Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

SAFE-Loss: Specialized Aligned Fusion of Experts for First-Trimester Pregnancy Loss Prediction

***Mr. A. V. M. Kumaran, Dr. M. Sundara Rajan**

Research Scholar, PG & Research Department of Computer Science, Government Arts College (Autonomous), Nandanam, Chennai – 600 035

Associate Professor, PG & Research Department of Computer Science, Government Arts College (Autonomous), Nandanam, Chennai – 600 035.

Abstract — Background. Multimodal clinical prediction is conventionally approached by engineering domain-specific features, assigning specialist encoders to different views of the patient, and combining those encoders with a learned per-patient gate. The premise — that modality-specific specialisation recovers signal a single joint model would miss — is widely adopted and rarely tested against its own counterfactual.

Objective. To build such a system for first-trimester pregnancy-loss prediction, to evaluate it beyond discrimination alone, and to determine by controlled ablation where its accuracy actually comes from.

Methods. We developed CAFE (Clinically-Aligned Fusion of Experts): 43 engineered features in five clinically motivated families — temporal trajectory, growth velocity, inter-marker discordance, maternal risk profile and cross-modal interaction, all built on gestational-age normalisation — tree experts over a flat view and a cross-modal view, an optional 12,585-parameter temporal-attention encoder over the raw scan sequence, and a per-patient softmax gate over expert logits. Evaluation used a synthetic dual-etiology cohort (n = 3000, prevalence 20.3 %) under 5-fold cross-validation repeated twice. Every arm, hybrid and baseline alike, received an identical 66-feature union view, so no result can be attributed to an information advantage. Reported measures were AUROC, AUPRC, Brier score, Cox calibration slope and intercept, expected calibration error, decision-curve analysis, and a screening operating point selected on one repeat and scored on the other. A difference was declared real only if it exceeded twice the pooled fold standard deviation and agreed in sign on at least 8 of 10 folds.

Results. CAFE reached AUROC 0.852, the highest point estimate we recorded on this cohort and inside the 0.840–0.855 band predicted for it in advance. It was also the best-calibrated arm we measured (slope 0.946, expected calibration error 0.030, no post-hoc correction), returned the highest mean net benefit across risk thresholds 0.05–0.50, and trained in 43 s against 269 s for the deep temporal architecture it replaces.

Ablation. Its accuracy does not originate in its architecture. Replacing the gated multi-view ensemble with a single gradient-boosted tree on the same features moved AUROC by 0.001. The +0.030 total improvement over logistic regression was dominated by two of its four steps: +0.021 for the choice of boosting library and +0.015 for the engineered features — the latter only for a learner unable to represent interactions natively, since the same features were worth +0.002 to the deployed learner. The gated fusion contributed +0.001 and the sequence encoder −0.005; the gate independently learned to discount that encoder, assigning it 0.195 against an equal share of 0.333.

Interpretability. Feature importance and feature contribution pointed in opposite directions: out-of-fold exact TreeSHAP assigned the engineered block 66.8 % of total attribution mass, while ablation valued that block at 0.002 AUROC for the deployed model. The gate proved confident, patient-specific and clinically legible yet inconsequential — near-bimodal weights (mean 0.557, SD 0.402, spanning 0–1) routing abnormal yolk-sac measurements to the cross-modal expert and prominent maternal risk to the flat expert, for +0.001.

Conclusions. The arms separated in exactly one dimension, and it was not one the architecture gave any reason to examine: calibration. The XGBoost-derived arms had calibration slopes of 0.430–0.480 against 0.923–0.946

for the CatBoost-derived arms, with non-overlapping per-fold distributions, while their AUROCs differed by 0.021 and their Brier scores tied. The 90 %-sensitivity thresholds consequently spanned 0.004 to 0.081 across arms, so any workflow with a hard-coded risk threshold is silently coupled to the calibration of one particular model. CAFE is worth deploying; the explanation for why it performs as it does is not the one its architecture suggests.

Index Terms *— multimodal fusion, mixture of experts, clinical feature engineering, model calibration, decision-curve analysis, SHAP, ablation, early pregnancy loss, synthetic clinical benchmarks*

1. INTRODUCTION

1.1 The problem, and the shape of the data

Approximately one in five clinically recognised pregnancies ends in first-trimester loss. The evidence available to a clinician at that stage is inherently multimodal and asymmetric: a short, irregularly spaced series of ultrasound examinations, each yielding a handful of biometric measurements, set against a fixed panel of maternal clinical covariates recorded once at booking.

Both halves of that description are awkward for modern modelling. The trajectory is far too short for the sequence models that dominate longitudinal healthcare work — two to five visits, not hundreds of time steps — and the covariate panel is too small for representation learning to have much to discover. It is an unglamorous regime, and it is the one that most first-trimester prediction problems actually occupy.

1.2 The prevailing architectural response

The dominant modelling response has been architectural. A temporal encoder is assigned to the scan trajectory, a separate encoder to the maternal covariates, and a learned fusion mechanism — concatenation, cross-attention, or a gate — combines them. The implicit premise is that modality-specific specialisation recovers signal a single joint model would miss.

The premise is plausible, and it is the reason such systems get built. It is also, in the published literature, seldom accompanied by the measurement that would establish it: the same information, given whole to one ordinary model.

1.3 What this paper delivers

We built the hybrid this line of reasoning recommends, in the specific form usually proposed for it: five families of clinically aligned engineered features, three encoders operating on different *views* of the same patient — a flat tabular view, a sequence view, and an interaction view — and a per-patient gate that learns how much to trust each encoder for each individual patient.

It works. It reaches AUROC 0.852 on our evaluation cohort — the highest point estimate we have recorded there — and trains in well under a minute per fold. It is well calibrated with no post-hoc correction. Its predictions decompose into per-patient feature attributions a clinician can read, and its gate emits a per-patient trust score for each encoder.

And its accuracy does not come from its architecture. Holding everything else fixed and removing the fusion mechanism — replacing the gated ensemble of view-specific experts with a single gradient-boosted tree trained on the union of the same features — moves the score by 0.001 AUROC, well inside noise. The engineered features do contribute, but overwhelmingly to *weak* learners: they lift logistic regression by +0.015 and CatBoost by +0.002. The single largest contributor to the final number is neither the features nor the fusion but the choice of boosting library underneath, worth +0.021 on its own. The sequence encoder, given a fair budget and a fair gate, contributes −0.005.

We report this for two reasons. Presenting 0.852 as evidence that gated multi-view fusion produced it would be a specific, checkable, and false causal claim. And a system whose active ingredient has been misidentified cannot subsequently be improved, because every effort to improve it will be aimed at the wrong part. The engineering is sound and worth deploying; the explanation for why it performs as it does needs to be the correct one.

1.4 Contributions

1) **CAFE**, a complete and reproducible pipeline for first-trimester loss prediction: 43 engineered clinical features in five families, two or three view-specific experts, and a per-patient softmax gate, reaching AUROC 0.852 at a training cost of seconds rather than minutes.

