

AI/ML-Driven Predictive Service Management for Telecom and ICT Infrastructure: A Vision 2030 Framework for Operational Resilience

Abstract:

The telecom and information and communication technology (ICT) infrastructure sector is shifting from a fault-management approach that focuses on components to one that is service-oriented and involves predicting potential degradation before it leads to an outage or a breach of service levels. This report brings together research carried out between 2020 and 2025 on the use of machine learning for traffic forecasting, anomaly and fault prediction, network slicing, digital twins, federated learning, explainability, and zero-touch orchestration. The objective is to examine how these capabilities can be combined into a predictive service-management architecture that meets the operational-resilience goals linked to Saudi Arabia's Vision 2030 digital transformation initiative. It has been found that the greatest operational benefit does not come from the individual accuracy of models, but from connecting the forecasts to controlled, closed-loop actions across the radio, transport, cloud, edge, and service layers. Deep temporal models enhance the ability to anticipate demand; anomaly detection systems and fault-prediction systems provide earlier opportunities for intervention; reinforcement learning aids in the adaptive allocation of resources; and digital twins allow for safer assessment of potential actions. Nevertheless, implementation is still limited by problems such as data drift, heterogeneous telemetry, the scarcity of failure labels, cross-domain causality, inference latency, energy consumption, explainability, and questions of organizational accountability. A five-stage framework is suggested: observe, predict, diagnose, decide, and assure. This framework incorporates a shared telemetry infrastructure, multi-horizon prediction, causal service-impact analysis, risk-aware orchestration, human-governed automation, and continuous measurement of resilience. For telecom and ICT environments in Saudi Arabia, the framework places particular emphasis on sovereign data handling, interoperable automation, energy-aware operation, and the assurance of measurable continuity for key digital services. The review ultimately concludes that predictive service management should be assessed in terms of the amount of service impact

avoided, the speed of recovery, the reduction in SLA risk, and the reliability of automation, rather than solely on predictive accuracy.

1. Introduction

Telecommunication networks have now been transformed into software-defined, virtualised, cloud-distributed systems in which the way in which services behave is determined by the interactions between the radio access network, the transport network, the core network, edge computing, the public and private clouds, the application platforms, and the operational support systems. Although this architectural separation enhances flexibility it also increases the number of possible points of failure and reduces the value of relying solely on device-specific assurance. In conventional network operations action is taken only after thresholds have been breached, alarms have been generated, or customers complain about a deterioration in service. These kinds of procedures are becoming ever less suitable for services which have to continuously meet requirements relating to latency, availability, and capacity. It is for this reason that zero-touch management is being developed, with prediction and closed-loop control being placed at the heart of future network operations [1,2].

The research in the field of AI and machine learning offers a number of components that are enabling this change. Traffic prediction is able to foresee congestion and capacity shortfalls [12-15]; network-fault prediction can detect risks to equipment or links before they fail [10,11]; anomaly detection can pick up on emerging behaviour that static thresholds are unable to identify [8,9,25]; and machine-learning-assisted slicing can assign shared resources in accordance with varying demand and service priorities [3-7]. Digital twins provide an additional level by allowing operators to simulate different network conditions and potential interventions before implementing them on the actual infrastructure [16-19]. Federated learning and edge-native intelligence deal with cases where it is difficult to centralise operational data due to constraints relating to privacy, bandwidth, latency, or ownership [20-24]. All of these approaches contribute to a move from simply monitoring network health to predicting the impact on services.

59

60 The issue is that research is still divided according to technical layer and algorithm. Even if a
61 very accurate load forecast is produced, this does not stop an SLA breach from occurring; a fault
62 classifier is operationally inadequate when it provides no lead time for taking action; and an
63 autonomous policy can increase risk if the decision it makes cannot be explained, bounded, or
64 reversed. Predictive service management thus has to involve the coordination of various kinds of
65 intelligence and must link together telemetry quality, prediction horizons, service topology,
66 causal diagnosis, resource decisions, action safety, and feedback. Investigations into network
67 automation and fault prediction also stress that intelligence should be incorporated into
68 management processes rather than be regarded as a separate analytics function [2,10].

69

70 The telecom infrastructure supports cloud adoption, smart services, the digitalisation of industry,
71 connected transport, public platforms, and data-intensive economic activity. In this context,
72 operational resilience goes beyond simple network availability; it means having the ability to
73 keep essential digital services running, to cope with abnormal situations, to restore performance
74 rapidly, and to make adaptations without incurring uncontrolled operational costs. Research into
75 digital transformation within the Saudi telecommunications sector also shows a connection
76 between technology modernisation and gains in both operational and marketing efficiency [30].
77 The framework associated with Vision 2030 should therefore convert predictive analytics into
78 tangible results regarding service continuity, customer experience, efficiency, and governance,
79 without assuming that automation is always advantageous.

80

81 Objectives. The purpose of this review is to create an evidence-based framework for predictive
82 service management driven by artificial intelligence and machine learning which enhances
83 operational resilience in the telecom and ICT infrastructure in Saudi Arabia. The synthesis is
84 guided by five objectives: to determine which predictive capabilities are most relevant to service
85 assurance; to look at the way in which prediction can be combined with orchestration and digital-
86 twin validation; to assess the technical and governance constraints that prevent its
87 implementation in production environments; to establish resilience measures that link model

88 outputs to operational value; and to suggest a step-by-step Vision 2030 implementation route that
89 maintains human accountability as it increases automation. The outcome is not simply another
90 list of algorithms. It is a service-management architecture that makes clear where prediction
91 generates operational value, where automation should be restricted, and how insights from
92 research into 5G, beyond-5G, and the emerging 6G can be used to inform current ICT
93 operations.

