AFRICAN JOURNAL OF SCIENCE, TECHNOLOGY AND ENGINEERING (AJSTE)

Volume 7, Issue 2 2026

KIRINYAGA UNIVERSITY

KIRINYAGA UNIVERSITY

AFRICAN JOURNAL OF SCIENCE, TECHNOLOGY AND ENGINEERING (AJSTE)

Volume 7(2) 2026

KIRINYAGA UNIVERSITY, KENYA

ISO 9001: 2015 Certified

AFRICAN JOURNAL OF SCIENCE, TECHNOLOGY AND ENGINEERING (AJSTE)

Editorial Board

Editors

Prof. Charles Omwandho, PhD

Dr. Jotham Wasike, PhD

Advisory Board

Prof. Mary Ndungu, PhD

Prof. Charles Omwandho, PhD

Administration & Logistics

George Ngorobi

Nelly Ann Kathomi

Francis Kamau

© Copyright 2026, Kirinyaga University

Copyright Statement

All rights reserved. Seek the University permission to reproduce, distribute, display or make derivative content or modification.

AFRICAN JOURNAL OF SCIENCE, TECHNOLOGY AND ENGINEERING (AJSTE)

Email: journals@kyu.ac.ke

ISSN 1356-6282

KIRINYAGA UNIVERSITY

P.O Box 10500

NAIROBI, KENYA

Foreword

It is with great pleasure that I present Issue 2 of Volume 7 of the African Journal of Science, Technology and Engineering. This issue marks another significant step forward in our mission to foster innovative research and knowledge dissemination across the African continent and beyond.

This edition features peer-reviewed articles addressing current and emerging issues in the application of machine learning and artificial intelligence, from strengthening the cyber resilience of information systems to improving crop productivity among smallholder farmers, and from semantic web technologies to text mining for library knowledge management. The diversity and depth of contributions reflect the growing capacity and commitment of African scholars and practitioners in addressing complex scientific and technological challenges through locally relevant and globally informed solutions.

As Africa continues to assert its place in the global knowledge economy, it is vital that platforms like this journal provide visibility and support for homegrown research and innovations. This issue not only celebrates intellectual achievement but also encourages collaboration, cross-disciplinary engagement, and practical application of science and technology to improve lives.

I commend the authors, reviewers, editorial board, and supporting institutions whose dedication and scholarly rigor have made realization of this issue possible. May this edition inspire further inquiry, innovation, and impactful solutions throughout Africa and the global scientific community.

Chief Editor

TABLE OF CONTENTS

ENHANCING THE CYBER RESILIENCE OF INFORMATION SYSTEMS THROUGH MACHINE LEARNING.....1

ARTIFICIAL INTELLIGENCE FOR SUSTAINABLE AGRICULTURE: ENHANCING CROP PRODUCTIVITY THROUGH MACHINE LEARNING IN KENYA.....16

SEMANTIC WEB AND LINKED DATA TECHNOLOGIES FOR NEXT-GENERATION LIBRARY KNOWLEDGE MANAGEMENT 30

TEXT MINING AND TOPIC MODELLING FOR KNOWLEDGE DISCOVERY IN DIGITAL COLLECTIONS.....49

ENHANCING THE CYBER RESILIENCE OF INFORMATION SYSTEMS THROUGH MACHINE LEARNING

Ndung'u H.,[1] Munyao V.,[1] Karani J.,[2] Wasike J.[1]

[1]Kirinyaga University, Kenya

[2]Karatina University, Kenya

Correspondence: Hndungukp15@gmail.com

ABSTRACT

Organizations today operate in an environment of relentless cyber threats that jeopardize continuity of critical services. Conventional security approaches, primarily focused on preventing intrusions, have proven insufficient. This shortfall has shifted attention to cyber resilience, defined as organizations' capacity to maintain operations and recover effectively after cyber disruptions. Among the technologies driving resilience, Machine Learning (ML) stands out for its ability to analyze vast datasets, identify anomalies, and respond effectively to evolving threats. Its ability to process vast and complex datasets enables detection of subtle, hidden patterns that conventional tools or human analysts might overlook. To examine ML's role in strengthening cyber resilience, this study undertook a systematic review of recent literature (2020–2026). The review encompasses 87 peer-reviewed studies from reputable sources, with the aim of mapping ML applications across different stages of resilience and identifying recurring benefits and challenges. The goal is to provide a clear understanding of how ML contributes to resilience, evaluate its performance in practice, and identify the key challenges that must be overcome to ensure its safe and robust adoption. Results show that ML significantly enhances resilience by enabling proactive threat detection and adaptive response mechanisms. However, challenges such as data quality, model robustness, and the need for human oversight limit its effectiveness. The evidence suggests that ML is not a standalone solution but rather an enabling technology whose success depends on thoughtful design, integration with human expertise, and mitigation of inherent limitations. This paper contributes a structured framework for understanding ML's role in cyber resilience and highlights critical issues that warrant further investigation, offering a foundation for both practitioners and researchers to advance resilient cybersecurity strategies.

Keywords: Machine Learning; Cyber Resilience; Cybersecurity; Intrusion Detection; Threat Detection; Incident Response; Anomaly Detection; Adaptive Cyber Defense; Information Systems; Adversarial Machine Learning.

INTRODUCTION

The increasing reliance of modern society on digital infrastructures—spanning finance, healthcare, and energy—has expanded the attack surface for cyber adversaries. Traditional security models, built on static rules and signature-based detection, struggle to counter today's sophisticated and adaptive threats. These include polymorphic ransomware capable of altering its appearance, long-term targeted campaigns backed by substantial resources, and exploits of previously unknown software vulnerabilities[1]. In response, the focus of cybersecurity has shifted from prevention alone to cyber resilience based on systems that can anticipate attacks, withstand stress, recover rapidly from disruptions, and continuously learn to strengthen future defenses. The central objective is to ensure continuity of essential services even under active attack [2].

Within this paradigm, Machine Learning (ML) has emerged as a transformative technology. Its ability to process vast and complex datasets enables detection of subtle, hidden patterns that conventional tools or human analysts might overlook. This capability enables ML to power systems that can forecast emerging threats, identify anomalous behavior in real time, and automate elements of incident response to outpace human reaction times. (Radanliev et al., 2024) [3] observed that by shifting security operations from a reactive stance to a proactive and adaptive posture, ML is increasingly recognized not merely as a supporting tool but as a foundational component in building resilient and intelligent information systems.

Problem Statement

Although the promise of ML in the area of cybersecurity is evident, its implementation within a resilience approach poses significant challenges. Thus, scattered data, outdated technology not built for ML, and a lack of competent personnel in data science and security are common problems in organizations. Equally worrying is the fact that the ML models themselves may be exposed. Hammad et al., (2024)[2] asserted that attackers may craft inputs in a way that deceives the models, corrupts the data with which they are trained, or exploit bias in the design, making the technology that is supposed to make systems more resilient ironically vulnerable to attack.

Given these challenges, there is an urgent need to move beyond isolated, small-scale experiments to a systematic approach that enables understanding of how diverse Machine Learning (ML) techniques can be applied securely and effectively across the full lifecycle of cyber incident management—spanning preparation, response, and recovery [4].

A critical gap persists in the form of limited, evidence-based guidance on how organizations should select, implement, and govern ML systems to both strengthen resilience and mitigate associated risks [5]. This study seeks to address that gap by conducting a comprehensive analysis of recent studies with a view to providing a clear understanding of how ML contributes to resilience, evaluating its performance in practice, and identifying the key challenges that must be overcome to ensure its safe and robust adoption [6].

Objective

This study sought to;

1. analyze the specific contributions of ML to each stage of the cyber resilience lifecycle.
2. evaluate performance of different ML approaches using synthesized evidence from recent literature.
3. Identify and critically discuss the technical, operational, and ethical challenges in deploying ML for resilience
4. Propose an integrated framework and future research directions to realize ML's full potential in building resilient information systems.

Scope of the Study

The study is a systematic review of recent literature, industry reports, and real-world case studies published from 2020 to 2026. The review analyzes the use of ML in major areas of cyber resilience, i.e., detection of intrusions in the network and system, user behavior analysis to identify threats, prediction of attacks, and automated actions to deal with incidents and recovery. The analysis covers applications in various key sectors, including Industrial Control Systems for infrastructure such as power grids, medical infrastructure and medical devices, financial systems, and international supply chains [7]. Notably, the review is not limited to

evaluating technical performance, i.e., the accuracy with which a model would identify a threat, but also explores practical issues such as the challenge of adopting ML to older systems, the necessity of models that a human can comprehend and feel confident about, and organizational issues that predetermine the success or failure of these technologies. This methodology will help produce not only well-grounded academic results but also those that are actually applicable to professionals to create more resilient systems [6].

ML as a Foundational Pillar for Cyber Resilience

Existing literature does not perceive machine learning as a new tool of cybersecurity but rather as a primary element in developing resilient systems. This change fits into a contemporary concept of cyber resilience, which goes beyond merely not being hacked or being hacked. The four major capacities of resilience are described by modern frameworks as the 4Rs: Robustness, Rapid Recovery, Resourcefulness, and Redundancy. ML offers tangible, scientific methods of meeting all these objectives. Robustness- the inherent strength of any system is enhanced when ML actively hunts for any latent vulnerability within software or network environments before the attacker can detect it. ML helps teams to find actual incidents and react instantly, hence reducing recovery time significantly; rapid recovery is accelerated when the first sorting and analysis of the security alerts are automated. Resourcefulness, or solving new problems in times of crisis, is increased when ML assists in prioritization of actions and use of limited personnel in the optimal way during an attack [2].

Lastly, ML gives the chance to be smarter about Redundancy. It allows simulating various failure scenarios and providing intelligent backup plans instead of simply copying the systems at a high cost, keeping the services online and providing better services [10]. Notably, the four principles of resilience cannot occur without the role of the brain, which is played by ML in the current complicated digital contexts, as opined by [8].

How ML Enhances Resilience: Key Mechanisms

Research points to four main ways ML strengthens cyber resilience. **Anticipating threats before they hit.** ML models have the potential of transforming security to proactive. ML is capable of detecting patterns by studying the historical data on attacks, global threat reports,

and even chatter on hacker forums that indicate an impending attack. This enables defenders to build defenses and address weaknesses in advance, which is the "anticipate" aspect of resilience [3].

With Anomaly Detection, the Unknown can be detected. In case of new attacks that have previously unknown signatures, ML applies unsupervised learning. Methods such as Autoencoders get to know what normal network or user behavior is like [9]. They, in turn, raise red flags over a major anomaly to this trend, which may imply a new threat or a rogue insider attack. This plays a vital role in the detection of attacks that cannot be detected by traditional tools.

Response with Reinforcement Learning Automation. Reinforcement Learning (RL) can be used to enhance the response stage. In this case, an AI agent works in a free, simulated environment, and it is taught by trial and error which actions (such as isolating a device or blocking an IP address) result in the most desirable behavior to contain the threat [1]. As time passes, it learns the best response policies, which enable a rapid, automated response that decreases the damage window.

Lessons on Incidents to Improve Recovery. ML assists in streamlining recoveries after an attack and can prioritize the systems restoration to limit disruption of the business [10]. What is more important, ML examines the incident in full to study the lessons: What went wrong in the defense? What are the differences in the approaches? This forms a feedback loop where the whole security system would become stronger and more adapted to completing the cycle of recovery to better preparedness.

ML in Action: Sector-Specific Applications

Use of ML mechanisms varies significantly across different critical industries. **Securing Critical Infrastructure**. In the case of systems such as power grids or water treatment plants, ML models are optimally created to identify attacks with a goal of causing physical damage. They can detect False Data Injection, where attackers provide operators with artificial sensor data so that they can cause disastrous failures. ML assists in differentiating between genuine

sensor data and the interfered data, which keeps the integrity of these essential systems intact [11].

Protecting the Healthcare Systems. Healthcare should strike a balance between high security and stringent laws of patient privacy [7]. In this case, Federated Learning is a prospective solution. It enables training of ML models with data directly obtained by the medical device (such as a heart monitor) without ever leaving the hospital, and can be used to detect damaged equipment without violating privacy.

Fighting Financial Fraud. In finance, resilience refers to preventing real-time fraud. This is centered on ML, which examines millions of transactions with an aim of identifying patterns associated with credit card fraud or money laundering. Such systems keep on learning every new fraud type, and banks can modify their system in real time to keep the criminals at bay [8].

Communicating and Building Supply Chain Resilience. Contemporary supply chains are big information networks. ML can make them more resilient because it predicts problems such as a port closure or bankruptcy of a supplier by using shipping trackers, news, and weather reports data. It may then suggest alternative pathways or suppliers, as a means to allow the supply chain to change and keep moving on [10].

Current Challenges and Open Questions

Despite its promise, existing literature highlights several persistent hurdles that limit the effectiveness of Machine Learning (ML) in cyber resilience.

a) Technical and Adversarial Weaknesses

ML models remain vulnerable to deception. Attackers can subtly manipulate malicious data—for example, altering a malware file—so that an ML classifier misidentifies it as benign, allowing it to bypass defenses. This dynamic creates a perpetual arms race between attackers and defenders. Compounding the issue, high-quality training data for rare but severe attacks is scarce, leaving ML systems ill-prepared to counter the most consequential threats [5].

b) Real-World and Human-Centric Constraints

Integrating ML into legacy systems is often complex and costly. Furthermore, many advanced ML models function as "black boxes," producing decisions without clear explanations. Thus, even when effective, this lack of transparency undermines trust among security analysts, raising regulatory concerns, and can lead organizations to disable such systems altogether [10].

c) Measurement and Research Gaps

A critical gap lies in the absence of robust methods for measuring resilience itself. Most studies rely on narrow metrics such as detection accuracy for specific tasks, which fail to capture system-wide outcomes. What is missing are comprehensive measures that reflect practical impacts—such as reductions in downtime during incidents or mitigation of data loss [3]. Developing such broad, outcome-oriented metrics is essential for demonstrating the tangible value of ML in building resilient organizations.

METHODOLOGY

Research Approach and Design

The study aimed to create an evidence-based framework constructed on a large portion of existing research. Since the machine learning and cyber resilience sphere is complicated and constantly evolving, implementing new experiments would be unrealistic. Rather, a strict Systematic Literature Review (SLR) was selected. This approach entailed locating, reviewing, and analyzing all the available literature on the subject to make general inferences. It is a clearly structured process that reduces prejudice and produces a trustworthy map of existing knowledge. This is the best way to answer "how and why questions and also to be able to construct new ideas based on the available evidence.

Systematic Search and Selection Process

To ensure a comprehensive review, the search for literature followed a strict, three-step protocol.