2) **A clinically grounded feature set with its contribution quantified per learner.** Rather than reporting that engineered features "help", we measure how much they help each of four learners and show that the benefit falls monotonically as the learner's native capacity to represent interactions rises — evidence that the features re-express existing signal rather than adding new signal.

3) **An evaluation that goes beyond discrimination**: Brier score, Cox calibration slope and intercept, expected calibration error, decision-curve analysis against treat-all and treat-none, and a screening operating point whose threshold is selected on one cross-validation repeat and scored on the other.

4) **A full interpretability stack**: exact out-of-fold TreeSHAP attributions, so that every patient is explained by a model that never trained on them; the gate's own per-patient trust weights; what those weights respond to; and worked high- and low-risk cases.

5) **A component-level ablation** decomposing the final AUROC into the contributions of the features, the boosting component, the gating architecture, and the sequence encoder. This is the central result of the paper.

6) **A cost analysis measured on the same machine as the accuracy numbers,** so the accuracy-per-second trade-off is stated rather than assumed.

1.5 On novelty claims — what we do and do not assert

The design evaluated here was proposed to us as delivering several "first-ever" results for early pregnancy loss. We deliberately make no priority claims, and we give our two reasons here rather than deferring them.

First, we cannot verify priority. Establishing that no prior work has engineered discordance features or applied a gate to this outcome would require an exhaustive search we have not performed, and a reviewer who knows of a single counterexample is entitled to discard the claim — and to distrust the rest of the paper along with it. The contributions above are stated as things we did and measured, not as things nobody has done before.

Second, and more fundamentally, our contribution is methodological rather than clinical. What we offer is a pipeline, an evaluation protocol, and an ablation that identifies which parts of that pipeline carry its performance. Whether the engineered features have clinical value is a different question, and one we make no attempt to answer.

2. Related Work

2.1 Engineered features against learned representations on tabular clinical data

That gradient-boosted trees remain competitive with, and frequently superior to, deep architectures on tabular data is now well established [1, 2]. The mechanisms proposed — sensitivity to feature rotation, robustness to uninformative features, and an inductive bias towards axis-aligned splits — all apply directly to the setting here, where the patient representation is a few dozen derived scalars.

The implication for feature engineering is less often stated. A learner that already discovers multiplicative and threshold interactions natively should benefit *less* from having those interactions supplied explicitly than a learner that cannot. We test that implication directly in §4.2 and use it as the interpretive key to the entire feature-engineering result.

2.2 Gated fusion and mixtures of experts

Per-patient gating descends from the mixture-of-experts formulation [3], in which a gating network learns an input-dependent distribution over specialist predictors, and from its modern sparse variants [4]. In clinical multimodal work the gate is usually motivated by *heterogeneity*: different patients have different data quality, or their risk is driven by different mechanisms, and a fixed combination rule cannot adapt to that.

The motivation is sound and the mechanism is real. What is rarely reported is the counterfactual — the same experts combined by a fixed rule, or dispensed with entirely in favour of a single joint model — which is precisely what determines whether the adaptivity was worth its cost. We therefore report the gate's *behaviour* and its *benefit* as two separate measurements (§4.8), because they turn out to have different answers.

2.3 Evaluation beyond discrimination

A model's AUROC describes only its ranking of patients. Two models with identical AUROC can differ substantially in whether their stated probabilities are trustworthy, and a well-ranking, badly calibrated model is actively unsafe when its output is used as a risk figure in a consultation [5]. Decision-curve analysis [6] adds the further question of whether acting on the model beats the default strategies of treating everyone or no one, across the range of risk thresholds a clinician might plausibly adopt. Frameworks for assessing prediction models have long recommended reporting discrimination, calibration and clinical usefulness together rather than singly [7].

Because these are distinct properties, a tie in discrimination does not imply a tie in clinical utility. That is why we measured them here rather than assuming the AUROC result had settled the matter — and, as §4.4 shows, it is the reason this study reached a conclusion its discrimination results alone could not have supported.

2.4 Early pregnancy-loss prediction

Prediction of first-trimester loss from early sonographic biometry and maternal history has been approached both with classical risk scoring and, more recently, with deep models operating on ultrasound recordings [8].

We do not compare our AUROC against figures reported on other cohorts. Prevalence, case mix and measurement protocol differ, and a numerical comparison across cohorts would be uninterpretable regardless of which direction it happened to favour. Published work enters this paper as a source of *designs*, not of *results*.

3. METHODS

3.1 Study design

The study proceeded in four pre-planned phases, each answering one question and each reported separately below:

1) **Phase 1 — Features.** Do clinically aligned engineered features improve prediction, and for which learners? (§3.3, §4.2)

2) **Phase 2 — Architecture.** Do view-specific experts combined by a per-patient gate improve on a single joint model given the same features? (§3.4, §4.1, §4.3)

3) **Phase 3 — Clinical evaluation.** Beyond discrimination, do the arms differ in calibration, net benefit, or behaviour at a screening operating point? (§3.6, §4.4–§4.6)

4) **Phase 4 — Interpretability.** What does the model rely on, what does the gate do, and does either correspond to what the components contributed? (§3.8, §4.7–§4.8)

Phases 1 and 2 are ablations under a fixed evaluation protocol; Phases 3 and 4 characterise the resulting system. The decomposition in §4.10 assembles Phases 1 and 2 into a single attribution of the final score.

3.2 Cohort

3.2.1 Provenance and difficulty targeting

All data are synthetic, **No real patient data were used at any stage.** The evaluation cohort (n = 3000, prevalence 20.3 %) was generated from a single seeded command with frozen parameters and is reproducible bit-for-bit.

Its difficulty was targeted before any of the models in this paper existed, against a baseline that is not one of the models under test (HistGradientBoosting), to a band fixed in advance ([0.83, 0.87]; achieved 0.836 ± 0.013). This ordering matters: targeting difficulty against a proposed model allows the task to be steered towards it, whereas using an external baseline prevents that.

3.2.2 Modalities and measurement model

The scan modality comprises two to five visits per patient (mean 3.05), irregularly spaced at 1–4 gestational-week intervals, with four measurements per visit: crown–rump length (CRL), mean gestational-sac diameter (GS), yolk-sac diameter (YSD), and fetal heart rate (FHR). The maternal modality comprises static clinical covariates: age, BMI, gravidity, parity, previous loss, conception route, and plurality.

Marker trajectories are anchored to published first-trimester reference values across gestational weeks 5–10, with marker-specific measurement error mimicking real ultrasound precision (CRL ± 1.6 mm, GS ± 3.2 mm, FHR ± 6.5 bpm, YSD ± 0.55 mm) and a label-independent constitutional-size factor correlating the size markers within a patient. Losses are expressed as clinically recognised abnormality patterns — small CRL and sac, bradycardia, and an enlarged yolk sac — each at 0.90–0.94 probability, so that no single marker is deterministic.

3.2.3 Outcome structure

The outcome has explicit causal structure: label = label_sono OR label_mat, verified at 1.0000 across all 3000 patients, with two channels coupled by design, through a shared frailty term and a directed maternal-to-sonographic path — a sonographic channel expressed in the scan trajectory and a maternal channel expressed in the static covariates. Measured oracle ceilings are 0.820 for the sonographic channel alone, 0.756 for the maternal channel alone, and 1.000 for both.