95 2. Review Methodology

96
97 A structured integrative review was chosen since the body of evidence includes surveys,
98 experimental network studies, operational datasets, architecture papers, and optimization
99 research, the results of which cannot be compared on a statistical basis. The time period covered
100 by the review was from January 2020 to December 2025. The searches were based on four
101 interconnected concepts: artificial intelligence and machine learning; telecom or ICT
102 infrastructure; prediction or proactive assurance; and service or network management. The
103 relevant terms used were "machine learning", "deep learning", "reinforcement learning",
104 "federated learning", "telecommunication", "5G", "6G", "ICT infrastructure", "traffic prediction",
105 "fault prediction", "anomaly detection", "predictive maintenance", "network slicing", "zero-
106 touch", "digital twin", "service assurance", and "orchestration". Scopus, Web of Science Core
107 Collection, and IEEE Xplore were used as the main sources for discovering the literature, and the
108 DOIs and bibliographic information were verified using the publishers' records and the official
109 indexing websites.

110
111 To be included, the publication had to have been carried out between 2020 and 2025, have
112 appeared in a peer-reviewed journal or in proceedings that had been subjected to rigorous review,
113 have a clear connection with communication networks or with other ICT operations, and have
114 provided substantial evidence regarding prediction, optimization, anomaly detection,
115 maintenance, orchestration, explainability, distributed learning, or digital twins. Studies were
116 excluded if AI was only indirectly involved, if the operational context could not be applied to

telecom service management, if the bibliographic details or DOI could not be verified, or if the work presented evidence that was already covered by a more comprehensive source. Thirty references were kept for the final synthesis. Review articles that were published outside the evidence window were used only for organising the methodology and were excluded from the evidential synthesis as well as from all final in-text citations.

The studies that were kept were coded according to six criteria: the operational problem, the infrastructure layer, the AI/ML method, the input data, the decision enabled, and the deployment limitation. In a further coding stage the intended operational effect was categorised as being anticipatory, diagnostic, optimising, assurance-related, or relating to governance. Quality was assessed by reference to the realism of the data or scenario, the clarity of the validation, the relevance of the performance metrics, the presence of a comparison with baseline figures, and the discussion of the operational constraints. Since the various studies make use of different datasets, time horizons, network scales, definitions of outcomes, and experimental conditions, no pooled effect estimate was carried out and instead a thematic synthesis was applied in order to spot recurring mechanisms and areas of disagreement. This method prevents the accuracy values from incompatible tasks from being treated as though they were examples of the same effect size.

The process took three stages. In the first stage, the studies were organised according to the following areas: traffic and capacity forecasting, anomaly and fault prediction, resource orchestration, digital twins, and trustworthy distributed intelligence. In the second stage, the cross-cutting operational dependencies were determined, these being telemetry quality, prediction lead time, topology awareness, model drift, policy constraints, latency, and human oversight. In the third stage, these dependencies were used to develop a closed-loop service-management framework as well as a staged implementation roadmap. Since the review is therefore not meta-analytic but interpretive, its aim is to show how individual technical advances can be brought together to form an operational system and to point out that in some cases the evidence is still not mature enough for autonomous deployment.

146

147 3. From Predictive Analytics To Predictive Service Management

148

149 Predictive service management starts by making a distinction between forecasting infrastructure
150 behaviour and forecasting service risk. Infrastructure prediction involves estimating various
151 factors such as traffic volume, radio-link failure, equipment health, resource demand, or unusual
152 KPI patterns. Service prediction poses a more important question: given those signals and taking
153 into account the dependency structure of a service, what is the probability, timing, and likely
154 severity of customer-visible degradation? It is important to make this distinction since many
155 network events are abnormally situated locally but are operationally harmless, while minor
156 degradations in a key shared dependency can impact thousands of sessions. Service management
157 thus needs a mapping from component signals to service topology, SLA exposure, customer
158 segments, and business criticality.

159

160 Deep neural networks have been used on a number of occasions to handle non-linear, seasonal,
161 and multi-scale network demand [12-16]. As Jiang's comparative study indicates, model
162 selection must be based on empirical evidence rather than on current trends since different
163 architectures perform variably depending on the temporal structure and the dataset [13]. The
164 example of long-horizon 5G forecasting also proves that non-stationarity becomes all the more
165 significant as the forecasting horizon increases, since network expansion decisions have to be
166 made on the basis of stable estimates that are developed months in advance rather than minutes
167 ahead [15]. For operational purposes, this means that a multi-horizon approach is required.
168 Forecasts at the minute level are useful for admission control and for enabling elastic scaling;
169 those at the hourly level help with staffing and the scheduling of maintenance; and day-to-month
170 forecasts are relevant to capacity procurement, energy planning, and investment. Therefore, one
171 forecasting model cannot be used to optimally serve all the different horizons.

172

The prediction of faults and anomalies deals with a different kind of uncertainty: rather than focusing on how much demand will arrive, it is about whether the infrastructure will deviate from its expected behaviour. Actual radio-link data indicate that by combining information on the weather with historical link parameters it is possible to predict the risk of failure, thus providing an intervention window which static monitoring does not have [11]. A wider survey shows that network-fault prediction is valuable whenever it leads to lower maintenance costs, less downtime, or shorter restoration times, not just when the accuracy of classification increases [10]. Studies of anomalies in the O-RAN environment show that supervised models are able to learn multivariate KPI patterns [9], while ensemble methods combine load prediction with anomaly detection in 5G data [8]. Nevertheless, recent reviews specific to the telecommunications sector highlight ongoing problems such as imbalanced classes, constantly evolving attack and failure patterns, and inconsistent benchmark datasets [25]. In practical use, these weaknesses can result in an overflow of alerts or in false confidence.

