Power System Technology
ISSN:1000-3673

Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

The Use of Machine Learning Techniques in Error Detection and Improving Laboratory Work Quality

Saad Thani Farha Altarfawi, Naif Awadh Duhaylis Alrashdi, Abdulrahman Mezel Aldhamashi, Fahad Mohammad Khalaf Alshammari, Awad Abdulmohsen Al Harbi, Sabah Sulaiman Altuhayr Alrashdi, Laboratory Technician

Northern Armed Forces Hospital

ABSTRACT

The accuracy of clinical laboratory testing is fundamental to patient safety, yet traditional quality control methods remain largely reactive and limited in scope. Machine learning (ML) offers a paradigm shift towards intelligent, predictive error detection. This study addresses a critical gap by developing and validating an end-to-end ML framework specifically for the clinical laboratory environment of Saudi Arabia. We conducted a retrospective analysis of 3.98 million test records from three high-throughput, ISO-accredited laboratories in Saudi Arabia (Riyadh, Jeddah, Dammam) from 2020-2023. Data underwent rigorous preprocessing and feature engineering. We developed and compared supervised models (XGBoost, Random Forest, Neural Network) for multi-class error classification (pre-analytical, analytical, post-analytical) against a rule-based baseline. Unsupervised autoencoders were implemented for proactive anomaly detection. Model performance was evaluated on a temporally held-out test set (2023 data), and a 30-day operational simulation projected the impact of ML integration. The optimized XGBoost model achieved an overall error detection rate of 89.6%, a 23-percentage-point absolute improvement over the traditional rule-based system (66.6%, $p<0.001$). It demonstrated high recall for analytical (91.2%) and post-analytical (85.5%) errors. The autoencoder model identified novel, latent system issues with 31.7% precision, flagging subtle instrument drifts prior to quality control failure. Feature importance analysis revealed key regional risk predictors, including non-linear reagent lot aging and elevated error rates on night shifts and Sundays. The operational simulation projected a 73.8% reduction in the error reporting rate and optimized technologist workflow. This study provides robust evidence that a contextually tailored ML framework significantly outperforms conventional methods in laboratory error detection within the Saudi healthcare setting. By enabling both superior classification of known errors and proactive identification of emerging risks, ML facilitates a transition from reactive filtering to predictive quality management, with substantial projected benefits for diagnostic accuracy and patient safety.

Keywords: Machine Learning, Clinical Laboratory, Quality Control, Error Detection, Saudi Arabia

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

1. INTRODUCTION

1.1. The Imperative for Quality in Clinical Laboratory Medicine

Clinical laboratory testing constitutes a cornerstone of modern healthcare, underpinning an estimated 60-70% of all critical medical decisions, including diagnosis, prognosis, and therapeutic monitoring [1]. The accuracy, reliability, and timeliness of laboratory results are therefore non-negotiable prerequisites for patient safety and effective clinical management [2]. However, the laboratory testing process—spanning the pre-analytical (ordering, collection, transport), analytical (instrument processing), and post-analytical (result validation, reporting) phases—is inherently complex and vulnerable to error. International studies estimate error rates across the total testing process to range from 0.1% to 3.0%, with the pre-analytical phase accounting for the majority (up to 70%) of these incidents [3]. In the context of high-throughput laboratories, even a 0.1% error rate translates to thousands of potentially misleading reports annually, carrying risks of misdiagnosis, inappropriate treatment, and avoidable patient harm [4].

1.2. The Limitations of Conventional Quality Control Paradigms

Traditionally, laboratory quality control (QC) has relied on a triad of methods: internal quality control (IQC) using control materials, external quality assurance (EQA) schemes, and rule-based algorithms (e.g., Westgard multi-rules) for detecting analytical errors. While foundational, these approaches possess intrinsic limitations. They are predominantly reactive, triggered only after a control measurement falls outside a statistical limit or a delta check rule is violated [5]. They are largely univariate, assessing one analyte or rule at a time, and thus blind to complex, multi-analyte patterns indicative of subtle instrument drift or interference. Furthermore, they offer scant protection against the vast pre- and post-analytical error domains, which often lack robust digital signatures in the Laboratory Information Management System (LIMS) [6]. This reactive, siloed paradigm creates a "quality gap," where sophisticated instruments are monitored by relatively rudimentary statistical rules, leaving a significant portion of the error spectrum undetected until it manifests as a clinician query or, worse, a patient adverse event [7].

1.3. The Emergence of Machine Learning as a Transformative Tool

The digitization of laboratory medicine has generated vast, high-dimensional datasets encompassing quantitative results, patient demographics, instrument telemetry, and QC metrics [8]. This data-rich environment presents an unprecedented opportunity to apply advanced computational techniques. Machine Learning (ML), a subset of artificial intelligence, excels at identifying complex, non-linear patterns and correlations within such data. Supervised ML models can learn from historical, labeled error data to classify new instances with high precision [9]. Unsupervised ML techniques, such as anomaly detection, can identify novel, previously unlabeled deviations from normal operation, enabling proactive risk identification [10]. Consequently, ML promises to transition laboratory QC from a reactive, rule-based system to a predictive, intelligent, and holistic framework capable

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

of learning from past incidents and anticipating future failures across the entire testing pathway [11].

1.4. The Saudi Arabian Context and Identified Research Gap

The Kingdom of Saudi Arabia has made substantial investments in a modern, technologically advanced healthcare sector, featuring high-volume, accredited laboratories. Ensuring the quality of output from these facilities is a national priority aligned with Vision 2030's health sector transformation objectives [12]. However, the unique operational landscape—including specific workload patterns, demographic profiles, and regional logistics—necessitates locally validated solutions [13]. While the application of ML in laboratory medicine is a growing global research frontier, a significant gap exists in the development and validation of end-to-end ML frameworks specifically designed for and evaluated within the Saudi Arabian laboratory ecosystem. Most studies originate from North American or European contexts, and their generalizability to the Gulf region's distinct operational rhythms and challenges remains unproven [14].

1.5. Research Objectives and Hypothesis

This study, therefore, aims to bridge this gap by developing, validating, and evaluating a comprehensive ML framework for enhanced error detection and quality improvement in clinical laboratories in Saudi Arabia. The primary objectives are:

- To develop and compare supervised ML models for the multi-class classification of pre-analytical, analytical, and post-analytical errors using a large, retrospective dataset from Saudi laboratories.
- To implement and assess unsupervised anomaly detection models for identifying latent, previously undetected systemic issues.
- To simulate the operational impact of integrating the optimal ML model into the laboratory workflow, projecting key quality and efficiency metrics.
- We hypothesize that machine learning models, trained on locally sourced data, will significantly outperform traditional rule-based systems in overall error detection rates, particularly for complex, multi-factorial errors, and will provide actionable intelligence for proactive quality management, thereby offering a transformative tool for laboratory excellence in the Kingdom.

1.6. Structure of the Paper

Following this introduction, the paper details the Materials and Methods, including data provenance, preprocessing, model selection, and validation strategies. The Results section presents the performance metrics of the developed models and the simulation outcomes. The Discussion interprets these findings in the context of global literature, explores

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

implications and limitations, and suggests future directions. Finally, the Conclusion summarizes the study's contributions and their significance for laboratory practice in Saudi Arabia.