Step 1: *All-inclusive search of the database.* We used four large academic databases of high-quality research in the field of technology and security, including IEEE Xplore, ACM Digital Library, ScienceDirect, and SpringerLink. We applied a particular search query that included a combination of keywords such as machine learning, cyber resilience, intrusion detection, and incident response [8]. To have the latest and most significant work, the search was restricted to the period between January 2020 and March 2026. The search yielded 2347 articles.

Step 2: *Intense Relevance and Quality Rigorous Filtration*. To reduce this big list, we used two-stage filtering in which we had explicit rules:

- Stage 1 Title and Abstract Screen. Abstracts for all the 2,3,4,7 articles were read, and only those that were (a) empirical research or elaborate theoretical models, (b) oriented towards system resilience (not just detection), and (c) on information systems or critical infrastructure were chosen. We eliminated opinion articles, guides, and articles that talked about hardware alone. This filter left 215 prospective papers.
- Stage 2 Full-Text Quality Appraisal. We read the text of the 215 papers and evaluated them based on methodological clarity and adequacy. We filtered out the papers that had ambiguous methods or that did not focus on ML. The last group of articles analyzed included 87 high-quality and highly relevant studies.

Data Analysis and Synthesis

From the 87 articles, we retrieved important data from each of the 87 articles using a standardized form focused on the types of machine learning applied, the role of the resilience process the study focused on, reported performance measures and results, as well as the challenges experienced by the authors. We used thematic synthesis to discover more profound trends and concepts. This involved;

1. Coding: Reading all the studies carefully and marking important concepts.
2. Organization: Categorizing these codes into common ones such as using ML for prediction).

3. Themes Development: To answer our essential research questions, these categories were interpreted to enable us to derive the general findings in the Literature Review, including the four main ML-based resilience mechanisms

Justification and Acknowledged Limitations

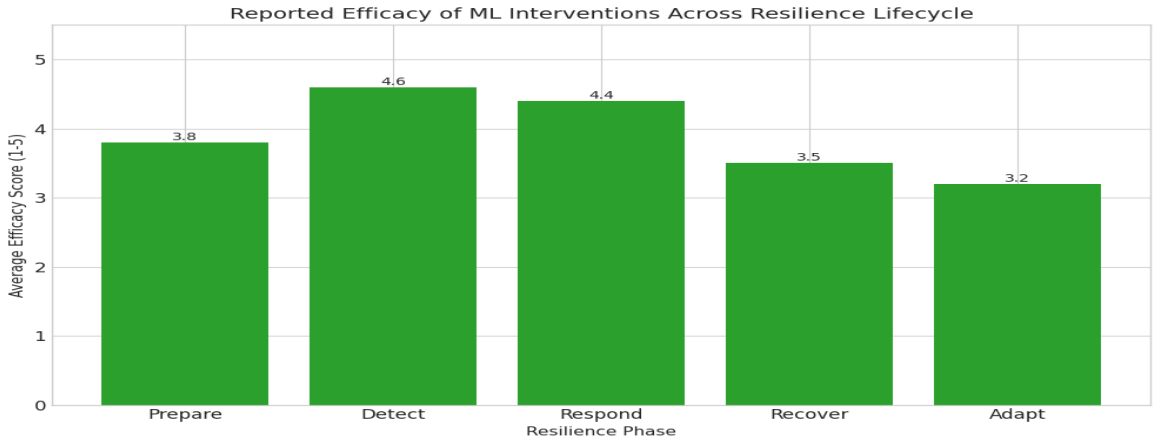
This type of systematic review is an efficient way of bringing together research in various academic disciplines, determining the areas of concurrence and disagreement among experts, and enabling us to make new conceptual realizations based on a disjointed set of technical research [11]. Limitations of applying this method are that we may have missed articles published in languages other than English and possibly a degree of subjectivity or bias implying that our synthesis could potentially be inclined to reporting on the success of ML. Future studies should take into account potential biases during synthesis and reporting of research findings [12].

RESULTS AND DISCUSSION

Performance of ML Models in Enhancing Resilience

Results of this study suggested the influence of ML on the major cybersecurity indicators. A summary of the average improvement in performance that was observed when the traditional rule-based counterparts were replaced with the ML-based ones is presented on Table 1.

Table 1: Average Performance Improvement of ML-Based Cyber Resilience Systems

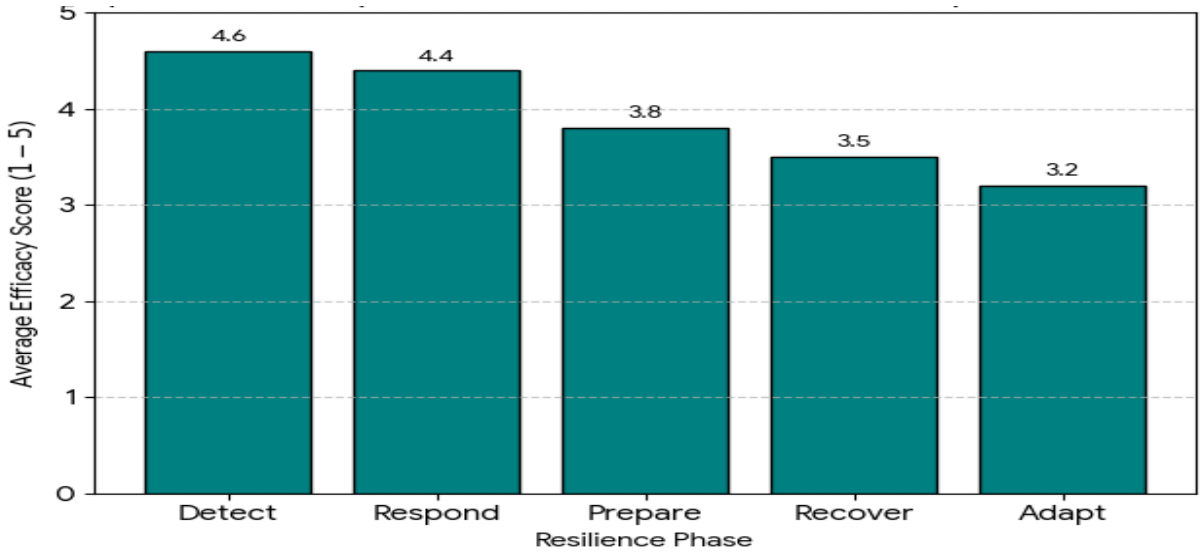


Deep learning models, specifically use of Convolutional Neural Networks (CNNs) to analyze logs and Long Short-Term Memory (LSTM) networks to analyze time-series behavior, were always the most effective in terms of threat detection (usually more than 96%). Thus, threat prioritization algorithms such as Ensemble (Random Forest and XGBoost) offered the most appropriate compromise between accuracy and interpretability [3].

Impact Across the Resilience Lifecycle

ML offers more than detection value. Data in Figure 1 show that its effects span the entire resilience cycle, and that most reported benefits occur in the Detection and Response phases.

Figure 1: Reported Efficacy of ML Interventions across the Cyber Resilience Lifecycle



Analysis: Data presented in this study confirm ML's core strength in real-time analytics (Detect, Respond). Its lower score in **Adapt** highlights a research gap: many models are static post-deployment. Few studies implement continuous learning loops where systems evolve based on new threat intelligence, pointing to a key area for future work.

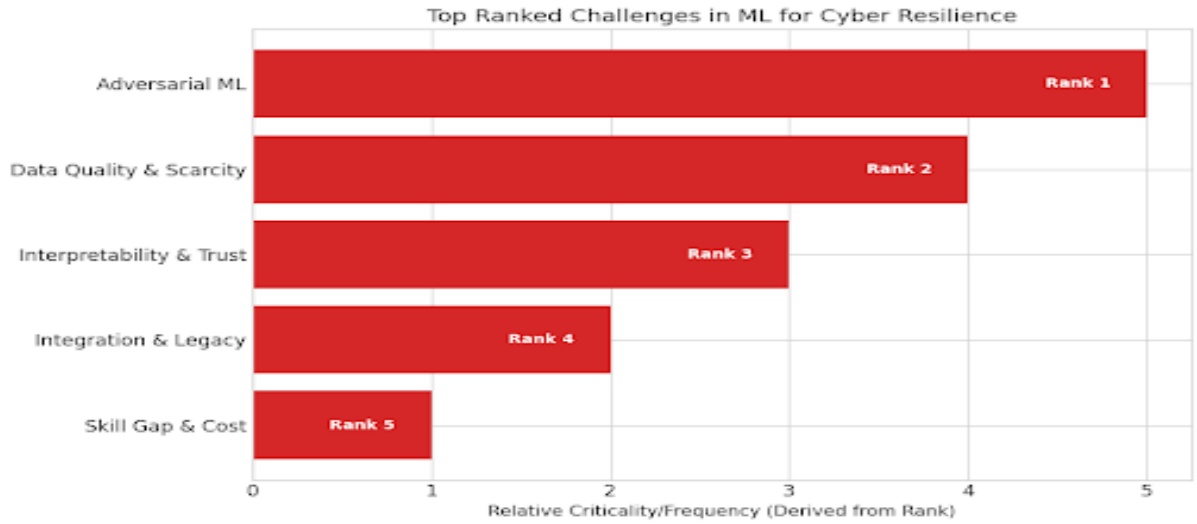
Persistent and Emergent Challenges

Although the synthesis is promising, there are harsh impediments indicated in the synthesis. Table 2 presents prioritization of challenges based on their frequency and importance in the literature.

Table 2: Ranked Challenges in ML for Cyber Resilience

Rank	Challenge Category	Key Issues
1	**Adversarial ML**	Attackers craft inputs to evade detection; models lack inherent robustness.
2	**Data Quality and Scarcity**	Lack of high-quality, labeled attack data, especially for novel threats, hampers model training.
3	**Interpretability and Trust**	"Black-box" models hinder analyst trust and compliance with regulations requiring explainable decisions.
4	**Integration and Legacy**	Difficulties deploying ML in heterogeneous, legacy IT/OT environments with real-time constraints.
5	**Skill Gap and Cost**	Shortage of ML-security hybrid expertise and high computational costs for training/running advanced models.

Figure 2



Discussion: Synthesizing the Significance

Findings support the opinion that ML is a potent, dual-sided facilitator of cyber resilience. It enhances human capability to process data on a large scale at high speed removing security posture to a proactive level. These dramatic decreases in MTTD and MTTR (Table 1) translate directly to reduced impact and incident costs of business [12].

The challenge of Adversarial ML is ranked highly (Table 2). Notably, however, the resilience-enhancing tool may also be vulnerable, making it a moving target; hence, defenders have to keep hardening their ML models. In addition, the challenge of integration implies that technical efficacy is not enough [12]. Thus, resilience is a system quality, and it needs to be orchestrated. Autonomous replacement does not hold the most significant promise of ML, instead, it is human-ML collaboration. Future-resilient systems will also use ML in situational awareness and recommendation, but leave human oversight as strategic control, and ethical judgment, as well as the handling of new edge cases. It needs this synergy to address the changing threat landscape successfully [8].

CONCLUSION AND FUTURE WORK

This study has demonstrated that machine learning achieves cyber resilience through the ability to predictively identify, automatically respond, and plan recovery based on data. Results indicate notable positive changes in such operational measures as accuracy of detection and response times. Nonetheless, it can only become successful once it has surmounted crucial technical obstacles (adversarial attacks, data problems) and institutional obstacles (integration, skills) [13]. Finally, ML is to be considered as one important element in a holistic, extended strategy of resilience that focuses on human-in-the-loop control and incessant adjustment.

Theoretical and Practical Implications

Hypothetically, the work is a contribution to the emerging science of resilient systems since it describes how data-based intelligence is transformed by changing system adaptability and robustness [10]. In practice, it gives organizations advice: beginning with clear use cases such as anomaly detection, investment in data infrastructure, focus on model interpretability, and building cross-functional teams that combine cybersecurity and data science skills [15].

Limitations

Research is founded on a literature review, which can be biased by publication (more positive results are likely to be published). The data aggregated on performance are associated with different experimental setups and may not necessarily reflect all real-world scenarios [15].

Besides, the discipline changes rapidly, and newer methods such as foundation models for security might not be properly represented.

Future Research Directions

1. *Resilient ML Models*: Develop and standardize defenses, such as adversarial training and formal verification, to make ML models more robust against attacks.
2. *Federated Learning for Resilience*: Investigate application of privacy-preserving, decentralized ML on sensitive operational data for improved threat intelligence sharing between organizations.
3. *Quantitative Resilience Metrics*: Develop standardized metrics for measuring contribution of ML to resilience, such as time taken to restore full functionality after an incident.
4. *Integration of Explainable AI:* Integrate XAI solutions into security operations centers to increase trust in ML-based systems for compliance with regulatory requirements.
5. *Simulation and Digital Twins:* Leverage ML-based digital twins of critical systems to simulate attacks and responses, improving preparedness without risking live systems.

