

AI-Driven Cybersecurity Architecture for Saudi Smart Cities: Enhancing Threat Detection and Digital Resilience under Vision 2030

MOHAMED SHAFRAZ FAREEGUL NIZAM
[0009-0003-6522-4662](https://doi.org/10.64388/IREV10I4-1723862)

Abstract- As part of Saudi Arabia's smart-city agenda, Internet of Things devices, cloud platforms, cyber-physical systems, edge computing and artificial intelligence are increasingly being used to run urban services. This convergence poses a security challenge that cannot be addressed by traditional perimeter controls alone: attacks can spread across heterogeneous domains, exploit model and data dependencies, and disrupt critical services even when individual components are technically available. This paper presents an evidence-based architecture for AI-driven cybersecurity in smart cities in Saudi Arabia with a focus on threat detection and digital resilience within the framework of Vision 2030. This is a systematic integrative review of peer-reviewed research primarily published between 2020 and 2025 that synthesises evidence on intrusion detection, federated learning, edge intelligence, adversarial robustness, explainable AI, digital twins, smart-city risk, and Saudi cybersecurity readiness. The synthesis highlights the need for distributed detection at the data source, cross-domain correlation, explainable risk scoring, policy-constrained automation, and a resilience feedback loop connecting incident recovery and model assurance as key to effective architectures. Existing work shows strong potential for detection, but remains fragmented by domain, dataset and evaluation method; accuracy is often privileged over latency, recovery, privacy, resource cost and organisational governability. The paper thus proposes a layered architecture and resilience lifecycle that combines technical detection with the goals of governance and service continuity. The overall implications for Saudi deployments are that AI should be considered as an adaptive control capability in a governed socio-technical system, not just as a classifier. Future work should validate cross-city generalisation, adversarial robustness, federated collaboration, human supervision, and recovery-oriented metrics using Saudi data that are representative of operations.

Keywords: Artificial Intelligence, Cybersecurity, Smart Cities, Saudi Arabia, Vision 2030, Intrusion Detection, Cyber Resilience, Federated Learning, Digital Twins

I. INTRODUCTION

Smart cities turn sensing, connectivity and data analytics into decisions that affect transportation, utilities, healthcare, public safety, buildings and digital government. Their value is in integration: data from distributed assets are integrated via communication networks and computational platforms to coordinate services at the scale of the city. That same integration changes the nature of cyber risk. Compromise of an Internet of Things endpoint can lead to cloud resources. Manipulation of telemetry can confuse automated control. Disruption of communications can affect multiple services concurrently. An attack on a machine learning pipeline can change the behaviour of systems that look operational. Reviews of smart-city security are therefore increasingly framing cybersecurity as an urban governance and resilience problem rather than a narrow network-protection task [1,25,26]. This is especially the case when critical services and platforms facing citizens are being digitised at a fast pace.

A consequential context for this transition is provided by Saudi Arabia. Vision 2030 has accelerated digital government, smart infrastructure and data-intensive urban development. Saudi studies also report progress in smart-city capabilities but continue to report concerns about cybersecurity adoption, competency, institutional coordination and citizen trust [17–22]. The national context is relevant, since architecture cannot be simply transplanted by choosing the best performing intrusion-detection algorithm. Urban systems vary in asset criticality, procurement arrangements, cloud and edge dependencies, legacy operational technology, data residency requirements, workforce maturity and tolerance for automated response. Thus, an architecture suitable for cities in Saudi Arabia is required to connect technical detection

to risk ownership, service continuity, and decision accountability.

Artificial intelligence is attractive because conventional signature-based controls are not well suited for highly variable and previously unseen activity. Machine learning and deep learning can learn traffic patterns, detect anomalies and prioritise big volumes of telemetry multivariate. Scalable IoT detection and deep-learning benchmarks [2,4-8], fog-cloud processing, federated collaboration and adversarially trained models. But these advances do not mean resilience is automatic. Models can fail under distribution shift, unbalanced attack classes or deliberate adversarial perturbation; centralised data collection can cause privacy and sovereignty problems; edge devices have memory and energy limits; opaque scores can diminish operator confidence. Thus, a high point accuracy on a benchmark is not enough evidence that an urban security architecture will maintain critical functions in a real incident. This review fills that gap by treating threat detection and digital resilience as coupled architectural goals. It integrates technical and governance literature into an interconnected design where data collection, distributed AI detection, contextual correlation, explainable risk assessment, orchestrated response and post-incident learning run as a controlled loop. The design does not claim Saudi implementation experimentally validated. Rather, it finds what is currently supported by evidence, where there is a lack of evidence, and what design requirements should inform empirical validation. The net contribution is a review-based architecture that shifts the focus from individual model performance to reliable urban security operations.

II. AIM OF THE STUDY

The study seeks to develop a critically synthesised cybersecurity architecture for Saudi smart cities based on evidence of how AI-enabled threat detection is integrated with governance, response orchestration and recovery processes to improve digital resilience under Vision 2030. The study is about relations in architecture and implementation requirements and does not intend to introduce a new classifier or empirical performance claims. It asks what capabilities need to coexist for AI detection to remain

useful, given heterogeneous, privacy-sensitive, resource-constrained and operationally critical urban systems.