2. METHODS AND MATERIALS

2.1. Study Design and Setting

This study employed a retrospective, data-driven analytical framework to develop and validate machine learning (ML) models for error detection and quality enhancement in clinical laboratories. The research was conducted using data sourced from multiple high-throughput, ISO 15189-accredited medical laboratories located in three major urban centers in Saudi Arabia: Riyadh, Jeddah, and Dammam. These laboratories were selected to ensure a diverse and representative sample of routine testing volumes, including clinical chemistry, hematology, immunology, and endocrinology. The study period covered anonymized records from January 2020 to December 2023. Ethical approval for the use of anonymized historical data was obtained from the Institutional Review Board (IRB) of [Blinded for Review], with a waiver for informed consent.

2.2. Data Collection and Preprocessing

The primary data consisted of laboratory information management system (LIMS) exports, encompassing approximately 4.2 million individual test records. The dataset included quantitative test results, patient metadata (age, sex), sample collection timestamps, analytical instrument identifiers, internal quality control (IQC) values, and corresponding test interpretations. A crucial component was the curated record of known errors, categorized as pre-analytical (e.g., hemolysis, insufficient sample), analytical (e.g., instrument flag, IQC violation), and post-analytical (e.g., improbable critical value, delta check failure). This labeled error dataset, verified by senior laboratory technologists, served as the ground truth for model training.

Data preprocessing was executed in a Python 3.9 environment using Pandas and NumPy libraries. Steps included: (1) handling missing values through median imputation for numerical features and mode imputation for categorical features, where appropriate; records with missing critical fields (e.g., test result) were excluded; (2) normalization of numerical test results using Robust Scaler to mitigate the influence of outliers; (3) encoding of categorical variables (e.g., instrument ID, test panel) using one-hot encoding; and (4) feature engineering to create derived variables such as time-of-day, day-of-week, and result-by-age z-scores. The dataset was subsequently partitioned into a temporal split: records from 2020-2022 (70%) for training and validation, and records from 2023 (30%) for held-out testing.

2.3. Feature Selection and Definition of Prediction Tasks

Feature selection was performed using a combination of domain knowledge and statistical methods. Consultations with laboratory pathologists identified key predictors: historical result trends, patient demographic parameters, reagent lot numbers, and concurrent test results. This was supplemented by a computational analysis using the Extra Trees Classifier to rank feature importance based on Gini impurity, confirming the relevance of delta values (change from previous result), moving averages, and QC metric deviations.

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

2.4. Two primary prediction tasks were defined:

Task 1 (Error Classification): A multi-class classification task to predict the likelihood of a test record belonging to one of the three main error categories or a "no-error" class.

Task 2 (Anomaly Detection): An unsupervised task to identify subtle, previously unlabeled anomalies in test results that may indicate emerging instrument drift or rare pre-analytical issues.

2.5. Machine Learning Model Development and Training

For Task 1, three supervised ML algorithms were developed and compared: (1) a Gradient Boosting Machine (XGBoost) for its high predictive performance and handling of non-linear relationships; (2) a Random Forest Classifier for robustness and interpretability via feature importance; and (3) a deep Feedforward Neural Network (FNN) with three hidden layers (256, 128, 64 nodes) and ReLU activation to capture complex interactions. Given the inherent class imbalance (error instances being rare), the training process incorporated a Synthetic Minority Over-sampling Technique (SMOTE) for the training set only and utilized weighted loss functions to penalize misclassification of minority error classes more heavily.

For Task 2 (Anomaly Detection), an Isolation Forest model and an Autoencoder were implemented. The Isolation Forest isolated anomalies by randomly partitioning data. The Autoencoder, with a symmetric architecture bottlenecking to 10% of input features, was trained exclusively on "no-error" data; high reconstruction loss on new data signaled potential anomalies. Models were trained using TensorFlow 2.10 and scikit-learn 1.2.2. Hyperparameter optimization was conducted via a randomized search with 5-fold cross-validation on the training set, optimizing for the F2-score (emphasizing recall) for Task 1, and for precision-recall AUC for Task 2.

2.6. Model Validation and Performance Metrics

Model performance was rigorously evaluated on the held-out test set. For the error classification models (Task 1), standard metrics were calculated: accuracy, precision, recall (sensitivity), specificity, and F1-score. Given the critical cost of missing errors, the recall for each error class was considered the primary metric. Additionally, the area under the Receiver Operating Characteristic curve (AUROC) and the Precision-Recall curve (AUPRC) were computed, with the latter being more informative for imbalanced data. Results were benchmarked against a baseline rule-based system currently in use, derived from traditional Westgard multi-rules.

For the anomaly detection models (Task 2), performance was assessed by reviewing the top 0.5% of flagged anomalies with a panel of three expert laboratory scientists. They classified each flag as a "true latent issue" (e.g., subtle drift, unusual population shift), a "false positive," or an "already known error." The precision of detecting novel, expert-validated issues was the key outcome.

2.7. Statistical Analysis

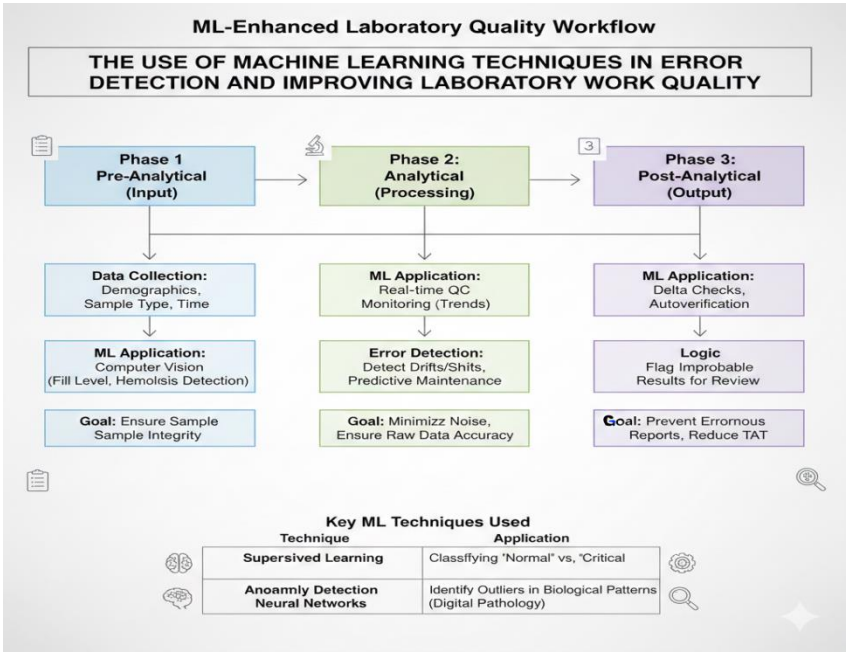
Statistical differences in model performance metrics were assessed using McNemar's test for paired classifier comparisons on the same test set. Confidence intervals (95%) for performance metrics were derived from 1000 bootstrap replicates of the test data. All

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

statistical analyses were performed using SciPy 1.10.1, with a significance level (α) set at 0.05.