Resource orchestration turns predictions into actions. Network slicing is a suitable example since different service types have to compete for a limited amount of radio, computing, and transport resources. To allocate or reconfigure these resources in response to changing demand, research has employed clustering, supervised learning, deep reinforcement learning, and multi-agent approaches [3-7,26-29]. ML-based forecasting can cause a slice to be reconfigured before its capacity is used up [7], and multi-agent reinforcement learning is able to coordinate resources end-to-end across a number of distributed decision points [28]. However, performance studies also indicate that learned approximations may breach SLAs when the load is high or when the channel conditions are changing, unless explicit enforcement mechanisms are put in place to constrain them [29]. It is necessary for prediction to reduce the range of possible decisions, not to eliminate operational controls.

A predictive service-management loop therefore consists of five functional stages. The Observe stage builds a synchronized telemetry fabric from network KPIs, logs, traces, alarms, configuration, topology, customer experience, cloud metrics, and external context. The Predict stage estimates future demand, anomaly probability, fault risk, and SLA exposure at various

horizons. The Diagnose stage links the predictions to the topology and service dependencies so that an operator can tell the difference between correlation and a plausible cause. The Decide stage ranks the interventions in terms of expected impact, cost, confidence, energy, and policy constraints. The Assure stage checks the results after an action has been taken and sends the outcomes back into the models. Figure 1 shows this architecture.

Table 1. Evidence synthesis for predictive service-management capabilities.

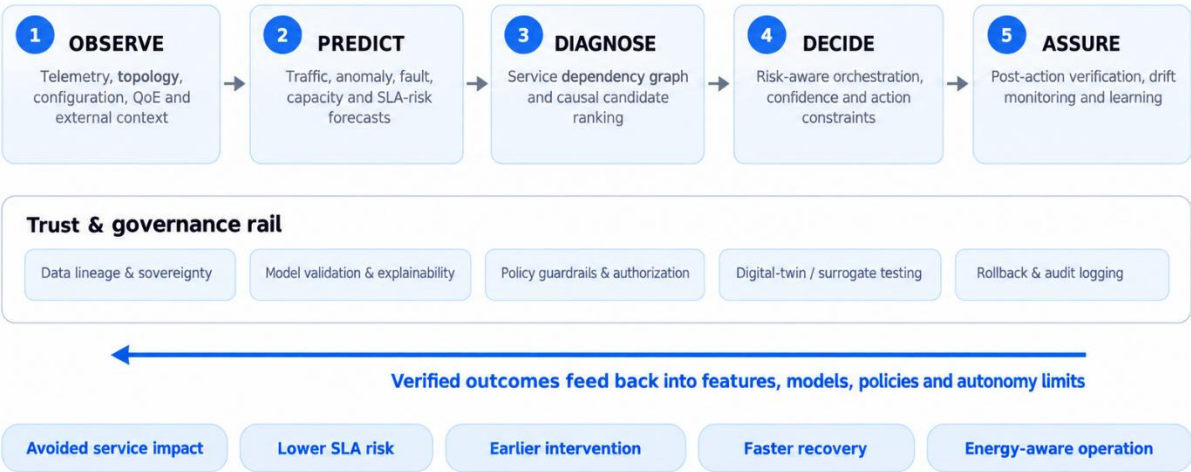
Predictive domain	Representative evidence / methods	Operational decision enabled	Resilience value	Main deployment constraint
Traffic & capacity	Deep temporal forecasting, non-stationary models, twin-oriented prediction [12-16]	Elastic scaling, admission control, capacity and energy planning	Earlier congestion avoidance; lower over-provisioning	Distribution shift, asymmetric peak error, horizon-specific accuracy
Fault & anomaly	Link-failure models, multivariate KPI classification, ensemble detection [8-11,25]	Maintenance prioritization, incident pre-emption, alarm consolidation	Longer intervention lead time; reduced downtime and restoration effort	Rare-event labels, false positives, weak cross-domain causality
Slicing & orchestration	Supervised approximation, deep RL, multi-agent RL [3-7,26-29]	Resource reallocation, slice reconfiguration, traffic steering	Adaptive service isolation and resource efficiency	SLA violations under drift; unsafe unconstrained policies
Digital twins	Traffic-aware twins, simulation integration, predictive	Pre-deployment action testing, scenario analysis,	Reduced intervention risk; safer	Twin synchronization, fidelity,

Predictive domain	Representative evidence / methods	Operational decision enabled	Resilience value	Main deployment constraint
	maintenance [16-19]	maintenance scheduling	optimization	computational latency
Distributed & trustworthy AI	Federated learning, edge intelligence, explainability [20-24]	Localized training/inference, governed decision support	Data-sovereign intelligence; lower latency; stronger accountability	Non-IID data, poisoning risk, communication overhead, explanation quality

210

Predictive service-management closed loop

From heterogeneous telemetry to verified, policy-governed service resilience



211

212

213

214

215

Figure 1. Closed-loop architecture for AI/ML-driven predictive service management. The framework connects service-aware observation and prediction to diagnosis, risk-aware decisions, verified actions, and governance controls.

