Consensus-Based Guardrails Improve Robustness of Deep Learning Driven Respiratory Phase Prediction for Gated Radiotherapy

Omar El-Sherif, PhD; Zachary S. Stottler, PhD; Jing Qian, PhD; Amanda J. Deisher, PhD; and Erik J. Tryggestad, PhD
Mayo Clinic, Rochester, MN

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Purpose
To evaluate consensus-based guardrails for improving the robustness of an in-house deep learning Long Short-Term Memory (LSTM) respiratory phase prediction model used for respiratory-gated radiotherapy. Consensus filter configurations were compared based on their impact on gate signal ringing, beam-on/off timing errors, and treatment delivery efficiency.

Methods
Ten anonymized patient respiratory traces were analyzed. Ground-truth respiratory phases were determined via expert retrospective peak identification and compared with predictions from an in-house LSTM model generating phase forecasts spanning 33–333ms. Two consensus-based guardrails were implemented, each requiring 2-of-3 agreement to change gate state. The Staggered method formed consensus across sequential predictions, requiring new amplitude data between votes, while the Convergent method combined three predictions converging to a single future time point (99ms ahead). Each method was evaluated independently and in combination with a 200ms dead-time filter applied to the gate signal. Simulated gating signals derived from ground truth were compared with consensus-filtered signals to quantify ringing prevalence, gate timing errors, and treatment delivery efficiency, assessed via duty cycle for a 20–80% phase-gating window.

Results
Both consensus methods significantly reduced gate signal ringing. Compared to 2.2% ringing without guardrails, ringing decreased to 1.2% with the Convergent method and 0.4% with the Staggered method, with further reductions to 0.2% and 0.1% following dead-time hybridization. Treatment delivery efficiency was minimally affected, with duty cycles of 59.6% (Convergent) and 60.5% (Staggered), compared to the expected 60%. Mean gate-ON timing errors increased modestly to approximately 10–11ms, and gate-OFF timing errors increased to approximately 51ms; all delays were clinically insignificant.

Conclusion
Consensus guardrails reduce gate signal ringing in respiratory phase prediction with negligible impact on delivery efficiency or timing accuracy. Dead-time provides additional robustness, offering a practical and generalizable strategy to improve the clinical reliability of deep learning–based respiratory gating systems.

Contact Info:
Omar El-Sherif, PhD
Mayo Clinic
200 First St, Rochester MN
Sherif.Omar@mayo.edu

INTRODUCTION

Deep learning is a novel approach to respiratory phase prediction for motion management, offering multi-horizon predictions that can be leveraged to improve gating stability. Our LSTM model inference scheme employs quasi-random dynamic annealing, intentionally producing nondeterministic outputs to avoid overfitting. We exploit this feature combined with multi-horizon predictions to establish statistical consensus before making final gate decisions.

To our knowledge, this is the first study to apply multiple predictions for improving respiratory gating system robustness in this context.

Motivation . Clinically, rapid Gate ON/OFF oscillations can potentially trigger delivery system pauses or aborts, significantly impacting treatment efficiency and prolonging the radiotherapy session. Longer treatments can lead to increased patient discomfort and likelihood of unwanted movement. Consensus-based guardrails effectively suppress these oscillations while preserving gating performance, offering a practical and simple solution to improve clinical reliability and protect delivery hardware.

OBJECTIVES

- Introduce two consensus-based guard rails on phase outputs.
- Compare system performance with and without the guard rail feature.
- Compare system performance with and without the guard rail feature + Dead-Time.

