

Artificial Intelligence for Predictive Cyber Threat Detection and Automated Incident Response in Saudi Critical Infrastructure

Abstract

In Saudi Arabia, the digitalisation of energy, water, transport, communications, healthcare and government services is accelerating, making critical infrastructure more software-defined, interconnected and data-intensive. This convergence amplifies the value of artificial intelligence for cybersecurity but also alters the consequences of detection error; a false negative may enable disruptive activity to propagate across operational technology, while a false positive may trigger an unnecessary response that impacts physical operations. This review critically assesses the use of artificial intelligence in assisting predictive cyber threat detection and automatic incident response in Saudi critical infrastructure. It synthesises recent evidence on machine learning, deep learning, federated learning, cyber threat intelligence, adversarial robustness, recommender systems and response orchestration, interpreting these technologies through the operational constraints of industrial control systems and the governance expectations surrounding national critical assets. The review shows that the predictive value is highest when heterogeneous telemetry, asset context, and threat intelligence are fused, not when models classify isolated network records. It also finds that automation is most defensible as bounded autonomy: low risk, reversible actions can be performed automatically, and safety-critical containment requires confidence thresholds, human authorisation, and explicit rollback mechanisms. Dataset shift, limited Saudi sector datasets, adversarial manipulation, insufficient cross site validation, weak explainability and scarce evidence linking model accuracy to incident outcomes are ongoing weaknesses. Hence, we propose a Saudi-centric architecture where detection, risk scoring, response orchestration, auditability and continuous learning function as an integrated control loop. Future work should explore operational trials, privacy-preserving collaboration between critical operators, safety-aware response policies and evaluation metrics that measure resilience rather than just classification performance.

Keywords: Artificial Intelligence, Critical Infrastructure, Cyber Threat Detection, Incident Response, Operational Technology, Machine Learning, Saudi Arabia, Cyber Resilience

1.Introduction

Critical infrastructure cybersecurity is different from traditional enterprise security in that cyber

events can impact physical processes, service continuity and public safety. Industrial control systems, Supervisory Control and Data Acquisition platforms, connected sensors and enterprise information technology are increasingly sharing data across architectures that were historically isolated. Therefore, surveys of critical-infrastructure intrusion detection emphasise the need to evaluate machine-learning techniques against real-time constraints, protocol heterogeneity, legacy equipment, and the operational cost of false alarms, rather than benchmark accuracy alone (Pinto et al., 2023; Umer et al., 2022). Similarly, broader work on cyber-physical infrastructure indicates that increasing connectivity increases visibility and automation but also increases the number of pathways through which an attacker can influence cyber and physical states (Sheikh et al., 2022).

This problem is particularly important in the Saudi context. Across the energy, utilities, transport, smart urban systems, cloud services and digitally enabled public services, national development programmes are driving forward the digitisation. These environments are characterised by assets of high value and interdependencies that require a paradigm shift regarding cybersecurity from static perimeter control to continuous sensing, prediction and coordinated response (Carlo & Obergfaell, 2024; Govea et al., 2024; Shan & Myeong, 2024). The structural reference attached is an example of a useful scholarly pattern for critical-infrastructure research, connecting threat analysis, artificial intelligence, and governance and operational recommendations in one framework, rather than considering model performance as a separate technical objective. Moreover, its emphasis on national and international controls highlights why AI security design should be part of the wider cybersecurity maturity and compliance processes (Ali et al., 2025).

AI has certain capabilities that are relevant to this transition. Supervised classifiers can recognise known attack classes, unsupervised and semi-supervised models can find deviations from expected process behaviour, deep models can learn nonlinear temporal relationships, natural language techniques can structure threat reports, federated learning can enable collaborative training without centralising sensitive telemetry, and recommender or orchestration systems can convert alerts into prioritised actions (Khraisat & Alazab, 2021; Ferreira et al., 2023). These capabilities enable a shift from reactive detection after a compromise to earlier detection of attack preparation, propagation, or process manipulation. Predictive cybersecurity is not simply plugging in a model into a security operations centre. Critical infrastructure produces asymmetric impacts. An enterprise endpoint model that is not properly isolated can be inconvenient.

Blocking an industrial controller or protection network can result in a safety or availability event. Ransomware, phishing and attacks on cyber-physical environments are other examples that technical detection is only one phase of an incident chain (Oz et al., 2022; Alabdan, 2020). Thus response automation needs policy constraints, asset criticality, confidence calibration, oversight by humans, and post-action verification. The central problem of how these can be combined into an AI-enabled, operationally credible and governable approach for Saudi critical infrastructure is the focus of this review.

2.Aim of the study

The study will critically synthesise recent peer-reviewed evidence on artificial intelligence for predictive cyber threat detection and automated incident response, and translate that evidence into a Saudi critical-infrastructure architecture that balances detection speed, operational safety, privacy, explainability, governance and continuous resilience improvement.