216

217

218 **4. Evidence Synthesis: Enablers Of Operational Resilience**

219

220 **4.1 Traffic forecasting and proactive capacity management**

221

222 The value of being able to predict traffic consists in delaying capacity decisions until earlier in
223 time. Instead of adding resources after usage has already gone up, predictive scaling sets aside or
224 turns on resources before the service reaches a state of congestion. Thanks to deep learning, the
225 representation of temporal dependence, multi-scale seasonality, and non-stationarity in both
226 Internet and 5G traffic has been improved [12-15]. Traffic prediction aimed at digital twins goes
227 a step further by offering a virtual environment in which policies driven by forecasts can be
228 tested before they are put into use [16]. The benefit of this approach is not just a reduction in
229 forecast error, but a lower probability of service degradation caused by capacity constraints
230 within a given cost range.

231

232 Forecast accuracy can however mislead in an operational sense when the error is averaged over
233 different periods or cells. The damage caused by under-prediction at a short critical peak may in
234 fact be greater than that resulting from larger over-prediction during calm periods. It is for this
235 reason that service management should make use of asymmetric loss functions which take into
236 account SLA exposure, scarcity, and the impact on customers. Prediction intervals are more
237 useful than point estimates since orchestrators need to know about the uncertainty involved.
238 Even if the mean of a highly variable forecast is below a threshold, it may still warrant the
239 allocation of reserve capacity or human review. In this way, forecasting is seen as a form of risk
240 estimation rather than demand estimation.

241

242 **4.2 Failure anticipation, anomaly detection, and causal diagnosis**

243

244 The aim of predictive maintenance is to take action before a failure occurs, but this is made
245 difficult in the case of telecom networks since faults are uncommon, diverse, and spread through
246 dependencies. A problem with the radio section might appear as increased application latency; a
247 bottleneck in cloud resources could be mistaken for transport congestion; and changes to the
248 configuration can lead to symptoms appearing in a number of different areas. Research into the
249 prediction of link failures and telecom faults shows the value of using time-series and contextual
250 features to enable earlier warnings [10,11]. Anomaly detection in 5G and O-RAN environments
251 also proves the usefulness of multivariate KPIs [8,9,25]. Nevertheless, anomaly scores by
252 themselves do not reveal the cause of a service failure.

253

254 What is missing is service-aware causal diagnosis. The models must integrate temporal evidence
255 with the network topology, the configuration history, the dependency graphs, and the change
256 events. Graph-based reasoning is appealing since telecom services can be naturally seen as a set
257 of relationships among users, cells, functions, paths, virtual resources, and applications. The
258 diagnosis engine should then evaluate the possible causes both in terms of statistical evidence
259 and topological plausibility. The aim in operations should be to minimise the mean time to
260 prevent or the mean time to isolate, not to achieve high model accuracy. This in turn also helps to
261 reduce the number of alerts: several anomalous signals can be combined into a single incident
262 narrative which is associated with the affected services.

263

264 4.3 Predictive orchestration and bounded autonomy

265

266 The literature concerning slicing shows how AI can help with making dynamic decisions about
267 resources, decisions that would be hard to compute by hand at network timescales [3-7,26-29].
268 Supervised models are able to approximate optimal decisions, reinforcement learning can
269 develop policies based on sequential feedback, and multi-agent methods can spread out the
270 decisions across different domains. While these techniques can lead to better use of resources,

their practical acceptability relies on having a limit on autonomy. A model should not be given the right to carry out the same kind of action just because it has a high level of confidence. Authorization has to take into account the importance of the service, the potential impact of a failure, whether the action can be reversed, and the level of maturity of the control loop.

There are four levels in a practical autonomy ladder. At the advisory level, the model assesses the risk and suggests an action. In the case of supervised execution, the human has to approve the recommendation. At the bounded automatic execution stage, actions that have already been pre-approved and are of low risk take place within set limits, for example by adding elastic capacity or by rerouting non-critical traffic. At the level of autonomous resilience, coordinated actions between different domains happen automatically, but only after policy verification, a rollback procedure, and post-action assurance has been carried out. This kind of ladder provides organisations with a way to migrate step by step while keeping accountability, and also avoids a typical mistake in automation programmes—that is, confusing the elimination of human involvement with a high degree of operational maturity.

4.4 Digital twins as a safety layer

Network digital twins produce virtual versions of operational environments and may thus act as a safety measure between prediction and action. Recent studies have combined deep traffic prediction, network simulation, and AI-enabled digital twins in order to aid planning, online optimisation, and maintenance [16-19]. In the case of predictive service management, the twin does not need to replicate all the physical details. What is important is decision fidelity – that is, whether it can accurately estimate the service effects of possible interventions for the particular decision in question.

This indicates the existence of a hierarchy among twins. The planning twin is able to assess long-term capacity scenarios by using data that occurs less frequently. The operational twin can

practice near-real-time actions in relation to the current topology and workload. The fault twin can simulate the process of component degradation and recovery. The service twin can estimate the effect on customers and on SLAs across the entire chain of dependencies. The closer the decision is to being immediate, the more closely the twin has to match the live state. It is also necessary that simulation errors be visible, since otherwise the digital twin might develop a false sense of confidence. Research into the integration of network simulation with digital twins stresses the importance of updating the model's credibility using field data and of making use of surrogate models when full simulation is too slow [17].

4.5 Distributed, explainable, and trustworthy intelligence

It is not always efficient or always appropriate to centralize all network data. Federated learning allows model training to take place across a number of distributed nodes without having to pool the raw data, thus reducing privacy risks and complying with data-sovereignty requirements [20-22,24]. Likewise, edge-native intelligence helps to decrease inference latency and reduce dependence on backhaul links [22]. These features are pertinent to multi-domain telecom environments in which operators, cloud platforms, enterprise networks, and equipment domains each retain ownership of their own data.

