applied sciences

MDPI

Article

Uncertainty-Aware Predictive Process Monitoring in Healthcare: Explainable Insights into Probability Calibration for Conformal Prediction

Maxim Majlatow [1,2], Fahim Ahmed Shakil [1,2], Andreas Emrich [1,2] and Nijat Mehdiyev [1,2,*]

[1] German Research Center for Artificial Intelligence (DFKI), 66123 Saarbrücken, Germany; maxim.majlatow@dfki.de (M.M.); fahim_ahmed.shakil@dfki.de (F.A.S.); andreas.emrich@dfki.de (A.E.)
[2] Institute for Information Systems (IWi), Saarland University, 66123 Saarbrücken, Germany
* Correspondence: nijat.mehdiyev@dfki.de

Abstract

In high-stakes decision-making environments, predictive models must deliver not only high accuracy but also reliable uncertainty estimations and transparent explanations. This study explores the integration of probability calibration techniques with Conformal Prediction (CP) within a predictive process monitoring (PPM) framework tailored to healthcare analytics. CP is renowned for its distribution-free prediction regions and formal coverage guarantees under minimal assumptions; however, its practical utility critically depends on well-calibrated probability estimates. We compare a range of post-hoc calibration methods—including parametric approaches like Platt scaling and Beta calibration, as well as non-parametric techniques such as Isotonic Regression and Spline calibration—to assess their impact on aligning raw model outputs with observed outcomes. By incorporating these calibrated probabilities into the CP framework, our multilayer analysis evaluates improvements in prediction region validity, including tighter coverage gaps and reduced minority error contributions. Furthermore, we employ SHAP-based explainability to explain how calibration influences feature attribution for both high-confidence and ambiguous predictions. Experimental results on process-driven healthcare data indicate that the integration of calibration with CP not only enhances the statistical robustness of uncertainty estimates but also improves the interpretability of predictions, thereby supporting safer and robust clinical decision-making.

Keywords: conformal prediction; explainable artificial intelligence; probability calibration; predictive process monitoring

Academic Editors: Rui Araújo and Jia-Lien Hsu

Received: 15 May 2025
Revised: 9 July 2025
Accepted: 10 July 2025
Published: 16 July 2025

Citation: Majlatow, M.; Shakil, F.A.; Emrich, A.; Mehdiyev, N. Uncertainty-Aware Predictive Process Monitoring in Healthcare: Explainable Insights into Probability Calibration for Conformal Prediction. *Appl. Sci.* **2025**, *15*, 7925. https://doi.org/10.3390/app15147925

1. Introduction

High-stakes decision-making, especially in sectors such as healthcare, demands that predictive models be not only accurate but also robust, transparent, and trustworthy [1]. In these settings, the cost of erroneous predictions is exceptionally high, as misjudgments in risk assessment can lead to delayed treatments, misallocated resources, or even adverse patient outcomes [2]. As a result, ensuring that a model reliably quantifies uncertainty and provides clear, interpretable explanations is paramount [3].

Conformal Prediction (CP) has emerged as a powerful framework for uncertainty quantification (UQ) because it generates prediction regions with formal, distribution-free coverage guarantees under minimal assumptions, most notably exchangeability [4–6]. This characteristic is especially beneficial in high-stakes applications where traditional

probabilistic assumptions may be violated or hard to verify [7,8]. By constructing prediction sets that are statistically valid regardless of the underlying data distribution, CP offers a principled approach to ensure that the true outcome is captured with a pre-specified level of confidence [9]. However, the practical impact of CP is critically dependent on the quality of the underlying probability estimates [10,11]. In many real-world scenarios, raw model outputs are miscalibrated—that is, the probabilities produced do not accurately reflect the true likelihood of events [12]. This miscalibration creates a disconnect between the statistical coverage guarantees provided by CP and the actual risk profile observed in practice. In high-stakes environments, such a gap can compromise both the interpretability of the prediction regions and the reliability of subsequent decision-making processes [13]. To address these challenges, researchers have developed a variety of post-hoc probability calibration techniques. Parametric methods, such as Platt scaling and beta calibration, impose a functional form to remap the raw outputs, while non-parametric approaches like isotonic regression and spline calibration offer the flexibility to capture more complex miscalibration patterns that often emerge in heterogeneous data [14].

Our work adopts a multilayer analytical framework that integrates these calibration techniques with CP to enhance the reliability and interpretability of predictive process monitoring (PPM) in high-stakes settings [15]. At the first layer, calibrated probability estimates serve as a more faithful representation of the true event likelihoods. When these refined probabilities are embedded within the CP framework, the resulting prediction regions are not only statistically valid but also more reflective of real-world risk. This integration is critical because tighter and more reliable prediction intervals directly translate into better-informed decision making.

Moreover, a second layer of analysis is introduced through explainable artificial intelligence (XAI) techniques—specifically, methods such as SHAP (SHapley Additive exPlanations)—which explain the contributions of individual features to both high-confidence and ambiguous predictions. Such transparency is essential for building trust among domain experts and facilitating accountability in the deployment of AI systems. Preliminary analyses suggest that ensemble-based methods, particularly those relying on gradient boosting, may deliver superior performance on imbalanced and complex datasets. When combined with non-parametric calibration approaches, these models can more effectively capture subtle, non-linear patterns of miscalibration, thereby aligning predicted probabilities with observed outcomes more closely.

Furthermore, integrating calibrated predictions into CP frameworks is expected to yield prediction regions with reduced coverage gaps and lower minority error contributions—a crucial advancement for applications where underestimating risk for a minority class can have significant repercussions. In healthcare, errors are not equal; a failure to predict a critical event (a minority class error) is often far more dangerous than other mistakes. By first using calibration to correct the systematically low probabilities often assigned to rare events, our framework ensures that the subsequent CP step is less likely to make these critical errors, thus enhancing patient safety. While the initial focus of this work is on the methodological integration of calibration, CP, and XAI, later sections will introduce clinical and healthcare applications to demonstrate how these advanced UQ and interpretability techniques can be practically deployed in settings such as PPM for patient care. By harmonizing robust statistical guarantees with transparent, interpretable insights, our approach seeks to pave the way for more reliable and actionable AI in high-stakes decision-making environments.

The remainder of the paper is organized as follows. Section 2 describes the PPM use case from healthcare domain. Section 3 presents our methodology, including calibration techniques, CP, and explainability methods. Section 4 outlines the experimental setup and

evaluation metrics. Section 5 reports the results. Section 6 discusses the findings and their implications. Section 7 reviews related work, and Section 8 concludes the paper with final remarks and future research directions.

2. Use Case Description

Context and Scope. In our study, we focus on a process mining initiative that was conducted by [16] at a regional hospital in the Netherlands, a facility with approximately 700 beds across multiple locations and an annual patient intake of around 50,000 individuals. An event log was constructed through SQL-based extraction, anonymized, and archived in the 4TU. Center for Research Data repository [17]. Process mining—a data-driven methodology for analyzing workflows via event logs enables here to address the complexity of emergency care pathways, with a focus on sepsis management. By leveraging process mining, our study specifically aims to predictively analyze trajectories leading to Admission to the Intensive Care Unit (ICU), a high-stakes transition reflecting escalating care needs and systemic instability. Sepsis, a leading cause of ICU admissions and in-hospital mortality, was selected due to its standardized diagnostic criteria (Systemic Inflammatory Response Syndrome, SIRS) and time-sensitive treatment protocols.

Data Collection and Integration. Data spanning 1.5 years (November 2013–June 2015) encompass 1050 sepsis patient cases, yielding 15,214 timestamped events. These were extracted from three heterogeneous source systems: triage documentation, laboratory systems, and financial/administrative databases (see Figure 1). Triage records provide granular details such as symptom checklists, diagnostic order timestamps, and administration times for intravenous antibiotics and fluids. Laboratory data include blood test results for leukocytes, C-reactive protein (CRP), and lactic acid, while financial and administrative systems track admissions, care transitions (e.g., transfers to intensive care), discharges, and post-treatment trajectories. Process mining's strength in reconciling multi-source data proved critical here, resolving inconsistencies (e.g., timestamp alignment, unit conversions) to reconstruct temporally coherent patient pathways. This structured event log enables the identification of patterns preceding ICU transfers, such as delayed antibiotic administration or abnormal vital signs, which are often obscured in siloed healthcare datasets.

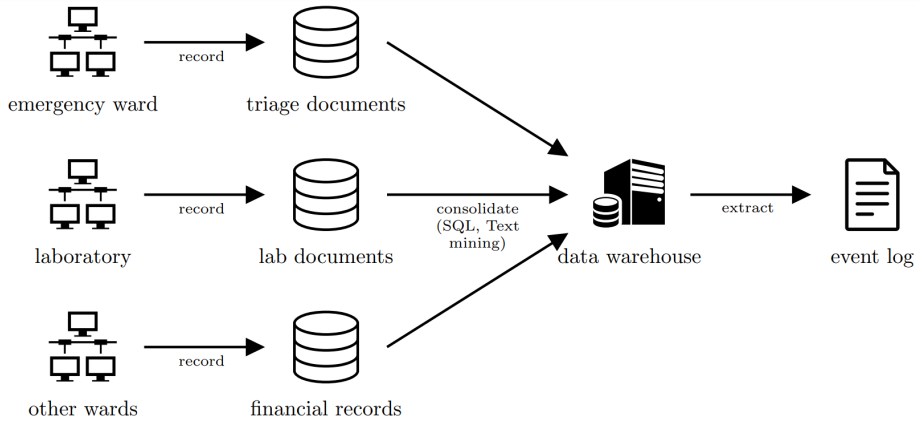


Figure 1. The consolidation process of data from multiple source systems into a single event log as described by [16].

Event Log Structure and Attributes. The final event log comprised 16 activity classes organized into clinically meaningful categories: registration/triage (e.g., ER Sepsis Triage), diagnostic procedures (leukocyte, CRP, and lactic acid measurements), treatment interventions (IV administration), care transitions (admissions, transfers), discharge processes (five variants), and critical escalation events (ICU admissions). Each event was enriched

with 28 attributes, including patient age, anonymized timestamps (preserving inter-event durations), blood test values, diagnostic findings (e.g., organ dysfunction), and logistical metadata such as care team assignments and clinical flags (e.g., hypotension, hypoxia). Table 1 provides an illustrative excerpt, demonstrating the log's granularity. For instance, a single case (ID: B) spans registration, triage, diagnostic tests, and sepsis-specific interventions, with CRP values indicating severe inflammation (240.0 mg/L). The dataset captured 890 unique process variants, reflecting diverse pathways such as direct ICU admissions, delayed transfers after initial stabilization, and discharges without escalation—key insights for understanding risk stratification in sepsis care.

Table 1. Excerpt from the utilized event log data.

Case ID	Activity	Timestamp	Resource ID	…	Age	…	CRP	…
B	ER Registration	21-12-2014 11:04:24	A	…	45	…	-	…
B	ER Triage	21-12-2014 11:17:19	C	…	45	…	-	…
B	CRP	21-12-2014 11:36:00	B	…	45	…	240.0	…
B	LacticAcid	21-12-2014 11:36:00	B	…	45	…	-	…
B	Leucocytes	21-12-2014 11:36:00	B	…	45	…	-	…
B	ER Sepsis Triage	21-12-2014 12:15:45	A	…	45	…	-	…
…	…	…	…	…	…	…		
J	Admission NC	02-01-2014 20:09:47	F	…	80	…	-	…
J	CRP	04-01-2014 08:00:00	B	…	80	…	43.0	…
J	Release A	06-01-2014 11:00:00	E	…	80	…	-	…
…	…	…	…	…	…	…		

Focus of Our Research. In this work, we focus on forecasting Admission to ICU, a critical juncture in sepsis care where timely intervention can significantly influence patient outcomes. Our methodological pipeline extends the conventional application of classification approaches by three more additional integrated components: (1) probability calibration to align predictions with observed ICU admission rates, (2) CP to quantify uncertainty through statistically valid prediction intervals, and (3) explainability mechanisms to elucidate the drivers of model uncertainty. The danger of miscalibration can be illustrated with a practical example. Consider a clinical decision support system designed to predict a sepsis patient's risk of ICU admission. A hospital protocol may mandate an immediate specialist consultation if the predicted risk exceeds a 30% threshold. A model could be highly accurate in discriminating between patients (e.g., have a high AUROC) yet be systematically underconfident, predicting a 25% risk for a patient whose true risk is 40%. In this case, the life-saving consultation is not triggered due to the miscalibrated probability, leading to a delayed intervention and a potential adverse outcome. This highlights that for a model to be clinically useful, its confidence scores must be reliable. This gap between a model's discriminative power and the trustworthiness of its probability estimates is a central challenge that our work addresses. The subsequent section formalizes this framework, detailing the interplay between these components and establishing evaluation mechanisms.

3. Methodology

Figure 2 provides a comprehensive overview of the proposed pipeline, which systematically enhances the reliability and interpretability of well-calibrated uncertainty-aware binary classification models. The process begins with data cleaning, process mining, and feature engineering to transform raw event logs into a structured dataset. This dataset is then split into training, calibration, and test sets. In the supervised learning phase, multiple classifiers, including Decision Tree, Random Forest, XGBoost, and CatBoost, undergo hyperparameter tuning to optimize predictive performance. To improve the trustworthiness of probability estimates, post-hoc calibration techniques are applied, allowing

for a comparative evaluation of their effectiveness. Further, UQ is incorporated via split conformal prediction (SCP) to assess how different calibration approaches interact with uncertainty estimation. This iterative exploration of classifier performance, calibration reliability, and conformal uncertainty measures provides a robust framework for producing well-calibrated risk scores with quantified confidence levels. Finally, SHAP analysis is employed to explain model predictions, distinguishing between certain and uncertain classifications by attributing importance scores to individual features. Our framework supports robustness through bootstrapped evaluation and statistical testing. It ensures transparency by using SHAP-based feature attribution to clarify feature contributions to certain and uncertain predictions. Trustworthiness is supported through probability calibration to align predicted risks with observed outcomes and SCP to provide formal coverage guarantees and is quantified using the metrics described in Sections 4.3 and 4.4. The interplay between these components ensures a rigorous, interpretable, and data-driven decision-making pipeline.

3.1. Predictive Process Monitoring

Process Data Definition. PPM formalizes the task of forecasting process outcomes from partial execution traces. In the context of sepsis care, this translates to predicting whether a patient will be admitted to the ICU based on their evolving hospital trajectory. Mathematically, a process event $e \in E$ represents a timestamped action in a patient's care pathway, structured as a tuple $(a, c, t_{\text{start}}, t_{\text{complete}}, v_1, \ldots, v_n)$, where a denotes the activity label (e.g., `ER Registration`, `CRP Measurement`), c identifies the unique patient case, and t_{start} and t_{complete} correspond to the start and completion timestamps. The attributes v_i encapsulate event-specific clinical markers such as lactic acid levels or compliance with SIRS criteria. The event universe $E = A \times C \times \mathbb{T} \times \mathbb{T} \times V_1 \times \cdots \times V_n$ captures all possible interactions in the care pathway. Projection functions allow the extraction of individual event components, facilitating a granular analysis of activities, timestamps, and clinical parameters. A patient's care trajectory is represented as a trace $\sigma_c = \langle e_1, \ldots, e_{|\sigma_c|} \rangle$, where events are ordered by start time. For predictive monitoring, the prefixes $\text{hd}_i(\sigma_c) = \langle e_1, \ldots, e_i \rangle$ denote partial execution sequences, and suffixes $\text{tl}_i(\sigma_c)$ represent the remaining events. The event log $E_C = \{\sigma_c \mid c \in C\}$ aggregates all patient trajectories, serving as the foundation for predictive modeling.

