Received: 7 May 2026 | Revised: 13 July 2026 | Accepted: 12 August 2026 | Published online: 2 September 2026

RESEARCH ARTICLE

Medinformatics
2026, Vol. 00(00) 1–13
DOI: 10.47852/bonviewMEDIN620210308

BON VIEW PUBLISHING

An Adaptive Dual-Loop Artificial Intelligence Framework for Integrated Disease Surveillance and Health Workforce Learning in Low-Resource Health Systems

Kenneth Goga Riany[1,*], Agnes Linus Muthoni[2], Marcellah Eucabeth Onsomu[3], Pauldy C. J. Otermans[4] and Dev Aditya[4]

[1] *Centre of Open and Distance Education (CODE), KCA University, Kenya*

[2] *School of Health Sciences, Meru University of Science and Technology, Kenya*

[3] *School of Communication, Daystar University, Kenya*

[4] *Otermans Institute, United Kingdom*

Abstract: Background: In many low-resource health systems, disease surveillance and workforce training run on separate rails, so outbreaks are caught late, and frontline capacity stays thin. Most artificial intelligence (AI) tools pick one side or the other; almost none couple them. Methods: We built and tested an Adaptive Dual-Loop AI Framework. One loop, for epidemiological intelligence, handles multi-source ingestion, anomaly detection, and time-series forecasting; the other, for workforce learning, runs a conversational AI tutor, contextual decision support, and interaction logging. Bidirectional feedback links them. We evaluated the system with a mixed-methods, implementation-science design in 24 facilities and 312 health workers across Nairobi (urban) and Kajiado (rural) counties, Kenya, over 10 months, benchmarking the forecaster against Autoregressive Integrated Moving Average (ARIMA) and rule-based baselines with rolling-window cross-validation on Integrated Disease Surveillance and Response and District Health Information Software 2 (DHIS2) data. Results: The dual-loop model reached an area under the curve of 0.94 (95% CI 0.91–0.96), against 0.81 for ARIMA and 0.72 for threshold rules. Mean lead-time gain was 3.6 days across cholera, malaria, and respiratory infections, and the outbreak-response cycle shortened from 12.9 to 6.2 days (52%). Competency climbed from 55.0% to 75.4% ($p < 0.001$), engagement tracked that gain ($r = 0.76$), and reporting completeness rose from 62% to 92%. Conclusion: Coupling prediction with embedded learning delivered gains neither half reached alone, a practical route to adaptive public-health intelligence where resources are scarce.

Keywords: artificial intelligence, bidirectional feedback, disease surveillance, health workforce learning, low-resource settings

1. Introduction

Threats tend to be spotted late in low-resource health systems and answered later still. The reason is not lack of investment. It is that surveillance, decision support, and workforce training sit on separate platforms that rarely exchange information. So outbreaks build quietly, responses lag, and the frontline staff who most need current information are usually the last to get it [1, 2].

On paper, the trend looks reassuring. Since the pandemic, more than 75% of countries say they have strengthened parts of their surveillance systems, yet only about 54% actually meet the core International Health Regulations capacities for real-time surveillance and response [1]. Detection has sped up across sub-Saharan Africa over the past decade. What has hardly moved is the lag between seeing a signal and acting on it [2, 3].

The burden falls unevenly. Low- and middle-income countries (LMICs) absorb most of the world's public-health emergencies, and Kenya is a familiar case, with cholera, malaria, and COVID-19 recurring against both endemic and emerging threats [4]. About 60% of known infectious diseases, and up to three-quarters of emerging ones, are zoonotic, so surveillance that joins human, animal, and environmental signals matters more each year [5]. Digitization has spread widely, but the data architectures remain fragmented, and a great deal of what is collected is never used, leaving a wide gap between what systems capture and what actually informs a decision [3]. A study across 15 Kenyan counties made the point plainly: District Health Information Software 2 (DHIS2) data are gathered routinely but used unevenly, mostly because analytical capacity is thin and trained records officers are scarce [6].

***Corresponding author:** Kenneth Goga Riany, Centre of Open and Distance Education (CODE), KCA University, Kenya. Email: kriany@kcau.ac.ke

AI now targets these gaps directly. For outbreak prediction, anomaly detection, and risk stratification, machine learning models already beat conventional statistics [7], and reviews from sub-Saharan Africa report the same edge, tempered by stubborn data and infrastructure limits [8]. The workforce problem runs the other way. Frenk et al. [9] put the global shortfall near 10 million health workers by 2030, concentrated in LMICs, and staff who are already stretched have little room for training or new tools; whether they take one up turns on how useful it seems, what peers do, and what local conditions allow [10]. Large language models (LLMs) have opened new ground in medical education and decision support, with promising early results in low-resource frontline care [11]. Even so, most of these tools stay disconnected from real-time, population-level data [12].

Both fields have moved fast, but along separate tracks. Surveillance platforms aggregate data, flag anomalies, and forecast outbreaks. Training platforms teach. The two seldom meet. And early detection is only half the task: someone still has to read the signal and act, and a worker schooled on last year's epidemiology will not read it well. That is where the losses compound, with thin data weakening the models and thin capacity leaving the models' output on the shelf. Few designs address both ends at once, so surveillance and learning keep running side by side instead of feeding one another.

Problems this entangled need an adaptive systems lens, one where surveillance and workforce learning keep informing each other instead of standing apart. Ignore the link, and the loop turns vicious: weak data drag the models down, and weak capacity wastes whatever the models produce.

1.1. Conceptual gap

AI has become routine in global public health, most visibly in surveillance, outbreak prediction, and systems optimization. For a decade now, researchers have forecast epidemiological trends from clinical records, environmental signals, mobility traces, and digital feeds. Recurrent neural networks, Long Short-Term Memory (LSTM) architectures, and hybrid ensembles tend to beat classical statistics, especially on nonlinear, time-dependent behavior [7, 13]. The snag is practical: these methods want clean, continuous data, and low-resource systems rarely supply it, which caps how far they scale.

In LMICs, the real limits are infrastructural and institutional, not algorithmic. Reporting systems run in parallel. Data entry is manual and slow. Records are half-digitized, and platforms like DHIS2 barely interoperate. Even where digital reporting exists, staff shortages, thin training, and competing clinical demands make it patchy, so models end up learning from gappy inputs that blunt accuracy and delay alerts [3, 14].

Workforce development has taken its own artificial intelligence (AI) path, through intelligent tutors, conversational agents, and simulation-based learning that pair language processing with adaptive algorithms to tailor training, mimic clinical reasoning, or coach in real time [9, 12]. LLMs have stretched this further into conversational, context-aware exchange [15], though careful evaluations warn that their clinical accuracy still swings with task and setting [16]. What most of these tools share is isolation from live epidemiology: their content is fixed, or updated on a schedule, rather than following the diseases actually circulating around the user.

One limitation runs deeper than the rest. Surveillance and learning systems do not answer each other. Surveillance platforms aggregate, visualize, and predict; training platforms hand down knowledge and skills. Stacked as separate layers, each quietly depends on the other, yet a gain on one side seldom registers as a gain overall. Table 1 sets out how earlier approaches have handled or sidestepped this integration problem.

Underneath sits a linear picture of public-health intelligence, not an adaptive one. Information travels one way, from collection

Table 1. Comparative synthesis of representative digital health and AI approaches against the integration criteria addressed by this study

System/approach	Surveillance function	Workforce-learning function	Bidirectional feedback	Tested in LMIC	Gap addressed here
DHIS2	Aggregation, reporting	None	No	Yes	Adds adaptive intelligence + learning layer
EWARS	Threshold-based alerts	None	No	Yes	Adds ML and bidirectional feedback
ProMED-mail	Event-based detection	None	No	Partial	Adds structured workforce coupling
Classical Autoregressive Integrated Moving Average (ARIMA)/threshold	Statistical forecasting	None	No	Yes	Adds nonlinear modeling + tutor
Intelligent Tutoring Systems	None	Adaptive learning	No	Limited	Adds surveillance integration
LLM clinical assistants	None	Conversational decision support	No	Limited	Adds population-level surveillance
Institute of Medicine (IOM) Learning Health System Learning Health System	Conceptual feedback	Conceptual learning	Conceptual only	Limited	Operationalizes for LMIC scale
Adaptive Dual-Loop (proposed)	ML + anomaly + forecast	AI tutor + decision support	Yes (operational)	Yes (Kenya)	n/a

to analysis to decision, with almost no feedback that would let the system learn from what it just did. Adaptive systems work differently: the output of one part loops back to reshape the others [17].