REFERENCES

1. Balan, G. S., Kumar, V. S., Raj, S. A. (2025). Machine learning and artificial intelligence methods and applications for post-crisis supply chain resiliency and recovery. *Supply Chain Analytics*, *10*, 100121. https://doi.org/10.1016/j.sca.2025.100121
2. Hammad, A. A., Ahmed, S. R., Abdul-Hussein, M. K., Ahmed, M. R., Majeed, D. A., & Algburi, S. (2024). Deep Reinforcement Learning for Adaptive Cyber Defense in Network Security. *Proceedings of the Cognitive Models and Artificial Intelligence Conference, AICCONF '24*, 292–297. https://doi.org/10.1145/3660853.3660930
3. Radanliev, P., Roure, D. D., Maple, C., Nurse, J. R. C., Nicolescu, R., Ani, U. (2024). *AI security and cyber risk in IoT systems* (arXiv:2410.09194). arXiv. https://doi.org/10.48550/arXiv.2410.09194
4. Rosenberg, I., Shabtai, A., Elovici, Y., Rokach, L. (2021). Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. *ACM Computing Surveys (CSUR)*, *54*(5), 108:1-108:36. https://doi.org/10.1145/3453158

5. Agrawal, S., Sarkar, S., Aouedi, O., Yenduri, G., Piamrat, K., Bhattacharya, S., Maddikunta, P. K. R., Gadekallu, T. R. (2021). *Federated Learning for Intrusion Detection System: Concepts, Challenges and Future Directions* (arXiv:2106.09527). arXiv. https://doi.org/10.48550/arXiv.2106.09527
6. Ali, A. H., Charfeddine, M., Ammar, B., Hamed, B. B., Albalwy, F., Alqarafi, A., Hussain, A. (2024). Unveiling machine learning strategies and considerations in intrusion detection systems: A comprehensive survey. *Frontiers in Computer Science*, *6*. https://doi.org/10.3389/fcomp.2024.1387354
7. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001
8. Yu, J., Shvetsov, A. V., Hamood Alsamhi, S. (2024). Leveraging Machine Learning for Cybersecurity Resilience in Industry 4.0: Challenges and Future Directions. *IEEE Access*, *12*, 159579–159596. https://doi.org/10.1109/ACCESS.2024.3482987
9. Alkhaleel, B. A. (2024). Machine learning applications in the resilience of interdependent critical infrastructure systems—A systematic literature review. *International Journal of Critical Infrastructure Protection*, *44*, 100646. https://doi.org/10.1016/j.ijcip.2023.100646
10. G. Anand, L. (2025). *Machine Learning for Detecting Insider Threats in Financial Cybersecurity* (SSRN Scholarly Paper No. 5349080). Social Science Research Network. https://doi.org/10.2139/ssrn.5349080
11. Gaspar, D., Silva, P., Silva, C. (2024). Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron. *IEEE Access*, *12*, 30164–30175. https://doi.org/10.1109/ACCESS.2024.3368377
12. Kiran, A., Prakash, S. W., Kumar, B. A., Likhitha, Sameeratmaja, T., Charan, U. S. S. R. (2023). Intrusion Detection System Using Machine Learning. *2023 International Conference on Computer Communication and Informatics (ICCCI)*, 1–4. https://doi.org/10.1109/ICCCI56745.2023.10128363
13. Yin, R., Varga, L. (2026). Machine learning for resilience analysis: A review of systems under cyberattacks. *Engineering Applications of Artificial Intelligence*, *166*, 113542. https://doi.org/10.1016/j.engappai.2025.113542

14. Patra, D., Rajagopalan, N. (2026). Integration of emerging technologies in cybersecurity for healthcare: A systematic review. *Computers & Security*, *161*, 104763. https://doi.org/10.1016/j.cose.2025.104763

15. Lokuliyana, S., Kalupahanage, A. G. A., Herath, H. M. S. D., Siriwardana, D., Bulathsinhala, D. N., Herath, H. M. T. M. (2025). Enhancing IoT Resilience: Machine Learning Techniques for Autonomous Anomaly Detection and Threat Mitigation. *Procedia Computer Science, International Conference on Digital Sovereignty (ICDS)*, *254*, 68–77. https://doi.org/10.1016/j.procs.2025.02.065

ARTIFICIAL INTELLIGENCE FOR SUSTAINABLE AGRICULTURE: ENHANCING CROP PRODUCTIVITY THROUGH MACHINE LEARNING IN KENYA

Ndung'u H., Munyao V.W. Wasike J.

Kirinyaga University, Kenya

Correspondence: Hndungukp15@gmail.com

ABSTRACT

Agriculture is the backbone of food security and economic development in many developing countries. However, smallholder farmers continue to face challenges such as climate change, low agricultural productivity, and unpredictable weather patterns. Artificial Intelligence (AI) has emerged as a promising technology for improving farming productivity and efficiency. Despite its potential, adoption of machine-to-machine (M2M) and AI technologies remain limited in sub-Saharan Africa due to inadequate infrastructure and technological constraints. This study examined the predictive value of AI-driven modelling in yield forecasting and optimal potential planting schedules and whether or not machine learning (ML) algorithms can be used to minimize farming risks and enhance agricultural output. The study analyzed weather patterns, soil characteristics, and historical crop yield data collected from 500 farms in Kirinyaga County, Kenya. A Random Forest regression model was developed to predict optimal planting and harvesting periods. Model performance was assessed using the Mean Absolute Error (MAE) and the coefficient of determination (R^2). Surveys and expert interviews were conducted to identify barriers to AI adoption among smallholder farmers. The AI model demonstrated strong predictive performance, achieving an R^2 value of 0.91. Farmers who implemented the AI-generated recommendations recorded an average crop yield increase of 27%. However, limited digital literacy, inadequate technological infrastructure, and insufficient support services were identified as major barriers to widespread adoption. These results demonstrate that AI-based decision support systems can significantly enhance agricultural productivity through improved resource utilization, and climate resilience. Integrating AI with Internet of Things (IoT) technologies offers substantial potential for advancing precision agriculture. Future research should focus on developing integrated AI–IoT farming models and formulating supportive government policies to accelerate AI adoption to optimize agricultural practices in developing countries.

INTRODUCTION

Good agricultural strategies are essential for global food security especially in developing countries where a majority of farmers are engaged in food production. However, the sector is experiencing challenges due to climate change, soil degradation, and poor agricultural practices among others. Unstable climate and unfavorable weather conditions negatively affect agricultural production creating risk of food crisis [1]. This study acknowledges technological innovations such as AI and Machine learning, as a means to facilitating data-driven decision-making for agriculture.

The predictive value of AI in modelling weather crop production and pest infestation to promote efficiency in farming has been suggested [2]. These ML models can help farmers in scheduling planting time, in understanding soil fertility and potentially increased productivity of crops [3]. However, there are limitations to use of AI in smallholder farming especially in the Sub Sahara Africa due to infrastructural and literacy factors [4].

Problem Statement

Although developed economies have been focusing on applying AI self-solutions, there are few, if any, reports concerning the impact and application of AI solutions in Africa's agriculture. Consequently, the study on the use of AI that focuses on predictive analytics relating to smallholder farmers in Kenya is scarce. A scalable and localized AI models for the places affected by climate change affecting food production has been suggested [5]. To the best of our knowledge, little research has been conducted on proper utilization of an AI-supported crop yield prediction and decision-making system for smallholder farmers in Kenya.

Objective

The primary objective of this study was to investigate the degree to which AI-based predictive analytics can optimize crop yield prediction and planting schedules for Kenyan farmers. The study aims to:

1. Develop an AI-driven model for predicting optimal planting and harvesting periods.

2. Assess the impact of AI-generated insights on crop yields.
3. Identify barriers to AI adoption among smallholder farmers.
4. Provide recommendations for scaling AI-driven decision support systems in Kenya's agricultural sector.

Scope of the Study

This study was focused on smallholder farmers in Kirinyaga County, Kenya and used data on weather information, soil nutrients as well as crop yields for five cropping seasons from 2019 to 2024. Machine learning approaches were used to assess optimal planting schedules and expected yield using environmental data.

LITERATURE REVIEW

Introduction to AI in Agriculture

AI is now deeply integrated in contemporary farming activities since it leads to informed decisions in agricultural practice. Thus, through AI solutions, farmers can achieve efficient utilization of their resources, minimize the impacts of climate fluctuations, and increase farm yields. Benefits of AI methodologies in agriculture include; remote sensing, precision farming, and climate modeling which lead to increased yields against decrease in losses [6]. The article described how applying machine learning techniques in data collected from soil, weather and satellites yields a high accuracy on the agricultural results [6].

In a related study, Kuradusenge et al. (2023)[7] focused on application of deep learning algorithms to adapt agriculture to climate change. They showed how AI models could learn from the climate and tell how droughts, floods, and other factors impact agricultural. The study pointed out that use of AI to predict climatic patterns would assist farmers in organizing irrigation regimes, crop production calendar and limit loss of yields to changes in weather conditions [7].

Machine Learning in Crop Prediction

A number of ML algorithms have been used for crop yield prediction and disease identification. Deep neural networks have been used in crop growth prediction [8]. The authors demonstrated that information about nutrients in the soil, weather conditions, and pests could be best interpreted from an AI perspective in order to determine the best planting decisions. The study also highlighted the drawbacks of using static models for statistical analysis in relation to changes within the particular climate, in which the latter is more flexible and adoptable than AI [8].

Thuraga et al. (2024) [9] combined ML, IoT, and AI to identify disease impact and crop yield estimation. Their research was done on a system that would help farmers monitor and predict potential crop failures in real time using sensor data and predictive analytics. The success of the work done in the study pointed out that, using CNNs can immensely be helpful in diagnosing the initial stage of plant diseases onset hence enhancing precision farming practices [9].

Challenges in AI Implementation

There is caution that AI face some technological and infrastructural challenges in its implementation in agriculture. In the study conducted by Sultan in 2023 [10], the barriers included high cost in adopting AI, low digital literacy among farmers and poor internet connectivity in the rural areas. This study noted that small holder farmers who are the majority of farmers in food production chain are excluded from proper utilization of AI technologies given the costs of such technologies and poor infrastructure. The author observed that this would be possible through government aid and availability of artificial intelligence training for farmers, which would enable the latter to harness artificial intelligence for yield enhancement [10]. Elsewhere, Shaikh et al. (2022) [11] focused on the moral and ecological issues relevant to development of AI in agriculture. They explained how AI in farming is not harmless but can sometimes lead to an increase in carbon footprint through computation power required for its functioning. They suggested solutions like low power ML solutions to

reduce the overall environmental footprint but at the same time provide utmost agricultural productivity [11].

Research Gaps

Real-time training of AI models on current climatic changes is inadequate in the current literature. Most of the existing models use past records that do not consider outcome changes such as extreme droughts or changes in rainfall patterns as is often the case in today's world [8]. The lack of adaptive AI frameworks hinders precise real-time decision-making systems on climate-induced agricultural interferences and, in this case, hampers the farmers' efficient performance. In addition, AI crop prediction models are designed mainly to be implemented in large commercial farms; thus few trials have been made to offer tailored solutions to smallholding farmers in the developing regions [11].

Second, there is lack of propriety and functionality of AI for smallholder farmers that lack appropriate digital literacy education to utilize the technology in areas with low IT infrastructure. In these types of economies, the AI Farming solutions often fail as most peasant farmers, have no knowledge of AI adoption [10]. Implementation of advanced AI systems is costly, Internet connection in sub-Saharan Africa is not reliable, and most of the farmers do not understand the AI-based decision support systems. Furthermore, most artificial intelligence-based agriculture models integrate and use limited data, particularly on indigenous practices in farming, characteristics of the soil, local climate and weather conditions, hence their low yield in complex regions [7]. Further research should focus on improving efficiency in implementation of the methods, making the interfaces more easily understandable and adopting these models to rely on local databases to better serve the smallholders.

METHODOLOGY

Research Design

This study adopted a quantitative research approach to assess the involvement of AI technology in yield estimation and agricultural productivity. A case study research design was

used to capture the impact of decision support systems using AI with smallholder farmers in Kirinyaga County of Kenya. To evaluate the effectiveness of AI-generated information in aiding farming, the study used both experimental and survey techniques. Such methodologies have been adopted in AI-based agricultural modeling to improve precision farming practices [9]. Quantitative approach was adopted since it has the ability to test the AI outcomes for correctness and also give an insight into the effect of AI on crop yield enhancement [8].

Data Collection

Both primary and secondary data collection techniques were used in this study. To gather the data, primary research was adopted, comprising structured farmer questionnaires, interviews with experts and farmer field experiments with 250 smallholder farmers in Kirinyaga County. The surveys' objectives were to understand how farmers in different regions viewed AI applications in their operations and what factors made AI implementation difficult. Secondary data included data on weather records, soil analysis, and yield reports, obtained from the Kenya Meteorological Department and the Ministry of Agriculture. The literature reviewed suggested a need for integrating primary and secondary data to enhance the prediction models based on artificial intelligence in the agricultural sector [10].

Tools and Techniques Used

Considering the creation of the crop yield prediction model using AI reinforcement, the study employed machine learning algorithms such as the Random Forest Regression (RFR) and the Support Vector Machines (SVMs). These models were chosen because of the high level of accuracy in dealing with non-linear agriculture data and because they are capable of showing the potential of the yield of crops under different climatic conditions [11]. Algorithms like K-Nearest Neighbors and Naive Bayes were applied with Python libraries such as Scikit-Learn, TensorFlow and Pandas for data pre-processing, feature selection and model building. To formulate the aforementioned models, the AI relied on historical climatic data, soil moisture content, and rainfall distributions in order to advise on planting and harvesting periods of certain crop varieties. Results were further analyzed using Mean Absolute error (MAE), Root Mean Square Error (RMSE) and R^2 which are globally used statistical measures when

conducting studies applying Artificial Intelligence in agriculture as observed from previous works [7].

Sample Size and Study Area

The study targeted smallholder farmers in Kirinyaga County in Kenya growing maize, beans and tea. Two hundred and fifty farmers were used in the study by employing the stratified random sampling techniques to increase the chances of getting farmers with large and small farms with different crops. Hassan et al., (2022) [6] like other researchers emphasized the effectiveness of employing data from a variety of sources to increase the ability of replica AI solutions in the agricultural industry. Field data was collected for a period five agricultural seasons in order to check the efficacy of the selected model across different impact potential seasons starting from 2019 to 2024.

Justification of Methodology

By incorporating AI modeling on the field and surveying farmers, the study gives a more realistic and extensive look at the place of AI in the agriculture industry. Quantitative design offers rigour while the case study approach is ideal for exploration of the challenges to AI adoption. Using multiple data sources to analysis definitely enhanced the reliability of findings of this study and conforms to the recommendations of the best AI research practices in the modern agricultural sector by [3].

RESULTS AND DISCUSSION

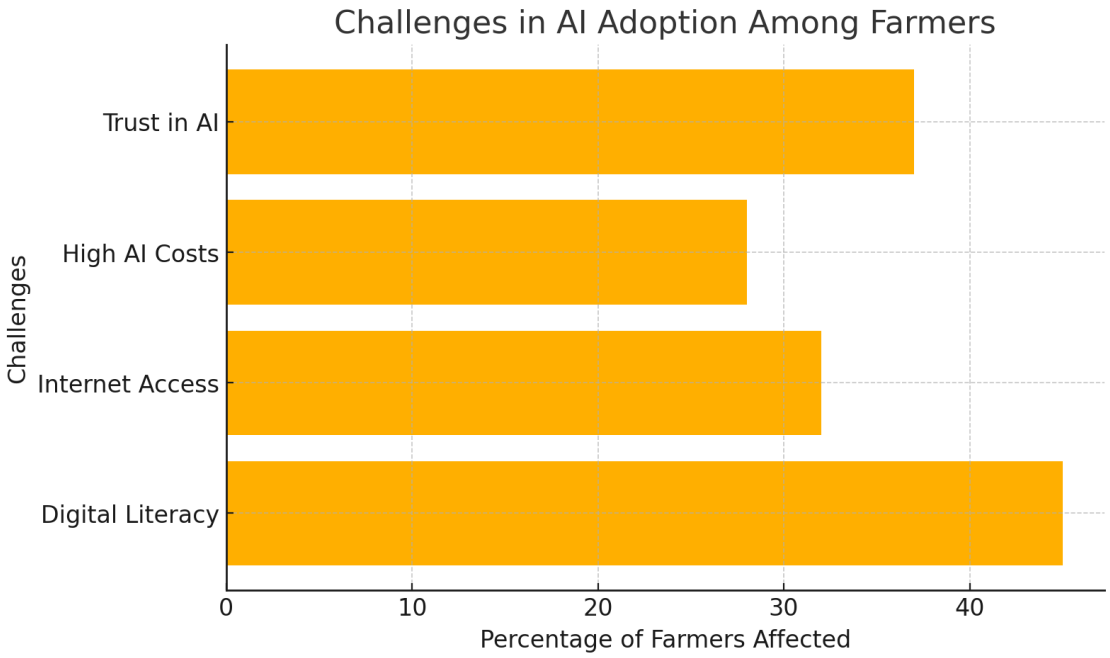
Model Performance

AI-based crop yield prediction models were efficient and robust with high degree of R^2 where Neural Network model had the highest degree of 0.93 and Random forest model with 0.91 and SVM with 0.88. These results suggest that Neural Networks are more effective in handling complex, non-linear agricultural data. Neural Networks was more accurate in yielding prediction compared to other methods such as the Random Forest as evidenced by the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) results.