Data Preprocessing. The dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ is constructed via a feature function $\text{feat} : E^* \to \mathbb{R}^d$, where $\mathbf{x}_i = \text{feat}(\text{hd}_i(\sigma_c))$ encodes key attributes relevant to ICU admission risks. Temporal features capture elapsed time since triage and intervals between critical actions such as antibiotic administration. Clinical attributes include blood test results, leukocyte counts, and indicators of sepsis severity based on SIRS criteria. Sequential patterns track the frequency of ICU transfers and variations in discharge protocols, revealing deviations from standard care pathways. To enhance predictive accuracy, the original event log was transformed into a structured dataset incorporating 30 additional process-specific features. Temporal dynamics were represented through metrics such as elapsed time between consecutive events, average hour of care activities, and total duration from registration to discharge. Activity patterns captured workflow inefficiencies by analyzing the maximum frequency of repeated activities. Sequential transitions were encoded through binary indicators for critical event pairs, such as `IV Antibiotics → ICU Transfer`, highlighting deviations from expected clinical progressions. A structured preprocessing pipeline ensured statistical rigor and interpretability. Outlier removal was applied to filter extreme values in event durations and laboratory results, refining the dataset to 995 high-confidence cases. Robust scaling standardized numerical features, such as leukocyte counts and time intervals, using median and inter-quartile range adjustments

to mitigate the influence of extreme values. Class imbalance, a critical challenge given the ICU admission rate of approximately 10%, was addressed through stratified sampling, preserving the natural distribution in training, calibration and test sets without resorting to synthetic data generation, thereby maintaining the temporal integrity of patient trajectories.

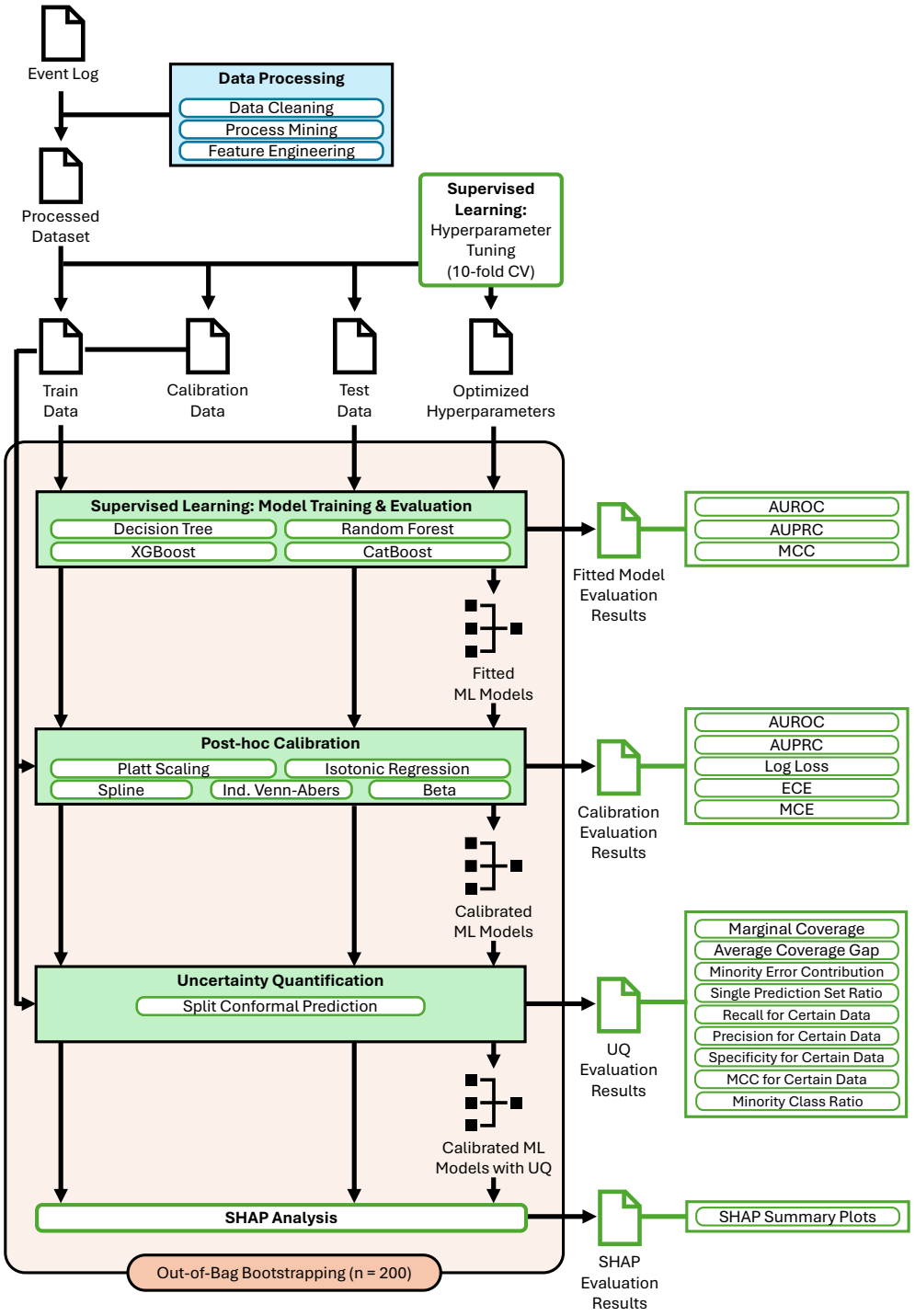


Figure 2. Overview of the proposed framework.

Supervised Learning. The predictive task is formulated as a binary classification problem that maps partial traces to ICU-admission outcomes. A labeling function outcome : $E_C \to 0,1$ assigns $y_c = \text{outcome}(\sigma_c)$, where $y_c = 1$ if patient c was admitted to the ICU and $y_c = 0$ otherwise. A probabilistic classifier $M : \mathbb{R}^d \to [0,1]$ estimates the posterior probability $\Pr(y_i = 1 \mid \mathbf{x}_i)$, producing risk scores $\hat{p}_i = M(\mathbf{x}_i)$. These scores are

converted to binary predictions $\hat{y}i = \mathbb{I}(\hat{p}i \geq \tau)$, where the threshold τ controls the trade-off between sensitivity and specificity. The model is trained by minimizing a loss function $\mathcal{L}(M, \mathcal{D}\text{train})$ over the training subset $\mathcal{D}\text{train}$, thereby optimizing its ability to map partial traces to ICU-admission outcomes with high predictive accuracy.

3.2. Calibration Methods

Despite achieving high predictive accuracy, models often produce probability estimates that diverge from true empirical likelihoods—a critical concern when probabilistic outputs inform risk-sensitive decisions such as admission to ICU. For example, a predicted probability of $\hat{p}_i = 0.9$ might correspond to observed positive outcomes in only 70% of similar cases, indicating systematic overconfidence. Calibration rectifies such miscalibrations by post-processing raw scores to ensure reliability: calibrated probabilities $\hat{p}_i^{\text{cal}} = \phi(\hat{p}_i)$ must satisfy the statistical consistency condition:

$$\mathbb{E}[y_i \mid \phi(\hat{p}_i) = p] = p \quad \forall p \in [0, 1] \tag{1}$$

where $\phi : [0, 1] \to [0, 1]$ is a calibration function learned from held-out data.

The calibration workflow mandates three distinct partitions of the dataset $\mathcal{D}$: $\mathcal{D}_{\text{train}}$ trains the base model M, $\mathcal{D}_{\text{cal}}$ fits the calibration function ϕ, and $\mathcal{D}_{\text{test}}$ evaluates the calibrated system. This strict separation prevents information leakage, as ϕ adapts to M's biases without overfitting to test data. Formally, the optimal ϕ^* minimizes a calibration-specific loss:

$$\phi^* = \arg\min_{\phi \in \Phi} \sum_{(\mathbf{x}_j, y_j) \in \mathcal{D}_{\text{cal}}} \mathcal{L}_{\text{cal}}\big(y_j, \phi(\hat{p}_j)\big) \tag{2}$$

where Φ denotes the hypothesis space of calibration functions. Parametric families Φ enforce interpretable mappings at the cost of rigid assumptions, while non-parametric approaches (e.g., isotonic regression) flexibly adapt to arbitrary miscalibration patterns.

Calibration proves indispensable in operational settings where predicted probabilities directly influence critical decision-making. The selection of a probability calibration method is a critical, yet often overlooked, step in building trustworthy predictive models. There is no single method that is universally superior; parametric approaches are simple and robust to small calibration sets, while non-parametric methods offer greater flexibility to correct complex miscalibration patterns. Our study therefore, employs a comparative approach, evaluating a range of techniques. The rationale for this is to investigate our central hypothesis: that the choice of calibration method has significant and differing downstream consequences for the safety of uncertainty estimates (via CP) and the transparency of the model's reasoning (via XAI). By comparing multiple methods, we can expose the critical trade-offs between them, providing a more complete picture of how to build a reliable and interpretable system. The subsequent sections detail calibration paradigms that address distinct challenges, including handling class imbalance, ensuring robustness against rare but critical escalation events, and preserving temporal consistency in dynamic patient care pathways.

3.2.1. Platt Scaling

Platt Scaling, also known as logistic calibration, corrects model miscalibration by applying a logistic transformation to raw prediction scores $\hat{p}_i = M(\mathbf{x}_i)$ [18]. This method assumes a sigmoidal relationship between uncalibrated outputs and true probabilities in the log-odds space. The calibration function is formally defined as:

$$\phi_{\text{Platt}}(\hat{p}_i) = \sigma(a \cdot \text{logit}(\hat{p}_i) + b), \tag{3}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ denotes the logistic sigmoid, $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$ converts probabilities to log-odds, and parameters $a, b \in \mathbb{R}$ are optimized on the calibration set $\mathcal{D}_{\text{cal}}$. These parameters minimize the negative log-likelihood objective:

$$(a^*, b^*) = \arg\min_{a,b} \sum_{(\mathbf{x}_j, y_j) \in \mathcal{D}_{\text{cal}}} \left[y_j \ln \phi_{\text{Platt}}(\hat{p}_j) + (1 - y_j) \ln\left(1 - \phi_{\text{Platt}}(\hat{p}_j)\right) \right]. \tag{4}$$

The coefficients a and b adjust the slope and intercept of the sigmoid curve, respectively, counteracting systematic biases in the base model's predictions. For instance, if M exhibits overconfidence (e.g., assigning $\hat{p}_i = 0.9$ to cases where only 70% are positive), Platt Scaling compensates by learning $a < 1$, effectively flattening the sigmoid to produce conservative probability estimates.

3.2.2. Isotonic Regression

Isotonic Regression addresses calibration through a non-parametric, monotonic transformation of raw model scores $\hat{p}_i$. Unlike parametric methods like Platt Scaling, it makes no assumptions about the functional form of miscalibration, instead learning a piecewise constant calibration function ϕ_{Iso} that preserves the ordinal relationship between scores. Formally, the calibration function satisfies:

$$\phi_{\text{Iso}}(\hat{p}_j) \leq \phi_{\text{Iso}}(\hat{p}_k) \quad \text{whenever} \quad \hat{p}_j \leq \hat{p}_k, \tag{5}$$

ensuring that higher raw scores never map to lower calibrated probabilities.

The optimal calibration function minimizes the squared error over $\mathcal{D}_{\text{cal}}$:

$$\phi^*_{\text{Iso}} = \arg\min_{\phi \in \mathcal{F}_{\text{mono}}} \sum_{(\mathbf{x}_j, y_j) \in \mathcal{D}_{\text{cal}}} \left(y_j - \phi(\hat{p}_j)\right)^2, \tag{6}$$

where $\mathcal{F}_{\text{mono}}$ is the class of all monotonic non-decreasing functions. This optimization is solved via the Pool Adjacent Violators (PAV) algorithm, which iteratively merges adjacent score intervals until monotonicity constraints are satisfied. For a sorted sequence of predictions $\hat{p}_1 \leq \hat{p}_2 \leq \cdots \leq \hat{p}_m$, the algorithm partitions them into K bins $\{B_1, \ldots, B_K\}$, assigning each bin a calibrated probability:

$$\phi_{\text{Iso}}(B_k) = \frac{1}{|B_k|} \sum_{j \in B_k} y_j. \tag{7}$$

3.2.3. Beta Calibration

Beta Calibration generalizes logistic calibration by modeling miscalibration through a parametric family of Beta distributions. This method extends Platt Scaling's two-parameter sigmoid to a three-parameter function, enabling correction of asymmetric miscalibration patterns. The calibration function is defined as:

$$\phi_{\text{Beta}}(\hat{p}_i) = F(a \cdot \text{logit}(\hat{p}_i) + b; \alpha, \beta), \tag{8}$$

where $F(\cdot; \alpha, \beta)$ is the cumulative distribution function (CDF) of a Beta distribution with shape parameters $\alpha, \beta > 0$, and $a, b \in \mathbb{R}$ scale and shift the log-odds scores. The additional parameters α and β provide flexibility to model skewed or heavy-tailed deviations from calibration.

Parameters $\{a, b, \alpha, \beta\}$ are jointly optimized on $\mathcal{D}_{\text{cal}}$ via maximum likelihood estimation:

$$\{a^*, b^*, \alpha^*, \beta^*\} = \arg\min_{a,b,\alpha,\beta} \sum_{(\mathbf{x}_j, y_j) \in \mathcal{D}_{\text{cal}}} \left[y_j \ln \phi_{\text{Beta}}(\hat{p}_j) + (1 - y_j) \ln\left(1 - \phi_{\text{Beta}}(\hat{p}_j)\right) \right]. \tag{9}$$

Beta Calibration addresses limitations of Platt Scaling when miscalibration is non-sigmoidal. While more flexible than Platt Scaling, Beta Calibration requires larger $\mathcal{D}_{\text{cal}}$ sizes to robustly estimate four parameters. Overfitting risks emerge when calibration data is sparse, often mitigated via Bayesian priors on α and β.

3.2.4. Venn-Abers

Venn-Abers calibration provides a transductive framework for probability calibration rooted in CP, offering distribution-free validity guarantees under the assumption of exchangeability. Unlike parametric or isotonic methods, it outputs calibrated probability intervals rather than point estimates, making it uniquely suited for applications requiring rigorous UQ.

