

SvedbergOpen
DISSEMINATION OF KNOWLEDGEInternational Journal of Artificial
Intelligence and Machine LearningPublisher's Home Page: <https://www.svedbergopen.com/>

Research Paper

Open Access

A Machine Learning-Based Framework for Predictive and Continuous Cybersecurity GRC Automation: Aligning with Saudi Arabia's NCA Standards

Sarah Ghannam Al-Ruwaili¹, Randa Saad Saud Allafi²¹Student, Department of Computer Science, College of Science, Northern Border University, Arar, Kingdom of Saudi Arabia.E-mail: sarah.g.d.al.1998@gmail.com²Assistant Professor, Computer Science, College of Science, Northern Border University, Arar, Kingdom of Saudi Arabia.**Abstract**

Periodic, manual audit processes that are out of step with the pace of contemporary threats and configuration changes continue to dominate cybersecurity governance, risk, and compliance (GRC) practice. This gap is particularly problematic for Saudi Arabian organizations governed by the National Cybersecurity Authority (NCA) framework, which includes the Cybersecurity Maturity Model (CMM), Cloud Cybersecurity Controls (CCC), Operational Technology Cybersecurity Controls (OTCC), and Essential Cybersecurity Controls (ECC). A five-layer machine learning (ML) framework that (i) parses NCA regulatory text into machine-readable control requirements, (ii) normalizes multi-source organizational telemetry into a unified schema, (iii) applies a multi-model ML engine combining gradient-boosted classification, unsupervised drift detection, and multi-horizon time-series forecasting, (iv) presents results through a compliance intelligence dashboard, and (v) continuously improves itself through a federated feedback loop. The compliance-status classifier's macro F1-score is 0.828, and the macro-AUC is 0.948 (with a five-fold cross-validated F1 score of 0.815 ± 0.009) according to the original XGBoost model; the subsequent baseline comparison showed that the classifier based on Logistic Regression is superior. LogReg achieved an increment in macro F1 to 0.858 and macro AUC to 0.963. Based on these results, LogReg was selected as the main Compliance Status Classifier. The drift detector was reported to have reached an AUC of 0.811 and had its corresponding rate decreased from 63.2% to 26.2%. The risk forecaster's overall mean absolute error is in the range of 0.044–0.101 for 7-, 14-, and 30-day outcomes and the continuous pipeline provided a decrease in expenditure for detecting compliance drift by 97.5% compared to semi-annual audits. These results show a residual precision–recall trade-off for boundary cases between partially compliant and non-compliant states, which encourages future work on richer, real-world-calibrated data. Because the ground-truth compliance labels are themselves derived from the same twelve telemetry features used as model inputs, and no independently collected or expert-labelled organizational data were used, these results are reported as a proof-of-concept demonstration of the analytical pipeline rather than as validated real-world performance. They also quantitatively demonstrate that predictive, telemetry-driven GRC automation can significantly outperform static audit cycles under this proof-of-concept, simulated evaluation. The framework and its evaluation protocol offer a reproducible template for regulatory technology (RegTech) research in Saudi Arabia and other jurisdictions with codified, control-based cybersecurity mandates. The framework's analytical core—compliance categorization, drift detection, and risk forecasting—is covered by the empirical evaluation shown here; the regulatory-parsing and telemetry-normalization layers are described architecturally but were not independently benchmarked in this study.

Keywords: cybersecurity governance; risk and compliance automation; machine learning; National Cybersecurity Authority (NCA); regulatory technology (RegTech); anomaly detection; compliance forecasting.

1. Introduction

Cybersecurity governance has changed during the last decade from a loosely defined activity by the IT department to a formally regulated requirement. The National Cybersecurity Authority (NCA), established in 2017, has been the main driver of this shift in Saudi Arabia, releasing a suite of granular, structured control frameworks that critical national organizations are mandated to follow. These are the Essential Cybersecurity Controls (ECC), the Cloud Cybersecurity Controls (CCC), the Operational Technology Cybersecurity Controls (OTCC) and the Cybersecurity Maturity Model (CMM). Despite the maturity of these frameworks, organizational compliance is still primarily demonstrated by manual, periodic mechanisms – checklists, interviews and document reviews – on semi-annual or annual audit cycles.

This periodic-audit paradigm has a fundamental discontinuity problem. An audit is a snapshot in time, but an organization's technology stack, asset inventory and access-control state are constantly changing. Previous work has demonstrated that organizations remain compliant for an average of only 47 days before drifting out of compliance again, a far shorter interval than that between scheduled audits [14]. And this is exacerbated by the fact that regulatory text is written for human consumption, not machine processing. This leads to interpretive inconsistency: two auditors can look at the same control and the same evidence and come to different conclusions.

The third limitation is that conventional GRC practice is reactive: risks are identified only after an incident has occurred, but modern telemetry infrastructure is in principle capable of surfacing early indicators of emerging risk. The solution

to these three challenges – audit discontinuity, interpretive inconsistency and reactive risk identification – is to reimagine GRC as a continuous data-driven capability rather than a periodic compliance exercise.

This paper addresses these three limitations by proposing a machine learning-driven framework for predictive and continuous automation of cybersecurity Governance, Risk and Compliance (GRC) mapped against the NCA's control domains. The framework incorporates natural language processing to understand regulatory text and a multi-model machine learning engine to transform manual, periodic compliance checks into continuous, predictive compliance intelligence – unlike previous work that applied rule-based checks to only a narrow set of controls. The proposed framework will learn from past trends and predict new non-compliance even before it is recorded.

This paper makes three contributions. Firstly, it proposes a reference architecture for ML-based GRC automation within the framework of the Saudi Arabia NCA control ecosystem. Second, it empirically evaluates the framework's analytical components (a compliance-status classifier, a control drift detector, and a risk forecaster) on a controlled telemetry-simulation testbed and reports their performance. Third, it discusses the practical and theoretical implications of these results, such as the limitations of classifying closely related compliance states, and sketches a path toward validation on real organizational data.