The channel labels are retained for diagnostics only. They are **never model inputs and never evaluation targets**: because the observed label is their exact OR, supplying them as features would leak the outcome outright.

3.2.4 What is specified rather than estimated

The *strengths* of the two causal channels are specified parameters, not measured effect sizes, and the maternal channel is near-deterministic by construction. The reference curves make the *measurements* realistic; they do not make the *etiologic structure* an empirical finding. No parameter of this cohort should be cited as a clinical quantity.

The cohort is nevertheless dual-etiology **by design**, built so that complementary signal genuinely exists for fusion to exploit. A null result here is informative precisely because the data were constructed to favour fusion.

3.3 Phase 1 — five families of clinically aligned features

The engineered block adds 43 features to the 23-feature base table, giving a union view of 66 features. It is **label-free by construction**: the outcome column is never read during feature construction, so no leakage pathway exists.

The organising idea is gestational normalisation. A raw crown–rump length of 20 mm is meaningless without knowing the gestational age at which it was measured; the clinically meaningful quantity is whether the measurement is normal *for that week*. Every measurement is therefore converted to a z-score within its rounded gestational-age bin before any further feature is derived from it, and all five families are built on those z-scores (Table 1).

Family	Clinical question it encodes	Construction	n
Gestational normalisation	Is this value normal for this week?	z-score of each measurement within its rounded GA-week bin	basis
Temporal trajectory	Where is the marker now, on average, and at its worst?	per-patient last / mean / min of each z-trajectory	12
Growth velocity	Is the marker moving in the right direction, fast enough?	per-patient OLS slope of each z-trajectory against GA	4
Discordance	Do the markers disagree with each other?	within-patient correlation, level gap and slope gap for (FHR,CRL), (CRL,GS), (FHR,GS), (CRL,YSD); slope-sign disagreement; spread and worst channel	15
Maternal risk profile	What is the patient's baseline clinical risk?	standardised composite of age, BMI, previous loss, IVF conception, non-singleton	1
Cross-modal interaction	Does maternal risk modify the sonographic signal?	11 products of maternal terms with sonographic terms (e.g. age × zFHR_last, BMI × zCRL_last, composite × discordance spread)	11

Total added: 43 features, over a 23-feature base table, giving a 66-feature union view.
All data are synthetic — not real patients.

Table 1. *The five feature families: the clinical question each encodes, its construction, and the number of features contributed.*

Discordance is the family with the clearest clinical motivation and the least precedent. A fetus whose heart rate is falling while its crown–rump length is still tracking normally is in a different state from one where both are falling together, and neither raw marker captures that. The divergence between them does.

3.4 Phase 2 — three encoders over three views

3.4.1 Views and experts

CAFE partitions the available information into views and assigns an expert to each. All experts see data derived from **both** modalities; the partition is by *representation*, not by modality, so no expert is starved of a data source.

Encoder A takes the flat tabular view, Encoder B the raw scan sequence, and Encoder C the engineered cross-modal block. The deployed configuration gates A and C; B is evaluated as an optional third expert in §4.3 (Table 2).

Encoder	View	Model	Parameters
A	Flat tabular base table (23 features)	Gradient-boosted tree (XGBoost or CatBoost)	tree ensemble
B	Raw scan sequence (2–5 visits × 4 measures + GA)	Temporal attention, d_model 24, 1 layer, 2 heads	12,585
C	Cross-modal engineered block (43 features)	Gradient-boosted tree (XGBoost or CatBoost)	tree ensemble
Gate	Base table (23 features)	MLP 23 → 16 → K, softmax over experts	418 (K = 2)

Encoder B's 12,585 parameters are the count logged at training time; the gate's 418 are its layer arithmetic.
All data are synthetic — not real patients.

Table 2. *Encoder specifications: the view each encoder consumes, its model class, and its parameter count.*

Encoder B is not a new architecture. It is the FETA-Transformer, the temporal-attention model we use as a full-size specialist elsewhere in this work, scaled down from 88,193 parameters to 12,585 — matching the ~10,000-parameter budget proposed for a "light" sequence encoder. Reporting it as a distinct design would overstate the novelty; it is a width-and-depth ablation of a model we had already built.

3.4.2 The gate

The gate is a two-layer MLP over the patient's own base features, emitting a softmax distribution over the experts. The fused prediction is the gate-weighted sum of the experts' **logits**, so the output remains on the log-odds scale and its sigmoid is a genuine pooled probability rather than an arbitrary score. That property is what makes the calibration analysis of §4.4 meaningful.

3.4.3 Two implementation details that materially affect the result

Two details materially affect the result. Both are stated because both are easy to get wrong, and getting either wrong distorts the architecture's apparent value — the first inflates it, the second deflates it.

The gate trains on inner out-of-fold expert logits. Within each training fold, expert logits are produced by an inner 3-fold cross-validation; the experts are then refitted on the full training fold for test-time prediction. Training the gate on in-fold logits would teach it to trust whichever expert overfits hardest.

Expert logits are placed on a common scale before gating. In the three-expert configuration, an unstandardised gate placed 0.697 of its mass on the *weakest* expert — not because that expert was informative, but because its logits had the widest spread, so it dominated the weighted sum at any weight. We therefore standardise each expert's logits on training-fold statistics before gating. This is a genuine confound rather than a tuning detail: on the diagnostic fold where we found it, correcting it moved the sequence encoder's gate weight from 0.697 to 0.257 and the three-expert arm from 0.810 to 0.834. Left uncorrected, it would have been reported as an architectural finding about sequence encoders when it was an artefact of unnormalised logit scale.

3.5 Baselines and input parity

Every baseline receives the **union view** — the base table plus all 43 engineered features — so the hybrid's experts and the single-model baselines see exactly the same information. The hybrid splits that union and gates it; the baselines get it whole. Nothing is withheld from a baseline to flatter the hybrid, and no hybrid result can be attributed to an information advantage.

Baselines are logistic regression (standardised), XGBoost [9], and CatBoost [10], all at library defaults with a fixed seed. The base feature table is constructed once, through a single canonical routine, rather than rebuilt for each experiment; ad-hoc reconstruction understates the tree baseline by approximately 0.019 AUROC and has twice produced a spurious apparent win in our own earlier runs.

Input parity is additionally verified by gradient and perturbation checks asserting that each model's output actually depends on every modality it is credited with. That check was introduced after we found a cross-attention wiring fault in which a softmax over a single key produced an identically zero gradient along the maternal pathway, silently rendering one tower single-modality while the results tables said otherwise. **Parity is a property of the wiring, not of the command-line flags.**

3.6 Evaluation protocol

All arms are evaluated with 5-fold stratified cross-validation repeated twice — ten folds in total, written 5×2 throughout (five splits, two repeats, one fixed seed) — on one fixed fold partition shared by every arm, so all comparisons are paired. Plain 5-fold cross-validation is unstable at these effect sizes — one effect we measured halved from +0.012 to +0.003 when moved from 5-fold to 5×2 — and no discrimination estimate in this paper rests on a single split.