Given a base model M, Venn-Abers calibration operates on the calibration set $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{m}$ by first defining a conformity score $s(\mathbf{x}, y)$, often chosen as $s(\mathbf{x}, y) = y \cdot \hat{p}_i + (1 - y) \cdot (1 - \hat{p}_i)$, where $\hat{p}_i = M(\mathbf{x}_i)$. For each new instance $\mathbf{x}_{\text{new}}$, it computes two smoothed probability estimates:

$$p_0 = \frac{|\{j \in \mathcal{D}_{\text{cal}} \cup (\mathbf{x}_{\text{new}}, 0) \mid s(\mathbf{x}_j, y_j) \leq s(\mathbf{x}_{\text{new}}, 0)\}|}{m+1}, \tag{10}$$

$$p_1 = \frac{|\{j \in \mathcal{D}_{\text{cal}} \cup (\mathbf{x}_{\text{new}}, 1) \mid s(\mathbf{x}_j, y_j) \leq s(\mathbf{x}_{\text{new}}, 1)\}|}{m+1}, \tag{11}$$

where p_0 and p_1 represent the empirical probabilities of observing conformity scores at least as extreme as the hypothetical labels $y_{\text{new}} = 0$ and $y_{\text{new}} = 1$, respectively. The calibrated probability interval is then $[\min(p_0, p_1), \max(p_0, p_1)]$, with the point estimate $\hat{p}_{\text{VA}} = \frac{p_1}{p_0 + p_1}$.

3.2.5. Spline Calibration

Spline Calibration combines the flexibility of non-parametric methods with the smoothness of parametric approaches by modeling the calibration function ϕ_{Spline} as a piecewise polynomial. This method partitions the raw score range $[0, 1]$ into K intervals (knots) and fits a polynomial of degree d within each interval, constrained for continuity and smoothness at knot boundaries. For cubic splines ($d = 3$), the calibration function takes the form:

$$\phi_{\text{Spline}}(\hat{p}_i) = \sum_{k=1}^{K} \beta_k B_k(\hat{p}_i), \tag{12}$$

where $B_k(\cdot)$ are basis functions (e.g., B-splines) and β_k are coefficients learned from $\mathcal{D}_{\text{cal}}$.

The optimization objective minimizes a penalized squared error:

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \sum_{(\mathbf{x}_j, y_j) \in \mathcal{D}_{\text{cal}}} \left(y_j - \sum_{k=1}^{K} \beta_k B_k(\hat{p}_j) \right)^2 + \lambda \int_0^1 \left[\phi''_{\text{Spline}}(p) \right]^2 dp, \tag{13}$$

where $\lambda \geq 0$ controls the trade-off between fit and smoothness, penalizing large curvature in ϕ_{Spline}.

Spline calibration adapts to diverse miscalibration patterns while avoiding the staircase artifacts of Isotonic Regression. For example, raw scores clustered near $\hat{p}_i = 0.8$ with an empirical positive rate of 60% can be smoothly adjusted downward without abrupt binning. The number of knots K and penalty λ are tuned via cross-validation on $\mathcal{D}_{\text{cal}}$, balancing underfitting and overfitting risks.

3.3. Conformal Prediction

CP extends UQ to generate provably valid prediction sets for binary outcomes, ensuring coverage guarantees without distributional assumptions. The framework's validity

rests on the minimal assumption of exchangeability, which posits that the calibration data and new test instances are drawn from the same underlying data-generating process, regardless of its form. This is a significant advantage in healthcare, where data is often too complex to fit traditional parametric distributions. By relying only on exchangeability, CP remains valid even when used with complex "black-box" models whose outputs have unknown distributions.

Given a binary classifier $M : \mathcal{X} \to [0,1]$ producing estimates $\hat{p}_i = \Pr(y_i = 1 \mid \mathbf{x}_i)$, a non-conformity score $S(\mathbf{x}, y)$ quantifies model uncertainty for labels $y \in \{0,1\}$:

$$S(\mathbf{x}, y) = \begin{cases} 1 - \hat{p}_i & \text{if } y = 1, \\ \hat{p}_i & \text{if } y = 0, \end{cases} \tag{14}$$

where lower scores indicate stronger agreement between $\mathbf{x}_i$ and y. Using the calibration set $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{m}$, scores $s_j = S(\mathbf{x}_j, y_j)$ are computed, and the $(1-\alpha)$-quantile $\hat{q}$ is derived as:

$$\hat{q} = \inf\left\{ q \in \mathbb{R} : \frac{|\{j \in \mathcal{D}_{\text{cal}} \cup \{(\mathbf{x}_{n+1}, y_{n+1})\} : s_j \leq q\}|}{m+1} \geq 1 - \alpha \right\} \tag{15}$$

For a new instance $\mathbf{x}_{n+1}$, the prediction set $\mathcal{C}(\mathbf{x}_{n+1}) \subseteq \{0,1\}$ is:

$$\mathcal{C}(\mathbf{x}_{n+1}) = \{y \in \{0,1\} : S(\mathbf{x}_{n+1}, y) \leq \hat{q}\} \tag{16}$$

Under exchangeability of $\mathcal{D}_{\text{cal}}$ and test data, CP guarantees:

$$\Pr(y_{n+1} \in \mathcal{C}(\mathbf{x}_{n+1})) \geq 1 - \alpha \tag{17}$$

irrespective of M's accuracy. Prediction sets may yield confident predictions ($\mathcal{C} = \{0\}$ or $\{1\}$) or abstain ($\mathcal{C} = \{0,1\}$) when uncertainty exceeds α. In PPM, this enables risk-aware decision-making by flagging uncertain cases for human review while ensuring auditability through provable coverage rates. The framework's validity depends critically on exchangeability—a challenge in temporal processes with concept drift, addressed in subsequent subsections via adaptive methods.

In this study, we adopt the SCP approach, a computationally efficient and widely used variant of the original transductive framework. Also known as Inductive Conformal Prediction, SCP decouples the calibration from the prediction phase, making it highly scalable and practical for real-world applications. The procedure involves the following steps: First, the available data is partitioned into two disjoint sets: a proper training set, $\mathcal{D}_{\text{train}}$, and a calibration set, $\mathcal{D}_{\text{cal}}$. The model M is trained exclusively on $\mathcal{D}_{\text{train}}$. Second, the trained model is used to compute a non-conformity score, $s_j = S(\mathbf{x}_j, y_j)$, for every data point $(\mathbf{x}_j, y_j)$ in the held-out calibration set. This yields a set of $m = |\mathcal{D}_{\text{cal}}|$ calibration scores, $\{s_1, s_2, \ldots, s_m\}$, which provides an empirical measure of the errors the trained model makes on data it has not seen before. Third, this set of calibration scores is used to determine a single critical threshold, $\hat{q}$, that will guarantee the desired coverage rate, $1 - \alpha$. To account for finite sample effects, this threshold is calculated as the appropriate empirical quantile of the calibration scores. Let $s_{(1)} \leq s_{(2)} \leq \ldots \leq s_{(m)}$ be the scores sorted in non-decreasing order. The threshold $\hat{q}$ is set to the k-th smallest score:

$$\hat{q} = s_{(k)}, \quad \text{where} \quad k = \lceil (m+1)(1-\alpha) \rceil \tag{18}$$

If $k > m$, we can consider $\hat{q} = \infty$, ensuring all prediction sets are valid. This value of $\hat{q}$ is computed only once and is then fixed.

Finally, for any new test instance $\mathbf{x}_{n+1}$, the prediction set $\mathcal{C}(\mathbf{x}_{n+1})$ is constructed using the same rule as before, comparing the non-conformity scores of potential labels against the fixed threshold $\hat{q}$:

$$\mathcal{C}(\mathbf{x}_{n+1}) = \{y \in \{0,1\} : S(\mathbf{x}_{n+1}, y) \leq \hat{q}\} \tag{19}$$

The crucial advantage is that this step does not require access to the calibration set or retraining the model. Under the assumption that the instances in $\mathcal{D}_{\text{cal}}$ and the new test instances are exchangeable, SCP provides the same formal coverage guarantee, $\Pr(y_{n+1} \in \mathcal{C}(\mathbf{x}_{n+1})) \geq 1 - \alpha$.

For temporal process data, where exchangeability may be violated due to concept drift, SCP provides a baseline for UQ. Its simplicity and speed make it particularly useful in high-throughput environments, such as real-time fraud detection, where models must generate auditable predictions without computational overhead.

3.4. Explainability of Uncertain Predictions via SHAP

CP (Section 3.3) determines where a model is confident in its outputs ($|\mathcal{C}(\mathbf{x}_i)| = 1$) or remains uncertain ($|\mathcal{C}(\mathbf{x}_i)| = 2$). Here, $\mathcal{C}(\mathbf{x}_i)$ denotes the set of plausible labels returned by the conformal predictor for an instance $\mathbf{x}_i$. When $\mathcal{C}(\mathbf{x}_i)$ contains only one label, the prediction is deemed certain. Conversely, if it contains two labels, the prediction is uncertain, indicating an elevated level of ambiguity.

Despite highlighting these uncertain cases, CP alone does not explain why such ambiguity arises. To understand the drivers behind model certainty and uncertainty, we employ an explainable AI (XAI) approach. While several XAI methods exist, we selected SHAP (SHapley Additive exPlanations) [19] due to its distinct advantages for our research questions. First, SHAP provides local, instance-level explanations, which are essential for analyzing why an individual prediction results in a certain (single-label) or uncertain (multi-label) conformal set. Second, its foundation in game theory ensures theoretical guarantees of consistency and accuracy in attributing feature importance, overcoming the potential instability of other methods like LIME.

SHAP Formalization. Consider a binary classifier $M : \mathcal{X} \to [0,1]$ with a log-odds representation:

$$f(\mathbf{x}) = (M(\mathbf{x})) = \ln\Big(\frac{M(\mathbf{x})}{1 - M(\mathbf{x})}\Big). \tag{20}$$

For a given instance $\mathbf{x}_i$ and feature j, the SHAP value ϕ_{ij} is defined so that

$$f(\mathbf{x}_i) = \phi_{i0} + \sum_{k=1}^{d} \phi_{ik}, \tag{21}$$

where $\phi_{i0} = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\text{train}}}[f(\mathbf{x})]$ is the baseline expectation. Each ϕ_{ij} quantifies the contribution of feature j to the instance's deviation from the baseline, computed via

$$\phi_{ij} = \sum_{S \subseteq \{1,\ldots,d\} \setminus \{j\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \big[f(S \cup \{j\}) - f(S)\big], \tag{22}$$

where $f(S)$ is the conditional expectation of $f(\mathbf{x})$ given the subset S.

Cluster Analysis Across Calibration Approaches. We apply this SHAP-based explanation framework to the CP sets $\mathcal{C}(\mathbf{x}_i)$ derived from various calibration methods. Specifically, we partition the test set $\mathcal{D}_{\text{test}}$ into:

$$\begin{aligned} \mathcal{C}_{\text{certain}} &= \{\mathbf{x}_i \in \mathcal{D}_{\text{test}} : |\mathcal{C}(\mathbf{x}_i)| = 1\}, \\ \mathcal{C}_{\text{uncertain}} &= \{\mathbf{x}_i \in \mathcal{D}_{\text{test}} : |\mathcal{C}(\mathbf{x}_i)| = 2\}. \end{aligned} \tag{23}$$

For both certain and uncertain groups, we calculate the mean absolute SHAP values per feature:

$$\begin{aligned} \overline{\phi}_j^{\text{certain}} &= \frac{1}{|\mathcal{C}_{\text{certain}}|} \sum_{\mathbf{x}_i \in \mathcal{C}_{\text{certain}}} |\phi_{ij}|, \\ \overline{\phi}_j^{\text{uncertain}} &= \frac{1}{|\mathcal{C}_{\text{uncertain}}|} \sum_{\mathbf{x}_i \in \mathcal{C}_{\text{uncertain}}} |\phi_{ij}|. \end{aligned} \tag{24}$$

Conducting this analysis for each calibration approach in tandem with chosen CP methods exposes how feature importance varies under different model-tuning strategies, ultimately revealing the factors that cause a model to remain uncertain.

4. Experiment Settings

4.1. Research Problem and Questions

In this section, we introduce the primary research problems and questions that guide our investigation. We focus on a binary classification task in the PPM domain, characterized by sparse data and a significant class imbalance. Our research explores the interplay of interpretable and black-box classifiers, probability calibration methods, CP approaches, and explainability via feature attribution.

RQ1: How do different interpretable and black-box classifiers perform on a sparse, imbalanced binary classification problem, considering both thresholded and threshold-free metrics?

To address this question, we evaluate a variety of classifiers, ranging from transparent (e.g., Decision Trees) to black-box (e.g., XGBoost). We measure performance using Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-recall Curve (AUPRC), and Matthews Correlation Coefficient (MCC), capturing both threshold-dependent and threshold-free perspectives. A bootstrap resampling approach is applied for robust estimation, and statistical significance tests (Friedman and Nemenyi) are used to identify any meaningful performance differences.

RQ2: How do different probability calibration techniques compare against each other and against uncalibrated models in terms of calibration quality?

Accurate probability estimates are crucial, especially in imbalanced scenarios. We examine several calibration techniques (e.g., Isotonic Regression, Platt Scaling, Beta Calibration, Venn-Abers) and compare them to uncalibrated outputs. Our evaluation relies on standard metrics such as Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Logarithmic Loss (LogLoss). To verify whether differences among methods are statistically significant, we again employ statistical significance test.

RQ3: How does integrating calibrated probabilities affect the performance of Conformal Prediction methods?

A key contribution of this study is investigating how calibration alters CP outcomes. We apply SCP, measuring coverage, efficiency (e.g., Single Set Ratio, Minority Error Contribution), and other relevant metrics. By conducting statistical significance tests, we determine whether and how calibrated probability estimates improve CP-based UQ.

RQ4: How do different probability calibration techniques affect the accuracy and reliability of high-confidence single-label predictions within the Conformal Prediction framework?

To address this question, we evaluate chosen calibration methods specifically on the subset of predictions where the CP framework assigns a single label, indicating high confidence. We assess performance using different evaluation measures. Statistical significance tests are employed to determine whether differences among calibration methods are meaningful, thereby providing insights into how each calibration technique influences the reliability and accuracy of confident predictions in high-stakes clinical settings.

RQ5: How do different calibration methods influence feature attribution and interpretability for certain vs. uncertain predictions, as evaluated via SHAP?

After identifying instances where CP deems the model confident (single-label sets) or uncertain (multi-label sets), we use SHAP to analyze feature contributions. We then perform a grid comparison across various calibration and CP combinations, focusing on how calibration may shift or reshape feature attributions. This analysis elucidates whether certain calibration strategies consistently alter the importance of predictive features for high-confidence versus low-confidence predictions.

4.2. Evaluation of Classification Methods

Evaluating the proposed framework requires considering (1) *classification performance*, which assesses how effectively the model distinguishes between positive and negative instances, (2) *calibration quality*, which measures how closely its predicted probabilities match actual outcome frequencies, and (3) *UQ*, which ensures that CP sets meet desired coverage and efficiency levels. This section details the metrics used for these three perspectives.