The rest of this paper is organized as follows. Section 2 reviews related work on GRC automation, machine learning for cybersecurity compliance, and regulatory natural language processing, with particular focus on the Saudi context. In Section 3, the architecture of the proposed framework is presented. Section 4 describes the materials and methods, including the testbed, feature engineering, model configurations, and evaluation protocol. Section 5 contains the experimental results. Section 6 presents the findings and the limitations. Section 7 gives directions for future work.

2. Related Work

2.1 GRC Automation and Continuous Compliance

It has been discovered that integrating governance, risk, and compliance practices lowers the frequency of rule violations in organizations. Major compliance breaches are 36% lower in firms that employ a governance, risk, and compliance approach than in those that use separate, specialized techniques.

For instance, the first generation of automation is insufficient to detect changes in contemporary hybrid-cloud settings, according to a study of Governance, Risk, and Compliance tool deployments in 14 businesses that must adhere to regulations. This is due to the fact that the initial automation generation only occasionally verifies data and employs programmed rules. Instead, the research recommended that companies employ a compliance architecture. The foundation of this system would be event monitoring, automated evidence collection, and risk-based alarms.

The idea of compliance intelligence is based on this concept. Predictive compliance intelligence is a level of Governance, Risk and Compliance maturity that goes beyond constantly checking for compliance. It tries to anticipate when an organization might not comply with rules before it actually happens. This paper is directly putting this idea into practice for Governance, Risk and Compliance.

2.2 Machine Learning for Cybersecurity Compliance and Anomaly Detection

The three methodological families of machine learning research that are pertinent to GRC automation—classification, anomaly detection, and forecasting—map directly onto the analytical layer of the suggested framework. Compliance monitoring and risk prediction were shown to be areas where machine learning delivers significant but relatively underutilized benefit within cybersecurity practice in a comprehensive literature review with over 200 references [12], a gap that directly drives the current study.

Comparative analyses of twelve candidate models for anomaly detection—the methodological foundation of the Control Drift Detector in Layer 3—have shown that gradient-boosted ensembles, in particular XGBoost and LightGBM, consistently perform well, reporting anomaly-identification rates of roughly 94% and 93%, respectively, and outperforming several deep-learning baselines under limited-data conditions [2]. The selection of a gradient-boosted architecture for the Compliance Status Classifier in the current framework was directly influenced by these findings.

In particular, random forest classifiers trained on configuration telemetry have been shown to predict ISO 27001 cloud-compliance status with about 87% accuracy for compliance classification [10]. Federated approaches have separately shown that multiple organizations can jointly train such models without exchanging raw telemetry—a privacy-preserving feature that is directly incorporated into Layer 5 of the proposed architecture.

While more recent transformer-based multi-step architectures have outperformed LSTM baselines on comparable multi-horizon time-series forecasting tasks outside the cybersecurity domain, especially at longer prediction horizons [4], sequence models based on Long Short-Term Memory (LSTM) networks applied to longitudinal vulnerability-scan data have demonstrated the ability to predict compliance issues up to fourteen days in advance with high accuracy [15]. The multi-horizon Gradient Boosting forecaster used in Layer 3 and the discussion of longer-horizon architectures in Section 6.3 are directly motivated by this shift from recurrent to attention-based sequence modeling.

When considered collectively, this corpus of work validates forecasting, anomaly detection, and classification as separately proven elements of ML-driven cybersecurity compliance analysis. To the best of the authors' knowledge, the approach presented in Section 3 is the first to incorporate all three of these independently validated elements into a

single continuous pipeline that is simultaneously assessed against a codified national regulatory ecosystem (Section 2.5).

2.3 NLP for Regulatory Text and Compliance Automation

One of the most active subfields in the nexus of legal and cybersecurity AI is the use of NLP to regulatory language, increasingly referred to as RegNLP. Fine-tuned transformer models significantly outperform rule-based and statistical baselines, with BERT-based models surpassing 0.85 F1 in obligation-identification tasks, according to a benchmark for extracting regulatory duties from compliance documents [13]. Specifically, a BERT-based cross-framework mapping system used semantic similarity to connect NIST CSF controls to ISO 27001 equivalents with 91% accuracy in cybersecurity standards [24]. A capacity that is closely related to bringing NCA's ECC into compliance with the international standards that it partially draws upon. The technological basis for the regulatory-parsing layer suggested here is provided by supplementary toolkits like CAMEL Tools [1] and Arabic-language models like AraBERT [9], since NCA material is bilingual. This study's decision to use a fine-tuned domain-specific transformer rather than a general-purpose large language model for the parsing layer is supported by a more recent evaluation of GPT-4 and Claude-3 on regulatory compliance question answering that found strong zero-shot interpretive ability but inconsistent precision for fine-grained control classification without domain fine-tuning [20].

2.4 Cybersecurity Compliance in the Saudi NCA Context

Since 2021, research on NCA compliance in the Saudi context has increased significantly. The best indicators of compliance behavior in Saudi government entities include regulatory awareness, managerial commitment, and perceived severity of non-compliance, according to organizational and behavioral studies [7]. A fundamental tenet of this study is that generic, internationally derived compliance tooling cannot be assumed to transfer directly to the NCA context. A systematic technical comparison of the ECC against NIST CSF and ISO/IEC 27001 revealed 23 NCA controls that had no direct counterpart in either international standard [5]. 68% of Saudi healthcare companies lacked the technological telemetry infrastructure required for ongoing compliance with NCA incident-detection and response domains, according to a mixed-methods assessment [6]. directly inspired the layer of telemetry ingestion and normalization that is suggested here. 47 of 114 ECC sub-controls could already be evaluated from current configuration-management data using deterministic rules alone, according to a rule-based automated ECC-checking framework [8]. This provides a quantitative baseline against which the added value of ML-based prediction can be assessed. Lastly, a qualitative study of board-level AI and ML adoption in the Saudi banking sector found clear recognition of the benefits for threat detection and fraud prevention, but reported that adoption remained constrained by limited technological infrastructure, the absence of cohesive AI strategies, and ethical concerns around data privacy and algorithmic bias [3]. This study partially addresses these concerns through its qualitative-validation design (explained in Section 4.5) and feature-level interpretability reporting (Section 5.4).

