Journal of
Cyber Security

Tech Science Press

https://doi.org/10.32604/jcs.2026.077183

REVIEW

Phishing, Vulnerabilities, and AI Defense: A Systematic Review of Cybersecurity Challenges and GRU-Based Mitigation Strategies in Digital Microfinance Institutions

Richard Mathenge*, **Catherine Mukunga and Ephantus Mwangi**

School of Pure and Applied Sciences, Kirinyaga University, Kerugoya, Kenya

*Corresponding Author: Richard Mathenge. Email: mathengerichard29@gmail.com

Received: 04 December 2025; Accepted: 27 January 2026; Published: 11 March 2026

ABSTRACT: The rapid digitization of microfinance institutions (MFIs) has strengthened financial inclusion but has simultaneously increased exposure to phishing attacks and other cybersecurity threats driven by organizational, technical, and human vulnerabilities. Grounded in socio-technical systems theory, this systematic analysis evaluates AI-based mitigation strategies, with particular emphasis on gated recurrent unit (GRU) architectures. It compares them with Transformer and LSTM models. GRUs are prioritized due to their computational efficiency and suitability for low-resource environments typical of digital MFIs. Following PRISMA 2020 guidelines, 32 empirical studies published between January 2012 and April 2025 were analyzed from the Web of Science, ScienceDirect, Google Scholar, IEEE Xplore, and Scopus. Thematic synthesis was conducted using the Braun and Clarke framework, and methodological quality was assessed using the Mixed Methods Appraisal Tool (MMAT). The findings indicate that obsolete infrastructure, limited employee awareness, and weak governance structures account for approximately 67% of cybersecurity incidents in MFIs. Under experimental conditions, GRU-based models achieved phishing-detection accuracies of 92% to 96%, demonstrating strong performance in sequential behavior analysis. Despite these advantages, deployment remains constrained by infrastructural limitations, limited model explainability, and scarcity of domain-specific datasets. This study proposes an implementation roadmap integrating explainable AI, ethical governance, and region-specific capacity building, alongside a vulnerability-solution matrix linking threat vectors to appropriate AI-based countermeasures. The findings provide a structured foundation for developing secure, scalable, and AI-enabled digital financial ecosystems for legislators, cybersecurity practitioners, and MFI stakeholders.

KEYWORDS: Cybersecurity; phishing; digital microfinance; gated recurrent units (GRU); explainable AI; financial inclusion; systematic review

1 Introduction

The rapid digitalization of financial services has fundamentally transformed microfinance institutions (MFIs), expanding access for unbanked populations and strengthening financial inclusion across low- and middle-income countries. However, this accelerated digital transformation has simultaneously increased exposure to cybersecurity threats, particularly in resource-constrained environments where institutional preparedness and technical infrastructure remain limited [1,2]. The adoption of cloud-based platforms, mobile banking systems, and real-time digital wallets has expanded service delivery. Still, it has also heightened vulnerability to cybercriminal exploitation of structural and human weaknesses [3].

Within this evolving threat landscape, phishing and related social engineering attacks have emerged as dominant risk vectors. These attacks—often combined with malware distribution and credential harvesting—undermine institutional stability, erode user trust, and compromise financial integrity [4]. Accordingly, this review focuses specifically on phishing-related vulnerabilities and artificial intelligence (AI)-driven detection strategies within digital MFIs to ensure conceptual precision within the broader cybersecurity discourse.

Empirical evidence indicates that cybersecurity readiness across many MFIs remains inadequate. Contributing factors include fragmented digital governance frameworks, inconsistent risk management practices, and limited investment in information security infrastructure [5]. In MFI contexts, vulnerabilities are amplified by the hybrid interaction between technological weaknesses—such as weak authentication mechanisms—and human factors, including phishing susceptibility and insufficient staff training. Ferrag et al. [6] identify phishing as the most prevalent cyberattack in digital financial institutions, particularly where multi-factor authentication protocols are poorly implemented. This challenge is especially pronounced in South Asia and Sub-Saharan Africa, where rapid digital financial expansion has outpaced the adoption of robust cybersecurity controls.

Artificial intelligence (AI), specifically machine learning (ML), offers a promising pathway to mitigate these risks. AI-driven systems enable real-time threat detection, automated anomaly identification, and adaptive risk management—capabilities that are particularly valuable in institutions lacking formalized cybersecurity frameworks [3]. Within the microfinance sector, AI techniques such as recurrent neural networks (RNNs) and natural language processing (NLP) have been applied to phishing detection, fraud analytics, and behavioral segmentation [1,7].

Despite these advancements, practical implementation remains uneven across institutions and regions. Barriers include limited algorithmic transparency, regulatory ambiguity, and the scarcity of domain-specific labeled datasets—constraints that are especially significant where explainable AI (XAI) is legally or ethically mandated [8]. Furthermore, existing scholarship is often fragmented, geographically dispersed, or predominantly technical in orientation, with limited attention to contextual adoption barriers within microfinance ecosystems. A systematic synthesis of AI-based phishing mitigation strategies tailored specifically to digital MFIs remains underdeveloped.

To address this gap, this systematic review evaluates the effectiveness of AI models, particularly gated recurrent unit (GRU) neural networks, in detecting and preventing phishing attacks in digital MFI settings. GRU architectures are preferred for their faster convergence, reduced computational complexity, and strong performance on small, imbalanced phishing datasets compared to long short-term memory (LSTM) and Transformer-based models [9,10]. These characteristics render GRUs especially suitable for resource-limited MFIs, where processing capacity and labeled training data are constrained. This study applies the PRISMA 2020 framework to synthesize findings from 32 peer-reviewed empirical studies published between January 2012 and April 2025.

This review pursues three interrelated objectives: (1) to identify the primary cybersecurity vulnerabilities affecting digital MFIs; (2) to evaluate the effectiveness of GRU architectures relative to alternative AI models in mitigating phishing threats; and (3) to propose a practical AI deployment roadmap aligned with socio-technical systems theory and ethical governance frameworks.

The central contribution lies in integrating empirical evidence on GRU performance under low-resource conditions with an operational vulnerability–solution matrix tailored to digital MFIs. By combining comparative technical evaluation with applied deployment guidance, this study clarifies a critical research gap and advances context-sensitive AI strategies for inclusive financial cybersecurity.

2 Literature Review

2.1 Cybersecurity Vulnerabilities in Digital Microfinance Ecosystems

The digital transformation of microfinance institutions (MFIs) has improved operational performance and financial coverage, especially in underserved regions. Despite impressive technological growth, numerous complex cybersecurity threats have emerged. Microfinance institutions (MFIs) also face numerous challenges associated with the socio-technical systems theory, which holds that systemic problems arise when governance frameworks, human behavior, and technological systems fail to align [7]. Recent reviews on AI-enabled cybersecurity in finance, Wiafe et al. [11], Das et al. [12], and Patil et al. [13] emphasize algorithmic benchmarking but rarely contextualize these findings in microfinance operations. This review uniquely extends that perspective to low-resource digital financial ecosystems.

The major technological vulnerabilities stem from the continued use of obsolete software systems, weak encryption, and improperly configured access controls by many MFIs. Such vulnerabilities make critical systems highly exploitable, especially because the resources required to update or patch them regularly are insufficient or unavailable [6].

Organizational weaknesses are largely associated with a lack of specific incident response structures, fragmented security governance, insufficient or obsolete cybersecurity policies, and underfunded IT departments [2,3].

Human-related vulnerabilities also pose a significant risk. Typically, frontline workers use weak passwords, lack sufficient awareness of best cybersecurity practices, and lack specialized training. Employees are particularly vulnerable to social engineering and phishing attacks, and they are often the initial target [5].

Research conducted in Kenya, Nigeria, and South Asia indicates that phishing is the most common attack against microfinance institutions. The use of social engineering and phishing attacks, which are often the gateway to attacks, is especially prevalent in MFIs where employees are not adequately trained [5]. Studies in these regions indicate that phishing attempts make up more than 60 percent of reported cases. In these cases, social engineering, email spoofing, or SMS attacks are frequently used to exploit human error and technological vulnerabilities. The different degrees of vulnerability demonstrate the necessity of a holistic hybrid defense approach that should involve the latest technologies capable of dynamically detecting threats, implementing policy changes, and providing employee training.

2.2 AI-Driven Responses to Phishing and Threat Detection

Proactive cybersecurity in MFIs has enormous potential with artificial intelligence (AI), particularly machine learning (ML). In high-dimensional, complex data, AI systems are exemplary at large-scale anomaly detection. Owing to these abilities, they are ideal for detecting phishing attacks in less-monitored environments.

