritance
Inputs considered	Control presence only	Risk-based approach
Computation method	Mean of binary indicators	A combination of binary score, AI integration, and human judgment
Final score	$T^{(0)} = 0.90$, 90% score, indicating that the tested controls were highly effective.	$S_i = 0.872$, 87% score, indicating that the inherent risks of the assessed cybersecurity control are very high.
Interpretation	Controls largely compliant	High residual risk remains
Sensitivity to threat landscape	None	High
Control maturity awareness	Ignored	Explicitly included
Risk differentiation	Low (coarse-grained)	High (fine-grained)
Decision accuracy	Superficial compliance	Risk-informed decision-making

Table 9 expands on the rationale supporting the proposed model by emphasizing the need to move beyond cybersecurity auditing as a tick-box exercise. A practical illustration of the limitations of binary checks arises in firewall rule management. A binary assessment may confirm that a firewall is implemented and active, yet fail to detect overly permissive any–any rules, unused legacy rules supporting decommissioned systems, or the absence of segmentation between critical business systems and less-trusted network zones. In such instances, the control formally exists and passes the audit, while simultaneously introducing significant exposure to lateral movement and data exfiltration risks that directly threaten business operations.

Over time, continued reliance on this approach adversely affects business continuity and operational stability, as latent control weaknesses, including zero-day attacks and internal lateral movement, remain undetected. These weaknesses can result in security incidents that disrupt core business processes and potentially affect revenue generation, despite the organization appearing compliant under conventional audit practices.

The proposed approach provides a foundational baseline for evolving cybersecurity audits from binary compliance verification toward risk-based evaluation. This transition enables the audit process to account for additional influential factors, including intelligence derived from public

 doi:10.20944/preprints202601.1703.v1

sources and structured human judgment. A critical evaluation of the proposed model is, therefore, required before identifying areas for future research and concluding the study.

4.4. Model Critical Evaluation

As emphasized in [34], adopting AI must account for human societal considerations, including human judgment and model explainability. Muthukrishnan et al. [33] similarly emphasize the importance of human-in-the-loop approaches. In the present study, the proposed model integrates AI capabilities with human judgment to enhance the effectiveness of technical cybersecurity controls and cybersecurity audit processes for compliance and business-strategy alignment. Although debate persists over the quantification of cybersecurity risk, the proposed model introduces a scoring mechanism that supports the quantitative assessment of control maturity and the overall cybersecurity audit process maturity, based on both human and AI inputs. This approach provides a bounded, explainable, and risk-based framework that clarifies the relationship between human expertise and AI collaboration in strengthening cybersecurity control postures.

As demonstrated in the experimental evaluation, a key strength of the proposed model lies in its dual-lens assessment of technical controls and compliance auditing against defined standards and internal organizational policies. By integrating technical control evaluation with audit oversight, the model supports a comprehensive assessment of cybersecurity posture. The model also incorporates human factors, reinforcing the Anti-Sherif audit philosophy. While the model demonstrates the capacity to quantify real-world control effectiveness beyond checklist-based compliance, a critical evaluation remains necessary to assess its robustness, validity, and limitations, as summarized in Table 8.

For clarity, fully satisfied evaluation criteria are marked with a tick (✓), partially satisfied criteria are indicated with a dash (–), and criteria that are not satisfied are denoted by a cross (×)

Table 10. Model critical evaluation.

Evaluation criteria	Rational	Results	Explanation
The proposed model should be validated using various other relevant models	The proposed model should align logically with international standards and benchmarks, such as CIS, NIST CSF, and ISO 27001.	✓	In the present study, the author cited international standards, such as the CIS. The cybersecurity controls adopted from the model are derived from the NIST CSF. This was conducted to ensure that the model is not interpreted as an alternative or replacement for the current recognized cybersecurity standards, but rather as a supplement to them
The proposed model data sources should be dependable	The data source used should be reliable and easily accessible to support study repeatability and future research.	–	The data used in the proposed model were obtained from an experimental setup meant specifically for the model demonstration. In an ideal world, a simulated environment with more than one server can provide a more realistic representation of the perception behind the proposed model
The model should be centered around the	Evaluate the proportionality of	✓	The model clearly demonstrated the importance of human-centered AI

 doi:10.20944/preprints202601.1703.v1

human judgment-influencing factor	human judgment toward the improvement of the model.		capabilities, in which the input from the expert played a vital role in the final risk score
The model should integrate AI enhancement	Evaluate the extent the which AI factors contribute to the model improvement.	–	Although the model integrates the AI capabilities through OSINT, more influential factors can be built using real-life threat detection AI capabilities, and user-centric input to enhance the AI insights
The proposed model should be scalable and automated	Examine if the model can be generalized, expanded, or automated for large-scale audits with a larger dataset.	–	The proposed model automated most of the entailed steps; however, the existence of cybersecurity for the binary check exercise was conducted using a manual approach. More automation can be considered for the concept of cybersecurity auditing
The model should be explainable, interpretable, and transparent	Measures the clarity of model outputs for non-technical stakeholders.	✓	The proposed model was presented in a way that could be understood by both technical and non-technical personnel
The applicability of the proposed model should be tested using empirical validation	Examine the adaptability of the model's outcomes against real-world audit or incident data, reflecting on the case scenario application adopted.	✓	A fictional case scenario was used to test and validate the applicability of the proposed model in a real-life scenario

The proposed model provides a foundational baseline from which more advanced Anti-Sherif models can be developed; however, it has several limitations. The subsequent section presents the study's limitations and outlines directions for future research.

 doi:10.20944/preprints202601.1703.v1

5. Model Limitations and Future Work

The challenges arising from the ongoing evolution of ICT and technological advancement cannot be resolved by a single study. Responding to these challenges requires more integrated, collaborative research across cybersecurity and related technological domains, including AI.