Work on human-centered and augmented intelligence does stress collaboration between clinicians and algorithms. But it mostly concerns bedside tasks such as diagnostic support, not population-level system design, and it assumes stable data and well-trained users, two things low-resource settings cannot promise. The upshot is a scattered ecosystem where surveillance, training, and decision support coexist but rarely connect.

So a real conceptual gap remains. Surveillance AI sharpens prediction; learning AI builds capability; neither is built as a single architecture that keeps co-evolving, where surveillance reshapes what the workforce learns while the workforce, in turn, improves surveillance. Without that two-way feedback, digital health systems can only learn so much at the system level. Bridging it means treating human expertise and machine learning (ML) not as separate tools but as parts of one system that reinforce each other.

1.2. Study contribution and objectives

This study addresses that gap. The Adaptive Dual-Loop AI Framework pulls disease surveillance and workforce learning into one co-evolving system, its two loops tied together by bidirectional feedback. The coupling is the whole idea: frontline interactions feed new data into the models, and the improved output then reshapes how practitioners work and learn. We built it and ran it in two very different Kenyan counties, which let us test its effect on outbreak detection, response time, data quality, and workforce performance against real conditions.

The specific objectives were to (1) design a dual-loop architecture integrating surveillance and workforce learning; (2) operationalize it within Kenyan national health information systems; (3) evaluate algorithmic, system-level, and workforce outcomes against predefined benchmarks; and (4) examine the bidirectional feedback effect on data quality.

1.2.1. Novelty and contributions

Four things set this work apart. The first is architecture: platforms like DHIS2 and Early Warning, Alert and Response System (EWARS), and the learning tools beside them, run as one-way pipelines, whereas our design closes the loop, feeding frontline interaction logs back into retraining so human behavior becomes an input to learning rather than an endpoint. The second is method: instead of tuning a forecaster in isolation, we bolt an LSTM–gradient-boosted ensemble to a workforce-learning loop and then measure the coupling itself. The third is evidence: much of the learning-health-system literature stops at concept, while we report a prospective, mixed-methods, two-site evaluation, with interrupted-time-series and ablation results that isolate each loop (Section 3.4.1). The fourth is fit: the system is engineered for scarcity, running offline-first, tolerating gaps in the data, and augmenting national DHIS2/Integrated Disease Surveillance and Response (IDSR) infrastructure rather than displacing it. We know of no earlier system that demonstrates a bidirectionally coupled, empirically validated surveillance-plus-learning loop in an LMIC deployment.

1.3. The Adaptive Dual-Loop AI Framework

At heart, the Adaptive Dual-Loop AI Framework brings epidemiological intelligence and workforce capability into one system that keeps learning. The premise is plain: in low-resource settings, what limits performance is often how data systems and human judgment interact, not the data or the algorithms on their own. So the framework treats public-health intelligence as an adaptive system with two interdependent loops, an Epidemiological Intelligence Loop and a Health Workforce Learning Loop. Figure 1 gives the overall architecture.

The Epidemiological Intelligence Loop does the sensing. It pulls in and reads data from many sources at once—electronic records, routine DHIS2 reports, laboratory confirmations, environmental indicators, and frontline clinical notes—whether structured or not. Everything moves through a layered pipeline: preprocessing, feature extraction, anomaly detection, then forecasting, where the LSTM and gradient-boosted ensemble catch unusual patterns and turn them into structured alerts, risk maps, and decision-support prompts.

Its output feeds straight into the Health Workforce Learning Loop, which turns prediction into capability through adaptive learning. Here, the conversational tutor, scenario simulations, and context-aware decision support all react to current conditions, so what a worker is taught shifts with live surveillance rather than waiting for the next scheduled refresh.

This loop is not just a delivery channel. It watches what users actually do—the questions they ask, the calls and diagnoses they make, and the reports they file— and folds those traces back through ingestion as fresh structured data. Human behavior becomes a learning signal. The system teaches its users and learns from them at once, adapting to local epidemiology and habits it could never have been told in advance.

Feedback moves both ways. In one direction, surveillance decides what the workforce learns and when, a cholera cluster, say, triggering targeted modules on case identification, treatment, and reporting. In the other, tighter reporting and steadier diagnosis lift the quality of incoming data, which sharpens the models. The whole thing compounds, machine and human capability climbing together.

The build is a closed loop in four layers. A data-acquisition layer gathers inputs from facilities, laboratories, environmental sensors, and community channels. A machine-intelligence layer cleans, transforms, models, and predicts, drawing on supervised, unsupervised, and reinforcement learning. A human-interaction layer carries training, decision support, and guidance through chat and mobile. And a feedback-integration layer shuts the circuit, catching user behavior, system performance, and field notes and routing them back into both the data and the models.

Most digital health systems still push information in a straight line. This one bends it into a cycle. Each pass tightens prediction, keeps training current, and improves responsiveness, which counts most where disease patterns shift, data quality wobbles, and capacity swings from one site to the next. It also stretches the usual human-in-the-loop idea. Where those designs stop at a single task, here workers help generate both the data and the intelligence, so their actions steer how the system evolves.

Retraining keeps the design honest. As new data arrive from the field and from workforce activity, the models refresh to match what is happening on the ground, which holds off the drift that quietly degrades static predictors when conditions change. And it is meant to ride on top of what already exists, sitting alongside DHIS2 and electronic medical records (EMRs) as an augmentation layer, not a replacement.

2. Materials and Methods

2.1. Study design and reporting standards

We chose a mixed-methods, implementation-science design because the intervention was three things at once—a piece of

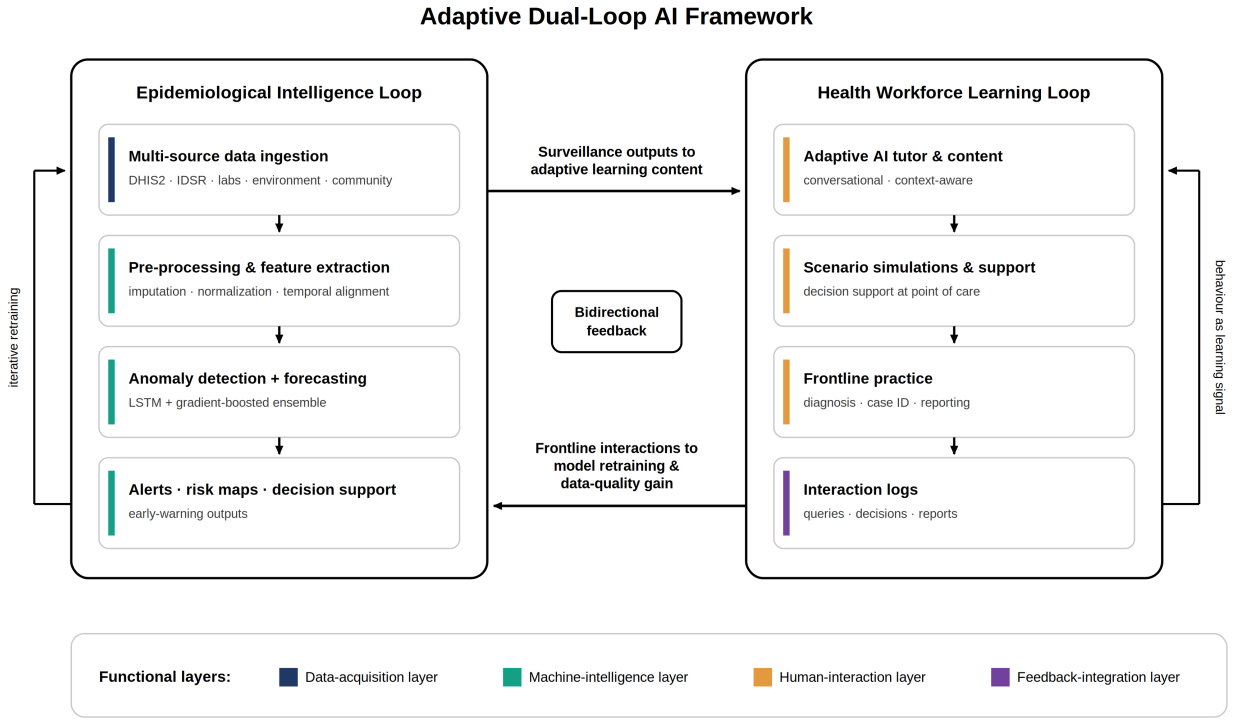

Figure 1. Conceptual architecture of the Adaptive Dual-Loop AI Framework. The Epidemiological Intelligence Loop and the Health Workforce Learning Loop operate as interdependent, bidirectionally coupled subsystems: frontline interactions generate data that refine predictive models, while predictive outputs shape adaptive learning and decision support, producing a self-reinforcing public-health intelligence system.

