eCART and Tree-Based Algorithms

Early warning scores for inpatient deterioration

AI in Medicine elective

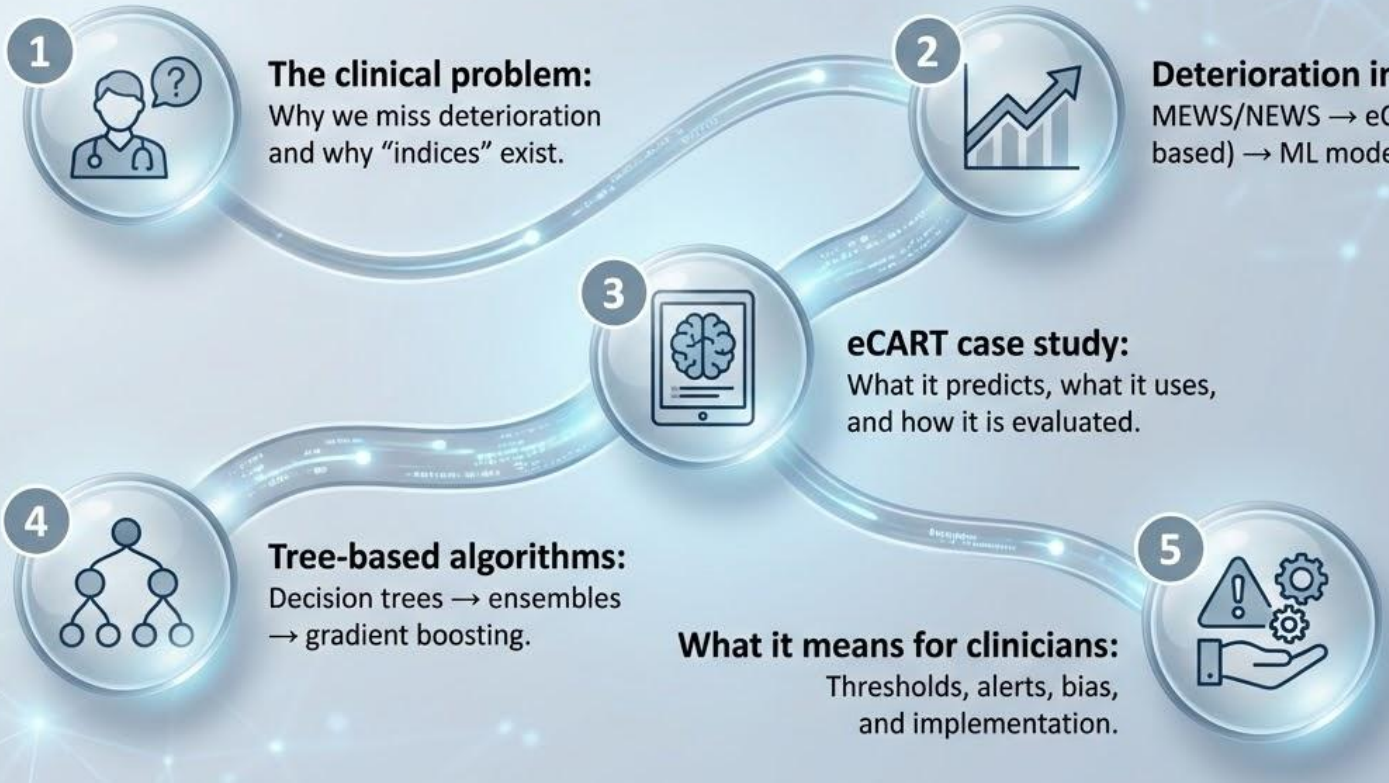
Key idea: Modern eCART versions use gradient-boosted decision trees (a tree-based ensemble).

Learning objectives

- Define “clinical deterioration” and what early warning scores are trying to predict.
- Contrast traditional scores (MEWS/NEWS) with EHR-based models like eCART.
- Explain decision trees and why ensembles (boosting) often work better.
- Interpret model outputs (risk, thresholds) and common evaluation metrics (AUROC, PPV, lead time, calibration).
- Identify practical and ethical issues: alert fatigue, fairness, prospective validation, and monitoring.