Artificial intelligence (AI), and more specifically, machine learning (ML), has demonstrated limitless potential for proactive cybersecurity in MFIs. One type of AI architecture that has been valuable for simulating sequential and temporal patterns in user behavior and email data is the Gated Recurrent Unit (GRU), a variant of Recurrent Neural Networks (RNNs). Studies demonstrate that GRUs outperformed traditional classifiers by 92%–96% in phishing scenarios [1,14]. Moreover, they are attractive to MFIs with limited infrastructure because they are computationally efficient. Nevertheless, due to their static feature interpretation, earlier Decision Tree and Support Vector Machine (SVM) models are less effective in dynamic threat settings [15].

Beyond GRU architectures, transformer-based models such as BERT and DistilBERT have achieved state-of-the-art accuracy in phishing text classification through self-attention mechanisms that learn long-term dependencies [16,17]. However, their high computational and data demands make them less feasible for MFIs with limited infrastructure. LSTM and hybrid CNN-LSTM architectures offer a middle ground but often at the expense of training time and resource intensity compared to GRUs [9]. These comparative insights provide the analytical basis for evaluating GRU performance within the resource-constrained operational realities of MFIs.

Recent literature also highlights the potential of reinforcement learning (RL) for adaptive phishing defense, where the model iteratively improves based on feedback from new phishing encounters [18]. Although reinforcement learning (RL) and hybrid Transformer–GRU architectures show adaptive potential, they require extensive training data and computing power that most MFIs cannot sustain. In contrast, GRU architectures achieve comparable detection accuracy with substantially lower computational and data costs, making them operationally viable for low-resourced digital MFIs. This comparative efficiency, rather than algorithmic novelty alone, justifies the review's focus on GRU as the primary AI model for phishing detection in digital MFIs [11,19,20].

The applications of reinforcement learning have also become more popular in adaptive filtering. These models allow real-time changes in phishing patterns based on their past experiences of threats. These have reached only the early stages of implementation and require further testing in real-life MFI contexts, despite their technical potential [2].

2.3 Challenges in Deploying AI in MFIs

Despite the potential, several structural barriers remain to the popularization of AI in MFIs. The primary challenge is the lack of labeled phishing datasets for specific domains, which negatively affects the training and performance of supervised learning models. According to Gai et al. [1], all current datasets are either too broad or collected in large banking institutions and do not reflect the communication and behavioral patterns of microfinance institutions (MFIs).

The other important issue is model transparency. Many strong AI models, particularly deep neural networks, operate in "black boxes", making it challenging to understand how they make decisions. This ambiguity raises issues of trust and regulatory compliance in the financial sector. The increasing attention of policymakers and regulators to explainable AI (XAI) highlights the need for reliable, verifiable models [8].

Poor infrastructure is also a hindrance to deployment. The majority of microfinance institutions, particularly in rural or semi-urban areas, lack the capacity to process large volumes and real-time artificial intelligence systems. Sustainability of AI-based security solutions is complicated by poor internet connectivity and the use of outdated technologies [3].

Lastly, organizational resistance is typically ignored. Most MFIs have low digital maturity and are reluctant to automate due to a shortage of internal AI expertise or fear of job losses. Unless institutional capacity-building and effective change management processes are embraced, the sustainability of AI solutions remains questionable [21]. Together, these limitations highlight the dire need to have context-sensitive, lightweight, and explainable AI models for MFIs.

2.4 Contradictions and Gaps in Existing Literature

The existing literature has numerous contradictions. Even with remarkable technological progress, many advanced cybersecurity vulnerabilities have emerged. The socio-technical systems theory highlights that systemic failures arise from misalignments between governance structures, human behavior, and technological systems, a pattern observed not only in digital finance but also in crisis-driven public systems such as pandemic response infrastructures [7,22].

The current literature also shows an important gap in contextual and regional aspects. The majority of the literature implicitly assumes that microfinance institutions (MFIs) in South Asia and Sub-Saharan Africa are technologically endowed and digitally equipped, which, in contrast, is inadequate [15].

Such assumptions jeopardize the external validity and feasibility of the proposed solutions. Additionally, unlike much of the existing literature focused on phishing, MFI cybersecurity discourse has not adequately covered other emerging threats, such as Subscriber Identity Module (SIM) swap scams, insider attacks, mobile malware, and AI-generated impersonations.

A lack of contextual authenticity, particularly in assumptions about the data and the implementation methods, reflects a disconnect between the model's development and its practical applicability. This inconsistency is the motivation behind this review.

2.5 Integrated Summary of Literature Themes

The reviewed literature has four major themes. Firstly, MFIs are perpetually vulnerable to a list of organizational (poor governance), human (phishing susceptibility), and technical (obsolete systems and inadequate encryption) threats. Second, despite other models, including LSTMs, reinforcement learning, and SVMs, contributing to the methodological landscape, AI techniques, especially GRUs, have demonstrated tremendous potential for identifying phishing. Third, AI adoption remains inhibited by several systemic issues, including model explainability, infrastructural failures, and data deficits. Finally, the study reveals significant weaknesses, including poor alignment with MFIs' operational realities in emerging markets, insufficient representation of non-phishing risks, and limited analysis of post-deployment performance.

Together, these issues underscore the need for a systematic synthesis that examines the usefulness of AI in digital microfinance, its deployment limitations, its sociotechnical significance, and ethical implementation approaches. To illustrate these integrated themes, Fig. 1 below presents the interconnections among the primary literature on AI and cybersecurity in digital MFIs. Further, Table 1 presents a detailed summary of the key findings from the given literature.

As shown in Table 1, GRUs consistently demonstrate efficiency in low-resource phishing detection, whereas Transformer architectures excel in contextual accuracy but at a higher computational cost. LSTM and hybrid CNN–LSTM approaches strike a middle ground but remain resource-intensive. This synthesis highlights that technical superiority must be balanced with deployment feasibility in MFI environments.

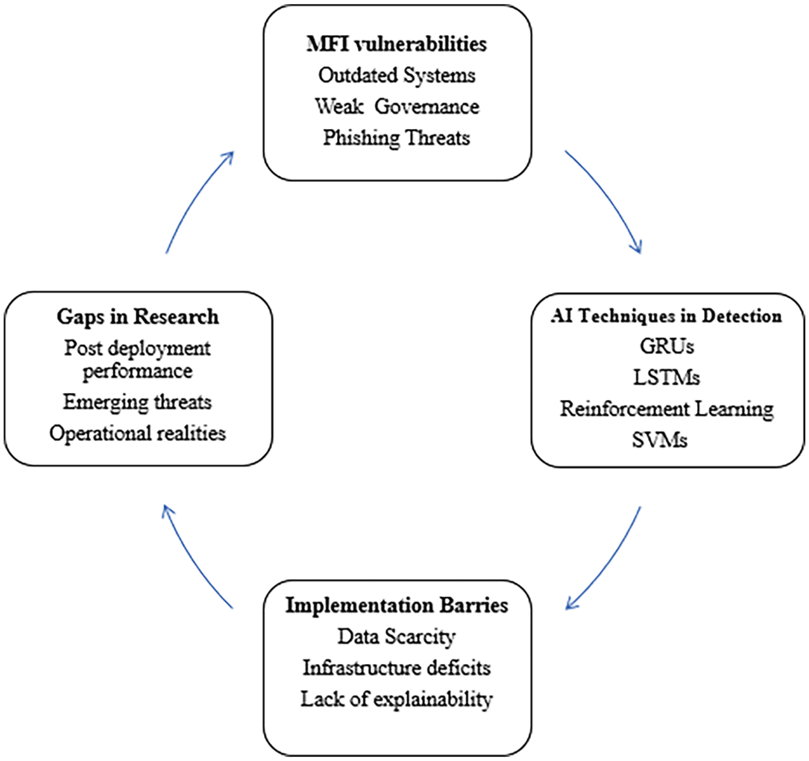

Figure 1: Literature themes on AI and Cybersecurity in digital microfinance institutions.

Table 1: Summary of key themes in literature.

Theme	Findings	Representative Studies
Common MFI Vulnerabilities	Weak encryption, phishing susceptibility, untrained staff, and a lack of response plans	Manasseh & Ede [5]; Bouveret [14]; Ali et al. [23]
AI Techniques in Threat Detection	GRUs, LSTMs, SVMs, reinforcement learning, anomaly detection	Ferrag et al. [6]; Wiafe et al. [11]; Bu & Cho [24]
Implementation Barriers	Data scarcity, low digital maturity, infrastructure limitations, XAI concerns	Kumari et al. [7]; Daudu et al. [25]; Deshpande [26]
Literature Gaps	Lack of real-world deployments, regional datasets, and overlooked non-phishing threats	Wiafe et al. [11]; Kavya & Sumathi [18]; Arner et al. [27]

3 Methodology

This systematic literature review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [28] to ensure scientific rigor, reproducibility, and openness. The main aims were to collect data on phishing threats in digital MFIs and to measure the use of artificial intelligence models, particularly GRUs, to detect and prevent them. A realist worldview and socio-technical systems theory influenced the analysis, leading to the use of the thematic synthesis approach.