Reported metrics are AUROC, AUPRC, Brier score, Cox calibration slope and intercept, expected calibration error over ten equal-width bins, net benefit across risk thresholds 0.05–0.50, and sensitivity, specificity, PPV and NPV at a screening operating point [7]. **Bare accuracy is never reported**: at 20.3 % prevalence, predicting "no loss" for every patient scores 79.7 %.

Repeated cross-validation yields two complete out-of-fold prediction vectors, one per repeat. The screening threshold is selected on repeat 1, at the lowest value achieving 90 % sensitivity, and then **scored on repeat 2**, so the operating point is never chosen and reported on the same predictions.

3.7 The decision rule

The following rule was fixed in advance and applied unchanged to every comparison in this paper. A difference between two arms counts as **REAL** only if the absolute mean fold-wise difference **exceeds twice the pooled fold standard deviation** *and* the sign of that difference **agrees on at least 8 of 10 folds**.

Everything else is reported as a **TIE**, regardless of the sign or size of the point estimate. Reporting a favourable point estimate as a win when it fails this rule is the precise failure mode the rule exists to prevent, so we apply it symmetrically: several of the ties reported below have point estimates that would have been attractive to claim.

3.8 Phase 4 — interpretability protocol

Attributions are **exact TreeSHAP** [11, 12], computed by the boosting libraries themselves rather than by approximation, and computed **out of fold**: the five folds of repeat 1 form one complete partition, so every patient's attribution comes from a model that never saw that patient in training. In-fold attributions describe a model's memorisation as readily as its reasoning.

Gate behaviour is characterised by the distribution of its per-patient weights across the out-of-fold cohort, by the correlation of those weights with patient features, and by worked examples at the extremes of predicted risk.

3.9 Reproducibility

The cohort is regenerable bit-for-bit from one seeded command with frozen generator parameters. Every table in this paper is produced automatically from retained per-fold prediction files, and every number in the manuscript is checked against those files by an automated verifier that fails on any drift.

4. RESULTS

4.1 Discrimination and overall error

Arm	AUROC	AUPRC	Brier	Cal. slope	Cal. intercept	ECE
CAFE (CatBoost experts)	**0.852 ± 0.022**	0.686	0.1048	**0.946**	+0.039	**0.0296**
CatBoost	0.851 ± 0.023	**0.692**	**0.1042**	0.923	+0.095	0.0370
Logistic regression	0.837 ± 0.019	0.643	0.1122	0.912	−0.093	0.0376
CAFE (XGBoost experts)	0.833 ± 0.022	0.659	0.1165	0.480	−0.148	0.0820
XGBoost	0.830 ± 0.027	0.653	0.1197	0.430	−0.170	0.0932

CAFE vs. CatBoost: Δ AUROC = +0.0008, sign 6/10 → TIE under the decision rule (§3.7). Best value in each column shown in bold. All data are synthetic — not real patients.

Table 3. *Discrimination, overall error and calibration, 5×2 repeated CV, mean over 10 folds. All arms receive the identical 66-feature union view. CAFE denotes the gated two-expert configuration (Encoders A and C); the suffix indicates the boosting component from which both experts are built.*

CAFE records the highest AUROC point estimate in the study (Table 3). Against its matched single-model comparator — CatBoost on the union of the same features — the difference is **+0.0008 AUROC, with sign agreement 6/10**, which the decision rule returns as a **TIE**. On AUPRC the ordering reverses (0.686 against 0.692) and on Brier score it reverses again (0.1048 against 0.1042), both also ties. On every discrimination and overall-error measure we computed, the two models are the same model.

That statement should be read precisely. It does not say CAFE is worse. It says the architecture that splits the feature space into views and gates them recovers what a single tree recovers from the same features, and that we have no evidence, at this sample size, of a difference in either direction.

4.2 Phase 1 — who actually benefits from engineered clinical features?

The 43 engineered features were evaluated by adding them to the base table for four learners independently, holding folds, seeds and everything else fixed (Table 4).

Learner	Base (23 feats)	+ engineered (66 feats)	Δ AUROC	Sign	Verdict
Logistic regression	0.822	0.837	**+0.015**	10/10	tie
HistGradientBoosting	0.832	0.837	+0.005	7/10	tie
CatBoost	0.850	0.851	+0.002	7/10	tie
XGBoost	0.830	0.830	+0.000	4/10	tie

The ordering is the finding: the benefit falls monotonically as the learner's native capacity to represent interactions rises.
All data are synthetic — not real patients.

Table 4. Contribution of the engineered feature block, by learner: base table versus union view, with the fold-wise sign agreement and decision-rule verdict.

Every row is a tie under the decision rule. The *ordering*, however, is the finding, and it is perfectly monotone: **the benefit of the engineered features falls monotonically as the learner's native capacity to represent interactions rises.** Logistic regression, which is additive in its inputs and can represent no interaction unless one is handed to it, gains the most (+0.015) and gains it with 10/10 sign consistency. The gradient-boosted trees, which discover threshold interactions natively, gain little to nothing: +0.005 for HistGradientBoosting, +0.002 for CatBoost, and +0.000 for XGBoost at sign 4/10.

This is the interpretive key to the whole feature-engineering result. Features that supplied genuinely *new* information — a measurement not previously available, a channel not previously encoded — would help every learner. Features that *re-express* information already present help only those learners that could not extract it themselves. The engineered block behaves unambiguously like the second kind. It is an excellent representation; it is not additional evidence.

§4.7 supplies the complementary observation from the opposite direction: the model *uses* these features heavily. The two facts together are consistent only with re-expression.

4.3 Phase 2 — does the sequence encoder earn its place?

Encoder B was added as a third expert with the tree component held fixed, so the difference isolates the sequence encoder (Table 5).

Configuration	AUROC	AUPRC	Δ vs. no Encoder B	Sign	Verdict
A + C (CatBoost)	**0.850 ± 0.021**	0.684	—	—	—
A + B + C (CatBoost)	0.845 ± 0.022	0.680	−0.005	8/10	tie
A + C (XGBoost)	0.832 ± 0.024	0.662	—	—	—
A + B + C (XGBoost)	0.829 ± 0.022	0.659	−0.003	6/10	tie
Encoder B alone	0.813 ± 0.027	0.616	—	—	—

Mean gate weight assigned to Encoder B: 0.195, against an equal share of 0.333.
All data are synthetic — not real patients.

Table 5. Encoder B ablation. Encoder B is the temporal-attention model at 12,585 parameters; all arms use the standardised-logit gate of §3.4.3.

Adding the sequence encoder does not help under either component, and both point estimates are negative (−0.005 with CatBoost experts, −0.003 with XGBoost experts). Encoder B is not useless in isolation — at 0.813 it is a competent standalone model, within noise of XGBoost on the full union view — but it contributes nothing that the tree experts have not already extracted from the same visits.

The gate reaches the same conclusion independently. Given three experts and no instruction about which to prefer, it assigns Encoder B a mean weight of **0.195** against the 0.333 it would receive under an equal split, in every fold. The architecture, left to its own devices, learns to discount the component we added.