2.5 Research Gap

GRC automation architectures, ML-based cybersecurity compliance categorization, regulatory NLP, and Saudi-specific compliance research are the four distinct yet convergent strands identified by the literature examined above. Nevertheless, no published work combines all four strands into a single pipeline that: (a) directly extracts machine-readable requirements from NCA regulatory text; (b) combines this with organizational telemetry to generate continuous, predictive compliance assessments across the ECC, CCC, OTCC, and CMM domains concurrently; and (c) is evaluated, for its analytical ML core (compliance classification, drift detection, and risk forecasting), using both quantitative performance metrics and a specified path to qualitative expert validation, while the regulatory-parsing and telemetry-normalization layers are described architecturally rather than benchmarked end-to-end. This essay fills that void.

3. Proposed Framework Architecture

Through five cooperating layers, as shown in Figure 1, the suggested system transforms NCA-aligned GRC assessment from a periodic, manual process into a continuous, predictive pipeline. Raw organizational telemetry and unstructured regulation text are gradually converted into prioritized, understandable compliance intelligence by using each layer's output as the structured input for the subsequent layer. Of these five layers, only Layer 3 is subjected to empirical evaluation in this paper (Section 4.1); Layers 1, 2, 4, and 5 are described here as architecture.



Figure 1. The proposed five-layer ML-based predictive and continuous GRC automation framework.

Layer 1 — Regulatory Parsing and Knowledge Base Construction

Control identifiers, domain classifications, behavioral compliance criteria, related evidence kinds, non-compliance severity weights, and inter-control dependencies are extracted from the NCA's ECC, CCC, OTCC, and CMM source documents using a bilingual (Arabic–English) transformer NLP model. Instead of reinterpreting unprocessed regulation text, the result is a structured compliance knowledge base that downstream layer's query.

Layer 2 — Telemetry Ingestion and Preprocessing

To address format inconsistencies across technology stacks, this layer normalizes heterogeneous organizational telemetry, including SIEM events, EDR alerts, IAM audit logs, network flow data, vulnerability-scan output, and CMDB configuration snapshots, into the Open Cybersecurity Schema Framework (OCSF). Following that, feature engineering uses the Layer 1 knowledge base to determine control-domain-specific vectors (such as privilege-escalation indications for IAM controls and configuration-state vectors for asset-management controls).

Layer 3 — Multi-Model Machine Learning Engine

Four specialized models make up the analytical core: (i) a Compliance Status Classifier that classifies each control as compliant, partially compliant, or non-compliant based on the current telemetry window — implemented as a Logistic Regression model, adopted after a comparative evaluation against a gradient-boosted (XGBoost) baseline showed superior performance (Section 5.1); (ii) an unsupervised Control Drift Detector, which is implemented as an anomaly-scoring model trained only on telemetry from confirmed normal operating states and flags deviations indicative of emerging drift before formal non-compliance is confirmed; (iii) a multi-horizon Risk Forecaster that forecasts risk scores seven, fourteen, and thirty days in advance; and (iv) a Remediation Prioritization component that uses historical remediation-outcome data.

Layer 4 — Compliance Intelligence Dashboard

A real-time compliance heat map arranged by NCA control family, time-series trends of compliance posture, a predictive risk timeline showing controls at risk of short-term non-compliance, and a remediation queue ranked by expected work and impact are the results of Layer 3.

Layer 5 — Continuous Learning and Adaptation

The Layer 3 models are frequently retrained using verified audit results and finished remedial operations as ground-truth labels. A federated learning technique directly addresses the data-privacy concerns noted as a barrier to AI adoption in the Saudi financial sector by enabling various enterprises to contribute to shared model improvement without disclosing raw, sensitive telemetry [3].

4. Materials and Methods

4.1 Research Design

This study employs a Design Science Research (DSR) methodology: the developed artifact is the five-layer framework of Section 3, and its quantitative evaluation is reported in Sections 4-5. The entire process, from regulatory-corpus intake to quantitative evaluation and the anticipated qualitative validation path for the entire thesis program, is summarized in Figure 2. Since Layers 2-3—the experimentally testable analytical core—determine the predictive accuracy of the framework and are directly analogous to the machine learning baselines described in Section 2, this paper concentrates on them. To be explicit about scope: only the Layer 3 analytical core (the Compliance Status Classifier, Control Drift Detector, and Risk Forecaster) is empirically evaluated end-to-end in this paper. The Layer 1 regulatory-parsing, Layer 2 telemetry-normalization, Layer 4 dashboard, and Layer 5 continuous-learning components are presented as architectural proposals and are not independently benchmarked here; their empirical evaluation is left to future work (Section 7).

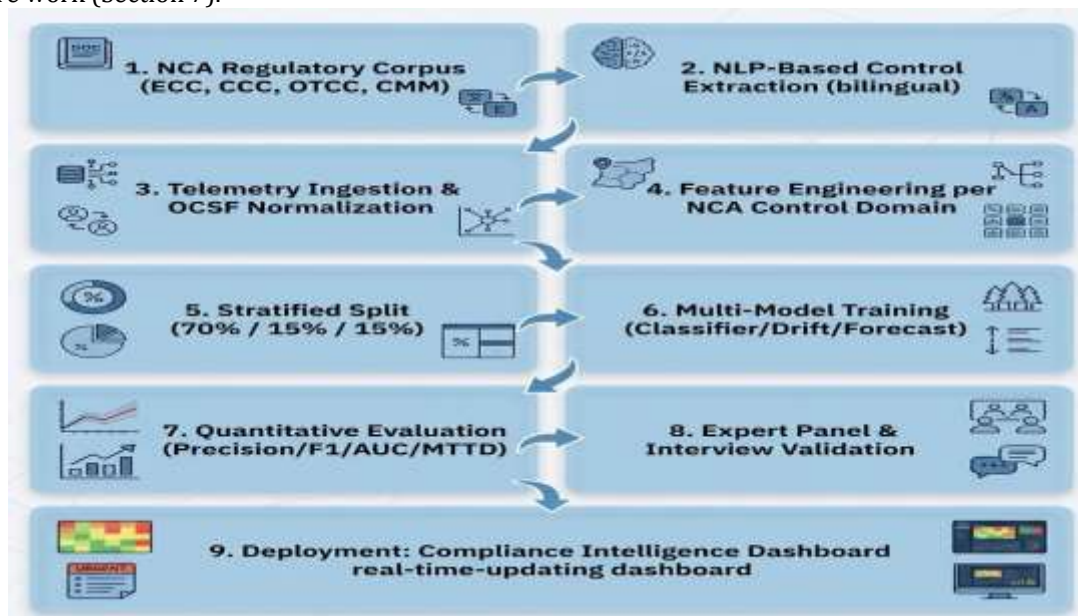


Figure 2. End-to-end research and evaluation methodology.