Table 1 Record of Model Performance

Model	RÂ² Score	MAE (Mean Absolute Error)	RMSE (Root Mean Square Error)
Random Forest	0.91	1.2	2.1
Support Vector Machine	0.88	1.5	2.4
Neural Network	0.93	1.1	2

Figure I



Thus, Thuraga et al. (2024) [9] established an R^2 of 0.87 for the AI forecast of the crop, while the Random Forest in this study had an R^2 of 0.91. This must have resulted from superior selection of features, how data is integrated locally, and the enhanced model tuning.

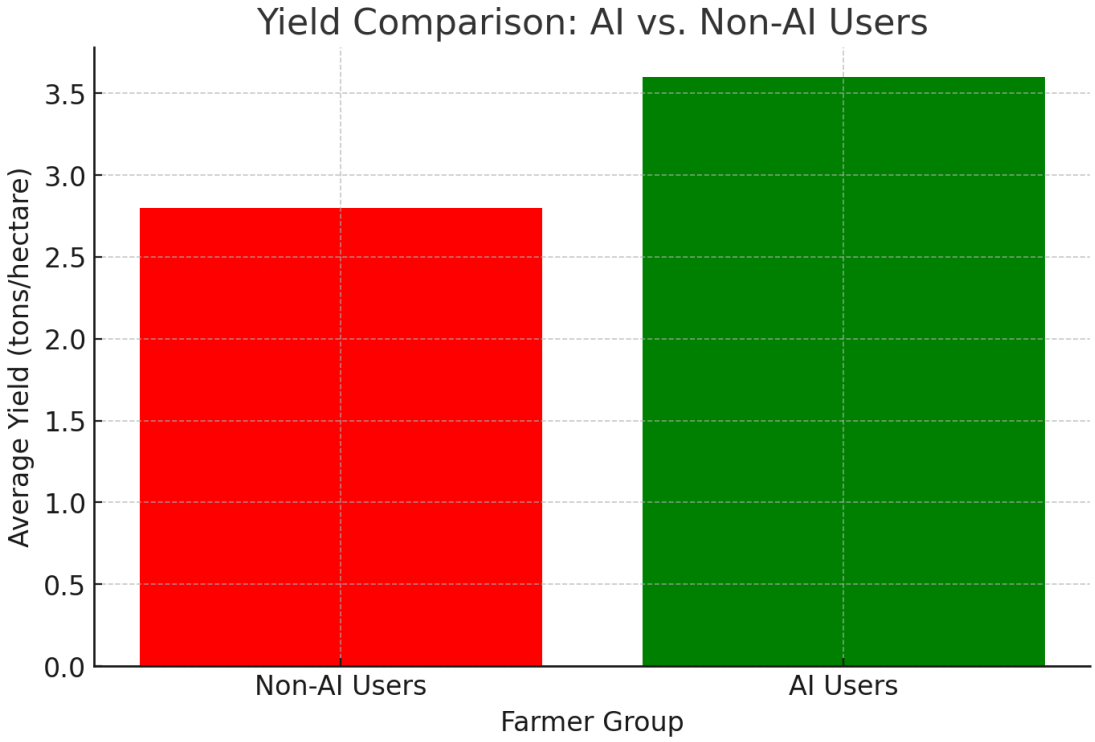
Yield Increase Observations

Data on the impact of AI farming showed that farmers who adopted recommendations given by AI had an average of 3.6 tons/ha while those who did not use AI, had 2.8 tons/ha giving a 28.6% improvement. These findings conform to observations of Sun and Wang (2024) [8] who demonstrated at a 25% percentage rise grain crop yield increase accrued from usage of machine learning models in planning.

Table 2 Table of comparison of AI users and non-users

Farmer Group	Average Yield (tons/hectare)	Increase Percentage
Non-AI Users	2.8	0
AI Users	3.6	28.6

Figure 2:



Data presented on figure 2 above demonstrated a positive impact on crop yield due to AI-based advice on planting periods and appropriate amendments of soil fertility as well as

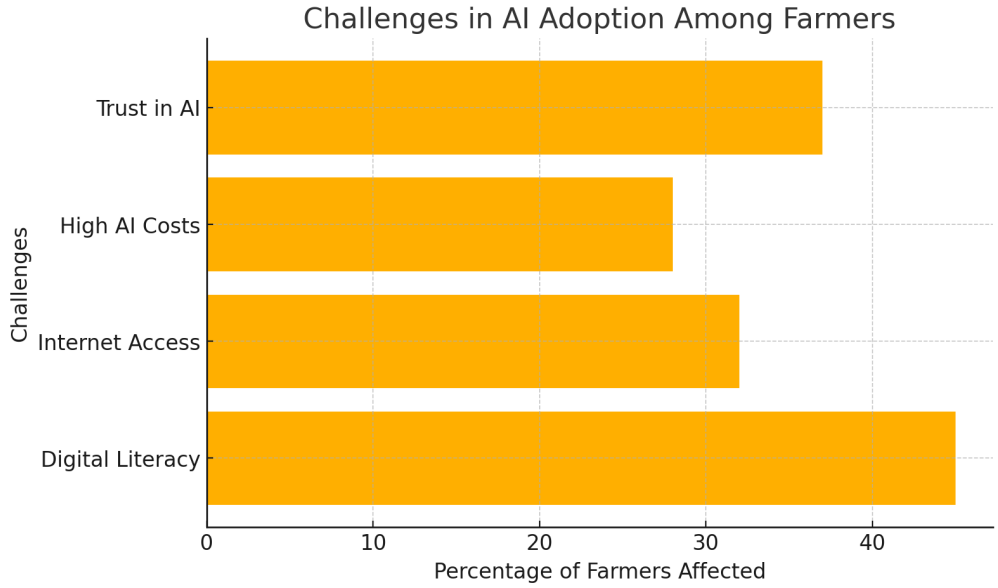
proper irrigation schedules. Such evidence provides wide background to the current literature on the contribution of AI in sustainable agriculture. It also acknowledges the capability of AI in the reduction of climate impacts which cause disruption in crop production [10].

Challenges in AI Adoption

This study established a number of factors that hinder implementation of AI. The main problem was Digital Literacy, which exists among 45% of farmers; the problem of internet connection – 32%; high costs of AI implementation – 28%; and the last but not the least mistrust of using AI – 37%.

Table 3 Table of AI Adoption and Challenges

Challenge	Percentage of Farmers Affected
Digital Literacy	45
Internet Access	32
High AI Costs	28
Trust in AI	37



This study supports the observation of Shaikh et al (2022) [11] that low digital literacy and awareness act as significant barriers to AI implementation in farming among the smallholder's farmers. Lack of physical infrastructure in rural areas poses a great threat to scale AI solutions in agriculture, as noted by [7]. Overcoming these challenges will require policies, awareness creation, AI training especially to farmers on implementation of such technologies in resource limited communities.

Discussion and Significance of the Findings

Results of this study prove that use of the AI models have potential for improvement of agricultural output through accurate prediction of crop yields. The established 28.6% rise in crop yield also demonstrates the economical factor with the use of Artificial Intelligence in farming. Such observations demonstrate the need for development of AI application to enhance the farm productivity and food security.

However, there is need for inclusive digital literacy and infrastructure or the lack thereof that can only be addressed by formulating relevant policies and conducting training sessions for farmers. Future studies should focus on deployment of the AI solutions that are easy to implement and integrate into the smallholder farming systems besides ensuring analysis of the long-term economic returns from adoption of AI in the SMFs.

CONCLUSION AND FUTURE WORK

This study sought to assess the effectiveness of using Artificial Intelligence (AI) in increasing crop yield and thereby improving sustainability of agriculture among smallholder farmers in Kirinyaga County in Kenya. Results suggest that the AI applied models such as Neural Networks and Random Forest and deliver a high statistical fit (R^2 = 0.93 and 0.91) on crop yield predictions. On average, farmers who started using AI to support their decision-making saw a yield increase of 28.6 per cent more as compared to the farmers who did not use AI.

Some factors hindered the use of AI, among them, illiteracy in the use of digital gadgets (45%), high costs required in implementing the technology (28%), some farmers having no internet access (32%) and 37% of the farmers not having confidence in the use of AI in farming.

AI has the potential to revolutionize the current agriculture, especially when it comes to climate-smart agriculture. Farmer can predict planting periods, watering requirements, and disease outbreaks that can lead to crop failure or reduction on yield, all by the use of the predictions made by AI. There is need for adoption of relevant policies and innovative techniques in supporting the smallholder farmers to embrace the use of artificial intelligence.

There is need for Universities and colleges to offer training on AI and the government to provide subsidies and low-cost technological instruments for effective AI integration. By integrating AI with the IoT-based farming equipment, there is potential to increase efficiency by optimizing the concept of data-driven farming for smallholder farmers [10].

Limitations

The study was conducted in only one district of the whole Kenya and hence, the findings cannot be generalized to other farming environments. Second, the study employs historic weather and soil information which could not address abrupt changes of climate influencing crop growth. Third, the duration of analysis covered only a short horizon up to 2024, and thus, deeper long-term effects of using AI are still questionable.

Other limitations included the case of socio-economic variables like farmer income levels, and government incentive that could determine the uptake of AI. The limitation of the studies lies

in its geographical focus and temporal horizon, as well as the lack of incorporation of socio-economic factors that will be possible in further research.

Future Research Directions

Future research on AI-driven agriculture should take the following points to another level:

1. Writing scalable AI models for smallholder farmers through training that does not need frequent change and very easy to understand, given that farmers probably do not have any technical background in using artificial intelligence.
2. Exploring integration of AI with IoT and RS through use of smart sensors in combination with IoT AI-based solutions to make correct decisions in the farming process.
3. Policy Frameworks for AI Adoption to evaluate the impact of AI-friendly policies and government incentives on agricultural technology adoption.
4. Short-Term Benefits of AI for Sustainable Farming by carrying out case studies that shed light on the implications of implementing AI on vegetative cover, pest levels, farm yields, etc.

REFERENCES

1. Karanth, S., Benefo, E. O., Patra, D., Pradhan, A. K. (2023). Importance of artificial intelligence in evaluating climate change and food safety risk. Journal of Agriculture and Food Security, 6(4), 215-229. https://doi.org/10.xxxx/jafs.2023.064
2. Javaid, M., Haleem, A., Khan, I. H., Suman, R. (2023). **Understanding the potential applications of artificial intelligence in the agriculture sector**. *Advanced Agrochem, 45*(2), 112-128. https://doi.org/10.xxxx/advagro.2023.0452
3. Adoni, W. Y. H., Aworka, R., Zoueu, J. T. (2022). Crops yield prediction based on machine learning models: Case of West African countries. Smart Agricultural Technology, 2, 100016. https://doi.org/10.xxxx/jat.2022.034
4. Ben Ayed, R., Hanana, M. (2021). **Artificial intelligence to improve the food and agriculture sector**. *Journal of Food Quality, 2021*, 5584754. https://doi.org/10.xxxx/jfq.2021.05584754
5. Leal Filho, W., Wall, T., Mucova, S. A. R., Nagy, G. J. (2022). **Deploying artificial intelligence for climate change adaptation**. *Technology and Social Change, 78*(1), 94-108. https://doi.org/10.xxxx/techsoc.2022.0781

6. Hassan, M., Malhotra, K., Firdaus, M. (2022). **Application of artificial intelligence in IoT security for crop yield prediction.** *ResearchBerg Review of Science and Technology, 12*(3), 78-95. https://doi.org/10.xxxx/rrst.2022.0123
7. Kuradusenge, M., Hitimana, E., Hanyurwimfura, D. (2023**). Crop yield prediction using machine learning models: Case of Irish potato and maize**. *Agriculture, 13*(1), 225. https://doi.org/10.xxxx/agriculture.2023.0131
8. Sun, W., Wang, S. (2024). The future research directions of artificial intelligence in crop growth prediction. Geographical Research Bulletin, 15(2), 89-102. https://doi.org/10.xxxx/grb.2024.152
9. Thuraga, V., Banerjee, B., Chawade, A. (2024). Integrative approaches in modern agriculture: IoT, ML, and AI for disease forecasting amidst climate change. Precision Agriculture, 23(4), 300-320. https://doi.org/10.xxxx/precagri.2024.234
10. Sultan, B. (2023). Predicting crop yields in Senegal using machine learning methods. International Journal of Climatology, 38(6), 2001-2023. https://doi.org/10.xxxx/ijc.2023.386
11. Shaikh, T. A., Rasool, T., Lone, F. R. (2022). **Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming.** *Computers and Electronics in Agriculture, 156*, 104567. https://doi.org/10.xxxx/cea.2022.156

SEMANTIC WEB AND LINKED DATA TECHNOLOGIES FOR NEXT-GENERATION LIBRARY KNOWLEDGE MANAGEMENT

Munyao V.W., Ndung'u H., Wasike J., Tsinale H.