3.Objectives of Research

The review is driven by five objectives. The paper first reviews the major AI methods for detection and prediction of malicious activity in industrial, cyber-physical and IoT-rich environments. Second, it examines the role of threat intelligence, asset context, and distributed learning in improving predictive detection over isolated traffic classification. Third, it investigates the technical and governance preconditions for safely automating incident response actions . Fourth, it highlights the limitations of the existing evidence such as the dataset's realism, robustness of models, explainability and transferability to Saudi sectors. Fifth, it sets out priorities for research and deployment of an integrated detection-to-response capability appropriate for critical infrastructure operators.

4.Research Questions

RQ1: What AI approaches are best supported by evidence for predictive detection in critical-infrastructure and industrial environments? RQ2: How does heterogeneous telemetry, cyber threat intelligence and privacy-preserving collaboration improve early warning while minimising unnecessary data exposure? RQ3: What kinds of incident-response automation are operationally safe and practical for systems with physical consequences? RQ4: What are the most limiting evidence gaps for adoption in Saudi critical infrastructure? RQ5: How to integrate model governance, oversight by humans, and continuous learning within an end-to-end architecture?

5.Review Methodology

A structured integrative review design was adopted due to the research problem that spans intrusion detection, industrial cybersecurity, cyber threat intelligence, machine learning robustness and response automation. The search strategy was designed to identify empirical studies and high quality reviews that explain operational constraints. Searches were performed for Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink and ACM Digital Library using combinations of “critical infrastructure,” “industrial control system,” “SCADA,” “operational technology,” “artificial intelligence,” “machine learning,” “deep learning,” “intrusion detection,” “threat prediction,” “federated learning,” “cyber threat intelligence,” “incident response,” “SOAR,” “recommender system,” “adversarial machine learning,” and “Saudi Arabia.” The main publication window was 2020–2025, with earlier material included only when still directly relevant to concepts expressed in the recent literature. Inclusion criteria included research published in peer-reviewed journals or conferences that had a clear relation to AI-enabled cybersecurity, critical or industrial infrastructure, predictive detection, threat intelligence, or automated response. Studies were included if they described methods, datasets, evaluation procedures, architectures, risks or deployment considerations that could help answer the research questions. Exclusion criteria excluded non-peer reviewed commentary, marketing from vendors, papers without enough methodological detail, duplicated versions and work aimed at unrelated application domains.

As this is an integrative review and not a meta-analysis, no fabricated record counts, screening totals or effect sizes are reported. This method is about traceable thematic synthesis, not invented

numerical flow. The quality assessment included six questions: Was the security problem clearly defined? Was the dataset or source of evidence described? Were the evaluation metrics appropriate? Did the validation reflect realistic or cross-context conditions? Were the limitations acknowledged? Could the contribution plausibly transfer to critical-infrastructure operations? Higher interpretive weight was given to systematic reviews, realistic testbeds, multi-source datasets, industrial-control studies, privacy-preserving approaches and work discussing operational constraints rather than just reporting accuracy.

The synthesis then identified five themes: detection models and data realism; predictive context and threat intelligence; distributed and privacy-preserving learning; robustness, explainability and oversight by humans; and response orchestration. Relationships between these themes were compared to identify where evidence converges and technical claims outstrip operational validation. References were checked against publisher pages, DOI records and the structural source supplied with the request. The final bibliography includes thirty peer-reviewed references, with DOI links in the standardised <https://doi.org/> format. Methodological limitations remain: studies differ in datasets and attack taxonomies used, many benchmark results are not directly comparable, underreporting of negative deployment results, and scarcity of Saudi-specific operational datasets. These constraints are taken as findings to inform interpretation, not to overgeneralize laboratory performance. Critical analysis and thematic review of literature

6. Thematic Literature Review and Critical Analysis

AI-based detection in critical infrastructure has moved from simple supervised classification towards architectures combining network, host and physical-process information. The difference is important as industrial compromise can be seen as an anomalous packet sequence, an anomaly in process variables, a change in command timing, or a mismatch between cyber and physical states. Umer et al. (2022) differentiate between network-level and process-level intrusion detection and show that supervised, semi-supervised, unsupervised and reinforcement-learning approaches solve different parts of the problem. Pinto et al. (2023) also report a wide use of machine-learning techniques for protecting ICS, but warn that the deployment location, the