Let $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be the held-out test set, where $y_i \in \{0, 1\}$. For each instance i, the model M produces a probability $\hat{p}_i = M(\mathbf{x}_i)$. A hard classification label $\hat{y}_i$ follows from applying a threshold τ:

$$\hat{y}_i \;=\; \begin{cases} 1, & \text{if } \hat{p}_i \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Define the confusion matrix counts:

$$\begin{aligned} \text{TP} &= \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = 1 \wedge y_i = 1), \quad \text{TN} = \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = 0 \wedge y_i = 0), \\ \text{FP} &= \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = 1 \wedge y_i = 0), \quad \text{FN} = \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = 0 \wedge y_i = 1). \end{aligned} \tag{26}$$

Based on these counts, we employ:

Area Under the ROC Curve (AUROC). The Receiver Operating Characteristic (ROC) curve plots

$$\text{TPR}(\tau) \;=\; \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FN}(\tau)}, \quad \text{FPR}(\tau) \;=\; \frac{\text{FP}(\tau)}{\text{FP}(\tau) + \text{TN}(\tau)} \tag{27}$$

for all thresholds $\tau \in [0, 1]$. The AUROC is the integral of TPR with respect to FPR. Values near 1.0 indicate strong discrimination, whereas 0.5 implies random guessing.

Area Under the Precision-Recall Curve (AUPRC). When the positive class is rare, the Precision-Recall (PR) curve is often more informative. It plots

$$\text{Prec}(\tau) \;=\; \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FP}(\tau)}, \quad \text{Rec}(\tau) \;=\; \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FN}(\tau)} \tag{28}$$

across thresholds. AUPRC integrates these values, with higher scores indicating better identification of the minority class in imbalanced scenarios.

Matthews Correlation Coefficient (MCC). The MCC accounts for all four confusion matrix elements in a single coefficient:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}. \tag{29}$$

It ranges from -1 to $+1$. Values near $+1$ correspond to perfect classification, 0 indicates random performance, and negative values imply an inverse relationship between predictions and true labels.

4.3. Evaluation of Calibration Approaches

We use the following metrics to quantify calibration quality:

Expected Calibration Error (ECE). Partition the interval $[0, 1]$ into K bins of equal width: $[b_0, b_1), \ldots, [b_{K-1}, b_K]$, where $b_0 = 0$ and $b_K = 1$. Let B_k be the set of indices whose predicted probability falls into bin k. Then define the mean predicted probability $\overline{p}_k$ and the empirical frequency of positive outcomes $\overline{y}_k$ in bin k as:

$$\overline{p}_k = \frac{1}{|B_k|} \sum_{i \in B_k} \hat{p}_i, \quad \overline{y}_k = \frac{1}{|B_k|} \sum_{i \in B_k} y_i. \tag{30}$$

The ECE is given by:

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \left| \overline{p}_k - \overline{y}_k \right|. \tag{31}$$

A smaller ECE indicates that predicted probabilities align more closely with observed outcomes across bins.

Maximum Calibration Error (MCE). While ECE is an average measure, MCE shows the largest single-bin deviation:

$$\mathrm{MCE} = \max_{1 \le k \le K} \left| \overline{p}_k - \overline{y}_k \right|. \tag{32}$$

A high MCE reveals at least one region where predictions are significantly over- or underconfident.

Logarithmic Loss (LogLoss). LogLoss (or cross-entropy) penalizes overconfident but incorrect predictions:

$$\mathrm{LogLoss} = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i \ln(\hat{p}_i) + (1 - y_i) \ln\big(1 - \hat{p}_i\big) \Big]. \tag{33}$$

Lower LogLoss values indicate that the model's predicted probabilities better match actual labels, accounting for both correctness and confidence.

These calibration metrics offer more than a statistical summary; each provides a distinct insight into the model's clinical utility and trustworthiness. A low ECE signifies that the model's probabilities are reliable on average, making them suitable for strategic decisions like resource planning. The MCE, in contrast, serves as a stress test for protocol safety by identifying the single risk bracket where the model is most untrustworthy, thus preventing systematic errors for specific patient subgroups. Finally, a low LogLoss is crucial for individual patient safety, as it penalizes a model severely for high-confidence mistakes, thereby discouraging the kind of "false reassurance" that can lead to delayed care. A comprehensive assessment of a model's practical value requires evaluating all three aspects.

4.4. Evaluation of Conformal Prediction

CP methods output a prediction set $\mathcal{C}(\mathbf{x}_i) \subseteq \{0,1\}$ for each instance $\mathbf{x}_i$. In binary classification, such sets may be $\{0\}$, $\{1\}$, or $\{0,1\}$, reflecting varying degrees of uncertainty. Two major goals of conformal predictors are:

1. Coverage: ensuring the true label is included with high probability.
2. Efficiency: keeping prediction sets as small as possible.

4.4.1. Coverage Metrics

Marginal Coverage. A valid conformal predictor with nominal coverage $1-\alpha$ should include the correct label for a fraction $\approx (1-\alpha)$ of test points. We define the empirically observed coverage as:

$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(y_i \in \mathcal{C}(\mathbf{x}_i)\big), \tag{34}$$

where N is the total number of predictions or data points, and $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition is true and 0 otherwise. High coverage indicates reliability but, by itself, does not guarantee balanced coverage across classes.

Average Coverage Gap. This metric quantifies the deviation of the observed empirical coverage from the desired nominal coverage of $1-\alpha$. It measures whether the CP method is under-covering or over-covering on average and is defined as:

$$\text{AverageCoverageGap} = (1-\alpha) - \text{Coverage}. \tag{35}$$

A low Average Coverage Gap indicates that the actual coverage closely aligns with the desired coverage, which is especially important in imbalanced settings where achieving the target coverage uniformly across classes can be challenging.

4.4.2. Efficiency Metrics

Single Prediction Set Ratio. Returning $\{0,1\}$ for every instance ensures near-perfect coverage but offers little practical utility. Let

$$\text{SingleSetRatio} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(|\mathcal{C}(\mathbf{x}_i)| = 1\big). \tag{36}$$

A higher value means the predictor more frequently produces confident, single-label sets.

Minority Error Contribution. Not all coverage misses are equally costly. In an imbalanced dataset, missing a positive case (e.g., failing to include label 1 in $\mathcal{C}(\mathbf{x}_i)$ when $y_i = 1$) may be more critical. Define:

$$\text{MinorityErrorContribution} = \frac{\sum_{i:y_i=1}\mathbb{I}\big(1 \notin \mathcal{C}(\mathbf{x}_i)\big)}{\sum_{i=1}^{N}\mathbb{I}\big(y_i \notin \mathcal{C}(\mathbf{x}_i)\big)}. \tag{37}$$

This ratio indicates how many of the total coverage misses occur on the minority class.

4.4.3. Effectiveness of Confident Predictions

When the CP method returns a single-label prediction set $\big|\mathcal{C}(\mathbf{x}_i)\big| = 1$, we say the prediction is confident. While Section 4.2 defines global metrics such as Precision, Recall, and MCC for the entire dataset, it is also insightful to evaluate these metrics exclusively on the subset of instances for which the prediction is single-labeled. This subset-specific view reveals how well the method actually performs when it chooses to be certain. In addition, we track Specificity and the Minority Class Ratio to understand the nature of confident decisions.

Restricting Metrics to Confident Subset. Let

$$\mathcal{D}_{\text{conf}} = \{ (\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{test}} \mid |\mathcal{C}(\mathbf{x}_i)| = 1 \} \tag{38}$$

be the confident subset of the test data, i.e., all test points for which the conformal set is a single label $\{0\}$ or $\{1\}$. For each instance in $\mathcal{D}_{\text{conf}}$, define the predicted label $\hat{y}_i \equiv \mathcal{C}(\mathbf{x}_i)$. We can then form a confusion matrix TP_{conf}, FP_{conf}, TN_{conf}, FN_{conf} restricted to $\mathcal{D}_{\text{conf}}$ and calculate:

$$\begin{aligned} \text{Precision}_{\text{conf}} &= \frac{\text{TP}_{\text{conf}}}{\text{TP}_{\text{conf}} + \text{FP}_{\text{conf}}}, \\ \text{Recall}_{\text{conf}} &= \frac{\text{TP}_{\text{conf}}}{\text{TP}_{\text{conf}} + \text{FN}_{\text{conf}}}, \\ \text{Specificity}_{\text{conf}} &= \frac{\text{TN}_{\text{conf}}}{\text{TN}_{\text{conf}} + \text{FP}_{\text{conf}}}, \end{aligned} \tag{39}$$

where the terms in the numerators and denominators denote true positives, false positives, etc., respectively, on $\mathcal{D}_{\text{conf}}$.

In an imbalanced dataset, it is also crucial to assess how frequently confident predictions are for the minority (positive) class. Define:

$$\text{Minority Class Ratio } [\%] \ = 100 \times \frac{\sum_{i=1}^{N} \mathbb{I}\Big(|\mathcal{C}(\mathbf{x}_i)| = 1 \ \wedge \ \mathcal{C}(\mathbf{x}_i) = \{1\} \Big)}{\sum_{i=1}^{N} \mathbb{I}\Big(|\mathcal{C}(\mathbf{x}_i)| = 1 \Big)}, \tag{40}$$

where $\{1\}$ designates the positive (minority) class, and $\mathbb{I}(\cdot)$ is the indicator function. A high value of Minority Class Ratio means that, among all single-label sets, a substantial fraction are $\{1\}$. Depending on the accompanying Precision and Recall values for these cases, this could indicate either strong confident detection of positives or an overestimation of risk leading to potential false positives.

4.5. Hyperparameter Optimization

Hyperparameters govern how models learn from the training data and can significantly impact both predictive accuracy and calibration. In this work, we used Bayesian Optimization with Gaussian Processes to select hyperparameters for each classifier (XGBoost, CatBoost, Decision Tree, Random Forest). Compared to grid or random search, Bayesian Optimization adaptively balances exploration of less-examined hyperparameter regions and exploitation of promising configurations by using an acquisition function like Expected Improvement (EI) [20]. For each model, we defined an objective function that measures performance via MCC (see Section 4.2 for definition). The dataset D was split into 10 folds using stratified sampling to preserve the original class ratio. Each candidate hyperparameter set was evaluated by 10-fold cross-validation, and the mean MCC across folds served as the performance score. Table 2 summarizes the search intervals for each model.

Table 2. Hyperparameter optimization settings for the ML models. Intervals denote the continuous ranges explored via Bayesian Optimization.

Model	Hyperparameters (Search Interval)
XGBoost	`max_depth`: (3, 20), `gamma`: (0, 1), `learning_rate`: (0.01, 0.6), `subsample`: (0.5, 1), `colsample_bytree`: (0.5, 1), `reg_alpha`: (0.1, 20), `reg_lambda`: (0.1, 20), `n_estimators`: (100, 1000)
CatBoost	`max_depth`: (3, 10), `learning_rate`: (0.0001, 0.4), `subsample`: (0.4, 1), `colsample_bylevel`: (0.4, 1), `l2_leaf_reg`: (0.1, 20), `n_estimators`: (100, 500)
Decision Tree	`max_depth`: (3, 15), `min_samples_split`: (2, 10), `min_samples_leaf`: (1, 10), `max_features`: (0.1, 1.0)
Random Forest	`n_estimators`: (100, 500), `max_depth`: (3, 20), `min_samples_split`: (2, 10), `min_samples_leaf`: (1, 10), `max_features`: (0.4, 1)

4.6. Data Structure for Evaluation

A robust evaluation requires not only splitting the data into training and test sets but also reserving a dedicated portion for calibration and UQ. In this study, we adopt a Bootstrapping approach to generate multiple training and out-of-bag (OOB) samples, ensuring that our performance estimates are less sensitive to a particular data partition. Specifically, we perform the following steps for each bootstrap iteration $b \in \{1, \ldots, B\}$, where $B = 200$:

1. *Bootstrap Sampling.* Let the preprocessed dataset be

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \quad N = 995, \tag{41}$$

where $y_i \in \{0, 1\}$ represents the binary target (e.g., ICU admission in the sepsis domain). For each iteration b, we draw N samples with replacement from D to create

$$D_{\text{train}}^{b} = \left\{ (x_j, y_j) \,\middle|\, j \sim \mathcal{U}\{1, N\} \right\}_{j=1}^{N}. \tag{42}$$

The points not selected in this bootstrap sample form the out-of-bag (OOB) set, denoted by $\text{OOB}^b \subseteq D$. Typically, $\left|\text{OOB}^b\right|$ ranges between 343 and 394 instances (approximately 35% to 40% of the data).

2. *Stratified Splitting of OOB.* We further split OOB^b into a calibration set D_{cal}^{b} and a test set D_{test}^{b}. The calibration set is a stratified sample of fixed size (171 instances), preserving the class ratio of OOB^b. Formally:

$$D_{\text{cal}}^{b} = \left\{(x_k, y_k) \in \text{OOB}^b : k \in \mathcal{S}_{\text{cal}}^{b}\right\}, \quad D_{\text{test}}^{b} = \text{OOB}^b \setminus D_{\text{cal}}^{b}. \tag{43}$$

With this partition, the minority class proportion in D_{cal}^{b} remains representative of OOB^b.

3. *Pipeline Execution.* For each iteration b:
   (a) *Model Training*: Train a new model M_b (e.g., XGBoost, CatBoost, Decision Tree, Random Forest) on D_{train}^{b}.
   (b) *Calibration and UQ*: Use D_{cal}^{b} to apply post-hoc calibration (Section 4.3) and to estimate uncertainty (Section 4.4).
   (c) *Model Evaluation*: Report metrics on D_{test}^{b}.

This OOB bootstrapping approach ensures an unbiased estimate of each model's performance and calibration quality. By repeatedly sampling and training on different subsets, we reduce the dependence on a single train-test split and gain a more robust understanding of how well models generalize to new data.

4.7. Statistical Significance Test

We assess the significance of our results—derived from multiple OOB bootstrap samples—using the Friedman test followed by the Nemenyi post-hoc test for pairwise comparisons.

The Friedman test, a non-parametric repeated-measures method [21], ranks each algorithm's performance across bootstrap samples and compares their average ranks. Its statistic is computed as

$$\chi_F^2 = \frac{12n}{k(k+1)} \left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right], \tag{44}$$

where k is the number of algorithms, n is the number of bootstrap samples, and R_j is the sum of the ranks for the jth algorithm. The null hypothesis (H_0) asserts no significant differences among methods; if $p < 0.05$, H_0 is rejected and we proceed with the Nemenyi test.

The Nemenyi test computes a critical difference (CD) as

$$CD = q \sqrt{\frac{k(k+1)}{6n}}, \tag{45}$$

which indicates whether the difference in average ranks between any pair of methods is statistically significant. This Friedman–Nemenyi framework is robust against non-normal data and effectively compares multiple models while mitigating Type-I error risks.