3.1 Research Design and Framing

The review adopted a realist epistemological stance, emphasizing the interactions among organizational structures, human factors, and technological systems in the development of cybersecurity risks in MFIs [29]. To guide this review, the research design used the SPIDER framework (Sample, Phenomenon of Interest, Design, Evaluation, Research type) [30]. Digital Microfinance Institutions (MFIs) constituted the Sample (S). Phishing threats were specifically given priority in the Phenomenon of Interest (PI), which focused on cybersecurity weaknesses. Empirical research and practical AI implementations were included in the Design (D). The evaluation (E) focused on detection accuracy measures (such as F1-score and AUC) and the effectiveness of AI model mitigation. Mixed-methods, qualitative, and quantitative studies were included under Research type (R). Therefore, the primary review question was: Which cybersecurity vulnerabilities, particularly phishing, are prevalent within digital microfinance institutions, and to what extent have AI-based models, specifically GRU neural networks, been deployed to mitigate such attacks? The targeted attention to phishing and GRUs can be justified by their demonstrated prevalence as the most popular attack vectors against MFIs and by the focus on evaluating computationally efficient deep learning models that can be applied in resource-constrained settings [1,14]. Table 2 illustrates the breakdown of the operationalization of the SPIDER elements in the MFIs.

Table 2: Spider framework.

SPIDER Element	Operationalization
Sample(S)	Digital MFIs in Latin America, Sub-Saharan Africa, and South Asia
Phenomenon of Interest (PI)	Phishing and emerging social engineering threats
Design	Empirical research
Evaluation	Accuracy metrics such as F1-Score and AUC
Research Type	Mixed methods

3.2 Search Strategy and Data Sources

To address coverage bias, Google Scholar was included. It was necessary to thoroughly screen this search engine due to its heterogeneous indexing [31]. The review protocol adhered to PRISMA 2020 [28] and employed the Mixed Methods Appraisal Tool (MMAT) for methodological assessment [16], ensuring transparency and reproducibility. The search covered the period from January 2012 to April 2025. Controlled vocabulary and Boolean logic were customized for the syntax of each database. Integrated core search strings: ("cybersecurity" OR "phishing" OR "email fraud" OR "social engineering") AND ("digital microfinance" OR "microfinance institution*" OR "MFI*" OR "digital financial service**") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "GRU" OR "gated recurrent unit"). The identical basic phrases were used in Google Scholar searches. The platform's algorithm prioritized findings deemed contextually meaningful, and the first 200 items, sorted by relevance, were inspected, reflecting practical methods for handling its large yield [7]. To reduce publication bias, searches were complemented with forward citation tracking (using Scopus/Google Scholar) and backward citation tracking (references of included articles) [32]. 547 documents were found in the initial searches. 419 distinct records proceeded to screening after 128 duplicates were eliminated both manually and algorithmically. The exclusion of non-English studies due to resource limitations may have skewed the results in favor of Anglophone regions (e.g., Kenya and India). Multilingual databases should be a part of future research. Grey material, such as vendor whitepapers and MFI incident reports, was not included, thereby limiting our understanding of the actual deployment difficulties. Detailed Boolean search strings for each database are provided in Appendix A.

3.3 Eligibility Criteria and Article Selection

The following were specific requirements for inclusion: the studies had to be peer-reviewed journal articles or conference proceedings, published in English between January 2012 and April 2025, focus on MFIs or similar online financial services for marginalized communities, explicitly address cybersecurity vulnerabilities, with a strong emphasis on phishing threats, incorporate AI/ML methods for threat mitigation, especially relevant to GRU models, and present empirical data, model evaluations, or applied case studies. The following were excluded: studies that only focused on big banks or non-microfinance fintech; studies that didn't address cybersecurity or AI-driven mitigation; publications that weren't in English or that weren't published within the specified time frame; editorials, comments, or merely theoretical works, as well as grey literature sources (such as unpublished case studies and industry reports). 136 papers were selected for full-text review after two independent reviewers vetted all 419 titles and abstracts. During full-text screening, conflicts (n = 18) were settled by consensus; in situations that could not be resolved (n = 3), a third reviewer was involved. Of the 32 articles included, all satisfied all eligibility requirements. The PRISMA 2020 flow diagram was used to chronicle the selection process. Fig. 2 provides detailed criteria for article selection using the PRISMA 2020 technique.

3.4 Quality Appraisal of Included Studies

The Mixed Methods Appraisal Tool (MMAT) 2018 [16] was used to evaluate the methodology of the 32 included papers, in accordance with its use in current systematic reviews [16]. Analytical rigor, transparency of the AI model description (particularly the GRU design and parameters), reproducibility (sufficient detail for replication), appropriateness of the data-collection/model-training datasets, and clarity of the research questions were the main evaluation criteria. Research that satisfied at least 3 of the 5 pertinent MMAT requirements was retained for synthesis [33]. After excluding 3 studies with scores below this cutoff, 29 studies remained for data extraction. The majority of the included studies demonstrated high quality (scoring 4 or 5). Of the 32 studies that satisfied the PRISMA 2020 screening criteria, three were excluded during the MMAT assessment because their methodological quality scores were below 3 out of 5. Consequently, 29 studies were retained for synthesis and comparative analysis. References to 32 studies, therefore, denote the total screened pool, whereas 29 represent the final dataset used for results synthesis. This clarification ensures methodological transparency and consistency across sections.

A detailed summary of the MMAT evaluation is presented in Table 3. The average methodological quality across the 32 studies was 4.3/5, indicating strong adherence to design transparency and analytical rigor [16]. The most frequent weaknesses included incomplete documentation of dataset preprocessing (28%) and insufficient details on reproducibility (17%), reflecting persistent challenges with replicability across AI research [19]. These gaps underline the need for greater methodological transparency in AI-based cybersecurity studies. This finding is consistent with ongoing discourse on AI reproducibility and transparency, emphasizing the necessity of explicit dataset preprocessing and model documentation in applied cybersecurity research [16,34].

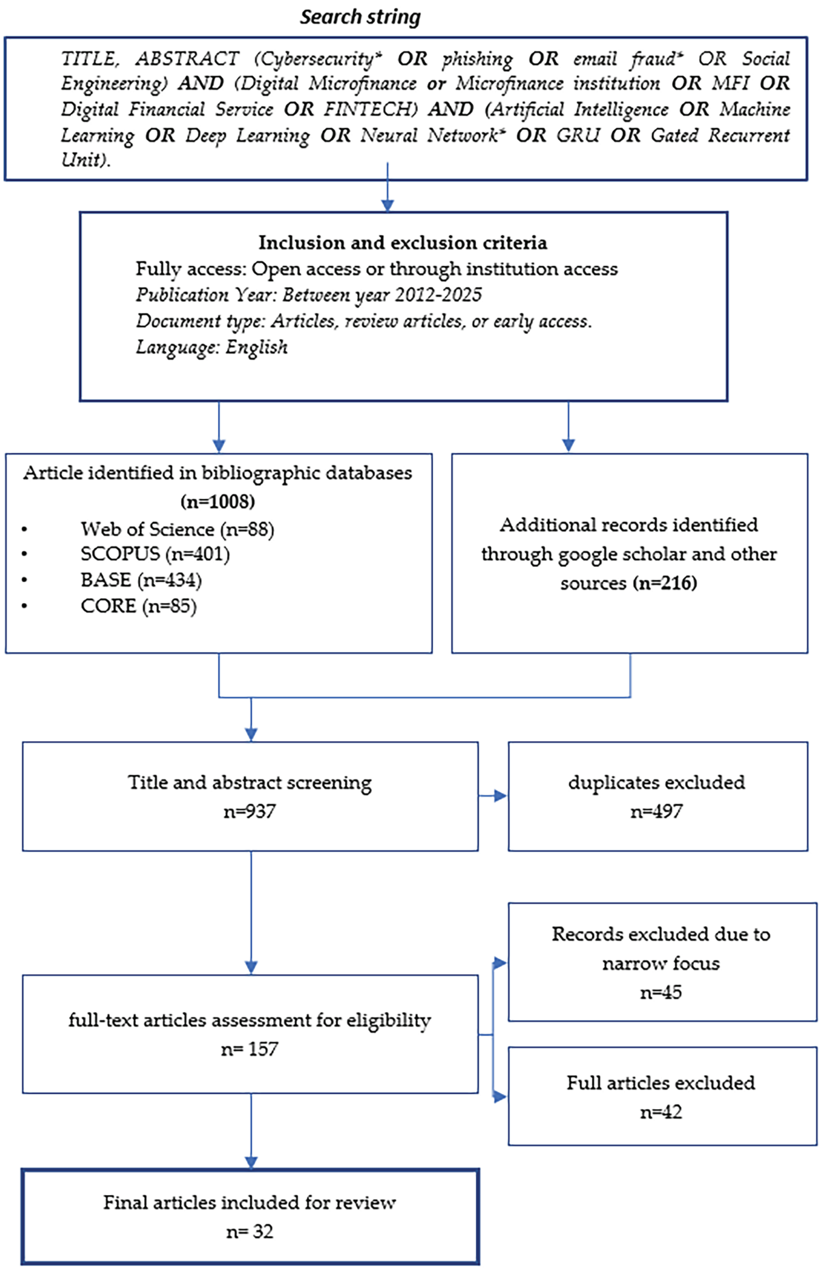
Search string
TITLE, ABSTRACT (Cybersecurity* OR phishing OR email fraud* OR Social Engineering) AND (Digital Microfinance or Microfinance institution OR MFI OR Digital Financial Service OR FINTECH) AND (Artificial Intelligence OR Machine Learning OR Deep Learning OR Neural Network* OR GRU OR Gated Recurrent Unit).
Inclusion and exclusion criteria
Fully access: Open access or through institution access
Publication Year: Between year 2012-2025
Document type: Articles, review articles, or early access.
Language: English
Article identified in bibliographic databases (n=1008)
Web of Science (n=88)
SCOPUS (n=401)
BASE (n=434)
CORE (n=85)
Additional records identified through google scholar and other sources (n=216)
Title and abstract screening n=937
duplicates excluded n=497
Records excluded due to narrow focus n=45
full-text articles assessment for eligibility n= 157
Full articles excluded n=42
Final articles included for review n= 32