Two tracks
You can follow the lecture without math. “Optional math” boxes are for students who want more detail.

Where we are going



Part 1

The clinical problem: deterioration on the wards

What do we mean by ‘clinical deterioration’?

Unexpected ICU transfer from a general ward

Death on the ward

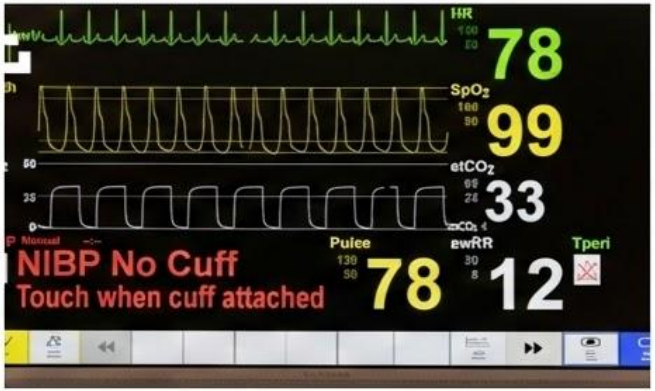
(In some scores)
In-hospital cardiac arrest or a composite outcome

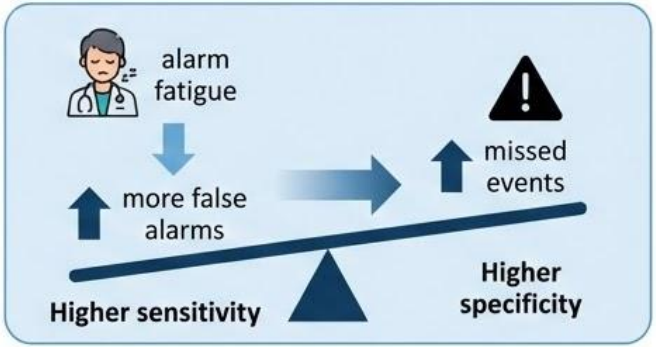
Why a 24-hour horizon?

Many tools predict whether an event will occur within the next 24 hours of an observation. This supports actionable escalation decisions and aligns with common benchmarking.

Why early warning indices exist

- Clinical deterioration occurs in a **meaningful minority** of inpatients (often ~3–5% depending on definition).
- **Delays** in escalation are common and are associated with worse outcomes (mortality, LOS).
- **Physiologic decline** is frequently detectable in vitals/labs hours before an event.
- Teams have **limited bandwidth**: the goal is earlier detection with fewer false alarms.





Deterioration indices: A quick taxonomy

Aggregated weighted scores

Examples: MEWS, NEWS/NEWS2

- Hand-calculable
- Discrete bins
- Transparent

Regression-based models

Examples: Logistic regression (often with splines)

- Needs electronic calculation
- Can model nonlinearity
- Still fairly interpretable

Machine-learning models

Examples: Gradient-boosted trees, neural nets

- Higher capacity
- Can use trends & interactions
- Harder to explain & govern

Important caveat:
More complex models are not automatically better. Data quality, implementation, and clinical workflow matter as much as the algorithm.

Traditional scores: MEWS and NEWS

MEWS (Modified Early Warning Score)

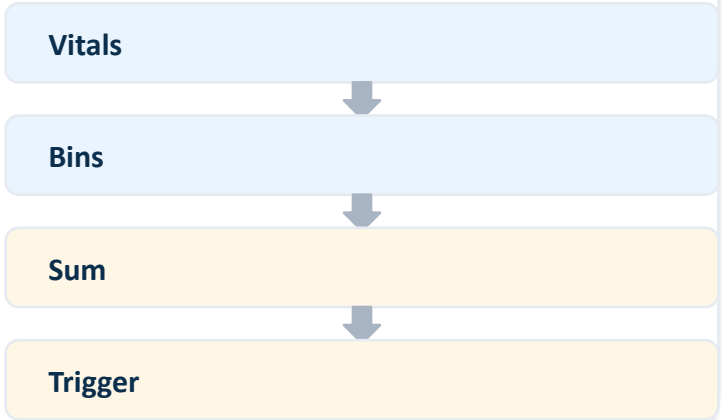
- Aggregates a small set of vital signs into a point score.
- Designed to flag patients at risk of “catastrophic deterioration”.
- Easy to teach and compute; limited nuance.