technology, a health-system strengthening effort, and a way to build human capacity—and a purely quantitative read would have missed how those pull on each other. Reporting followed TRIPOD-AI for the predictive-modeling work [18] and the Standards for Reporting Qualitative Research (SRQR) and Consolidated Criteria for Reporting Qualitative Research (COREQ) standards for the qualitative work, with completed checklists provided as **Supplementary File S1**.

2.2. Setting and ethics

Fieldwork spanned two counties chosen for contrast, urban Nairobi and rural Kajiado, 12 facilities each, from sub-county hospitals down to dispensaries. Altogether 312 health workers took part (168 in Nairobi, 144 in Kajiado): clinical officers, nurses, laboratory technologists, and community health volunteers. The institutional Ethics and Research Committee approved the study (Approval No. KNH/UoN-ERC/A/000-2024; February 14, 2024), which followed the Kenya Data Protection Act (2019) and WHO guidance for AI in health [19]. Everyone gave written informed consent. Running two sites (Figure 2) let us probe external validity and scalability.

2.3. Data sources and preprocessing

Several streams fed the Epidemiological Intelligence Loop. Routine case data came from the Kenya Health Information System (DHIS2) and the IDSR system, alongside laboratory confirmations, environmental indicators (rainfall, temperature, water access), and community-based reports. A standard pipeline of imputation, normalization, and temporal alignment prepared them, as laid out in Figure 3 and Table 2.

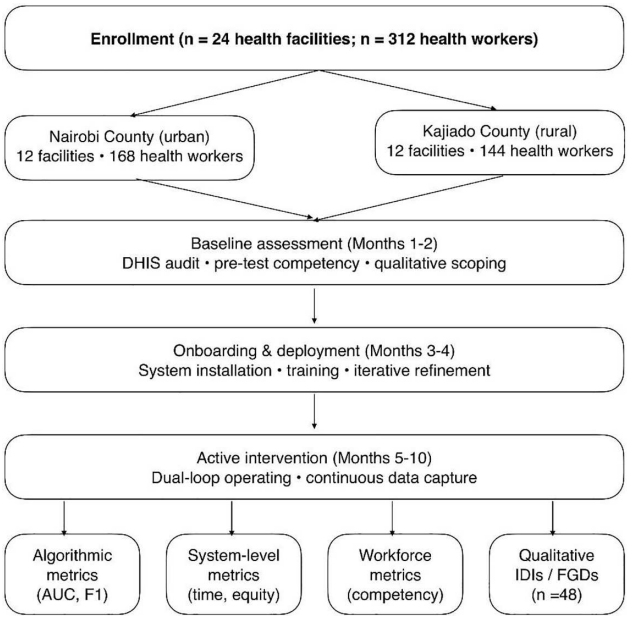

Figure 2. Mixed-methods evaluation design across two contrasting Kenyan counties, tracing participant flow, intervention phases, and parallel quantitative and qualitative evaluation streams

2.4. Machine learning pipeline and benchmarks

For training and validation, we used rolling-window cross-validation, which keeps time in order and avoids the leakage that random splits cause in time-series work. Three approaches went head-to-head: the dual-loop ensemble, pairing LSTM forecasting

with a gradient-boosted anomaly detector; classical ARIMA; and the rule-based threshold detector that IDSR practice uses today. We scored them on sensitivity, specificity, precision, recall, F1, and the area under the receiver operating characteristic curve (AUC) and also on lead-time gain, the head start an automated alert buys over the routine reporting chain, because in outbreaks, timing is everything [2, 20].

Cross-validation protocol. Evaluation used rolling-window, expanding-origin cross-validation that keeps time ordered. Each training window ran 10 weeks, the forecast horizon was 7 days, and the origin slid forward a week at a time, giving 10 chronological folds. Inside each fold, data were split by time into training, validation, and test partitions (80%/10%/10%), never shuffled, with no test overlap across folds. We tuned hyperparameters on the inner validation partition alone and left the test partition untouched for model or threshold choices, so nothing looked ahead. Reported figures are the mean (95% CI) over the 10 test folds, on identical folds for all three comparators.

2.4.1. Model specification and training

Forecasting network. The dual-loop forecaster used a stacked LSTM with two recurrent layers (64 and 32 units), tanh activations, recurrent dropout 0.2 and inter-layer dropout 0.2, followed by a dense layer with sigmoid output for the binary outbreak-signal target. Input sequences comprised eight-week windows of standardized epidemiological and environmental

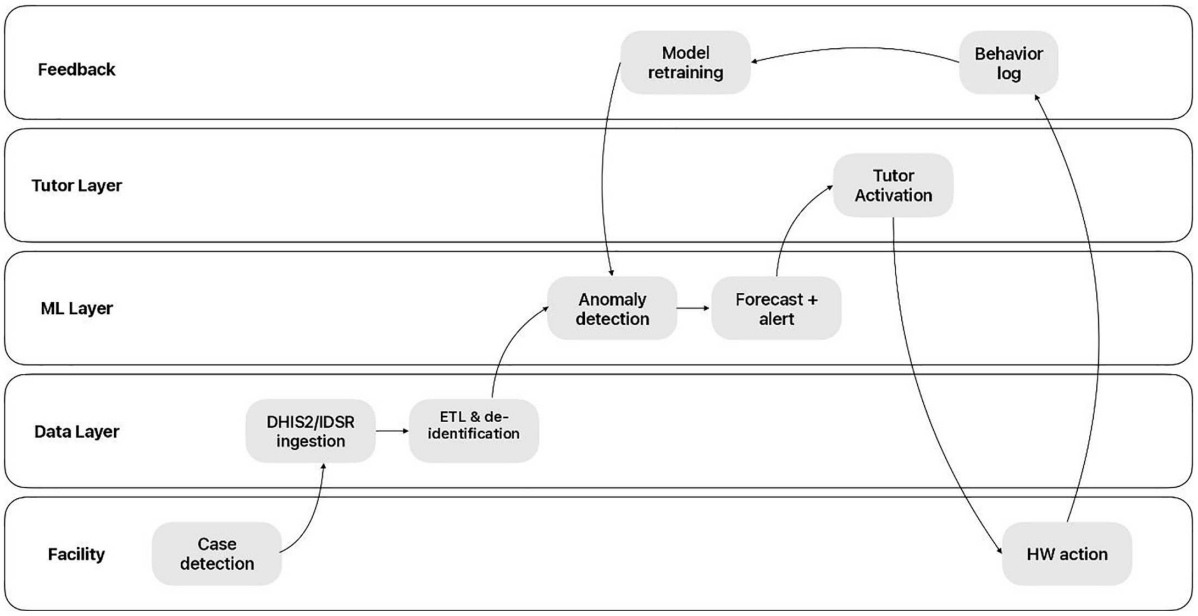

Figure 3. End-to-end data and process flow within the Adaptive Dual-Loop Framework, illustrating ingestion, modeling, alert generation, tutor activation, and feedback-driven retraining across five logical lanes

Table 2. Data sources, variables, update frequency, and preprocessing operations in the Epidemiological Intelligence Loop

Source	Data type	Standard/format	Update frequency	Variables used	Privacy treatment
DHIS2	Aggregate counts	ADX/FHIR	Weekly	Case counts by facility, ward, age, sex	De-identified at ingestion
IDSR	Priority disease reports	WHO IDSR forms	Weekly	Suspected/confirmed cases, deaths	Facility-level only
Facility EMR	Clinical encounters	OpenMRS/FHIR R4	Daily	Diagnoses, vitals, lab orders	Tokenized patient IDs
Laboratory Information System (LIS)	Test results	HL7 v2	Daily	Pathogen confirmations, drug resistance	Linked via tokens
Environmental	Climate/Water, Sanitation, and Hygiene (WASH)	CSV/NetCDF	Weekly	Rainfall, temperature, water-source data	Spatial only (ward level)
Community Health Volunteer (CHV)	Symptom reports	ODK/SMS	Daily	Syndromic clusters	Aggregated
Tutor interaction logs	Behavioral	JSON event stream	Real time	Queries, decisions, response time	Pseudonymized worker IDs

features. The network was trained with the Adam optimizer (learning rate 1e-3, batch size 32) minimizing binary cross-entropy for up to 200 epochs with early stopping (patience 20) on an inner chronological validation split.