III. RESEARCH OBJECTIVES

The review has five objectives: (1) to highlight the main cyber-risk properties of interconnected IoT, cloud, edge and cyber-physical smart-city systems; (2) to compare AI-based detection approaches for scalability, privacy, robustness, interpretability and deployment location; (3) to establish how threat intelligence, explainable risk assessment and automated response can be combined so as not to compromise governance or human accountability; (4) to translate the evidence into a layered cybersecurity architecture that meets Saudi smart-city resilience needs; and (5) to specify research and evaluation priorities for validating the architecture using operationally representative data, cross-domain scenarios and recovery-oriented metrics.

IV. RESEARCH QUESTIONS

RQ1. Which smart-city attack surface components are most influential in the design of AI-based threat detection? RQ2. How to combine centralised, edge, fog and federated learning approaches to balance the detection quality, latency, privacy and resource constraints? RQ3. What controls are necessary to make AI-assisted response explainable, adversarially robust, and operationally accountable? RQ4. How to translate detection outputs to measurable digital resilience of critical Saudi urban services? RQ5. What empirical gaps must be filled before an architecture derived from a review can be considered deployable at city scale?

V. REVIEW METHODOLOGY

The study uses a structured integrative review. The approach was chosen because the research question covers heterogeneous evidence: experimental intrusion-detection studies, architectural frameworks, smart-city risk analyses, digital-twin studies, privacy-preserving learning, and Saudi studies concerned with adoption, capability, and governance. The studies use different datasets, threat taxonomies, performance measures and units of analysis, hence a narrow meta-analysis is not appropriate. The method rather

emphasises transparent selection, critical comparison, and architectural synthesis. The attached reference paper was used only for guidance in terms of structure and academic writing. Its conceptual framework, wording, tables and figures were not reproduced.

Searches were built for Scopus, Web of Science Core Collection, IEEE Xplore, ScienceDirect, SpringerLink and ACM Digital Library. Search expressions combined four concept groups: smart-city context (smart city, urban infrastructure, cyber-physical system, IoT, smart grid); cybersecurity function (intrusion detection, anomaly detection, threat detection, cyber resilience, incident response); AI technique (machine learning, deep learning, federated learning, explainable AI, adversarial learning, edge AI); and Saudi relevance (Saudi Arabia, Vision 2030, Saudi smart cities, digital government). The search was restricted to publications from 2020-2025, however closely related peer-reviewed work was included if it provided a needed architectural concept. Reference chaining was used to identify directly relevant studies that might have been missed because of differences in terminology. Inclusion criteria were research published in peer-reviewed journals or conferences with a clear contribution to the field of smart-city cybersecurity, IoT/CPS intrusion detection, AI robustness, distributed learning, digital resilience, or Saudi smart-city implementation. Studies also had to provide enough methodological or conceptual detail to allow for comparison. Exclusion removed non-scholarly commentary, papers that were purely cryptographic in nature with no architectural relevance, generic AI studies with no relation to cyber defence, duplicate accounts of the same work and publications in which claims could not be linked to identifiable methods or design implications. We do not present a numerical screening flow as the review does not claim to be exhaustive and systematic, and fabricated counts would create false precision.

The quality assessment was qualitative and multi-dimensional. The methodological transparency, data or scenario realism, reproducibility of the described method, appropriateness of evaluation metrics, external validity, treatment of privacy or adversarial conditions, and relevance to urban operational resilience of each candidate study were evaluated. A study based on a single benchmark reporting accuracy

without class sensitive metrics, omitting computational cost or generalising from one IoT domain to an entire city was interpreted with more caution. Saudi context studies were reviewed for technical evidence, organisational evidence, or adoption/governance evidence. The categories were not mutually exclusive. Synthesis was in two stages: First, we coded studies by theme: attack surface and risk, distributed detection, model robustness, privacy and collaboration, explainability and operator trust, digital twins and cyber-physical context, Saudi readiness, and resilience measurement. Second, repeated capabilities were abstracted into architectural layers, which were then mapped to operational outcomes. Contradictions were retained instead of averaging out. Centralised learning may ease the model management but increases data concentration, while federated learning reduces raw-data movement but brings non-IID data, model poisoning and coordination problems [4,7]. This comparative logic is the basis for the proposed architecture.

VI. THEMATIC LITERATURE REVIEW AND CRITICAL ANALYSIS

6.1 Smart-city architecture attack surface and risk.

It is hard to secure a smart city since the networks on which urban services are built have been developed under different assumptions. Reviews suggest that interdependencies between IoT endpoints, communications, data platforms and physical services are increasing, creating pathways where local compromise can result in system-wide impacts [1,26]. Risk is fluid: devices come and go, firmware ages, cloud services change, and traffic baselines shift with events and seasons. Kalinin et al. show that smart city risk assessment has to deal with dynamic device-to-device infrastructures instead of static inventories [25]. Similarly, Bibri et al. point out that the convergence of AI, IoT and big-data infrastructures offers opportunities for sustainability but also governance costs [30]. The implication is architectural. Identical anomalies can have very different consequences in a public Wi-Fi network and an energy-control path, so threat detection should consume asset criticality and service context, not just packet features.