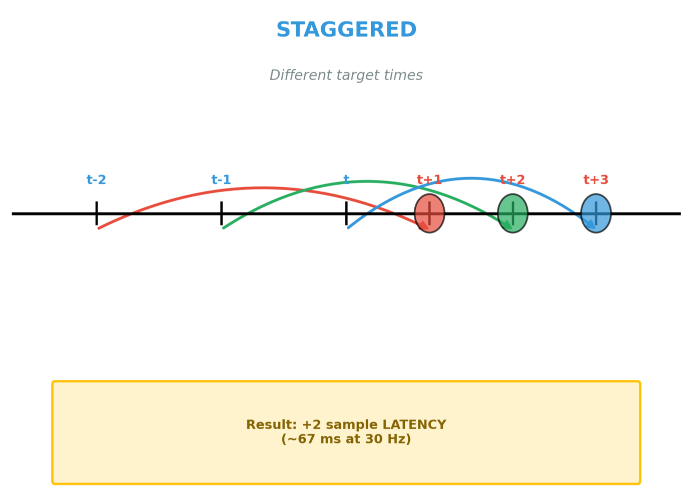
METHODS

Data: 10 anonymized patient respiratory amplitude traces

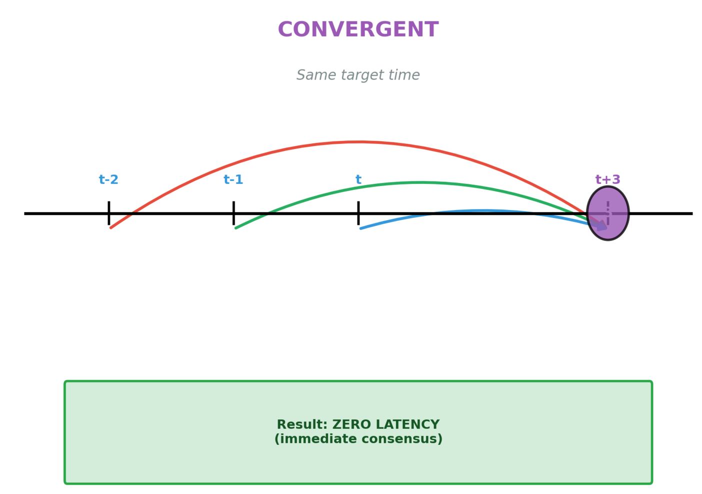
Ground truth: Manual retrospective peak identification

PhaseRx: In-house LSTM-based phase prediction model

Consensus Guardrails
- 2 of 3 voting required to change gate state
- Staggered consensus
  - Sequential predictions



- Convergent Consensus
  - Three predictions converging to a single future time point



Dead-Time Hybridization
- Optional 200 ms dead-time applied to gate signal
- Evaluated with and without dead-time

Evaluation:
- Gate signal ringing (rapid ON/OFF oscillations)
- Beam-on and beam-off timing errors vs ground truth
- Treatment delivery efficiency via duty cycle
  - Assessed for a 20–80% phase-gating window