The introduction of distribution leads to new governance issues since data that is not uniformly distributed can bias the local models, the communication overhead can decrease efficiency, and if some participants are compromised then they can corrupt the shared learning [20-24]. Explainability is thus not merely an optional feature of the user interface. When it comes to important network decisions, operators need to know which signals led to a particular prediction, what uncertainty is still present, and whether the recommendation is in line with policy [23]. The level of explainability should correspond to the nature of the decision. Engineers will require feature attribution and counterfactual behaviour; service managers will need to know which services are affected, the predicted impact, and the level of confidence; auditors will need model

lineage, approval records, and a record of actions. Trust is established through this combination of technical evidence and an accountable process.

5. A Vision 2030 Framework For Operational Resilience

The framework suggested works towards achieving service-resilience as its main aim, rather than focusing on a model portfolio. In its first stage there is a sovereign telemetry fabric which standardises data from radio, transport, core, cloud, edge, security, application, customer experience, and configuration areas. Data contracts specify issues of ownership, quality, retention, timing, and permitted use. Streaming capabilities enable real-time inference, while curated historical databases are used for training and auditing. A topology and a service-dependency graph provide the common context required to turn network observations into service impact.

The second layer consists of a multi-horizon intelligence plane, which has forecasting models that estimate traffic, calculate demand, determine energy load, and assess capacity saturation; anomaly models that pick up any deviations from the normal behaviour that has been learned; fault models that estimate the probability and time to failure; and service-risk models that combine these predictions with information on topology, redundancy, criticality, and SLA thresholds; risk abstraction is then used to reduce cognitive load and to make the model outputs usable for orchestration systems.

The third layer involves a decision-making process that is guided by policy. When assessing possible actions, account is taken of their technical feasibility, the expected benefit to services, cost, energy usage, security restrictions, and whether the action is reversible. In the case of decisions of high impact, a digital twin or a fast surrogate model is used to test the proposed action before it is carried out. The actions are then allocated to an autonomy tier, with low-risk and reversible interventions being automated; those whose cross-domain consequences are

uncertain needing approval; and actions which affect critical services requiring more substantial evidence and the readiness to roll back. In this approach, human oversight is seen as a control mechanism rather than as a temporary shortcoming to be removed.

The fifth stage is that of assurance and learning. Following each intervention, the system checks the actual results against the expected ones. For example, if it had been predicted that a reroute would reduce latency, the platform checks whether latency has indeed decreased without causing congestion in other areas. When a predicted fault does not take place, the event is used as evidence to help with recalibration rather than being treated as a silent false positive. Monitoring of model performance continues to be carried out for changes over time with respect to geography, vendor, service type, and traffic conditions. The operational results are then fed back into the training process and used in policy reviews. Thus, a resilience learning cycle is established whereby the automation becomes safer as more evidence is gathered.

The framework is in line with Vision 2030 since it focuses on four practical results. Firstly, service continuity enables digital public services, enterprise platforms, connected infrastructure, and cloud-dependent economic activity. Secondly, proactive resource management boosts both capital and operating efficiency by matching capacity to predicted demand rather than relying on static over-provisioning. Thirdly, sovereign and distributed learning methods allow for the acquisition of intelligence without there being a need to move large amounts of raw data. Fourthly, energy-aware optimisation can ensure that resilience is compatible with sustainability rather than involving the addition of a computationally expensive automation layer. Research into the transformation of the Saudi telecommunications sector shows that digital strategies can lead to improved operational efficiency, thus providing a local reason for linking the deployment of AI to tangible management outcomes [30].

Figure 2 presents the framework in the form of a step-by-step roadmap. The sequence starts with observability and data quality, followed by validated prediction, then by cross-domain orchestration, bounded autonomy, and finally adaptive resilience. Each stage should be advanced

only on the basis of evidence and not according to calendar dates. An operator should not move from prediction to autonomous execution simply because the models have achieved a specified level of accuracy; instead, the advance should occur when the predictions remain stable even in the face of drift, the actions can be audited and reversed, the service-impact models are reliable, and the post-action assurance shows a repeatable benefit. Table 2 gives a summary of the capability priorities and outcome measures for this maturity pathway.

Table 2. Indicative Vision 2030 capability pathway and operational gates.

Stage	Primary capability	Automation boundary	Resilience measures	Governance gate
Foundation - 2026	Unified telemetry, service graph, data contracts, lineage	Human-led operations with analytics support	Coverage, data freshness, topology accuracy, baseline MTTD/MTTR	Ownership, access control, data quality and retention rules
Predictive assurance - 2027	Multi-horizon traffic, fault, anomaly and SLA-risk models	Advisory recommendations; selected supervised actions	Lead time, calibration, false-negative cost, avoided SLA exposure	Independent validation, drift thresholds, explanation requirements
Cross-domain diagnosis -	Causal/service-impact reasoning plus digital-twin tests	Human approval for cross-domain interventions	Isolation time, diagnosis precision, simulated action	Twin fidelity, change evidence, rollback design

Stage	Primary capability	Automation boundary	Resilience measures	Governance gate
2028			risk	
Bounded orchestration - 2029	Policy-aware closed loops across network and cloud domains	Automatic low-risk, reversible actions within explicit limits	Prevented incidents, recovery acceleration, resource and energy efficiency	Authorization tiers, blast-radius limits, immutable action logs
Adaptive resilience - 2030	Continuous assurance, distributed learning, policy adaptation	Measured autonomy with escalation for uncertainty or criticality	Continuity, SLA risk, cost, energy, trust, safe-action success rate	Periodic policy review, model retirement, auditability and human accountability

Vision 2030 operational-resilience maturity pathway

Indicative capability sequence: advancement is gated by evidence, not by date alone



Figure 2. Indicative Vision 2030 maturity pathway for operational resilience. Progress is evidence-gated: greater autonomy follows validated data quality, calibrated prediction, cross-domain diagnosis, controlled orchestration, and measurable assurance.