6.2 Resource-aware and Distributed AI Detection.

Early evidence for scalable urban intrusion detection emphasises placement of analytics close to resource constrained IoT devices. Rahman et al. [2] propose a scalable machine-learning IDS based on collaborative edge/fog processing, while Roy et al. [6] suggest distribution of detection at fog and cloud layer to reduce service delay and better utilisation of constrained resources. Deep learning benchmarks illustrate that the choice of model and dataset characteristics matter, and no architecture is always superior [5]. Dina et al. also show that the imbalanced attack classes can distort the conventional training and loss design for the minority attacks [9]. In concert, these studies advocate a tiered detection plane: lightweight local inference for immediate signals, more powerful regional or cloud analytics for complex correlation, and model-selection rules based on latency and criticality, rather than a single, city-wide model.

Resource efficiency is not just an engineering convenience. In urban deployments you can have very large numbers of devices and heavy models can compete with the applications they protect. For strong IoT monitoring, Zakariyya et al. deal with memory and execution cost directly and rationalise the need to test security models against computational constraints [11]. Saudi architecture should categorise assets according to inference capacity and consequence. Network aggregation points may support rich models at a high rate, while sensors need light-weight feature extraction or delegated analysis. This arrangement also allows graceful degradation: if cloud connectivity is lost, critical edge detection can continue instead of having a complete loss of visibility.

6.3 Privacy-preserving collaboration and federated learning

Cross-district and cross-operator learning enhances citywide detection, but centralising raw telemetry can violate privacy, data minimisation, and organisational boundaries. Campos et al. point out the challenges of non-IID data and aggregation and identify federated learning as a promising approach for IoT IDS [4]. Collaborative detection is extended by Sarhan et al. with hierarchical federated learning and blockchain-based coordination [7]. Priyadarshini explores federated and split learning for smart-city cyberattack

anomaly detection [22]. These approaches support a design principle of moving model updates instead of unrestricted raw data. But federated learning should not be positioned as a privacy guaranty, in and of itself. Model updates may reveal information, compromised participants can poison training and heterogeneous traffic can lead to unstable global models. Federated collaboration therefore demands participant authentication, secure aggregation, update validation, provenance, and rollback.

6.4 Adversarial Robustness and Model Reliability

Security decisions based on AI components are part of the attack surface. Rashid et al. show the potential for adversarial examples to greatly decrease the effectiveness of the detectors in IoT smart-city applications and the use of adversarial retraining to improve the resistance [8]. Lilhore et al. [12] studied the problem of intrusion detection in IoT and 5G environments, and emphasised the need for models that can handle dynamic communication contexts. An important socio-technical dimension is given by Zeng et al. who link AI anomaly detection to technological readiness and human engagement [29]. Model assurance should therefore include drift monitoring, adversarial testing, calibration, version control, dataset provenance and safe fall back behaviour. A robust system needs to identify when model confidence is unreliable. Otherwise automation can convert uncertainty into damaging response actions.

6.5 Explainability, correlation and decision quality.

Alerts from different technologies and domains are ingested into a city security operations centre (city SOC). Often, the bottleneck in the architecture is not the detection, but the decision on what signals describe the same incident and what intervention is justified. Protic et al. [27] study contradictory anomaly decisions in smart-city network behaviour and expose the problem of inconsistent detectors. Explainability matters here not as a cosmetic feature but as an operational interface: analysts need contributing features, affected assets, confidence, temporal context, and plausible attack paths. Contextual correlation can also help to alleviate alert fatigue by aggregating low-level anomalies into service-level hypotheses. Thus the proposed architecture decouples detection from risk adjudication. Models can generate scores, but a correlation layer should combine model output with

asset criticality, identity events, threat intelligence and policy constraints before response.

6.6 Contexto ciberfísico y gemelos digitales.

Why do smart-city attacks matter? Because digital manipulation can affect physical outcomes. Mohammadpourfard et al. directly demonstrate this in smart grids, where an LSTM detector is designed to distinguish attacks from legitimate dynamic changes [3]. [15] Meng and Zhu integrate IoT, blockchain and digital-twin concepts in smart urban energy cybersecurity. [16] El-Hajj considers digital twins in parallel with IDS for smart-city infrastructure protection. Alnaser et al. review AI-powered digital twins and IoT for sustainable built environments [24]. These studies show that digital twins can enhance security analytics with expected physical state and plausibility checks that pure network monitoring cannot offer. However a twin is a data dependent model itself and poor synchronisation may produce false confidence. Therefore, security architectures should consider twins as contextual evidence, not as a source of truth.

6.7 Saudi smart city readiness and governance.

Saudi-specific literature finds that technical capability must be matched with the capacity to adopt and

institutional capacity. Alhalafi and Veeraraghavan conceptualise the challenges of cybersecurity adoption in smart cities of Saudi Arabia [17]. Alhakami assesses the development of cybersecurity competency and capability [20]. Aldegheishem [21] reports progress in Saudi smart-city dimensions.. Alsabt et al. synthesise the role of AI in smart-city optimisation as well as highlight challenges of integration relevant to Saudi urban planning [18]. The studies do not validate any specific IDS, but they are critical for architecture because they show that deployment outcomes depend on skills, governance, trust, and coordination. AI security that is not auditable, staffed or integrated with service owners is not resilient, regardless of whether it's 99% accurate in the lab or not. Table 1 summarises evidence across architectural capabilities. There is a consistent pattern: the technical literature is strongest on detection methods and weakest on cross-domain operational recovery, whereas Saudi literature is strongest on context, adoption and governance, but offers less city-scale validation of AI cybersecurity pipelines. The proposed architecture is intended to bridge those evidence sets rather than to treat one as a surrogate for the other.