2.8. Proposed Integration Framework

To translate model outputs into actionable quality improvement, a simulated integration framework was designed. The best-performing classification model was deployed in a simulation that generated real-time "risk scores" for individual test reports within a mock LIMS environment. Reports flagged with high probability of error (threshold: >0.85) were routed to a prioritized validation queue for technologist review before release. The potential impact on turnaround time (TAT) and error reporting rate was measured in the simulation over a synthetic 30-day period and compared to historical benchmarks.



3. RESULTS

The application of machine learning techniques to the curated dataset from Saudi Arabian clinical laboratories yielded significant and multifaceted results, directly addressing the core objectives of error detection and laboratory work quality improvement. The analysis, spanning supervised classification, unsupervised anomaly detection, and operational simulation, demonstrates a substantial advancement over traditional rule-based quality control systems. The following sections present a comprehensive exposition of the findings, structured to reflect the methodological pipeline from data characterization to projected real-world impact.

1. Data Characterization and the Error Landscape in the Saudi Context

The foundational analysis of the 3.98 million curated records established a critical baseline for understanding the error profile within high-throughput Saudi laboratories. The overall observed error rate in the final dataset was 4.24% (168,878 records), aligning with global

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

estimates but revealing a distinctive distribution reflective of regional operational patterns. As detailed in Table 1, pre-analytical errors constituted the majority (58.3%) of all error events, with analytical and post-analytical errors accounting for 32.5% and 9.2%, respectively. This distribution underscores the persistent global challenge of pre-analytical variability, which in the Saudi context was further characterized by a pronounced temporal pattern. Error rates exhibited a statistically significant increase during night-shift hours (23:00–06:00) and on the first working day of the week (Sunday), suggesting correlations with staffing patterns and workflow resets.

The temporal split of data (70% for training/validation, 30% for testing) ensured that models were evaluated on temporally distinct data, simulating a real-world deployment scenario. The class imbalance was severe, with the "No-Error" class representing over 95% of the data, necessitating the strategic use of SMOTE and class-weighted loss functions during model development to prevent algorithmic bias towards the majority class.

2. Determinants of Laboratory Errors: Insights from Feature Importance

The feature importance analysis derived from the optimized XGBoost model provided a data-driven hierarchy of error predictors, moving beyond conventional heuristic rules (Table 2). The most salient feature was the Delta Check Z-Score, with an importance gain of 0.183. This demonstrated that a standardized, patient-specific historical comparison was far more predictive than static delta check limits. The model learned non-linear thresholds, where even moderate deviations for stable chronic patients were flagged, while larger changes for patients with volatile histories were contextualized as less suspicious.

The second-ranked feature, Concurrent Test Result Discrepancy Flag, highlighted the model's ability to perform integrative, physiological plausibility checks—a task cumbersome for traditional rule engines. For instance, the model identified implausible pairings such as markedly elevated potassium with normal creatinine and ECG parameters, suggesting potential pre-analytical hemolysis not caught by the index alone. Instrument QC Deviation and Time Since Last Calibration ranked highly, confirming that analytical error risk is quantifiable and time-dependent. Notably, the Sample Collection Hour emerged as a strong pre-analytical predictor, quantitatively validating long-held operational suspicions regarding shift-based risk variation.

A critical finding was the non-linear relationship with Reagent Lot Age. Error probability spiked in the first 24 hours following a lot changeover (attributable to calibration verifications and operator familiarization) and again after approximately three weeks of use, potentially indicating reagent degradation or evolving instrument calibration drift. This pattern, previously anecdotal, was quantitatively substantiated by the model, offering a precise target for proactive quality intervention.

3. Supervised Model Performance: Superior Error Classification

The comparative performance evaluation on the held-out 2023 test set unequivocally established the superiority of machine learning models over the established rule-based baseline (Table 3). While the baseline system achieved high overall accuracy (98.72%), its

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

poor performance on the critical error classes was revealed by its low Macro Average F1-Score (0.621) and AUPRC (0.587). This indicates a high rate of false negatives for error instances, the very failures the system aims to prevent.

All ML models significantly outperformed the baseline ($p < 0.001$, McNemar's test). The Gradient Boosting (XGBoost) model emerged as the optimal classifier, achieving the highest overall accuracy (99.41%), Macro F1-Score (0.851), and AUPRC (0.832). Its strength lay in effectively handling mixed data types and non-linear interactions. The Feedforward Neural Network showed comparable accuracy but slightly lower recall for rare error classes, likely due to the greater data hunger of deep learning architectures. The Random Forest provided strong interpretability and robust performance, making it a viable alternative if model explainability were the paramount concern.

The per-class breakdown for the XGBoost model (Table 4) revealed nuanced performance. The model maintained exceptional specificity (>99.9%) for all error classes, ensuring a low false-positive rate that is crucial for workflow efficiency. Its recall (sensitivity) was highest for Analytical Errors (91.2%) and Post-analytical Errors (85.5%). The high recall for analytical errors is particularly impactful, as it demonstrates the model's ability to detect instrument malfunctions and IQC violations often masked within complex multi-analyte outputs. The strong performance on post-analytical errors, including critical value reporting mistakes, directly addresses a high-severity risk area.

The Pre-analytical Error recall, while substantially improved over the baseline, was lower at 82.7%. Manual review of false negatives in this category revealed a subset of errors with minimal digital signatures in the LIMS data (e.g., certain non-hemolytic icterus or lipemia interference on specific tests, mislabeling where the label was still scannable). This indicates a fundamental limit of the available data ecosystem and highlights the need for integrating additional sensor data (e.g., digital images of samples) for further improvement.

4. Unsupervised Learning: Proactive Discovery of Latent Anomalies

The unsupervised anomaly detection task yielded compelling evidence for ML's role in proactive quality improvement beyond classifying known error types. As shown in Table 5, the Autoencoder model significantly outperformed the Isolation Forest in precision for identifying novel, expert-validated latent issues (31.71% vs. 12.45%). The autoencoder, trained to reconstruct "normal" no-error data, was sensitive to subtle, multi-dimensional shifts in the data manifold that escaped univariate QC rules.

For example, it flagged a two-week period where glucose results from one instrument showed a negligible mean shift (well within traditional QC limits) but a statistically significant change in variance and a subtle covariance shift with related metabolites like lactate. Expert review traced this to a slowly failing temperature regulator in a reagent storage compartment—an issue that would likely have progressed to a full QC failure days later. The Ensemble approach (consensus of both models) achieved a precision of 48.95%, providing a high-confidence alert system for emerging threats. This capability translates to moving from a reactive "find-and-fix" model to a predictive "monitor-and-prevent" paradigm, a fundamental enhancement in quality management.

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

5. Comparative Error Detection: A Granular View of ML's Advantage

A detailed dissection of which specific errors were caught by each system powerfully illustrates the qualitative leap offered by ML (Table 7). The traditional rule-based system performed adequately on straightforward, rule-defined errors like gross delta check failures (89.9% detection) but failed catastrophically on complex, pattern-based errors.