6. Governance, Security, And Deployment Constraints

The reason why predictive service management is difficult is not because algorithms are not available but because of the problem of operational integration. Telecom data are spread out among systems that are specific to each vendor, with different time resolutions, naming conventions, and ownership divisions. If the data are not properly synchronized then the models may deduce false relationships, and if there are changes to the configuration the baselines that have been learned can be rendered invalid. Data quality therefore demands the implementation of engineering controls such as timestamp alignment, lineage, handling of missing data, schema governance, and the explicit management of topology changes. Without these basic foundations more advanced models will increase rather than reduce operational uncertainty.

Model drift is just as important; the behaviour of the network changes with software updates, changes in the spectrum, the introduction of new devices, seasonal occurrences, commercial campaigns, and as customer applications evolve. A model that has been validated in one region or for one vendor might fail in another. Continuous monitoring therefore needs to include calibration, stability, the cost of false negatives, lead time, and service-impact error as well as the usual measures of accuracy. Research into RAN slicing by Green-ML also indicates that computational efficiency should be regarded as an objective, especially in cases where optimisation takes place continuously at the edge [27]. A model that is superior in an operational context may be only slightly less accurate but can be faster, more stable, cheaper, and easier to explain.

As the ability of automation increases, so does the need for security. It is possible for training data to be corrupted, for telemetry to be altered, for the interfaces of the model to become points of attack, and for automated actions to amplify malicious inputs. Although federated architectures reduce the need to move large amounts of raw data, they do so at the cost of introducing risks relating to the trust of the participants and to the aggregation process [20-24]. Because of this, every high-impact control loop must include identity verification, authorization, checks on model integrity, protected feature pipelines, policy enforcement, and immutable logging of decisions. The basic idea is that the level of authority given to a model must never go beyond the strength of the controls which verify its inputs and limit its outputs.

7. Discussion And Research Agenda

What the synthesis shows is a change in the unit of optimization: in the early days of AI applications attention was focused on a single KPI or network function; whereas predictive service management has to optimize the probability and severity of service disruption over a range of interacting domains. This means that the way in which research is designed must also change. A paper which reduces the RMSE in traffic forecasting may be useful, but it does not show that it has resilience unless the forecast leads to a decision and that decision results in a

measurable protection of service. Evaluation in the future should therefore link prediction to interventions, interventions to service outcomes, and service outcomes to cost, energy, and risk.

There are four research gaps which are particularly significant. The first of these is concerned with cross-domain causality. Present-day datasets tend to focus on either radio, core, or traffic variables, whereas actual incidents involve multiple layers. In order to address this, research should aim at creating multimodal datasets that correlate telemetry, topology, configuration changes, tickets, and customer experience. These kinds of datasets would enable the use of causal and graph-based methods to tell the difference between symptoms and the conditions that cause them. Since no single organization has a complete view of the end-to-end process, privacy-preserving data collaborations might be required.

The second issue is uncertainty-aware orchestration. Although learned policies typically aim to maximise expected reward, service resilience must take tail risk into account. An action which has a low probability of causing a widespread outage should be handled in a different way from one whose worst possible outcome is only minor inefficiency. Future research should integrate prediction intervals, calibrated probabilities, constrained reinforcement learning, risk-sensitive optimisation, and digital-twin stress testing. The output of the decision-making process should comprise uncertainty together with a safe fallback, not just an action label.

The third area concerns operational explainability. Although current research into explainable AI offers some useful methods, telecom control needs explanations that are linked to the workflow [23]. An explanation must state why the risk was raised, which dependencies are involved, what evidence could disprove the diagnosis, why one course of intervention was chosen, and what the rollback condition is. This kind of explanation can help build engineer trust as well as assist in regulatory audits. It can also show whether the model is relying on unstable proxies rather than on causal indicators.

The fifth area concerns resilience economics. While studies usually give figures on accuracy, latency, throughput, or resource utilisation, it is the operators who make decisions about investments on the basis of avoided downtime, reduced truck rolls, capacity deferral, energy savings, customer impact, and the cost of restoration. Murphy et al. clearly advocate for fault-prediction measures that are related to maintenance value [10]. This line of reasoning should also be applied to service management. Model selection should then be made on the basis of the expected operational value under uncertainty, which would allow a simpler model to beat a more accurate one when it results in earlier, safer, or cheaper interventions.

For Saudi operators and ICT providers, the existence of these gaps indicates that a portfolio strategy should be adopted. In the short term, programs should focus on high-quality telemetry, service dependency mapping, traffic forecasting, and advisory fault prediction since these capabilities generate value without needing a high degree of autonomy. The following step should involve the integration of digital-twin testing and supervised orchestration when making capacity and maintenance decisions. Cross-domain bounded automation should only be introduced where failure modes, rollback procedures, and accountability have all become mature. This sequence enables rapid modernisation without involving a technology-first approach in which the sophistication of the models exceeds the operational controls.

The framework suggested also alters the meaning of zero-touch in that it is not necessary nor desirable for all services to have a complete lack of human involvement. What constitutes mature automation is that human attention is directed in accordance with the level of uncertainty and the potential consequences. Decisions which are routine, reversible and well understood can be left to automation; those that are ambiguous or have a high impact should be escalated together with clear evidence. This approach is more scalable than relying on manual operations or allowing unrestricted autonomy and it also provides a practical way of learning: as the system gathers in its records successful actions and verified results, policy can then authorize a broader range of automatic interventions.