Table 1. Evidence synthesis by architectural capability

Capability	Representative evidence	Strength of evidence	Principal limitation	Saudi smart-city relevance
Edge/fog detection	[2,6,10,11,13,14]	Strong technical evidence for lower-latency, resource-aware inference	Benchmark and device heterogeneity limits transferability	Supports distributed visibility when cloud connectivity or bandwidth is constrained
Federated collaboration	[4,7,22]	Promising privacy-aware collaboration across data owners	Non-IID data, poisoning, update leakage and coordination overhead	Enables shared learning without routine centralization of raw operator telemetry
Adversarial/model robustness	[8,12,28]	Clear evidence that model robustness and architecture choice affect IDS reliability	Robustness often evaluated in narrow scenarios	Requires red-teaming, drift monitoring, rollback and model assurance before automation

Capability	Representative evidence	Strength of evidence	Principal limitation	Saudi smart-city relevance
Contextual/XAI decision support	[17,20,27,29]	Strong rationale for explainability, readiness and trust	Limited evidence linking explanations to better SOC outcomes	Important for accountable escalation across municipal and critical-service owners
Cyber-physical/digital twins	[3,15,16,24]	Useful for linking cyber signals to expected physical behavior	Twin fidelity and synchronization can themselves fail	Relevant to energy, mobility, buildings and other safety-sensitive urban domains
Governance and urban readiness	[18-21,23,26,30]	Strong contextual evidence that governance shapes smart-city outcomes	Less direct validation of city-scale AI detection pipelines	Connects technical controls to capability, procurement, digital government and resilience

VII. DISCUSSION: PROPOSED ARCHITECTURE FOR SAUDI SMART CITIES

The synthesis supports a layered architecture rather than a monolithic intelligent IDS. Figure 1 begins with the urban attack surface that is a source of diverse security events from IoT, cyber-physical and operational technology, cloud services, edge platforms and digital applications. A telemetry and trust layer normalises network flows, logs, identity events and sensor state with provenance. Detection is then distributed: local models detect time sensitive anomalies close to assets, while regional or cloud components provide deeper analysis. The design combines the latency benefits of edge/fog processing [2,6] with the collaborative nature of federated learning [4,7,22]. It also puts limits on the premise that all security data aggregated is valid or necessary. The next layer turns detection into defensible decisions. Cross domain correlation should resolve conflicting outputs, relate anomalies to asset criticality, and provide explanation of risk scores. This addresses the trust issues embedded in the Saudi adoption literature [17,20] as well as the operational problem identified by Protic et al. [27]. First, contextual assessment. Second, response orchestration. Automation is useful for bounded actions such as rate limiting, session revocation, or

quarantining a non-critical endpoint when confidence is high. Evidence, pre-set safety boundaries and human approval should be required for actions that could disrupt critical physical services. This graduated autonomy is critical, as a false positive in city infrastructure can itself be a service disruption. A resilience plane exists across the architecture. Resilience is about more than just blocking attacks, it's about keeping essential functions running, restoring trusted state, learning from validated incidents and improving future response. This is illustrated in Figure 2 as a lifecycle of sensing, detecting, assessing, containing, recovering, learning and governing. This loop should only see model updates once incident labels and remediation outcomes have been validated. Otherwise, future models could be polluted by poisoned or misclassified events. Digital twins are good in comparing observed behaviour with expected physical states [3,15,16,24]. However, assumptions and the quality of the synchronisation should be validated as any other analytical component.

The architecture also changes the way effectiveness should be measured. Accuracy, precision, recall and F1-score are still needed for detection, but they don't tell us if the city can sustain or restore services. Evaluation should therefore address: false positive load, detection and decision latency, mean time to contain, mean time to recover, service availability

during attack, analyst workload, model drift, compute and energy cost and privacy exposure. For a time-critical service a detector that improves recall slightly but increases response latency threefold may be worse. In contrast, a lightweight edge model with lower standalone accuracy may add substantial resilience by maintaining minimum detection capability during network partition.

The multi-objective view is particularly important for Vision 2030. Smart-city cybersecurity should enable trusted digital services, not innovation based on obscure centralised control. Cities in Saudi Arabia may share common architectural principles but also allow domain-specific implementations for energy, mobility, buildings, healthcare and government platforms. So the architecture specifies interfaces and assurance obligations, not a single algorithm. Models can evolve as new evidence emerges, as long as telemetry provenance, validation, explainability, policy constraints and recovery procedures are explicit.