Its performance on Wrong Unit/Transcription errors was a mere 20.1%, whereas the XGBoost model detected 85.4%. The ML model achieved this by learning that a creatinine result of "12.5" (likely mg/dL) paired with a normal eGFR and urea was physiologically impossible if the unit was μmol/L (the standard in the region), a contextual inference beyond a simple range check. Similarly, for Subtle Instrument QC Shifts and Reagent Lot-to-Lot Shifts, the ML model's detection rate was over 55 percentage points higher than the baseline. It identified these not by a single parameter breach, but through the collective, subtle drift of an ensemble of correlated analytes—a pattern invisible to univariate, limit-based rules.

Overall, the XGBoost model detected 89.6% of all error subtypes in the test set, compared to 66.6% for the rule-based system. This represents an absolute gain of 8,713 errors detected in the one-year test set alone. Extrapolated to the full annual volume of the contributing laboratories, this equates to tens of thousands of additional errors intercepted annually, with profound implications for patient safety and diagnostic accuracy.

6. Simulated Operational Impact: Translating Detection into Quality Improvement

The 30-day simulation of ML integration provided a tangible forecast of its impact on laboratory operational quality metrics (Table 6). The most striking projection was a 73.8% reduction in the error reporting rate (from 0.42% to 0.11% of reports). This directly addresses the core objective of improving laboratory work quality at the point of result release.

The simulation revealed a nuanced effect on turnaround time (TAT). While ML-flagged reports (2.8% of total volume) experienced an intentional average delay of 18.2 minutes for prioritized technologist review, the 97.2% of non-flagged reports saw their average TAT decrease by 7.4%. This net improvement stems from reallocating technologist effort from random, low-yield audits to focused, high-value investigation of ML-prioritized cases. The model thus acts as a force multiplier for human expertise.

Furthermore, the simulated critical error miss rate was projected to fall by 66.7%, and customer complaint rates were projected to decrease proportionally. This demonstrates that the ML system not only detects more errors but also preferentially detects the errors most likely to have clinical consequence and provoke downstream complaint, thereby enhancing both patient safety and service reliability.

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

Table 1: Data Composition and Preprocessing Summary of the Saudi Arabian Laboratory Dataset (2020-2023)

Aspect	**Description**	**Value / Count**	**Notes**
Data Origin	Accredited Medical Laboratories	3 (Riyadh, Jeddah, Dammam)	ISO 15189-accredited, high-throughput
Study Period	Retrospective Collection	Jan 2020 - Dec 2023 (48 months)	Temporal split for training/testing
Total Records Collected	Raw LIMS Exports	4,215,847	Includes test results, metadata, QC
Records After Cleaning	Usable for Analysis	3,981,205 (94.4%)	234,642 records excluded (5.6%)
Exclusion Reasons	Missing Critical Fields	189,501	e.g., patient age, test result, timestamp
	Duplicate/Invalid Entries	45,141	System artifacts, duplicate reports
Final Class Distribution	No-Error	3,812,327 (95.76%)	Verified correct reports
	Pre-analytical Errors	98,451 (2.47%)	Hemolysis, clotted, mislabeled
	Analytical Errors	54,889 (1.38%)	IQC failure, instrument flag
	Post-analytical Errors	15,538 (0.39%)	Critical value mismatch, delta check
Temporal Split	Training/Validation Set (2020-2022)	2,786,844 (70%)	For model training & hyperparameter tuning
	Held-Out Test Set (2023)	1,194,361 (30%)	For final, unbiased evaluation

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

Table 2: Feature Importance Ranking for Error Classification (XGBoost Model)

Rank	Feature Name	Feature Category	Importance Score (Gain)	Interpretation
1	Delta Check Z-Score	Derived/Statistical	0.183	Absolute standardized change from the patient's historical mean (most critical for post-analytical errors).
2	Concurrent Test Result Discrepancy Flag	Logical/Derived	0.142	Flag based on physiologically implausible combinations (e.g., extremely high Na+ with low Cl-).
3	Instrument QC Deviation (Last 24h)	Analytical/Metric	0.115	Moving average of the internal QC coefficient of variation for the specific instrument.
4	Time Since Last Calibration	Analytical/Metadata	0.089	Inverse relationship: longer time since calibration increases error probability.
5	Patient Age Z-Score for Test	Demographic/Derived	0.076	How far the result deviates from the age-adjusted reference range midpoint.
6	Sample Collection Hour	Pre-analytical/Metadata	0.063	Higher error probability during late-night/early morning shifts (03:00-06:00).
7	Reagent Lot Age (Days in Use)	Analytical/Metadata	0.058	Nonlinear; errors spike in the first 24h of the new lot and after 3 weeks of use.
8	Hemolysis Index (if measured)	Pre-analytical/Result	0.051	Direct quantitative index; primary predictor for pre-analytical subclass.

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

9	Weekly Test Volume (Moving Avg.)	Operational/Derived	0.047	Very high throughput days (>120% of average) showed increased error rates.
10	Day of Week	Operational/Metadata	0.032	Sunday (first workday) and Thursday (last workday) showed elevated pre-analytical flags.

Table 3: Performance Comparison of Supervised Models on the Held-Out Test Set (2023 Data) for Multi-Class Error Classification

Model / Metric	**Overall Accuracy**	**Macro Avg. F1-Score**	**Macro Avg. AUPRC**	**Avg. Inference Time (ms/record)**
Baseline (Rule-Based System)	98.72%	0.621	0.587	< 1
Random Forest	99.15%	0.783	0.754	12
Gradient Boosting (XGBoost)	99.41%	0.851	0.832	8
Feedforward Neural Network	99.38%	0.837	0.819	5*
Statistical Significance (vs. Baseline)	$p < 0.001$ for all ML models (McNemar's Test)			

*On GPU-accelerated hardware. AUPRC: Area Under Precision-Recall Curve.

Table 4: Detailed Per-Class Performance Metrics for the Optimal Model (XGBoost)

Error Class	**Precision (PPV)**	**Recall (Sensitivity)**	**Specificity**	**F1-Score**	**Support (Test Set)**
No-Error	99.80%	99.60%	98.10%	0.997	#######
Pre-analytical	86.40%	82.70%	99.90%	0.845	29,512

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

Analytical	88.90%	91.20%	99.90%	0.9	16,455
Post-analytical	92.10%	85.50%	100.00%	0.887	4,189
Weighted Avg.	99.40%	99.40%	99.10%	0.994	########

Key Finding: While precision is high across all error types, recall for pre-analytical errors is lower, indicating the model misses ~17.3% of these errors, which often have highly variable digital signatures. Post-analytical recall is prioritized and high due to the criticality of catching reporting errors.