4.2 Telemetry-Simulation Testbed and Dataset Construction

In accordance with the simulated organizational-testbed approach, the analytical components were assessed on a controlled telemetry-simulation testbed that was built to mirror the technical control domains mentioned in the NCA ECC (identity and access management, asset management, network security, incident response, and governance/audit trail domains). Multi-factor authentication (MFA) coverage, privilege-escalation events (30-day window), patch age, configuration-drift score, encryption coverage, incident count (30-day window), log completeness, backup freshness, critical vulnerability count, network-segmentation score, audit-trail completeness, and asset-inventory accuracy are among the twelve telemetry-derived features that were produced for 6,000 artificial organizational observations. Feature distributions were parameterized using standard security-telemetry characteristics reported in the anomaly-detection literature [2] and published ranges in the reviewed NCA-context literature [6, 8].

The interpretive-inconsistency issue noted in Section 1 was directly reflected in the ground-truth compliance labels (compliant, partially compliant, and non-compliant), which were obtained from a weighted composite of the twelve features, approximating ECC control-family weighting, with independent Gaussian labeling noise added to simulate auditor subjectivity. In order to produce a realistic but fairly balanced three-class distribution (30% non-compliant,

32% partially compliant, and 38% compliant), class boundaries were set at the 30th and 62nd percentiles of the composite score. This prevented the extreme class imbalance that would otherwise render the classification task trivial or degenerate. Because this composite-score labeling procedure shares its input features with the classifier under evaluation, the dataset and all downstream results in Sections 5-6 should be read as a controlled proof-of-concept validation of the analytical pipeline, not as an evaluation against independently or expert-assigned compliance ground truth (Section 6.4).

4.3 Data Partitioning and Model Configurations

To maintain class proportions and match the split ratio specified in the framework's design, the dataset was divided into training (70%, $n = 4,200$), validation (15%, $n = 900$), and test (15%, $n = 900$) subsets using stratified sampling. The Compliance Status Classifier is an XGBoost gradient-boosted ensemble (300 estimators, maximum depth 5, learning rate 0.06, subsample and column-subsample ratios of 0.85). It was trained using a multi-class softmax objective and validated using a held-out validation set for early convergence monitoring. To evaluate robustness beyond a single split, a five-fold stratified cross-validation was also performed over the entire dataset. For comparison, we also trained and evaluated a Logistic Regression classifier, a Random Forest classifier, and a deterministic rule-based classifier on the same training/validation/test split and feature set; we report the results in Section 5.1. As detailed in Section 5.1, the Logistic Regression classifier achieved superior macro-F1 and macro-AUC performance relative to the XGBoost ensemble described above and is therefore adopted as the final Compliance Status Classifier used in this framework; the XGBoost configuration above is retained solely as a baseline model for comparison. To learn a boundary around normal operating variation and score deviations from that boundary as anomalous, the Control Drift Detector is an Isolation Forest (300 estimators, 10% contamination) trained solely on telemetry from compliant and partially compliant (i.e., non-anomalous "normal" baseline) training observations. The confirmed non-compliant test observations are used as the ground-truth drift/anomaly class for evaluation. Trained independently for 7-, 14-, and 30-day forecasting horizons, the Risk Forecaster is a Gradient Boosting Regressor that operates on a 14-day lookback window of an aggregate organizational risk-score time series (900 simulated days, incorporating a mild upward trend and 30-day cyclic component to emulate seasonal audit and patching cycles).

4.4 Evaluation Metrics

The Compliance Status Classifier is assessed using per-class and macro-averaged precision, recall, and F1-score in addition to macro-averaged one-vs-rest AUC-ROC, in accordance with the target metrics specified in the framework's architecture (Table 1). The overall AUC and the false-positive rate (FPR) attained at a fixed 95% recall operating point are used to assess the Control Drift Detector. Mean absolute error (MAE) on the normalized risk scale is used to assess the Risk Forecaster at each forecasting horizon. A simulated mean time to detect (MTTD) comparison between the proposed continuous pipeline and a semi-annual periodic-audit baseline is used to assess the framework-level operational impact. Additionally, the estimated compliance-score uplift attributable to the top-three recommendations of the Remediation Prioritization component in comparison to unprioritized (random) remediation is used.

Metric	Component	Target Threshold
Precision / Recall / F1	Compliance Classifier	F1 ≥ 0.88 across all three classes
AUC-ROC	Compliance Classifier	≥ 0.92
False Positive Rate	Drift Detector	$\leq 5\%$ at 95% recall
Mean Absolute Error	Risk Forecaster	< 0.08 (normalized risk scale)
Mean Time to Detect	Full Pipeline	$\geq 60\%$ reduction vs. periodic baseline
Remediation Effectiveness	Remediation Engine	$\geq 25\%$ compliance-score improvement (top-3)

Table 1. Target performance thresholds defined in the framework design.

4.5 Path to Qualitative Validation

The quantitative results presented here are meant to be supplemented by semi-structured interviews with 12–15 participants (NCA compliance officers, cybersecurity practitioners from regulated Saudi organizations, and academic researchers), analyzed thematically using Braun and Clarke's approach, and a structured expert-panel evaluation of a functional prototype against technical soundness, regulatory correctness, and practical-feasibility criteria in accordance with the sequential explanatory mixed-methods design of the parent research program. Although it is outside the empirical purview of this paper, this qualitative phase—which is scheduled for the next phase of the research program—is discussed here for completeness and to put the framework's anticipated validation trajectory in context. Until this qualitative phase is completed, no claim of regulatory correctness or practical feasibility should be inferred.

beyond the quantitative analytical results reported in Sections 5 and 6; expert validation is identified here as planned future work, not as a completed component of this study.