Another architectural implication is management of dependencies. Smart-city security controls rely on time services, identity providers, configuration repositories, software supply chains, and communication links that may all fail together during a coordinated incident. Resilience analysis should therefore identify common dependencies and security functions that need local fallback. When external enrichment or cloud analytics are not available, detection models must have a minimum operating profile. Response playbooks should also separate loss of visibility from lack of attack. If the telemetry disappears suddenly, the system should treat that condition as an observable resilience event, not silently reduce risk. Using fault injection and controlled cyber ranges in operational testing allows exploration of whether detection, escalation and recovery survive degraded connectivity. This approach also helps procurement, since vendors can be asked to show behaviour under partial failure, not just normal operation. The architecture should preserve tamper evident audit trails linking model version, evidence, analyst decision and action. This traceability supports incident reconstruction, accountability and safe learning after recovery. It also helps organisations determine whether an adverse outcome was the

consequence of bad data, model error, policy design, integration failure or operator override. These differences are required to improve the system without assuming that every failure indicates the need for a more complex model. In many cases, resilience may be improved more by simplifying dependencies, strengthening identity controls, or redesigning escalation paths than by increasing model depth. Architecture governance should therefore consider the total control system as an evolving socio-technical service and phase out analytical components that do not prove their operational value. The second implication is that resilience testing needs to mirror the order of decisions taken in an incident. Standard validation is typically to freeze a dataset and see if a model gets the right label. Operational cyber defence is sequential. Alerts change the intensity of monitoring. Containment changes the pattern of traffic. Recovery changes the state of assets. Those interventions change the data available to subsequent models. Therefore, evaluation should include scenario-based tests where detectors, analysts, response playbooks and service owners interact over time. Metrics should measure not only whether an attack was detected, but also whether the right asset was isolated, whether critical functions were available, whether the restoration introduced new vulnerabilities, and whether subsequent model updates captured the verified lessons of the incident. This sequential view can expose automation cascades — where a technically correct alert causes a needlessly disruptive response. And it can show where human approval leads to unacceptable delay. Saudi smart-city programmes could use cyber ranges and digital-twin environments for such rehearsals before controls go into operational networks. Scenario libraries need to include mobility peaks, major public events, utility disturbances, cloud outages, compromised credentials, supply-chain updates and coordinated attacks that bridge the digital and physical domains. Each scenario should set safety boundaries and desired service priorities, but not prescribe a single technical response. It lets vendors and operators show resilience, while maintaining architectural diversity. Repetition of exercises can also provide better evidence for governance decisions by identifying controls that reduce recovery time, analyst workload or cross domain uncertainty. In time, those observations can inform a maturity model based on demonstrated

operational outcomes, not nominal possession of security technologies. The evidence should guide continuous improvement.

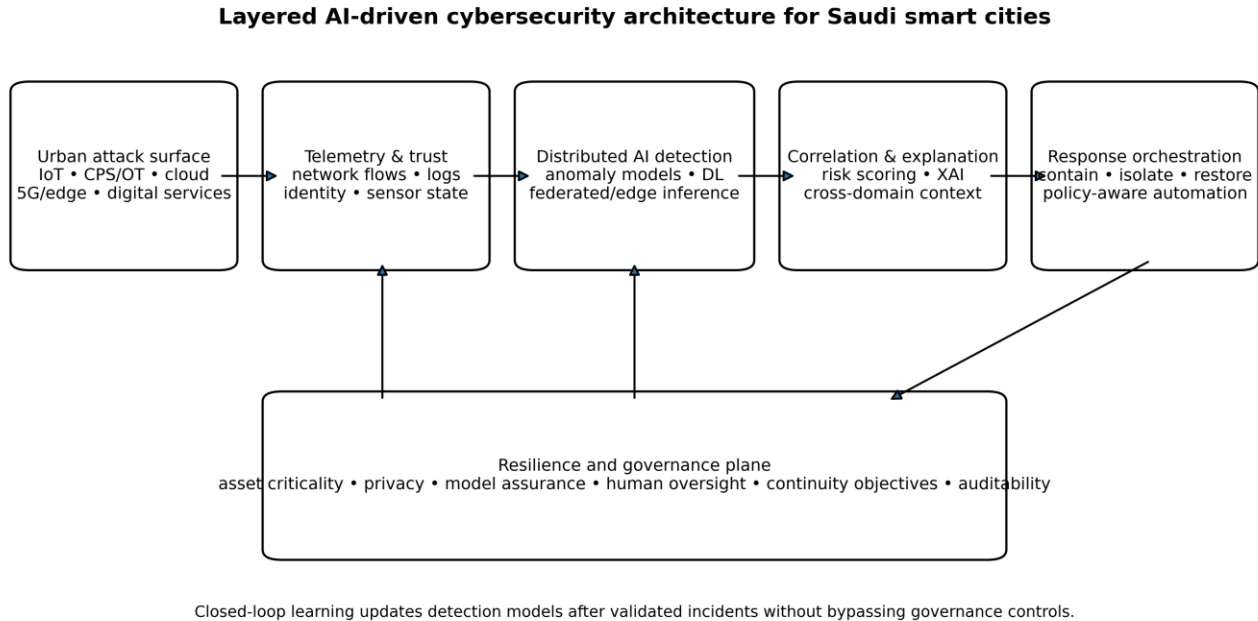


Figure 1. Review-derived layered AI-driven cybersecurity architecture for Saudi smart cities.