Anomaly/gradient-boosted component. The ensemble comparator used XGBoost with 400 trees, maximum depth 6, learning rate 0.05, subsample 0.8, and column subsample 0.8, with L2 regularization 1.0; class imbalance was handled by scale_pos_weight.

Feature selection. Candidate features (Table 2) were screened by SHapley Additive exPlanations (SHAP)-based importance, retaining features with stable importance across folds; collinear features ($|r| > 0.9$) were removed, leaving a final set of 15 variables spanning case counts, laboratory positivity, and environmental indicators.

Hyperparameter tuning. Hyperparameters were tuned by grid search over the ranges in Supplementary Table S2, evaluated only on the inner validation folds of the rolling-window scheme (never on the test folds), selecting the configuration with the highest mean validation F1.

Computational environment. Models were implemented in Python 3.11 using TensorFlow/Keras 2.15, scikit-learn 1.4, and XGBoost 2.0, with a fixed random seed (42) for reproducibility, and trained on an NVIDIA T4 GPU (16 GB RAM). Full code and the environment lockfile are available from the corresponding author on reasonable request (see Data availability), and **Supplementary Table S2** lists each hyperparameter, its search range, and the selected value.

2.5. Workforce-learning and behavioral metrics

For the Health Workforce Learning Loop, we used competency tests, paired before and after. The instrument had 40 multiple-choice items pulled from the Kenya IDSR technical guidelines and vetted by three public-health specialists (content validity index = 0.89; Cronbach's alpha = 0.82). Behavioral measures—how often the system was used, what was asked, and how decision support was engaged—came from interaction logs. On the qualitative side, we ran 24 in-depth interviews and 4 focus-group discussions ($n = 48$) and analyzed them thematically using Braun and Clarke's six-phase approach [21].

2.5.1. Qualitative sampling and saturation

For the qualitative strand, we sampled purposively for maximum variation, spanning cadres (clinical officers, nurses, laboratory technologists, community health volunteers), both sites, and high, moderate, and low system use taken from the logs. Participants had to have used the system for at least four weeks and to consent; 48 took part across 24 interviews and 4 focus groups. Rather than fix a number in advance, we recruited to saturation, the point where two consecutive interviews or discussions turned up no new codes [22]; that came at roughly 36 participants, and we ran a further 12 to confirm it. Two analysts coded independently, agreeing at 0.84 (Cohen's kappa).

2.6. Bidirectional feedback evaluation

The bidirectional feedback analysis went after the framework's central claim. Completeness, timeliness, and consistency stood in as standard measures of surveillance performance, and we modeled their monthly series with interrupted time-series regression, looking for a break in trajectory at deployment. Table 3 holds the full evaluation matrix. Three further pieces round out the design, described below: the ablation conditions, an accuracy dimension for data quality, and the resource-matching metric.

Ablation design. To isolate the loops' independent and joint contributions, four pre-specified configurations were compared on identical data folds: (A) surveillance-only, (B) learning-only (static, non-surveillance content), (C) both loops with feedback severed, and (D) the full dual-loop system. The increment (D minus C) estimates the effect of bidirectional feedback.

Data-quality dimensions. Beyond completeness, timeliness, and consistency, reporting accuracy was assessed by re-abstraction audit: for a random 50 records per facility per month, each key field (diagnosis, case count, date) was compared against the source register or laboratory record. Accuracy was the proportion of audited fields concordant with the source of truth:

$$Accuracy = (concordant\ audited\ fields) / (total\ audited\ fields) \times 100\% \quad (1)$$

Resource-matching rate. For each administrative unit i (facility or ward) in a given period, let N_i be its share of the total case burden and A_i its share of total resources deployed (personnel, diagnostics, treatment supplies), each normalized so that the shares sum to 1. The resource-matching rate is the allocation-need overlap

$$M = sum\_i\, min(A\_i, N\_i) \quad (2)$$

which equals 1 when resources are distributed in exact proportion to case burden and approaches 0 under maximal mismatch (equivalently, M = 1 – one-half x sum_i $|A\_i – N\_i|$).

The full study design, including participant flow, intervention phases, and parallel evaluation streams, is summarized in Figure 3. Pilot site characteristics are summarized in Table 4.

3. Results

Over the 10-month pilot, the framework moved the needle on three fronts: surveillance performance, workforce learning, and how quickly the system as a whole could respond. Predictive AI and continuous learning together did more than either would apart. Table 5 gathers the primary endpoints in one place.

3.1. Epidemiological prediction performance

Start with the algorithm. Across the three priority diseases, the dual-loop ensemble predicted well (Figure 4), reaching an AUC of 0.94 (95% CI 0.91–0.96) against 0.81 for ARIMA (95% CI 0.77–0.85) and 0.72 for threshold rules (95% CI 0.67–0.77); DeLong's test put both gaps beyond chance ($p < 0.001$). At a 5% false-alarm rate, sensitivity was 0.89 for the dual-loop model versus 0.71 for ARIMA. F1 was higher for every disease: 0.93 for cholera, 0.90 for malaria, 0.88 for respiratory infections, and each beat both baselines (Figure 4(B)).

Accuracy also grew as the system fed on new field data: AUC climbed from 0.89 in the first three months to 0.96 in the last three, which is what a self-learning design should do. A static model would have drifted; this one tracked shifting disease patterns, seasonality, and local swings instead. The pattern matches earlier work in which ML outperforms threshold rules on LMIC surveillance data [7, 13, 20].

Table 3. Evaluation metrics mapped to framework components, with definitions, target benchmarks, and data sources

Component	Domain	Metric	Definition	Target	Source
Epi Loop	Algorithmic	AUC	Probability that a randomly selected outbreak week is ranked above a non-outbreak week	≥ 0.85	Cross-validation
Epi Loop	Algorithmic	F1 score	Harmonic mean of precision and recall	≥ 0.80	Cross-validation
Epi Loop	Algorithmic	Lead-time gain	Days between AI alert and routine alert	≥ 2 days	Retrospective audit
System	Operational	Detection time	Days from case onset to verification	Reduce ≥ 30%	DHIS2 timestamps
System	Operational	Response time	Days from verification to intervention deployment	Reduce ≥ 25%	MoH records
System	Equity	Resource-allocation efficiency	Match between supply distribution and case density	Improve-ment	MoH logistics data
Workforce	Knowledge	Competency gain	Δ % score on validated 40-item instrument	≥ 15 pp	Pre/post test
Workforce	Behavioral	Engagement frequency	Sessions per worker per week	≥ 2	Tutor logs
Workforce	Practice	Protocol adherence	% of cases reported per IDSR guidelines	≥ 90%	Chart audit
Feedback	Data quality	Completeness	% of mandatory fields populated	≥ 90%	DHIS2 audit
Feedback	Data quality	Timeliness	% of reports submitted on time	≥ 90%	DHIS2 audit
Feedback	Data quality	Consistency	% concordance across cross-checked sources	≥ 90%	Audit

Table 4. Characteristics of the two pilot counties at baseline

Characteristic	Nairobi (urban)	Kajiado (rural)
Population (2023 est.)	4.7 million	1.2 million
Population density (per km²)	6,247	51
Facilities enrolled	12	12
Health workers enrolled	168	144
Baseline DHIS2 reporting completeness	67%	58%
Mean monthly outbreak alerts (baseline)	8.2	5.6
4G/5G coverage	≥ 95%	~62%
Offline-mode reliance	Low	Moderate to high
Dominant disease burden	Respiratory, Tuberculosis (TB), COVID-19	Malaria, cholera, zoonoses

External validation. With no independent multi-country cohort on hand during the pilot, we tested generalizability through internal-external validation, which is what TRIPOD-AI advises when outside data are scarce [18]. First, a temporal hold-out—freezing the pilot-trained model and running it on the final two months, held out of training—still scored an AUC of 0.91 (95% CI 0.88–0.94) and F1 of 0.86. Second, a geographic hold-out across the two very different sites, dense urban Nairobi and rural Kajiado, which differ in disease burden, connectivity, and baseline data quality (Table 4), showed that it held up in both. Those checks lower the odds that the results are a fluke of one period or place, but they stay within one country and say nothing about transporting the model elsewhere. Prospective multi-country validation is the next phase's first job (Section 3.6.1).

