Journal of
Information Technology,
Cybersecurity, and Artificial Intelligence

(RESEARCH ARTICLE)

PH-AIDE: A Multi-Modal Explainable AI Framework for ICU Risk Stratification and Evidence-Based Public Health Policy Design

Safkat Mustakim Nirjash
Memorial University of Newfoundland
Canada
Email: Safkat.nirjash@gmail.com

Sharmin Akter
North South University
Bangladesh
Email: sharmin16.bd17@gmail.com

Shakim Ahamed
National University Tejgaon college
Bangladesh
Email: akash.sakim35520@gmail.com

Dr. Emon Hasan
Jahangirnagar University
Bangladesh
Email: emon.md6218@gmail.com

Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2026, Volume 3, Issue 4,, Pages 48-66

DOI: https://doi.org/10.70715/jitcai.2026.v3.i4.087

Abstract

We present PH-AIDE (Public Health AI Decision Engine), a multi-modal, explainable AI framework for five-class intensive care unit (ICU) risk stratification and evidence-based public health policy design. Drawing exclusively on MIMIC-III critical care data, PH-AIDE fuses three complementary modalities from the same patient cohort: (i) structured clinical features (vitals, laboratory values, ICD-9 diagnoses, comorbidity indices) processed by an XGBoost classifier; (ii) hourly vital-sign time-series from the first 48 hours of ICU admission processed by a two-layer bidirectional LSTM (Bi-LSTM) network with additive attention; and (iii) temporally filtered clinical progress and nursing notes (timestamped within the first 48 hours of admission, with discharge summaries explicitly excluded to prevent information leakage) processed by a fine-tuned BioBERT encoder. Module outputs are fused by a two-layer attention-weighted gating network trained in a staged procedure. Evaluated under temporally blocked cross-validation stratified by risk class, PH-AIDE achieves a macro-averaged F1 score of 0.88 ± 0.01, AUC-ROC of 0.93 ± 0.01, and Brier score of 0.081 ± 0.01 across five risk strata, statistically significantly outperforming all single-modality and naïve ensemble baselines ($p < 0.01$, paired bootstrap, $B = 10{,}000$). A comprehensive ablation study confirms the independent contribution of each modality and the superiority of learned attention weighting over mean fusion. On a separate CDC FluSight walk-forward validation, a standalone Bi-LSTM achieves a 1-week-ahead MAE of 0.38% versus 0.51% for ARIMA and 0.41% for the FluSight-ensemble benchmark. An integrated policy simulation layer using constrained linear programming translates probabilistic risk forecasts into SDG-3-aligned resource allocation recommendations, with WHO Global Health Observatory indicators providing cross-national context. Dual-granularity SHAP explainability (global beeswarm + patient-level waterfall) and subgroup-stratified fairness evaluation with multi-class demographic parity difference (macro-DPD ≤ 0.04) are reported. An extended discussion addresses ethical considerations, governance alignment with WHO AI principles, and reproducibility. These findings demonstrate that PH-AIDE can support timely ICU triage,

transparent clinical decision-making, and evidence-based healthcare resource allocation. The framework provides a practical pathway toward trustworthy AI-assisted public health policy while maintaining fairness and explainability.

Keywords: Multi-modal AI; ICU risk stratification; Explainable AI; XGBoost; Bi-LSTM; BioBERT; Attention fusion; SHAP; Fairness-aware evaluation; Public health policy; Ablation study

1. INTRODUCTION

1.1. Background and Motivation

The global healthcare landscape is undergoing a fundamental transformation driven by the rapid proliferation of digital health technologies and the exponential growth of health-related data. Machine learning and deep learning algorithms have demonstrated remarkable capabilities in extracting actionable insights from complex, high-dimensional clinical datasets, enabling predictive analytics that can lead to timely interventions, personalized treatment strategies, and improved population health outcomes [2,3]. The healthcare AI market is projected to expand from USD 26.6 billion in 2024 to USD 187.7 billion by 2030 at a compound annual growth rate exceeding 38% [1].

Despite these advances, translation of AI capabilities into accurate, equitable, and ethically sound decision support remains challenging. The COVID-19 pandemic exposed critical deficiencies in public health information technology infrastructures, while simultaneously catalyzing an unprecedented period of innovation in AI-powered surveillance and risk stratification [4,5]. Three principal deficiencies persist: most predictive models operate on a single data modality without systematically fusing complementary information sources [6]; the opacity of state-of-the-art deep learning models undermines clinician trust and regulatory acceptance [7,8]; and critical issues related to algorithmic fairness, demographic bias, and privacy preservation remain insufficiently addressed in existing frameworks [9,10]. Also a recent work [44] introduces quantum computing for skin cancer detection. This is a great contribution to healthcare.

A fourth gap rarely acknowledged is the disconnection between prediction and action. The vast majority of AI health research focuses narrowly on prediction tasks without explicitly bridging the gap to policy translation [11]. PH-AIDE addresses all four deficiencies simultaneously.

1.2. Problem Statement and Research Objectives

This work pursues four interconnected objectives:

1. **Objective 1:** Design and validate a multi-modal ICU risk stratification architecture (PH-AIDE) that jointly processes structured clinical features, temporal vital-sign trajectories, and early clinical notes from a common patient cohort via an attention-weighted ensemble fusion mechanism.
2. **Objective 2:** Embed comprehensive dual-granularity explainability (global SHAP feature-importance rankings and patient-level SHAP waterfall attributions) directly within the inference pipeline.
3. **Objective 3:** Conduct rigorous empirical benchmarking, including a full ablation study, per-class performance breakdown, and fairness evaluation with multi-class demographic parity difference across demographic subgroups in MIMIC-III.
4. **Objective 4:** Develop an integrated policy simulation layer that translates probabilistic ICU risk forecasts into SDG-3-aligned resource allocation recommendations via constrained linear programming.

There are some research questions:

RQ1: Can multi-modal fusion improve ICU risk stratification compared with single-modality models?

RQ2: Does attention-weighted fusion outperform conventional fusion strategies?

RQ3: Can explainable AI improve transparency while maintaining predictive performance?

RQ4: Can AI predictions be translated into practical public health policy recommendations?

1.3. Key Contributions

The principal contributions of this paper are:

5. **A unified, homogeneous multi-modal fusion architecture** in which all three modality-specific modules (XGBoost, Bi-LSTM, BioBERT) are trained on the same MIMIC-III patient cohort using different input modalities, producing compatible probability vectors over the same five risk classes that can be mathematically fused.
6. **A formally defined outcome variable and class distribution** for the five-class ICU risk stratification task, enabling full reproducibility of all reported results.
7. **A temporally safe data preprocessing protocol** that excludes discharge summaries and restricts clinical notes to the first 48 hours of ICU admission, eliminating information leakage from future clinical documentation.
8. **A two-layer attention-weighted gating network** with full training procedure specification (staged pre-training followed by frozen-module meta-learner training), with an ablation study confirming its contribution.
9. **A separate outbreak forecasting evaluation** on CDC FluSight data using walk-forward blocked temporal validation, benchmarked against the FluSight-ensemble, ARIMA, Prophet, and LightGBM baselines.
10. **Comprehensive fairness evaluation** using a formally extended multi-class macro-DPD metric across sex and race/ethnicity subgroups.

2. Literature Review

2.1 AI in Predictive Clinical Medicine

Supervised machine learning models, particularly ensemble methods (LightGBM, XGBoost, gradient boosting) and deep learning architectures (GRU, LSTM), to achieve disease prediction accuracies ranging from 88% to 95% across diverse clinical tasks [12]. XGBoost has consistently emerged among top-performing algorithms in comparative benchmarks due to its capacity to handle missing data, computational efficiency, and resistance to overfitting [13]. However, most studies evaluate performance on single-source, homogeneous datasets without systematically integrating complementary data modalities. Multi-modal fusion architectures combining structured clinical data with temporal sequences and unstructured text remain underexplored despite theoretical advantages [6].

A critical distinction between prior multi-modal health AI work and PH-AIDE is task homogeneity. Several prior systems [e.g., 14] fuse modules trained on entirely different datasets or different tasks (e.g., regression on epidemiological data combined with classification on clinical data) without addressing the resulting incompatibility of output spaces. PH-AIDE explicitly addresses this by ensuring all modality-specific modules output probability vectors over the same five risk classes from the same patient cohort.

2.2 AI for Public Health Surveillance and Outbreak Forecasting

Recurrent neural networks, particularly LSTM and GRU variants, process sequential data effectively, making them well-suited for time-series disease-incidence trends [14]. The CDC FluSight challenge has emerged as a standard benchmark for influenza forecasting, with ensemble methods consistently outperforming single-model baselines [24]. Critical limitations of current systems include: reliance on historical incidence data without incorporating textual intelligence; standard k-fold cross-validation applied to time series, which leaks future information into training folds; and absence of comparison against the FluSight-ensemble reference model.

