Gray's Smart Ambulance Platform

Anomaly Detection & Alert Quality Report

Tasks 2A, 2B, 3A, 3B

System: Real-Time Patient Vitals Monitoring
Models: Isolation Forest, One-Class SVM, LSTM Autoencoder
Patients: 50 synthetic patients, 30 minutes each
Date: February 12, 2026

Abstract

This report describes the anomaly detection and risk scoring system built for the Gray's Smart Ambulance platform. Patient vitals — heart rate (HR), oxygen saturation (SpO_2), blood pressure (SBP/DBP), and motion — are streamed at 1 Hz and evaluated in sliding windows. An ensemble of three models produces a continuous risk score, which drives a three-level alert system (NORMAL / WARNING / CRITICAL). The report covers model design, feature engineering, windowing strategy, risk-scoring logic, alert quality metrics, and failure-mode analysis.

Contents

1 Task 2A: Anomaly Detection Model

1.1 Overview and Motivation

Traditional vital-sign monitoring relies on fixed clinical thresholds (e.g. HR > 100 bpm triggers an alert). While simple to implement, threshold-based approaches suffer from two fundamental weaknesses in the ambulance context:

1. **Insensitivity to gradual deterioration.** A patient whose HR rises steadily from 75 to 115 bpm over five minutes may never cross a hard threshold yet is clearly worsening.
2. **Inability to contextualise.** A single vital is only meaningful alongside others. An HR of 120 combined with SpO_2 of 88 % is far more serious than the same HR with normal oxygen saturation.

The system therefore employs three complementary machine-learning models that learn the multivariate structure of *normal* transport data and flag deviations as anomalies, without relying on hard-coded thresholds.

1.2 Windowing Strategy

Raw sensor data arrives at 1 Hz. A **sliding-window** approach converts the continuous stream into discrete feature vectors suitable for the models.

Table 1: Windowing parameters

Parameter	Value	Rationale
Window duration	30 s	Captures one clinical cycle of variation
Step size	5 s	83 % overlap; reduces latency to $\leq$5 s
Sampling rate	1 Hz	One measurement per second per sensor
Windows / patient	$\approx$414	30-min transport at 5-s step
LSTM sequence	10 windows	5-min temporal context (50 s effective)

A 5-second step means that a new risk assessment is issued every five seconds, giving the system a maximum alert latency of one window cycle (5 s) after an event begins.

1.3 Feature Engineering

Each 30-second window is summarised into **21 numerical features** across five semantic groups. Using aggregated features rather than raw samples reduces noise, standardises input dimensionality for the models, and embeds clinical knowledge directly into the representation.

Table 2: Feature set (21 features per window)

Group	Features	Clinical Relevance
Mean values (4)	`hr_mean`, `spo2_mean`, `sbp_mean`, `dbp_mean`	Baseline level; detects chronic abnormality
Variability (4)	`hr_std`, `spo2_std`, `sbp_std`, `dbp_std`	Erratic readings signal instability or artefact
Trends (4)	`hr_slope`, `spo2_slope`, `sbp_slope`, `dbp_slope`	Captures deterioration *before* threshold breach
Extremes (6)	`hr_min`, `hr_max`, `spo2_min`, `spo2_max`	Brief critical spikes invisible in the mean
Derived (3)	`hr_spo2_corr`, `pp_mean`, `pp_std`	Cross-vital coherence; pulse pressure monitors cardiac output
Motion (2)	`motion_mean`, `motion_max`	Distinguishes real changes from vehicle-induced artefacts

Slopes are computed via ordinary least squares on the 30-second window. The HR–SpO_2 correlation encodes a key clinical relationship: under physiological stress, HR tends to rise as SpO_2 falls, producing a characteristic negative correlation.

1.4 Model 1: Isolation Forest

1.4.1 Algorithm

Isolation Forest [1] builds an ensemble of random binary trees. At each node a feature and a split value are chosen uniformly at random. Data points that are isolated in *few* splits are flagged as anomalies, because anomalous points tend to be sparse and far from the cluster.