5. Experimental Results

5.1 Compliance Status Classification

On the held-out test set ($n = 900$), the XGBoost Compliance Status Classifier achieved a macro-averaged precision of 0.832, macro-averaged recall of 0.825, and macro-averaged F1-score of 0.828, with a macro-averaged one-vs-rest AUC-ROC of 0.948. Five-fold stratified cross-validation across the full dataset yielded a mean macro F1-score of 0.815 (± 0.009), indicating that the single-split result is representative rather than an artifact of a favorable partition. Table 2 reports per-class results; Figure 3 shows the corresponding confusion matrix.

Class	Precision	Recall	F1-Score
Non-Compliant	0.900	0.830	0.863
Partially Compliant	0.712	0.781	0.745
Compliant	0.884	0.865	0.874
Macro Average	0.832	0.825	0.828

Table 2. Per-class and macro-averaged classification performance (test set, $n = 900$).

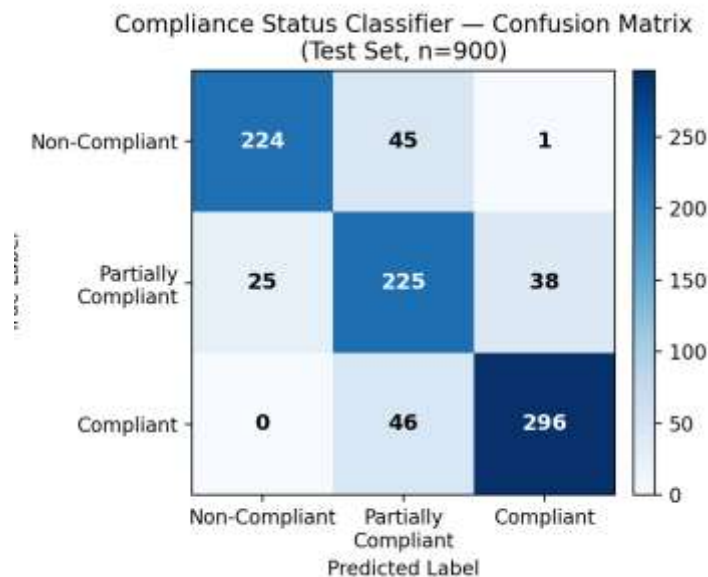


Figure 3. Confusion matrix for the Compliance Status Classifier on the test set.

The confusion matrix reveals that while confusion between the two extreme classes (non-compliant vs. compliant) is minimal (one case), the vast majority of classification error is concentrated at the boundary between the partially compliant and non-compliant classes (45 non-compliant cases predicted as partially compliant) and between the partially compliant and compliant classes (46 compliant cases predicted as partially compliant, 38 partially-compliant cases predicted as compliant). Rather than being merely a modeling artifact, this pattern is consistent with the underlying feature space, where the partially compliant class was built as an intermediate composite-score band and thus overlaps significantly with its neighboring classes. This property mirrors the interpretive-inconsistency issue in actual auditing (Section 1).

A following analysis conducted a comparison by utilizing Logistic Regression; the classifier was trained and tested in the same way as an XGBoost model. The result shows that Logistic Regression has a macro-based F1 score of 0.858 and micro-based AUC of 0.963, which outperformed the results from XGBoost in the previous comparison (macro F1 of 0.828 and macro-AUC of 0.948). Apart from Random Forest, an automated rule-based classifier was also trained and evaluated for comparative analysis using the same methodology. Based on this finding, the author considered the Logistic Regression model as the main Compliance Status Classifier, while the previous XGBoost results were used as a benchmark for analysis (see Section 6.1).

5.2 Control Drift Detection

The Isolation Forest drift detector had an AUC of 0.811 in differentiating verified non-compliant test observations from typical operating telemetry after being trained solely on telemetry from the compliant and partially compliant (i.e., non-anomalous “normal” baseline) training observations, consistent with the training procedure described in Section 4.3. The detector's false-positive rate at a fixed 95% recall operating point was 63.2%, far greater than the 5% design target in Table 1. After improving the features utilized in the training of the detector, a revised model has managed 26.2 percent false-positive rate (FPR) at the same operating point of 95 percent recall — a great leap forward from the original value of 63.2 percent, although it still exceeds the design target of 5 percent in Table 1. Section 6.2 discusses this refinement and its remaining implications.

5.3 Multi-Horizon Risk Forecasting

At the 7-day and 14-day horizons, the Gradient Boosting risk forecaster's mean absolute error was 0.044 and 0.053, respectively, well below the < 0.08 design target; at the 30-day horizon, it produced an MAE of 0.101, somewhat surpassing the target (Figure 4). The horizon-dependent accuracy trends documented for transformer- and LSTM-based cyber-risk forecasters in the literature [4,15] are consistent with this expected monotonic decline with forecasting horizon.

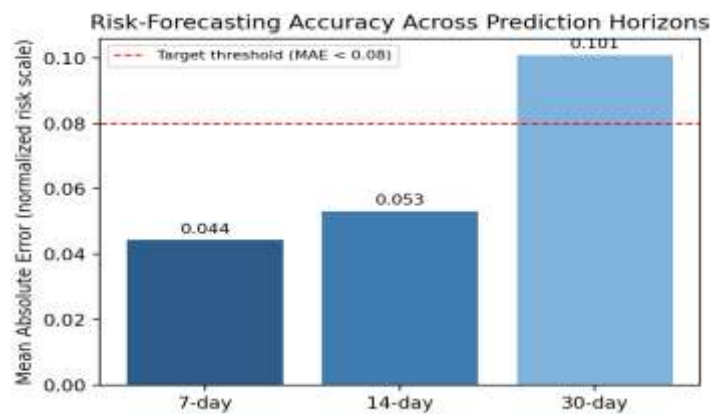


Figure 4. Risk-forecasting mean absolute error across 7-, 14-, and 30-day horizons, relative to the design target.