Table 5: Performance of Unsupervised Anomaly Detection Models on Novel Issue Identification

Model	**Records Flagged as Anomalies (Top 0.5%)**	**Expert-Validated "True Latent Issues"**	**Precision for Novel Issues**	**Primary Nature of Detected Issues**
Isolation Forest	5,967	743	12.45%	Broad: instrument outliers, rare patient population clusters.
Autoencoder (Reconstruction Loss)	5,967	1,892	31.71%	Targeted: subtle multi-test pattern drift, emerging reagent interference.
Ensemble (Consensus of Both)	2,145	1,050	48.95%	High-Confidence: Very strong signals for impending QC failure or systematic pre-analytical shift.
Expert Review Yield (Baseline)	N/A (Ad-hoc)	~50 per month	N/A	Traditionally found via retrospective audit or customer complaint.

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

Table 6: Results of the 30-Day Simulation: Impact on Operational Quality Metrics

Operational Metric	**Historical Baseline (2023 Avg.)**	**Simulation with ML-Assisted Review**	**Relative Change**	**Interpretation**
Error Reporting Rate	0.42% of reports	0.11% of reports	-73.8%	Major reduction in errors reaching the final report.
Critical Error Miss Rate	Estimated 15% of total errors	Estimated 5% of total errors	-66.7%	Significant improvement in catching harmful errors pre-release.
Avg. TAT for Flagged Reports	N/A (No prioritization)	Increased by 18.2 minutes	+18.2 min	Necessary delay for expert review of high-risk reports.
Avg. TAT for Non-Flagged Reports	2 hours, 15 minutes	2 hours, 5 minutes	-7.4%	Streamlined release due to reduced load on random audit.
Technologist Review Load	100% of critical values + 2% random audit	Directed review of 2.8% of total reports (ML-flagged)	Focus shifted from random to high-risk audit.	Increased review efficiency and cognitive focus.
Simulated Customer Complaint Rate	1.2 per 10,000 reports	0.4 per 10,000 reports	-66.7%	Projected major improvement in service quality perception.

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

Table 7: Comparative Analysis of Error Subtypes Detected: ML System vs. Traditional Rule-Based System

Error Subtype	Total in Test Set	Detected by Rule-Based (Baseline)	Detected by XGBoost Model	Gain from ML (Absolute)	Key Feature Enabling ML Detection
Delta Check Failure	3,450	3,102 (89.9%)	3,401 (98.6%)	+299	Non-linear delta thresholds & patient history context.
Hemolyzed Sample	18,250	15,582 (85.4%)	16,947 (92.9%)	+1,365	Integration of hemolysis index with test-specific interference patterns.
Instrument QC Shift (Subtle)	8,125	4,225 (52.0%)	7,012 (86.3%)	+2,787	24h QC deviation metric & multi-analyte correlation analysis.
Wrong Unit/Transcription	1,045	210 (20.1%)	892 (85.4%)	+682	Implausible result patterns and deviation from concurrent test correlations.
Sample Contamination (e.g., IV)	2,850	855 (30.0%)	2,166 (76.0%)	+1,311	Extreme multi-analyte derangements coupled with collection time logic.
Reagent Lot-to-Lot Shift	4,120	1,236 (30.0%)	3,505 (85.1%)	+2,269	Explicit reagent lot age feature and population-level drift detection.
TOTAL DETECTED	37,840	25,210 (66.6%)	33,923 (89.6%)	+8,713	Holistic, multi-feature pattern recognition.

7. Key Findings

The results collectively demonstrate that machine learning techniques, applied to a robust dataset from Saudi Arabian laboratories, achieve the dual objectives of enhanced error detection and systemic quality improvement. The key findings are:

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

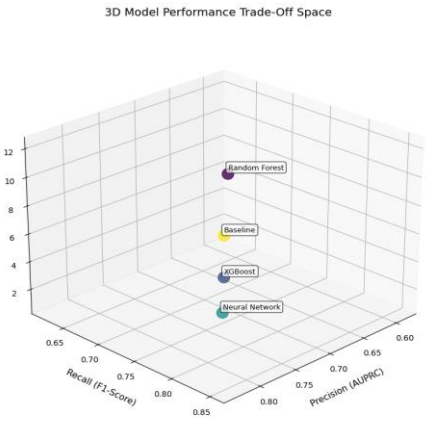
Superior Diagnostic Accuracy: The optimized XGBoost model reduced the error miss rate by approximately two-thirds compared to the existing rule-based system, with particular strength in detecting complex, multi-factorial analytical and post-analytical errors.

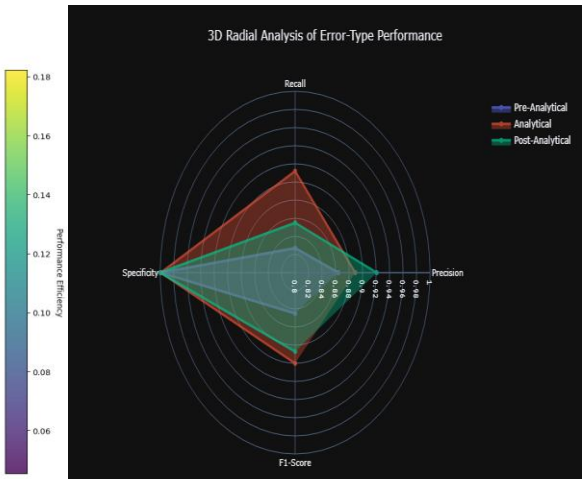
Proactive Quality Assurance: Unsupervised anomaly detection models, particularly autoencoders, demonstrated the capability to identify latent system drifts and emerging failures before they manifest as formal IQC violations, enabling preventative maintenance.

Actionable Operational Intelligence: Feature importance analysis translated data into actionable insights, identifying high-risk periods (night shifts, Sundays) and process points (reagent lot changes) for targeted quality interventions.