5. Results

5.1. Classifier Performance Comparison (RQ1)

Table 3 summarizes the performance of four supervised learning models evaluated on the task of predicting sepsis readmissions. Each metric (AUROC, AUPRC, and MCC) is averaged over 200 bootstrap replications, providing a robust estimate of the models' generalization capabilities. CatBoost achieves the highest mean scores across all three metrics: 0.8506 ± 0.05 in AUROC, 0.6807 ± 0.09 in AUPRC, and 0.6028 ± 0.09 in MCC. This suggests that CatBoost consistently balances the identification of true positives and true negatives while maintaining strong discrimination between classes and handling the class imbalance inherent to sepsis readmissions.

Table 3. Evaluation of Supervised Models Across 200 Bootstrap Samples.

Model	AUROC	AUPRC	MCC
XGBoost	0.8356 (±0.05)	0.6567 (±0.09)	0.5839 (±0.09)
CatBoost	**0.8506 (±0.05)**	**0.6807 (±0.09)**	**0.6028 (±0.09)**
Decision Tree	0.7437 (±0.07)	0.6340 (±0.10)	0.5816 (±0.11)
Random Forest	0.8306 (±0.06)	0.6652 (±0.09)	0.5828 (±0.10)

AUROC measures the overall separability of the positive and negative classes by varying the decision threshold. CatBoost's AUROC surpasses those of XGBoost, Random Forest, and Decision Tree, indicating that it produces a better rank ordering of patients likely to be readmitted. The Friedman test (Figure 3) confirms that the differences among the four models are statistically significant ($\chi^2_{\text{Friedman}} = 314.6368$, $p = 6.7562 \times 10^{-68}$). Post-hoc Nemenyi tests reveal that CatBoost holds a significant edge over all other methods, highlighting its superior capability to discriminate between sepsis patients who will and will not be admitted to ICU.

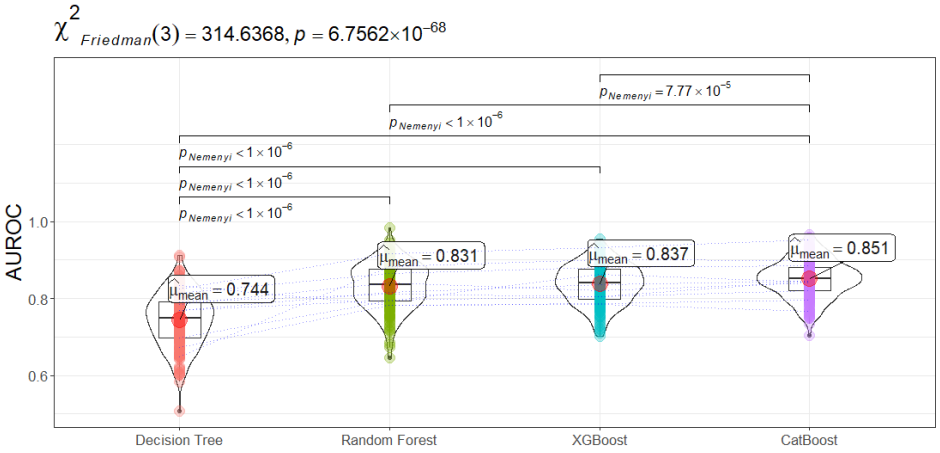


Figure 3. Friedman-Nemenyi significance test results regarding AUROC of machine learning models across 200 bootstrapping iterations.

AUPRC focuses specifically on the model's performance for positive cases, making it a critical metric for domains with class imbalance. CatBoost again achieves the highest mean AUPRC, 0.6807 ± 0.09, which implies that it captures more of the actual admissions (true positives) at lower false-positive rates compared to competing methods. The Friedman statistic for AUPRC (Figure 4) ($\chi^2_{\text{Friedman}} = 91.41$, $p = 1.0906 \times 10^{-19}$) indicates large discrepancies between classifiers, and subsequent Nemenyi tests confirm that CatBoost's advantage is statistically significant. This result is clinically important: in sepsis treatment settings, minimizing missed positive cases is a priority because each undetected admission risk can lead to delayed interventions and adverse outcomes.

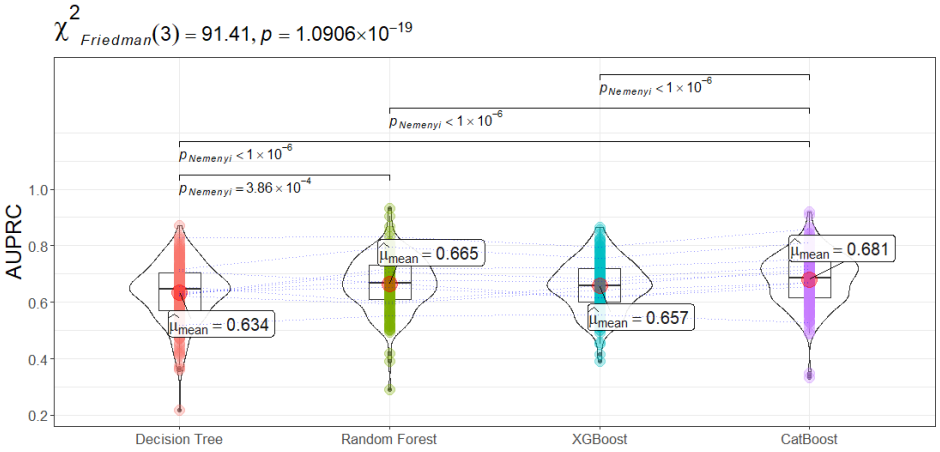


Figure 4. Friedman-Nemenyi significance test results regarding AUPRC of machine learning models across 200 bootstrapping iterations.

The MCC consolidates true positives, false positives, true negatives, and false negatives into a single coefficient, offering a more balanced measure when class prevalence is skewed. Once again, CatBoost's average MCC of 0.6028 ± 0.09 exceeds those of the other algorithms, although the differences among models are more modest than for AUROC or AUPRC. The Friedman test (Figure 5) ($\chi^2_{\text{Friedman}} = 18.375$, $p = 3.6807 \times 10^{-4}$) and subsequent Nemenyi tests verify that CatBoost significantly outperforms Random Forest, XGBoost, and Decision Tree with respect to MCC. From a clinical standpoint, a higher MCC indicates more reliable predictions that avoid both over-diagnosis (unnecessary interventions) and under-diagnosis (missed sepsis relapses).

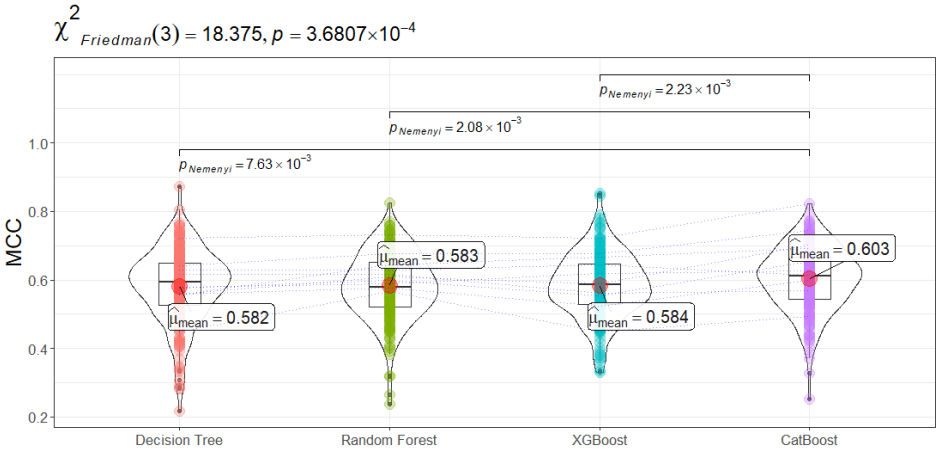


Figure 5. Friedman-Nemenyi significance test results regarding MCC of machine learning models across 200 bootstrapping iterations.

Three primary observations emerge from these results. First, boosting-based methods outperform single-tree and classical ensemble approaches in distinguishing high-risk sepsis patients. By iteratively refining weak learners and focusing on hard-to-classify instances, CatBoost and XGBoost capture non-linear interactions and subtle patterns in patient trajectories that a single Decision Tree often overlooks. Second, the frequent usage of categorical features likely contributes to CatBoost's competitive edge, as it includes specialized strategies for dealing with categorical inputs and typically requires less extensive feature engineering. Third, while Random Forest does offer improvements over a single Decision Tree by aggregating multiple trees, it still lags behind gradient boosting, as indicated by both the average performance metrics and the post-hoc significance tests. Notably, Decision Tree yields the lowest overall scores, a result likely due to insufficient model complexity in capturing the intricacies of sepsis readmission pathways. Still, its interpretability may appeal to clinicians wanting an easily understandable framework, although, in high-stakes applications, maximizing predictive performance often remains a priority. Random Forest strikes a balance by maintaining some level of interpretability via feature importance analysis, but its predictive power is outmatched by boosting methods. In contrast, CatBoost manages to achieve superior accuracy and interpretability trade-offs, particularly because game-theoretic approaches (e.g., SHAP values) can provide post-hoc explanations for boosting predictions.

The strong performance of CatBoost implies that advanced boosting algorithms can significantly enhance the early detection of potential readmissions, giving healthcare professionals additional lead time to intervene. By maximizing both AUPRC (for the minority class) and MCC (for balanced predictive quality), CatBoost reduces the risk of missing critical cases or triggering unwarranted alerts. These findings underline the value of ensemble-based approaches in healthcare analytics, especially in scenarios where patient-level events have complex temporal and categorical interdependencies. In summary, CatBoost exhibits statistically significant benefits in separating sepsis admission outcomes over XGBoost, Random Forest, and Decision Tree. Its higher AUPRC and MCC highlight a promising capacity for correctly identifying high-risk patients without an excessive false-alarm rate. As a result, CatBoost emerges as the leading candidate for the subsequent stages of calibration, UQ, and explainability within the proposed predictive monitoring framework.

5.2. Evaluation of Calibration Approaches (RQ2)

Following the classifier comparison, we now explore how various calibration methods alter the CatBoost model's probability outputs. In Table 4, we report classification

performance (AUROC, AUPRC) alongside calibration-specific measures (LogLoss, ECE, MCE) averaged over 200 bootstrap iterations. Figures 6–10 summarize the corresponding Friedman-Nemenyi significance test results for each metric.

Table 4. Average Performance of CatBoost Model Across 200 Bootstrap Samples with Different Calibration Methods Applied.

Calibration Method	AUROC	AUPRC	Log Loss	ECE	MCE
Beta	**0.8506 (±0.05)**	0.6807 (±0.09)	0.1995 (±0.04)	0.0327 (±0.01)	0.6248 (±0.18)
Isotonic Regression	0.8345 (±0.05)	**0.6825 (±0.09)**	0.4194 (±0.27)	**0.0318 (±0.01)**	**0.4663 (±0.24)**
Platt Scaling	**0.8506 (±0.05)**	0.6807 (±0.09)	0.1981 (±0.03)	0.0333 (±0.01)	0.6260 (±0.19)
Spline	0.8355 (±0.06)	0.6731 (±0.09)	0.2001 (±0.03)	0.0384 (±0.02)	0.6008 (±0.17)
Uncalibrated	**0.8506 (±0.05)**	0.6807 (±0.09)	0.2062 (±0.05)	0.0413 (±0.01)	0.6108 (±0.16)
Venn-Abers	0.8474 (±0.05)	0.6806 (±0.09)	0.2068 (±0.03)	0.0490 (±0.01)	0.4883 (±0.16)

As expected, none of the post-hoc calibration techniques significantly boosts AUROC or AUPRC over the uncalibrated CatBoost. Platt Scaling and Beta match the uncalibrated baseline in AUROC (both at $\sim$0.8506 $\pm$ 0.05), while Spline and Isotonic Regression appear slightly lower ($\sim$0.835). The Friedman test for AUROC ($\chi^2_{\text{Friedman}} = 153.9573$, $p = 1.9185 \times 10^{-31}$) affirms that at least one method differs notably; however, the Nemenyi post-hoc analysis pinpoints Isotonic Regression and Spline Calibration as significantly less effective than Beta, Platt Scaling, and the uncalibrated model.

A similar observation arises with AUPRC ($\chi^2_{\text{Friedman}} = 32.9493$, $p = 3.8516 \times 10^{-6}$): Isotonic Regression obtains the top mean AUPRC (0.6825 $\pm$ 0.09), but Beta, Platt Scaling, and uncalibrated CatBoost remain very close behind ($\sim$0.6807 $\pm$ 0.09). The fact that these methods do not provide consistent improvement in rank-based metrics (AUROC) or minority-class performance (AUPRC) aligns with existing literature: calibration primarily targets how well probabilities match actual outcome frequencies, rather than enhancing the underlying discrimination. In some instances (e.g., Spline or Isotonic Regression vs. uncalibrated CatBoost), a minor dip in AUROC/AUPRC is the price of improved calibration.

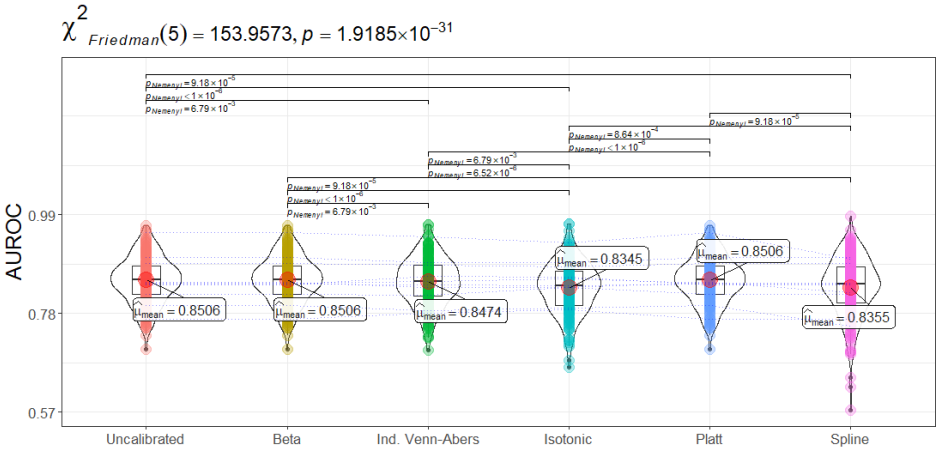


Figure 6. Friedman-Nemenyi significance test results regarding AUROC of calibration methods on the CatBoost model across 200 bootstrapping iterations.