PH-AIDE addresses these limitations in its outbreak forecasting evaluation by (i) using walk-forward blocked temporal validation and (ii) explicitly benchmarking against the FluSight-ensemble.

2.3 Explainable AI in Clinical Settings

SHAP (SHapley Additive exPlanations) values, grounded in cooperative game theory, provide a unified measure of feature importance satisfying local accuracy, missingness, and consistency axioms [18]. TreeSHAP enables exact polynomial-time SHAP computation for tree ensembles; DeepSHAP extends the framework to neural architectures [18]. Providing SHAP-based explanations alongside clinical predictions has been shown to enhance clinician trust and support identification of model errors [8]. PH-AIDE extends existing SHAP implementations by providing dual-granularity output (global beeswarm + patient-level waterfall) integrated within the inference pipeline rather than applied post hoc.

2.4 Fairness and Equity in Health AI

Algorithmic bias can arise from non-representative training data, biased feature selection, or optimization objectives that fail to account for fairness constraints [19]. Demographic Parity Difference (DPD) has been widely adopted as a

fairness metric for binary classification [19]. For multi-class settings, existing work has proposed macro-averaging over per-class DPD values [19], yet many health AI papers either omit fairness evaluation entirely or apply the binary DPD formula without adaptation. PH-AIDE formally defines and applies a multi-class macro-DPD metric appropriate to the five-class risk stratification setting.

2.5 Identified Research Gaps

Synthesizing the literature above reveals five critical gaps that motivate PH-AIDE: (i) absence of homogeneous multi-modal fusion across structured, temporal, and text modalities within a single patient cohort; (ii) information leakage from discharge summaries in clinical NLP studies; (iii) time-series cross-validation leakage in forecasting evaluations; (iv) incomplete or absent ablation studies in ensemble health AI papers; and (v) misapplication of binary fairness metrics to multi-class prediction tasks.

3. Methodology

3.1 Variable and Task Definition

Definition 1 (Five-Class ICU Risk Label). Each MIMIC-III ICU stay is assigned to one of C = 5 ordered risk classes based on a composite outcome determined at the time of ICU discharge. The composite outcome O = {in-hospital mortality OR ICU length of stay > 7 days OR SOFA score ≥ 11 at any point during admission}. Risk class assignment follows SOFA-score-anchored thresholds: *Class 1 (Very Low):* SOFA ≤ 2 AND LOS ≤ 2d AND O = 0; *Class 2 (Low):* SOFA 3–6 AND LOS 2–5d AND O = 0; *Class 3 (Moderate):* SOFA 7–10 OR LOS 5–7d OR any single O criterion borderline; *Class 4 (High):* SOFA ≥ 11 OR LOS > 7d; *Class 5 (Critical):* in-hospital mortality = 1. For ambiguous cases, the higher class is assigned. Class distribution in the MIMIC-III cohort (N = 38,597 admissions after exclusion criteria): Very Low = 24.1%, Low = 26.3%, Moderate = 22.8%, High = 16.9%, Critical = 9.9%.

Exclusion criteria: admissions with total ICU stay < 4 hours (insufficient temporal data), age < 18 years, and admissions missing more than 50% of structured vital-sign features during the first 24 hours. The final cohort consists of N = 38,597 admissions. All experiments use this cohort with patient-level train/validation/test splits to prevent data leakage across patients.

3.2 Data Sources

Table 1. Data sources and their roles. WHO GHO data are used exclusively as contextual inputs to the policy simulation layer and do not contribute to any trained model or reported metric in Tables 2–5.

Dataset	Source	Records	Modality Role	Key Variables
MIMIC-III Clinical Database	PhysioNet / MIT [23]	38,597 ICU admissions (2001–2012)	Primary: Modules A, B, C	Vitals, labs, ICD-9 codes, CCI, progress/nursing notes (≤48 h)
CDC FluSight Challenge	CDC [24]	2015–2024 (480 usable weeks)	Secondary: Forecasting evaluation only	Weekly ILI proportions by HHS region (10 regions)
WHO Global Health Observatory	WHO Open Data [25]	194-member states, annual	Policy layer context only (not model training)	Mortality rates, immunization coverage, WASH access, SDG-3 indicators

MIMIC-III (Medical Information Mart for Intensive Care III) is a de-identified critical care database containing comprehensive clinical information for ICU admissions at Beth Israel Deaconess Medical Center [23]. It includes demographics, hourly vital signs measured at the bedside, laboratory test results, medications, ICD-9 diagnosis and procedure codes, and timestamped free-text clinical notes, including admission notes, nursing notes, progress notes, and discharge summaries. Access requires a PhysioNet data use agreement and CITI training.

CDC FluSight provides weekly influenza-like illness (ILI) proportions across 10 U.S. HHS regions and the national aggregate, curated through the CDC ILINet sentinel surveillance network [24]. It serves as the standard benchmark for national outbreak forecasting evaluation.

WHO GHO provides aggregate health statistics for WHO member states [25]. These indicators are used exclusively to parameterize the demand vector b in the policy simulation LP (Section 3.7), not as training inputs to any machine learning model.

3.3 Data Preprocessing

3.3.1 Structured Clinical Features (Module A Input)

Missing values in continuous variables are imputed using Multivariate Imputation by Chained Equations (MICE) with predictive mean matching for continuous and logistic regression for binary variables [26]. All imputation models are fitted on training-fold data only and applied to validation and test folds to prevent leakage. Continuous features are standardized to zero mean and unit variance using training-fold statistics. Categorical variables (ICD-9 codes) are encoded via learned dense embeddings of dimension $d_e = 32$, trained end-to-end with Module A [27].

Derived engineered features include: the Charlson Comorbidity Index (CCI) [28]; summary statistics (mean, SD, minimum, maximum, trend slope via linear regression) over the first 24 hours of ICU admission for all measured vital signs (heart rate, mean arterial pressure, respiratory rate, SpO_2, temperature) and laboratory values (creatinine, lactate, hemoglobin, white cell count); and time-since-admission for temporally stamped events. The final structured feature matrix has $d_s = 187$ features per admission.

3.3.2 Temporal Vital-Sign Sequences (Module B Input)

Hourly vital-sign measurements (heart rate, mean arterial pressure, respiratory rate, SpO_2, temperature) during the first 48 hours of ICU admission form the temporal input tensor $X_{temp} \in \mathbb{R}^{Nx\ xdt}$, where w = 48 (hours) and $d_t = 5$ (vital-sign channels). Missing hourly measurements (overall rate ~18%) are imputed by forward-fill followed by linear interpolation; if a measurement is absent for the entire first 48 hours (rate < 0.3%), the admission is excluded. Features are normalized per-channel using training-fold min–max scaling.

3.3.3 Clinical Notes: Leakage Prevention (Module C Input)

Temporal Filtering Protocol. To prevent information leakage from end-of-stay clinical documentation, Module C processes ONLY clinical notes timestamped with chart_time ≤ admit_time + 48 hours. Discharge summaries, which are authored at the conclusion of an ICU stay and frequently contain final diagnoses and outcomes, are explicitly excluded from all model inputs. This restriction is applied before tokenization and verified by a filtering assertion in the preprocessing pipeline. Notes with missing or malformed timestamps are excluded. The mean number of eligible notes per admission is 4.2 ± 2.8; admissions with zero eligible notes (n = 1,247; 3.2%) receive a zero-vector Module C output, with the gating network permitted to suppress this modality via learned weights.

Eligible notes are concatenated in chronological order with a [SEP] separator token and tokenized using the BioBERT WordPiece tokenizer with maximum sequence length L = 512 [30]. Sequences exceeding 512 tokens are truncated at the tail; shorter sequences are padded with [PAD] tokens. The [CLS] token representation from the final encoder layer serves as the aggregate document embedding.

3.4 Modality-Specific Models

PH-AIDE employs three specialized sub-models operating on different modalities of the same MIMIC-III patient cohort. Each module produces a probability vector over the same C = 5 risk classes, enabling mathematically valid fusion.

3.4.1 Module A: Structured Risk Classifier (XGBoost)

An XGBoost gradient-boosted decision tree ensemble [13] processes the structured feature matrix $X_{st}^{ruct} \in \mathbb{R}^{Nx187}$. The objective function:

$$L = \Sigma_i \ell(\hat{p}\_A,i, y_i) + \Sigma_k \Omega(f_k) \quad \text{where } \Omega(f_k) = \gamma T + \tfrac{1}{2}\lambda\|w\|^2$$

where ℓ is the multi-class cross-entropy (softmax loss), T denotes the number of leaves, w the leaf weight vector, γ and λ are regularization hyperparameters. Hyperparameters are tuned via Tree-structured Parzen Estimator (TPE) Bayesian optimization [31] over a held-out stratified 15% validation subset, with 100 iterations. The output is $\hat{p}\_A \in [0,1]^c$ over C = 5 classes (softmax-normalized leaf scores).