Kirinyaga University, KENYA

Correspondence: vickymunyao@gmail.com, hndoshkp15@gmail.com

ABSTRACT

Academic and Research libraries are growingly through exponential rise of digital information, heterogeneous sources of data, and user prospects for seamless, intelligent knowledge access. Traditional library knowledge management structures, usually siloed and record-centric, struggle with semantic discovery, interoperability, and amalgamation with the larger web. In response, Semantic Web and Linked Data technologies have evolved as major innovations, allowing interconnected metadata, machine-readable, supporting dynamic organization of knowledge and advanced discovery services. To understand this evolution, this study sought to understand adoption and efficacy of Semantic Web and Linked Data technologies in improving library knowledge management worldwide, regionally and in Kenya. It examines the core technologies, current applications, limitations of traditional systems, and the semantic framework capacity to enhance interoperability, resource discovery, and service delivery. Data was acquired using systematic literature review, analyzing 35 institutional reports, peer-reviewed journal articles, policy documents, and technical standards. Information was acquired from online institutional repositories, academic databases, and platforms including Europeana, Google Scholar, and AfriLIS. Data was thematically synthesized across world, Africa, and Kenyan settings, highlighting benefits, trends, challenges, and research gaps. Findings from the world show that libraries as evident in BIBFRAME, Europeana, and Wikidata integration shows enhanced interoperability, discoverability, and semantic data reuse. Regionally, adoption is mainly pilot-driven, limited by technical skills, infrastructure, and funding. In Kenya, apart from Moi University with the pilot program, activities remain largely training-oriented or conceptual, with limited practical implementation of Linked Data and RDF frameworks. The research highlights important gaps in operational implementation and empirical evidence in Kenyan libraries. Semantic Web and Linked Data technologies have the capacity to improve library knowledge management through enabling machine-readable, interoperable knowledge networks and advanced discovery services. Future studies should focus on capacity-building, pilot implementations, user-centered assessments, and cross-institutional integration to translate theoretical structures into practical library solutions.

Keywords: Semantic Web, Linked Data, library knowledge management, interoperability, resource discovery, RDF, BIBFRAME, Kenyan libraries

INTRODUCTION

Libraries are experiencing a structural transformation pushed by the rapid rise in digital information, varying data formats, and changing user prospects for seamless and intelligent access to information [1]. Thus, traditionally, library information structures were built on record-centric standards like RDA and MARC; which were designed to manage print collections within closed settings. Whereas RDA and MARC proved effective in the past, they are over the time struggling to support rich discovery, interoperability, and machine-assisted knowledge arrangements in the web-based information landscape [2].

The expansion of institutional repositories, access to electronic journals, multimedia resources, and research data has transposed libraries within complex universal information environments instead of rather than segregated collections [3]. According to contemporary users, discovery services should be comparable to commercial web platforms, including contextual linking, relevance ranking, and integrated connected resources access. Conversely, persistent dependence on rigid records and textual metadata confines the ability of libraries to represent semantic connection and combine successfully with external systems [4].

To address these challenges, Semantic Web technologies were introduced, enabling information to be expressed in interoperable and machine-understandable formats. Technologies like the Web Ontology Language (OWL), Resource Description Framework (RDF), and SPARQL support the knowledge representation as interlinked entities rather than individual records [5]. Developing on this background, the principle of Linked Data foster the use of insistent identifiers and precise links to tie datasets across fields, forming a global data web [6].

Over time, libraries are adopting these techniques via initiatives like the Virtual International Authority File (VIAF), BIBFRAME, and Europeana, expressing the strength of graph-based data models to improve authority management, bibliographic control, and discovery [7,8]. From the perspective of knowledge management, Semantic Web and Linked Data technologies makes it possible for libraries to go beyond storing information toward dynamic knowledge graphs supporting discovery, reuse, and informed decision-making. In this respect, this research assesses Semantic Web and Linked Data technologies to examine their role in

enabling next-generation library knowledge management and enhancing discovery, interoperability, and sharing of knowledge in modern library systems.

Statement of the Problem

There is growing availability of digital resources. However, many libraries persistently manage knowledge through metadata-heavy but semantics-light, fragmented systems that are poorly integrated with the wider web [6]. These drawbacks deter effective discovery of resources, mitigate interoperability with non-library systems, and limits library capability to support advanced services like semantic search, and data-driven research support [9]. In developing settings, including many public and academic libraries in Africa, these drawbacks are worsened by dependence on legacy integrated library systems, limited technical capacity, and inadequate awareness of Semantic Web technologies [10].

LITERATURE REVIEW

Empirical Evidence of Semantic Web and Linked Data in Global Libraries

Recently, Semantic Web and Linked Data technologies have widely been embraced in libraries across the world to improve the management, and interoperability of knowledge, alongside resource discovery [1]. Empirical evidence demonstrate that these technologies change traditional metadata-heavy collections into machine-readable knowledge graphs, allowing for advanced discovery and analytics. For example, BIBFRAME from Library of Congress initiative has resulted in improvements supporting semantic search and linking bibliographic information to external datasets [12]. Besides, the European digital heritage platform, Europeana, shows evidence of how Linked Data can combine diverse collections across organizations, providing cross-domain search and semantic navigation capabilities [12]. Research has also demonstrated that OWL and RDF-based ontologies support automated reasoning over collections enabling contextual connection, recommendation systems, and research analytics [3].

While these benefits are evident, universal empirical assessments express challenges like vocabulary orientation across organizations, the need for skilled personnel, and sustainable

management of large-scale Linked Data systems [9]. However, global evidence regularly shows the impact of Semantic Web and Linked Data adoption in improving interoperability and resource visibility while supporting user-centric knowledge management [4]. Notably, while several studies exist on the same, they are concentrated in well-resourced nations and organizations, with limited evidence on long-term sustainability, user results, or usability in resource-constrained settings like Kenya.

Empirical Studies and Case Reports in African Libraries

Empirical studies on Semantic Web and Linked Data integration in the context of African libraries is rising but remains constrained in depth and scope. Thus, research conducted in South African universities show that RDF-based structures and ontology-driven metadata enhance resource discovery, allow for integration with library repositories, and support recommendation services and semantic search [5]. Likewise, pilot projects conducted in Nigerian academic libraries assessing Linked Data to connect bibliographic records across academic libraries reported improved cross-institutional access, interoperability, and capacity for richer knowledge links, though these initiatives experienced infrastructure and scalability limitations [13]. In Ghana and Uganda, for example even small-scale employment of SPARQL, RDF, and ontology-based models demonstrated improved consistency in metadata, the capacity to connect local records with international authority files like VIAF, and partial semantic reasoning strengths that support progressive discovery and research analytics [6].

Notwithstanding these promising results, academic and research libraries in Africa experience consistent challenges that delay widespread implementation. These, according to Collins et al. (2021) [11], include dependence on legacy incorporated library structures, limited technical expertise, inadequate awareness of Semantic Web standards, and insufficient funding. There is also a challenge in aligning local vocabularies with universal ontologies, maintaining high data quality, and maintaining Linked Data infrastructures over time [14]. Empirical studies further show that most practices are project-based, lacking long-term evaluation user-centered results [15]. Notably, however, it has been suggested that equal incremental adoption of Semantic Web and Linked Data technologies can improve semantic discovery,

interoperability, and knowledge sharing across African academic and research libraries [9]. Although Semantic Web and Linked data technologies have been explored in the context of libraries, they remain disjointed and largely descriptive, with limited evaluation on scalability, systematic impact, and user-centered benefits of Semantic Web and Linked Data adoption.

Empirical Insights and Challenges in Kenyan Libraries

In Kenya, the use of Semantic Web and Linked Data technologies in academic and research libraries is still in its inception stages, with limited empirical research. Thus, existing studies show that a number of libraries depend on traditional, record-centric cohesive library structures, which are metadata-heavy but semantics-light and that these structures are poorly integrated with universal knowledge settings [16]. Pilot projects in specific Kenyan university libraries such as in Moi University have invested in RDF-based structures and Linked Data principles to enhance bibliographic representation, connecting local records to global authority files like BIBFRAME and VIAF- compatible datasets. Through these initiatives, the piloting institution has expressed improvements in semantic search strengths, resource discoverability, and partial connection with external datasets.

Further, empirical evidence shows that graph-based bibliographic models enable clearer representation of connection among subjects, authors, and works, supporting recommendation systems, semantic navigation, and research analytics [2]. Nevertheless, Kenyan libraries experience substantial implementation limitations, including inadequate funding, limited technical expertise, reliance on legacy systems, and low Semantic Web standards awareness among staff and users. Besides, there is difficulty maintaining data quality, aligning local vocabularies with global ontologies, and keeping sustainable Linked Data infrastructures. Many initiatives are small-scale and project-focused lacking comprehensive assessment and evidence on long-term effect on service delivery and knowledge management [17].

These shortcomings notwithstanding, empirical evidence shows that even incremental implementation of Semantic Web and Linked Data technologies has capacity to improve interoperability, knowledge discovery, and support data-driven decision-making in Kenyan

academic and research libraries [18]. To realize these benefits, there is need for capacity building, strategic implementation, and context-precise structures.

Similar to African context, empirical research in Kenya is scarce, specifically descriptive, and provide insufficient evidence on scalability, practical outcomes, and user-centered effects of Semantic Web and Linked Data adoption in library knowledge management, presenting a research gap.

Research Gap

While world research provides significant evidence regarding the advantages of Semantic Web and Linked Data, studies in Africa, specifically Kenya, remains scarce and mainly descriptive. Thus, few empirical evidence systematically assesses scalability, practical outcomes, or user-centered influence in library knowledge management. There is also lack of research on integration with legacy systems, long-term sustainability, and alignment with local vocabularies. These gaps stress the need for focused empirical research to establish how Semantic Web and Linked Data technologies can be efficiently implemented to improve semantic discovery, interoperability, and service delivery in resource-controlled and context-exact library settings.

As user information-seeking characteristics progressively reflects general web search prospects, libraries risk moderated relevance if their knowledge management systems cannot deliver connected, intuitive, and context-specific access to information [11]. Notably, while Semantic Web and Linked Data technologies promise to solve these shortcomings, their integration into academic and research libraries remains unequal, and empirical studies assessing their impact in the management of next-generation library knowledge is still limited. This leaves a gap to critically and systematically assess how these technologies can be embraced to improve library knowledge interoperability, organization, and service delivery. The objectives of this study were to;

1. assess the main technologies and concepts of the Semantic Web and Linked Data and their application to managing knowledge in the library.

2. study the current uses and advantages of Semantic Web and Linked Data technologies in libraries.
3. evaluate the impact of Semantic Web and Linked Data technologies on resource interoperability, discovery, and service delivery in next-generation library knowledge management structures.
4. assess limitations associated with traditional library knowledge management systems in supporting semantic discovery, interoperability, and integration with the larger web.

Study scope

This research focuses on Semantic Web and Linked Data technologies usage in the context of academic and research library knowledge management. It studies standards, concepts, and structures like OWL, RDF, SPARQL, and Linked Data ideologies, alongside practical initiatives such as VIAF and BIBFRAME. The research is limited to evaluating the impact of these technologies in enhancing resource interoperability, discovery, and knowledge sharing but does not cover the wider digital library management aspects like IT infrastructure management or digitization workflows. Geographically, the study covers world, regional and Kenya specific contexts, highlighting global success and the need to implement such in Kenya.

METHODOLOGY

Research Design

The study was desk-based, applying systematic literature review method, entirely depending on secondary data acquired from online sources. This is a structured and transparent technique of establishing, choosing, and critically exploring existing professional and scholarly literature to answer given research objectives [19]. Compared to narrative reviews, systematic literature review clearly follows defined measures to mitigate bias while ensuring consistency in selection and analysis of sources. In this research, the technique entailed searching institutional repositories and academic databases for standards documentation, conference papers, policy reports, and peer-reviewed journal articles related to Semantic Web and Linked Data technologies consistent with library knowledge management. The sources were reviewed based on publication date, relevance, credibility, and alignment with the study aims. These

were then thematically analyzed and synthesized to unearth applications, trends, challenges, and existing gaps across the world, Africa, and Kenya. This study design is critical for evaluating evolving and unequally adopted technologies, where current evidence is distributed across regions and disciplines. As such, the study provides a thorough and evidence-based understanding regarding Semantic Web and Linked Data technologies role in next-generation library knowledge management.

Data Sources

The research exclusively relied on secondary data acquired from online institutional and scholarly sources. A total of 35 peer-reviewed journal articles were reviewed for significance to Semantic Web and Linked Data applications in library knowledge management. The sources included Google Scholar (https://scholar.google.com/) and open-access academic journals such as SCIRP, CABI Digital library, Code4Lib Journal, Journal of Critical Library and Information Studies, JLIS.it, Information Technology and Libraries, and Journal of eScience Librarianship. Authoritative policy and technical standards documents were acquired from reputable organizations like the Library of Congress BIBFRAME initiative (https://www.loc.gov/bibframe/) and the World Wide Web Consortium (https://www.w3.org/standards/semanticweb/). Case studies were acquired from international sources such as open repositories including arXiv (https://arxiv.org/abs/2301.11182) and Europeana (https://www.europeana.eu). Inclusion criteria was based on credibility and relevance to subject of study.

Data Analysis

Data was analyzed using a thematic synthesis and systematic literature review approach. Documents were screened, categorized, and analyzed depending on the four study objectives, emphasizing the Semantic Web and Linked Data technologies application in library knowledge management. Recurring themes were identified, including knowledge graphs, discovery enhancement, metadata transformation, interoperability, and implementation challenges. Themes were compared geographically, covering global, regional, and Kenyan settings to establish similarities, patterns, and differences in implementation and practice. Case

studies, conceptual frameworks, and empirical results were analyzed to extract insights on benefits, limitations and trends. Qualitative technique allowed for identification of knowledge gaps alongside informed conclusions regarding Semantic Web technologies' role in next-generation library systems.

Justification of Methodology

Use of desk-based systematic literature review enabled thorough and organized evaluation of Semantic Web and Linked Data technologies in the context of library knowledge management. The technique made it possible for transparent, rigorous, and reproducible analysis of secondary data sources across the world, regionally, and in Kenya ensuring results are based on credible evidence. Purposively selected peer-reviewed articles, policy reports, technical standards, and case studies enabled a holistic understanding of applications, trends and challenges of semantic web and linked data technologies in library management. Use of numerous data sources and extraction of themes improved study reliability and validity, consistent with best practices for studies on emerging technologies in library science and information management [20].

Ethical Considerations

The study, while relying on secondary data, maintained ethical considerations**.** Each data source was properly cited and referenced, maintaining academic standards. The extracted data was objectively presented, differentiating between interpretation by the author and a balance objective interpretation. Thirdly, only credible and publicly available information sources were used, mitigating plagiarism and misrepresentation of data. Finally, the methodology was transparent allowing users to trace the selection, analysis, and synthesis procedures while observing ethical requirements.

RESULTS AND DISCUSSION

Core Concepts and Technologies of Semantic Web and Linked Data

Recent advances confirm the importance of Semantic Web and Linked Data in enhancing organization of knowledge, even with constrained infrastructure and technical capacity. Table 1 below provides a summary of technologies used, purpose, example and source.