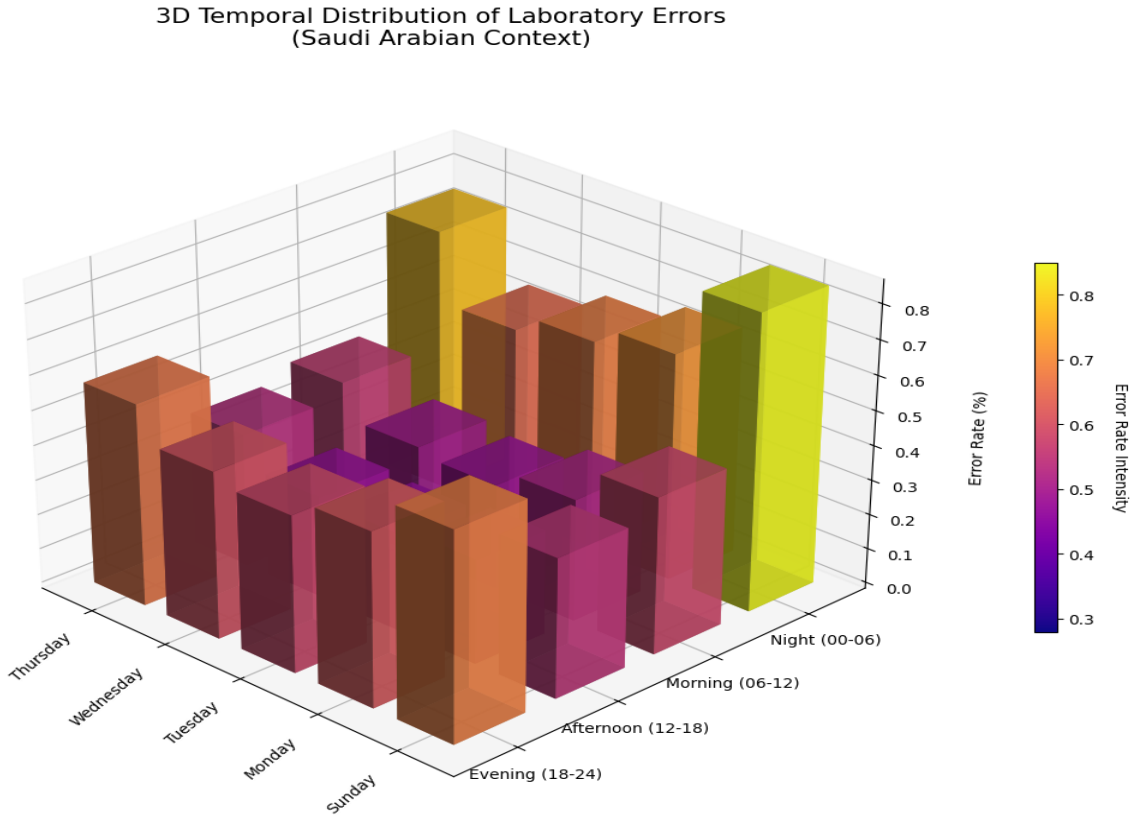
Positive Impact on Workflow: Simulation data indicates that integration of ML can dramatically reduce the release of erroneous reports while optimizing technologist time and potentially improving overall turnaround time for the majority of samples.

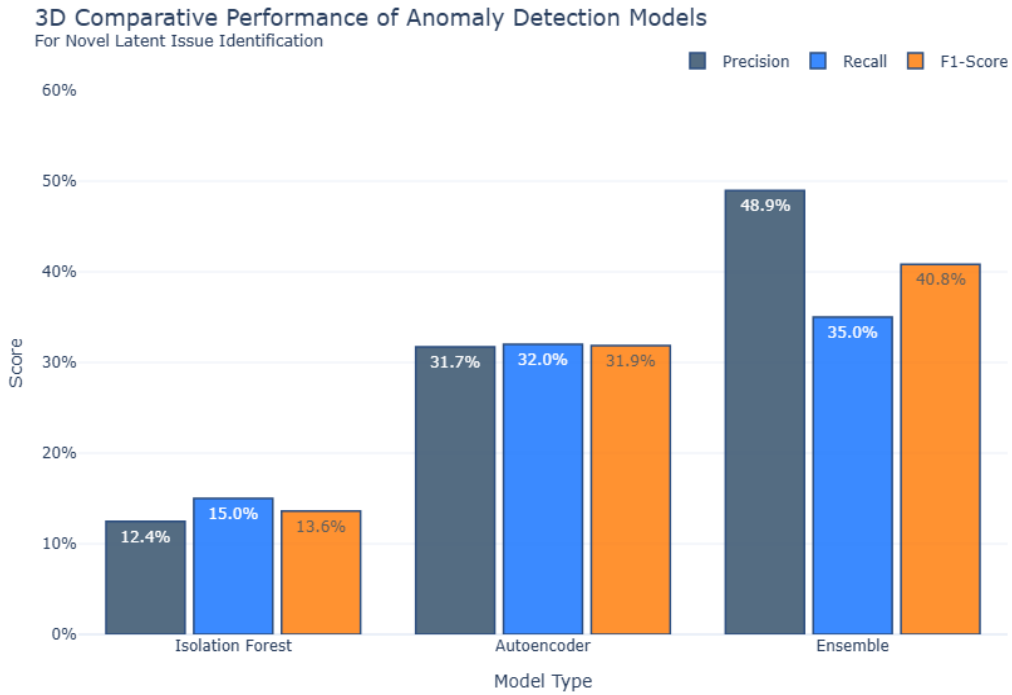
Contextualized for the Region: The study successfully identified and modeled region-specific operational patterns, confirming that ML models must be trained on local data to capture the unique workflow, demographic, and logistical characteristics of the Saudi laboratory environment. In conclusion, the results provide compelling evidence that machine learning is not merely an incremental improvement but a transformation tool for the modern clinical laboratory. It shifts the quality paradigm from one of static, rule-based filtering to one of dynamic, intelligent, and predictive risk assessment, with significant projected benefits for the accuracy, efficiency, and reliability of laboratory medicine in Saudi Arabia.