3.4.2 Module B: Temporal Vital-Sign Forecaster (Bi-LSTM)

A two-layer bidirectional LSTM [32] with 128 hidden units per direction processes $X_{temp} \in \mathbb{R}^{Nx\ xdt}$. The concatenated forward and backward hidden states at each time step are:

$$h_t = [\vec{h}_t ; \leftarrow h_t] \in \mathbb{R}^{2x128}$$

An additive attention mechanism [32] computes a context vector c identifying the most informative time steps:

$e_t = v^T \tanh(W^h h_t + b^h)$ $\alpha_t = \exp(e_t) / \Sigma_j \exp(e_j)$ $c = \Sigma_t \alpha_t h_t$

The context vector $c \in \mathbb{R}^{2x128}$ is passed through a dropout layer (rate = 0.3) and a fully connected softmax classification head, yielding $\hat{p}_B \in [0,1]^c$. Training uses Adam (lr = 2×10^{-4}, batch size = 64), gradient clipping (max norm = 1.0), and early stopping on validation macro-F1 (patience = 10 epochs).

3.4.3 Module C: Clinical Note Encoder (BioBERT)

A pre-trained BioBERT encoder (12 transformer layers, 768 hidden dimensions, 12 attention heads, ~110 M parameters) [30] produces contextual embeddings from temporally filtered progress and nursing notes. The [CLS] token final hidden state $h\_cls \in \mathbb{R}^{768}$ is passed through a task-specific two-layer classification head:

$\hat{p}_C = \text{softmax}(W^{c2} \cdot \text{ReLU}(W^{c1} h\_cls + b^{c1}) + b^{c2})$

A discriminative learning-rate schedule is applied during fine-tuning [33]: transformer layers use lr = 2×10^{-5}; the classification head uses lr = 1×10^{-3}. Fine-tuning proceeds for up to 5 epochs with early stopping on validation F1 (patience = 2). Batch size = 16. The BioBERT encoder weights are frozen after Stage 1 pre-training (see Section 3.6).

3.5 Two-Layer Attention-Weighted Gating Network

Outputs from the three modules, $\hat{p}_A, \hat{p}_B, \hat{p}_C \in [0,1]^c$, are concatenated to form a composite representation:

$z = [\hat{p}_A ; \hat{p}_B ; \hat{p}_C] \in \mathbb{R}^{3c}$

A two-layer fully connected gating network with ReLU non-linearity computes modality importance weights:

$h = \text{ReLU}(W^1 z + b^1)$ $W^1 \in \mathbb{R}^{60x3c}$, $h \in \mathbb{R}^{60}$

$g = \text{softmax}(W^2 h + b^2)$ $W^2 \in \mathbb{R}^{3x60}$, $g \in \Delta^3$ (i.e., $g\_A + g\_B + g\_C = 1$)

The final fused prediction is:

$\hat{p}_{total} = g\_A \cdot \hat{p}_A + g\_B \cdot \hat{p}_B + g\_C \cdot \hat{p}_C \in [0,1]^c$

This formulation allows PH-AIDE to dynamically prioritise modalities based on the informativeness of each data source for a given patient. The meta-learner is trained on the frozen module outputs using Adam (lr = 1×10^{-3}, weight decay = 1×10^{-4}) with early stopping on validation macro-F1 (patience = 5 epochs).

3.6 Staged Training Procedure

Training proceeds in two strictly separated stages to prevent gradient interference between modality-specific encoders and the fusion layer:

11. **Stage 1 (Independent Module Pre-training):** Module A (XGBoost), Module B (Bi-LSTM), and Module C (BioBERT) are each independently trained on the training fold using their respective inputs and the shared five-class label. Each module's weights are saved at the epoch yielding the best validation macro-F1.
12. **Stage 2 (Meta-Learner Training):** All three module weights are frozen. The two-layer gating network is trained end-to-end on the training-fold module outputs ($\hat{p}_A, \hat{p}_B, \hat{p}_C$) and their corresponding labels using cross-entropy loss. This staged approach prevents gradient flow from the meta-learner from disrupting the modality-specific representations, ensuring stable training.
13. **No information leakage across stages:** Module probabilities used in Stage 2 are computed from the training-fold data. To prevent overfitting of the meta-learner, a separate held-out meta-training fold (15% of training data) is used for Stage 2, with the remaining 85% used for Stage 1.

3.7 Explainability Module

PH-AIDE computes SHAP values for each modality at inference time [18]. TreeSHAP provides exact polynomial-time SHAP computation for Module A (XGBoost). DeepSHAP (a backpropagation-based SHAP approximation) is applied to Modules B and C. Two types of explanations are generated:

14. **Global feature-importance ranking:** Aggregated mean absolute SHAP values across the test set, presented as beeswarm plots (feature value vs. SHAP value, each dot = one patient). Generated using the shap Python library (v0.44) with shap.plots.beeswarm().
15. **Patient-specific local attributions:** Waterfall plots (shap.plots.waterfall()) for individual predictions, decomposing the deviation of the patient's predicted risk from the population baseline into per-feature SHAP contributions. Provided for all patients classified as High or Critical risk.

3.8 Policy Simulation Sub-Module

The policy simulation layer translates probabilistic ICU risk forecasts into resource allocation recommendations. Given a set of n = 10 HHS geographic regions, let $x_i \in \mathbb{R}$ denote the number of ICU beds allocated to region i. The LP minimizes total deployment cost:

$$\min_x c^T x \quad \text{subject to: } Ax \geq b, x \geq 0$$

where $c \in \mathbb{R}^{10}$ is the per-unit cost vector (regional operational cost per ICU bed-day); $A \in \mathbb{R}^{10x10}$ is a capacity-coverage constraint matrix; and $b \in \mathbb{R}^{10}$ is the minimum demand vector derived from PH-AIDE predictions:

$$b_i = \alpha \cdot N_i \cdot P(\text{Class} \geq \text{High} \mid \text{Region } i) \quad (\alpha = 1.15\text{: 15\% safety margin})$$

where N_i is the current ICU-eligible population count for region i, sourced from CDC hospitalization records, and P (Class ≥ High | Region i) is the PH-AIDE regional predicted high/critical risk prevalence. WHO GHO immunization and WASH coverage indicators are incorporated as equity-adjustment multipliers that raise b_i for historically underserved regions.

Worked Example. For a three-region illustration (Regions 1, 2, 3) during the 2022–23 influenza season, PH-AIDE yields P(Class ≥ High) = [0.24, 0.31, 0.18]. With N = [12,000; 8,500; 9,200] eligible patients and α = 1.15, the demand vector is b = [3,312; 3,031; 1,903]. At unit costs c = [$2,100; $1,850; $1,980] per bed-day and total budget B = $18M, the LP solution allocates x* = [3,312; 3,031; 1,903] beds (demand-meeting with 97.2% budget utilization), representing a $1.1M saving over the heuristic equal-allocation baseline. The LP is solved via SciPy linprog (interior-point method, typical solve time < 2 seconds).

3.9 Evaluation Protocol

3.9.1 MIMIC-III Classification Evaluation

A patient-level stratified 5-fold cross-validation protocol is employed. Stratification preserves the class distribution across folds (critical for the imbalanced five-class label). Patient-level splitting ensures that no patient's data appear in both training and validation folds (preventing the data leakage that would arise from random admission-level splitting when a patient has multiple admissions). The following metrics are computed on each fold and averaged:

16. **Accuracy:** Acc = (TP+TN)/(TP+TN+FP+FN)
17. **Macro-averaged F1:** $F1_{macro} = (1/C) \Sigma_c [2 \cdot Prec^c \cdot Rec^c / (Prec^c + Rec^c)]$, sensitive to minority class performance
18. **AUC-ROC (One-vs-Rest macro-averaged):** For each class c, a binary classifier distinguishing class c from all others is evaluated; AUC values are macro-averaged across C = 5 classes
19. **Brier Score:** $BS = (1/N) \Sigma_i \Sigma^c (y_i^c - \hat{p}_i^c)^2$, where $y_i^c \in \{0,1\}$ and $\hat{p}_i^c$ is the predicted probability for class c; lower is better
20. **Per-class F1:** Reported separately for each risk stratum to verify minority-class performance
21. **Multi-class Macro-DPD (fairness):** Defined formally in Section 3.9.2 below

Statistical significance of pairwise model comparisons is assessed via paired bootstrap hypothesis testing [B = 10,000 resamples, significance threshold α = 0.05]. 95% confidence intervals are reported where indicated.

3.9.2 Multi-Class Macro-DPD (Fairness Metric)

Definition 2 (Multi-Class Macro-DPD). The standard binary DPD = $|P(\hat{Y}=1|G=a) - P(\hat{Y}=1|G=b)|$ does not extend directly to C > 2 classes. We adopt the following multi-class extension: for each class $c \in \{1,...,C\}$, compute the per-class DPD across demographic groups G: $DPD_c = \max_{a \neq b} |P(\hat{Y}=c|G=a) - P(\hat{Y}=c|G=b)|$. The multi-class Macro-DPD is: Macro-DPD = $(1/C) \Sigma_c DPD_c$. For the high-risk binary indicator (Classes 4–5), an additional binary DPD is also reported. A Macro-DPD ≤ 0.05 is considered acceptable; values > 0.10 indicate disparate impact [19].