NEWS / NEWS2 (National Early Warning Score)

- Standardizes scoring/response using RR, SpO_2, O_2 supplementation, temp, SBP, HR, and consciousness.
- Widely adopted (UK NHS) as a surveillance system for inpatients and acute presentations.

How a typical bedside score works

Measure vitals → put each into bins → add points → compare to thresholds



Common limitations (and why EHR models emerged)

- Discretization loses information (e.g., RR 23 vs 24 can change points).
- Limited ability to model interactions (e.g., tachycardia “means different things” with fever vs bleeding).
- Often uses only current values (not trajectories), despite trends being clinically meaningful.
- Workflow and data capture issues: RR is often measured inaccurately; missingness is common.
- False alarms can contribute to alarm fatigue and reduced trust in alerts.

Takeaway
Simple scores are valuable baselines. When we have high-quality EHR data, we can often do better by modeling continuous variables, trends, and interactions.

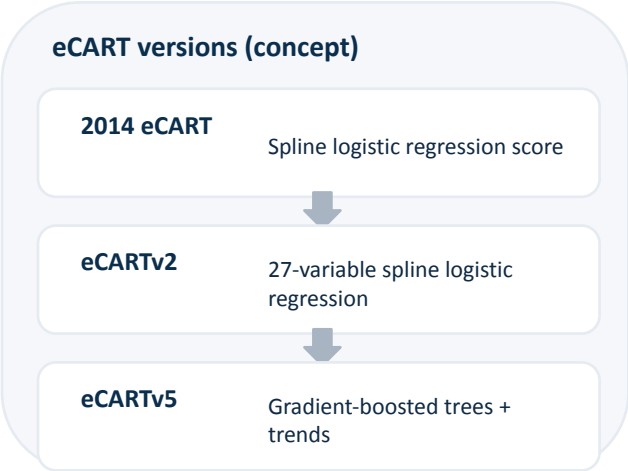
Part 2

eCART as a case study (and why “versions” matter)

What is eCART?

eCART = electronic Cardiac Arrest Risk Triage score

- Designed for adult patients on general medical–surgical wards.
- Uses routinely collected EHR data (vitals ± labs ± documentation).
- Early versions targeted cardiac arrest, ICU transfer, and ward death (often within 24 hours).
- Later versions focus on ICU transfer or death within 24 hours and incorporate trends + ML.



Teaching note
"eCART" is a family of models. When you read a paper or see a hospital implementation, confirm which version and outcome definition are being used.

What goes in, what comes out

Inputs (examples)

- Demographics (e.g., age)
- Vital signs (RR, HR, SBP, SpO_2, Temp, etc.)
- Labs (when available)
- Documentation signals (e.g., mental status via AVPU)
- Trends over time (e.g., max RR over prior 24h)