RESULTS

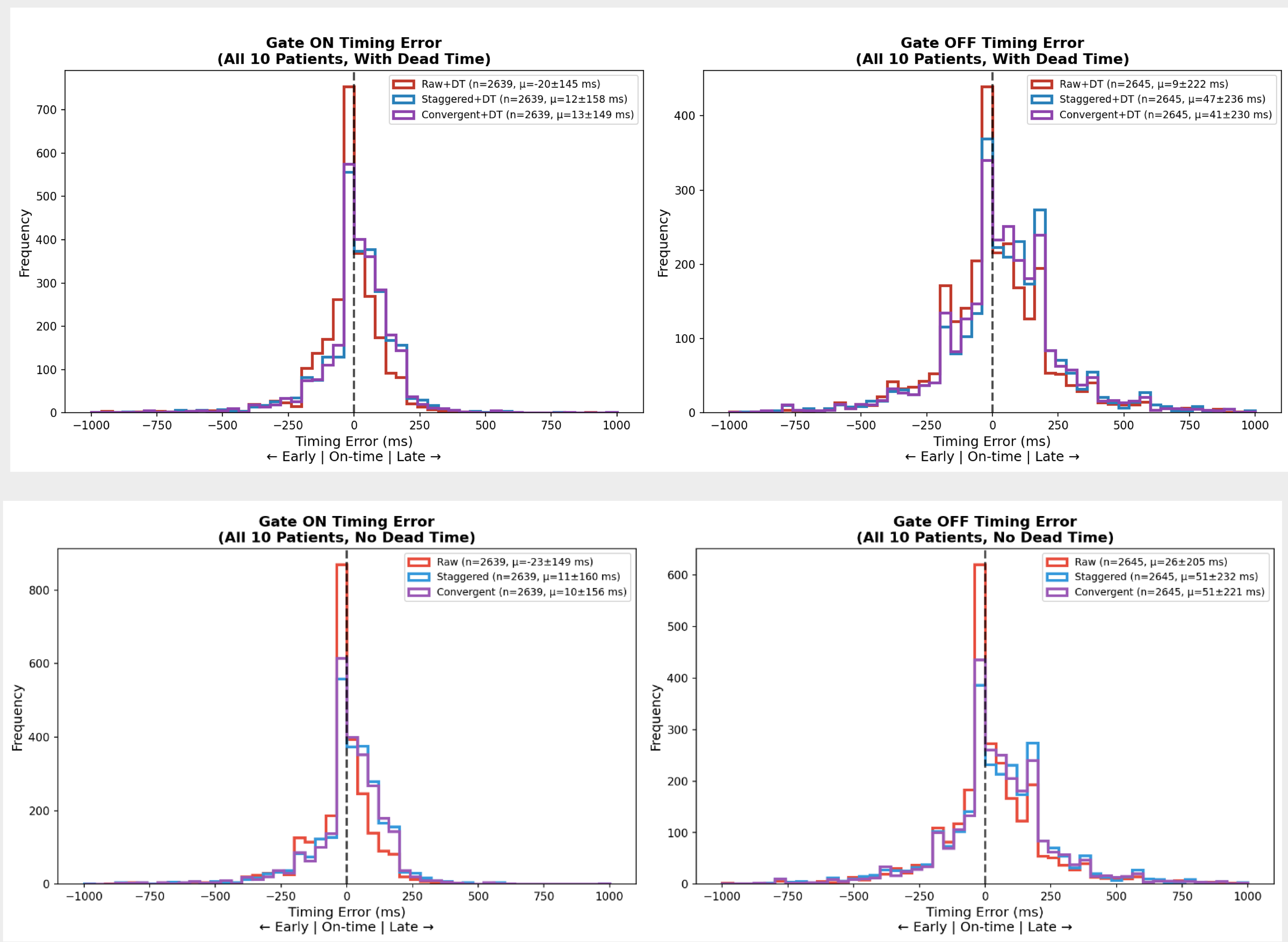


Figure 1. Distribution of gate timing errors for Raw, Staggered Consensus, and Convergent Consensus methods with and without dead time enforcement. Timing error is calculated as the difference between predicted and ground truth gate transition times (negative = early, zero = on-time, positive = late). Data aggregated across all 10 patients (n=2639 Gate ON transitions, n=2645 Gate OFF transitions). Vertical dashed line indicates zero timing error (on-time). Raw prediction shows a symmetric distribution centered near zero, while both consensus methods shift the distribution toward positive (late) timing due to the 2-of-3 voting requirement. Gate ON timing error (with deadtime) for Raw, Staggered, and Convergent are -20 ± 145 ms, 12 ± 158 ms, 13 ± 149 ms (mean ± SD) respectively. Gate OFF timing error for Raw, Staggered, and Convergent are 9 ± 222 ms, 47 ± 236 ms, 41 ± 230 ms (mean ± SD) respectively. Gate ON timing error (without deadtime) for Raw, Staggered, and Convergent are -23 ± 149 ms, 11 ± 160 ms, 10 ± 156 ms (mean ± SD) respectively. Gate OFF timing error for Raw, Staggered, and Convergent are 26 ± 205 ms, 51 ± 232 ms, 51 ± 221 ms (mean ± SD) respectively.