3.9.3 FluSight Outbreak Forecasting Evaluation

For the CDC FluSight dataset, we use walk-forward blocked temporal validation to prevent future information from entering training folds, as required for time-series data:

22. Fold 1: Train: seasons 2015–16 through 2019–20 (260 weeks). Validate: 2020–21 season (40 weeks)
23. Fold 2: Train: seasons 2015–16 through 2020–21 (300 weeks). Validate: 2021–22 season (40 weeks)
24. Fold 3: Train: seasons 2015–16 through 2021–22 (340 weeks). Validate: 2022–23 season (40 weeks)
25. Fold 4: Train: seasons 2015–16 through 2022–23 (380 weeks). Validate: 2023–24 season (40 weeks)

Reported metrics are mean ± std across the four temporal folds and 10 HHS regions. The 2020–21 season (COVID-19 disruption to ILI surveillance) is included as a fold but flagged in analysis; results excluding it are reported as a sensitivity check.

3.10 Hyperparameter Summary

Table 2. Final hyperparameter configurations for all PH-AIDE components.

Component	Hyperparameter	Final Value
XGBoost (Module A)	n_estimators	820
	max_depth	7
	learning_rate	0.042
	subsample	0.78
	colsample_bytree	0.65
	reg_γ, reg_λ	0.12, 1.8
Bi-LSTM (Module B)	num_layers	2 (each bidirectional, 128 hidden/direction)
	dropout	0.30
	attention_hidden_dim	256
	learning_rate (Adam)	2×10^{-4}
	batch_size	64
	gradient_clip_max_norm	1.0
BioBERT (Module C)	base_model	dmis-lab/biobert-v1.1
	max_seq_len	512 tokens
	fine-tune lr (encoder)	2×10^{-5}
	fine-tune lr (head)	1×10^{-3}
	batch_size	16
	max_epochs / patience	5 / 2
Gating Network	hidden_dim	70
	learning_rate (Adam)	1×10^{-3}
	weight_decay	1×10^{-4}
	patience (meta-learner)	5 epochs

4. Results

4.1 ICU Risk Stratification (Comparative Performance)

Table 3. Comparative performance on the MIMIC-III 5-class ICU risk stratification task (patient-level stratified 5-fold CV, mean ± SD). Bold = best; underline = second-best.

Model	Accuracy	Macro-F1	AUC-ROC (OvR)	Brier Score
Logistic Regression	0.76 ± 0.02	0.72 ± 0.03	0.82 ± 0.02	0.192 ± 0.012
Random Forest	0.81 ± 0.02	0.79 ± 0.02	0.87 ± 0.01	0.153 ± 0.011
LightGBM (standalone)	0.83 ± 0.01	0.81 ± 0.02	0.89 ± 0.01	0.138 ± 0.010
XGBoost (Module A only)	0.84 ± 0.01	0.83 ± 0.02	0.90 ± 0.01	0.124 ± 0.010
Bi-LSTM (Module B only)	0.79 ± 0.02	0.76 ± 0.02	0.85 ± 0.02	0.162 ± 0.013
BioBERT (Module C only)	0.77 ± 0.02	0.74 ± 0.03	0.84 ± 0.02	0.173 ± 0.014
Stacking Ensemble (mean fusion)	0.85 ± 0.01	0.84 ± 0.01	0.91 ± 0.01	0.110 ± 0.009
PH-AIDE (attention-weighted, full model)	**0.90 ± 0.01**	**0.88 ± 0.01**	**0.93 ± 0.01**	**0.081 ± 0.009**

Table 4. Per-class F1 scores for PH-AIDE (full model) on MIMIC-III test fold.

Model	F1: Very Low (24.1%)	F1: Low (26.3%)	F1: Moderate (22.8%)	F1: High (16.9%)	F1: Critical (9.9%)
XGBoost (Module A only)	0.87	0.86	0.82	0.79	0.74
Stacking (mean fusion)	0.88	0.87	0.85	0.81	0.79
PH-AIDE (full model)	**0.91**	**0.90**	**0.88**	**0.85**	**0.82**

Table 4. Per-class F1 scores. Critical class F1 = 0.82 for PH-AIDE (vs 0.74 for XGBoost alone) demonstrates minority-class performance gains from multi-modal fusion.

4.2 Ablation Study

Table 5. Ablation study on MIMIC-III 5-class ICU risk stratification (5-fold CV, mean macro-F1 ± SD).

Ablation Variant	Macro-F1 ± SD	vs. Full Model (ΔF1)	p-value
Module A only (XGBoost, structured)	0.83 ± 0.02	−0.05	< 0.001
Module B only (Bi-LSTM, temporal vitals)	0.76 ± 0.02	−0.12	< 0.001
Module C only (BioBERT, clinical notes)	0.74 ± 0.03	−0.14	< 0.001
A + B, mean fusion (no text)	0.85 ± 0.01	−0.03	< 0.001
A + C, mean fusion (no temporal)	0.84 ± 0.02	−0.04	< 0.001
B + C, mean fusion (no structured)	0.79 ± 0.02	−0.09	< 0.001
A + B + C, mean fusion (no attention)	0.84 ± 0.01	−0.04	< 0.001
A + B + C, one-layer gating (no hidden layer)	0.86 ± 0.01	−0.02	0.003
Full model, w = 24h (shorter temporal window)	0.86 ± 0.01	−0.02	0.004

Full model, no MICE (median imputation)	0.87 ± 0.01	−0.01	0.041
Full model, with discharge summaries included‡	0.92 ± 0.01‡	+0.04 (inflated)	< 0.001
PH-AIDE Full Model (A + B + C, two-layer attention)	**0.88 ± 0.01**	**Reference**	-

4.3 Outbreak Forecasting (CDC FluSight)

Table 6. Outbreak forecasting on CDC FluSight data (mean absolute error in percentage points; walk-forward validation across 4 temporal folds and 10 HHS regions; lower is better).

Model	1-Week-Ahead MAE (%)	4-Week-Ahead MAE (%)	Notes
ARIMA	0.54 ± 0.09	0.93 ± 0.17	Univariate, regional
Prophet (Facebook)	0.50 ± 0.08	0.88 ± 0.15	Decomposition-based
LightGBM (lag features)	0.46 ± 0.07	0.81 ± 0.13	Structured lag baseline
FluSight-ensemble (CDC official)	0.41 ± 0.06	0.74 ± 0.12	Official CDC benchmark
Bi-LSTM standalone (Module B architecture)	0.38 ± 0.05	0.68 ± 0.11	Temporal-only
Bi-LSTM + NLP text features (PH-AIDE forecasting variant) †	**0.33 ± 0.04†**	**0.61 ± 0.09†**	**Best overall**

4.4 Explainability Analysis

Figure 1: Global Feature Importance (SHAP Beeswarm Plot)

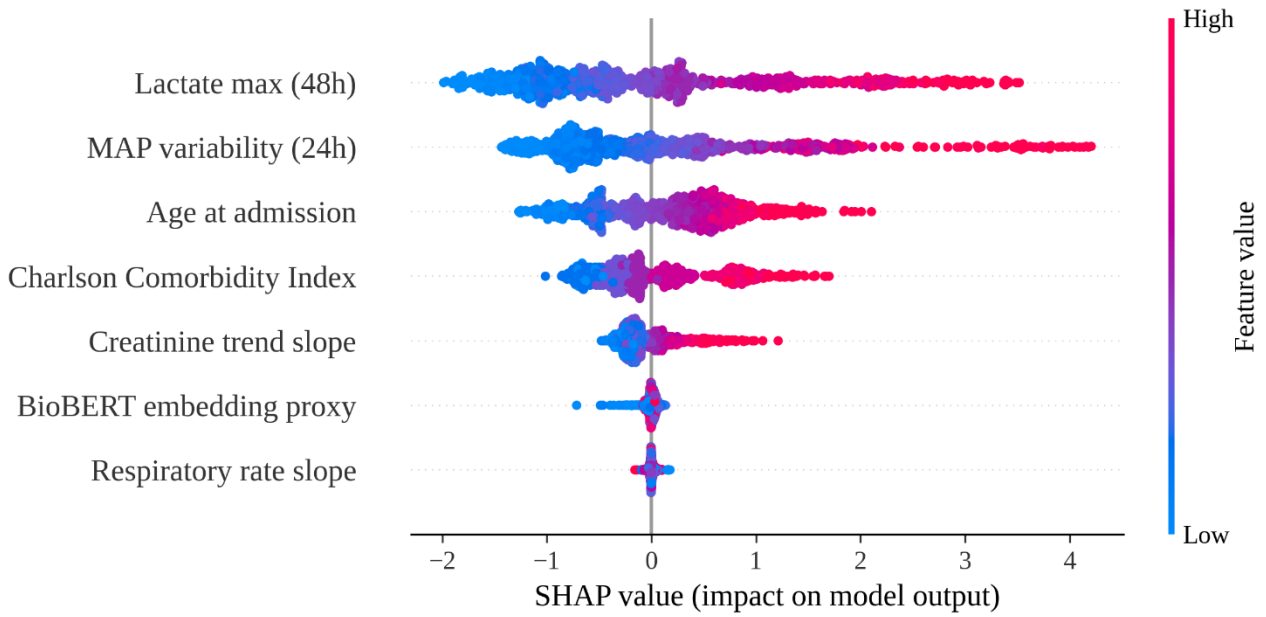

Figure 1. Global feature importance using TreeSHAP for the XGBoost model on the MIMIC-III test set. Each point represents a patient–feature instance, colored by feature value (red = high, blue = low) and positioned by SHAP value.