Output

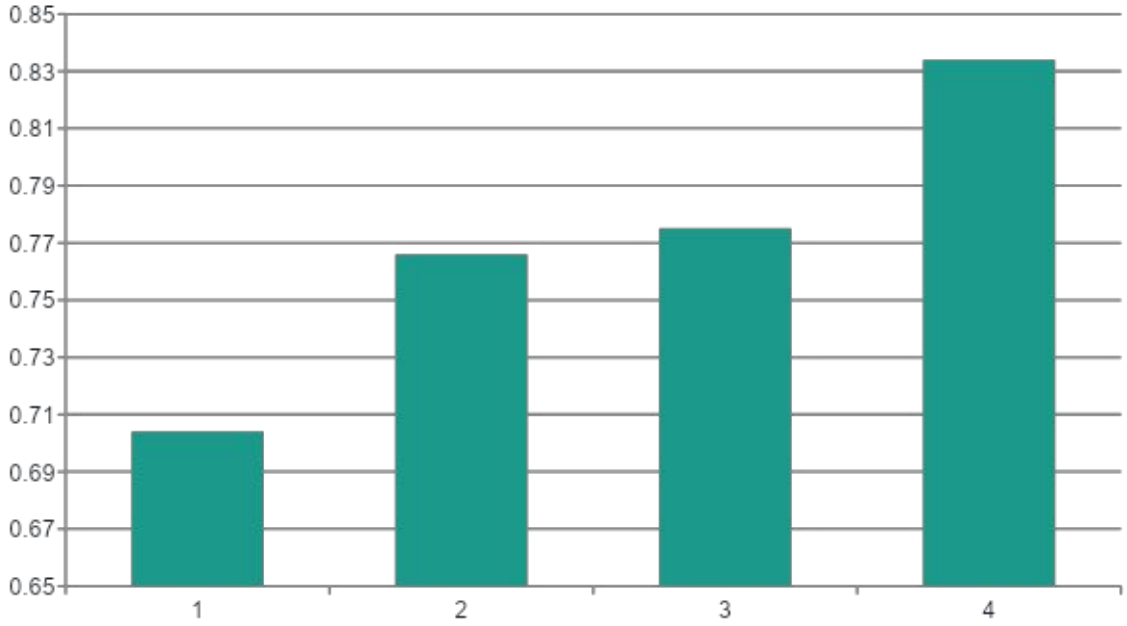
A risk estimate (score or probability):
“ICU transfer or death within 24 hours”

Operational step: thresholds

- Moderate-risk trigger (more sensitive)
- High-risk trigger (more specific)
- Threshold choice should match staffing & workflow

Performance snapshot: eCARTv5 vs common baselines

External retrospective validation (21 hospitals): AUROC for ICU transfer or death within 24h



Interpretation
AUROC summarizes rank-order discrimination, but bedside usefulness also depends on thresholds, PPV, lead time, calibration, and workflow.

Prospective validation
eCARTv5 maintained AUROC ≥ 0.80 in prospective external validation in the same study.

Head-to-head comparisons: AUROC is not the whole story

- In a large comparison of 6 early warning scores, performance varied widely.
- Even “good” AUROC values can translate into low PPV when the event rate is low.
- Matching thresholds on sensitivity or specificity helps compare tools fairly.
- Lead time matters: detecting earlier is only helpful if it changes care.

Back-of-the-envelope PPV

If 5% deteriorate, then even a “high-risk” alert stream will include many false positives unless specificity is very high.

Clinical implication

- Design tiered responses
- Measure alert burden
- Monitor downstream actions and outcomes

Part 3

Tree-based algorithms: from intuition to gradient boosting

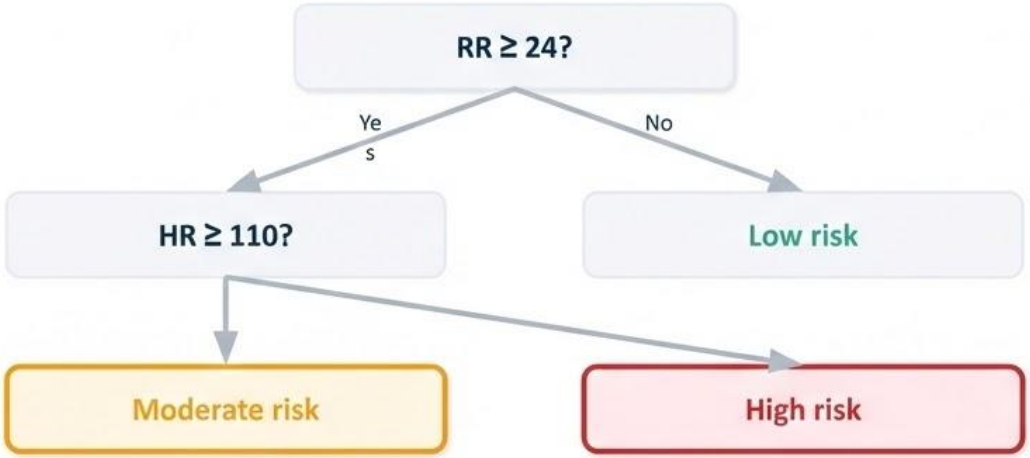
Why decision trees show up in clinical prediction