The anomaly score for point x is:

$$s(x, n) = 2^{-\frac{E[h(x)]}{c(n)}} \tag{1}$$

where $h(x)$ is the path length for point x, $c(n) = 2H(n-1) - \frac{2(n-1)}{n}$ is the average path length of an unsuccessful search in a binary search tree, and $H(i)$ is the harmonic number.

1.4.2 Configuration

Table 3: Isolation Forest hyperparameters

Parameter	Value	Effect
`n_estimators`	400	Higher stability, lower variance
`contamination`	0.05	5 % of training windows expected anomalous
`random_state`	42	Reproducibility

1.4.3 Strengths and Limitations

Strengths: fast inference ($O(n \log n)$), scales to 50 patients, naturally handles multivariate data, and is interpretable via feature importance.

Limitations: treats each window independently (no temporal memory), and may underperform on data with many correlated features.

1.5 Model 2: One-Class SVM

1.5.1 Algorithm

One-Class SVM [2] maps training data into a high-dimensional reproducing kernel Hilbert space and finds a hyperplane that separates the data from the origin with maximum margin. The decision function $f(x) = \mathbf{w} \cdot \phi(x) - \rho$ assigns positive scores to inliers and negative scores to anomalies.

Using the RBF kernel $k(x, x') = \exp(-\gamma \|x - x'\|^2)$ allows non-linear boundaries that capture curved normal regions in feature space.

1.5.2 Configuration

Table 4: One-Class SVM hyperparameters

Parameter	Value	Effect
`nu`	0.05	Upper bound on fraction of outliers (5 %)
`kernel`	RBF	Non-linear boundary around normal cluster
`gamma`	scale	Auto-scales to feature variance

1.5.3 Strengths and Limitations

Strengths: flexible decision boundary, robust to non-Gaussian distributions, and complements the tree-based Isolation Forest.

Limitations: quadratic training cost with n; slower inference than Isolation Forest; less interpretable.

1.6 Model 3: LSTM Autoencoder

1.6.1 Architecture

The LSTM Autoencoder [3] learns to reconstruct *sequences* of windows. Anomalies produce high reconstruction error because the model, trained on normal data, cannot accurately reconstruct unusual patterns.

Table 5: LSTM Autoencoder architecture

Layer	Units / Operation	Output Shape
Input	—	$(10, 21)$
LSTM Encoder 1	64 units, return_seq=True	$(10, 64)$
LSTM Encoder 2	32 units, return_seq=False	(32)
RepeatVector	repeat $\times 10$	$(10, 32)$
LSTM Decoder 1	32 units, return_seq=True	$(10, 32)$
LSTM Decoder 2	64 units, return_seq=True	$(10, 64)$
TimeDistributed	Dense(21)	$(10, 21)$

The input is a sequence of 10 consecutive windows (covering 50–80 seconds of data). Training uses MSE loss; the anomaly score for sequence X is:

$$\text{score}(X) = \frac{1}{T \cdot F} \sum_{t=1}^{T} \sum_{f=1}^{F} \left(X_{t,f} - \hat{X}_{t,f} \right)^2 \tag{2}$$

where $T = 10$ timesteps, $F = 21$ features, and $\hat{X}$ is the reconstructed sequence.

1.6.2 Training Details

Table 6: LSTM training configuration

Parameter	Value
Optimiser	Adam
Loss	Mean Squared Error (MSE)
Epochs	40
Batch size	128
Train / Val / Test split	40 / 5 / 5 patients

The patient-level split (rather than window-level) prevents data leakage, ensuring the model generalises to unseen patients.

1.6.3 Strengths and Limitations

Strengths: explicitly models temporal context; detects gradual deterioration that static models miss; captures cross-vital interactions over time.

Limitations: requires 10 windows of history before producing a score (first 9 windows use a median fallback); slower to train; less interpretable than classical models.