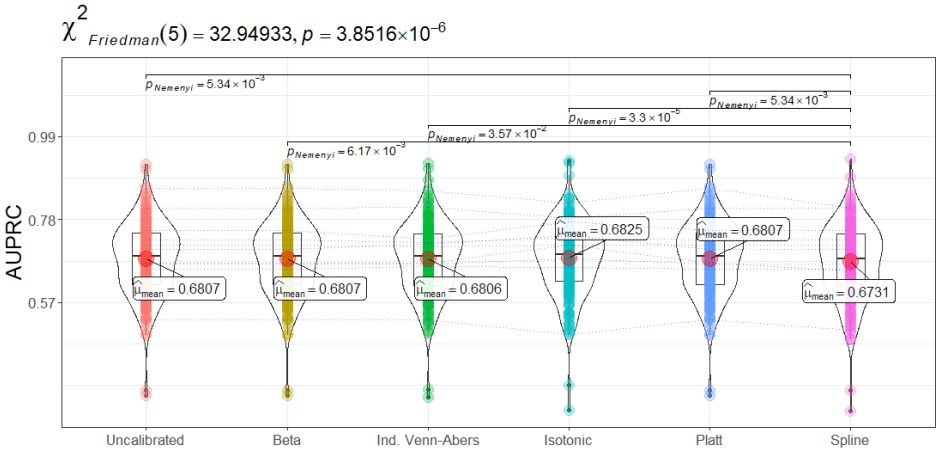


Figure 7. Friedman-Nemenyi significance test results regarding AUPRC of calibration methods on the CatBoost model across 200 bootstrapping iterations.

LogLoss gauges the magnitude of misalignment between predicted probabilities and actual outcomes—severely penalizing instances assigned incorrect high-confidence scores. Notably, Platt Scaling yields the lowest mean LogLoss (0.1981 ± 0.03), followed closely by Beta (0.1995 ± 0.04) and Spline (0.2001 ± 0.03). By contrast, Isotonic Regression exhibits a substantially higher mean LogLoss (0.4194 ± 0.27). The Friedman test ($\chi^2_{\text{Friedman}} = 178.8486$, $p = 9.4262 \times 10^{-37}$) indicates these differences are highly significant, and the subsequent Nemenyi comparisons identify Isotonic as significantly worse than nearly all other calibration methods in this regard. From a clinical perspective, lower LogLoss translates into more reliable estimation of readmission risk across the entire probability spectrum. For example, an overly confident model might assign probabilities close to 1.0 for patients who ultimately do not get readmitted, incurring large penalization. If healthcare decisions hinge on probability thresholds to, say, intensify monitoring or allocate ICU beds, a method with a low LogLoss is valuable because it avoids severe misclassifications. Hence, Platt or Beta might be more attractive if one seeks a stable, precise reflection of readmission likelihood without excessively skewing the probability scale.

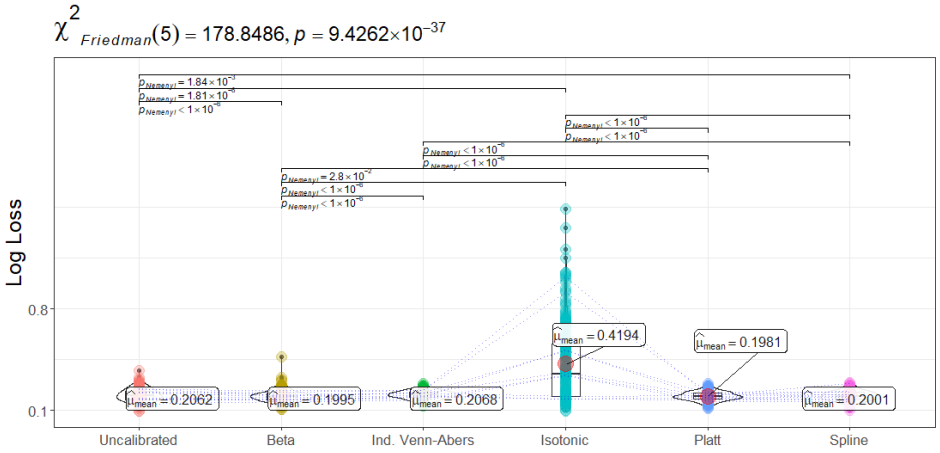


Figure 8. Friedman-Nemenyi significance test results regarding log loss of calibration methods on the CatBoost model across 200 bootstrapping iterations.

While LogLoss focuses on penalizing incorrect high-confidence assignments, ECE and MCE capture how close predicted probabilities are to empirical event frequencies. This distinction is crucial in domains like sepsis management, where calibrating risk estimates for threshold-based interventions can directly affect patient outcomes. ECE measures the

average gap between predicted probability and the true proportion of positives. A low ECE means that if the model predicts, for instance, a 40% chance of readmission, then roughly 40% of those patients do indeed return. MCE captures the largest such deviation across all probability bins, highlighting worst-case miscalibrations that could be critical in a high-stakes clinical workflow.

Results show that Isotonic Regression achieves the lowest ECE (0.0318 ± 0.01) and lowest MCE (0.4663 ± 0.24). Beta and Platt Scaling do yield moderate improvements compared to the uncalibrated model, but they cannot match Isotonic Regression in terms of minimizing average and worst-case calibration error. The Friedman tests for ECE ($\chi^2_{\text{Friedman}} \approx 236$) and MCE ($\chi^2_{\text{Friedman}} \approx 148$) both yield p-values far below 0.01, indicating statistically significant differences among methods. Pairwise Nemenyi tests highlight that Isotonic Regression's ECE and MCE are significantly lower than those of Venn-Abers, the uncalibrated baseline, and sometimes Spline.

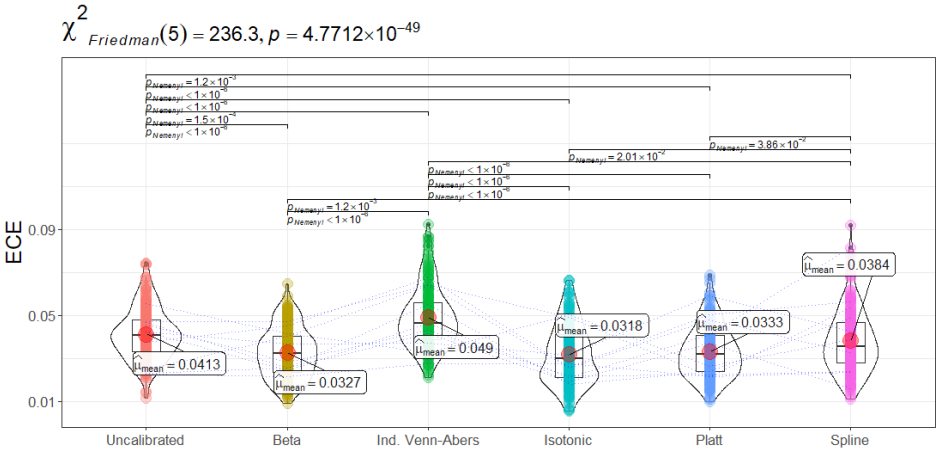


Figure 9. Friedman-Nemenyi significance test results regarding ECE of calibration methods on the CatBoost model across 200 bootstrapping iterations.

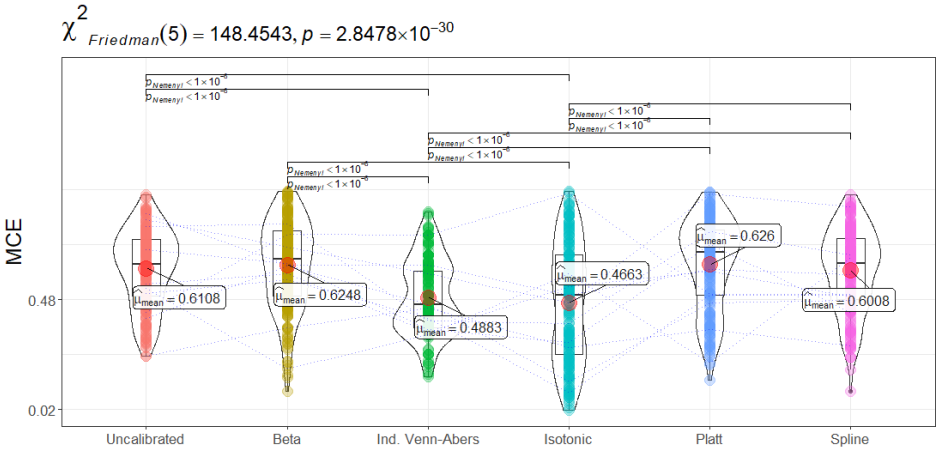


Figure 10. Friedman-Nemenyi significance test results regarding MCE of calibration methods on the CatBoost model across 200 bootstrapping iterations.

Clinically, lower ECE and MCE signify more consistent alignment between the numeric score and real-world ICU admission risk. For instance, an Isotonic Regression-calibrated model that assigns an 80% admission probability might be especially trustworthy for mobilizing time-sensitive interventions. On the flip side, the relatively high LogLoss of Isotonic Regression suggests that such piecewise-constant adjustments can become extreme

in certain probability bins, creating large penalty spikes. This means the model is, on average, well-calibrated but can mispredict certain individual cases rather sharply.

In sum, each calibration method presents a distinct trade-off. Isotonic Regression dominates in calibration error metrics (ECE, MCE), ensuring a tight alignment of predicted versus observed risk. However, it incurs higher LogLoss and occasionally lowers AUROC or AUPRC. Platt Scaling and Beta Calibration preserve near-top discrimination (AUROC, AUPRC) while also achieving consistently low Log Loss, indicating more balanced updates to probability scores. Spline calibration is moderately effective across all metrics, offering flexible piecewise corrections but does not stand out as best in any one category. Venn-Abers provides interval predictions with theoretical validity guarantees—a unique advantage in safety-critical settings—but has higher calibration errors. From a healthcare standpoint, the choice of calibration hinges on balancing the need for accurate high-risk identification (AUROC, AUPRC) with the demand for trustworthy probability statements (LogLoss, ECE, MCE). Figure 11's reliability plots illustrate these findings. The uncalibrated model slightly overestimates risk in mid- to high-probability bins. Platt Scaling, Beta, and Spline each temper this overconfidence, more closely aligning with the diagonal "perfect calibration" line. Although Isotonic stands out for particularly low calibration errors overall, its stepwise adjustments sometimes cause spikier changes in probability assignment, visible as flatter regions on the reliability curve. For contexts where stable, smoothly varying scores are preferred, such discontinuities might be less desirable despite the strong ECE performance.

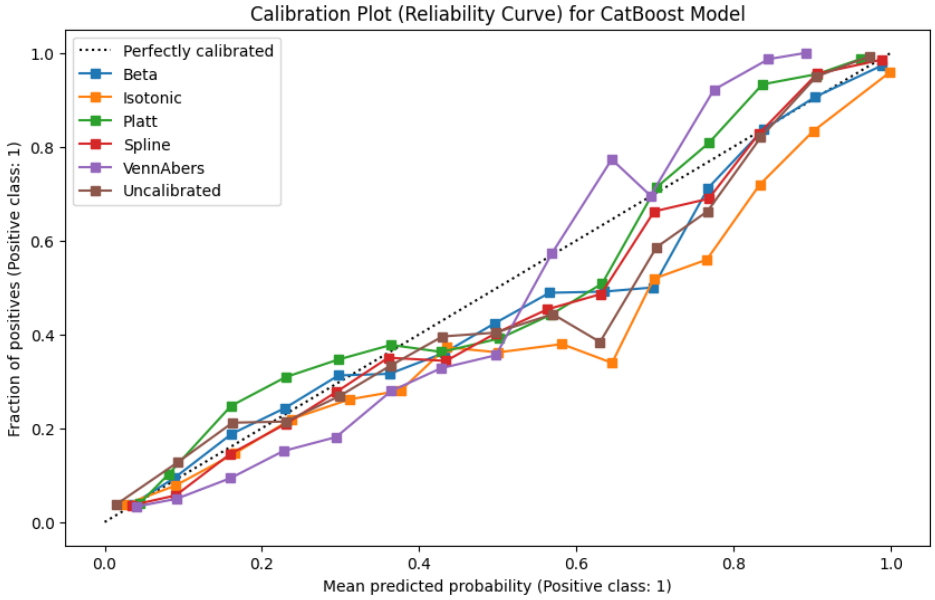


Figure 11. Calibration plots of post-hoc calibration methods across 200 bootstrap iterations.

Overall, there is no single "best" calibration method for every clinical situation. If strict accuracy of the probability estimate is paramount—so that a predicted 70% readmission risk closely matches actual outcomes—then Isotonic Regression or Spline might be worth of any drop in LogLoss or AUPRC. Conversely, if the clinical workflow demands a stable probability distribution with minimal penalization for misclassifications, Platt Scaling or Beta may be most appealing. In our sepsis context, where critical resources (ICU beds, antibiotics) hinge on balancing over- and under-treatment risks, Platt Scaling and Beta Calibration emerge as especially strong candidates. They keep classification metrics intact while substantially correcting probability estimates, supporting more nuanced risk-based decisions and potentially leading to better patient outcomes.

5.3. Evaluation of Conformal Prediction Methods

This section examines SCP applied to the CatBoost classifier, along with various calibration strategies. Table 5 and Figures 12–15 summarize the main findings for marginal coverage, average coverage gap, minority error contribution, and the single prediction set ratio. Together, these results reveal how often the conformal procedure yields valid and precise predictions, and how misclassification risk is distributed across the minority (admission to ICU) class.

Table 5. Average Conformal Prediction Results for Split Conformal with CatBoost Model Across 200 Bootstrap Samples.

Calibration Method	Marginal Coverage	Average Coverage Gap	Minority Error Contribution (%)	Single Prediction Set Ratio (%)
Beta Calibration	0.9599 (±0.02)	19.0294 (±7.99)	93.4913 (±12.11)	90.6204 (±10.02)
Isotonic Regression	0.9578 (±0.02)	19.4813 (±8.27)	**90.8993 (±14.13)**	**91.1105 (±13.55)**
Platt Scaling	0.9608 (±0.02)	19.1753 (±8.11)	95.3790 (±10.44)	89.2572 (±10.87)
Spline Calibration	0.9599 (±0.02)	19.3481 (±7.73)	94.3152 (±11.21)	90.8709 (±9.62)
Uncalibrated CatBoost	**0.9615 (±0.02)**	**18.6658 (±7.61)**	95.5935 (±10.37)	89.7549 (±10.80)
Venn-Abers	0.9578 (±0.02)	19.6807 (±8.29)	92.0942 (±14.88)	89.8766 (±12.73)

Figure 12 shows that marginal coverage for the CP converges near or slightly above 0.95, indicating that the true label is included in the model's prediction set roughly 95–96% of the time. Uncalibrated CatBoost achieves the highest average coverage (0.9615 ± 0.02), narrowly surpassing Platt Scaling and Beta/Spline calibrations (around 0.9599–0.9608). Isotonic Regression and Venn-Abers trail slightly at 0.9578. Statistical tests ($\chi^2_{\text{Friedman}} = 80.5488$, $p = 6.4421 \times 10^{-16}$) confirm significant differences among the calibration methods. Post-hoc comparisons suggest that although Uncalibrated CatBoost has a minor coverage edge, several pairwise differences (e.g., uncalibrated vs. Beta or Spline) are subtle. Clinically, these marginal coverage levels imply that across repeated samples, the prediction sets produced by CP include the correct label at or beyond the intended 95% frequency—a reassuring result for risk-sensitive healthcare environments.