In the end, predictive service management ought to be seen as an organizational capability not a software product. The various functions involved—data engineering, network operations, service management, cybersecurity, cloud teams, and governance—should all agree on the meanings of service health, risk, and acceptable automation. Those who have ownership of the models must also be responsible for dealing with drift and for retiring them, while the people in charge of policy must specify who is authorised to take actions and on what evidence. Without such institutional arrangements, technical intelligence will remain disconnected from resilience; but with them, AI/ML becomes a disciplined way of turning network data into operational decisions that are identified earlier, are safer, and can be defended more economically.

8. Conclusion

The use of AI and machine learning can greatly enhance the resilience of telecom and ICT systems if prediction is incorporated into the service-management control loops rather than being used as standalone analytics. The evidence examined indicates that each of these approaches has its own advantages: deep learning models are better at forecasting traffic over multiple time horizons; methods for detecting faults and anomalies allow for earlier detection; learned orchestration adjusts resources in response to changing demand; digital twins offer a safer environment for testing actions; and federated and explainable methods deal with the issues of distributed data and trust in decisions. No single one of these capabilities is adequate on its own.

The Vision 2030 framework in question therefore includes five stages: observing, predicting, diagnosing, deciding, and ensuring. Under multi-horizon intelligence it features a sovereign telemetry network and a service-dependency graph, and automated action is restricted by means of policy, digital-twin validation, autonomy levels, rollback mechanisms, and post-action verification. Evaluation is shifted from model accuracy to avoided service impact, reduction of SLA risk, shorter intervention lead times, faster recovery, greater energy efficiency, and accountable automation. This approach is more appropriate for critical digital infrastructure since

it acknowledges that a prediction has operational value only if it allows for a timely and safe decision.

The next stage of research should focus on resilience that is cross-domain, uncertainty-aware and based on economic considerations. Telecom AI should be increasingly evaluated in terms of its ability to prevent incidents, contain the spread of problems, reduce the effort required for restoration, and safely learn from the results when real network changes take place. For Saudi Arabia, such an approach will be able to meet the demands for scale and reliability associated with a digital economy while at the same time maintaining governance, data sovereignty and operational accountability. Predictive service management must therefore be developed in a step-by-step manner: first by establishing trustworthy observation, then by making validated predictions, followed by bounded orchestration, and only afterwards by allowing broader autonomy on the basis of evidence.

References

- [1] Benzaid C, Taleb T. AI-Driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions. *IEEE Network*. 2020;34(2):186-194. <https://doi.org/10.1109/MNET.001.1900252>
- [2] El Rajab M, Yang L, Shami A. Zero-touch networks: Towards next-generation network automation. *Computer Networks*. 2024;243:110294. <https://doi.org/10.1016/j.comnet.2024.110294>
- [3] Salhab N, Langar R, Rahim R. 5G network slices resource orchestration using Machine Learning techniques. *Computer Networks*. 2021;188:107829. <https://doi.org/10.1016/j.comnet.2021.107829>
- [4] Wu W, Zhou C, Li M, Wu H, Zhou H, Zhang N, Shen XS, Zhuang W. AI-Native Network Slicing for 6G Networks. *IEEE Wireless Communications*. 2022;29(1):96-103. <https://doi.org/10.1109/MWC.001.2100338>

553 [5] Azimi Y, Yousefi S, Kalbkhani H, Kunz T. Applications of Machine Learning in Resource
554 Management for RAN-Slicing in 5G and Beyond Networks: A Survey. IEEE Access.
555 2022;10:106581-106612. <https://doi.org/10.1109/ACCESS.2022.3210254>

556 [6] Xie Y, Kong Y, Huang L, Wang S, Xu S, Wang X, Ren J. Resource allocation for network
557 slicing in dynamic multi-tenant networks: A deep reinforcement learning approach. Computer
558 Communications. 2022;195:476-487. <https://doi.org/10.1016/j.comcom.2022.09.015>

559 [7] Camargo JS, Coronado E, Ramirez W, Camps D, Sanchez Deutsch S, Perez-Romero J,
560 Antonopoulos A, Trullols-Cruces O, Gonzalez-Diaz S, Otura B, Rigazzi G. Dynamic slicing
561 reconfiguration for virtualized 5G networks using ML forecasting of computing capacity.
562 Computer Networks. 2023;236:110001. <https://doi.org/10.1016/j.comnet.2023.110001>

563 [8] Haider U, Waqas M, Hanif M, Alasmay H, Qaisar SM. Network load prediction and
564 anomaly detection using ensemble learning in 5G cellular networks. Computer Communications.
565 2023;197:141-150. <https://doi.org/10.1016/j.comcom.2022.10.017>

566 [9] Alves PVA, Goldbarg MASS, Barros WKP, Rego ID, Filho VJMT, Martins AM, et al.
567 Machine Learning Applied to Anomaly Detection on 5G O-RAN Architecture. Procedia
568 Computer Science. 2023;222:81-93. <https://doi.org/10.1016/j.procs.2023.08.146>

569 [10] Murphy K, Lavignotte A, Lepers C. Fault Prediction for Heterogeneous Telecommunication
570 Networks Using Machine Learning: A Survey. IEEE Transactions on Network and Service
571 Management. 2024;21(2):2515-2538. <https://doi.org/10.1109/TNSM.2023.3340351>

572 [11] Aktas S, Alemdar H, Ergut S. Towards 5G and beyond radio link diagnosis: Radio link
573 failure prediction by using historical weather, link parameters. Computers & Electrical
574 Engineering. 2022;99:107742. <https://doi.org/10.1016/j.compeleceng.2022.107742>