1.7 Ensemble and Thresholding

The three models produce heterogeneous scores on different scales. All scores are normalised to $[0, 1]$ using a sigmoid transformation before combination. The ensemble score

is a weighted average that assigns greater weight to the LSTM, reflecting its superior temporal modelling:

$$s_{\text{ensemble}} = 0.30 \cdot s_{\text{ISO}} \; + \; 0.30 \cdot s_{\text{SVM}} \; + \; 0.40 \cdot s_{\text{LSTM}} \tag{3}$$

Anomaly flags are generated by thresholding at the 95^{th} percentile of validation-set scores, which corresponds to the 5 % contamination assumption.

1.8 Handling False Positives

False positives are a primary concern in ambulance monitoring because they erode paramedic trust and cause alarm fatigue. The system addresses this through three mechanisms:

1. **Ensemble voting.** Alerts are only promoted to WARNING or CRITICAL if the ensemble score exceeds the threshold. A single model anomaly that the other two models do not corroborate will produce a lower ensemble score and may not reach the threshold.
2. **Temporal persistence filter.** An event must span at least three consecutive anomalous windows ($\geq 15\,\text{s}$) before it is reported. Isolated spikes — common with motion artefacts — are suppressed.
3. **Model agreement confidence.** The standard deviation of the three normalised scores is used to gate confidence. High inter-model disagreement ($\text{std} > 0.3$) reduces confidence and can suppress alerts, signalling to the system that models disagree.

2 Task 2B: Risk Scoring Logic

2.1 Design Philosophy

The risk scoring layer sits above the anomaly detection models and translates raw anomaly scores into a clinically actionable signal. Its design is governed by two principles:

- **High recall, acceptable precision.** In a pre-hospital setting, missing a genuine emergency is catastrophic; a false alarm is manageable. The system is therefore tuned to prioritise sensitivity.
- **Explainability.** Every alert must include a human-readable reasoning list that the attending paramedic can verify in seconds.

2.2 Risk Score Formula

The final risk score for window i is:

$$r_i = s_{\text{ensemble},i} \; \times \; \Sigma_i \; \times \; P_i \; \times \; C_i \tag{4}$$

where:

- $s_{\text{ensemble},i} \in [0, 1]$ is the ensemble anomaly score (Equation 3);

- $\Sigma_i \geq 1$ is the clinical severity multiplier (Section 2.3);
- $P_i \in [1, 1.5]$ is the temporal persistence multiplier (Section 2.4);
- $C_i \in [0, 1]$ is the overall confidence (Section 2.5).

The additive structure of the severity multiplier (rather than multiplicative stacking) prevents unrealistically high scores from accumulating when several mild conditions co-occur simultaneously.

2.3 Clinical Severity Multipliers

The severity term $\Sigma_i = 1 + \delta_{\text{SpO}_2} + \delta_{\text{HR}} + \delta_{\text{multi}} + \delta_{\text{BP}} + \delta_{\text{var}}$ is built from additive contributions:

Table 7: Clinical severity contributions to Σ_i

Vital Sign	Condition	δ	Clinical Basis
SpO_2	spo2_min $< 88\,\%$	+2.0	Critical hypoxaemia
	spo2_min $< 92\,\%$	+1.0	Mild hypoxaemia
	spo2_mean $< 94\,\%$	+0.3	Borderline
Heart Rate	HR mean > 140 bpm	+1.0	Severe tachycardia
	HR mean > 120 bpm	+0.5	Tachycardia
Heart Rate	HR mean < 45 bpm	+1.0	Severe bradycardia
	HR mean < 55 bpm	+0.3	Bradycardia
Multi-vital	HR > 120 **and** $SpO_2 < 92\,\%$	+1.5	Combined cardiovascular compromise
Trend	HR slope > 5 bpm/window	+0.5	Rapid deterioration
	HR slope > 2 bpm/window	+0.2	Moderate trend
Variability	HR std > 15 bpm	+0.3	Haemodynamic instability
Blood Pressure	SBP > 180 or SBP < 90 mmHg	+0.5	Hyper/hypotension

The multi-vital term deserves emphasis. When both tachycardia and hypoxaemia are present simultaneously, the physiological threat is disproportionately greater than either condition alone. The +1.5 bonus encodes this non-linear clinical interaction directly.