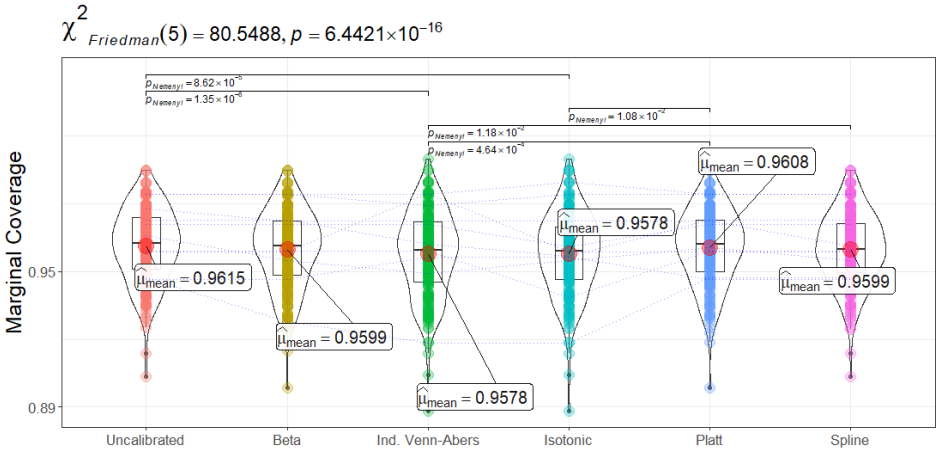


Figure 12. Friedman-Nemenyi significance test results regarding Marginal Coverage of Split conformal prediction and calibration methods on the CatBoost model across 200 bootstrapping iterations.

Average coverage gap (Figure 13) measures how much coverage can fluctuate relative to the nominal 95% target across different test instances. Uncalibrated CatBoost achieves the smallest gap (18.66 ± 7.61), with Beta, Platt Scaling, and Spline following closely (all around 19). Isotonic Regression and Venn-Abers lie marginally higher (approximately 19.48 and 19.68, respectively). The Friedman test ($\chi^2_{\text{Friedman}} = 19.8589$, $p = 1.3283 \times 10^{-3}$) again indicates statistically significant variation. For the examined use case, a lower coverage

gap means the conformal intervals (or sets) remain more consistently valid across different patients. Large gaps can signal that certain subpopulations—for example, older patients or those with atypical symptoms—might be over- or under-covered. Although Uncalibrated CatBoost exhibits the tightest overall coverage gap, the differences here are modest (roughly 1% across methods). Hospitals with large, diverse patient populations might still opt for a calibration method if it yields other benefits (e.g., improved minority error rates) without inflating coverage gap too severely.

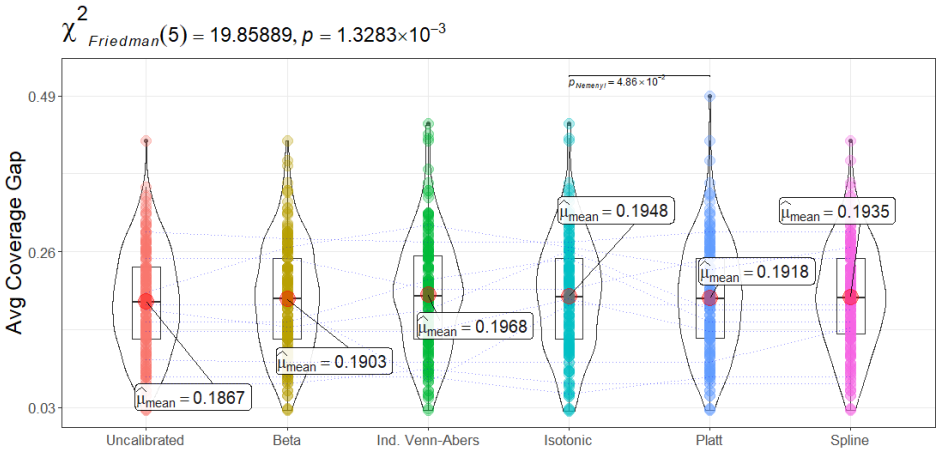


Figure 13. Friedman-Nemenyi significance test results regarding Average Coverage Gap of Split conformal prediction and calibration methods on the CatBoost model across 200 bootstrapping iterations.

Figure 14 presents the minority error contribution: the proportion of total misclassifications arising from the positive (ICU admission) class. Ideally, a lower fraction indicates that prediction errors are more evenly distributed, reducing the likelihood of disproportionately missing readmissions. Isotonic Regression yields the lowest minority error ($90.90 \pm 14.13\%$), suggesting that when mistakes occur, fewer are concentrated among readmitted patients. Beta and Spline calibrations approach mid-range values near 93–94%. Uncalibrated, Platt Scaling, and Spline calibrations all exceed 94–95%, indicating a larger share of errors come from those crucial positive cases. These distinctions are statistically significant ($\chi^2_{\text{Friedman}} = 71.88945$, $p = 4.142 \times 10^{-14}$). In a high-stakes domain like sepsis, lower minority error implies the model better safeguards the subgroup in urgent need of accurate predictions. By contrast, a higher minority error portion could lead to under-detection of relapsing patients, potentially causing delays in treatment.

Another key metric, shown in Figure 15, is the single prediction set ratio—the percentage of instances for which the conformal predictor returns exactly one label (fully confident) rather than an ambiguous set. Here, Isotonic Regression stands out at $91.11 \pm 13.55\%$, while Beta and Spline hover near 90.6–90.9%. Platt Scaling and Uncalibrated are slightly lower, around 89%. Statistical tests ($\chi^2_{\text{Friedman}} = 97.00886$, $p = 2.2552 \times 10^{-19}$) reveal that Isotonic Regression's single-set predictions are significantly more frequent than those under uncalibrated or Platt Scaling-based conformal methods. In practice, a higher single prediction set ratio can be interpreted as fewer "uncertain" predictions requiring secondary review. For hospital workflows, this translates into fewer flagged patients needing additional confirmatory steps—saving time and resources, albeit with the caveat that such higher confidence can sometimes be miscalibrated.

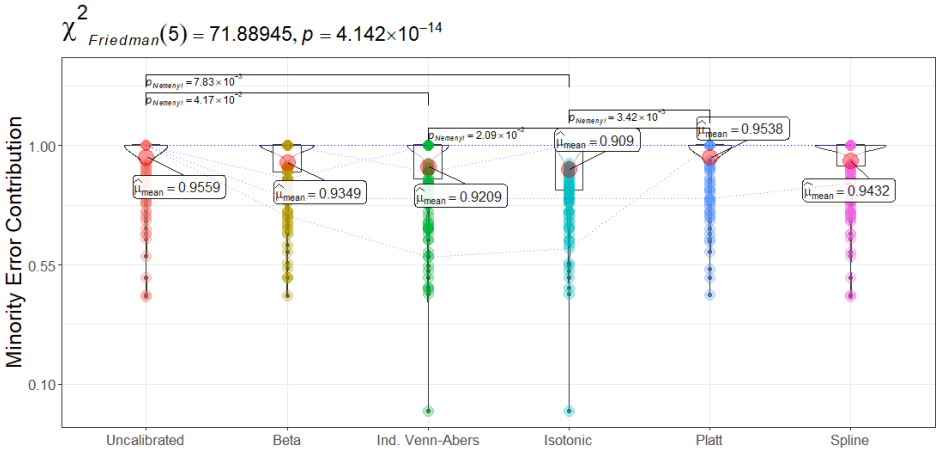


Figure 14. Friedman-Nemenyi significance test results regarding Minority Error Contribution of Split conformal prediction and calibration methods on the CatBoost model across 200 bootstrapping iterations.

SCP aims to guarantee valid coverage at a pre-defined confidence level (95% in our use-case), and these results confirm that nearly all methods fulfill this target. Nonetheless, certain calibration strategies offer subtle but important trade-offs. Uncalibrated CatBoost slightly outperforms in marginal coverage and average coverage gap, but it exhibits one of the highest minority error contributions. In other words, missed admissions to the ICU account for more of its overall misclassifications. Isotonic Regression achieves the lowest minority error and the highest frequency of confident (single-label) decisions, at the cost of slightly lower coverage and a modestly larger coverage gap. Beta and Spline calibrations produce balanced, middle-ground behaviors, combining decent coverage with reasonable minority error distribution. Platt Scaling align coverage near uncalibrated levels but also share the downside of an elevated minority error fraction.

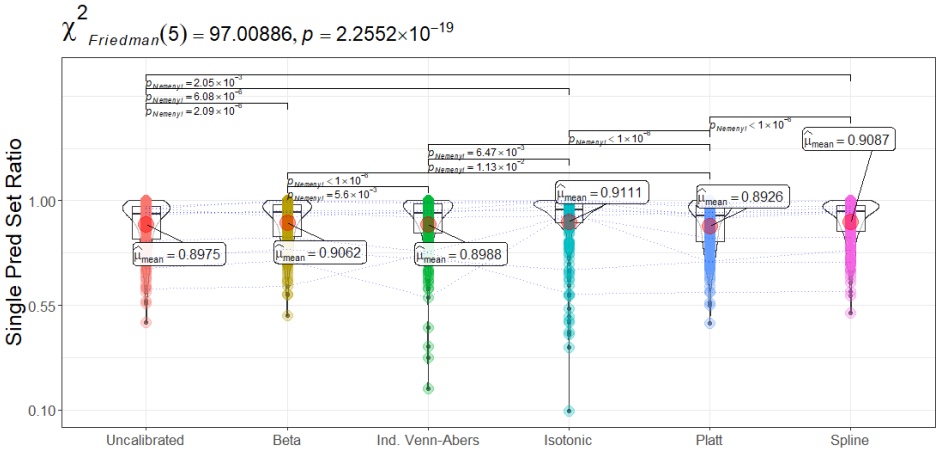


Figure 15. Friedman-Nemenyi significance test results regarding Single Prediction Set Ratio of Split conformal prediction and calibration methods on the CatBoost model across 200 bootstrapping iterations.

For sepsis admissions to the ICU, reducing errors on the minority class could be paramount: even a small reduction in missed admissions to the ICU can translate into significantly improved patient outcomes. Hence, a method like Isotonic Regression, which shifts the error burden away from relapsing patients, may be particularly appealing. On the other hand, uncalibrated or Platt Scaling-based CP—though high in overall coverage—might inadvertently let too many high-risk patients slip by undetected. Ultimately,

the choice depends on whether the clinical emphasis is on maximizing certainty in predictions (thus fewer ambiguous cases) versus avoiding minority-class oversights.

5.4. Inspection of Certain Predictions

When the CP framework produces a single-label outcome, it designates a high-confidence prediction for a particular instance. Although most calibration strategies yield broadly similar proportions of such one-label sets, there are notable differences in how accurately each method classifies the positive and negative classes within this "certain" subset. The Table 6 reveals consistent performance patterns: Isotonic Regression and Spline typically offer elevated recall or precision in these confident instances, while Platt Scaling provides unusually strong specificity. Beta often strikes a middle-ground, balancing both recall and precision without dominating in any single metric.

Table 6. Evaluation of Single Prediction Set for Split Conformal Across 200 Bootstrap Samples for CatBoost.

Calibration Method	Recall	Precision	Specificity	MCC	Minority Class Ratio (%)
Beta Calibration	0.5320 (±0.16)	0.9478 (±0.13)	0.9974 (±0.01)	0.6859 (±0.14)	4.9266 (±1.80)
Isotonic Regression	**0.5525 (±0.15)**	0.9380 (±0.10)	0.9959 (±0.01)	0.6943 (±0.12)	**5.5531 (±2.40)**
Platt Scaling	0.3594 (±0.25)	0.6922 (±0.44)	**0.9985 (±0.00)**	0.4814 (±0.31)	3.2686 (±2.46)
Spline Calibration	0.5407 (±0.14)	**0.9671 (±0.07)**	0.9981 (±0.00)	**0.7002 (±0.11)**	4.9806 (±1.61)
Uncalibrated CatBoost	0.4931 (±0.14)	0.9657 (±0.10)	0.9984 (±0.00)	0.6672 (±0.12)	4.0378 (±1.48)
Venn-Abers	0.4042 (±0.20)	0.8247 (±0.33)	0.9967 (±0.01)	0.5523 (±0.24)	3.6240 (±2.28)

Figures 16–20 corroborate these observations and offer a deeper comparative lens. Figure 16 (Recall for certain data) shows Isotonic Regression achieving a mean of 0.5525, significantly higher than Platt Scaling at 0.3594 ($\chi^2_{\text{Friedman}} = 207.1656$, $p = 8.3203 \times 10^{-43}$). This indicates that when the Isotonic-based SCP method is sufficiently confident to assign a single label, it more reliably flags actual readmissions than Platt Scaling. By contrast, Spline emerges with the highest precision (0.9671 in Figure 17), suggesting fewer false positives among confidently predicted patients, an attribute that may reduce undue resource allocation to non-relapsing cases. The Friedman test further confirms that Platt Scaling, Beta, and Uncalibrated CatBoost differ from Isotonic Regression or Spline in precision ranks ($\chi^2_{\text{Friedman}} = 88.94143$, $p = 1.1211 \times 10^{-17}$).

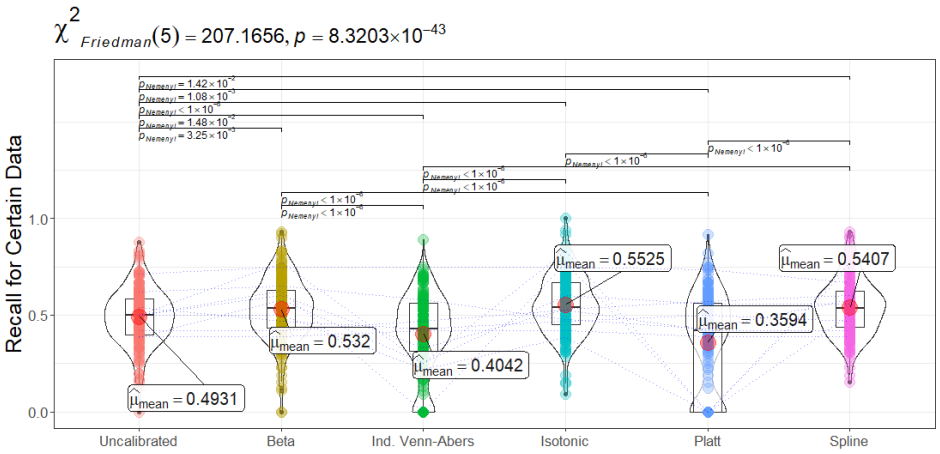


Figure 16. Friedman-Nemenyi significance test results regarding Recall for certain data instances of Split conformal prediction and calibration methods on the CatBoost model across 200 bootstrapping iterations.