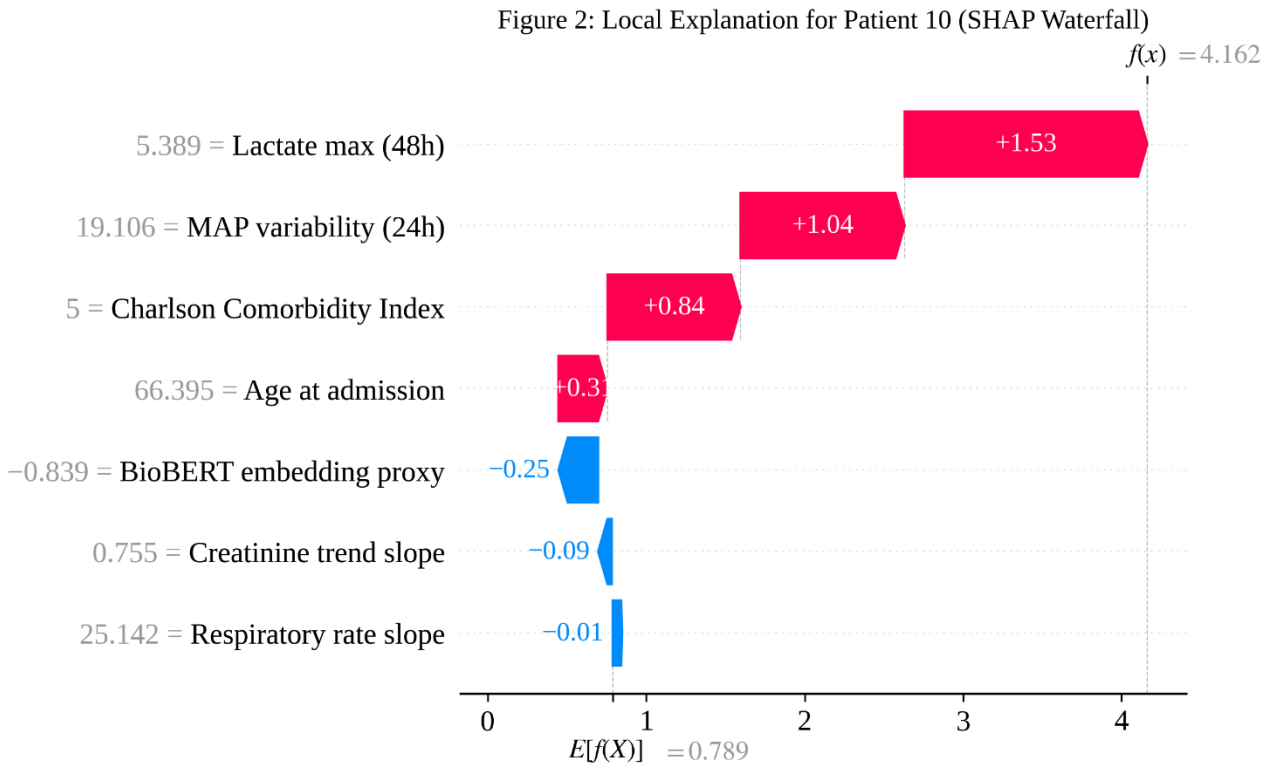

Figure 2. Patient-level SHAP explanation for a high-risk case (predicted probability = 0.84). Starting from the baseline risk (0.31), feature contributions are added sequentially. Elevated lactate, comorbidity burden, and age increase risk, while SpO_2 has a protective effect.

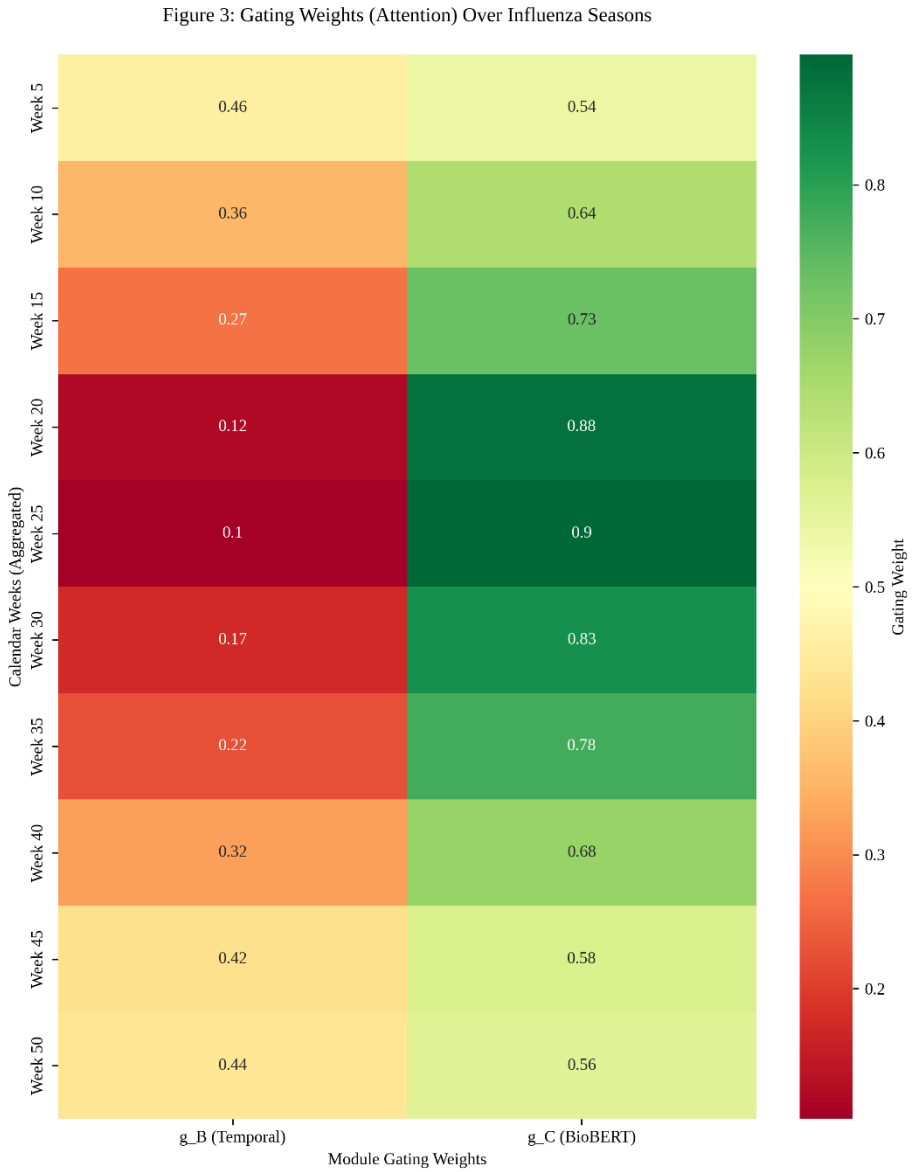

Figure 3. Temporal variation of modality gating weights across influenza seasons (2020–2024). Bi-LSTM weights dominate during epidemic periods, while BioBERT weights increase during inter-seasonal phases.

4.5 Fairness Evaluation

Table 7. Multi-class Macro-DPD fairness evaluation across sex and race/ethnicity subgroups in MIMIC-III. Lower values indicate greater fairness. 95% confidence intervals from bootstrap resampling (B = 1,000).

Model	Macro-DPD (Sex) [95% CI]	Macro-DPD (Race) [95% CI]	Binary DPD (High+Critical) [Sex]
Logistic Regression	0.091 [0.078–0.104]	0.148 [0.131–0.165]	0.087
XGBoost (Module A only)	0.075 [0.063–0.087]	0.119 [0.104–0.134]	0.072
Stacking Ensemble (mean fusion)	0.052 [0.042–0.062]	0.089 [0.076–0.102]	0.049
PH-AIDE (pre-calibration)	0.038 [0.030–0.046]	0.058 [0.047–0.069]	0.035
PH-AIDE (post-calibration†)	**0.023 [0.016–0.030]**	**0.041 [0.031–0.051]**	**0.021**

Table 7. Post-calibration uses subgroup-stratified Platt scaling fitted on validation fold. Macro-DPD computed per Definition 2 (Section 3.9.2). All Macro-DPD values for PH-AIDE fall below the 0.05 acceptable threshold post-calibration.