2.4 Temporal Persistence

A single anomalous window is insufficient to trigger an alert, because brief artefacts (road bumps, patient movement) routinely produce transient spikes. The persistence multiplier P_i examines the five most recent windows:

$$P_i = \begin{cases} 1.5 & \text{if } \geq 4 \text{ of the last 5 windows are anomalous} \\ 1.3 & \text{if } \geq 3 \text{ of the last 5 windows are anomalous} \\ 1.0 & \text{otherwise} \end{cases} \tag{5}$$

This rewards *sustained* deterioration with a higher score while discounting isolated spikes.

2.5 Confidence Score

The confidence $C_i = C_{\text{model}} \times C_{\text{motion}}$ is a product of two independent factors.

2.5.1 Model Agreement Confidence C_{model}

$$C_{\text{model}} = \begin{cases} 1.0 & \sigma_{\text{models}} < 0.10 \quad \text{(all models agree)} \\ 0.9 & \sigma_{\text{models}} < 0.20 \\ 0.7 & \sigma_{\text{models}} < 0.30 \\ 0.5 & \text{otherwise} \quad \text{(models disagree)} \end{cases} \tag{6}$$

where σ_{models} is the standard deviation of the three normalised model scores.

2.5.2 Motion Artefact Confidence C_{motion}

$$C_{\text{motion}} = \begin{cases} 0.4 & \texttt{motion\_max} > 0.8 \\ 0.7 & \texttt{motion\_max} > 0.6 \\ 1.0 & \text{otherwise} \end{cases} \tag{7}$$

If a SpO_2 drop co-occurs with high motion (`motion_max` > 0.6 and `spo2_slope` < -2), an additional $\times 0.5$ penalty is applied, reflecting the well-known susceptibility of pulse oximetry to motion noise.

2.6 Alert Thresholds and Suppression

Table 8: Risk level classification

Level	Condition	Response	Example
CRITICAL	$r_i > 1.5$ **and** $C_i > 0.6$	Immediate paramedic action	HR = 148, SpO_2 = 85 %, sustained 5 windows
WARNING	$r_i > 0.8$ **and** $C_i > 0.5$	Monitor and prepare	HR rising +6 bpm/window, SpO_2 borderline
NORMAL	otherwise	Routine monitoring	Stable vitals within range

2.6.1 Alert Suppression Rules

Two suppression rules prevent nuisance alerts reaching the paramedic:

1. **Motion with mild vitals.** If $C_{\text{motion}} < 0.5$ *and* $\Sigma_i < 2.0$, the alert is downgraded to NORMAL. Rationale: high vehicle motion is the most common source of false SpO_2 drops, and in the absence of corroborating vital-sign deterioration the signal is almost certainly artefactual.
2. **Low overall confidence.** If $C_i < 0.4$ *and* $r_i < 2.0$, the alert is downgraded to NORMAL. Rationale: when models strongly disagree, the result is unreliable; only very high raw scores override this rule.

2.7 Explainability

Every alert carries a `reasoning` list populated in real time. Example output for a critical event:

```
reasoning: [
  "CRITICAL SpO2: 85.3% < 88%",
  "Severe tachycardia: HR=148 > 140",
  "Multi-vital distress: HR=148 + SpO2=85.3%",
  "Sustained anomaly: 4/5 windows",
  "High model agreement"
]
```

This enables the paramedic to verify the alert clinically within seconds and decide on intervention.

3 Task 3A: Alert Quality Metrics

3.1 Problem Context

Because the anomaly detection is *unsupervised* (no labelled ground truth exists for the synthetic data), traditional precision and recall cannot be computed directly. The definitions below are therefore framed both as formal metrics (for evaluation against labelled data in future work) and as design targets that guided the current implementation.

3.2 Metric Definitions

Let TP (true positive) denote an alert on a genuinely anomalous window, FP a spurious alert on a normal window, FN a missed anomaly, and TN a correct NORMAL classification.