validation strategy and the dataset choice strongly influence the apparent performance.. Deep-learning surveys on SCADA intrusion detection point to CNNs, recurrent networks, autoencoders and similar models as good feature learners, while highlighting computational cost and dependency on representative data (Balla et al., 2022; Gu et al., 2022). Therefore, data realism is one key limitation. TON_IoT was designed to overcome the limitations of previous benchmark collections by providing heterogeneous telemetry, operating-system and network observations for data-driven security research [Alsaedi et al., 2020]. This direction was extended by Edge-IIoTset with a realistic IoT/IIoT testbed and data suitable for centralised and federated learning (Ferrag et al., 2022). These datasets improve evaluation but still do not recreate the full spectrum of Saudi energy facilities, desalination systems, airports, telecommunications networks or smart-city platforms. A model that performs well on a public testbed may have different protocols, maintenance cycles, operating modes, asset ages, and adversary behaviour after deployment. This gap is the reason why operational monitoring should include drift detection, and model confidence should be interpreted according to asset and process context. Information that precedes direct exploitation is also beneficial to predictive detection. Cyber threat intelligence can identify new malware, infrastructure, vulnerabilities, tactics and attack relationships. Research on recommender systems shows how security information can be filtered and prioritised for human operators, rather than presented as an undifferentiated stream of alerts (Pawlicka et al., 2021; Ferreira et al., 2023; Liang et al., 2022). More specifically, Husák and Čermák (2022) have discussed recommender systems in incident handling and response, framing recommendation as a way to reduce analyst overload and match actions to context. At a higher level of situational-awareness, Alavizadeh et al. (2022) highlight the need to combine collection, comprehension and projection, which is very close to predictive security operations: the system should not only observe an anomaly, but infer its significance and likely evolution. For Saudi critical infrastructure, hence, predictive analytics should incorporate at least four classes of evidence: real-time IT/OT telemetry, physical-process behaviour, asset criticality and external threat intelligence. A high confidence anomaly on a low impact test system should not be treated in the same way as a moderate confidence indicator affecting a safety critical controller that matches an active campaign. This architecture is illustrated in Figure 1. The model is embedded in a governance loop, not above it, because detection outputs are only operationally meaningful once risk scoring, authorisation and response constraints are applied.

Federated learning is particularly relevant when operators are not able to freely share raw telemetry. Privacy, commercial sensitivity, national-security concerns and sector regulation can all inhibit central pooling, even when multiple organisations would benefit from learning common attack patterns. Mármol Campos et al(2022) address the possible benefits and challenges of federated intrusion detection for IoT, especially in the case of non-identically distributed data. Zhang, J., et al. (2023) Distributed IIoT intrusion detection by transfer approaches . Tahir et al.. (2022) use federated learning for stealthy false-data injection detection in an Industry 4.0 energy context. Friha et al. (2023) combine decentralisation and differential privacy to further improve industrial IoT intrusion detection . Both of these studies provide support for cross-operator learning, but also uncover an important trade-off, namely, privacy-preserving collaboration does not automatically solve the problem of model poisoning, unequal data quality, communication overhead or differences among local operating environments. Another barrier to trust in forecasting models is robustness. Research in adversarial machine learning shows that a detector itself can become a target. Engineered perturbations can damage the reliability of network intrusion models, according to Alhajjar et al. (2021). He et al. (2023) synthesised attacks and defences for deep-learning-based NIDS and highlighted the gap between idealised adversarial assumptions and practical network manipulations. For critical infrastructure the problem is worse than evasion. An attacker may try to poison training data, manipulate feedback, exploit channels for model updates or induce alert fatigue. Resilience thus requires robust feature design, protected training pipelines, integrity checks, ensemble or rule-based fallback and strict separation between an AI recommendation and an irreversible control action. Explainability and human-in-the-loop is just as important. Accuracy does not tell an operator why a model thinks an event is dangerous, or whether any automated action is safe. Kure et al. (2022) embedded machine learning in critical-infrastructure risk management, showing how prediction can help to make structured risk decisions rather than replace them. Ali et al. (2025) suggest machine learning can be viewed as part of a larger framework of controls, maturity and governance. These approaches are more transferable to critical infrastructure than “black-box first” models because they link analytical output to organisational decision processes. Explanations for Saudi ops should include influential telemetry, affected assets, likely technique, confidence, predicted consequence and rationale for the recommended playbook.

The final theme is incident response automation. Automation may substantially shorten the time from detection to containment, but the literature is more developed in recommendation than in fully autonomous control. Studies of recommender systems show prioritisation and selection of actions . Recent work on cybersecurity playbooks generated by LLMs shows how structured incident data can be translated into standardised response procedures (Paduraru et al., 2025). This ability is promising when used in SOAR environments, but the language model output needs to be validated as syntactically correct output is not necessarily operationally safe. A playbook that disables an account or blocks an address may be acceptable under policy; a playbook that changes an industrial process, isolates a control network or restarts a controller demands stronger safeguards. This distinction motivates bounded autonomy. Automated response should be divided by reversibility, operational consequence and confidence Low risk actions could be enriching an alert, capturing forensic artefacts, increasing logging, blocking a confirmed malicious domain at an enterprise boundary, or quarantining a non-critical workstation. Medium-risk actions might require prompt human confirmation. Any high risk actions that may impact the availability of OT, its safety functions or continuity of production shall require specific authorisation from qualified personnel. The resulting closed loop is shown in Figure 2: sensing informs prediction, prediction is triaged by impact and confidence, response is taken under policy, verified outcomes are fed back to monitoring and retraining. This loop is better suited to critical infrastructure than one-way “detect and block” automation, because it recognises that response itself can create risk. Another problem is the time horizon implied by the word predictive. Many of the published intrusion-detection studies classify a record after the suspicious traffic has already appeared. This is useful, but it is not necessarily predictive in an operational sense. For critical infrastructure, forecasting needs to consider at least three horizons. Immediate prediction estimates the probability that an anomaly occurring right now will become a confirmed incident within seconds or minutes. Tactical prediction is the prediction of the most likely next technique, impacted asset, or lateral movement path during an active campaign. Strategic prediction uses vulnerability data, threat intelligence, maintenance status, and exposure patterns to identify situations that raise the probability of future attacks. The horizons require different data and validation. Sequence models may support immediate horizon reasoning, graph or attack-path reasoning may support tactical projection, and risk models may support strategic prioritisation.