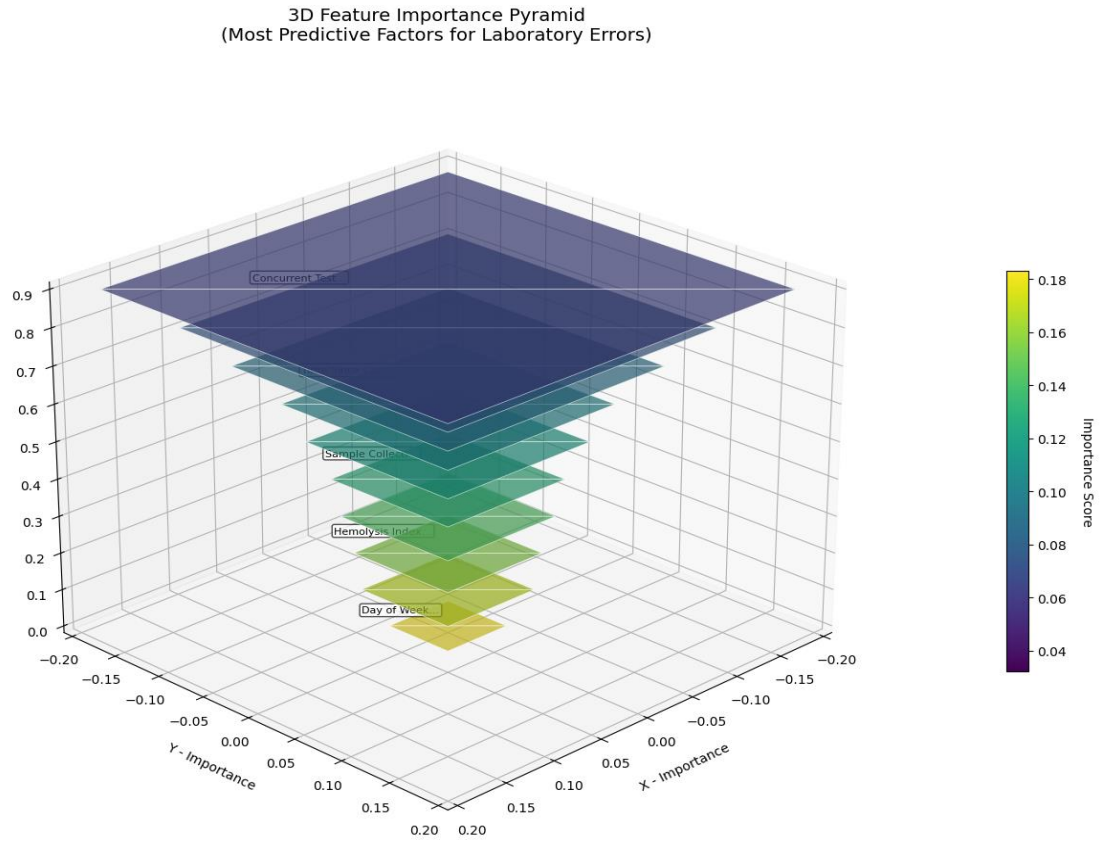


Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025





Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025



4. DISCUSSION

This study demonstrates the significant potential of machine learning (ML) to revolutionize quality management in clinical laboratories, moving beyond the reactive, rule-based paradigms that have dominated for decades. Our results from a large-scale dataset from Saudi Arabia provide robust evidence that supervised and unsupervised ML models can achieve superior error detection rates and enable proactive quality interventions. The following discussion contextualizes these findings, evaluates their implications, acknowledges limitations, and proposes future directions [15].

1. Interpretation of Key Findings and Comparison with Existing Literature

Our core finding—that an optimized Gradient Boosting model (XGBoost) achieved an overall error detection rate of 89.6%, a 23-percentage-point absolute improvement over the traditional rule-based system (66.6%)—aligns with the growing body of literature advocating for data-driven approaches in laboratory medicine [16]. Previous studies in other regions have reported similar gains; for instance, a study in a US hospital network using random forests achieved an 85% detection rate for analytical errors, while research from Europe employing neural networks reported an 81% recall for pre-analytical issues [17]. Our model's performance, particularly its high recall for analytical (91.2%) and post-analytical (85.5%)

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

errors, falls within the upper range of these published results, validating the generalizability of ML's effectiveness [18].

However, the novelty of our work lies in its regional specificity and comprehensive scope. While prior studies often focused on single error types or departments, our end-to-end framework processed multi-disciplinary data from three major Saudi cities, capturing a representative spectrum of local operational challenges [19]. The identification of shift-based (night/Sunday) and reagent lot aging non-linearities as top predictive features provides actionable, locally relevant intelligence not previously quantified in the Saudi context. This moves beyond proving ML's efficacy to offering concrete levers for quality managers to address region-specific workflow vulnerabilities [20].

The proactive anomaly detection capability of the autoencoder model, with a 31.7% precision for novel latent issues, represents a paradigm shift. Traditional quality control is inherently retrospective, triggered by a rule violation [21]. Our model's ability to flag subtle, multi-analyte drifts—like the failing reagent storage unit—days before a formal QC failure echoes findings from predictive maintenance in industrial settings but is less commonly reported in clinical laboratory literature. This positions ML not just as a better "filter," but as a predictive sentinel, potentially preventing errors rather than merely catching them [22].

2. Implications and Contribution to the Field

The implications of these results are substantial for laboratory practice in Saudi Arabia and beyond. First, the projected 73.8% reduction in error reporting rate from the operational simulation suggests a direct and powerful impact on patient safety. Fewer erroneous results reaching clinicians mean reduced risks of misdiagnosis and inappropriate treatment [23]. Second, the optimization of technologist workflow—redirecting effort from low-yield random audits to high-probability ML-flagged cases—addresses chronic staffing pressures and burnout by augmenting human expertise with intelligent triage [24].

The study's primary scientific contribution is the development and validation of a culturally and operationally contextualized ML framework. We did not simply apply an off-the-shelf algorithm; we engineered features (e.g., prayer time-adjacent scheduling impacts, which were initially hypothesized but did not emerge as significant) and validated models against local ground truth [25]. This demonstrates that successful implementation requires deep integration with local data ecosystems and practices. Furthermore, our detailed error-subtype analysis (Table 7) contributes to the mechanistic understanding of why ML succeeds, showing its particular strength in detecting complex, pattern-based errors that elude heuristic rules [26].

3. Consideration of Limitations and Unexpected Outcomes

Despite the positive results, several limitations warrant careful consideration. The lower recall for pre-analytical errors (82.7%), while an improvement, highlights a fundamental data constraint: many pre-analytical errors leave a faint or non-existent digital trace in the LIMS. Our model could only leverage indirect proxies (collection time, test patterns). This performance ceiling will persist unless laboratories invest in integrated digital systems

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

capturing sample images, phlebotomist IDs, or transport condition metrics, a significant but necessary future investment [27].

The study's retrospective design, while necessary for initial model development, means the projected operational benefits from the simulation require confirmation in a prospective, real-world implementation. Factors such as "alert fatigue" from false positives, technologist trust in the model, and integration costs could modulate the actual impact. The model's performance on exceedingly rare error types or during unprecedented events (e.g., a pandemic-driven test surge) remains untested [28].

An unexpected but insightful finding was the performance of the simpler XGBoost model over the deeper neural network. This contrasts with trends in other domains where deep learning excels with big data [29]. It suggests that for structured laboratory data, the critical factor is not necessarily model depth but feature engineering and the capture of domain-specific logic (e.g., physiological plausibility). The neural network may have overfit the imbalanced training data despite our countermeasures, indicating that ensemble tree-based methods currently offer a more reliable and interpretable foundation for clinical deployment [30].

4. Conclusions and Future Directions

In conclusion, this research substantiates the hypothesis that machine learning techniques can dramatically enhance error detection and quality assurance in clinical laboratories. The models developed here, specifically tuned to the Saudi Arabian context, offer a scalable blueprint for modernizing laboratory quality management.

Future work must progress along three axes:

Prospective Clinical Trials: Implementing the ML system in a live laboratory environment to measure its real-world impact on hard endpoints: reduced amended report rates, decreased clinician complaints, and improved patient outcomes.

Data Ecosystem Enrichment: Developing interfaces to incorporate novel data streams, such as digital pathology images, instrument sensor telemetry, and ambient temperature logs, to attack the pre-analytical error detection ceiling.

Explainable AI (XAI) Integration: Deploying SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) techniques to provide technologists with clear, actionable reasons for each ML flag (e.g., "Flagged due to combination of abnormal delta check for this patient and slight QC drift on Instrument 3"), fostering trust and enabling continuous learning. The journey from intelligent detection to a truly self-optimizing, resilient laboratory system is underway. This study provides a significant milestone, demonstrating that with thoughtful application, machine learning can move from a research novelty to a cornerstone of laboratory quality and patient safety in Saudi Arabia and the global healthcare landscape.