Table 2. Architecture design requirements and evaluative KPIs

Architecture layer	Control objective	Illustrative AI/control method	Resilience KPI	Governance concern
Telemetry and trust	Preserve reliable, attributable observations	Normalization, identity correlation, provenance checks	Telemetry completeness; time integrity	Data minimization; retention; ownership
Distributed detection	Identify known and unknown malicious behavior	Lightweight ML, deep models, edge/fog inference	Detection latency; recall; FPR	Model scope; versioning; resource cost
Collaborative learning	Share defensive knowledge across operators	Federated/hierarchical learning	Cross-site generalization; update latency	Poisoning resistance; privacy leakage
Correlation and explanation	Convert alerts into contextual risk	Ensembles, XAI, attack-path correlation	Analyst decision time; alert reduction	Auditability; confidence communication
Response orchestration	Contain threats without unsafe disruption	Policy engine, bounded automation, SOAR playbooks	Mean time to contain; service availability	Human authorization; rollback; safety
Recovery and learning	Restore trusted state and improve future controls	Post-incident validation, drift monitoring, retraining	MTTR; recurrence rate; model drift	Evidence quality; approval of model updates

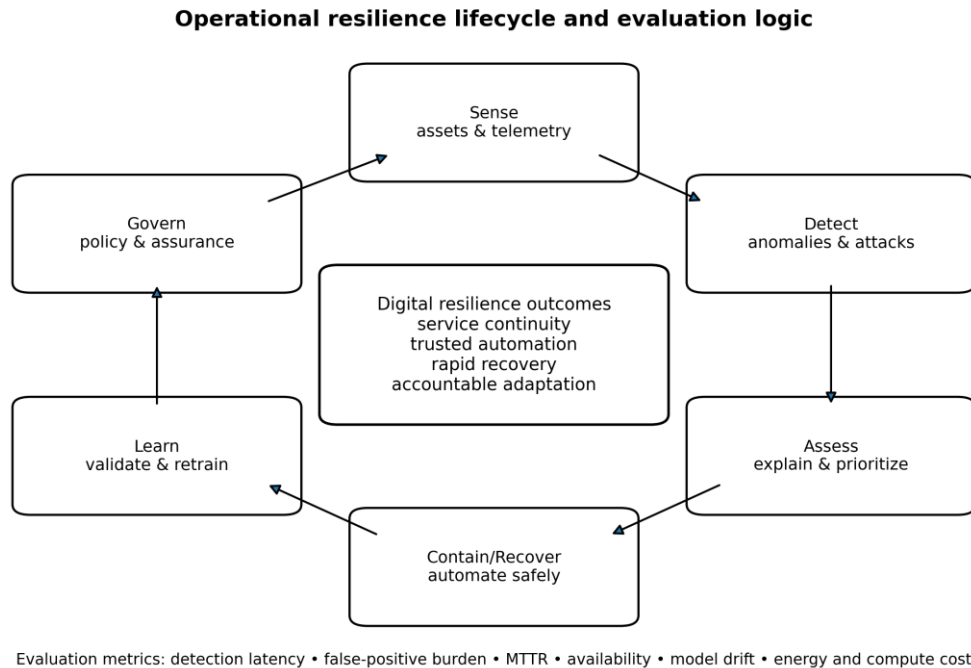


Figure 2. Operational digital-resilience lifecycle linking detection, governed response, recovery and learning.

VIII. RESEARCH GAPS AND FUTURE RESEARCH

Five research gaps should be prioritised. First, the cross-domain datasets are still limited. Most IDS studies evaluate a network or IoT benchmark, but urban incidents can propagate across identity, cloud, operational, and physical layers. Future Saudi testbeds should collect synchronised multi-domain telemetry with defensible anonymization. Second, generalisation across cities and operating conditions is under-tested. Models should be tested against seasonal traffic changes, special events, network reconfiguration, and new, previously unseen device types. Third, adversarial robustness needs to be upgraded from an optional stress test to a standard evaluation dimension, including evasion, poisoning, model extraction and compromised federated participants [8]. Fourth, research should evaluate explainability not only in terms of whether it produces feature rankings, but rather whether it improves analyst decisions. Fifth, resilience experiments must measure containment and recovery, not just detection. Federated and edge learning deserve their own Saudi validation as they can bridge the gap between distributed ownership and fast local response.

Experiments should compare centralised, federated and hierarchical designs with non-IID traffic, intermittent connectivity, malicious clients and bandwidth constraints. Rather than inferring privacy leakage from the absence of raw-data transfer, model updates should be explicitly tested for privacy leakage. Resource-aware studies should report memory, inference time, energy use and update cost, as well as security metrics [11]. Future work on digital twins should quantify the impact of model fidelity on cyber-physical anomaly detection, and identify failure modes when the twin is stale or manipulated. Finally, architectural research should be about organisational experiments. Security operations teams, municipal service owners, technology vendors and regulators need common processes for alert ownership, automated action, evidence retention and post-incident learning. Human-in-the-loop studies can help identify when explanation appropriately fosters trust and when it fosters automation bias. This is crucial because resilient AI needs both technical correctness and disciplined institutional use.