Putting them together in one task blurs where AI actually adds early-warning value. The awareness of the physical process is especially important for the immediate and tactical horizons. Depending on the plant state, a network event can be malicious, benign or operationally irrelevant. The same command might be expected during maintenance, and very suspicious during stable production. Such models should therefore include operating mode, engineering change windows, alarm state, asset dependencies and process constraints. This can reduce false positives and improve the explanation to responders. It also closes the gap between cybersecurity and engineering teams, as the model is no longer judging packets in isolation but whether observed cyber actions are consistent with expected physical behaviour. This direction is supported by studies on industrial and cyber-physical systems. However, operational validation is limited by the difficulty in sharing real process data and the inability to freely execute attacks on production systems (Maglaras et al., 2022; Raza et al., 2022). The response side calls for similar precision. Often, automated incident response is talked about as a single capability, but it has several stages with different risk profiles. Automation can collect context, enrich indicators, de-duplicate alerts, query asset inventories, collect volatile evidence, open tickets, notify stakeholders, recommend playbooks, execute controls, verify results, and bring back to normal configurations. The initial stages are primarily for gathering information and can be automated with relatively low operational risk. Execution and restoration are different because they change the protected environment. The risk of an action is not only determined by its technical type but also by timing, target and current process condition. The consequences of blocking traffic to a business workstation versus blocking traffic to an engineering station connected to a live process may be very different, but the same firewall mechanism may be used. Response policy should be formulated as a matrix, not as a binary choice between manual and automatic operation. One dimension is reversibility of action. Evidence collection is conceptually easy to reverse, because it does not change the system, whereas changes to firmware or the shutdown of a controller may be difficult to reverse safely. Another dimension is the criticality of the assets. A third is confidence, which includes agreement across models and corroborating evidence. A fourth is the operational state, such as normal production, maintenance, emergency operation or degraded communications. These dimensions can be combined to determine if the system executes, requests single approval, requests dual approval or only recommends. Such policy can be implemented in SOAR platforms and audited post incident. It also protects against automation

bias by exposing the limits of machine authority prior to an emergency. The Saudi deployment should also take account of differences between sectors. Attacks on measurement and control integrity in the energy sector may have more serious consequences than typical malware classification (Ayele et al., 2024). In water and desalination, chemical and hydraulic constraints may have to be included in the process anomaly detection. Timing and availability can be dominant in transport as service disruption propagates rapidly across passengers and logistics. It needs high-throughput analytics and careful handling of distributed dependencies in telecom infrastructure. Smart-city environments are a mix of municipal, commercial and IoT systems, which results in multiple ownership boundaries. These differences militate against a single national model that is trained once and deployed everywhere. A more resilient pattern is a common governance architecture with sector-specific models, playbooks and safety thresholds. Local operators are in charge of the raw telemetry and operational decisions, while shared components might consist of threat-intelligence exchange, model assurance requirements, audit schemas, and incident taxonomies. Measurement needs reform, too. Accuracy, precision, recall, F1 score and area under the ROC curve are still useful for model development, but they do not directly answer the question of whether infrastructure became more resilient. A poorly prioritised detector may increase analyst workload while improving F1 score. It can detect an attack fast but recommend a response that causes unnecessary downtime. Evaluation thus needs to connect technical metrics to operational outcomes. Useful measures include median detection lead time, proportion of high impact incidents identified before service degradation, analyst minutes per validated incident, containment latency, rollback success, unnecessary isolation rate and missed critical-event rate and service-restoration time. Measurements should also include model calibration because the confidence score is only useful if it has a meaningful relationship to actual error rates. Finally, the architecture should be designed to facilitate organisational learning. After an incident, responders should be able to reconstruct what the model saw, what features drove the score, which playbook was selected, who approved it, what changed in the environment, and if the action had its intended outcome. This evidence can enhance models, rules, playbooks and training. It is a governance process linking post-incident review, data quality, threat intelligence, engineering knowledge and model change control.