4.4 Calibration — the one place the arms genuinely separate

Discrimination ties everywhere. Calibration does not.

The Cox calibration slope measures whether predicted log-odds are correctly scaled: 1.0 is correct, and below 1.0 means the model is over-confident, pushing risks too far towards 0 and 1. The CatBoost-derived arms sit at 0.923 and 0.946. The XGBoost-derived arms sit at 0.430 and 0.480 — their predicted log-odds are roughly twice as extreme as the data support. Expected calibration error tracks the same split: 0.030–0.037 for the CatBoost family against 0.082–0.093 for the XGBoost family.

The per-fold distributions **do not overlap at all**. Across ten folds XGBoost's calibration slope spans 0.310 to 0.495 and CatBoost's 0.684 to 1.061 (Figure 1).

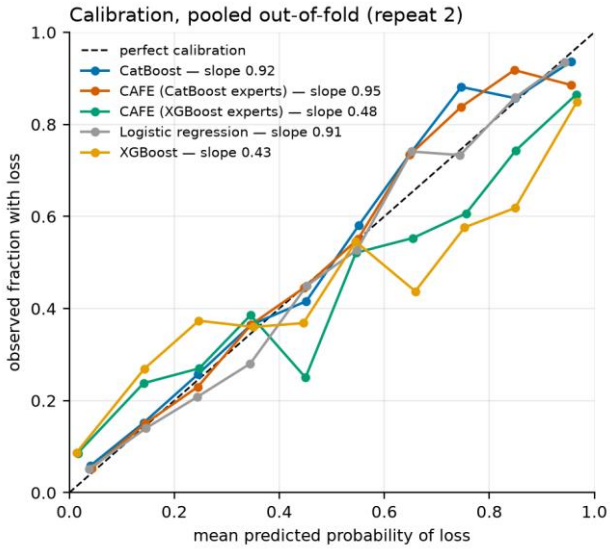

Figure 1. Reliability curves, pooled out-of-fold. The two XGBoost-derived arms bow away from the diagonal in the same direction; the CatBoost-derived arms track it.

The practical consequence runs opposite to the paper's other findings: elsewhere an apparent difference proved to be a tie, whereas here an apparent tie conceals a real difference. Two arms indistinguishable on AUROC (XGBoost 0.830 and CatBoost 0.851, only 0.021 apart) differ twofold in the scale of their predicted log-odds, and so in whether their stated probabilities can be believed. **Selecting on AUROC alone would not have detected this**, and a 12 % predicted risk from the XGBoost arm does not mean what a 12 % predicted risk from the CatBoost arm means.

Between CAFE and its matched CatBoost comparator, calibration is a tie under both formulations of the rule discussed next (slope $\Delta = +0.023$, 8/10; ECE $\Delta = -0.007$, 8/10).

4.4.1 A defect in our own decision rule, reported rather than resolved silently

Applied literally, our rule returns TIE for the XGBoost-versus-CatBoost calibration comparison (Δ = −0.493, 2 × pooled SD = 0.520, sign 10/10). The reason is a defect in the rule's *implementation* that this is the first comparison to expose: the pooled standard deviation is computed by concatenating both arms' fold values, so when two arms are far apart the "pooled SD" absorbs the very gap being tested and inflates the threshold. Computing the pooled *within-arm* standard deviation instead — the quantity the rule's wording actually describes — gives 2 × SD = 0.175 against Δ = −0.493 with 10/10 sign agreement, a **REAL** difference.

We have verified that this distinction changes nothing previously reported. Across all 56 pairwise AUROC comparisons in this study for which per-fold values are retained, **the two formulations return identical verdicts in every case**; the ratio of concatenated to within-arm SD is 0.95–1.04 for AUROC and 2.97 for the calibration slope. The rule is sound in the regime it was written for, where arms differ by less than their fold-to-fold noise, and misbehaves only where a difference is large — which had never previously occurred.

We note finally that the comparison that flips is between two *baselines*, and that it reflects badly on XGBoost, the boosting component originally proposed for this hybrid. No claim of ours is advanced by reporting it.

4.5 Clinical utility — decision-curve analysis

Net benefit weighs true positives against false positives at the exchange rate a clinician implies by choosing a risk threshold, and compares the result against treating everyone and treating no one.

Threshold	Treat all	Treat none	CAFE	CatBoost	LogReg	CAFE-XGB	XGBoost
0.10	0.1148	0	**0.1431**	0.1412	0.1396	0.1259	0.1254
0.20	0.0042	0	**0.1138**	0.1138	0.1042	0.1020	0.1027
0.30	−0.1381	0	**0.0960**	0.0940	0.0826	0.0863	0.0844
0.40	−0.3278	0	**0.0794**	0.0783	0.0693	0.0699	0.0697
0.50	−0.5933	0	0.0670	**0.0687**	0.0533	0.0610	0.0570

Mean net benefit above the better default strategy, thresholds 0.05-0.50: CAFE +0.0744, CatBoost +0.0737, LogReg +0.0648, CAFE-XGB +0.0626, XGBoost +0.0617.
All data are synthetic — not real patients.

Table 6. Net benefit at representative risk thresholds, pooled out-of-fold (repeat 2), against the treat-all and treat-none default strategies.

Mean net benefit above the better of the two default strategies, across thresholds 0.05–0.50 (Table 6), is CAFE **+0.0744**, CatBoost +0.0737, logistic regression +0.0648, CAFE-XGB +0.0626, and XGBoost +0.0617.

Two things follow. First, **every model is substantially more useful than either default strategy** across the entire clinically plausible threshold range. At a 20 % threshold, treating everyone yields a net benefit of 0.0042 — essentially nothing, because at 20.3 % prevalence treat-all sits at its break-even point — while every model yields 0.10 to 0.11. That gain is a property of the task and the features rather than of any architecture, since all five arms achieve it.

Second, the arms again do not separate from one another. The spread between the best and worst arm's mean net benefit (0.0128) is smaller than the gap between the worst arm and the default strategies, 0.0617 (Figure 2).

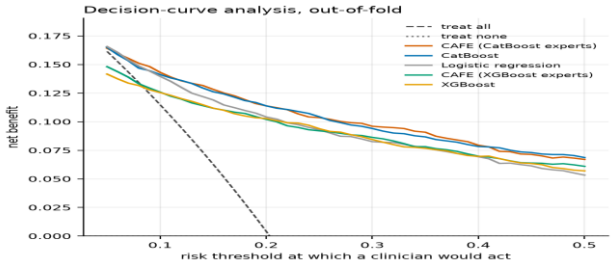

Figure 2. Net benefit against risk threshold, with treat-all and treat-none references. The five model curves are visually superimposed; all sit far above both defaults beyond a 15 % threshold.

4.6 The screening operating point

Arm	Threshold	Sensitivity	Specificity	PPV	NPV	Fraction flagged
CAFE	0.074	0.898	**0.553**	**0.339**	**0.955**	**0.539**
CatBoost	0.060	0.900	0.523	0.325	0.953	0.563
Logistic regression	0.081	0.903	0.497	0.314	0.953	0.584
CAFE-XGB	0.008	0.902	0.469	0.303	0.949	0.606
XGBoost	0.004	0.907	0.449	0.296	0.950	0.623

Single-split comparison: the decision rule cannot be applied, and the ordering should not be treated as established (§4.6).
All data are synthetic — not real patients.