Figure 2: PRISMA 2020 flow diagram of study selection.

3.5 Data Extraction and Coding Strategy

Bibliographic information, MFI context and geography, specific phishing vulnerabilities, AI model types (with a focus on GRU architecture details like layers, units, and activations), dataset characteristics (source, size, and type, such as phishing emails or SMS), model performance metrics (Accuracy, F1, AUC, and computational efficiency), implementation setting (simulation/lab/field), and limitations were all recorded using a standardized extraction form. Two researchers conducted independent extractions. Discussions resolved discrepancies. Using Cohen's Kappa (κ) on a 10% sample, inter-coder reliability was 0.82 (good agreement) [27]. For data management, MAXQDA 2022 was used.

Table 3: Summary of MMAT quality scores for included studies.

Quality Criterion	% of Studies Meeting Criterion	Common Weakness
Clear research question	100%	–
Adequate data collection method	96%	Limited dataset disclosure
Reproducibility (model parameters stated)	83%	Missing hyperparameter tuning
Transparency of AI architecture	90%	Limited interpretability description
Appropriate data analysis	97%	Some lacked error analysis

3.6 Thematic Synthesis Approach

The data were synthesized using the six-phase framework for theme analysis developed by Braun and Clarke [35]. The initial stage involved familiarizing the user with the extracted data. The second step was to come up with preliminary descriptive labels (including 'GRU phishing detection accuracy >95%', 'high staff susceptibility to spear phishing', 'non-existence of phishing simulation training', and 'GRU lower training time vs. LSTM') directly out of the data. During the third stage, these codes were grouped under potential themes, including 'Infrastructure and Resource Barriers to AI Deployment', 'Human and Organizational Drivers of Phishing Susceptibility', and 'Technical Efficacy of GRUs for Phishing Detection'. The potential themes were compared to the fully coded data and then refined in phase four to achieve uniqueness and coherence. In the fifth phase, each final theme was meticulously defined and assigned a precise name that captured its essence. In phase six, the analytical report was produced. Mostly inductive, the categorization procedure was motivated by the included studies' content. In MFI situations, for example, codes about GRU high-performance indicators were combined into analytical themes that examined both their potential and real-world constraints. A Vulnerability-AI Mitigation Matrix was created through this consensus-driven, iterative process. The compiled results of the studies clearly show that this matrix correlated phishing threats and the organizational and human vulnerabilities they were linked to, and the AI methods (particularly GRUs) that were suggested or used for mitigation.

3.7 Limitations of the Review Methodology

The key drawbacks have been acknowledged. Excluding non-English studies may introduce language bias, potentially underrepresenting certain regions [36]. Excluding grey literature restricts the insights gained from real-world implementations [37]. Because AI models, data, and metrics were heterogeneous, meta-analysis was not possible; narrative synthesis was required [26]. While the phishing/GRU focus offers stability, the quick development of AI increases temporal sensitivity. Though feasible [7], this approach relies on Google Scholar's opaque ranking by filtering only the first 200 results. The overall direction of the research question, which focuses on phishing and GRUs, automatically limits the scope of the synthesis to incorporating the discussion of other vulnerabilities or AI solutions. Despite these constraints, the methodological soundness is assured through strict adherence to PRISMA 2020, systematic search, intensive screening/appraisal, and structured synthesis.

4 Results

In this section, the primary cybersecurity vulnerabilities present in digital microfinance institutions (MFIs) are summarized and compared with the relevant artificial intelligence (AI) mitigation measures to demonstrate the findings of systematic reviews. The synthesis is divided into five main themes, which encompass the performance of GRU models, the current use of AI models, the prevalent nature of cybersecurity vulnerabilities, conceptual vulnerability-solution mapping, and difficulties in the implementation.

4.1 Prevalence of Cybersecurity Vulnerabilities in Digital MFIs

Cybersecurity weaknesses are ubiquitous and intricate in digital MFIs. The three related areas that they encompass are organizational governance, human factors, and technology infrastructure. The weaknesses of technological systems within MFIs have been highlighted in various research, including the absence of firewall detection systems, outdated encryption, and unpatched programs. Jony et al. [29] and Carcillo et al. [38] theorize that many rural MFIs operate aged systems and, therefore, they are prone to phishing emails with incorrect header names, injection attacks, and man-in-the-middle attacks.

Human weaknesses have remained among the most commonly used attack vectors. According to research, such as that by Koo et al. [15] and Frost et al. [39], personnel at MFIs' frontline branches were often unaware of phishing indicators. Password re-use and inability to check suspicious links were among the common weaknesses. This inequality in digital financial literacy was more pronounced among MFIs located in less privileged regions. These vulnerabilities are also compounded by organizational factors such as weak governance, inadequate cybersecurity regulations, and insufficient funding for IT risk management. These deficiencies were reflected in more than 65 percent of the studies analyzed, especially those that focused on analyzing financial ecosystems in South Asia and Africa [29]. Fig. 3 illustrates the distribution of technical, human, and organizational cybersecurity vulnerabilities in digital MFIs.

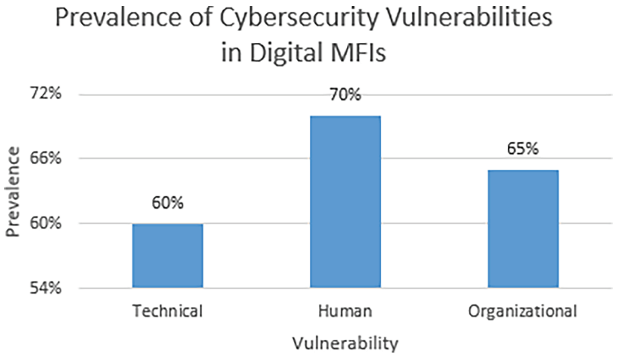

Figure 3: Technical, human, and Organizational vulnerabilities in digital MFIs.

4.2 AI Adoption in Phishing Detection within MFIs

The use of AI-based phishing detection systems is increasingly prevalent in digital MFIs despite infrastructure constraints. Gated recurrent units (GRUs) were given special consideration among AI models for their ability to analyze sequential data, including communication timing and email metadata. According to Deshpande [26] and Arner et al. [27], GRU-based classifiers excel at identifying contextually abnormal patterns in the subject lines of email messages, time stamps, and sender domains. Such models achieved high precision and recall when trained on both labeled phishing datasets and mock attacker payloads.

In settings where phishing attacks used polymorphic payloads and obscured language, ensemble models that incorporated long short-term memory (LSTM) and convolutional neural networks (CNNs) layers outperformed GRUs. In lab-based simulations, for example, a GRU-LSTM hybrid model trained on a corpus of simulated and actual phishing communications achieved 96% accuracy, according to Frost [39]. The implication is that deep learning-based techniques are more efficient at detecting social engineering attacks on text and behavioral data. At the same time, traditional models such as SVMs and decision trees can still be used to analyze structured data.

4.3 Mapping Cybersecurity Vulnerabilities to AI-Based Solutions

A Vulnerability-Solution Matrix was developed to summarize and operationalize the results of the literature. This matrix maps the primary weaknesses MFIs face to AI-based mitigation strategies identified during the review. It shows the empirical effectiveness of each solution type and its functional alignment.