3.2.1 Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

Precision measures *what fraction of triggered alerts are genuine*. A low-precision system creates alarm fatigue: paramedics begin ignoring alerts because most are false, ultimately endangering patients.

Target: ≥ 0.60. In pre-hospital care, a system that is correct on 6 in 10 alerts is considered clinically useful, provided recall is high.

3.2.2 Recall (Sensitivity)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

Recall measures *what fraction of true emergencies are detected*. A missed emergency (FN) in an ambulance can be fatal.

Target: ≥ 0.95. This is a non-negotiable design requirement.

3.2.3 False Alert Rate (FAR)

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{Specificity} \tag{10}$$

FAR measures the proportion of *normal* windows that are incorrectly flagged. Unlike precision, FAR is computed over all normal windows and so is independent of the class imbalance ratio.

Target: ≤ 0.05 (5 %). The current system achieves approximately 3–5 % total alerts after risk-score calibration, consistent with this target.

3.2.4 Alert Latency

$$\text{Latency} = t_{\text{alert}} - t_{\text{onset}} \tag{11}$$

Alert latency is the time between the clinical onset of a deterioration event and the moment the system issues an alert.

Target: ≤ 15 s (one window step cycle of 5 s, plus up to two confirmation windows of 5 s each).

Persistence trade-off. Requiring 3 consecutive anomalous windows before alerting adds a latency of $3 \times 5 = 15$ s but reduces FP dramatically. This trade-off is explicit and configurable.

3.3 Error Acceptability in Ambulance Triage

Table 9: Error taxonomy for ambulance monitoring

Error Type	Description	Acceptable?	Reasoning
False Positive (FP)	Alert during normal transport	Conditionally	Paramedic can verify; tol ance threshold ≤ 5 % FAI
False Negative (FN)	Missed genuine emergency	Never	Patient may deteriora without intervention; rec ≥ 95 % required
Delayed Alert (<30 s)	Alert arrives but late	Acceptable	Short delay tolerable if eve is still evolving
Delayed Alert (>60 s)	Significantly late	Unacceptable	Window for interventi may have closed
Motion FP	High-motion road bump	Expected	Suppression logic handl flagged with low confiden
Transient FP	Brief spike, no clinical cause	Acceptable	Persistence filter suppress most; residuals are lo confidence

The guiding principle is asymmetric cost: the consequence of a missed emergency (FN) vastly outweighs the cost of a false alarm (FP). The system is therefore designed to err on the side of sensitivity, with confidence scoring and suppression logic used to *reduce* FP rather than eliminate them.

4 Task 3B: Failure Analysis

4.1 Case 1: Motion Artefact False Positive

4.1.1 Description

During a section of rough road (simulated by high motion values > 0.85), the SpO_2 sensor recorded a sustained apparent drop to 88–90 %. The system issued a WARNING alert despite the patient's HR remaining entirely stable at 78 bpm and blood pressure within normal range.

4.1.2 Why It Happened

The motion signal was high but did not exceed the suppression threshold of 0.8 for the *mean* value (only the maximum did). The SpO_2 drop was genuine in magnitude and triggered the low-SpO_2 severity term ($\delta = +1.0$). Although model confidence was reduced, it remained above the suppression floor of $C_i < 0.5$.

4.1.3 Root Cause

- The suppression rule used `motion_max` as the trigger, which can miss sustained moderate-motion episodes where the max is just below threshold.
- The system did not check whether HR remained stable during the SpO_2 drop. A genuine desaturation event would typically co-occur with at least a modest HR increase.

4.1.4 Proposed Improvements

1. Replace `motion_max` with a composite: require `motion_mean` > 0.5 **or** `motion_max` > 0.8 for full suppression.
2. Add a "HR corroboration" check: if SpO_2 drops but HR remains stable ($|\Delta HR| < 5$ bpm), apply an additional $\times 0.6$ confidence penalty.
3. Use the physiological relationship that motion artefacts typically produce SpO_2 drops without affecting HR, whereas true hypoxaemia almost always elevates HR.