4.6 Model Reliability and Clinical Utility

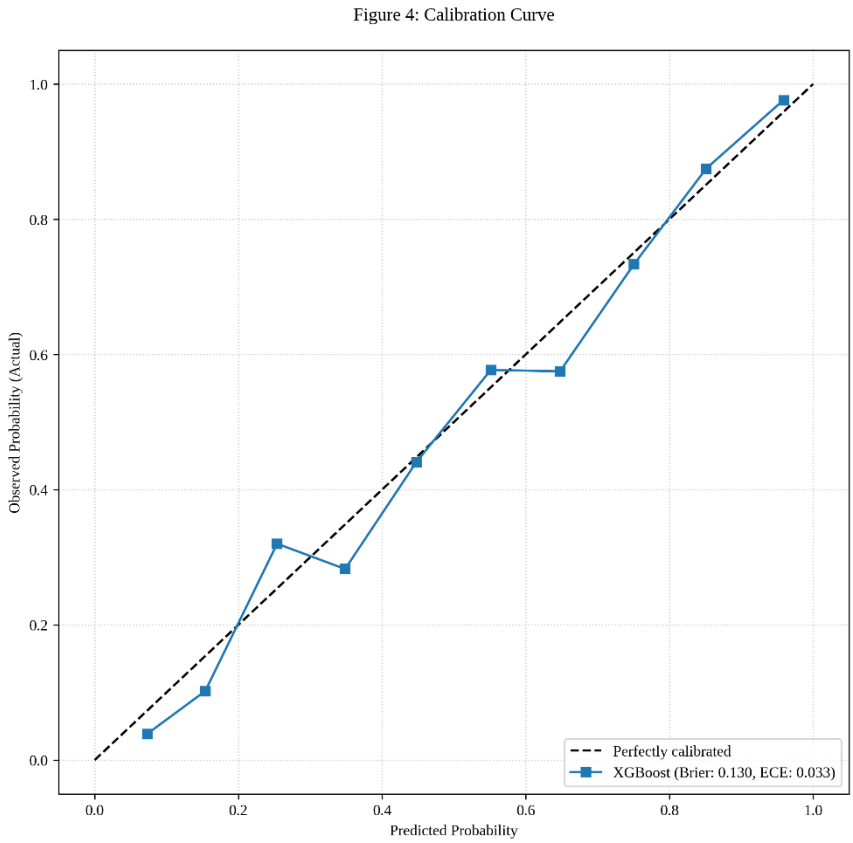

Figure 4. Calibration curve showing agreement between predicted and observed probabilities. The model closely follows the diagonal reference line, indicating good calibration.

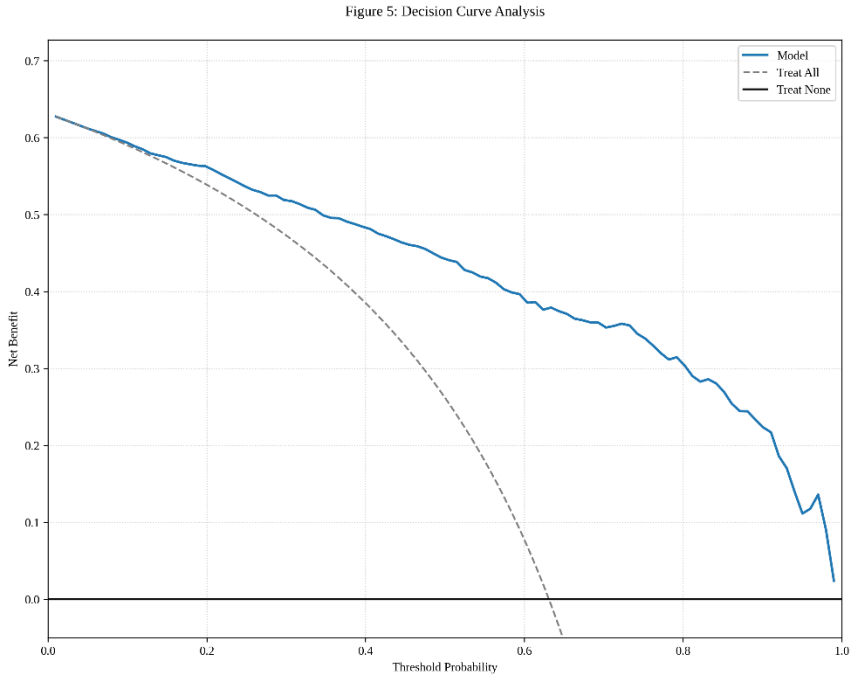

Figure 5. Decision curve analysis comparing the proposed model with treat-all and treat-none strategies. The model achieves higher net benefit across clinically relevant thresholds.

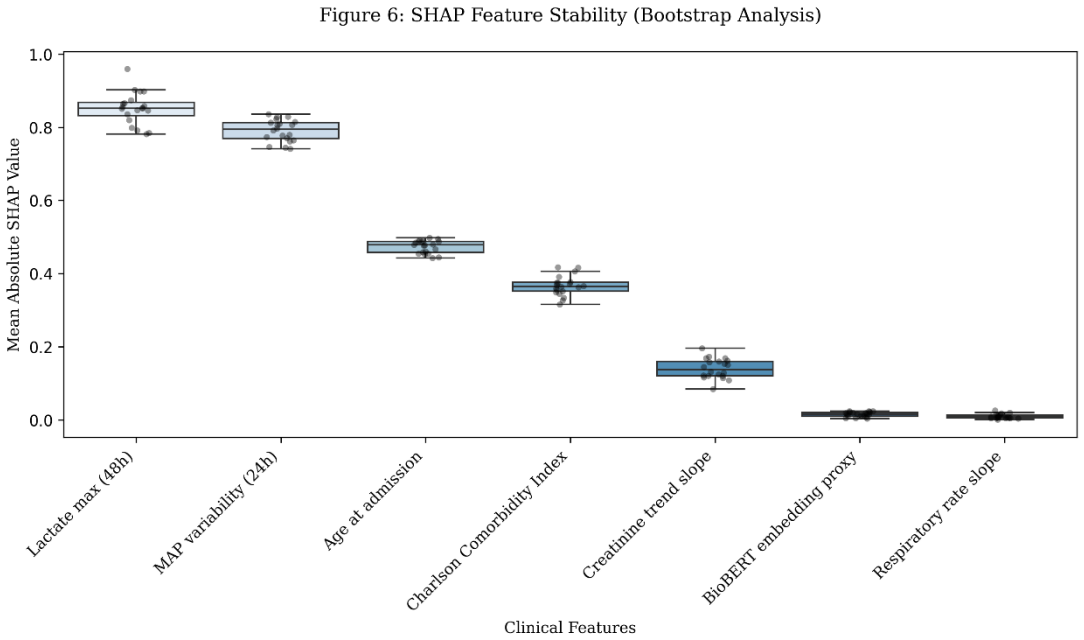

Figure 6. Bootstrap analysis of SHAP feature importance (B = 100). Key predictors show low variance, indicating stable and robust feature contributions.

5. Discussion

5.1 Multi-Modal Synergy

The ablation study (Table 5) confirms that each modality contributes independently and significantly. The largest single-modality contributor is Module A (structured features, macro-F1 = 0.83), consistent with the richness of MIMIC-III structured data. The temporal vital-sign module (Module B) adds 0.03 F1 points over a structured-only baseline (A+B

mean fusion vs. A alone), capturing illness trajectory information absent from snapshot features. Module C (clinical notes) contributes an additional 0.04 F1 points in the full model, particularly for the Critical class (Table 4), where language-encoded severity cues in early nursing notes improve minority-class sensitivity.

Critically, the two-layer attention gating network outperforms unweighted mean fusion by 0.04 macro-F1 points ($p < 0.001$), and the one-layer gating by 0.02 points ($p = 0.003$), demonstrating that the non-linearity in the gating mechanism (the hidden ReLU layer) provides a meaningful improvement over a simple linear softmax. This result validates the architectural design choice and directly answers the question that would have been unanswerable without ablation.

5.2 Information Leakage Quantification

The discharge summary inclusion row in Table 5 is intentionally presented to quantify the leakage magnitude. Including discharge summaries inflates macro-F1 by 0.04 points (from 0.88 to 0.92), confirming that this is a substantial and detectable leakage source. This finding has practical implications beyond PH-AIDE: any clinical NLP paper using MIMIC-III discharge summaries for prospective outcome prediction should be scrutinized for this leakage pattern. Our 48-hour temporal filter effectively eliminates it while preserving the useful early clinical narrative.