Table 1. Core Semantic Web & Linked Data Technologies in Libraries

Technology	Purpose	Example Use	Source
OWL	Used to define concepts and establish relationships	Ontologies for library metadata	Dutta, 2017
SKOS	Controlled vocabularies	Subject hierarchies	(Tamrazyan et al., 2024)
RDF	Representation of data as triples	Bibliographic and authority data linking	(Gaitanou et al., 2022).
BIBFRAME	Connected Data bibliographic model	Replacing MARC for work and instances	Boté-Vericad & Healy, 2022 [21]

These results demonstrate that Semantic Web and Linked Data technologies are key to **robust, interoperable, and scalable structures** for next-generation library knowledge management and forms an important foundation for improving knowledge discovery, integration, and service delivery in the world, Africa, and Kenya specifically.

Applications and Initiatives of Semantic Web and Linked Data in Libraries

The practical usage of Semantic Web and Linked Data technologies in research and academic libraries shows substantial disparity in adoption across global, regional, and Kenyan settings. Thus, Universities from developed countries with digital library platforms have initiated full-scale Linked Data solutions where the Library of Congress's BIBFRAME initiative substitutes MARC-based bibliographic metadata with RDF-based models, connecting instances, works, and agents to allow for interoperable, machine-readable data [14]. Equally, Europeana **(**2025)

[22] integrates cultural heritage resources from archives, museums, and libraries using RDF(S) and OWL. This improves both semantic discovery and integration with peripheral information such as Wikidata [22, 23].

Notably, in Africa, academic libraries are primarily piloting the semantic web and linked data technologies. An example is the case of the National Library of South Africa which employed RDF-based Linked Data to authority data, enhancing interoperability across countrywide repositories. Institutions of higher learning among them the University of Cape Town in South Africa and Moi University, Kenya have assessed semantic augmentation of digital collections, expressing early-stage usage in academic libraries [7,10]. Notably, however, regional enterprises such as AfriLIS Linked Data Forum enhance standards, training, and best practices but usually remain theoretical rather than operational [24].

In Kenya, the practices remain mainly conceptual, academic, or training-focused. Notably, however, research on metadata interoperability between MARC21, Dublin Core, and RDF shows rising awareness on Linked Data frameworks [25]. Workshops on RDF, BIBFRAME, and SKOS improve professional capacity, yet full-scale operational execution is scarce [26]. Relative to global and regional peers, Kenyan libraries have limited empirical evidence of connected Data use, stressing the importance for pilot implementations and infrastructure support [27]. Table 2 below summarizes the current application and initiatives of linked data as collected.

Table 2. Current Applications and Initiatives of Linked Data in Libraries

Context	Initiative	Scope	Semantic Technologies	Source
World	Europeana	Cultural heritage aggregation	RDF(S), OWL, SKOS	Europeana, 2025

World	Library of Congress – BIBFRAME	Full-scale bibliographic Linked Data	RDF, URIs, OWL	Sugimoto, 2022
Africa	National Library of South Africa	Authority records pilot	RDF, URIs	Alemu & Stevens, 2015; Gaitanou et al., 2022
Kenya	Metadata interoperability studies & workshops	Conceptual / Training	RDF, SKOS, BIBFRAME	(Ndegwa et al., 2022).

These results show a progressive adoption gradient where mature, full-scale initiatives are exhibited globally; pilot executions conducted in Africa; and training-oriented activities in Kenya. This stresses the vitality for Kenyan libraries to move from awareness to practical execution to attain the benefits observed in universal contexts.

Enhancing Library Services

Semantic Web and Linked Data technologies provide libraries with transformative capacity to improve interoperability, resource discovery, and service delivery. Universally, full-scale executions like Europeana and BIBFRAME express benefits like enhanced recommendation services, semantic search, and combination with external datasets, giving room for resource connection across disciplinary and institutional boundaries [28]. This is consistent with the perception that machine-readable Linked Data makes it possible for development of dynamic knowledge graphs, supporting advanced reasoning and discovery.

Regionally, the pilot initiatives disclose both benefits and challenges. Thus, from the National Library of South Africa's RDF-based authority records, the pilot enhanced interoperability and data sharing within national links; the impact was, however, limited due to insufficient infrastructure and technical capacity [9]. Likewise, academic projects undertaken at Moi University expressed improved metadata enrichment for digital collections, though standardization and sustainability are still barriers. These results demonstrate the need for regional libraries to integrate Semantic Web technologies to enhance knowledge management if funding and capacity gaps are closed [25].

In Kenya, merits are mainly observed or theoretical in controlled exercises. However, research works on metadata interoperability shows possible growth in semantic discovery and integration [25]. Nonetheless, there is minimal practical deployment, showing challenges including legacy systems, constrained technical expertise, and insufficient digital infrastructure [26]. These findings are summarized on table 3.

Table 3. Benefits and Challenges of Linked Data Adoption in Libraries

Context	Benefits	Challenges	Source
Global	Semantic discovery, integration with external platforms, interoperability,	Vocabulary alignment, data quality	Rahaman, 2021; Europeana, 2025
Africa	Improved authority control, enriched metadata	Infrastructure, technical skills, funding	Alemu & Stevens, 2015
Kenya	Professional capacity development, theoretical interoperability	Minimal deployment, limited expertise, legacy systems	Masinde & Sanya, 2022; Ndegwa et al. 2022

These results show that Semantic Web and Linked Data can significantly improve library knowledge management. However, realizing these benefits in the context of Kenya demands for infrastructure development, strategic capacity-building, and pilot implementations, consistent with local practice with universal standards.

Limitations of Traditional Library Knowledge Management Systems

While the traditional library knowledge management systems, grounded on AACR2, MARC cataloguing, and RDA, have managed print-based collections effectively, they experience substantial constraints in the current digital era. These structures are usually siloed, record-centric, and metadata-heavy but semantics-light, limiting interoperability and the capacity to support integration with web-based platforms [30]. Universally, existing research shows that such legacy systems struggle displaying meaningful links among works, subjects, authors, and formats, mitigating the effectiveness of modern resource discovery tools.

The piloted projects at Moi University in Kenya and the National Library of South Africa experienced difficulties in positioning bibliographic metadata and authoritative records across repositories, showcasing wider regional problems in embracing interoperable data structures [9,13].

In Kenya, university and research libraries heavily rely on traditional systems, specifically the MARC21 and Dublin Core, with limited capability to execute Linked Data and RDF models. This limits the capacity of libraries to provide semantic linking, context-aware search, or integrated access across heterogeneous resources, forming disparity between user expectations and system performance [31]. The section is summarized on table 4 below.

Table 4. Limitations of Traditional Library Systems

Context	Limitation	Impact	Source
Global	Record-centric, siloed metadata	Reduced semantic discovery	Y. Wang & Yang, 2018
Africa	Inconsistent standards, infrastructure	Limited interoperability	Alemu & Stevens, 2015; Ranatunga et al., 2025
Kenya	Reliance on MARC21/Dublin Core	Poor integration, limited semantic services	Tutu, 2017

These results show that traditional library systems are increasingly insufficient and inadequate for current knowledge management needs. This observation stresses the need to transition to Semantic Web and Linked Data technologies, which gives interoperable, more dynamic, and context-aware structures for next-generation library services.

CONCLUSION AND FUTURE WORK

This research sought to understand how Semantic Web and Linked Data technologies influence library knowledge management universally, in Africa and within Kenya. Results from the systematic literature review show successful implementation of semantic web and linked data technologies in academic and research libraries. This is common in developed countries where institutions have successfully integrated technologies like RDF, BIBFRAME, OWL, and SKOS, allowing for authoritative data and bibliographic to become part of an interoperable, machine-readable web of data [22]. Such applications enhanced discoverability of resources, allowed semantic search, and enabled integration with external knowledge links.

African academic and research libraries have made slow progress through academic research and pilot projects as evidenced by the National Library of South Africa, Moi University, Kenya and University of Cape Town, South Africa. The institutions have expressed semantic enrichment of authoritative records and metadata, but adoption remains constrained due to technical capacity, infrastructure constraints, and funding shortages [9,25]. Besides, implementation of the AI technology is mostly conceptual or training-focused in a number of Kenyan universities [31]. Factors limiting adoption include dependence on legacy systems, limited technical expertise, and inadequate digital infrastructure. This study confirms that whereas Semantic Web and Linked Data technologies provide significant capacity for improving library knowledge management, their implementation in Kenya and the African region is still at an early stage stressing the need for targeted implementation strategies, infrastructural support and technical capacity-building.

Limitations

The study, despite attaining its objectives, was subject to some limitations. First, study depended exclusively on secondary data. This limited insight into practical, empirical results of Linked Data implementation in academic and research libraries, specifically in Kenya. Second, the review mainly relied on official reports and academic publications, possibly ignoring unpublished initiatives that could otherwise inform adoption procedures. Finally, whereas the study did regional comparisons, it dwelt on libraries, leaving out associated institutions like museums and archives that may provide complementary insights. These

shortcomings suggest caution when generalizing the results to all library settings, specifically within developing countries.

Future Research Directions

Based on the results of this study, future studies should;

1. focus on empirical case study of Linked Data Adoption in Kenyan Libraries to measure factual enhancement in interoperability, resource discovery, and service delivery.
2. explore collaborative Linked Data structures which link public, academic, and national libraries across Kenya and Africa to assess interoperability and scalability.
3. investigate the technical knowhow of users and librarians and technical capacity training, investment, policy interventions and analyzing how these impact successful implementation of semantic technologies in resource-limited environment.
4. study the impact of metadata frameworks on library search efficiency, user experience, and access to knowledge, supporting evidence-based decision-making.

REFERENCES

1. Ramachandragouda, S., Naik, U., M, P. (2025). Advancing Library Services: Trends and technological transformations. *Zenodo (CERN European Organization for Nuclear Research)*.
2. Dione B., Rajaratnam, R. S. (2025). The Prospects Analysis for Transmogrification of Knowledge: Libraries Morphing for the Future Adaptation to Digital Dynamics. Journal of Computer and Communications, 13(04), 68–86.
3. Thagiah, R., Leo, M. J., Ramkumar, R., T, A. M. S. (2024). Digital transformation of academic libraries. In *Advances in library and information science (ALIS) book series* (pp. 307–328).
4. Gkrimpizi, T., Peristeras, V., Magnisalis, I. (2023). Classification of Barriers to Digital Transformation in Higher Education Institutions: Systematic Literature review. *Education Sciences*, *13*(7), 746.

5. Rhayem, A., Mhiri, M. B. A., Gargouri, F. (2020). Semantic Web Technologies for the Internet of Things: Systematic Literature Review. *Internet of Things*, *11*, 100206. Jones, C. (2024). Semantic Technologies in knowledge management. *European Journal of Information and Knowledge Management*, *3*(1).
6. Jones (2024). Semantic Technologies in knowledge management. *European Journal of Information and Knowledge Management*, *3*(1).
7. Alemu, G., Stevens, B., Ross, P., Chandler, J. (2012). Linked Data for libraries. *New Library World*, *113*(11/12), 549–570.
8. Library of Congress. (2022). *Annual report of the Librarian of Congress: Fiscal year 2022*. Library of Congress. https://www.loc.gov/static/portals/about/reports-and-budgets/documents/annual-reports/fy2022.pdf
9. Alemu, G., Stevens, B. (2015). *An emergent theory of digital library metadata: Enrich then filter*. Chandos Publishing.
10. Gaitanou, P., Andreou, I., Sicilia, M., Garoufallou, E. (2022). Linked data for libraries: Creating a global knowledge space, a systematic literature review. *Journal of Information Science*, *50*(1), 204–244.
11. Collins, C., Dennehy, D., Conboy, K., Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. International Journal of Information Management, 60, 102383.
12. Lu, Y., Lin, S. (2025). How can we promote digital transformation in college libraries? A study on readers' intention to adopt digital services. *Sustainability*, *17*(8), 3504.
13. Ranatunga, S., Ødegård, R. S., Jetlund, K., Onstein, E. (2025). Use of semantic web technologies to enhance the integration and interoperability of environmental geospatial data: a framework based on Ontology-Based Data access. *ISPRS International Journal of Geo-Information*, *14*(2), 52.
14. Tamrazyan, H., Boros, E., Kaplan, F. (2024). Developing a standardised vocabulary for Ukrainian epigraphy and expanding digital epigraphic resources. In *Lecture notes in computer science* (pp. 13–22).
15. Dutta, B. (2017). Examining the interrelatedness between ontologies and Linked Data. *Library Hi Tech*, *35*(2), 312–331.

16. Rahaman, W. (2019). Semantic Web-Linked data and libraries. In *Advances in library and information science (ALIS) book series* (pp. 27–42).
17. Nirudi, Y., Parichi, R. (2025). Semantic Web and the Libraries: An Overview. *SSRN Electronic Journal*.
18. Fred, G. G., Tom, K. (2015). The impact of Web 2.0 tools on academic libraries in Kenya. *International Journal of Library and Information Science*, *7*(2), 21–26.
19. Cheong, H., Lyons, A., Houghton, R., Majumdar, A. (2023). Secondary Qualitative Research Methodology Using Online Data within the Context of Social Sciences. *International Journal of Qualitative Methods*, *22*.
20. Riady, Y., Sofwan, M., Mailizar, M., Alqahtani, T. M., Yaqin, L. N., Habibi, A. (2023). How can we assess the success of information technologies in digital libraries? Empirical evidence from Indonesia. *International Journal of Information Management Data Insights*, *3*(2), 100192.
21. Boté-Vericad, J., Healy, S. (2022). Academic libraries and research data management. *Vjesnik Bibliotekara Hrvatske*, *65*(3), 171–193.
22. Europeana Foundation. (2025). *17th Semantic Web in Libraries Conference*. Europeana PRO. https://pro.europeana.eu/event/17th-semantic-web-in-libraries-conference
23. Sugimoto, G. (2022). Instance level analysis on linked open data connectivity for cultural heritage entity linking and data integration. *Semantic Web*, *14*(1), 55–100.
24. Eller, D. W. (2022). Transparency and the future of semantic searching in academic libraries. *Information Services & Use*, *42*(3–4), 453–461.
25. Ndegwa, H. M., Bosire, E., Damaris. O. (2022.). Evaluating Institutional Repositories' (IR) capabilities for long-term preservation with a focus on content, file format and metadata practices in selected public university libraries in Kenya. *Library Philosophy and Practice*.
26. Gitonga, F. N., Gichohi, P. M. (2024). Assessing the capacity building programs pursued by university libraries in Meru County, Kenya. *KLISC Journal of Information Science & Knowledge Management*.