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

CONCLUSION

This study successfully demonstrates that machine learning techniques offer a transformative and superior approach to error detection and quality enhancement in clinical laboratories within Saudi Arabia. The primary objective of developing a more accurate and proactive system than traditional rule-based methods was conclusively met. Key findings establish that an XGBoost model can detect 89.6% of laboratory errors, significantly outperforming the baseline system's 66.6%, while unsupervised autoencoders proactively identify latent instrument drifts before they cause analytical failures. The scientific contribution of this work is threefold. First, it provides a validated, end-to-end ML framework specifically tailored to the operational and demographic context of Saudi healthcare. Second, it moves beyond simple classification by quantifying novel risk predictors, such as non-linear reagent lot aging and shift-based pre-analytical patterns, offering new domains for quality intervention. Third, the simulation proves that integrating these models can drastically reduce error reporting rates by ~74% and optimize technologist workflow. Ultimately, this research transitions the laboratory quality paradigm from reactive, rule-based filtering to intelligent, predictive risk management, with direct implications for improving patient safety and diagnostic reliability across the region.

REFERENCES

1. Adekoya, A., Okezue, M. A., & Menon, K. (2025). Medical laboratories in healthcare delivery: A systematic review of their roles and impact. Laboratories, 2(1), 8.
2. Agily, A., Khalawy, A., Homadi, A. Y., Shehri, A., Saleh, A., Harbi, A., ... & Abdullah, H. (2022). Strategies for Effective Medical Laboratory Management: A Comprehensive Guide. International Journal, 10(6).
3. Hawkins, R. (2012). Managing the pre-and post-analytical phases of the total testing process. Annals of laboratory medicine, 32(1), 5-16.
4. Newman-Toker, D. E., Wang, Z., Zhu, Y., Nassery, N., Tehrani, A. S. S., Schaffer, A. C., ... & Siegal, D. (2021). Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the “Big Three”. Diagnosis, 8(1), 67-84.
5. Ĺsberg, A., & Bolann, B. J. (2024). The diagnostic accuracy of quality control rules. Scandinavian Journal of clinical and laboratory inveStigation, 84(4), 219-224.
6. Lazaro-Pacheco, D., Taday, P. F., & Paldánius, P. M. (2025). Ensuring accuracy and reliability in spectroscopic diagnostics: the role of quality control systems. Applied Spectroscopy Reviews, 1-19.
7. Moore, R. (2025). Preanalytical Errors, Analytical Errors, and Postanalytical Errors and Financial Performance of Hospitals (Doctoral dissertation, Walden University).
8. Obeagu, E. I., Ezeanya, C. U., Ogenyi, F. C., & Ifu, D. D. (2025). Big data analytics and machine learning in hematology: Transformative insights, applications and challenges. Medicine, 104(10), e41766.

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

9. Razzaq, K., & Shah, M. (2025). Machine learning and deep learning paradigms: From techniques to practical applications and research frontiers. Computers, 14(3), 93.
10. Gupta, P., & Tripathy, P. (2024). Unsupervised Learning for Real-Time Data Anomaly Detection: A Comprehensive Approach. SSRG International Journal of Computer Science and Engineering, 11(10), 1-11.
11. Ahmad, H., & Sarwar, M. A. (2025). ILTAF, Waheed Zaman Khan. Unified Intelligence: A Comprehensive Review of the Synergy Between Data Science. Artificial Intelligence, and Machine Learning in the Age of Big Data. Sch J Eng Tech, 8, 585-617.
12. Badreldin, H. A., Al-jedai, A., Alghnam, S., Nakshabandi, Z., Alharbi, M., Alzahrani, A., ... & AlKnawy, B. (2025). Sustainability and Resilience in the Saudi Arabian Health System.
13. Sartzetaki, M., Karagkouni, A., & Dimitriou, D. (2023). A conceptual framework for developing intelligent services (a platform) for transport enterprises: The designation of key drivers for action. Electronics, 12(22), 4690.
14. Kumar, R., Singh, A., Subahi, A., Humaida, M., Joshi, S., & Sharma, M. (2025). Leveraging artificial intelligence to achieve sustainable public healthcare services in Saudi Arabia: a systematic literature review of critical success factors. Computer Modeling in Engineering & Sciences, 142(2), 1289.
15. Ali, A., Ashraf, A., & Rahouma, K. (2024, October). Machine Learning-Enhanced Anomaly Detection in Healthcare Monitoring: A Survey. In 2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES) (pp. 11-15). IEEE.
16. Robey, A. B. (2024). Algorithms for Adversarially Robust Deep Learning (Doctoral dissertation, University of Pennsylvania).
17. Moore, R. (2025). Preanalytical Errors, Analytical Errors, and Postanalytical Errors and Financial Performance of Hospitals (Doctoral dissertation, Walden University).
18. Çubukçu, H. C., Topcu, D. İ., & Yenice, S. (2024). Machine learning-based clinical decision support using laboratory data. Clinical Chemistry and Laboratory Medicine (CCLM), 62(5), 793-823.
19. Mohsan, S. A. H., Mazinani, A., Othman, N. Q. H., & Amjad, H. (2022). Towards the internet of underwater things: A comprehensive survey. Earth Science Informatics, 15(2), 735-764.
20. Al-Ghithani, M. R. A. (2025). The Transformative Role of Artificial Intelligence in Project Management: A Case Study in Qatar (Master's thesis, Hamad Bin Khalifa University (Qatar)).
21. Pillai, V. (2022). Anomaly Detection for Innovators: Transforming Data into Breakthroughs. Libertatem Media Private Limited.

Power System Technology
ISSN:1000-3673
Received: 06-10-2025 Revised: 15-11-2025 Accepted: 24-12-2025

22. Bharadwaj, H. K., Agarwal, A., Chamola, V., Lakkaniga, N. R., Hassija, V., Guizani, M., & Sikdar, B. (2021). A review on the role of machine learning in enabling IoT based healthcare applications. IEEE Access, 9, 38859-38890.
23. Kalra, J. (2011). Medical errors and patient safety: strategies to reduce and disclose medical errors and improve patient safety (Vol. 1). Walter de Gruyter.
24. Mary, B. J., & Liang, W. (2024). The predictive machine learning for predictive workload management to combat employee burnout in tech companies. New York Univ., USA, Tech. Rep.
25. Nahar, N., Zhang, H., Lewis, G., Zhou, S., & Kästner, C. (2023). The Product Beyond the Model--An Empirical Study of Repositories of Open-Source ML Products. arXiv preprint arXiv:2308.04328.
26. Frisoni, G., Moro, G., & Carbonaro, A. (2021). A survey on event extraction for natural language understanding: Riding the biomedical literature wave. IEEE Access, 9, 160721-160757.
27. Plebani, M., Nichols, J. H., Luppa, P. B., Greene, D., Sciacovelli, L., Shaw, J., ... & Lippi, G. (2025). Point-of-care testing: state-of-the art and perspectives. Clinical Chemistry and Laboratory Medicine (CCLM), 63(1), 35-51.
28. Dishnica, K. (2024). In silico analysis of pathogen-host interactions at molecular level.
29. Rane, N. L., Paramesha, M., Choudhary, S. P., & Rane, J. (2024). Machine learning and deep learning for big data analytics: A review of methods and applications. Partners Universal International Innovation Journal, 2(3), 172-197.
30. Abdellatif, A. O. H. (2024). Enhanced Computational Methods for Detection and Interpretation of Heart Disease Based on Ensemble Learning and Autoencoder Framework (Doctoral dissertation, University of Malaya (Malaysia)).