- They naturally represent “if–then” logic and non-linear effects.
- They can capture interactions without explicitly writing interaction terms.
- They handle mixed data types (continuous + categorical).
- Tree ensembles (random forests, gradient boosting) often perform well on tabular EHR data.
- With appropriate tools, you can extract feature importance and partial dependence for interpretability.

Key point
A single tree is easy to understand but can be unstable. Ensembles trade some interpretability for better accuracy and robustness.

Decision trees: the intuition

A decision tree repeatedly asks a yes/no question that best splits patients into more “pure” risk groups.



Clinical interpretation:
This toy example is not “eCART.” Real models use many variables, continuous thresholds, and repeated splits.

Strength

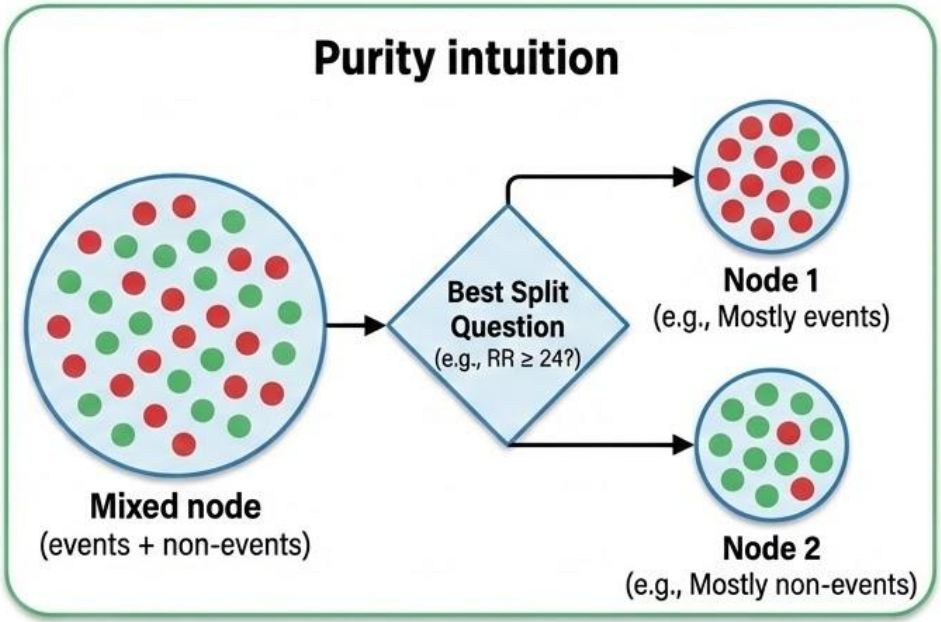
- Transparent rule structure; aligns with many clinical heuristics.

Weakness

- Small data changes can produce a different tree (high variance).

How does a tree choose a split? (high level)

- At each node, the algorithm tests candidate questions like “RR ≥ 24?”
- It picks the split that best separates outcomes (e.g., event vs no event).
- Common criteria: Gini impurity, entropy (classification) or variance reduction (regression).
- The process repeats until stopping rules are met (e.g., max depth, min samples per leaf).

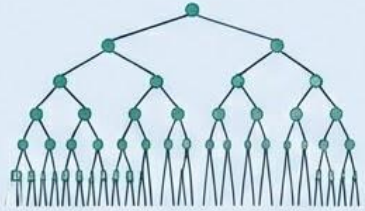


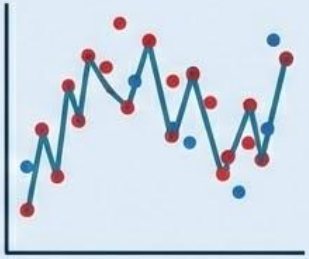
Optional math

Gini impurity for a node: $1 - \sum p(k)^2$. A good split reduces impurity (more homogeneous children).