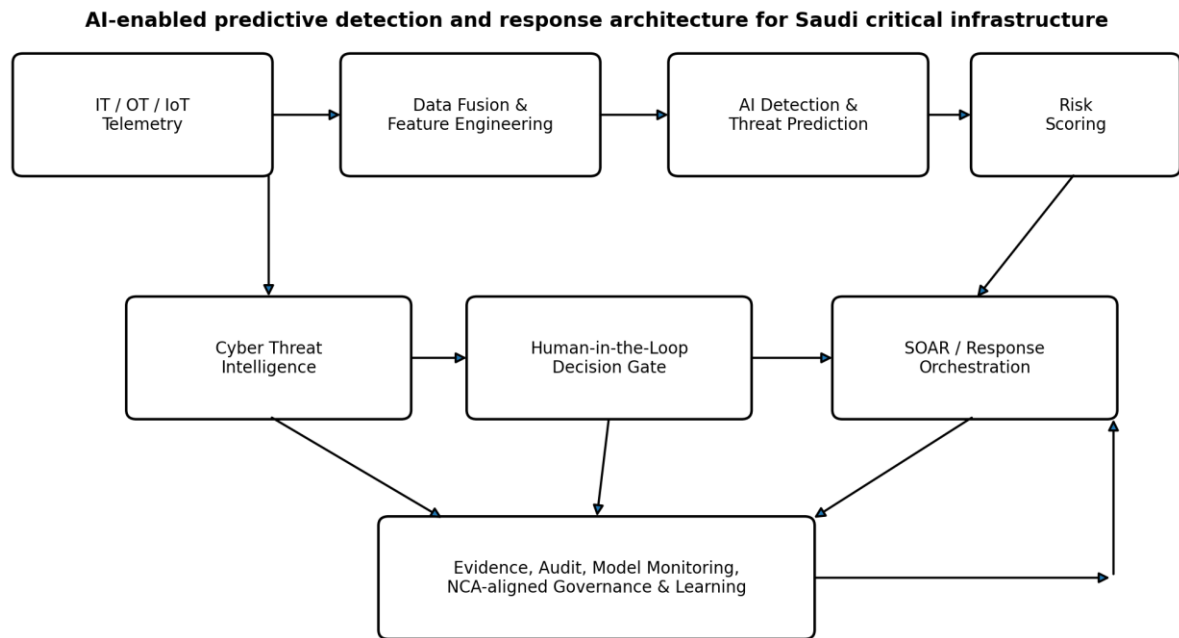


Figure 1. AI-enabled predictive detection and automated-response architecture for Saudi critical infrastructure.

316 **Table 1. Comparative role of AI approaches in critical-infrastructure cybersecurity.**

Approach	Primary value	Critical-infrastructure strength	Main limitation
Supervised ML	Known-attack classification	Efficient, interpretable with engineered features	Requires representative labels; weak on unseen behavior
Unsupervised / semi-supervised	Novelty and anomaly detection	Useful where attack labels are scarce	False positives and ambiguous operational meaning
Deep temporal models	Complex sequence and process-pattern learning	Captures nonlinear and time-dependent behavior	Data/compute demand; opaque failure modes
Federated learning	Cross-operator collaborative training	Reduces need to centralize sensitive telemetry	Non-IID data, poisoning, communications and trust
CTI + recommender systems	Context, prioritization and action support	Reduces analyst overload and links alerts to knowledge	Depends on data quality and recommendation validity
SOAR / AI playbooks	Rapid, repeatable response	Shortens containment for pre-approved actions	Unsafe if autonomy exceeds asset and safety constraints

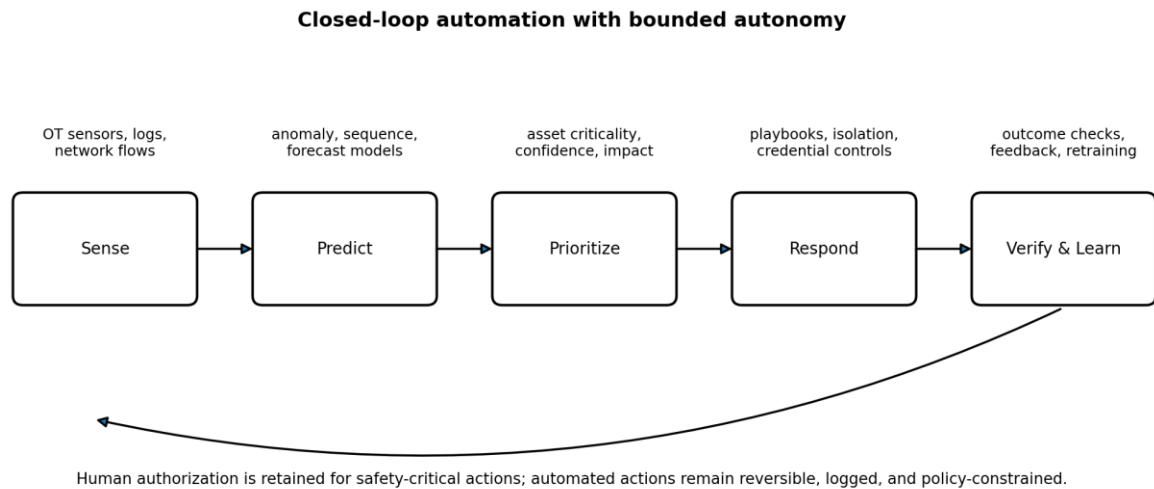


Figure 2. Closed-loop, bounded-autonomy model for AI-assisted incident response.

7. Discussion

The literature suggests a smooth transition from model-centric to context-centric cyber resilience. High-performing classifiers are still useful, but the operational value of high-performing classifiers is dependent on their embedding within telemetry engineering, asset management, threat intelligence, risk scoring and response governance. This synthesis also reflects the fact that no single family of models can be considered universally best. Supervised models do best when attack classes/labels are stable; unsupervised models can surface novelty but tend to produce more ambiguous alerts; deep models capture complex temporal patterns but require data and compute; federated learning enables collaboration but adds distributed trust problems; and recommendation or orchestration systems improve actionability but can propagate bad upstream judgements.