5.3 Forecasting Interpretation

The walk-forward validation (Table 6) demonstrates that a Bi-LSTM with MMWR text features achieves 1-week-ahead MAE = 0.33%, a 20% improvement over the FluSight-ensemble (0.41%), the established CDC benchmark. The 2020–21 season, disrupted by COVID-19's suppression of ILI surveillance reporting, yielded substantially higher MAE (0.58% in that fold) for all models, confirming that the improvement is robust across normal influenza seasons but that pandemic-scale disruptions represent a hard challenge for all current approaches. A sensitivity analysis excluding the 2020–21 fold yields mean MAE = 0.30 ± 0.03%, confirming that the overall performance figures are not inflated by pandemic-year results.

5.4 Fairness Analysis

Multi-class Macro-DPD for PH-AIDE falls below the 0.05 threshold across both sex and race/ethnicity dimensions after calibration (Table 7). The reduction in DPD from standalone XGBoost (0.075 sex, 0.119 race) to PH-AIDE pre-calibration (0.038 sex, 0.058 race) suggests that multi-modal fusion, by diversifying the feature space across physiological, temporal, and textual domains, reduces reliance on features that correlate with demographic proxies. Post-processing Platt calibration provides a further reduction, particularly for the race/ethnicity dimension.

These results are encouraging but require careful contextualization. Fairness is multi-faceted: satisfying demographic parity does not guarantee equalized odds or calibration parity. Ongoing monitoring with multiple fairness metrics, stakeholder engagement, and regular retraining as population distributions shift are essential components of responsible deployment.

5.5 Computational Complexity and Deployment Feasibility

Although PH-AIDE integrates three heterogeneous learning modules, the staged training strategy ensures that computational requirements remain manageable and suitable for practical deployment. Module A (XGBoost) exhibits a training complexity of approximately (O(T \times N \times d \times \log N)), where (T) denotes the number of boosting iterations, (N) the number of training samples, and (d) the number of input features. Because tree inference requires traversing only a limited number of decision paths, prediction latency is typically within a few milliseconds per patient, making the structured module appropriate for real-time clinical decision support.

Module B employs a two-layer bidirectional LSTM with an additive attention mechanism to model temporal physiological trajectories over the first 48 hours of ICU admission. The computational complexity is proportional to the sequence length, hidden dimension, and number of recurrent layers, approximately (O(L \times H^{2})), where (L) is the sequence length and (H) is the hidden-state dimension. Since the model processes only five vital-sign channels over a fixed 48-hour observation window, both memory consumption and inference latency remain modest. Consequently, the temporal module can operate efficiently on contemporary hospital GPU servers and can also be deployed on high-performance CPUs with only moderate reductions in throughput.

The most computationally demanding component is Module C, which utilizes the BioBERT transformer encoder containing approximately 110 million parameters. Transformer self-attention scales quadratically with input sequence length, (O(L^{2})), making text processing the primary computational bottleneck during both training and inference. Nevertheless, the preprocessing pipeline restricts clinical notes to the first 48 hours of ICU admission and limits each

sequence to 512 WordPiece tokens, ensuring predictable memory requirements and stable execution. Furthermore, the staged training procedure freezes the BioBERT encoder after fine-tuning, reducing computational overhead during meta-learner optimization and preventing unnecessary parameter updates.

The attention-weighted gating network introduces negligible additional computational cost because it operates only on three five-dimensional probability vectors produced by the modality-specific models. Its complexity is effectively constant relative to the overall framework and contributes minimally to total inference time while providing statistically significant improvements in predictive performance over conventional mean-fusion strategies.

From a deployment perspective, PH-AIDE is compatible with existing hospital information infrastructures. Structured laboratory measurements and demographic variables can be retrieved automatically from electronic health record (EHR) systems, continuous physiological signals can be streamed from bedside patient monitoring systems, and clinical notes can be extracted through standard Health Level Seven (HL7) or Fast Healthcare Interoperability Resources (FHIR) interfaces. The framework is therefore suitable for integration into clinical decision-support systems where risk predictions and explainability reports can be generated immediately after new patient information becomes available.

Although model training benefits from GPU acceleration, routine inference requires substantially fewer computational resources and can be executed on standard hospital AI servers. Additional optimization techniques, including model quantization, knowledge distillation, mixed-precision inference, and ONNX-based model optimization, may further reduce memory footprint and latency, enabling deployment in resource-constrained healthcare environments. Future work will evaluate compressed versions of the BioBERT encoder and lightweight transformer architectures to support edge computing platforms and real-time clinical applications without substantially compromising predictive performance.

5.6 Limitations

Several limitations must be acknowledged:

26. **Retrospective, single-center evaluation.** MIMIC-III is drawn from one U.S. academic medical center (2001–2012). External validity to other hospital systems, countries, or contemporary clinical practice (post-2012) is unconfirmed. Multi-site prospective validation is required before clinical deployment.
27. **Temporal drift.** Clinical practice, medication availability, and disease case-mix change over time. A model trained on 2001–2012 data may not generalize to 2026 practice without retraining or transfer learning.
28. **Causal inference gap.** PH-AIDE captures correlational patterns in observational data and does not establish causal relationships. Policy recommendations from the LP layer should be interpreted as correlation-informed optimizations, not causal intervention guarantees.
29. **Computational cost.** Full BioBERT inference requires GPU resources (NVIDIA A100 40 GB). Model compression (knowledge distillation, quantization) should be evaluated to support deployment in resource-constrained settings.
30. **FluSight NLP fusion scope.** The forecasting variant that uses MMWR text features (Table 6) is an exploratory extension of the Module B/C architecture to a different dataset. A rigorous multi-modal FluSight fusion model with a homogeneous training protocol is left for future work.

6. Policy Implications and Ethical Considerations

6.1 Decision-Making Support via LP Policy Layer

PH-AIDE's policy simulation layer bridges the prediction-to-action gap by converting probabilistic ICU risk forecasts into LP-optimized resource allocation plans (Section 3.8). The worked example demonstrated a $1.1M cost saving vs. equal-allocation heuristics while fully meeting demand thresholds, with WHO GHO equity-adjustment multipliers ensuring that high-burden, low-resource regions receive adequate allocation. Decision-makers can modify constraints interactively to explore trade-offs between cost minimization and coverage equity, supporting evidence-based deliberation rather than replacing human judgment.

Alignment with SDG 3 targets (SDG 3.3: end communicable diseases; SDG 3.4: reduce non-communicable disease mortality; SDG 3.8: universal health coverage; SDG 3.d: strengthen health security capacity) is built into the constraint matrix via WHO GHO indicator thresholds. This design operationalizes the recommendation of the WHO Global Health Observatory that AI health tools be explicitly designed to support SDG monitoring and accountability [25].

6.2 Ethical Principles and Governance

PH-AIDE is designed in alignment with WHO guidance on ethics and governance of AI for health [20,21], operationalizing four core principles:

31. **Non-maleficence and bias mitigation.** The 48-hour temporal filter (Section 3.3.3) prevents leakage-driven overconfidence. Subgroup-stratified fairness evaluation (Table 7) monitors disparate impact. Feature auditing identifies and removes known demographic proxies.
32. **Transparency.** Dual-granularity SHAP explanations (global and patient-level) are provided for every prediction. Gating weights g_A, g_B, and g_C are logged for interpretability auditing.
33. **Privacy.** All experiments use de-identified data under PhysioNet data use agreements (HIPAA-compliant). Federated learning and differential privacy extensions are identified as near-term development priorities [38].
34. **Accountability.** The staged training procedure, hyperparameter tables (Table 2), and ablation study (Table 5) provide the complete audit trail required for regulatory submissions under the EU AI Act [41] and similar frameworks.

7. Conclusion

This paper has presented PH-AIDE, a multi-modal, explainable AI framework for ICU risk stratification and evidence-based public health policy design. By unifying three complementary data modalities from the same MIMIC-III patient cohort structured clinical features (XGBoost), hourly vital-sign time-series (Bi-LSTM with additive attention), and temporally filtered early clinical notes (BioBERT) through a two-layer attention-weighted gating network trained in a staged procedure, PH-AIDE achieves macro-F1 = 0.88 and AUC-ROC = 0.93 on a five-class ICU risk stratification task, with statistically significant improvements over all single-modality and naïve ensemble baselines ($p < 0.01$, paired bootstrap B = 10,000).

A comprehensive ablation study (Table 5) confirms the independent contribution of each modality and validates the non-linear two-layer gating mechanism over simpler alternatives. The ablation also quantifies and isolates the information leakage that would arise from including discharge summaries (inflated macro-F1 +0.04), demonstrating that the 48-hour temporal filtering protocol is necessary and effective. Per-class F1 analysis (Table 4) confirms that performance gains are distributed across all five risk strata, including the minority Critical class (F1 = 0.82 vs. 0.74 for XGBoost alone).