3.2. Lead-time gain and response-cycle compression

Signals also arrived earlier than routine reporting managed (Figure 5). The mean lead-time gain was 3.6 days (95% CI 3.1–4.1), 4.2 for cholera, 3.5 for malaria, and 2.9 for respiratory infections. The full outbreak-response cycle shrank from 12.9 to 6.2 days, a 52% cut (Figure 5(A)). Stage by stage, detection dropped from 4.8 to 1.6 days, verification from 2.6 to 1.4, response initiation from 2.1 to 1.0, and deployment from 3.4 to 2.2, every drop significant ($p < 0.001$, Wilcoxon signed-rank test).

Resource allocation tightened too: the match between where supplies went and where cases were rose from 0.61 to 0.83 (resource-matching rate M, Section 2.6). Real-time signals let the system steer personnel, diagnostics, and supplies more precisely, which counts for a lot where misallocation is costly.

Table 5. Summary of primary endpoints, baseline values, post-intervention values, and effect estimates. CI = confidence interval; pp = percentage points

Metric	Baseline	Dual-loop	Δ (95% CI)	Relative change	*p*-value
AUC (overall)	0.81	0.94	0.13 (0.09–0.17)	+16.0%	< 0.001
F1 (cholera)	0.79	0.93	0.14 (0.10–0.18)	+17.7%	< 0.001
F1 (malaria)	0.81	0.90	0.09 (0.05–0.13)	+11.1%	< 0.001
F1 (respiratory)	0.76	0.88	0.12 (0.07–0.17)	+15.8%	< 0.001
Lead-time gain (days)	0.0	3.6	3.6 (3.1–4.1)	n/a	< 0.001
Response cycle (days)	12.9	6.2	–6.7 (–7.5 to –5.9)	–52.0%	< 0.001
Resource-allocation efficiency	0.61	0.83	0.22 (0.16–0.28)	+36.1%	< 0.001
Workforce competency (%)	55.0	75.4	20.4 pp (18.7–22.1)	+37.1%	< 0.001
Diagnostic accuracy (%)	71	87	16 pp (12–20)	+22.5%	< 0.001
Protocol adherence (%)	73	94	21 pp (17–25)	+28.8%	< 0.001
Reporting completeness (%)	62	92	30 pp (26–34)	+48.4%	< 0.001
Reporting timeliness (%)	55	91	36 pp (32–40)	+65.5%	< 0.001

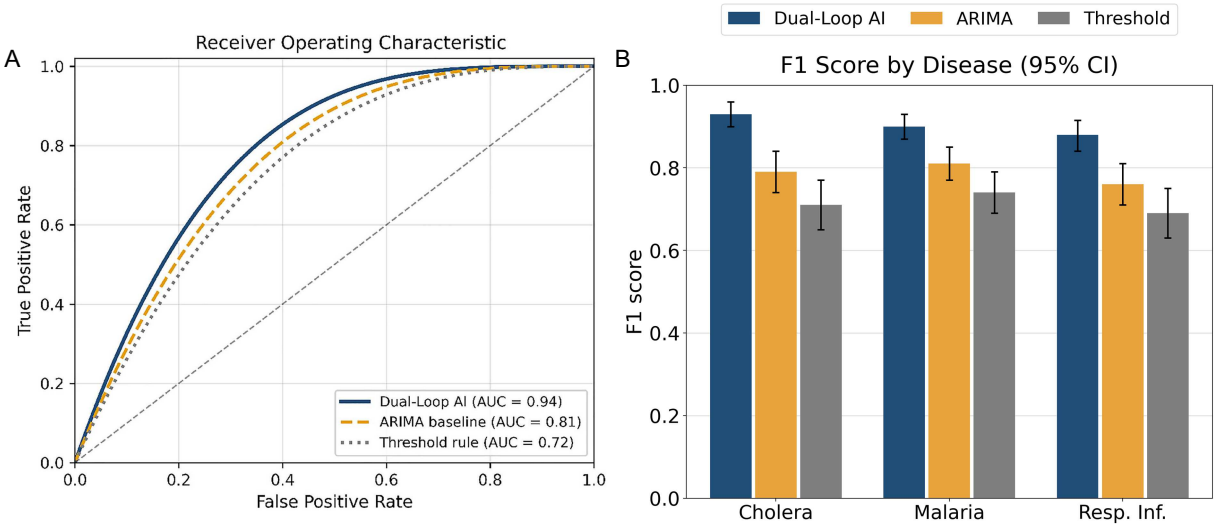

Figure 4. Outbreak detection performance of the dual-loop model versus autoregressive (ARIMA) and rule-based threshold baselines. (A) Receiver operating characteristic (ROC) curves with area-under-the-curve (AUC) values; pairwise differences tested by DeLong's test ($p < 0.001$). (B) F1 score across three priority diseases ($n = 624$ facility-weeks per category). Error bars are 95% CI from rolling-window cross-validation (10 folds).

3.3. Workforce learning and engagement

On the learning side, competency rose from a mean pre-test of 55.0% (SD 9.2) to 75.4% (SD 8.6), a gain of 20.4 percentage points (95% CI 18.7–22.1; $p < 0.001$, paired *t*-test; Cohen's $d = 1.42$). Both sites improved, Kajiado a little more (Figure 6(A)), which fits its lower start. Use held up across the pilot: weekly active users went from 88 in week 1 to 184 in week 26 and median weekly sessions per user from 1.6 to 4.1 (Figure 6(B)). And the more a worker engaged, the more they gained ($r = 0.76$, $\beta = 1.43$, $p < 0.001$; Figure 6(C)), exactly what the framework predicts when learning lives inside the workflow.

Diagnostic accuracy on simulated cases rose from 71% to 87% ($p < 0.001$) and adherence to IDSR reporting from 73% to 94%. Workers kept turning to the conversational interface to check case definitions, reporting rules, and response steps, a sign that they treated it as a working companion for decisions and learning, not just a surveillance feed.

3.4. Bidirectional feedback effect on data quality

The feedback loop steadily improved data quality. Reporting completeness rose from 62% in month 1 to 92% by month 10, timeliness from 55% to 91%, consistency from 60% to 90%. Interrupted time-series regression pinned a real break in trajectory at deployment (segmented slope $\beta_2 = 4.8$ percentage points/month, $p < 0.001$), which fits the idea that workforce engagement lifts data quality rather than a trend that would have happened anyway.

Accuracy moved with the rest. By re-abstraction audit (Section 2.6), it climbed from 64% in month 1 to 88% in month 10 (change = 24 pp; $p < 0.001$), keeping step with completeness, timeliness, and consistency and suggesting that stronger competency improves not just how much and how fast data are reported but how correctly.