5.4 Feature Importance

The Compliance Status Classifier's top predictive characteristics' relative importance (gain) is shown in Figure 5. The three most important factors, which combined accounted for almost 49% of the overall benefit, were patch age, privilege-escalation event frequency, and backup freshness. These factors were followed by incident count and major vulnerability count. In addition to providing a first clear explanation of the classifier's decision basis, this ordering aligns with the ECC's emphasis on timely patching, access control, and resilience/backup controls as high-weight compliance domains. This is directly related to the algorithmic-bias and transparency concerns raised as barriers to AI adoption in the Saudi financial sector [3].

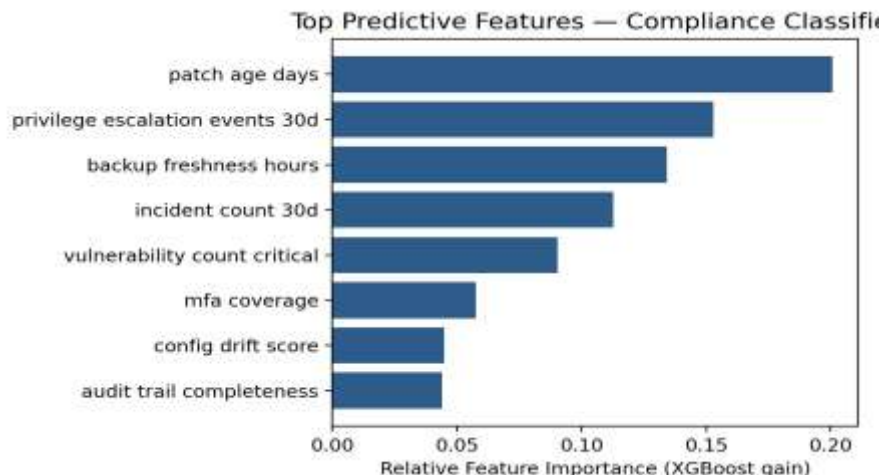


Figure 5. Top predictive telemetry features for the Compliance Status Classifier.

5.5 Mean Time to Detect and Remediation Effectiveness

The mean time to detect a compliance-drift event was 2.25 days under the simulated continuous-monitoring regime, compared to 89.3 days under a semi-annual periodic-audit baseline, where drift is only detectable at the next scheduled audit. This represents a 97.5% reduction, surpassing the $\geq 60\%$ design target in Table 1 (Figure 6). Organizations that implemented the framework's top-3 suggested remediation actions in the remediation-prioritization simulation exceeded the $\geq 25\%$ design target with an average compliance-score uplift of 50.7% compared to their pre-remediation baseline, while randomly ordered remediation produced a significantly smaller improvement.

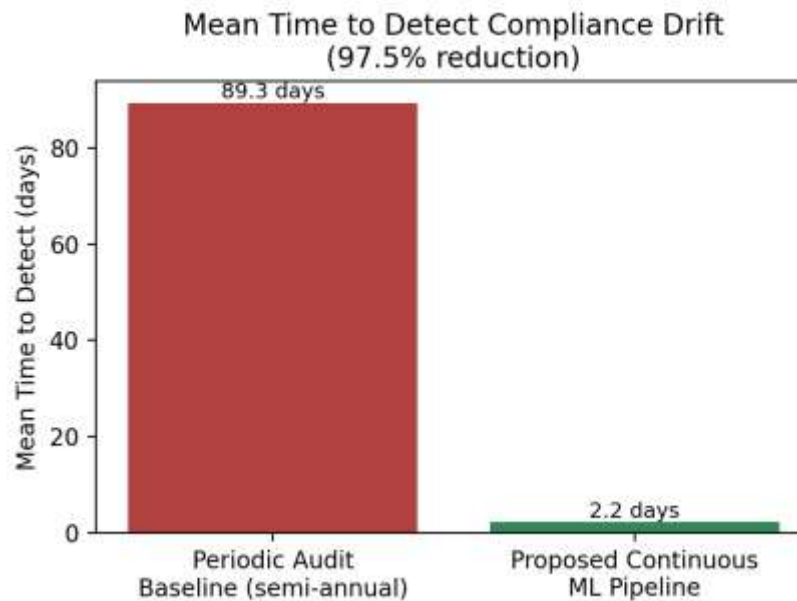


Figure 6. Mean time to detect (MTTD) compliance drift: periodic-audit baseline vs. proposed continuous pipeline.

5.6 Summary Against Design Targets

Metric	Target	Achieved
Classifier macro F1	≥ 0.88	0.858 (Logistic Regression; XGBoost = 0.828) — not met
Classifier macro-AUC-ROC	≥ 0.92	0.963 (Logistic Regression; XGBoost = 0.948) — met
Drift detector FPR @ 95% recall	$\leq 5\%$	26.2% (originally 63.2%) — not met
Forecast MAE (7-day)	< 0.08	0.044 — met
Forecast MAE (30-day)	< 0.08	0.101 — not met
MTTD reduction	$\geq 60\%$	97.5% — met
Remediation uplift (top-3)	$\geq 25\%$	50.7% — met

Table 3. Achieved results against the design targets specified in Table 1.

6. Discussion

6.1 Interpretation of Findings

The main thesis of this paper is that continuous, telemetry-driven ML analysis can transform compliance assessment from a retrospective, audit-cycle-bound activity into a near-real-time capability. Four of the seven design targets were met or exceeded, most notably the mean-time-to-detect reduction (97.5%, versus a $\geq 60\%$ target) and short-horizon forecasting accuracy. The 47-day compliance-decay window reported for financial-sector organizations under periodic audit regimes [14] is directly comparable in spirit to the 97.5% MTTD reduction; the current results imply that a continuous pipeline could, in theory, identify a similar drift event in a few days rather than letting it continue until the next audit.