5.1. Study Limitations

A key limitation of the present study relates to its defined scope. The study focuses on the technical assessment of cybersecurity controls, enabling an in-depth evaluation but excluding a broader analysis of business, operational, and strategic risks associated with modern technological advancement. Expanding the scope to incorporate these dimensions would support a more comprehensive assessment of organizational risk in future research.

A second limitation concerns using fictional or simulated scenario data. While such data adequately support experimental validation and PoC demonstration, they may not fully reflect the complexity and unpredictability of real-world environments. The robustness of the proposed approach could be strengthened by applying the model in operational organizational contexts or by using data derived from historical cybersecurity audit engagements.

5.2. Future Direction

Future studies can expand the scope beyond technical cybersecurity control assessments to examine cybersecurity governance risks and the role of modern technologies, including AI, in mitigating such risks. Future work may also adopt a broader range of research methodologies, including qualitative approaches, such as interviews and questionnaires with cybersecurity experts, to assess the practical applicability of AI in advancing cybersecurity practice. Mixed-method designs could further enrich understanding by integrating empirical technical evaluation with practitioner perspectives.

With the study's limitations and future research directions defined, the present study concludes.

6. Conclusion

The proposed Anti-Sherif audit approach presents both a conceptual model and a practical mechanism for integrating AI with human governance to enhance cybersecurity audit processes. By combining automated intelligence with expert oversight, the approach mitigates limitations inherent in traditional binary audit methods and strengthens audit depth, adaptability, and effectiveness. As cybersecurity threats continue to grow in complexity and sophistication, this approach becomes essential not only for safeguarding information exchange but also for maintaining stakeholder trust in digital transformation initiatives. The experimental evaluation demonstrates the potential of AI-assisted auditing to support more robust, context-aware, and forward-looking cybersecurity assurance practices.

7. Disclaimer

The authors acknowledge the limited use of AI-generated content in the study titled "A Detailed Anti-Sherif AI-Driven Cybersecurity Audit Model: Beyond AI Adoption in Cybersecurity Auditing." ChatGPT was used to improve the readability of selected concepts; all such content was subsequently reviewed, reworded, and edited to align with the study's objectives and scholarly standards. Grammarly was used to identify spelling and grammatical issues before submission to a professional language editor.

Abbreviations

The following abbreviations are used in this manuscript:

 doi:10.20944/preprints202601.1703.v1

AI	Artificial intelligence
APT	Advanced persistent threats
CCM	Continuous control monitoring
CIA	Confidentiality, integrity, and availability
CIS	Center for Internet Security
CISA	Certified Information Systems Auditor
CMM	Capability maturity model
CSF	Cybersecurity Framework
CVE	Common vulnerabilities and exposures
DLP	Data loss prevention
DR	Disaster recovery
DSR	Design science research
DSRM	Design science research methodology
EDR	Endpoint detection and response
EPSS	Exploit prediction scoring system
ICT	Information and communication technology
IDS	Intrusion detection systems
IoT	Internet of Things
IPS	Intrusion Prevention Systems
IR	Incident response
IS	Information systems
ISACA	Information Systems Audit and Control Association
KPI	Key performance indicators
MFA	Multi-factor authentication
MI	Material irregularities
NIST	National Institute of Standards and Technology
NVD	National Vulnerability Database
OSINT	Open-source intelligence
PoC	Proof-of-concept
POPIA	Protection of Personal Information Act
PR	Pull requests
RPO	Recovery point objective
RTO	Recovery time objective
SIEM	Security information and event management
SME	Small and medium-sized enterprises
SOAR	Security orchestration, automation, and response
UEBA	User and entity behavior analytics

References

1. D. Craigen, N. Diakun-Thibault, and R. Purse, "Defining Cybersecurity," *Technol. Innov. Manag. Rev.*, vol. 4, no. 10, pp. 13–21, 2014, doi: 10.22215/timreview835.
2. NIST, "Computer Security Resource Center." Accessed: Aug. 12, 2024. [Online]. Available: https://csrc.nist.gov/glossary/term/cybersecurity
3. J. vom Brocke, A. Hevner, and A. Maedche, "Introduction to Design Science Research," no. November, pp. 1–13, 2020, doi: 10.1007/978-3-030-46781-4_1.

 doi:10.20944/preprints202601.1703.v1

4. A. R. Hevner, "Design science research," *Comput. Handb. Two-Volume Set*, pp. 1–23, 2022, doi: 10.1201/b16768-26.
5. M. Cai, J. Yang, J. Gao, and Y. J. Lee, "Matryoshka Multimodal Models," no. 2, pp. 1–16, 2024, [Online]. Available: http://arxiv.org/abs/2405.17430