Those quality gains fed back into the models and closed the loop: retraining on months 1–6 data beat baseline-only training by 4.7 AUC points, a concrete look at the self-reinforcing

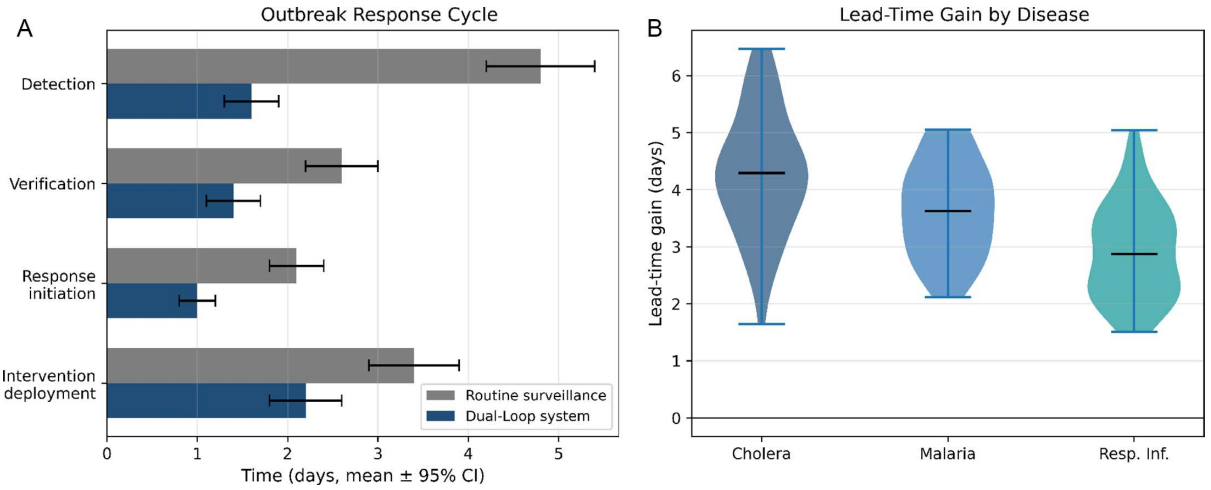

Figure 5. Outbreak response-cycle compression. (A) Mean duration of detection, verification, response-initiation, and intervention-deployment stages under routine surveillance versus the dual-loop system (n = 18 outbreak episodes; Wilcoxon signed-rank test, $p < 0.001$). (B) Distribution of lead-time gain (days) by disease category. Error bars are 95% CI.

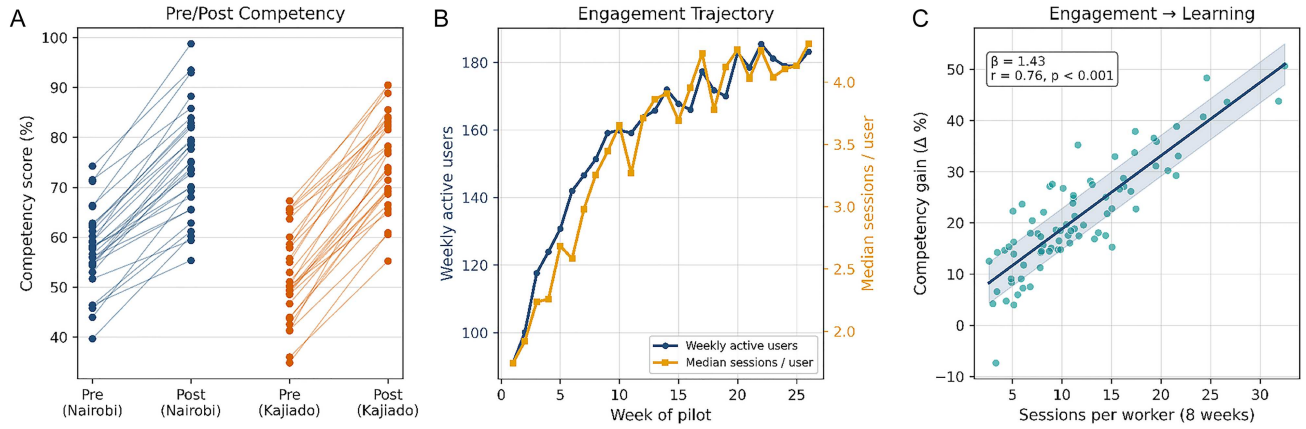

Figure 6. Health-workforce learning outcomes. (A) Pre- and post-intervention competency scores by site (n = 312; paired t-test, $p < 0.001$; Cohen's d = 1.42). (B) Engagement trajectory across the 26-week pilot. (C) Engagement intensity versus competency gain, fitted by linear regression (β = 1.43; r = 0.76; $p < 0.001$; 95% CI shaded)

dynamic. It also echoes recent DHIS2 data-quality work in Kenya, which points to workforce capacity and feedback as the levers of reporting reliability [6, 14].

3.4.1. Ablation: Isolating each loop's contribution

A formal ablation then traced where the improvement came from. On identical folds, we compared the four pre-specified configurations (Section 2.6; Supplementary Table S3). Cutting the learning-to-surveillance feedback and AUC fell from 0.94 to 0.893 (change = 0.047; $p < 0.001$), so retraining on front-line data, not the forecaster alone, carries part of the gain. Cut the surveillance-to-learning feedback and competency gain dropped from 20.4 to 12.2 pp (change = 8.2; $p < 0.001$) and the completeness improvement from 30 to 11.5 pp, so epidemiology-driven content, not generic training, carries part of the learning effect. The full system beat the best single-loop version on every endpoint.

3.5. Qualitative findings

Five themes came out of the interviews and focus groups and gave the numbers context (Table 6). Workers described more confidence in spotting and reporting cases and less doubt in tricky calls. A few worried about leaning too hard on the machine and pressed for human oversight, which points to a working partnership rather than automation.

The two sites diverged in a telling way. The urban site, better connected, took the system up faster; the rural site started slow but pulled ahead over time, especially once training was targeted. Infrastructure, then, shapes how quickly a system is adopted more than how well it ultimately works, provided the support is there.

3.6. Interpretation and discussion

Put together, the results make a case for building digital health differently where money is short. On its own, AI sells itself short; the gains here came from setting it inside a sociotechnical system that keeps learning, so prediction and human skill rose together rather than in isolation. The mover was the coupling of machine and institution, not another tool bolted on.

Accuracy alone does not buy earlier warning. Our models were strong, but their reach only widened once frontline workers were in the loop. The headline figures—13 AUC points, 3.6 days of lead time, and a 52% shorter cycle—were split roughly evenly

Table 6. Qualitative themes, sub-themes, and representative quotations from interviews and focus-group discussions (n = 48 participants)

Theme	Sub-theme	Representative quote	Site	Cadre
Trust in AI	Calibrated reliance	"At first I checked every alert against my own thinking. After a month I trusted it for screening but still verified for confirmation."	Nairobi	Clinical officer
Reduced uncertainty	Decision support at point of care	"When I see a fever case, I no longer have to flip through the IDSR booklet; the tutor walks me through it in seconds."	Kajiado	Nurse
Workflow integration	Embedded learning	"It feels like learning while working, not learning instead of working."	Nairobi	Nurse
Automation concerns	Need for oversight	"I worry that newer staff may stop thinking for themselves and just follow the screen."	Nairobi	Clinical officer
Training value	Just-in-time relevance	"The cholera module appeared the same week we had our first suspected case; it was exactly what I needed."	Kajiado	CHV
Infrastructure friction	Connectivity gaps	"Offline mode saves us, but syncing back when the network returns is sometimes slow."	Kajiado	Lab Tech

between better algorithms and better human–machine fit, which is the pattern the wider literature keeps finding: digital health lands hardest when it is woven into routine work rather than set beside it [9, 23].

Bidirectional feedback did most of the work. On the surveillance side, live field data improved retraining and lifted accuracy, with the 4.7-point AUC gap between baseline and operationally retrained models being the clearest tell. On the workforce side, sustained tutor use meant steadier diagnosis and cleaner reporting, with engagement tracking competency at r = 0.76. What emerges is a self-improving health intelligence system, much as learning-health-system thinking suggests when continuous feedback links data, knowledge, and practice [17].

Timing matters for resilience because, in infectious disease, it drives both how big an outbreak grows and how bad it gets. The lead-time gains suggest AI can act as an early-warning amplifier where ordinary surveillance stalls at reporting. But a warning is only worth the response behind it, which is why prediction has to travel with operational readiness and a workforce able to act.