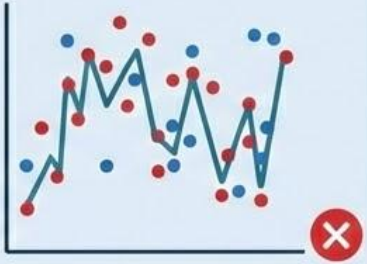
A single tree can overfit

The Problem: **Overfitting (High Variance)**

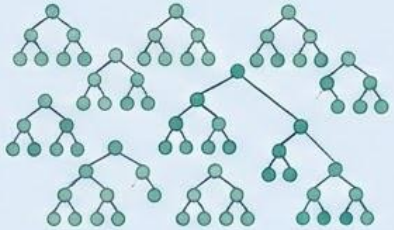


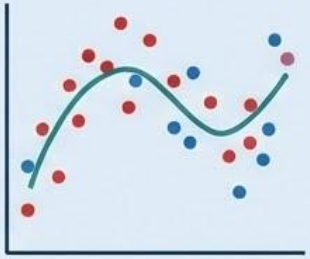


- Deep trees memorize noise and outliers.
- Perfect performance on training data (0% error).
- Poor performance on new, unseen data (high test error).

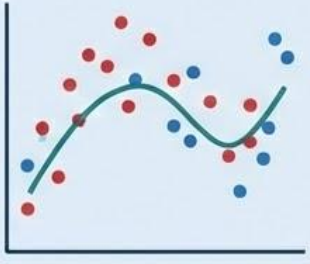


The Goal: **Generalization (via Ensembles)**





- Combine many “weak” trees to smooth out noise.
- Good, but not perfect, on training data.
- Better performance and robustness on new data.

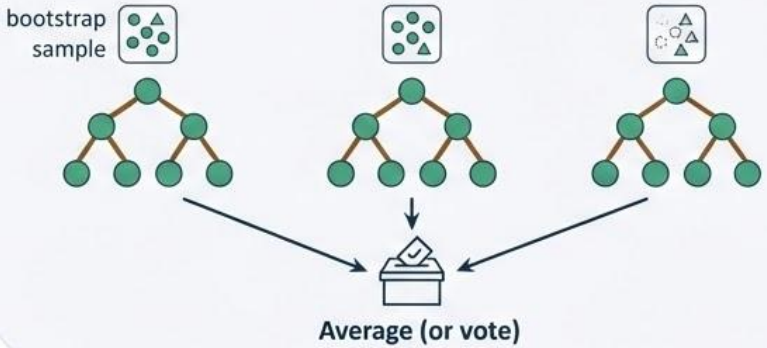


Clinical Analogy: A rule that perfectly fits one hospital’s unique patient population and workflows (like a deep tree) will likely fail in a different hospital. This is why external and prospective validation is crucial. Ensembles aim to learn general, transferable patterns.

Ensembles: two common patterns

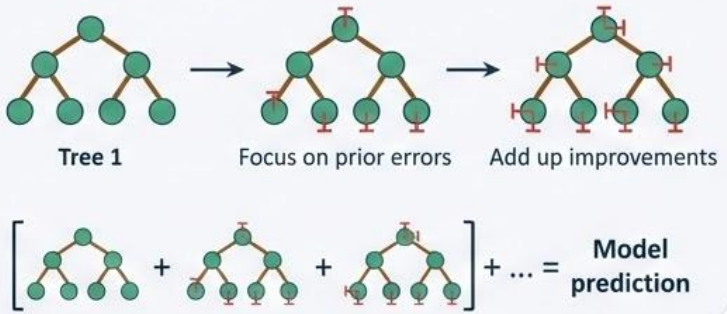
Bagging (Random Forest)

- Train many trees in parallel
- Each tree sees a bootstrap sample
- Average (or vote) the predictions
- Main effect: reduces variance



Boosting (Gradient Boosting)

- Train trees sequentially
- Each new tree focuses on prior errors
- Add up many small improvements
- Main effect: reduces bias (often with some variance control)



Gradient boosting in plain language

1. Start with a baseline guess

Example: the average event rate (or log-odds).