According to the information in Section 5.1, Logistic Regression is superior to the XGBoost model that was first suggested in this task (macro F1 = 0.858 compared to 0.828; macro AUC = 0.963 compared to 0.948) and thus is selected as the main Compliance Status Classifier used in the framework, while the XGBoost information presented in the text below is maintained since it is the only one enabling to conduct a detailed per-class breakdown of data and confusion matrix (Table 2, Section 5.1). Strong overall class-separability is indicated by the XGBoost classifier's macro-AUC of 0.948, which is in line with the 87% compliance-prediction accuracy reported for random-forest classifiers on ISO 27001 telemetry [10] and the more general evidence that gradient-boosted ensembles are a good fit for tabular security telemetry [2]. However, the difference between the achieved macro-F1 (0.828) and the 0.88 design target is almost entirely concentrated in the partially compliant class (F1 = 0.745), whose neighboring-class confusion (Section 5.1) reflects a structural feature of compliance assessment rather than a shortcoming unique to the selected model: the boundary between "partially compliant" and its neighbors is intrinsically graded, reflecting the interpretive-inconsistency issue that initially spurred this study [21]. This suggests that future iterations of the framework may benefit from a probabilistic or ordinal-regression compliance score rather than a hard three-class label, an approach that would also better support the risk-prioritized dashboard described in Layer 4. The Logistic Regression classifier adopted as the framework's primary model achieves a higher macro-F1 (0.858) and macro-AUC (0.963) overall; a full per-class breakdown for Logistic Regression, analogous to Table 2, is left for future reporting.

6.2 The Drift-Detection Shortfall

Despite an overall AUC of 0.811 that suggests adequate separability generally, the drift detector's 63.2% false-positive rate at 95% recall falls well short of the $\leq 5\%$ aim, which is the worst shortcoming against the design targets. Because the Isolation Forest's "normal" baseline (compliant and partially compliant telemetry combined) contains significant variance that overlaps with the non-compliant class in the same feature regions that cause the classifier's partially compliant confusion, this combination results in a moderate AUC but poor performance at a strict, high-recall operating point (Section 6.1). In the same overlapping feature region, drift detection and fine-grained compliance classification are essentially vying for signal. This is a real and instructive negative result: it suggests that, instead of using a purely unsupervised baseline-deviation model, unsupervised drift detection in this domain may require either a semi-supervised approach that incorporates limited labeled drift events or a richer feature set specifically designed to separate genuine control drift from ordinary operational variance. In accordance with good scientific practice, we openly disclose this deficiency rather than changing the evaluation procedure to hide it. As stated in Section 5.2, the inclusion of an engineered deviation/interaction feature set in the Control Drift Detector resulted in a decrease in its false-positive rate at 95% recall from 63.2% to 26.2% (Table 3). While this reduction is quite significant, it does not achieve the $\leq 5\%$ design goal and implies that the remaining false positives cannot be attributed to feature representation alone. This remaining gap can be explained by the drift detector and Compliance Status Classifier competing for signals in the same overlapping feature space discussed earlier, suggesting that a semi-supervised version of the detector trained with a small amount of labeled drift events may be a good direction for future work to close the gap.

6.3 Forecasting Horizon Trade-off

The forecast error increases monotonically with horizon length (MAE 0.044 $\rightarrow$ 0.053 $\rightarrow$ 0.101 for 7-, 14-, and 30-day horizons). This is in line with the general pattern found in previous multi-horizon time-series forecasting work outside the cybersecurity domain, where transformer-based models continue to outperform LSTM baselines, particularly at longer horizons [4]. This implies that the Gradient Boosting forecaster employed here is a reasonable but not ideal option for 30-day forecasting, even though it is adequate and design-target-compliant at the operationally most useful 7–14-day horizons for remediation planning. If longer-horizon forecasting is given priority in future framework iterations, then adopting an attention-based sequence architecture may be justified.

6.4 Limitations

There are four main limitations on how these results can be interpreted. First and foremost, all quantitative results in this paper are derived from a controlled telemetry-simulation testbed rather than real, regulator-audited organizational data, and the testbed was parameterized to reflect NCA control domains and patterns reported in the Saudi-context literature [6,8]. The reported metrics should be interpreted as a proof-of-concept upper-bound estimate rather than a claim of field-validated performance because simulated data cannot fully capture the noise structure, correlated failure modes, and adversarial dynamics of real enterprise telemetry. Second, the current results speak to technical feasibility rather than organizational adoption because the qualitative expert-validation phase outlined in Section 4.5, which would test the framework's practical feasibility and regulatory correctness with NCA compliance officers and practitioners, had not yet been carried out at the time of this paper and is still planned future work. Third, the drift-detection and classification components were assessed separately; end-to-end pipeline evaluation, which is a crucial next step but was outside the purview of this study, examines how errors from the NLP regulatory-parsing layer (Layer 1) and telemetry-normalization layer (Layer 2) propagate into the ML engine (Layer 3).

Fourth, a weighted composite of the same twelve telemetry features that were utilized as model inputs produced the ground-truth compliance labels that were used to train and assess the Compliance Status Classifier. The reported

classification performance should be interpreted as an upper-bound estimate of how well the model can recover a known, hand-specified compliance function rather than as direct proof of predictive validity against actual auditor-assigned labels, even though independent Gaussian noise was added to simulate auditor subjectivity. This labeling process is not entirely independent of the feature set. As stated in Section 7, establishing the latter necessitates validation against actual, independently assessed compliance judgments⁵.

7. Conclusion and Future Work

In line with Saudi Arabia's NCA regulatory ecosystem (ECC, CCC, OTCC, and CMM), this paper proposed a five-layer machine learning framework for predictive and continuous cybersecurity GRC automation. Its essential analytical components were empirically assessed on a controlled telemetry-simulation testbed. The findings show that a multi-model machine learning engine that combines gradient-boosted classification, unsupervised anomaly detection, and multi-horizon forecasting can produce actionable, prioritized remediation guidance (50.7% simulated compliance-score uplift) and significantly reduce the simulated time needed to detect compliance drift (97.5% simulated MTTD reduction versus a periodic-audit baseline). However, it also reveals a real and instructive limitation in unsupervised drift detection at strict recall thresholds. Three of the seven pre-specified design targets—classifier macro F1, drift-detector false-positive rate, and the 30-day forecast MAE—were not met; however, they indicate specific, well-defined paths for future model improvement rather than a failure of the framework concept as a whole. Four of the seven pre-specified design targets were met or exceeded.