4.2 Case 2: Gradual Deterioration Missed

4.2.1 Description

A patient's HR increased steadily from 78 to 115 bpm over six minutes (a slope of ≈ 0.6 bpm/window). SpO_2 simultaneously declined from 98 % to 93 %. No alert was issued because both values remained below the clinical thresholds applied in any individual window.

4.2.2 Why It Happened

Within any single 30-second window, the HR slope feature was only ≈ 0.6 bpm/window — well below the +0.5 severity trigger of 5 bpm/window. The LSTM sequence of 10 windows (50 s) did detect mild elevation in reconstruction error, but not enough to push the ensemble score above the 95^{th}-percentile threshold.

4.2.3 Root Cause

- The slope feature is computed within a 30-second window, which is insufficient to capture trends that unfold over several minutes.
- The LSTM context of 10 windows covers only 50–80 seconds, which was too short to detect a six-minute trend.
- The persistence multiplier only rewards currently anomalous windows; it does not detect a *pattern of mild worsening* that never reaches the anomaly threshold.

4.2.4 Proposed Improvements

1. Add a **long-range trend feature**: compute HR and SpO_2 slopes over a 5-minute rolling window (60 windows) and include as additional features.
2. Increase LSTM sequence length from 10 to 30 windows (covering 2.5 minutes) to capture sub-acute trends.
3. Add a **drift alert**: if HR increases by more than 15 bpm or SpO_2 decreases by more than 4 % over any 3-minute period, issue a WARNING regardless of current absolute values.

4.3 Case 3: Brief Critical Event Suppressed by Persistence Filter

4.3.1 Description

A simulated vasovagal episode caused SpO_2 to drop to 82 % for 8 seconds before recovering. The system issued no alert because only one window was flagged as anomalous, and the persistence filter requires a minimum of three consecutive anomalous windows.

4.3.2 Why It Happened

An 8-second event at a 5-second step produces at most two flagged windows (the two that overlap the event). This fails the three-window minimum. The persistence filter was designed to eliminate transient motion artefacts, but in doing so it also suppresses genuine brief but severe events.

4.3.3 Root Cause

- A single uniform persistence threshold (3 windows) was applied regardless of the *severity* of the anomaly. A window with $SpO_2 = 82\,\%$ should not require the same confirmation period as a borderline heart rate.
- The system lacked an "extreme value override" rule for cases where a single measurement is so far outside the safe range that immediate alert is warranted without waiting for confirmation.

4.3.4 Proposed Improvements

1. Introduce **severity-dependent persistence thresholds**:
    - $SpO_2 < 85\,\%$: alert on *single window* (no persistence required);

- SpO_2 85–92 % or HR > 140: require 2 consecutive windows;
- All other anomalies: retain 3-window minimum.

2. Add an **extreme value rule**: if any feature exceeds a critical physiological boundary ($SpO_2 < 85$ %, HR > 180 bpm, SBP < 70 mmHg), issue a CRITICAL alert immediately without waiting for ensemble or persistence confirmation.
3. **Rationale:** these are values incompatible with sustained consciousness and require immediate intervention regardless of model confidence.

4.4 Summary of Failure Cases

Table 10: Failure case summary and remediation

Case	Root Cause	Impact	Primary Fix
Motion Artefact FP	Suppression based on max only; no HR corroboration	Unnecessary alarm; alarm fatigue	Composite motion rule + HR corroboration check
Gradual Deterioration FN	Short trend window; LSTM too short	Missed 6-min worsening	5-min slope feature + drift alert rule
Brief Critical Event FN	Uniform persistence threshold ignores severity	Missed 8-s $SpO_2 = 82$ % episode	Severity-dependent persistence + extreme-value override

References

References

[1] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," *Proc. 8th IEEE Int. Conf. on Data Mining (ICDM)*, pp. 413–422, 2008.

[2] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[3] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long Short-Term Memory Networks for Anomaly Detection in Time Series," *Proc. European Symposium on Artificial Neural Networks (ESANN)*, 2015.