IX. PRACTICAL, MANAGERIAL AND POLICY IMPLICATIONS

For practitioners, the architecture points to deploying detection as a portfolio of capabilities. Prior to model selection, asset inventories must identify criticality, connectivity and possible inference sites. Telemetry pipelines should maintain time synchronisation and provenance for alert correlation across domains. Security teams should manage model registries that contain training data lineage, version, purpose operating conditions, validation results, and rollback procedures. High impact automated responses should be coded as policy constrained playbooks with clear human escalation paths.

Procurement should move away from requiring managers to achieve a headline accuracy score towards resilience service levels. Smart-city platform contracts can require detection latency, auditability, integration interfaces, update governance, export and recovery support for incident evidence. Workforce planning is no less important. The Saudi competency analysis by Alhakami highlights the need for continuous human capacity in technical systems as well [20]. Cross-functional exercises should involve cybersecurity analysts, network engineers, operational-technology teams, privacy specialists and service owners as incident decisions often go across organisational boundaries.

Policymakers have more durable levers than prescribing a specific AI model: interoperability and assurance. Common requirements can establish minimums for logging, identity assurance, model accountability, privacy protection, secure update processes and recovery testing across smart-city vendors. Policy should also promote managed mechanisms for data sharing that enable collaborative defence and limit excessive centralisation. This opens up a path for federated threat learning without the assumption that all operators have to disclose raw operational data. Governance should include periodic adversarial and resilience testing of AI components affecting critical services, with evidence commensurate with operational consequence.

X. LIMITATIONS

There are four limitations to this review. First, it is an integrative and not an exhaustive systematic review, thus the selected literature supports architectural reasoning but does not represent a complete census of all smart city cybersecurity publications. Second, heterogeneity of studies prevents direct quantitative comparison. Datasets, labels, attack types, class balance, compute platforms, and evaluation metrics are all over the map, so reported performance numbers shouldn't be taken to be interchangeable. Third, limited Saudi-specific peer-reviewed evidence exists on city-scale AI threat-detection deployments, so contextual studies inform governance and readiness rather than algorithmic effectiveness. Fourth, the proposed architecture is conceptual in nature. No claims are made that the architecture improves detection, recovery or cost until comparative validation is performed; it has not been implemented over an operational Saudi city. These limitations are intentionally reflected in the future-research agenda.

XI. CONCLUSION

This review concludes that AI-based cybersecurity for Saudi smart cities must be envisioned as a governed resilience architecture rather than a set of high-accuracy classifiers. Data from IoT intrusion detection, fog and edge analytics, federated learning, adversarial robustness, cyber-physical security and digital twins, show that artificial intelligence can improve detection, but each technique also leads to dependencies that must be managed. Saudi research shows that competency, adoption, governance and trust also affect the sustainability of technical capability as operational protection [17-21]. The proposed architecture is woven together by six functions: trustworthy telemetry, distributed AI detection, cross-domain correlation, explainable risk assessment, policy-constrained response, and resilience-oriented learning. Its main contribution is that it explicitly couples threat detection with continuity and recovery. In this model, success is not only defined in terms of classification metrics, but also in terms of the ability to detect quickly, act safely, sustain essential services, recover trusted state and improve without compromising privacy or accountability. This framing connects cybersecurity

investment to the larger digital-resilience goals of Vision 2030, but it also allows for flexibility for other Saudi urban domains. Empirical work should be done to test the architecture on representative multi-domain data, adversarial conditions, federated collaboration and recovery metrics before deployment claims are made.