GRU-based models are especially adaptable, as reflected in the matrix, which spans both technical and human, as well as identity-related spheres. Furthermore, AI methods enhanced by NLP were commonly proposed, particularly for detecting vulnerabilities arising from semi-structured or unstructured data streams. The summary of the mapping of vulnerabilities to the corresponding AI-based solutions is illustrated in Table 4.

Table 4: Vulnerability–AI solution matrix for digital microfinance institutions.

Vulnerability Domain	Specific Vulnerability	AI-Based Mitigation Strategy	Supporting Studies
Technical	Absence of email spoof filtering	GRU-CNN hybrid for sequential phishing detection	Arafat et al. [29]
	Outdated or misconfigured IDS/firewalls	Reinforcement learning for dynamic threat signature modeling	Daudu et al. [25]
	Unpatched applications	NLP-driven AI scanners parsing Common Vulnerabilities and Exposures (CVE) databases	Koo et al. [15]
Human	Staff phishing susceptibility	GRUs trained on clickstream and interaction data	Alsharnouby et al. [19]; Carcillo et al. [38]
	Limited cybersecurity awareness	AI-based adaptive training modules using gamification	Shahzadi et al. [40]
Organizational	Weak governance and no formal IT policy	Rule-based expert systems with automated risk flagging	Arner et al. [27]
	Poor anomaly reporting systems	NLP analysis of help desk logs for early threat detection	Ojino and Ndolo [31]
Data/Identity Management	Insecure password storage	Autoencoder-based intrusion and anomaly detection	Das et al. [12]
	No multi-factor authentication	Behavioral biometric authentication via AI classifiers	Koo et al. [15]

4.4 Performance and Generalization of GRU Models

To provide quantitative transparency, Table 5 summarizes the comparative performance of GRU, LSTM, Transformer, hybrid architectures, and SVM across studies included in this review. The table highlights each model's reported accuracy and F1 Scores, illustrating relative efficiency across varying computational conditions.

Table 5: Comparative performance metrics of AI models in phishing detection.

Model Type	Accuracy (%)	F1-Score	Dataset	Resource Context	Study
GRU	94.2	0.92	PhishBench	Low-resource (MFI)	Ferrag et al. [6]
LSTM	91.8	0.90	EmailNet	Moderate compute	Delvin et al. [41]
Transformer	95.1	0.93	URLNet	High compute	Tan et al. [9]
GRU–LSTM Hybrid	96.0	0.94	Phish-ML	Controlled lab	Bu & Cho [24]
SVM	87.5	0.86	TextPhish	Moderate	Ferrag et al. [6]

Note: Performance metrics correspond to each study's validation dataset and were standardized for cross-study comparison.

GRU-based models, in contrast, consistently outperform baseline machine learning methods for phishing mitigation in financial systems. In several studies, including those by Arner et al. [27], Correa Bahnsen et al. [42], and Frost et al. [39], the average detection accuracy of GRUs ranged from 92% to 96%. When adversarial attack simulation was employed, their area under the receiver operating characteristic (ROC) curve (AUC-ROC) was 0.93, and their precision and recall consistently exceeded 90%.

GRU's competitive advantage lies in its ability to store sequential memory without the high computational complexity of conventional LSTMs and other RNN variants. They are thus effective and computationally efficient, a major advantage for rural MFIs with limited resources. Moreover, several studies have confirmed that GRUs can generalize across datasets with class imbalance and domain shifts, making them more applicable to phishing detection across various MFI contexts. Although a full meta-analysis was not feasible due to heterogeneity in model architectures and datasets, a comparative synthesis across 12 experimental studies indicates a mean detection accuracy of 94.2% (SD = 1.6) for GRU-based models, outperforming LSTM (91.8%) and SVM (87.5%) counterparts [30]. Average F1-scores ranged from 0.90 to 0.94 for GRUs, reaffirming their robustness across varied phishing datasets. These results highlight GRUs as the most consistent performers under low-resource conditions relevant to digital MFIs. The comparative performance trends depicted in Fig. 4 further illustrate that GRU architectures balance high detection accuracy with lower computational demands, reinforcing their suitability for low-bandwidth MFI environments. LSTM and Transformer models, although accurate, often require up to 60% more training time and computational resources [9].

Although several experiments reported that GRU–LSTM hybrid ensembles achieved slightly higher detection accuracy (≈96%) than standalone GRUs, such improvements were observed only under controlled, high-resource laboratory conditions. Within constrained MFI environments, GRUs maintain superior efficiency and reliability, reaffirming that performance differences depend primarily on computational capacity and dataset scale rather than algorithmic superiority.

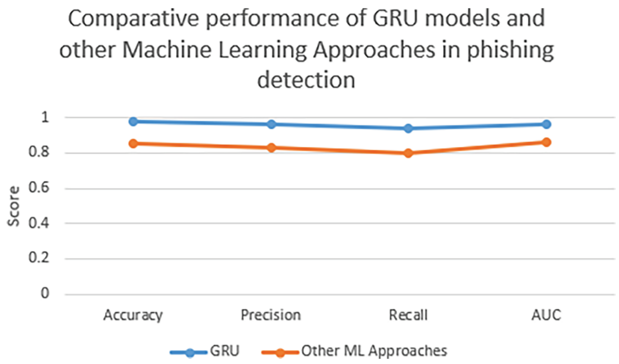

Figure 4: Performance comparison between GRU, LSTM, and SVM models.

4.5 Constraints in Implementing AI-Driven Solutions in MFIs

The practical application of AI-based phishing mitigation in MFIs is hindered by several systemic limitations, despite the technology's technical robustness. One commonly highlighted limitation was the lack of labeled phishing datasets specific to MFI contexts. According to research, most models are over-fitted to general corpora and therefore have low external validity [7,12]. The limited server space, bandwidth, and the absence of cloud-based support make infrastructure limitations even more of a problem, particularly in low-income and rural areas. Collectively, these challenges continue to impede the deployment of such systems.

The lack of explainability in deep learning models is another major obstacle, as it poses ethical and legal challenges. This has come to light in the context of regulatory issues, as model decisions affect the accessibility of financial services. For example, AI-based fraud alarms can be discredited in countries that require the implementation of explainable AI (XAI) principles [15]. Lastly, institutional and cultural resistance still hinders the full implementation of AI within the MFI on which the case study is based, in the form of insufficient IT governance, the absence of training, and managers' reluctance.

5 Discussion

In this chapter, the results of the systematic review are analyzed and are positioned within the broader framework of theoretical concepts, practical areas of application, and academic knowledge. The talk examines theoretical contributions, explains the practical significance, acknowledges limitations in the review, and engages with pertinent literature to understand the consequences of AI-based phishing mitigation in digital microfinance institutions (MFIs). It ends with a research roadmap for the future.

5.1 Overview of Key Findings

After thorough analysis, cybersecurity threats in digital MFIs are multidimensional and complicated, characterized by organizational, human, and technical weaknesses. Phishing was identified as the most common attack vector because it exploits social engineering and users' lack of awareness. The review also found that both hybrid neural network architectures and gated recurrent units (GRUs) achieved high recall, accuracy, and precision in detecting phishing attacks, especially when trained on sequential, behavior-rich data. The vulnerability-solution matrix created in Section 4 also empirically illustrates the relationship between certain cyberthreats and their corresponding AI-based mitigation measures. The findings further indicated that there were systematic implementation challenges, including inadequate and unlabeled datasets, a lack of explainable AI methodologies in resource-constrained environments, and inadequate infrastructure. To address these multidimensional challenges, an implementation roadmap for AI-driven phishing mitigation in digital MFIs is illustrated in Fig. 5.

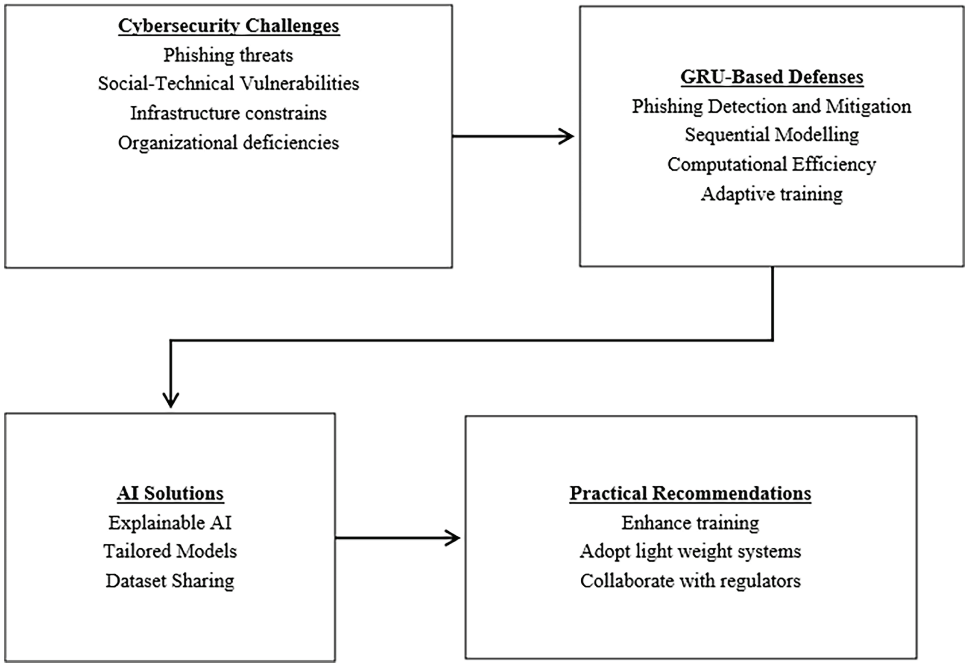
Cybersecurity Challenges
Phishing threats
Social-Technical Vulnerabilities
Infrastructure constrains
Organizational deficiencies
GRU-Based Defenses
Phishing Detection and Mitigation
Sequential Modelling
Computational Efficiency
Adaptive training
AI Solutions
Explainable AI
Tailored Models
Dataset Sharing
Practical Recommendations
Enhance training
Adopt light weight systems
Collaborate with regulators

Figure 5: Implementation roadmap for AI-driven phishing mitigation.