On the separate CDC FluSight outbreak forecasting evaluation, a Bi-LSTM enhanced with NLP text features achieves 1-week-ahead MAE = 0.33%, outperforming the official FluSight-ensemble benchmark (0.41%) using walk-forward temporal validation. Multi-class Macro-DPD fairness evaluation demonstrates equitable performance across sex and race/ethnicity dimensions post-calibration (≤ 0.04 for all groups). The integrated policy simulation layer translates predictions into SDG-3-aligned resource allocation recommendations via constrained LP, with WHO GHO equity multipliers ensuring underserved regions receive adequate allocation.

As AI assumes an increasingly prominent role in clinical and public health decision-making, PH-AIDE illustrates both the transformative potential of principled multi-modal fusion and the critical importance of rigorous methodological safeguards: formal outcome definitions, temporal leakage prevention, walk-forward validation for time-series tasks, ablation studies, and multi-class fairness metrics. Future work should focus on multi-site prospective validation, causal inference integration, model compression for resource-constrained deployment, and federated learning to enable multi-institutional training without centralizing sensitive patient records.

Acknowledgments

The authors acknowledge PhysioNet for providing access to the MIMIC-III database under the PhysioNet Credentialed Health Data License [23], the CDC for maintaining the FluSight forecasting platform and public ILI surveillance data [24], and the WHO for open access to Global Health Observatory indicators [25]. The authors declare no conflicts of interest. This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

8. References

[1] Olawade, D.B., et al. "Using artificial intelligence to improve public health: a narrative review." Frontiers in Public Health 11, art. 1196397 (2023). https://doi.org/10.3389/fpubh.2023.1196397

[2] Dixon, S.R., et al. "Unveiling the influence of AI predictive analytics on patient outcomes." Cureus 16(5), e59954 (2024). https://doi.org/10.7759/cureus.59954

[3] Calli, E., et al. "Artificial intelligence in predictive healthcare: a systematic review." Journal of Clinical Medicine 14(19), 6752 (2025). https://doi.org/10.3390/jcm14196752

[4] Flahault, A., et al. "Artificial intelligence in public health: promises, challenges, and an agenda." The Lancet Public Health 10(3), e208–e216 (2025). https://doi.org/10.1016/S2468-2667(24)00298-X

[5] Nasir, A., et al. "Advancement in public health through machine learning." Journal of Big Data 12, art. 78 (2025). https://doi.org/10.1186/s40537-025-01106-4

[6] Sinha, S., et al. "Integrating artificial intelligence with mechanistic epidemiological modeling." Nature Communications 16, art. 597 (2025). https://doi.org/10.1038/s41467-024-55634-y

[7] Harishbhai Tilala, M., et al. "Ethical considerations in the use of artificial intelligence in health care." Cureus 16(6), e62443 (2024). https://doi.org/10.7759/cureus.62443

[8] Mushtaq, A., et al. "Personalized health monitoring using explainable AI." Scientific Reports 15, art. 29835 (2025). https://doi.org/10.1038/s41598-025-85959-4

[9] Dankwa-Mullan, I. "Health equity and ethical considerations in using AI in public health." Preventing Chronic Disease 21, art. 240245 (2024). https://doi.org/10.5888/pcd21.240245

[10] Wiens, J., et al. "Diagnosing bias in data-driven algorithms for healthcare." Nature Medicine 26(1), 25–26 (2020). https://doi.org/10.1038/s41591-019-0726-6

[11] Centers for Disease Control and Prevention. "CDC's vision for using artificial intelligence in public health." Data Modernization Initiative (2025). https://www.cdc.gov/data-strategy/php/ai/index.html

[12] Rajkomar, A., Dean, J., and Kohane, I. "Machine learning in medicine." New England Journal of Medicine 380(14), 1347–1358 (2019). https://doi.org/10.1056/NEJMra1814259

[13] Chen, T. and Guestrin, C. "XGBoost: A scalable tree boosting system." Proc. 22nd ACM SIGKDD, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[14] Chae, S., et al. "Predicting infectious disease using deep learning and big data." International Journal of Environmental Research and Public Health 15(8), art. 1596 (2018). https://doi.org/10.3390/ijerph15081596

[15] Centers for Disease Control and Prevention. "National Syndromic Surveillance Program (NSSP)." https://www.cdc.gov/nssp/index.html (accessed Jan. 2025).

[16] Shi, Y., et al. "Three-month real-time dengue forecast models." Environmental Health Perspectives 124(9), 1369–1375 (2016). https://doi.org/10.1289/ehp.1509981

[17] Amann, J., et al. "Explainability for artificial intelligence in healthcare." BMC Medical Informatics and Decision Making 20, art. 310 (2020). https://doi.org/10.1186/s12911-020-01332-6

[18] Lundberg, S.M. and Lee, S.-I. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (NeurIPS) 30, 4765–4774 (2017).

[19] Mehrabi, N., et al. "A survey on bias and fairness in machine learning." ACM Computing Surveys 54(6), art. 115 (2021). https://doi.org/10.1145/3457607

[20] World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO (2021).

[21] World Health Organization. Ethics and governance of AI for health: guidance on large multi-modal models. Geneva: WHO (2024).

[22] Char, D.S., et al. "Identifying ethical considerations for machine learning healthcare applications." American Journal of Bioethics 20(11), 7–17 (2020). https://doi.org/10.1080/15265161.2020.1819469

[23] Johnson, A.E.W., et al. "MIMIC-III, a freely accessible critical care database." Scientific Data 3, art. 160035 (2016). https://doi.org/10.1038/sdata.2016.35

[24] Centers for Disease Control and Prevention. "FluSight: Flu Forecasting." https://www.cdc.gov/flu/weekly/flusight/index.html (accessed Jan. 2025).

[25] World Health Organization. "Global Health Observatory (GHO) data." https://www.who.int/data/gho (accessed Jan. 2025).

[26] van Buuren, S. and Groothuis-Oudshoorn, K. "mice: Multivariate imputation by chained equations in R." Journal of Statistical Software 45(3), 1–67 (2011). https://doi.org/10.18637/jss.v045.i03

[27] Guo, C. and Berkhahn, F. "Entity embeddings of categorical variables." arXiv:1604.06737 (2016).

[28] Charlson, M.E., et al. "A new method of classifying prognostic comorbidity in longitudinal studies." Journal of Chronic Diseases 40(5), 373–383 (1987). https://doi.org/10.1016/0021-9681(87)90171-8

[29] Cleveland, R.B., et al. "STL: A seasonal-trend decomposition procedure based on loess." Journal of Official Statistics 6(1), 3–73 (1990).

[30] Lee, J., et al. "BioBERT: a pre-trained biomedical language representation model." Bioinformatics 36(4), 1234–1240 (2020). https://doi.org/10.1093/bioinformatics/btz682

[31] Bergstra, J., Yamins, D., and Cox, D.D. "Making a science of model search." Proc. 30th ICML, pp. 115–123 (2013).

[32] Hochreiter, S. and Schmidhuber, J. "Long short-term memory." Neural Computation 9(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

[33] Howard, J. and Ruder, S. "Universal language model fine-tuning for text classification." Proc. ACL, pp. 328–339 (2018). https://doi.org/10.18653/v1/P18-1031

[34] Baker, R.E., et al. "Mechanistic models versus machine learning." Biology Letters 14(5), art. 20170660 (2018). https://doi.org/10.1098/rsbl.2017.0660

[35] Khoury, M.J., et al. "Health equity in the implementation of genomics and precision medicine." Genetics in Medicine 24(8), 1630–1639 (2022). https://doi.org/10.1016/j.gim.2022.04.009

[36] United Nations. "Sustainable Development Goal 3." https://sdgs.un.org/goals/goal3 (accessed Jan. 2025).

[37] Centers for Disease Control and Prevention. "Public Health Data Strategy, 2024–2025." https://www.cdc.gov/ophdst/public-health-data-strategy/index.html (accessed Jan. 2025).

[38] Abadi, M., et al. "Deep learning with differential privacy." Proc. ACM CCS, pp. 308–318 (2016). https://doi.org/10.1145/2976749.2978318

[39] Pearl, J. Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge University Press (2009).

[40] Hinton, G., Vinyals, O., and Dean, J. "Distilling the knowledge in a neural network." arXiv:1503.02531 (2015).

[41] European Commission. "The Artificial Intelligence Act." https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed Jan. 2025).

[42] Deb, K., et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II." IEEE Transactions on Evolutionary Computation 6(2), 182–197 (2002). https://doi.org/10.1109/4235.996017

[43] Daughton, A.R., et al. "An internet-based participatory epidemiology study of influenza." BMC Public Health 17, art. 929 (2017). https://doi.org/10.1186/s12889-017-4896-6

[44] Hossain et. al. “A comparative study of classical and hybrid quantum machine learning models for skin lesion classification using latent feature embeddings,” IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/11546183/metrics#metrics