REFERENCES

- [1] Laufs J, Borrion H, Bradford B. Security and the smart city: A systematic review. *Sustainable Cities and Society*. 2020;55:102023. [ScienceDirect](#)
- [2] Rahman MA, Asyhari AT, Leong LS, Satrya GB, Tao MH, Zolkipli MF. Scalable machine learning-based intrusion detection system for IoT-enabled smart cities. *Sustainable Cities and Society*. 2020;61:102324. [ScienceDirect](#)
- [3] Mohammadpourfard M, Khalili A, Genc I, Konstantinou C. Cyber-Resilient Smart Cities: Detection of Malicious Attacks in Smart Grids. *Sustainable Cities and Society*. 2021;75:103116. [ScienceDirect](#)
- [4] Campos EM, Fernandez Saura P, Gonzalez-Vidal A, Hernandez-Ramos JL, Bernal Bernabe J, Baldini G, Skarmeta A. Evaluating Federated Learning for intrusion detection in Internet of Things: Review and challenges. *Computer Networks*. 2022;203:108661. [ScienceDirect](#)
- [5] Ahmad R, Alsmadi I, Alhamdani W, Tawalbeh L. A comprehensive deep learning benchmark for IoT IDS. *Computers & Security*. 2022;114:102588. [ScienceDirect](#)
- [6] Roy S, Li J, Bai Y. A Two-layer Fog-Cloud Intrusion Detection Model for IoT Networks. *Internet of Things*. 2022;19:100557. [ScienceDirect](#)
- [7] Sarhan M, Lo WW, Layeghy S, Portmann M. HBFL: A hierarchical blockchain-based federated learning framework for collaborative IoT intrusion detection. *Computers & Electrical Engineering*. 2022;103:108379. [ScienceDirect](#)
- [8] Rashid MM, Kamruzzaman J, Hassan MM, Imam T, Wibowo S, Gordon S, Fortino G. Adversarial training for deep learning-based cyberattack detection in IoT-based smart city applications. *Computers & Security*. 2022;120:102783. [ScienceDirect](#)
- [9] Dina AS, Siddique AB, Manivannan D. A deep learning approach for intrusion detection in Internet of Things using focal loss function. *Internet of Things*. 2023;22:100699. [ScienceDirect](#)
- [10] Wei C, Xie G, Diao Z. A lightweight deep learning framework for botnet detecting at the IoT edge. *Computers & Security*. 2023;129:103195. [ScienceDirect](#)
- [11] Zakariyya I, Kalutarage H, Al-Kadri MO. Towards a robust, effective and resource efficient machine learning technique for IoT security monitoring. *Computers & Security*. 2023;133:103388. [ScienceDirect](#)
- [12] Lilhore UK, Dalal S, Simaiya S. A cognitive security framework for detecting intrusions in IoT and 5G utilizing deep learning. *Computers & Security*. 2024;136:103560. [ScienceDirect](#)
- [13] Houichi M, Jaidi F, Bouhoula A. Cyber Security within Smart Cities: A Comprehensive Study and a Novel Intrusion Detection-Based Approach. *Computers, Materials & Continua*. 2024;81(1):393-441. [Tech Science Press](#)
- [14] Ali J, Singh SK, Jiang W, Alenezi AM, Islam M, Daradkeh YI, Mehmood A. A deep dive into cybersecurity solutions for AI-driven IoT-enabled smart cities in advanced communication networks. *Computer Communications*. 2025;229:108000. [ScienceDirect](#)
- [15] Meng X, Zhu L. Augmenting cybersecurity in smart urban energy systems through IoT and blockchain technology within the Digital Twin framework. *Sustainable Cities and Society*. 2024;106:105336. [ScienceDirect](#)
- [16] El-Hajj M. Leveraging Digital Twins and Intrusion Detection Systems for Enhanced Security in IoT-Based Smart City Infrastructures. *Electronics*. 2024;13(19):3941. [MDPI](#)
- [17] Alhalafi N, Veeraraghavan P. Exploring the Challenges and Issues in Adopting Cybersecurity in Saudi Smart Cities: Conceptualization of the Cybersecurity-Based UTAUT Model. *Smart Cities*. 2023;6(3):1523-1544. [MDPI](#)

- [18] Alsabt R, Adenle YA, Alshuwaikh HM. Exploring the Roles, Future Impacts, and Strategic Integration of Artificial Intelligence in the Optimization of Smart City-From Systematic Literature Review to Conceptual Model. *Sustainability*. 2024;16(8):3389. [MDPI](#)
- [19] Hashim H. E-government impact on developing smart cities initiative in Saudi Arabia: Opportunities & challenges. *Alexandria Engineering Journal*. 2024;96:124-131. [ScienceDirect](#)
- [20] Alhakami W. Enhancing Cybersecurity Competency in the Kingdom of Saudi Arabia: A Fuzzy Decision-Making Approach. *Computers, Materials & Continua*. 2024;79(2):3211-3237. [Tech Science Press](#)
- [21] Aldegheishem A. Assessing the Progress of Smart Cities in Saudi Arabia. *Smart Cities*. 2023;6(4):1958-1972. [MDPI](#)
- [22] Priyadarshini I. Anomaly Detection of IoT Cyberattacks in Smart Cities Using Federated Learning and Split Learning. *Big Data and Cognitive Computing*. 2024;8(3):21. [MDPI](#)
- [23] Wolniak R, Stecula K. Artificial Intelligence in Smart Cities-Applications, Barriers, and Future Directions: A Review. *Smart Cities*. 2024;7(3):1346-1389. [MDPI](#)
- [24] Alnaser AA, Maxi M, Elmousalami H. AI-Powered Digital Twins and Internet of Things for Smart Cities and Sustainable Building Environment. *Applied Sciences*. 2024;14(24):12056. [MDPI](#)
- [25] Kalinin M, Krundyshev V, Zegzhda P. Cybersecurity Risk Assessment in Smart City Infrastructures. *Machines*. 2021;9(4):78. [MDPI](#)
- [26] Almeida F. Prospects of Cybersecurity in Smart Cities. *Future Internet*. 2023;15(9):285. [MDPI](#)
- [27] Protic D, Gaur L, Stankovic M, Rahman MA. Cybersecurity in Smart Cities: Detection of Opposing Decisions on Anomalies in the Computer Network Behavior. *Electronics*. 2022;11(22):3718. [MDPI](#)
- [28] Mohammadpour L, Ling TC, Liew CS, Aryanfar A. A Survey of CNN-Based Network Intrusion Detection. *Applied Sciences*. 2022;12(16):8162. [MDPI](#)
- [29] Zeng H, Yunis M, Khalil A, Mirza N. Towards a conceptual framework for AI-driven anomaly detection in smart city IoT networks for enhanced cybersecurity. *Journal of Innovation & Knowledge*. 2024;9(4):100601. [ScienceDirect](#)
- [30] Bibri SE, Alexandre A, Sharifi A, Krogstie J. Environmentally sustainable smart cities and their converging AI, IoT, and big data technologies and solutions: an integrated approach to an extensive literature review. *Energy Informatics*. 2023;6:9. [Springer](#)