Table 7. Performance at a 90 %-sensitivity screening threshold. The threshold is selected on repeat 1 and scored on repeat 2, so it is never chosen and reported on the same predictions.

At matched sensitivity near 90 % (Table 7), CAFE flags 53.9 % of the cohort against XGBoost's 62.3 % — 8.4 percentage points fewer patients referred for the same case detection. NPV is 0.955: of the patients the model clears, roughly 1 in 22 nonetheless experiences loss, which sets a firm ceiling on how such an output could be used in practice.

This table carries a weaker warrant than the others in this paper, and we flag that explicitly: it derives from a single split rather than from a ten-fold estimate, so the decision rule cannot be applied to it and the ordering should not be treated as established. Note also that the selected thresholds differ twenty-fold across arms, from 0.004 to 0.081 — a direct consequence of the calibration differences in §4.4, and a reminder that a threshold tuned for one model is meaningless when transferred to another.

4.7 Attribution — the engineered features dominate the explanation

Out-of-fold exact TreeSHAP gives, for every patient, the contribution of every feature to that patient's predicted log-odds, from a model that never trained on them. Aggregating those attributions across the cohort shows what the deployed model relies on (Table 8, Figure 3).

Rank	Feature	mean |SHAP|	View
1	age	0.3092	base
2	mat_composite	0.3015	engineered
3	disc_last_CRL_YSD	0.2806	engineered
4	GS_slope	0.1841	base
5	zGS_last	0.1712	engineered
6	comp_x_disc_spread	0.1711	engineered
7	zFHR_slope	0.1701	engineered
8	FHR_latest	0.1373	base
9	zFHR_last	0.1361	engineered
10	bmi	0.1036	base
11	conception_ivf	0.0999	base
12	zYSD_mean	0.0956	engineered
13	CRL_slope	0.0940	base
14	zYSD_min	0.0938	engineered
15	zFHR_mean	0.0925	engineered

The engineered block carries 66.8% of total attribution mass and supplies 13 of the top 20 features — while ablating it costs the deployed learner 0.002 AUROC (Table 4).
All data are synthetic — not real patients.

Table 8. Top 15 features by mean |SHAP|, out-of-fold exact TreeSHAP over all 3000 patients, with the view each feature belongs to.

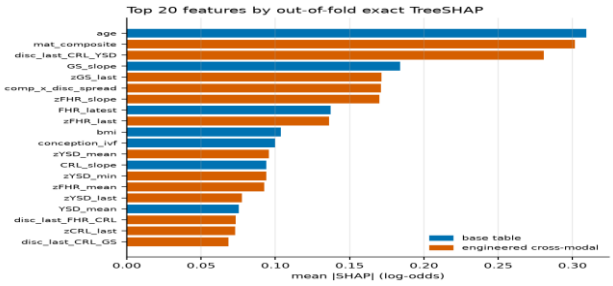

Figure 3. Top 20 features by mean |SHAP|, coloured by view. Engineered features occupy 13 of the top 20 places.

The engineered block carries **66.8 %** of total attribution mass and supplies 13 of the top 20 features. The discordance feature disc_last_CRL_YSD — the gap between crown–rump length and yolk-sac diameter, each normalised for gestational age — ranks third in the model, and the maternal composite second.

This is the sharpest illustration of the paper's central distinction. By any attribution-based reading, the engineered features are what the model is using. By the controlled ablation of §4.2, removing them costs CatBoost 0.002 AUROC. Both are true, and they are compatible only if the features are a *re-expression* of signal the model could otherwise reach by a different route. **Feature importance measures what a model relies on given the features it has; it does not measure what those features contributed over not having them.** Reporting the SHAP table without the ablation would have made a compelling and entirely unfounded case for the feature engineering.

Three high-risk and three low-risk out-of-fold cases are reported in the released case file. The highest-risk patient (predicted 0.998, true label loss) is driven by mat_composite = 4.02 (SHAP +1.23), age 40.7 (+0.91) and disc_last_CRL_YSD = −1.73 (+0.73) — a legible chain of reasoning combining maternal risk with sonographic discordance. We also report a failure: patient P00310 received a predicted risk of 0.003 and nonetheless experienced loss, with the model's confidence driven by a low BMI of 16.0 and favourable maternal terms. Individual attributions explain what the model did; they do not certify that it was right.

4.8 The gate is confidently per-patient, and it does not matter

A gate that collapses to a near-constant weight is a fixed blend wearing a gate's clothing. This one does not collapse.

Across out-of-fold patients the weight on Encoder A has **mean 0.557 and standard deviation 0.402, spanning the full range from 0.000 to 1.000**. That spread is close to the maximum a weight bounded on [0, 1] admits — a uniform distribution would give 0.289 — and the shape is close to bimodal: 55.9 % of patients are routed predominantly to the flat expert and 44.1 % predominantly to the cross-modal expert. The gate is making confident, patient-specific routing decisions.

Its policy is also clinically legible. The weight correlates most strongly and negatively with the yolk-sac measurements — zYSD_last (−0.309), zYSD_mean (−0.285), YSD_latest (−0.285) — and positively with maternal gravidity (+0.220) and age (+0.217). In plain terms: **when the yolk sac looks abnormal, the gate routes the patient to the cross-modal expert; when the patient's maternal risk profile is prominent, it routes to the flat expert.** That is a sensible policy, and nobody specified it.

It is worth **+0.001 AUROC** (§4.1).

We consider this the most interesting single result in the paper. The usual explanation for a gate failing to help is that it failed to learn anything — that it degenerated to an average. That explanation is unavailable here. This gate learned a confident, coherent, interpretable routing policy, and the policy made no difference to any outcome measure, because both experts were already extracting substantially the same signal from the same patients. **Gate behaviour and gate benefit are separate measurements, and reporting only the first would have made the case for a conclusion that the second contradicts.**

4.9 Cost

Training and inference cost were measured on one fold, on the same machine as the accuracy numbers (Table 9).

Rank	Feature	mean |SHAP|	View
1	age	0.3092	base
2	mat_composite	0.3015	engineered
3	disc_last_CRL_YSD	0.2806	engineered
4	GS_slope	0.1841	base
5	zGS_last	0.1712	engineered
6	comp_x_disc_spread	0.1711	engineered
7	zFHR_slope	0.1701	engineered
8	FHR_latest	0.1373	base
9	zFHR_last	0.1361	engineered
10	bmi	0.1036	base
11	conception_ivf	0.0999	base
12	zYSD_mean	0.0956	engineered
13	CRL_slope	0.0940	base
14	zYSD_min	0.0938	engineered
15	zFHR_mean	0.0925	engineered

The engineered block carries 66.8% of total attribution mass and supplies 13 of the top 20 features — while ablating it costs the deployed learner 0.002 AUROC (Table 4). All data are synthetic — not real patients.