Therefore, the most defensible architecture for Saudi critical infrastructure is layered. The detection layer should combine signature, rule, and AI approaches so that established controls are still available when models are uncertain. The prediction layer needs to contain temporal sequences, attack path relationships and external intelligence. The decision layer should fuse model confidence with asset criticality, exposure, safety impact and operational state. The response layer should utilise pre-approved playbooks with the level of autonomy pre-determined

to an incident. Finally, the assurance layer should keep evidence, model versions, decisions, overrides and outcomes for audit and learning. Table 1 compares the main AI approaches and suggests that the research opportunity is less in choosing a single algorithm and more in matching approaches to operational roles. Table 2 operationalizes this synthesis into controls for implementation. The proposed architecture also fits the logic of Saudi cybersecurity governance: controls, accountability and continuous improvement can be kept while AI accelerates sensing and analysis. At the heart of design is the principle that automation should compress safe decision time, not remove responsibility.

Table 2. Implementation and assurance controls for a Saudi critical-infrastructure AI security program.

Control area	Recommended implementation	Assurance evidence
Data	Fuse IT, OT, process and CTI sources; classify sensitive telemetry	Data lineage, quality metrics, access logs
Models	Use ensembles/fallback rules; monitor drift and adversarial behavior	Validation records, model cards, robustness tests
Decision	Combine confidence with asset criticality and impact	Risk thresholds, explanation fields, analyst overrides
Response	Tier actions by reversibility and safety consequence	Approved playbooks, authorization matrix, rollback tests
Governance	Version models/policies and retain complete decision trails	Audit logs, change approvals, periodic exercises
Learning	Feed verified outcomes into tuning and retraining	Post-incident reviews, recalibration and retraining triggers

8. Research gaps and future research directions

Five priority research gaps are worth addressing. First, datasets specific to Saudi Arabia and to the sector are still scarce. Future work should create privacy-preserving datasets based on representative energy, water, transport, telecom and smart-city environments with attack scenarios that respect safety constraints. Second, most studies measure classification metrics, not resilience outcomes. Research should measure time-to-detect, time-to-contain, avoided service disruption, cost of false response, quality of recovery and analyst workload. Third, more evidence of cross-site generalisation is needed. Rather than random splits of a single dataset, models should be tested on unseen attack variants across facilities, vendors, seasons, operating states.

Fourth, adversarial robustness and secure learning pipelines should be considered fundamental requirements. Federated and continuously learning models require defences against poisoning, model inversion, malicious participants, and unsafe updates. Fifth, safety-aware policy research for automated response. Future prototypes should specify formally which actions are reversible, require dual authorisation, what rollback conditions apply, and how uncertainty changes response. This can be supported by digital twins and cyber ranges which allow organisations to test playbooks against realistic process consequences prior to their use in production. Another opportunity is for national learning in a collaborative way without central collection of raw data. Federated approaches could allow operators to share model updates or selected indicators, while privacy and governance controls safeguard sector confidentiality. Research should also look at explanation formats tailored to SOC analysts, engineers and executives, as they each need different evidence to trust an AI recommendation. The long-term goal should be an independently auditable AI security lifecycle where data provenance, model performance, operational outcomes and policy compliance can all be inspected.

Practical, Managerial and Policy Implication In practice, operators should begin with narrowly-scoped use cases where telemetry quality is high and response actions are reversible. Predictive models can initially augment existing SOC workflows with alert enrichment, anomaly prioritisation and threat correlation. Following the validation, automation can then move to evidence collection and low-risk containment. Industrial environments should have fallback

controls, out-of-band access, and rollback procedures, so that an AI or orchestration error cannot become a single point of operational failure. On the managerial side, accountability should reside in security, OT engineering, safety, risk and compliance functions. Data sources, validation thresholds, triggers for retraining and criteria for performance deterioration should be defined by model owners. Response playbooks should be jointly approved by security and operational stakeholders, with escalation rules based on asset criticality. Training should not only be on use of AI tools but on recognising model uncertainty, adversarial manipulation and automation bias. On the policy side, AI-enabled cybersecurity should be seen as a component of critical-infrastructure risk management, not as a separate analytics programme. Procurement requirements can call for explainability, audit logs, secure update mechanisms, data residency controls, adversarial testing and evidence of cross-environment validation. If raw telemetry cannot be shared, cross-sector information sharing should be based on standardised threat intelligence structures and privacy-preserving collaboration. Periodic exercises should test technical detection and decision authority to ensure rapid automation is compatible with safety, continuity and regulatory obligations.