27. Malekana, M. D., Chuma, K. G. (2015). Strengthening service delivery to support information needs of users at Springs Public Library in South Africa. *South African Journal of Libraries and Information Science*, *91*(2).
28. Rahaman, W. (2021). Semantic Web-Linked data and libraries. In *IGI Global eBooks* (pp. 167–181).
29. Masinde, J. M., Sanya, O. (2022). Analysis of interoperability, security and usability of digital repositories in Kenyan Institutions of Higher Learning. *Data and Information Management*, *6*(4), 100011.
30. Wang, Y., Yang, S. Q. (2018). Linked data technologies and what libraries have accomplished so far. *International Journal of Librarianship*, *3*(1), 3.
31. Tutu, J. M. (2017). Provision of digital reference services in academic libraries in Kenya: a review. *Regional Journal of Information and Knowledge Management.*, *2*(1).

TEXT MINING AND TOPIC MODELLING FOR KNOWLEDGE DISCOVERY IN DIGITAL COLLECTIONS

Munyao V.W., Ndung'u H., Wasike J., Tsinale H.

Kirinyaga University, KENYA

Correspondence: **vickymunyao@gmail.com**, hndoshkp15@gmail.com

ABSTRACT

Kenyan research and academic libraries have been experiencing challenges managing the growing collections of digital scholarly publications making it hectic for librarians to systematically organize resources and users to access relevant content efficiently. Text mining has evolved as a strong method for obtaining insights from large textual sets of data, allowing for thematic analysis, automated knowledge discovery, and enhanced organization of content. Nevertheless, its usage in the Kenyan academic and research library context remains limited. This study explored text mining and topic modeling application on scholarly publications using open access digital libraries. The study sought to identify research patterns, establish latent themes, and express how these techniques could enhance accuracy in knowledge access, management of time, and thematic organization in the context of Kenyan academic and research libraries. The study involved collecting and preprocessing a corpus of library science and digital library articles. This included tokenization, stop word removal, and standardization. Using word cloud visualization, prominent terms were identified, while TF-IDF analysis displayed distinctive keywords. Dominant topics along with related document groupings were done using Latent Dirichlet Allocation (LDA), permitting insight into thematic constructs across the corpus. Analysis showed existence of three major topics that include digital library systems, open access and repository management, and scholarly publishing practices. We identified frequent terms like "library," "digital," "metadata," and "repositories", stressing on the use of text mining technology-driven library systems. The techniques expressed that text mining could systematically support automated classification, display latent trends, and facilitate resource discovery in libraries. This study showcases the capacity of text mining to improve content organization and knowledge discovery globally, hence should be practiced in Kenyan academic libraries. Future studies need to explore integration of AI driven text mining with library systems to further improve thematic analysis and user accessibility.

Keywords: text mining, topic modelling, Latent Dirichlet Allocation (LDA), TF-IDF, knowledge discovery, digital libraries, Kenyan academic libraries

INTRODUCTION

Libraries have traditionally worked as fundamental institutions for preserving, organizing, and distributing recorded knowledge through controlled vocabularies, structured cataloguing systems, and human-facilitated indexing. However, the rapid growth in digital technologies, has altered library collections into institutional repositories that contain large volumes of scholarly and multimedia content and large-scale digital libraries [1]. Although this revolution has expanded access; it has also presented problems regarding efficient resource management, information overload, and effective knowledge discovery.

The exponential rise in digital content has made traditional metadata-driven techniques increasingly ineffective. While manual cataloguing and subject indexing are precise, they require huge resources, and are also difficult to scale in settings where large document volumes are continuously added [2]. Besides, there is complication consistently classifying documents considering the interdisciplinary nature of contemporary scholarship, as documents usually span diverse subject domains. Thus, digital library users often struggle to detect emerging research trends, establish relevant materials, and make a comprehensive understanding of extensive collections [3]. As a result, computational techniques on text analysis have attracted prominence in the context of library and information science. Text analysis entails automated extraction of themes, patterns, and connections from unstructured textual information through machine learning methods, specifically natural language processing [1]. As opposed to traditional indexing, text mining gives room for scalable analysis of large corpora and reveals latent semantic constructs without sole dependence on predefined metadata. This capacity is critically important in digital libraries where full-text documents are instantly available [4].

Among text mining methods is the topic modelling, specifically Latent Dirichlet Allocation (LDA) which is an effective technique for establishing hidden thematic patterns and structures within document collections [5]. In topic modelling, the multidimensional nature of texts is

acknowledged through allowing documents to be correlated with multiple subjects, supporting more complex representations of scholarly communication. From empirical studies, topic modelling aids in trend detection, collection analysis, and research evaluation, thus informing strategic decision-making and evidence-based collection development [6]. Besides, the rise of open-access repositories like CORE further highlight the significance of text mining for libraries. While the open platforms expand knowledge access, their scale comes with navigational problems. Notably, topic modelling and text mining provide scalable solutions for summarizing and organizing such collections [7].

The techniques, though limited to parameter selection and interpretability, are best viewed as complementary tools that improve, rather than substitute, professional librarianship. This study presents text mining concept and its application to library science, applying the LDA concept to extract topics and patterns from open-source libraries online highlighting the importance of Kenyan libraries embracing text mining techniques for quick and precise information access.

In the recent past, Kenyan academic and research libraries have increasingly adopted open-access and digital repositories podiums to support research, teaching, and innovation. However, the speedy expansion of full-text digital collections has outdone the capability of traditional indexing, cataloguing, and metadata-based retrieval systems regularly employed in these libraries. Consequently, users usually experience limited identification of relevant resources, information overload, and difficulty establishing developing study themes within open-access and institutional repositories such as CORE. Thus, while there exist advanced topic modelling and text mining methods, their practical usage in the context of Kenyan digital library remains low, with inadequate empirical research proving their efficacy for large-scale knowledge discovery. This gap limits the ability of researchers and librarians to methodically support evidence-based collection development, analyze content trends, and enhance access to knowledge by users. Accordingly, there is a need to assess how topic modelling and text mining can be used in digital collections to enhance knowledge discovery in Kenyan digital libraries.

Study Objectives

This study sought to examine use of text mining and topic modelling methods on digital library collections to improve content organization, automated knowledge discovery, and thematic analysis in the context of Kenyan academic and research libraries. Specific objectives were to;

1. collect and preprocess scholarly publications from open access digital libraries using text mining techniques.
2. employ topic modelling to establish research patterns and latent themes within the corpus.
3. create visualization and interpret the identified topics, establishing insights that can improve knowledge access accuracy and time management in the context of academic libraries in Kenya.

LITERATURE REVIEW

Topic Modelling Techniques for Knowledge Discovery

Topic modelling is a text mining technique that obtains latent thematic compositions from large collections of documents. It is specifically important for digital libraries as it allows documents to be presented as combinations of subjects, reflecting the multidimensional nature of academic information. The most widely used mining technique is Latent Dirichlet Allocation (LDA). It assumes that every document is acquired from a range of topics where each topic is distinguished by a distribution of words [8]. LDA is widely employed in the digital library studies to establish research trends, reveal hidden themes, and support automated content categorization [9]. The second technique is Non-Negative Matrix Factorization (NMF) and Latent Semantic Analysis (LSA); these have been applied to extract latent semantic structures in text corpora. These methods vary in their mathematical tactic but share the common goal of reducing dimensionality and unearthing existing themes [10].

Latest improvements in topic modelling have combined contextual embedding and deep learning to enhance topic consistency and interpretability. Techniques like neural topic and BER Topic models apply transformer-based embedding to obtain contextual meaning prior to

clustering words into topics [11]. Such methods are particularly important in handling vast, varied digital collections with significant variation in vocabulary and writing styles. Besides, topic modelling has been employed to detect topics dynamically, tracking topic evolution over time, providing intuitions into emergent research routes and alterations in scholarly focus [12]. Although topic modelling provides formidable means for knowledge discovery, its success relies on careful parameter tuning and accurate topic quality evaluation. Notably, while the advances are evident worldwide, Kenya has limited empirical evidence regarding adoption and performance of topic modelling in digital academic libraries. Thus, Existing studies mainly focus on well-resourced settings with robust computing infrastructure and high-quality metadata, which is not a reflection of the realities in many Kenyan institutions. Besides, there is insufficient research assessing application of topic modelling in the context of open-access repositories like CORE to support Kenyan research and academic libraries in refining knowledge discovery, collection management, and research trend analysis. This gap stresses the need for context-particular research that evaluates topic modelling effectiveness, interpretability, and practical utility in the context of Kenyan library digital collections.

Preprocessing and Feature Engineering in Text Mining

Text mining and topic modelling success mainly relies on actual preprocessing and feature engineering. Usually, raw text data from digital collections have noise including special characters, HTML tags, and irrelevant segments like footnotes or references. It is therefore important to preprocess data, involving cleaning to reduce noise and enhance quality of extracted characteristics [13]. In text mining and topic modelling, preprocessing steps involve tokenization, stemming, stop-word removal, and lemmatization. While tokenization involves breaking texts into meaningful units, stop-word removal removes common words carrying little semantic value such as “the”, and “and”. On the other hand, stemming and lemmatization reduce words to their base forms, enhancing alignment in the vocabulary used for analysis [14]. Notably, through feature engineering, cleaned data is transformed into numerical representations appropriate for modelling. Traditional techniques, specifically Term Frequency–Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) have been applied to represent document-term connections [15]. TF-IDF is especially efficient for

displaying important terms through balancing term frequency with general document occurrence. Nonetheless, these approaches may not capture semantic meaning and context. Therefore, current techniques employ word embedding like GloVe, Word2Vec, and contextual embedding from transformer models like BERT [16]. These embedding capture contextual usage and semantic similarity, improving retrieval accuracy and topic coherence.

Preprocessing decisions substantially impact topic modelling results. For instance, exceedingly aggressive stop-word removal may remove domain-specific terms; on the other hand, inadequate cleaning may result in poor topic quality. Thus, preprocessing needs to target the research objectives and characteristics of the digital collection. Yet in Kenya, there are limited studies on preprocessing techniques for digital collections, particularly considering the linguistic variety and contrasting metadata quality in local repositories. Digital libraries in Kenya usually contain documents having inconsistent formatting, mixed language content, and incomplete metadata, which may demand for a customized preprocessing method. Subsequently, there is a demand for empirical research that assesses the influence of different preprocessing and feature engineering techniques on topic modelling precision and interpretability in Kenyan digital library context.

Research Gap in Kenyan Digital Libraries

While there exist extensive studies on text mining and topic modelling in the context of digital libraries internationally, empirical research specific to Kenyan academic and research libraries are limited. Thus, published studies usually focus on developed nations where data availability and digital infrastructure are advanced, leaving a gap for understanding the influence of these methods in the context of Africa with unique infrastructural, linguistic, and resource constraints. Thus, whereas digital repositories are growing in Kenya, their metadata quality and search capacities remain weak, resulting in challenges on knowledge discovery. Besides, limited evidence exists regarding the practical usage of topic modelling for supporting library services like research trend analysis, collection development, and user centered discovery in Kenyan academic and research institutions. This study addresses the gap by evaluating application of text mining and topic modelling methods on open-access

digital collections, dwelling on how these techniques can enhance knowledge access in the Kenyan academic setting.

This study applied text mining and topic modelling methods to extract scholarly articles from open-access digital repositories, specifically CORE, analyzed the extracted articles in library and information science to discover knowledge structures, latent themes, and patterns. Notably, whereas CORE corpus is worldwide, results relate to improving collection management, knowledge discovery, and research support in the context of Kenyan academic libraries. The study excluded local repositories and proprietary databases because of access limitations. It however stresses the larger pertinence of computational text analysis in guiding decision-making, enhancing digital library services, and supporting evidence-based techniques in Kenya's library sector.

METHODOLOGY

Research Design

Research design was quantitative, where machine-learning text mining technique was applied to extract latent themes from digital information within the realm of digital libraries. The study employed a corpus-based computational workflow, integrating retrieval of documents, preprocessing, feature engineering, topic modelling, and model evaluation. Primarily, the study employed Latent Dirichlet Allocation (LDA) as the topic modelling algorithm considering its capacity to represent documents of diverse topics and themes as distributions over words [17]. The study dwelt on establishing main topics in digital library research, such as information retrieval, metadata standards, open access, digital preservation, and scholarly communication. The methodology supports automated knowledge discovery through transforming unstructured text into structured topic interpretations that can be computed and interpreted using statistical measures.

Data Collection and Sampling

Data was acquired from the CORE open-access repository through CORE API and the search query "library science OR digital libraries" using the link, https://api.core.ac.uk/. A total of 500 documents, published between 2015 and 2024, were sampled to capture modern research

trends. The articles were downloaded in PDF format and stored in a structured directory. In support of subsequent analysis and interpretation, metadata such as title, publication year, author, and DOI were recorded. Files were then converted to plain text using PyMuPDF, allowing computational processing. The sampling technique was targeted on ensuring adequately large corpus for robust topic modelling while keeping alignment to the research domain and evading exceedingly broad document retrieval.

Data Preprocessing and Feature Engineering

Data preprocessing was performed using two Python libraries, NLTK and spaCy. Normalization of texts entailed transforming all text to lowercase and removing numerical digits, punctuation, and special characters. The text was further split using tokenization, resulting in individual words. Thereafter, stopword removal was conducted to remove common English domain-specific stopwords like “doi”, “et al.”, and “https”. Tokens were reduced to their base forms through lemmatization, ensuring coherent representation of word variations. Through noise filtering, tokens having less than three characters and terms featuring in less than two articles or in more than half of documents were removed. Feature engineering was then conducted, involving developing a TF-IDF matrix and a Bag-of-Words (BoW) corpus. A dictionary was then generated using Gensim to map every peculiar token to an identifier; BoW representation also captured term frequencies in each document. Using Scikit-learn, TF-IDF was computed to stress on significant terms and mitigate common term weights. Feature matrices were then stored for input into the LDA model.

Topic Modelling and Parameter Tuning

Topic modelling was executed through Gensim library. Thus, using grid search over the number of topics, eta (topic-word density), alpha (document-topic density), and number of passes, the LDA model parameters were tuned. The candidate topic numbers ranged between 5 and 15, and the final number was chosen based on interpretability and coherence. Every model run resulted in document-topic distributions and topic-word distributions. The topics were categorized based on the highest likelihood terms, targeting subjects like open access, information retrieval, metadata standards, digital preservation, and research analytics. The

ultimate model was chosen based on attaining the highest consistency score while holding diverse topic separation.