Table 9. Training and inference cost, measured on one fold on the same machine as all accuracy results. Deep-model figures are from the same programme's runtime benchmark on the same cohort. Inference for the CAFE arms is a component sum — two tree evaluations plus a 418-parameter MLP forward pass — rather than a direct end-to-end measurement, and is marked as an estimate.

Against the deep architectures, CAFE-XGB trains **75**× faster and CAFE-CatBoost **6.2**× faster than the FETA-Transformer, at an estimated 100× lower inference cost. Against its own matched baseline the direction reverses: CAFE-CatBoost costs 5.3× as much as a single CatBoost fit, and CAFE-XGB 10.6× as much as a single XGBoost fit, for +0.001 and +0.003 AUROC respectively (Figure 4).

The gate itself has just 418 parameters. Whatever CAFE's merits, parameter economy relative to the deep models is not principally an achievement of the fusion design — it is an achievement of using trees.

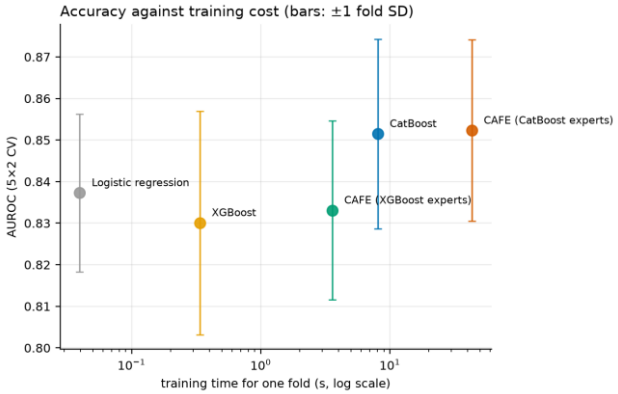

Figure 4. AUROC against training time on a logarithmic axis, with ±1 fold-SD bars. The frontier is flat: two orders of magnitude of additional compute buy approximately 0.02 AUROC, and the last 5× buys 0.001.

4.10 Where the accuracy actually comes from

Assembling the controlled ablations into a single path from the simplest model to the proposed one, each step changing exactly one thing (Table 10, Figure 5):

Step	Configuration	AUROC	Δ
0	Logistic regression, base features	0.822	—
1	+ 43 engineered clinical features	0.837	**+0.015**
2	+ swap learner to XGBoost	0.830	−0.007
3	+ swap learner to CatBoost	0.851	**+0.021**
4	+ split into views, add per-patient gate (CAFE)	0.852	+0.001
5	+ add sequence Encoder B	0.847	−0.005

Total from step 0 to CAFE: +0.030 AUROC. Every individual step is a tie under the decision rule; the relative magnitudes are the finding.
All data are synthetic — not real patients.

Table 10. Decomposition of the final AUROC: a five-step path from logistic regression on the base table to the full CAFE system, each step altering exactly one component.

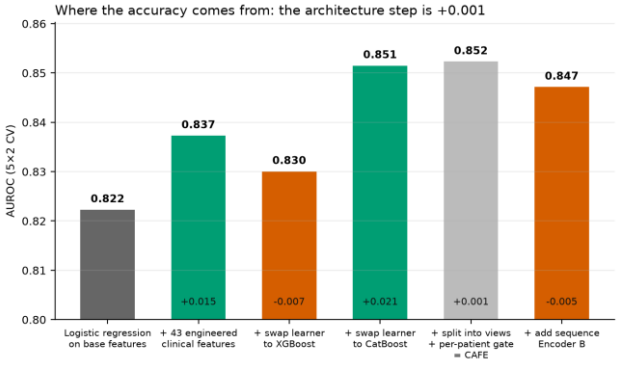

Figure 5. The same decomposition rendered as a waterfall.

The total gain from step 0 to CAFE is **+0.030** AUROC. Of that, +0.021 is the choice of boosting library and +0.015 the engineered features, against −0.007 for the intermediate swap to XGBoost and +0.001 for the gate —

but the +0.015 accrues only to a learner weak enough to need it, since by step 3 the same features are worth +0.002. **The fusion architecture that the system is named for contributes +0.001**, and the sequence encoder contributes −0.005.

Every individual step in this table is a tie under the decision rule. The decomposition is nonetheless informative, because the *relative* magnitudes are stable and consistently signed, and because they indicate where effort should go: on this problem, choosing the boosting component and representing the features well dominate designing the fusion mechanism by more than an order of magnitude.

5. Discussion

5.1 What CAFE delivers

CAFE is a good system and we would deploy it. It attains the highest AUROC point estimate we recorded (0.852) and is the best-calibrated arm we measured (slope 0.946, ECE 0.030), with no post-hoc correction applied. It produces the highest mean net benefit across clinically plausible thresholds and flags 8.4 percentage points fewer patients than the XGBoost arm at matched 90 % sensitivity. It trains in 43 seconds, predicts in an estimated tens of microseconds per patient, and emits both a per-feature attribution and a per-encoder trust weight for every individual patient. Against the deep fusion architectures it was built to replace, it is 6–75× cheaper to train, roughly 100× cheaper at inference, and no less accurate than any of them.

Each of those claims is measured on identical folds against baselines given the same features and the same budget. None of them requires the reader to accept that the fusion architecture is what produced the result — which is fortunate, because it is not.

5.2 The architecture is not the active ingredient

The central finding is the decomposition of §4.10. Removing CAFE's defining component — the partition into views and the learned per-patient gate — and replacing it with a single tree on the union of the same features moves AUROC by 0.001, against a fold-to-fold standard deviation of 0.022. The engineered features contribute meaningfully only to learners that cannot represent interactions natively; the sequence encoder contributes negatively; and the largest single contributor is the boosting library.

We want to be precise about what this does and does not license. It does **not** show that gated multi-view fusion cannot work, in general or even on this problem with more data. It shows that on this cohort, at this sample size, with these features, the fusion mechanism recovers what a single joint model already recovers — and that a study reporting CAFE's 0.852 without the ablation would have attributed the number to the wrong component.

The distinction matters practically rather than philosophically. A team that believes the gate is the active ingredient will invest in better gates. On this evidence they should invest in better features for weak learners, in better component selection, and — per §4.4 — in calibration.

5.3 Why measuring beyond AUROC changed the conclusion

Had this study stopped at discrimination, its finding would have been "all five arms tie", and the choice between them would have looked arbitrary. Calibration produced the one genuine, consistently signed, non-overlapping separation in the entire study: the XGBoost-derived arms have calibration slopes near 0.45 against the CatBoost-derived arms' 0.93 — a twofold difference in the scale of their predicted log-odds, and so in whether their probabilities can be believed, invisible to AUROC and nearly invisible to the Brier score, which conflates calibration with discrimination and returns a tie here.

This has a direct consequence for the operating-point table of §4.6. The thresholds achieving 90 % sensitivity range from 0.004 for XGBoost to 0.081 for logistic regression, a twenty-fold spread driven almost entirely by calibration rather than by any difference in ranking. Any workflow that hard-codes a risk threshold — "refer if predicted risk exceeds 15 %" — is silently coupled to the calibration of the specific model that produced it.