2. Measure the errors

Who did we under-predict vs over-predict? (residuals)

3. Fit a small tree to the errors

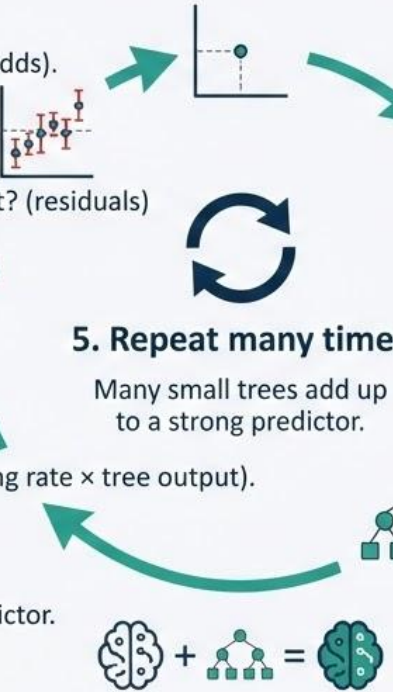
The tree learns patterns in the mistakes.

4. Update the model

New prediction = old prediction + (learning rate × tree output).

5. Repeat many times

Many small trees add up to a strong predictor.



5. Repeat many times

Many small trees add up to a strong predictor.

Optional math

"Gradient" refers to optimizing a loss function by moving in the direction that reduces error.
For classification, a common loss is log-loss.

Practical knobs

Number of trees, max depth, learning rate, subsampling.
These control overfitting and performance.

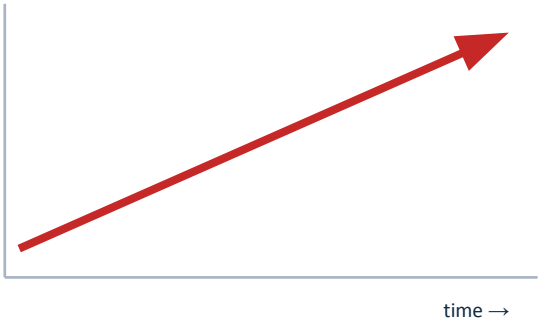
eCARTv5: gradient-boosted trees + trends

- eCARTv5 is a gradient-boosted trees model trained on large multicenter ward datasets.
- Predictors include demographics, vital signs, documentation, and lab values.
- A key design choice is incorporating trends (e.g., prior 24h min/max/changes), which often improves discrimination.
- Evaluated across many patient subgroups and validated retrospectively and prospectively.

Why trends help

Clinicians rarely react to a single value in isolation. A rising RR or dropping SBP is often more concerning than a stable value.

Example: RR trend



Interpretability: what can we explain in boosted trees?

- Global importance: which variables the model uses most (aggregate view).
- Local explanations: why this patient is high-risk right now (e.g., SHAP-style contributions).
- Partial dependence: how risk changes as one variable changes (holding others fixed).
- Caution: explanations can be misleading when variables are correlated or missingness is informative.

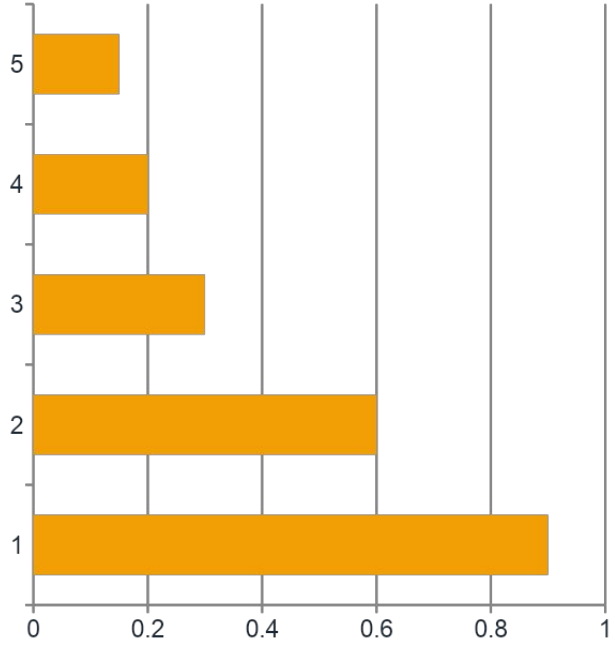


Illustration only (not an eCART output)