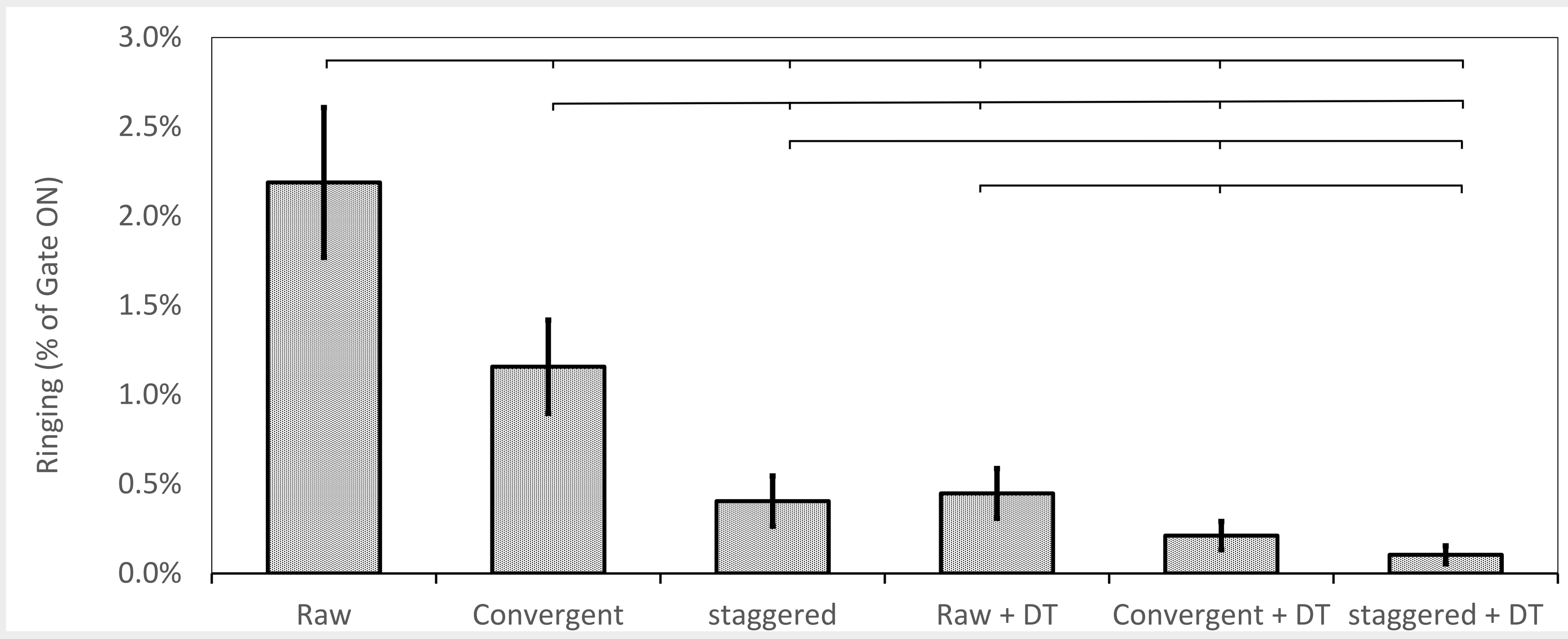


Figure 2. Percentage of gate-ON ringing events for raw LSTM gating and consensus-filtered methods, with and without 200-ms dead-time (DT). Consensus filtering markedly reduces ringing, with the lowest levels observed for the Staggered + DT configuration. Error bars indicate standard deviation (n = 10) and horizontal bars indicate statistical significance ($p < 0.05$).

CONCLUSIONS

Consensus guardrails reduce gate signal ringing in respiratory phase prediction with negligible impact on delivery efficiency or timing accuracy. Dead-time provides additional robustness, offering a practical and generalizable strategy to improve the clinical reliability of deep learning–based respiratory gating systems.