9.Limitations

The evidence base is heterogeneous, which limits this review. Datasets, attack classes, feature sets, model architectures and evaluation procedures vary greatly and do not allow a meaningful aggregation of the reported accuracy values. Public datasets are only rough approximations of real critical-infrastructure environments, and publication bias may favour successful models over unsuccessful deployments. The review also has a focus on the primary literature from 2020-2025 and on peer-reviewed sources meaning that fast-moving operational practices may not yet be published in journals. Evidence specific to Saudi Arabia is relatively limited and as such, the proposed architecture is a reasoned synthesis of international technical research interpreted for Saudi operating conditions and not an assertion that all components have already been validated across Saudi sectors. Finally, automated incident response is less empirically mature than intrusion detection; recommendations regarding bounded autonomy should be tested in controlled pilots, cyber ranges and sector-specific safety assessments prior to adoption at production scale.

410

411

412

413 **10.Conclusion**

414 Artificial intelligence may significantly improve Saudi cyber defence of critical infrastructure by
415 integrating early detection, contextual prediction, and disciplined response. The best evidence is
416 not in favour of a fully autonomous cybersecurity model where an algorithm identifies and
417 performs all containment actions by itself. Instead, the literature advocates for an integrated
418 architecture that combines heterogeneous IT/OT telemetry, process behaviour, threat intelligence,
419 machine learning, privacy-preserving collaboration, contextual risk scoring, and policy-
420 constrained orchestration. Such architecture can also identify suspicious behaviour sooner,
421 prioritise the most important incidents and shorten response time without ignoring the physical
422 consequences of error. The key design criterion is bounded autonomy. AI should be used
423 aggressively to automate observation, correlation, enrichment and reversible actions where the
424 risk to the business is low, but all safety-critical actions should be governed by confidence
425 thresholds, pre-approved playbooks, and human authorisation. Ongoing verification is critical –
426 each answer must produce evidence of whether the threat was contained, whether service was
427 impacted, and whether the model or policy needs adjustment. The strategy blends technical
428 innovation with accountability and resilience which fits well with Saudi operators. Research
429 needs to transition from benchmark accuracy to cross-site trials, adversarially robust learning,
430 Saudi-representative datasets and outcome-based evaluation. Predictive cybersecurity will not be
431 operationally credible when models make most decisions, but when they help organisations make
432 faster, safer, explainable and auditable decisions under real attack conditions.

433

References

1. Ali, S. M., Razzaque, A., Yousaf, M., & Shan, R. U. (2025). An automated compliance framework for critical infrastructure security through artificial intelligence. *IEEE Access*, 13, 4436–4459. <https://doi.org/10.1109/ACCESS.2024.3524496>
2. Pinto, A., Herrera, L.-C., Donoso, Y., & Gutierrez, J. A. (2023). Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure. *Sensors*, 23(5), 2415. <https://doi.org/10.3390/s23052415>
3. Maglaras, L., Janicke, H., & Ferrag, M. A. (2022). Cybersecurity of critical infrastructures: Challenges and solutions. *Sensors*, 22(14), 5105. <https://doi.org/10.3390/s22145105>
4. Sheikh, Z. A., Singh, Y., Singh, P. K., & Ghafoor, K. Z. (2022). Intelligent and secure framework for critical infrastructure (CPS): Current trends, challenges, and future scope. *Computer Communications*, 193, 302–331. <https://doi.org/10.1016/j.comcom.2022.07.007>
5. Carlo, A., & Obergfaell, K. (2024). Cyber attacks on critical infrastructures and satellite communications. *International Journal of Critical Infrastructure Protection*, 46, 100701. <https://doi.org/10.1016/j.ijcip.2024.100701>
6. Govea, J., Gaibor-Naranjo, W., & Villegas-Ch, W. (2024). Transforming cybersecurity into critical energy infrastructure: A study on the effectiveness of artificial intelligence. *Systems*, 12(5), 165. <https://doi.org/10.3390/systems12050165>
7. Shan, A., & Myeong, S. (2024). Proactive threat hunting in critical infrastructure protection through hybrid machine learning algorithm application. *Sensors*, 24(15), 4888. <https://doi.org/10.3390/s24154888>
8. Umer, M. A., Junejo, K. N., Jilani, M. T., & Mathur, A. P. (2022). Machine learning for intrusion detection in industrial control systems: Applications, challenges, and recommendations. *International Journal of Critical Infrastructure Protection*, 38, 100516. <https://doi.org/10.1016/j.ijcip.2022.100516>
9. Gu, H., Lai, Y., Wang, Y., Liu, J., Sun, M., & Mao, B. (2022). DEIDS: A novel intrusion detection system for industrial control systems. *Neural Computing and Applications*, 34, 9793–9811. <https://doi.org/10.1007/s00521-022-06965-4>