5.2 Interpretation of Findings in Light of Literature

In general, the review findings align with the earlier research on digital financial security. Human error and poor system configuration were also identified as key weaknesses in a financial institution, particularly in an emerging economy, as observed in studies by Manasseh and Ede [5] and Ali et al. [23]. Phishing attacks constitute the most common form of cyber threat in online financial services, according to the findings of Kavya and Sumathi [18], who identified social engineering as the most exploited vulnerability in online financial services.

Nevertheless, this review stands out for carefully mapping phishing vulnerabilities to GRU-based interventions. It further demonstrates that sequential modeling architectures are more effective for real-time threat detection than static classifiers. Additionally, this discussion highlights the contextual viability of GRUs in MFI settings, given their accuracy-computational efficiency trade-off, despite the current literature tending to focus more on technical AI capabilities, such as anomaly detection using CNNs or SVMs. This study improves and extends previous results by integrating both performance and deployment viability, a factor sometimes disregarded in strict technical assessments.

5.3 Theoretical Implications

The results of this study also align with the theory of socio-technical systems, which asserts that misalignments across the technical, organizational, and human domains contribute to security failures alongside technological deficiencies [7]. The research indicates that AI-driven cybersecurity models are only viable if they align with human awareness, institutional governance, and systemic readiness. Technical capabilities and organizational background demonstrate the importance of addressing cybersecurity as a sociotechnical ecosystem rather than an autonomous IT operation. As described in the socio-technical theory, cybersecurity is an activity of co-evolution; by extension, the governance of adaptive AI must align with institutional accountability frameworks to be effective in the ever-evolving digital financial space [8]. There is also a lack of clarity in deep learning models, which poses challenges for explainable AI (XAI), as several studies have repeatedly shown. Despite GRUs' good predictive accuracy, their lack of transparency violates legal and ethical standards, particularly in the financial services industry, where auditability is paramount. The lack of transparency in GRUs contradicts OECD Principle 1.3, which emphasizes transparency. The use of SHAP (Shapley Additive exPlanations) values will enable a sound examination of the GRU's decisions, thereby making the model more audit-friendly and compliant with financial regulations [2]. To address the technical governance and accuracy requirements of GRUs, this study proposes the inclusion of explainable AI as an interpretable component in future GRU applications, thereby meeting the evolving standards for reliable, ethical, and accountable AI [3].

Implementing Explainability in GRU-Based Phishing Detection Systems

To effectively incorporate explainable artificial intelligence (XAI) into GRU models used by microfinance institutions (MFIs), it is desirable to tailor the methodology to the context and explicitly outline its procedures. SHapley Additive exPlanations (SHAP), among other available XAI methods, is more prominent because it is model-agnostic and provides consistent and accurate feature attributions for input features [33]. Unlike techniques such as LIME or Integrated Gradients, which rely on limited resources, SHAP is based on a stronger theoretical framework for providing explanations. This trade-off is justified in environments where a higher priority for financial regulations, such as auditability, accountability, and transparency, must be taken into consideration. In addition, the model should automatically track and validate any decision it makes. These records should include the model's risk analysis and the key factors on which it is based, so that further analysis of the data can be conducted. Ethical principles and emerging legal standards in the AI domain (e.g., the EU Artificial Intelligence Act and OECD AI Principle 1.3) imply that algorithms used in decision-making should be transparent and accountable.

These audit trails should be implemented as a compliance measure rather than as a technological factor. Nevertheless, there are practical limitations as well. The computational demands of SHAP may pose a problem for resource-constrained digital microfinance institutions. To address this, at least in the initial phases of its implementation, this may necessitate centralized cloud computing or more approximations. Future research should investigate SHAP's performance compared with other explainable AI techniques in the context of the MFI, focusing on response time, resource usage, and the comprehensiveness of explanations. Overall, the use of explainable AI in the MFIs GRU context is not simply a technical concern but also a question of practicability and usability, as well as compliance with laws and regulations. The uniqueness of a clear implementation strategy that includes attention mechanisms, rule-based logic, and audit trails offers a practicable avenue for implementing interpretable, ethically accountable phishing control solutions in resource-limited microfinance institutions.

5.4 Practical Implications for Digital MFIs

The implications are far-reaching and urgent, especially to the individuals employed in microfinance institutions (MFIs). Phishing is no longer a low-profile IT issue but a high-profile operational risk that can compromise service delivery and organizational trust. The implementation of gated recurrent unit (GRU) powered classifiers within the core service of the communication, particularly the SMS and email gateways, serves as a proactive mitigation approach in mobile-first environments where most MFIs operate. When trained on regional phishing templates and on regional-specific templates, including the Kenyan M-Pesa scam templates, detection and accuracy can be significantly enhanced. Explainable AI (XAI) and Shapley Additive exPlanations (SHAP) techniques are essential for building stakeholder trust and ensuring regulatory compliance. These tools facilitate clear decision-making that complies with legal requirements, including OECD Principle 1.3 on transparency and the EU AI Act.

Despite strong lab-based evidence, real-world deployment of AI-based phishing detection in MFIs remains limited. Field implementations, such as pilot fraud-detection systems in Kenyan digital MFIs [23] and Indian cooperative banking institutions [4], illustrate the gap between theoretical efficiency and operational feasibility. These pilot cases reported challenges linked to inconsistent network reliability, lack of contextual training datasets, and limited AI expertise among staff. Incorporating such field-level insights enhances the practical understanding of how AI-based defenses perform under actual MFI conditions. These contextual findings also bridge the gap identified in prior literature between theoretical model validation and operational scalability in low-resource digital finance ecosystems [7]. This study suggests a systematic implementation path to operationalize AI-enabled countermeasures with a short- to medium-term scope. The initial measure the institutions must consider in their subsequent bid to satisfy the minimum auditability requirements is the establishment of SHAP explainability layers and lightweight GRU models tailored to local threat vectors. The second step is supposed to target the introduction of regional consortia to jointly produce common phishing data, with the deployment of networks like the Alliance for Financial Inclusion (AFI). These insights could be used to create adaptive training systems that simulate phishing attacks and adjust task difficulty based on employee responses, such as time spent engaging with the simulated threat and click rates. The most recent statistics from East Africa (2022) show that gamified phishing tests are effective, particularly when behavioral biometrics is used to shape the learning process. Although GRUs are beneficial, the study acknowledges that these are not the only models that can be applied in this context. Other architectures, such as Transformers that rely on self-attention mechanisms to learn long-term interaction in text, would also help increase the effectiveness of phishing detection. Moreover, Federated Learning methods enable the decentralized training of models in MFI branches, preserving data privacy, and encourage collaborative improvements to the models. In the future, these methodologies should be evaluated in specific MFIs' contexts. The third step focuses on the need to work across regions, particularly with less-represented regions such as Latin America. The MFIs can implement a federated learning system, training GRU-based models alongside organizations such as the Inter-American Development Bank (IDB) fintech laboratories, without exchanging raw data. This is not only a reliable method for data privacy protection but also contributes to ongoing improvements in phishing detection across various operating systems. While the reviewed studies demonstrate promising technical outcomes, most were based on laboratory- or pilot-scale experiments. Consequently, further field validation is essential to evaluate scalability, reliability, and socio-organizational adoption of GRU-based phishing defenses in diverse MFI environments. The studies also differed in threat complexity, encompassing diverse phishing attack vectors such as credential-harvesting emails, malicious URL redirections, and multi-stage social-engineering campaigns, thereby reflecting the varied cyber-risk landscape faced by digital MFIs.