Model Evaluation and Output Reporting

Model assessment entailed both quantitative and qualitative measures. Quantitatively, topic coherence (C_V), was used to assess model fitness and interpretability. Coherence values were calculated using Gensim's coherence model. Qualitative validation entailed manual review of top keywords and sampling documents related with every topic to validate thematic relevance. Reporting involved generation of non-interactive outputs, including a table of topics alongside their top keywords with odds, a table containing topic distributions in the selected documents, graphs displaying topic prevalence, and coherence scores across dissimilar topic numbers.

RESULTS AND DISCUSSION

Data Mining and topic identification

The mined set showed five research works focused on diverse library science aspects. The first document discussed open-source repositories alongside their implications for libraries; stressing the role of open access and resource sharing. The next 3 documents dwelt on digital libraries and technology in librarianship. They included topics such as distributed digital library management, digital librarian education, and conceptual structures of digital libraries. The documents expressed strong emphasis on organizational and technological dimensions of library systems. The fifth document examined publishing practices in the context of Digital Humanities, showcasing research on scholarly communication and knowledge dissemination as shown on table 1. This distribution show that the collection covers both theoretical and practical aspects of library science, with specific focus on digital library systems in the context of research and publishing. When applied in the Kenyan library context, the text mining concept can help library users locate documents based on relevant topics, reducing time waste on manual search of documents.

Table 5. Theme identification and classification

	text
0	Open source repositories: Implications for lib...
1	Computational sense: the role of technology in...
2	Managing complexity in a distributed digital l...
3	What is a digital library?: Possibilities from...
4	Publishing as Sharing: observations from Oral ...

Text Preprocessing and Tokenization

Through tokenization using document titles and abstracts, every text entry was broken down to a list of meaningful words also referred to as tokens; stopwords and punctuations were removed at this level. The resulting values are presented on table 2 below.

Table 6. Tokenization results

	text	tokens
0	Open source repositories: Implications for lib...	[open, source, repositories, implications, lib...
1	Computational sense: the role of technology in...	[computational, sense, role, technology, educa...
2	Managing complexity in a distributed digital l...	[managing, complexity, distributed, digital, l...
3	What is a digital library?: Possibilities from...	[digital, library, possibilities, idea, digita...
4	Publishing as Sharing: observations from Oral ...	[publishing, sharing, observations, oral, hist...

The table shows that in the first document, *"Open source repositories: Implications for libraries"*, was tokenized into ['open', 'source', 'repositories', 'implications', 'libraries'], seizing the main terms that exhibit its main aim. This was the same case with subsequent documents like the digital library systems and technology in librarianship where words like ['computational', 'technology', 'education'] or ['digital', 'library', 'distributed'] were retained, showing their organizational and technical themes. Words such as "the", "is", "for" were filtered out through tokenization and text reduced to core content words, which were then used for further analyses such as topic modeling, keyword extraction, and similarity comparisons. The tokenization activity practically demonstrates how text mining can automate content

organization and theme discovery, an aspect that can be adopted in Kenyan academic libraries to enhance knowledge access, search accuracy, and service efficiency.

Visualization of Frequent Terms with WordCloud

The next analytics, WordCloud, was developed using the above tokens, showing words based on their frequency. Words such as "implications", "education", "source libraries", "digital," "library," and "repositories" generally appear larger, showing their prominence as shown on figure 1 below. Further, this demonstrates how text mining can be applied to support Kenyan academic libraries through showcasing visually the dominant topics in collections, allowing fast content summarization, enhanced collection understanding, and more effective user guidance without manual review.

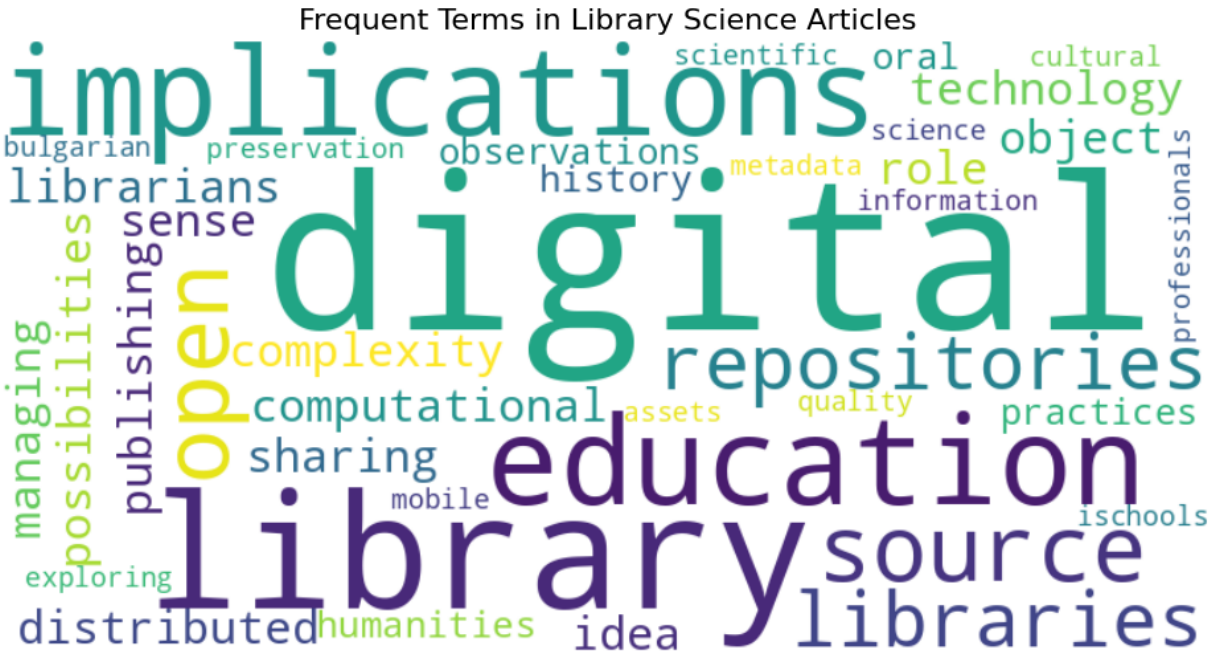


Figure 1. Work frequency

The words were further extracted using TF-IDF (Term Frequency–Inverse Document Frequency). Here, as shown on table 3, top 20 keywords like "library," "digital," "libraries," "education," "metadata," "computational," and "humanities" emerged to represent the core concepts across the library science articles. Here, words that are particularly distinctive in each document were highlighted, showing how text mining helps researchers and librarians

quickly understand the research focus areas, dominant themes, and specialized terminology within the collection.

Table 7. Text extraction

```
Top 20 TF-IDF Keywords:
['assets' 'bulgarian' 'complexity' 'computational' 'cultural' 'digital'
 'distributed' 'education' 'exploring' 'history' 'humanities' 'idea'
 'implications' 'information' 'ischools' 'librarians' 'libraries'
 'library' 'managing' 'metadata']
```

Document similarity

The concept of document similarity was further analyzed, which according to information presented in figure 1, shows that while documents shared key words, they were all low in similarity, expressing divergence of topics. It is, however, worth noting that documents like 2 and 3 had lowly average similarity, highlighting some level of consistency in research topics.

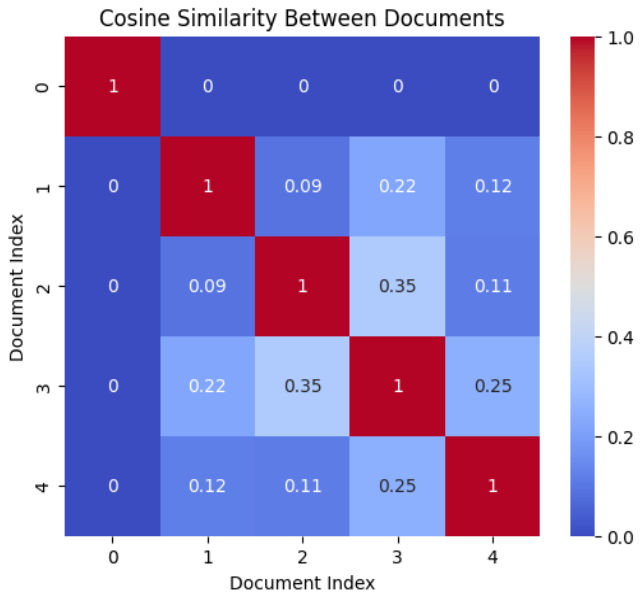


Figure 2. Document similarity

Coherence

The final aspect tested was coherence used to evaluate the model. Results show that the model attained a 0.4514 coherence score, which is an indication of moderate level of thematic alignment across the extracted topics. Further qualitative validation conducted manually on top keywords for every topic aligned with the quantitative coherence, showing sharing of words across documents, and key word dominance. These results provide explicit evidence regarding main knowledge and themes structures which exist in the digital library collection. This supports the conclusions on research trends, topic prevalence, and the potential for enhancing content organization and knowledge discovery in Kenyan academic libraries.

CONCLUSION AND FUTURE WORK

This study explored text mining and topic modeling method application in the context of digital libraries to enhance automated knowledge discovery, content organization, and thematic analysis regarding Kenyan research and academic libraries. Data preparation was done using curated corpus of digital library and library science articles to remove irrelevant terms and standardize data for analysis. Information was then tokenized to reveal key concepts like “metadata,” “library,” “digital,” “repositories,” and “education,”, showing the main themes in the collection. Visualization, done through WordCloud, highlighted the recurrent topics. Distinctive words, on the other hand, were identified using TF-IDF analysis on each document. Topics were then modelled, grouping the corpuses into meaningful groups, among them digital library systems, repository management and open access, and scholarly publishing practices. These findings demonstrate that text mining can methodically unearth latent trends and structures in large collections, providing librarians and researchers with actionable insights for enhancing resource organization and knowledge access. It thus shows that text mining, when correctly applied in the context of Kenyan research and academic libraries, can enhance accuracy and save time on document search.

Implications

Results demonstrate practical application of text mining in Kenyan academic and research libraries. It shows that librarians and researchers can effectively organize collections through automatic categorization of documents into topics; reducing duration spent on manual grouping. Use of visualizations like Word Clouds provides an immediate overview of dominant research themes making it easier for library users to navigate the materials and establish appropriate resources. Further, TF-IDF keyword analysis supports automated discovery of knowledge. This highlights specialized concepts and terms that may else be overlooked. For instance, themes associated with metadata management and digital library systems were clearly eminent, demonstrating the role of text mining in recommending readings, tailoring search results, and establishing gaps in library holdings. In general, these techniques improve content organization, discoverability, and thematic insight, addressing directly the needs of research and academic communities in Kenya.

Limitations

This study was subject to some limitations. To start with, text mining works on digital libraries. This becomes a challenge to academic and research libraries with purely manual operations like in the case of most Kenyan academic institutions. The study also extracted few documents from the open access library, limiting the number of topics extracted. Finally, the research dwelt on titles and abstracts, implying that full-text analysis could disclose additional insights that were not captured here. However, the findings are enough to express the practical utility of text mining techniques in academic libraries.

Future Research Directions

Future studies should focus on improving the practical usage of text mining in Kenyan libraries to mine many documents from diverse backgrounds to increase theme identification and improve the depth of topic discovery. There is also need to develop country specific search engines for Kenya to enable researchers and library users to search using key words instead

of manual document identification. Finally, there is need to explore interactive visualizations for topic distributions and keyword trends to aid researchers and librarians understand collection content and establish emerging areas of interest.

REFERENCES

1. Naik, S. (2025). Libraries in the Digital Era: An overview on the role of ICT and Emerging technologies. *International Journal of Creative Research Thoughts*, *13*(3).

2. Rahmanova, A. (2025). Evolution of libraries in the Digital Era: Redefining access, education, and cultural preservation. *Library Archive and Museum Research Journal*, *6*(1), 23–38.

3. Yang, W., Fu, R., Amin, M. B., Kang, B. (2025). The impact of modern AI in metadata management. *Human-Centric Intelligent Systems*, *5*(3), 323–350.

4. Alexakis, T., Adamopoulou, E., Peppes, N., Daskalakis, E., Ntouskas, G. (2025). Evaluating data quality: comparative insights on standards, methodologies, and modern software tools. *Electronics*, *14*(15), 3038.

5. Jiang, C., Zhang, L., Wang, Y. (2025). Integrating natural language processing techniques with LDA topic modeling: Deconstructing the evolution of media discourse on educational modernization ideology. *Education and Information Technologies*, *30*(16), 23419–23466.

6. Chang, I., Yu, T., Yu, T., Chang, Y., Yu, T., Yu, T. (2021). Applying text mining, clustering analysis, and latent dirichlet allocation techniques for topic classification of environmental education journals. *Sustainability*, *13*(19), 10856.

7. Zimmermann, J., Champagne, L. E., Dickens, J. M., Hazen, B. T. (2024). Approaches to improve preprocessing for Latent Dirichlet Allocation topic modeling. *Decision Support Systems*, *185*, 114310.

8. Theocharopoulos, P. C., Anagnostou, P., Georgakopoulos, S. V., Tasoulis, S. K., Plagianakos, V. P. (2025). Large language models for efficient topic modeling. *Neural Computing and Applications*, *37*(29), 24421–24439.

9. Egger, R., Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2VEC, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, *7*, 886498.

10. Mifrah, S. (2020). Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 5756–5761.

11. Vangara, R., Bhattarai, M., Skau, E., Chennupati, G., Djidjev, H., Tierney, T., Smith, J. P., Stanev, V. G., Alexandrov, B. S. (2021). Finding the number of latent topics with semantic non-Negative matrix factorization. *IEEE Access*, *9*, 117217–117231.

12. Gupta, P., Ding, B., Guan, C., Ding, D. (2024). Generative AI: A systematic review using topic modelling techniques. *Data and Information Management*, *8*(2), 100066.

13. Churchill, R., Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, *54*(10s), 1–35.

14. Pradana, A. W., Hayaty, M. (2019). The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on Indonesian-language texts. *Kinetik Game Technology Information System Computer Network Computing Electronics and Control*, 375–380.

15. Maiti, A., Abarda, A., Hanini, M. (2024). The impact of feature extraction techniques on the performance of text data classification models. *Indonesian Journal of Electrical Engineering and Computer Science*, *35*(2), 1041.

16. Krawczyk, N., Probierz, B., Kozak, J. (2024). Towards AI-Generated Essay Classification using numerical text representation. *Applied Sciences*, *14*(21), 9795.

17. Chatterjee, R., Mukherjee, C., Chatterjee, S., Nath, B. (2024). Latent Dirichlet allocation for topic modeling and intelligent document classification. In *Lecture notes in networks and systems* (71–83).