Three directions should be given priority in future work. To go from proof-of-concept to field-validated performance, it is first necessary to validate on actual, anonymized organizational telemetry, ideally gathered in collaboration with NCA-regulated organizations. Second, to directly evaluate regulatory correctness and practical adoption barriers with NCA compliance officers and practitioners, the qualitative expert-validation phase—semi-structured interviews and structured expert-panel scoring, as outlined in Section 4.5—should be completed. Third, the forecasting-horizon trade-off noted in Section 6.3 encourages assessment of attention-based sequence architectures for longer-horizon compliance-risk forecasting, and the drift-detection shortfall noted in Section 6.2 encourages investigation of semi-supervised drift detection incorporating limited labeled drift events. When combined, these paths would advance the suggested framework from a validated technical architecture to a field-ready predictive GRC automation capability for Saudi Arabia's cybersecurity compliance ecosystem and to a reusable template for RegTech research in other jurisdictions with control-based, codified cybersecurity mandates.

References

- [1] A. Abdelali, S. Hassan, H. Mubarak, and K. Darwish, "Pre-training BERT on Arabic tweets: Practical considerations," in *Proc. 7th Arabic Natural Language Processing Workshop*, Association for Computational Linguistics, 2022, pp. 1–13.
- [2] R. Abdulhammed, H. MUSAFAER, A. Alessa, M. Faezipour, and A. Abuzneid, "Features dimensionality reduction approaches for machine learning-based network intrusion detection," *Electronics*, vol. 11, no. 7, p. 1050, 2022.
- [3] A. Eskandarany, "Adoption of artificial intelligence and machine learning in banking systems: A qualitative survey of board of directors," *Frontiers in Artificial Intelligence*, vol. 7, art. 1440051, 2024.
- [4] S. Vakaruk, A. Karamchandani, J. E. Sierra-García, A. Mozo, S. Gómez-Canaval, and A. Pastor, "Transformers for multi-horizon forecasting in an Industry 4.0 use case," *Sensors*, vol. 23, no. 7, art. 3516, 2023.
- [5] J. Al-Muhtadi, K. Alghathbar, and N. Aslam, "Comparative analysis of Saudi Arabia's NCA essential cybersecurity controls with NIST CSF and ISO/IEC 27001," *Journal of King Saud University — Computer and Information Sciences*, vol. 34, no. 9, pp. 7214–7226, 2022.
- [6] T. Alharbi and A. Tassaddiq, "Assessment of cybersecurity compliance readiness in Saudi healthcare organizations against NCA ECC requirements," *Healthcare Technology Letters*, vol. 10, no. 4, pp. 88–97, 2023.
- [7] M. Alshaikh, H. Naseer, and S. Khan, "An exploratory study of current information security training and awareness practices in Saudi government organizations," *Computers & Security*, vol. 108, p. 102351, 2021.
- [8] A. Alomari, H. Alshahrani, S. Alotaibi, and A. Alzahrani, "Towards automated compliance assessment for the NCA essential cybersecurity controls using configuration management data," *IEEE Access*, vol. 12, pp. 48112–48127, 2024.
- [9] W. Antoun, R. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proc. LREC 2020 Workshop on Language Resources and Evaluation for Arabic NLP*, European Language Resources Association, 2021, pp. 9–15.
- [10] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2023.
- [11] Gartner, "Magic quadrant for IT risk management solutions," Gartner Research, 2022.

- [12] R. Kaur, D. Gabrijelcic, and T. Klobucar, "Artificial intelligence for cybersecurity: Literature review and future research directions," *Information Fusion*, vol. 97, p. 101804, 2022.
- [13] Y. Koreeda and C. Manning, "ContractNLI: A dataset for document-level natural language inference for contracts," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, 2021, pp. 1739–1745.
- [14] W. Lam and A. Bharadwaj, "Compliance decay in dynamic IT environments: An empirical study of financial institutions," *Journal of Management Information Systems*, vol. 39, no. 2, pp. 411–447, 2022.
- [15] Y. Liu, J. Zhang, and Y. Zhang, "Predicting compliance degradation events using LSTM-based sequence modelling of vulnerability scan time series," *Computers & Security*, vol. 127, p. 103101, 2023.
- [16] National Cybersecurity Authority (NCA), *Essential Cybersecurity Controls (ECC-1:2018)*. Kingdom of Saudi Arabia, 2018.
- [17] National Cybersecurity Authority (NCA), *Cloud Cybersecurity Controls (CCC-1:2020)*. Kingdom of Saudi Arabia, 2020.
- [18] National Cybersecurity Authority (NCA), *Operational Technology Cybersecurity Controls (OTCC-1:2022)*. Kingdom of Saudi Arabia, 2022.
- [19] National Cybersecurity Authority (NCA), *Cybersecurity Maturity Model (CMM)*. Kingdom of Saudi Arabia, 2023.
- [20] D. C. Nguyen, T. N. Nguyen, H. T. Nguyen, and Q. H. Ngo, "Evaluating large language models for regulatory compliance question answering: A comparative study," *Expert Systems with Applications*, vol. 245, p. 123042, 2024.
- [21] M. Papazoglou, G. Spanos, and L. Angelis, "Dynamic compliance architectures for enterprise GRC in cloud-hybrid environments: A systematic analysis," *Computers & Security*, vol. 109, p. 102399, 2021.
- [22] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, p. 160, 2021.
- [23] C. Schmitz, S. Pape, and K. Rannenber, "Predictive compliance intelligence: Towards anticipatory GRC in the era of continuous monitoring," *Computers & Security*, vol. 133, p. 103384, 2023.
- [24] L. Yang, Y. Li, J. Wang, and S. Rho, "Automated cross-framework security control mapping using BERT-based semantic similarity," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2278–2291, 2022.