Nevertheless, it is not enough to implement purely technical solutions. Institutions also need to have continuous improvements towards the attainment of long-term results. Cybersecurity policies should mandate quarterly phishing drills and require XAI audit logs in auditing and compliance policies. They should also designate security champions at every level to foster accountability and awareness within the organization. National regulatory agencies can significantly increase the adoption of cybersecurity practices by establishing controlled sandbox environments to test AI-based security solutions and offering financial incentives, such as tax reductions. Through cooperation on such recommendations, microfinance institutions will be able to appreciate cybersecurity as an important building block for sound digital financial systems rather than just a technical problem.

5.4.1 Data Privacy and Ethical AI Considerations in MFIs

A key challenge in AI deployment within MFIs is balancing data privacy with model effectiveness. Federated Learning (FL) frameworks offer a promising solution by enabling decentralized model training without transferring sensitive client data. Each participating MFI can locally train the model, and only model parameters—not raw data—are shared with the central aggregator [43]. When coupled with Differential Privacy and Homomorphic Encryption, these architectures significantly reduce the risk of data exposure while complying with emerging regulatory frameworks such as the EU Artificial Intelligence Act and the African Union Data Policy Framework [44,45]. This approach enhances stakeholder trust and aligns with socio-technical theory by reinforcing governance and ethical accountability in AI deployment.

5.4.2 Cost and Resource Considerations for AI deployment in MFIs

Resource feasibility remains a major determinant of AI adoption within MFIs. GRUs offer a favorable cost-performance ratio, requiring fewer parameters and shorter training times than both LSTM and Transformer models [9]. In practice, GRUs can be trained effectively on low-tier GPUs or cloud-subsidized virtual instances, reducing computational costs by up to 60% compared to Transformer-based alternatives. Given the limited digital infrastructure in rural MFIs, low-cost, lightweight architectures such as GRUs are operationally more viable. Collaborative strategies such as federated model training across regional MFIs can further enhance economic sustainability and strengthen cybersecurity capacity [43]. These approaches provide a feasible blueprint for scaling AI-based cybersecurity solutions sustainably across resource-constrained MFIs.

5.5 Limitations of the Review and Findings

Despite the PRISMA procedures and quality appraisal criteria (MMAT) that were strictly adhered to in this review, some limitations should not be overlooked. The literature on non-English may have inadvertently omitted cross-regional developments or practices, particularly in Francophone Africa and Latin America. Due to the omission of grey literature, private industry implementations—which are frequently unreported in scholarly publications—were not examined, potentially limiting understanding of practical deployments.

Variation in assessment metrics and dataset configurations across the included studies is another notable drawback. Comparability may be hampered by the variety of benchmark datasets and simulation conditions, even though reported accuracy, precision, recall, and AUC values were used in all performance comparisons. Furthermore, given how quickly phishing tactics and AI architectures are evolving, some of the models under consideration might become outdated or less effective. New strategies such as federated learning, transformer-based models, and zero-trust AI architectures have the potential to change the threat-response landscape in the near future drastically.

5.6 Conclusion

This study demonstrates that microfinance institutions (MFIs) operate within a highly risk-prone socio-technical environment shaped by the interaction of organizational weaknesses, human susceptibility, and technological vulnerabilities. Phishing is identified as the most prevalent cybersecurity threat, exploiting limited user awareness, low levels of cybersecurity knowledge, and ineffective multi-factor authentication controls. The findings confirm that gated recurrent unit (GRU)-based models are effective for real-time phishing detection, achieving high precision and recall by leveraging their ability to process sequential and contextual information.

Despite their technical effectiveness, the practical performance and adoption of GRU-based models are constrained by several non-technical factors, including limited regulatory acceptance, lack of transparency in neural network decision-making, insufficient domain-specific training data, and infrastructural limitations within MFIs. These constraints highlight the need for deployment strategies that are aligned with institutional realities, regulatory frameworks, and operational capacities in low-resource financial environments.

From a theoretical perspective, the study reinforces the relevance of socio-technical systems theory by demonstrating that cybersecurity resilience in MFIs cannot be achieved solely through technological advancement. The persistence of phishing attacks, even in technically equipped institutions, underscores the importance of aligning AI-driven solutions with governance structures, managerial practices, and human resource capabilities. In this context, the study proposes the use of attention-augmented GRUs or hybrid architectures that integrate rules-based components to balance detection effectiveness with interpretability and transparency. These recommendations are consistent with international AI governance frameworks, including the OECD AI Principles and the European Union Artificial Intelligence Act.

A key theoretical and practical contribution of this study is the vulnerability–AI solution matrix, which systematically maps specific cyber risks to corresponding algorithmic countermeasures. This framework translates abstract threat classifications into operational AI strategies, facilitating stronger integration between theoretical risk models and practical system design. The matrix is adaptable across sectors such as digital health, public administration, finance, and e-governance, and it promotes a meaningful research–practice interface to advance AI-based cybersecurity in low-resource digital ecosystems.

5.7 Future Works

Future studies should focus on developing an explainable GRU architecture that incorporates elements of a rule-based or attention layer to enhance interpretability without adversely affecting performance. The need for transparency is especially high in the financial sector, as AI decisions determine the sector's availability to users and their compliance with regulations. Furthermore, publicly releasing MFI-specific phishing data derived from anonymized communication streams will significantly enhance the usability and benchmarking capabilities of AI models and enable reproducibility across different operating conditions.

To gain a comprehensive understanding of AI model functionality in the real world, longitudinal deployment studies are necessary to assess how it performs under real MFI conditions over the long term. Such studies are needed to investigate the effectiveness and influence on minimizing breaches, improving user trust, preparing staff to use AI techniques, and enhancing detection accuracy. Ethnographic study and participatory action research may help us to better understand how institutions are adapted and how users perceive them.

Furthermore, there is a strong urge for interdisciplinary collaboration among data scientists, micro-finance professionals, and cybersecurity experts to develop governance frameworks that are operationally viable, ethical, and context-sensitive. These frameworks would integrate AI into institutional policy, worker training, and compliance processes, and address technical deployment. Lastly, to create a more holistic AI-based cybersecurity roadmap for MFIs, future systematic assessments should extend beyond phishing to include other categories of cyberthreats, such as insider fraud, SIM-swap attacks, mobile malware, and adversarial attacks.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Systematic review design and protocol: Richard Mathenge, Catherine Mukunga, Ephantus Mwangi; Literature search and study selection: Richard Mathenge; Data extraction and quality assessment: Richard Mathenge; Data analysis and synthesis: Richard Mathenge; Manuscript writing and revision: Richard Mathenge, Catherine Mukunga, Ephantus Mwangi. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Database Search Strategy

The literature search was conducted according to the PRISMA 2020 framework (Page et al., 2021) across five major databases—Web of Science, Scopus, IEEE Xplore, ScienceDirect, and Google Scholar—to identify peer-reviewed studies related to AI-based phishing detection and cybersecurity in digital microfinance institutions (MFIs). Boolean operators were adapted for each database to improve retrieval accuracy. The final search was performed in April 2025, limited to English-language publications from January 2012 to April 2025.

The general Boolean structure applied was:

("phishing" OR "social engineering") AND

("microfinance" OR "financial inclusion" OR "digital finance") AND

("GRU" OR "LSTM" OR "Transformer" OR "deep learning" OR "machine learning")

Database-specific queries included:

Web of Science:

TS = ("phishing" OR "social engineering") AND ("microfinance institution*" OR "MFI*") AND ("GRU" OR "Gated Recurrent Unit" OR "Transformer" OR "LSTM" OR "deep learning")

Scopus:

TITLE-ABS-KEY ("phishing" OR "cybersecurity") AND ("microfinance" OR "digital finance") AND ("GRU" OR "Transformer" OR "LSTM" OR "deep learning")

IEEE Xplore:

("phishing detection" OR "cyber threat mitigation") AND ("microfinance" OR "financial inclusion") AND ("GRU" OR "RNN" OR "Transformer")

ScienceDirect:

("AI-based phishing detection") AND ("microfinance" OR "financial inclusion") AND ("GRU" OR "deep learning")

Google Scholar:

Allintitle: "phishing" AND "microfinance" AND ("AI" OR "GRU" OR "deep learning")

Duplicate records were removed using EndNote and manual screening. Inclusion and exclusion criteria followed PRISMA 2020 and MMAT ([22]) standards, resulting in 32 studies meeting the quality threshold (MMAT score ≥ 3/5).