575 [12] Gao Z. 5G Traffic Prediction Based on Deep Learning. Computational Intelligence and
576 Neuroscience. 2022;2022:3174530. <https://doi.org/10.1155/2022/3174530>

577 [13] Jiang W. Internet traffic prediction with deep neural networks. Internet Technology Letters.
578 2022;5(2):e314. <https://doi.org/10.1002/itl2.314>

579 [14] Lohrasbinasab I, Shahraki A, Taherkordi A, Jurcut AD. From statistical- to machine
580 learning-based network traffic prediction. Transactions on Emerging Telecommunications
581 Technologies. 2022;33(4):e4394. <https://doi.org/10.1002/ett.4394>

582 [15] Yang Y, Geng S, Zhang B, Zhang J, Wang Z, Zhang Y, Doermann D. Long term 5G network
583 traffic forecasting via modeling non-stationarity with deep learning. Communications
584 Engineering. 2023;2:33. <https://doi.org/10.1038/s44172-023-00081-4>

585 [16] Lai J, Chen Z, Zhu J, Ma W, Gan L, Xie S, Li G. Deep Learning Based Traffic Prediction
586 Method for Digital Twin Network. Cognitive Computation. 2023;15(5):1748-1766.
587 <https://doi.org/10.1007/s12559-023-10136-5>

588 [17] Si-Mohammed S, Bardou A, Begin T, Guerin Lassous I, Vicat-Blanc P. NS+NDT: Smart
589 integration of Network Simulation in Network Digital Twin, application to IoT networks. Future
590 Generation Computer Systems. 2024;157:124-144. <https://doi.org/10.1016/j.future.2024.03.038>

591 [18] Muteba F. Digital twin (DT)-based predictive maintenance of a 6G communication network.
592 Procedia Computer Science. 2024;238:544-549. <https://doi.org/10.1016/j.procs.2024.06.058>

593 [19] Tao Z, Xu W, Huang Y, Wang X, You X. Wireless Network Digital Twin for 6G: Generative
594 AI as a Key Enabler. IEEE Wireless Communications. 2024;31(4):24-31.
595 <https://doi.org/10.1109/MWC.002.2300564>

596 [20] Yang Z, Chen M, Wong KK, Poor HV, Cui S. Federated Learning for 6G: Applications,
597 Challenges, and Opportunities. Engineering. 2022;8:33-41.
598 <https://doi.org/10.1016/j.eng.2021.12.002>

599 [21] Shome D, Waqar O, Khan WU. Federated learning and next generation wireless
600 communications: A survey on bidirectional relationship. Transactions on Emerging
601 Telecommunications Technologies. 2022;33(7):e4458. <https://doi.org/10.1002/ett.4458>

602 [22] Al-Quraan M, Mohjazi L, Bariah L, Centeno A, Zoha A, Arshad K, et al. Edge-Native
603 Intelligence for 6G Communications Driven by Federated Learning: A Survey of Trends and
604 Challenges. IEEE Transactions on Emerging Topics in Computational Intelligence.
605 2023;7(3):957-979. <https://doi.org/10.1109/TETCI.2023.3251404>

- [23] Wang S, Qureshi MA, Miralles-Pechuan L, Huynh-The T, Gadekallu TR, Liyanage M. Explainable AI for 6G Use Cases: Technical Aspects and Research Challenges. IEEE Open Journal of the Communications Society. 2024;5:2490-2540. <https://doi.org/10.1109/OJCOMS.2024.3386872>
- [24] Sirohi D, Kumar N, Rana PS, Tanwar S, Iqbal R, Hijjii M. Federated learning for 6G-enabled secure communication systems: a comprehensive survey. Artificial Intelligence Review. 2023;56(10):11297-11389. <https://doi.org/10.1007/s10462-023-10417-3>
- [25] Edozie E, Shuaibu AN, Sadiq BO, John UK. Artificial intelligence advances in anomaly detection for telecom networks. Artificial Intelligence Review. 2025;58:100. <https://doi.org/10.1007/s10462-025-11108-x>
- [26] Abidi MH, Alkhalefah H, Moiduddin K, Alazab M, Mohammed MK, Ameen W, Gadekallu TR. Optimal 5G network slicing using machine learning and deep learning concepts. Computer Standards & Interfaces. 2021;76:103518. <https://doi.org/10.1016/j.csi.2021.103518>
- [27] Oliveira A, Vazao T. Towards green machine learning for resource allocation in beyond 5G RAN slicing. Computer Networks. 2023;233:109877. <https://doi.org/10.1016/j.comnet.2023.109877>
- [28] Ejaz MA, Wu G, Iqbal T. Dynamic and efficient resource allocation for 5G end-to-end network slicing: A multi-agent deep reinforcement learning approach. International Journal of Communication Systems. 2024;37(17):e5916. <https://doi.org/10.1002/dac.5916>
- [29] Yarkina N, Gaydamaka A, Moltchanov D, Koucheryavy Y. Performance Assessment of an ITU-T Compliant Machine Learning Enhancements for 5G RAN Network Slicing. IEEE Transactions on Mobile Computing. 2024;23(1):719-736. <https://doi.org/10.1109/TMC.2022.3228286>
- [30] Zahid AF. The Role of Digital Transformation Strategies in Raising the Operational and Marketing Efficiency of Telecommunications Services in the Kingdom of Saudi Arabia: An Applied Study on the Saudi Telecom Company. The Arab Journal of Administration. 2025;45(2):239-256. <https://doi.org/10.21608/aja.2023.237426.1527>

UNDER PEER REVIEW IN IJAR