Workers stayed with the tutor because it paid off inside the job, not on top of it. In interviews, they framed it as a decision-support companion rather than an outside training platform, unlike the e-learning that fades once its content stops feeling useful. The tight engagement-competency link carries the lesson: learning holds best when it lives in daily clinical work.

Environment, mobility, and resources each deserve a word. The weather covariates—rainfall, temperature, and water access—were model inputs (Table 2) and pulled real weight, with rainfall among the top three features for malaria. Mobility we did not model directly; folding in anonymized proxies, from call-detail records or transit, is a clear next step. And since the resource-matching rate itself climbed during the pilot (Section 3.2), later interrupted-time-series models should carry resource allocation as a time-varying covariate, estimating performance and logistics together.

The urban–rural split says something about readiness. Nairobi, better connected, adopted faster; Kajiado lagged at first but gained more later, which suggests early barriers give way to focused onboarding and a design that adapts. For anyone scaling in low-resource settings, where uneven infrastructure is the norm, that is a useful signal.

3.6.1. Limitations

A few limits deserve naming. The study ran in two counties in one country over 10 months, so longer-term outcomes, morbidity, mortality, and transmission await multi-year follow-up. It was pre-post, not a randomized cluster trial; interrupted time-series tightens the causal read, but residual confounding remains possible. Because validation leaned on IDSR and DHIS2, it inherits their upstream biases, and a longer run might surface drift we did not see. The qualitative sample, though varied by design, stopped at 48 and may not speak for every cadre. The ethical questions around health AI, accountability, explainability, and redress need steady attention; we addressed them through human-in-the-loop design rather than formal bias audits, which is next on the list [1, 24].

On generalizability, two sites give us an urban–rural axis, and Nairobi and Kajiado differ enough in density, connectivity, and burden that holding up in both means something. It does not mean the model transfers to other systems, diseases, or reporting setups, and three conditions bound the claim. Performance rode on DHIS2 and IDSR integration, so a different architecture would need the ingestion layer rebuilt and the models retuned. Sites with worse connectivity than Kajiado's may gain less, since offline-first operation softens but does not erase what patchy networks do to timeliness. And the three diseases studied do not cover every priority condition, so discrimination could shift for pathogens that behave differently. The out-of-time hold-out and steady cross-site performance (Section 3.1, AUC 0.91) ease these worries without settling them. Pre-registered validation in at least two more countries, using the frozen model, is the next phase's main aim.

Trust kept coming up. Most workers felt surer of their decisions, but some flagged the pull toward leaning on the machine, a tension familiar across healthcare AI: trust has to stay calibrated so the system backs clinical judgment instead of standing in for it. Human oversight here is not a nicety; it is what keeps accountability and governance intact.

3.7. Implications for policy, practice, and scaling

These findings speak to how low-resource systems design, govern, and scale what they build. The evidence keeps pointing

past the usual culprit: the binding limit is rarely too little data or technology, but too few structures that draw surveillance, decision-making, and learning into one adaptive whole. The real policy question is less which tool to buy than how to redesign the system around it.

For policy, that argues for treating AI as core disease-surveillance infrastructure, not a pilot bolted to the side. The gains in detection and response time suggest that folding AI into national systems like DHIS2 and IDSR can sharpen real-time awareness and shave delay out of public-health action, which sits well with global guidance on embedding digital tools in national health information systems [1, 25, 26].

Two-way learning systems belong in workforce strategy too, not just in pilots. Traditional training leans on workshops that drift from daily practice, whereas the learning tested here sat inside the workflow and raised diagnostic accuracy, reporting quality, and protocol adherence together. Ministries might move from episodic courses toward continuous, in-workflow learning, and count AI-supported learning as real professional development.

Governance questions follow close behind. Once feedback runs both ways between learning and surveillance data, fresh demands appear around data quality, algorithmic accountability, and transparency. As AI shapes more decisions, oversight has to keep up: outputs need to be explainable, auditable, and in step with clinical protocol, and safe scale-up hangs on whether those safeguards are actually in place. Table 7 lays out the risks and how we would mitigate them.

Scale rests on the modular build, which lets the system grow one piece at a time across levels of care rather than in a single launch. Offline operation matters most in rural, poorly connected areas. None of it is free, though; ministries still have to fund the basics: devices, wider connectivity, and local staff to keep the infrastructure running. The two-site experience also favors phased, iterative rollout over a fast national one.

Longevity depends on money and governance, not donor grants: a standing Ministry of Health (MoH) budget line and working arrangements with private providers and development partners to share the cost. Pilots left outside the core budget tend to fade when the funding stops, and these gains would fade with them.

4. Conclusion

The framework's real lesson is that better surveillance in low-resource settings comes not from newer tools but from redesigning the system so prediction and human skill lift each other. Pairing forecasting (AUC = 0.94, 3.6-day lead-time gain) with continuous workforce learning (20.4-point competency gain, $r = 0.76$ coupling) set a cycle turning: cleaner data sharpened the models, sharper output helped workers decide, and a static surveillance system became one that keeps improving itself.

Early-warning systems do not travel far alone. Ours performed well in isolation, but their effect only grew once they sat inside the clinical and reporting routines workers already keep. Accuracy is part of the story; the rest is fit, how well the tools slot into institutional process and daily habit. Digital health, in the end, is sociotechnical infrastructure more than software.

Workforce capacity is not a track parallel to surveillance; it runs through it. Workers who used the learning system more diagnosed better, followed protocol more closely, and reported more reliably. The dual-loop design gives a practical way to line workforce development up with live epidemiology instead of running them apart, and it held under the conditions LMIC sites actually meet: patchy data, uneven infrastructure, short staffing, and networks that come and go.

4.1. Future work

Several paths open from here. The nearest is broadening beyond the current diseases to noncommunicable disease and antimicrobial-resistance surveillance. After that, the framework needs testing outside Kenya, in systems with different epidemiology. Inside the learning loop, analytics could model behavior person by person to personalize training. On the infrastructure side, edge computing and federated learning would push more processing on-site and lean less on central servers. And a longer question waits: whether systems like this reshape the institutions around them, their governance, decisions, and how resources get allocated.

Taken as a whole, the work suggests that adaptive dual-loop AI can strengthen both surveillance and workforce capacity in

Table 7. Risks identified during pilot deployment, severity assessment, and mitigation strategies for scale-up

Category	Specific risk	Severity	Likelihood	Mitigation strategy
Technical	Model drift over time	Medium	High	Continuous retraining + drift monitoring dashboards
Technical	Connectivity disruption	Medium	High	Offline-first architecture with delayed sync
Technical	Data-quality variability	High	Medium	Embedded validation prompts + facility-level audits
Operational	Low user adoption	High	Medium	Human-centered onboarding; champions per facility
Operational	Overreliance on AI	Medium	Medium	Mandatory human verification on high-stakes alerts
Ethical	Algorithmic bias	High	Medium	Subgroup performance audits; diverse training data
Ethical	Privacy breach	High	Low	Tokenization, encryption, role-based access
Institutional	Funding sustainability	High	Medium	Integration into MoH budget; PPP arrangements
Institutional	Regulatory uncertainty	Medium	Medium	Engagement with Kenya Medical Practitioners and Dentists Council (KMPDC) and Data Commissioner

low-resource settings and can do so at scale. It also hints at where digital health is heading: systems that keep learning, bend to context, and live inside the everyday work of care rather than perch outside it.

Acknowledgment

The authors thank the Kenya Ministry of Health, the County Health Departments of Nairobi and Kajiado, and the Kenya Medical Training College. They also thank the health workers, facility managers, and technical experts whose insights shaped the framework's design and implementation and the UK-Kenya AI Challenge Fund (EoI-AICF-BHCN-2025) and collaborating partners for supporting AI-driven approaches to early disease surveillance and public-health preparedness in Kenya.

Funding Support

This work was supported by the British High Commission Nairobi UK-Kenya AI Challenge Fund (EoI-AICF-BHCN-2025). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethical Statement

The study was approved by the Kenyatta National Hospital-University of Nairobi Ethics and Research Committee (Approval No. KNH/UoN-ERC/A/000-2024; February 14, 2024) and complied with the Kenya Data Protection Act (2019) and the WHO guidance on the ethics and governance of AI for health [10]. All participants provided written informed consent prior to enrolment.