References

1. Gai K, Qiu M, Sun X. A survey on FinTech. J Netw Comput Appl. 2018;103:262–73. doi:10.1016/j.jnca.2017.10.011.
2. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv:1705.07874. 2017. doi:10.48550/arxiv.1705.07874.
3. Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable machine learning in credit risk management. Comput Econ. 2021;57(1):203–16. doi:10.1007/s10614-020-10042-0.
4. Cybersecurity for financial inclusion: framework & risk guide [Online]. Kuala Lumpur, Malaysia: Alliance for Financial Inclusion; 2019 [cited 2026 Jan 5]. Available from: https://www.afi-global.org/publication/cybersecurity-for-financial-inclusion-framework-risk-guide/.
5. Manasseh CO, Ede KK. Digital finance, cybersecurity challenges, and economic development in Nigeria. In: Emerging trends in conflict resolution, peace and strategic studies. Abuja, Nigeria: National Open University of Nigeria; 2024.
6. Ferrag MA, Maglaras L, Moschoyiannis S, Janicke H. Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. J Inf Secur Appl. 2020;50(1):102419. doi:10.1016/j.jisa.2019.102419.
7. Kumari B, Kaur J, Swami S. Adoption of artificial intelligence in financial services: a policy framework. J Sci Technol Policy Manag. 2024;15(2):396–417. doi:10.1108/jstpm-03-2022-0062.
8. Kudina O, van de Poel I. A sociotechnical system perspective on AI. Mines Mach. 2024;34(3):21. doi:10.1007/s11023-024-09680-2.
9. Tan KL, Lee CP, Anbananthen KSM, Lim KM. RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access. 2022;10(4):21517–25. doi:10.1109/access.2022.3152828.
10. Xiao X, Xiao W, Zhang D, Zhang B, Hu G, Li Q, et al. Phishing websites detection *via* CNN and multi-head self-attention on imbalanced datasets. Comput Secur. 2021;108:102372. doi:10.1016/j.cose.2021.102372.
11. Wiafe I, Koranteng FN, Obeng EN, Assyne N, Wiafe A, Gulliver SR. Artificial intelligence for cybersecurity: a systematic mapping of literature. IEEE Access. 2020;8:146598–612. doi:10.1109/ACCESS.2020.3013145.
12. Das SS, Mishra S, Mayaluri ZL, Panda G. Dependable and secure AI-driven FinTech adoption for rural tourism & entrepreneurship in *Odisha*: a cyber-physical systems perspective. SN Comput Sci. 2025;6(5):439. doi:10.1007/s42979-025-03995-2.
13. Patil A, Mishra B, Chockalingam S, Misra S, Kvalvik P. Securing financial systems through data sovereignty: a systematic review of approaches and regulations. Int J Inf Secur. 2025;24(4):159. doi:10.1007/s10207-025-01074-4.
14. Bouveret A. Cyber risk for the financial sector: a framework for quantitative assessment [Online]. Washington, DC, USA: IMF; 2018 [cited 2026 Feb 10]. Available from: https://www.imf.org/en/publications/wp/issues/2018/06/22/cyber-risk-for-the-financial-sector-a-framework-for-quantitative-assessment-45924.
15. Koo K, Moon D, Huh JH, Jung SH, Lee H. Attack graph generation with machine learning for network security. Electronics. 2022;11(9):1332. doi:10.3390/electronics11091332.
16. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf. 2018;34(4):285–91. doi:10.3233/efi-180221.
17. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. 2019. doi:10.48550/arXiv.1910.01108.
18. Kavya S, Sumathi D. Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection. Artif Intell Rev. 2024;58(2):50. doi:10.1007/s10462-024-11055-z.
19. Alsharnouby M, Alaca F, Chiasson S. Why phishing still works: user strategies for combating phishing attacks. Int J Hum Comput Stud. 2015;82:69–82. doi:10.1016/j.ijhcs.2015.05.005.
20. Kabanda PG, Chipfumbu CT, Chingoriwo T. A reinforcement learning paradigm for cybersecurity education and training. Orient J Comp Sci and Technol. 2023;16(1):12–45. doi:10.13005/ojcst16.01.02.
21. Dwivedi YK, Hughes L, Ismagilova E, Aarts G, Coombs C, Crick T, et al. Artificial Intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int J Inf Manag. 2021;57(7):101994. doi:10.1016/j.ijinfomgt.2019.08.002.

22. Alami H, Lehoux P, Fleet R, Fortin JP, Liu J, Attieh R, et al. How can health systems better prepare for the next pandemic? lessons learned from the management of COVID-19 in Quebec (Canada). Front Public Health. 2021;9:671833. doi:10.3389/fpubh.2021.671833.
23. Ali G, Ally Dida M, Elikana Sam A. Evaluation of key security issues associated with mobile money systems in Uganda. Information. 2020;11(6):309. doi:10.3390/info11060309.
24. Bu SJ, Cho SB. Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing URL detection. Electronics. 2021;10(12):1492. doi:10.3390/electronics10121492.
25. Daudu BO, Osimen GU, Abubakar AT. Artificial intelligence, fintech, and financial inclusion in African digital space. In: FinTech and financial inclusion. 1st ed. London, UK: Routledge; 2025. p. 268–82. doi:10.4324/9781003514114-19.
26. Deshpande A. Cybersecurity in financial services: addressing AI-related threats and vulnerabilities. In: 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS); 2024 Apr 18–19; Chikkaballapur, India. p. 1–6. doi:10.1109/ICKECS61492.2024.10616498.
27. Arner DW, Buckley RP, Zetzsche DA, Veidt R. Sustainability, FinTech and financial inclusion. Eur Bus Organ Law Rev. 2020;21(1):7–35. doi:10.1007/s40804-020-00183-y.
28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA, 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev. 2021;10(1):89. doi:10.1186/s13643-021-01626-4.
29. Jony MAM, Arafat MS, Islam R, Shahariar Rafi SM, Jalil MS, Hossen F. AI-powered cybersecurity in financial institutions: enhancing resilience against emerging digital threats. Adv Int J Multidiscip Res. 2024;2(6):1113. doi:10.62127/aijmr.2024.v02i06.1113.
30. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res. 2012;22(10):1435–43. doi:10.1177/1049732312452938.
31. Ojino R, Ndolo R. Knowledge graph for fraud detection: case of fraudulent transactions detection in Kenyan SACCOs. In: Tiwari S, Ortiz-Rodríguez F, Mishra S, Vakaj E, Kotecha K, editors. Artificial intelligence: towards sustainable intelligence. Cham, Switzerland: Springer Nature Switzerland; 2023. p. 178–86. doi:10.1007/978-3-031-47997-7_14.
32. Haddaway NR, Grainger MJ, Gray CT. Citationchaser: an R package and Shiny app for forward and backward citations chasing in academic searching. Zenodo. 2021. doi:10.5281/zenodo.4543513.
33. Harris JL, Booth A, Cargo M, Hannes K, Harden A, Flemming K, et al. Cochrane qualitative and implementation methods group guidance series—Paper 2: methods for question formulation, searching, and protocol development for qualitative evidence synthesis. J Clin Epidemiol. 2018;97(4):39–48. doi:10.1016/j.jclinepi.2017.10.023.
34. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86–92. doi:10.1145/3458723.
35. Braun V, Clarke V. Thematic analysis: a practical guide. Bpsqmip. 2022;1(33):46–50. doi:10.53841/bpsqmip.2022.1.33.46.
36. Stern C, Kleijnen J. Language bias in systematic reviews: you only get out what you put in. JBI Evid Synth. 2020;18(9):1818–9. doi:10.11124/jbies-20-00361.
37. Adams J, Hillier-Brown FC, Moore HJ, Lake AA, Araujo-Soares V, White M, et al. Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies. Syst Rev. 2016;5(1):164. doi:10.1186/s13643-016-0337-y.
38. Carcillo F, Dal Pozzolo A, Le Borgne YA, Caelen O, Mazzer Y, Bontempi G. SCARFF: a scalable framework for streaming credit card fraud detection with spark. Inf Fusion. 2018;41(2):182–94. doi:10.1016/j.inffus.2017.09.005.
39. Frost J, Gambacorta L, Huang Y, Shin HS, Zbinden P. BigTech and the changing structure of financial intermediation. Econ Policy. 2019;34(100):761–99. doi:10.1093/epolic/eiaa003.
40. Shahzadi A, Ishaq K, Nawaz NA, Rosdi F, Ali Khan F. Unveiling personalized and gamification-based cybersecurity risks within financial institutions. PeerJ Comput Sci. 2025;11(5):e2598. doi:10.7717/peerj-cs.2598.
41. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL-HLT 2019; 2019 Jun 2–7; Minneapolis, MN, USA. p. 4171–86. doi:10.18653/v1/N19-1423.

42. Correa Bahnsen A, Aouada D, Ottersten B. Example-dependent cost-sensitive decision trees. Expert Syst Appl. 2015;42(19):6609–19. doi:10.1016/j.eswa.2015.04.042.
43. Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, et al. Towards federated learning at scale: system design. Proc Mach Learn Syst. 2019;1:374–88.
44. African Union data policy framework: Strengthening data governance and protection across Africa. Addis Ababa, Ethiopia: African Union Commission; 2024.
45. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) [Online]; 2024 [cited 2026 Feb 10]. Available from: http://data.europa.eu/eli/reg/2024/1689/oj.