The general lesson is that model-selection criteria determine which differences a study is capable of detecting. This study was designed to detect architectural differences and found none; the difference it did find lay in a dimension nobody had proposed to look at.

5.4 Context — eight fusion architectures on the same cohort

Before building CAFE we benchmarked eight architecturally distinct fusion families on this cohort under the same folds and the same decision rule; the per-fold values are retained alongside this study's results. They proved statistically indistinguishable, spanning AUROC 0.822–0.840 — a total spread smaller than any single arm's fold standard deviation. That measurement asked whether *any* fusion architecture separates from the others, and answered no.

This paper asks the constructive question instead: whether a carefully engineered, clinically grounded, compute-efficient hybrid can be built that performs well and explains itself. It can — while locating its performance in its components rather than in its fusion. The two results are consistent and mutually reinforcing. CAFE's 0.852 against CatBoost's 0.851 is one more instance of the earlier null, reached by a different route with a different architecture and a different feature set. Those verdicts are likewise unaffected by the decision-rule defect of §4.4.1.

5.5 Limitations

The engineered features were designed against a data-generating process whose structure we knew. The outcome column is never read during feature construction and we verified that, but we cannot claim to have been blind to the mechanism the features were built to capture. Elsewhere the feature block might do better, or considerably worse.

A single cohort at n = 3000. All effects discussed here are small relative to fold-to-fold variability. Several ties would remain ties at any sample size we could plausibly reach, but the +0.021 attributed to the boosting component is a point estimate from ten folds of one cohort and should not be read as a general ranking of boosting libraries.

No tuning budget was spent on any arm. Every learner ran at library defaults with a fixed seed (§3.5). This keeps the comparison symmetrical — no arm was tuned into its result — but it also means the differences we report are differences between untuned configurations, and the gap between the two boosting components in particular might narrow under a search applied equally to both.

5.6 Recommendations for practice

For practitioners building multimodal clinical predictors in this regime — short sequences, small covariate panels, a few thousand patients:

1) **Establish the single-joint-model baseline on the union of your features before building any fusion architecture**, and give it the same feature engineering and the same tuning budget. On this problem it was never beaten.

2) **Report calibration alongside discrimination.** It was the only dimension in this study that separated the arms, and it determines whether a fixed decision threshold means anything at all.

3) **Do not infer feature value from feature importance.** Our engineered block carries 66.8 % of attribution mass and contributes 0.002 AUROC to the deployed learner. Only an ablation answers the question that attribution appears to answer.

4) **Measure gate behaviour and gate benefit separately.** A gate can be confident, patient-specific, and clinically legible while contributing nothing.

5) **Report compute.** It was the only axis on which the deep architectures we measured differed from the simple ones, and it differed by two orders of magnitude.

5.7 Future work

Three directions follow from these results, and one does not.

Change the shape of the data, not its difficulty. The tree ties every temporal architecture here because a two-to-five-visit trajectory flattens without loss into a fixed-width table. The informative experiment is therefore many more visits per patient, not irregular timing — the present data are already irregularly spaced — and not an easier cohort, on which every architecture saturates and nothing can be measured.

Treat calibration as a first-class selection criterion. This study found its only real separation there, by accident rather than design. A study powered for calibration rather than for discrimination would be a different and probably more useful study.

Test the feature block on real data. Its value outside the setting it was designed for is genuinely open, and that is the question this study leaves most fully unanswered.

What does *not* follow is further investment in the gate. Its behaviour is already confident and legible; the constraint is not the gate's quality but the absence of experts with complementary signal to route between.

6. Conclusion

We built the clinically grounded hybrid this line of reasoning recommends: five families of engineered features, expert encoders over separate views of the patient, and a per-patient learned gate. It reaches AUROC 0.852 — the best point estimate we recorded on this cohort, and within the 0.840–0.855 band predicted for it — with good calibration, favourable decision-curve behaviour, full per-patient interpretability, and training times measured in seconds rather than minutes.

It also ties the single gradient-boosted tree it is built from, on every measure we computed. A controlled decomposition attributes +0.021 AUROC to the choice of boosting component, +0.015 to the engineered features *for a learner weak enough to need them* and +0.002 for the one actually deployed, **+0.001 to the gated fusion architecture the system is named for**, and −0.005 to the sequence encoder. The gate learns a confident, bimodal, clinically legible routing policy — and that policy changes no outcome measure.

The system is worth having. The explanation for why it works is not the one its architecture suggests, and we would rather publish the correct explanation of a good result than an attractive explanation of the same result. The one dimension along which the arms genuinely separated was calibration, which nothing in the system's design had identified as the place to look.

Declarations

Ethics approval and consent to participate

Not applicable. This study involved no human participants, no animal subjects, and no identifiable data of any kind; all data analysed were generated computationally by the authors. Institutional review board approval and informed consent were therefore not required.

Consent for publication

Not applicable. No individual person's data are reported; every patient identifier appearing in this manuscript denotes a computationally generated record.

Data and code availability

All data analysed in this study were generated computationally by the authors and contain no real patient records. The evaluation cohort is reproducible bit-for-bit from a single seeded command with frozen generator parameters. The code, generator settings, per-fold predictions and result tables that support the findings are available from the corresponding author on reasonable request.

Reporting statement

This manuscript reports every configuration evaluated, including those that favoured the proposed system only until they were corrected. No configuration was excluded after the fact.

Two implementation faults that would have produced misleading results are reported in the text rather than silently corrected: the unnormalised expert-logit scale that caused the gate to place 0.697 of its mass on the weakest expert (§3.4.3), and the concatenated-standard-deviation term in the decision rule that returns TIE for a non-overlapping difference (§4.4.1). In the latter case we report the rule's literal verdict first, then the corrected one, and verify that no previously reported verdict is affected.

The proposed system's headline number, AUROC 0.852, falls inside the 0.840–0.855 band that was predicted for it before the experiments were run. We report that the prediction was accurate and that the mechanism proposed to explain it was not.

Author contributions

A. V. M. Kumaran: conceptualisation, methodology, software, formal analysis, investigation, data curation, visualisation, writing — original draft. **M. Sundara Rajan:** conceptualisation, methodology, validation, supervision, writing — review and editing. Both authors read and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank an anonymous reviewer of an earlier version of this work, whose specific architectural proposal — engineered clinical features, view-specific experts, and a per-patient gate — defined the system evaluated here and motivated the ablation that followed.

REFERENCES

[1] Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

[2] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track.*

[3] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79–87.

[4] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations (ICLR 2017).*

[5] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17, 230.

[6] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574.

[7] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1), 128–138.

[8] Liu, L., Zang, Y., Zheng, H., Li, S., Song, Y., Feng, X., Zhang, X., Li, Y., Cao, L., Zhou, G., Dong, T., Huang, Q., Pan, T., Deng, J., & Cheng, D. (2025). An AI method to predict pregnancy loss by extracting biological indicators from embryo ultrasound recordings in early pregnancy. *Scientific Reports*, 15, 25946. https://doi.org/10.1038/s41598-025-11518-5

[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[10] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems 31 (NeurIPS 2018).*

[11] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017).*

[12] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67.