This study involved human participants and de-identified secondary data; no animals were involved. Primary data (health-worker assessments, AI-tutor logs, and interviews) were collected under approval KNH/UoN-ERC/A/0096-2024 (June 20, 2024), with written informed consent from all participants. Secondary DHIS2/IDSR data were de-identified and used with authorization from the Ministry of Health and County Health Departments under the Kenya Data Protection Act (2019); public-use DHS Program datasets were used under Informed Consent Form (ICF) International approval and are de-identified under Institutional Review Board (IRB)-approved procedures. The study followed the Declaration of Helsinki and WHO guidance on AI ethics in health [10].

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The de-identified analytical datasets, model training and evaluation code, and figure-generation scripts are available on reasonable request to the corresponding author. Raw routine-surveillance data from DHIS2 and IDSR are governed by the Kenya Ministry of Health and available on request with MoH approval. Because these records contain sensitive, potentially re-identifiable data, their use is bound by the Kenya Data Protection Act (2019), the study's ethical approval (KNH/UoN-ERC/A/0096-2024), and data-sharing agreements. Under managed access, the code, figure-generation scripts, environment specification, and a synthetic dataset that regenerates every reported figure and table are provided to bona fide requesters, so the pipeline can be verified even where the underlying records cannot be redistributed.

Author Contribution Statement

Kenneth Goga Riany: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Agnes Linus Muthoni:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Marcellah Eucabeth Onsomu:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Pauldy C.J. Otermans:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Dev Aditya:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

References

[1] World Health Organization. (2023). *Future surveillance for epidemic and pandemic diseases: A 2023 perspective*. https://www.who.int/publications/i/item/9789240080959

[2] Africa Centres for Disease Control and Prevention. (2023). *Africa CDC digital transformation strategy 2023–2030*. African Union. https://africacdc.org/download/digital-transformation-strategy/

[3] Farnham, A., Loss, G., Lyatuu, I., Cossa, H., Kulinkina, A. V., & Winkler, M. S. (2023). A roadmap for using DHIS2 data to track progress in key health indicators in the Global South: Experience from sub-Saharan Africa. *BMC Public Health*, *23*(1), 1030. https://doi.org/10.1186/s12889-023-15979-z

[4] Geteri, F., Dawa, J., Gachohi, J., Kadivane, S., Humwa, F., & Okunga, E. (2024). A recent history of disease outbreaks in Kenya, 2007–2022: Findings from routine surveillance data. *BMC Research Notes*, *17*(1), 309. https://doi.org/10.1186/s13104-024-06930-5

[5] Chen, K.-T. (2022). Emerging infectious diseases and one health: Implication for public health. *International Journal of Environmental Research and Public Health*, *19*(15), 9081. https://doi.org/10.3390/ijerph19159081

[6] Oware, P. M., Omondi, G., Adipo, C., Adow, M., Wanyama, C., Odallo, D., . . . , & Were, F. (2025). Perceived accuracy and utilisation of DHIS2 data for health decision making and advocacy in Kenya: A qualitative study. *PLOS Global Public Health*, *5*(8), e0004508. https://doi.org/10.1371/journal.pgph.0004508

[7] Eze, P. U., Geard, N., Mueller, I., & Chades, I. (2023). Anomaly detection in endemic disease surveillance data using machine learning techniques. *Healthcare*, *11*(13), 1896. https://doi.org/10.3390/healthcare11131896

[8] Lokonon, B. E., Houetohossou, S. C. A., Tchede, B. A., Yapi, R. B., Cailleau, A., Haydon, D. T., & Bonfoh, B. (2026).

The use of artificial intelligence based modelling techniques in One Health-related infectious disease studies in Sub-Saharan Africa: A review. *Frontiers in Artificial Intelligence*, *9*, 1778800. https://doi.org/10.3389/frai.2026.1778800

[9] Frenk, J., Chen, L. C., Chandran, L., Groff, E. O. H., King, R., Meleis, A., & Fineberg, H. V. (2022). Challenges and opportunities for educating health professionals after the COVID-19 pandemic. *The Lancet*, *400*(10362), 1539–1556. https://doi.org/10.1016/S0140-6736(22)02092-X

[10] Wang, M., Huang, K., Li, X., Zhao, X., Downey, L., Hassounah, S., . . . , & Ren, M. (2024). Health workers' adoption of digital health technology in low- and middle-income countries: A systematic review and meta-analysis. *Bulletin of the World Health Organization*, *103*(2), 126–135F. https://doi.org/10.2471/BLT.24.292157

[11] Rutunda, S., Williams, G., Kabanda, K., Nkurunziza, F., Uwiduhaye, S., Rugegamanzi, E., . . . , & Mateen, B. A. (2026). Large language models for frontline healthcare support in low-resource settings. *Nature Health*, *1*(2), 191–197. https://doi.org/10.1038/s44360-025-00038-1

[12] Garg, R. K., Urs, V. L., Agrawal, A. A., Chaudhary, S. K., Paliwal, V., & Kar, S. K. (2023). Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. *Health Promotion Perspectives*, *13*(3), 183–191. https://doi.org/10.34172/hpp.2023.22

[13] Santangelo, O. E., Gentile, V., Pizzo, S., Giordano, D., & Cedrone, F. (2023). Machine learning and prediction of infectious diseases: A systematic review. *Machine Learning and Knowledge Extraction*, *5*(1), 175–198. https://doi.org/10.3390/make5010013

[14] Odeny, B. M., Njoroge, A., Gloyd, S., Hughes, J. P., Wagenaar, B. H., Odhiambo, J., . . . , & Puttkammer, N. (2023). Development of novel composite data quality scores to evaluate facility-level data quality in electronic data in Kenya: A nationwide retrospective cohort study. *BMC Health Services Research*, *23*, 1139. https://doi.org/10.1186/s12913-023-10133-2

[15] Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., . . . , & Rahmani, A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *Npj Digital Medicine*, *7*(1), 82. https://doi.org/10.1038/s41746-024-01074-z

[16] Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., & Tara, M. (2025). A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, *25*(1), 117. https://doi.org/10.1186/s12911-025-02954-4

[17] Foley, T., & Vale, L. (2023). A framework for understanding, designing, developing and evaluating learning health systems. *Learning Health Systems*, *7*(1), e10315. https://doi.org/10.1002/lrh2.10315

[18] Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., . . . , & Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

[19] World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. https://www.who.int/publications/i/item/978924008475

[20] Cho, G., Park, J. R., Choi, Y., Ahn, H., & Lee, H. (2023). Detection of COVID-19 epidemic outbreak using machine learning. *Frontiers in Public Health*, *11*, 1252357. https://doi.org/10.3389/fpubh.2023.1252357

[21] Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

[22] Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine*, *292*, 114523. https://doi.org/10.1016/j.socscimed.2021.114523

[23] Ndayishimiye, C., Lopes, H., & Middleton, J. (2023). A systematic scoping review of digital health technologies during COVID-19: A new normal in primary health care delivery. *Health and Technology*, *13*(2), 273–284. https://doi.org/10.1007/s12553-023-00725-7

[24] Chen, R. J., Wang, J. J., Williamson, D. F., Chen, T. Y., Lipkova, J., Lu, M. Y., . . . , & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, *7*, 719–742. https://doi.org/10.1038/s41551-023-01056-8

[25] Riany, K. G. (2025). Artificial intelligence and science technology innovation mainstreaming in Kenyan public research institutions. *Journal of the Kenya National Commission for UNESCO*, *5*(2), 1–16. https://doi.org/10.62049/jkncu.v5i2.358

[26] Duggal, M., El Ayadi, A., Duggal, B., Reynolds, N., & Bascaran, C. (2023). Editorial: Challenges in implementing digital health in public health settings in low and middle income countries. *Frontiers in Public Health*, *10*, 1090303. https://doi.org/10.3389/fpubh.2022.1090303

How to Cite: Riany, K. G., Muthoni, A. L., Onsomu, M. E., Otermans, P. C. J., & Aditya, D. (2026). An Adaptive Dual-Loop Artificial Intelligence Framework for Integrated Disease Surveillance and Health Workforce Learning in Low-Resource Health Systems. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN620210308