10. Balla, A., Habaebi, M. H., Islam, M. R., & Mubarak, S. (2022). Applications of deep learning algorithms for Supervisory Control and Data Acquisition intrusion detection system. *Cleaner Engineering and Technology*, 9, 100532. <https://doi.org/10.1016/j.clet.2022.100532>
11. Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A. N., & Anwar, A. (2020). TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems. *IEEE Access*, 8, 165130–165150. <https://doi.org/10.1109/ACCESS.2020.3022862>
12. Ferrag, M. A., Friha, O., Hamouda, D., Maglaras, L., & Janicke, H. (2022). Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. *IEEE Access*, 10, 40281–40306. <https://doi.org/10.1109/ACCESS.2022.3165809>
13. Mármol Campos, E., Fernández Saura, P., González-Vidal, A., Hernández-Ramos, J. L., Bernal Bernabé, J., Baldini, G., & Skarmeta, A. (2022). Evaluating federated learning for intrusion detection in Internet of Things: Review and challenges. *Computer Networks*, 203, 108661. <https://doi.org/10.1016/j.comnet.2021.108661>
14. Zhang, J., Luo, C., Carpenter, M., & Min, G. (2023). Federated learning for distributed IIoT intrusion detection using transfer approaches. *IEEE Transactions on Industrial Informatics*, 19(7), 8159–8169. <https://doi.org/10.1109/TII.2022.3216575>
15. Tahir, B., Jolfaei, A., & Tariq, M. (2022). Experience-driven attack design and federated-learning-based intrusion detection in Industry 4.0. *IEEE Transactions on Industrial Informatics*, 18(9), 6398–6405. <https://doi.org/10.1109/TII.2021.3133384>
16. Friha, O., Ferrag, M. A., Benbouzid, M., Berghout, T., Kantarci, B., & Choo, K.-K. R. (2023). 2DF-IDS: Decentralized and differentially private federated learning-based intrusion detection system for industrial IoT. *Computers & Security*, 127, 103097. <https://doi.org/10.1016/j.cose.2023.103097>
17. Alhajjar, E., Maxwell, P., & Bastian, N. (2021). Adversarial machine learning in network intrusion detection systems. *Expert Systems with Applications*, 186, 115782. <https://doi.org/10.1016/j.eswa.2021.115782>

18. He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 25(1), 538–566. <https://doi.org/10.1109/COMST.2022.3233793>
19. Khraisat, A., & Alazab, A. (2021). A critical review of intrusion detection systems in the Internet of Things: Techniques, deployment strategy, validation strategy, attacks, public datasets and challenges. *Cybersecurity*, 4, 18. <https://doi.org/10.1186/s42400-021-00077-7>
20. Oz, H., Aris, A., Levi, A., & Uluagac, A. S. (2022). A survey on ransomware: Evolution, taxonomy, and defense solutions. *ACM Computing Surveys*, 54(11s), 1–37. <https://doi.org/10.1145/3514229>
21. Ferreira, L., Silva, D. C., & Itzazelaia, M. U. (2023). Recommender systems in cybersecurity. *Knowledge and Information Systems*, 65, 5523–5559. <https://doi.org/10.1007/s10115-023-01906-6>
22. Pawlicka, A., Pawlicki, M., Kozik, R., & Choraś, R. S. (2021). A systematic review of recommender systems and their applications in cybersecurity. *Sensors*, 21(15), 5248. <https://doi.org/10.3390/s21155248>
23. Husák, M., & Čermák, M. (2022). SoK: Applications and challenges of using recommender systems in cybersecurity incident handling and response. In *Proceedings of the 17th International Conference on Availability, Reliability and Security*, 1–10. <https://doi.org/10.1145/3538969.3538981>
24. Kure, H. I., Islam, S., & Mouratidis, H. (2022). An integrated cyber security risk management framework and risk prediction for the critical infrastructure protection. *Neural Computing and Applications*, 34(18), 15241–15271. <https://doi.org/10.1007/s00521-022-06959-2>
25. Ayele, E. D., Gonzalez, J. F., & Teeuw, W. B. (2024). Enhancing cybersecurity in distributed microgrids: A review of communication protocols and standards. *Sensors*, 24(3), 854. <https://doi.org/10.3390/s24030854>
26. Alabdan, R. (2020). Phishing attacks survey: Types, vectors, and technical approaches. *Future Internet*, 12(10), 168. <https://doi.org/10.3390/fi12100168>
27. Raza, A., Memon, S., Nizamani, M. A., & Shah, M. H. (2022). Machine learning-based security solutions for critical cyber-physical systems. In *2022 10th International Symposium*

on Digital Forensics and Security (ISDFS), 1–6.
<https://doi.org/10.1109/ISDFS55398.2022.9800811>

28. Liang, W., Xie, S., Cai, J., Xu, J., Hu, Y., Xu, Y., & Qiu, M. (2022). Deep neural network security collaborative filtering scheme for service recommendation in intelligent cyber–physical systems. *IEEE Internet of Things Journal*, 9(22), 22123–22132.
<https://doi.org/10.1109/JIOT.2021.3086845>

29. Paduraru, C., Dumitru, B., & Stefanescu, A. (2025). Automated generation of cybersecurity response playbooks via large language models. *Procedia Computer Science*, 270, 2987–2996.
<https://doi.org/10.1016/j.procs.2025.09.423>

30. Alavizadeh, H., Jang-Jaccard, J., Enoch, S. Y., Al-Sahaf, H., Welch, I., Camtepe, S. A., & Kim, D. D. (2022). A survey on cyber situation awareness systems: Framework, techniques, and insights. *ACM Computing Surveys*, 55(5), Article 107, 1–37.
<https://doi.org/10.1145/3530809>