Fairness and subgroup performance

- A model can look “good overall” but fail in specific subgroups.
- Subgroup checks commonly include: age, sex, race/ethnicity, comorbidities, surgical vs medical populations.
- Evaluate both discrimination (AUROC) and calibration (are predicted risks accurate?).
- Monitor over time: data drift and care process changes can degrade performance.

What eCARTv5 reported

Performance remained high (AUROC ≥ 0.80) across a range of demographics and clinical conditions, including during prospective validation.

If you deploy locally

Re-check subgroup performance in your own system and align thresholds to your resources and patient mix.

Implementation: what matters beyond the algorithm

Operational

- Data plumbing: when is each variable available (latency)?
- Missingness: is it random or informative?
- Alerting strategy: tiered thresholds, routing, escalation paths.
- Human factors: alarm fatigue, trust, and workflow fit.
- Evaluation after go-live: alert burden, response times, outcomes.

Clinical / governance

- Prospective validation beats retrospective simulations.
- Monitor drift, recalibrate when needed.
- Document intended use + limitations.
- Consider regulatory status for clinical decision support tools.
- Governance: who “owns” the model lifecycle?

Evidence level and regulation (high-level)

- Many deterioration scores are validated retrospectively; fewer are prospectively evaluated.
- For high-stakes clinical decision support, evidence should include external and prospective validation.
- Some early warning scores have been cleared by the U.S. FDA as medical devices; eCARTv5 is reported as FDA-cleared in peer-reviewed work.
- Local implementation still requires local validation and governance.

For students
When you see an “AI deterioration index” in a hospital, ask: What outcome? What time horizon? What validation? What thresholding? What monitoring?

Case discussion: choosing a trigger

Your hospital is considering a ward deterioration alert. You can pick ONE of these implementations:

Option A: Low threshold

High sensitivity (catches more events)
…but generates many alerts per day.

Option B: High threshold

High specificity (fewer alerts)
…but misses more events.

Option C: Tiered thresholds

Moderate-risk → nurse review
High-risk → rapid response evaluation

Prompt
Which option would you choose and why? What additional information would you request before go-live?

Key takeaways

- Deterioration indices exist to detect early physiologic decline with enough lead time to change care.
- eCART is a family of models: early versions used spline logistic regression; newer versions (eCARTv5) use gradient-boosted trees with trends.
- Tree ensembles are powerful for tabular EHR data, but require careful validation and governance.
- Model value is not just AUROC—thresholds, PPV, lead time, calibration, and workflow integration drive clinical impact.
- Responsible deployment includes prospective evaluation, subgroup testing, monitoring for drift, and transparency about limitations.

Selected references (open access when possible)

- Churpek MM et al. Multicenter Development and Validation of a Risk Stratification Tool for Ward Patients (eCART). Am J Respir Crit Care Med. 2014. (PMC)
- Churpek MM et al. Multicenter Development and Prospective Validation of eCARTv5: A Gradient-Boosted Machine-Learning Early Warning Score. Crit Care Explor. 2025. (PMC)
- Edelson DP et al. Early Warning Scores With and Without Artificial Intelligence. JAMA Netw Open. 2024. (Open access)
- Gerry S et al. Early warning scores for detecting deterioration in adult hospital patients: systematic review. BMJ. 2020. (PMC)
- Royal College of Physicians. National Early Warning Score (NEWS2) report. 2017.
- Subbe CP et al. Validation of a modified Early Warning Score (MEWS). QJM. 2001.
- Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001.