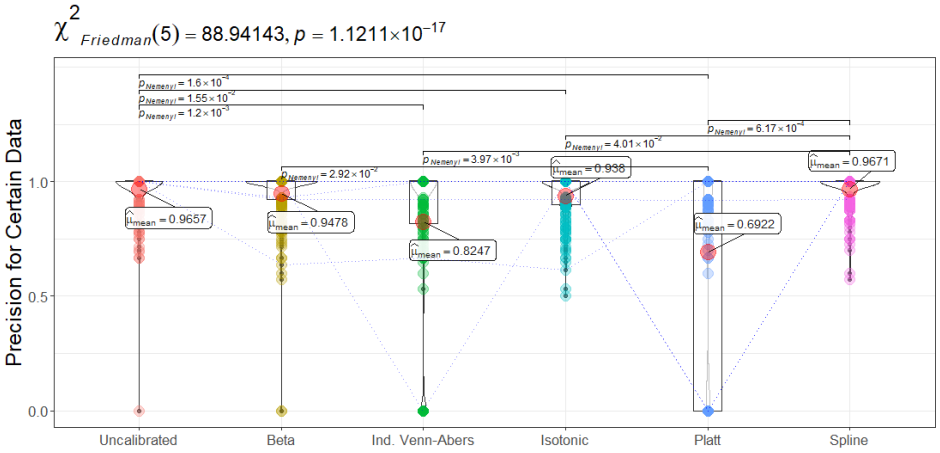


Figure 17. Friedman-Nemenyi significance test results regarding Precision for certain data instances of Split conformal prediction and calibration methods on the CatBoost model across 200 bootstrapping iterations.

Similar distinctions manifest in specificity and MCC (Figures 18 and 19). Platt Scaling delivers near-perfect specificity (0.9985) for certain predictions, whereas Beta and Uncalibrated CatBoost range closer to 0.9974–0.9984, with Spline also above 0.9980. Higher specificity reduces false alarms but can curb recall by excluding borderline positive patients. The MCC values reveal a parallel story: Spline peaks at 0.7002, surpassing both Beta and Isotonic Regression (0.6859 and 0.6943, respectively), attesting to its consistent balance of true positives and true negatives within single-label decisions. The Friedman statistic ($\chi^2_{\text{Friedman}} = 167.216, p = 2.8643 \times 10^{-34}$) and subsequent pairwise comparisons confirm Spline's significant advantage over Platt Scaling's more conservative approach (0.4814 MCC).

Clinical Relevance and Technical Considerations. From a clinical perspective, these findings highlight critical trade-offs in patient classification strategies. The recall and precision distributions across calibration methods directly influence patient triage decisions, particularly in high-stakes medical scenarios such as early sepsis detection, post-operative monitoring, and hospital readmission risk assessment. The model's ability to confidently predict a single-label classification directly impacts clinical workflows and intervention timing. Isotonic Regression's superior recall implies that more high-risk patients would be flagged with a certain positive classification, ensuring that at-risk individuals receive necessary monitoring or intervention. This attribute is particularly crucial in conditions where early warning signs are subtle yet predictive, such as sepsis onset or cardiac decompensation. However, high recall at the expense of specificity may increase unnecessary hospital admissions or treatments, leading to higher resource utilization and potential patient burden from false positives. Conversely, Platt Scaling and Spline, with their higher specificity and precision, minimize unnecessary interventions, favoring a more conservative approach. This calibration choice is preferable in cases where false positives carry substantial costs, such as invasive procedures or intensive resource allocation (e.g., ICU admission). For instance, a high specificity system ensures that only those with a true likelihood of relapse are assigned aggressive therapeutic strategies, reducing the risk of overtreatment. The MCC further contextualizes these trade-offs by offering a more holistic measure of classifier quality, incorporating all four confusion matrix components (true positives, false positives, true negatives, and false negatives). The MCC scores indicate that while Isotonic Regression and Beta provide strong recall-driven certainty, Spline calibration optimizes overall balance. A high MCC ensures that the classifier is not disproportionately favoring one class over the other, which is critical in clinical settings where both false

positives and false negatives can be costly. For example, underdiagnosing a condition like post-surgical infection (false negative) can result in complications, whereas overdiagnosing it (false positive) leads to unnecessary antibiotic usage and increased resistance concerns.

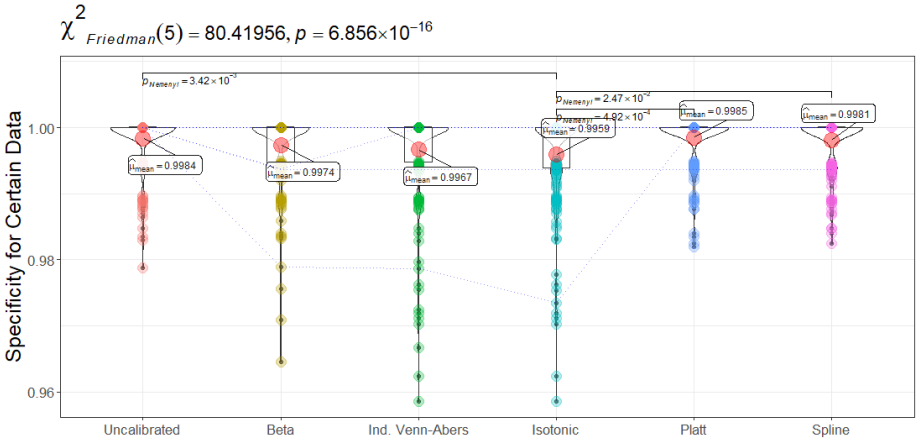


Figure 18. Friedman-Nemenyi significance test results regarding Specificity for certain data instances of Split conformal prediction and calibration methods on the CatBoost model across 200 bootstrapping iterations.

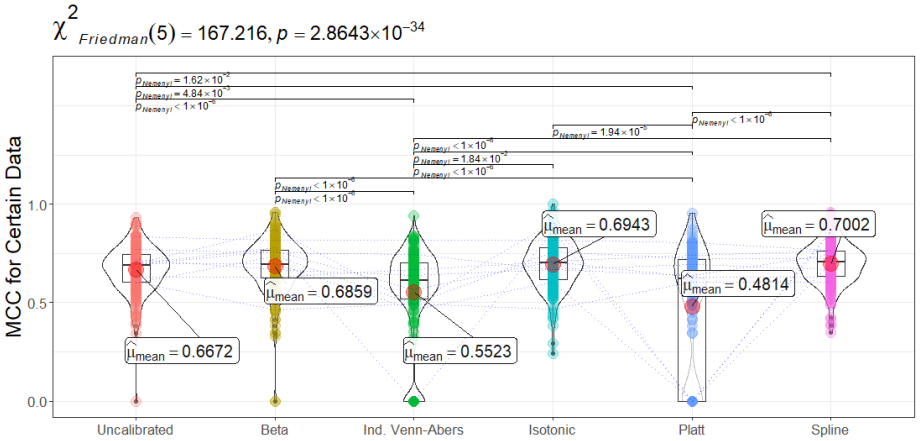


Figure 19. Friedman-Nemenyi significance test results regarding MCC for certain data instances of Split conformal prediction and calibration methods on the CatBoost model across 200 bootstrapping iterations.

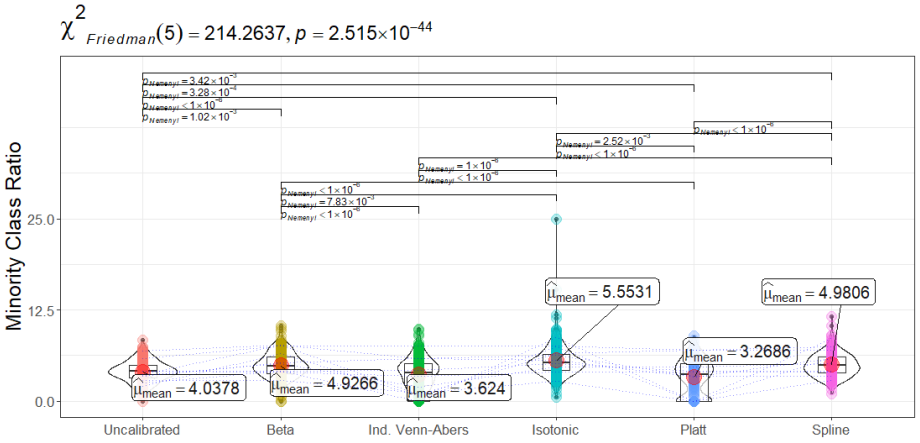


Figure 20. Friedman-Nemenyi significance test results regarding Minority Class Ratio of Split conformal prediction and calibration methods on the CatBoost model across 200 bootstrapping iterations.

Figure 20 (Minority Class Ratio) sheds light on another important consideration—the proportion of high-confidence positive predictions assigned to the minority class. A well-calibrated system should maintain a balance where underrepresented but clinically critical cases are neither overlooked nor overly emphasized. Isotonic Regression assigns approximately 5.6% of confidently labeled instances to positive cases, while Spline's 5.0% indicates a slightly more restrictive but stable classification boundary.

Overall, the results underscore the necessity of calibration-aware decision-making in predictive clinical modeling. The choice of calibration method should align with the clinical setting's tolerance for false negatives versus false positives. If the primary objective is to ensure that no high-risk patients are overlooked, Isotonic Regression may be preferable. If the aim is to maintain high precision and reduce overdiagnosis, Spline or Platt Scaling should be considered. Given the variations observed, hybrid strategies—such as dynamically adjusting calibration approaches based on incoming patient profiles or using ensemble calibration techniques—may offer additional improvements in real-world deployments.

5.5. SHAP-Based Explainability Analysis

The integration of calibration methods with CP not only refines probabilistic outputs but also reshapes the interpretability of predictive models, particularly in distinguishing high-confidence predictions from uncertain ones. To address RQ5, SHAP analysis was employed to dissect feature attributions across calibration strategies, revealing how post-hoc adjustments influence the drivers of certainty and uncertainty in ICU admission predictions. By comparing SHAP values for instances classified as certain (single-label sets) versus uncertain (both-label sets) under SCP, this analysis elucidates the interplay between calibration techniques and model interpretability in clinically actionable terms (see Figures 21–26).

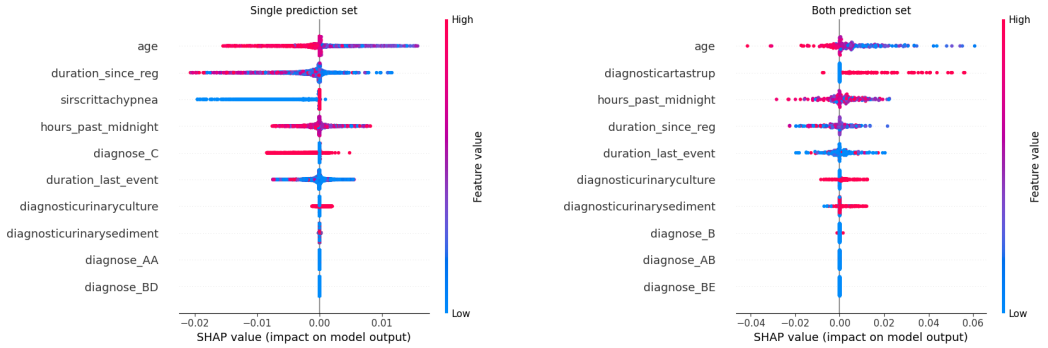
Feature Attribution Patterns Across Calibration Methods. For certain predictions, administrative and temporal features—such as age, duration_since_registration, and hours_past_midnight—consistently emerged as dominant contributors across all calibration methods. These features, which encode patient demographics and care timeline metadata, serve as robust anchors for high-confidence predictions, reflecting their stability in capturing systemic risk factors for sepsis progression. Clinical markers like diagnose_C (a diagnostic code for sepsis) and diagnosticurinaryculture (urinary tract infection indicators) also retained prominence, underscoring their established relevance in sepsis care pathways. This consistency suggests that calibration methods preserve the model's reliance on well-understood predictors when confidence is high, aligning with clinical intuition.

In contrast, uncertain predictions exhibited marked divergence in feature importance depending on the calibration approach. Under Beta calibration, uncertainty was primarily linked to rare diagnostic codes (e.g., diagnose_BE) and biochemical markers like diagnosticartastrup (arterial blood gas analysis), which are less frequently observed in sepsis cases. This pattern implies that ambiguity arises when the model encounters atypical clinical profiles, where sparse or conflicting laboratory results complicate risk assessment. Isotonic Regression, however, tied uncertainty to deviations from standard diagnostic pathways, emphasizing interactions between temporal features (duration_last_event) and less common lab tests (diagnosticurinarysediment). Such shifts highlight how non-parametric calibration amplifies the salience of edge-case clinical signals, potentially flagging patients whose trajectories defy conventional sepsis criteria.

Spline calibration introduced further nuance: while administrative features remained pivotal for certain predictions, uncertain cases saw heightened contributions from lab-result transitions (e.g., Leucocytes $\rightarrow$ LacticAcid), reflecting instability in interpreting serial biomarker trends. This suggests that Spline's piecewise adjustments, while smoothing

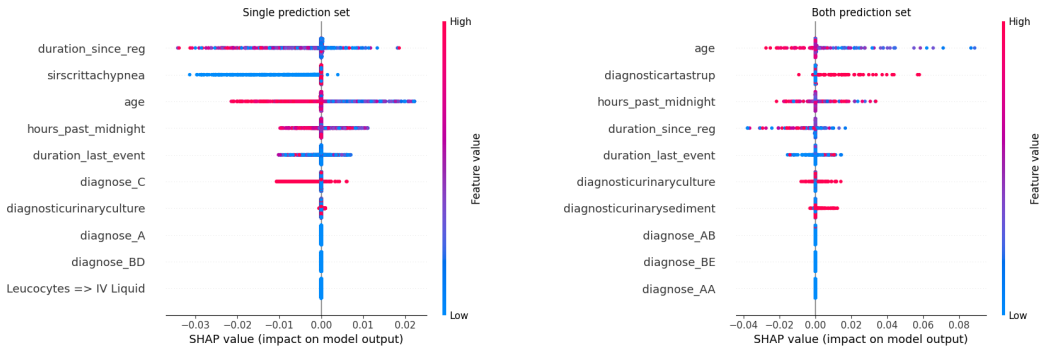
probability outputs, may inadvertently magnify the impact of transient or noisy clinical measurements. Venn-Abers calibration, despite its theoretical guarantees for validity, produced less coherent explanations for uncertain predictions, with paradoxical retention of administrative features (duration_since_reg) as key drivers even amid ambiguity. This discordance between feature importance and prediction uncertainty could undermine clinician trust, as administrative timestamps lack direct pathophysiological relevance to sepsis severity.

Parametric methods like Platt Scaling exhibited a hybrid behavior: certain predictions mirrored uncalibrated models in prioritizing age and diagnose_C, but uncertain predictions disproportionately emphasized temporal features (hours_past_midnight), divorcing explanations from clinical context. This misalignment indicates that logistic adjustments, while effective in probability correction, may obscure the biological rationale for uncertainty, rendering explanations less actionable. The uncalibrated model, unsurprisingly, displayed erratic attributions for uncertain cases, with rare diagnostic codes (diagnose_AB) and administrative artifacts dominating SHAP values—a consequence of unregulated probability overconfidence amplifying noise in feature space.



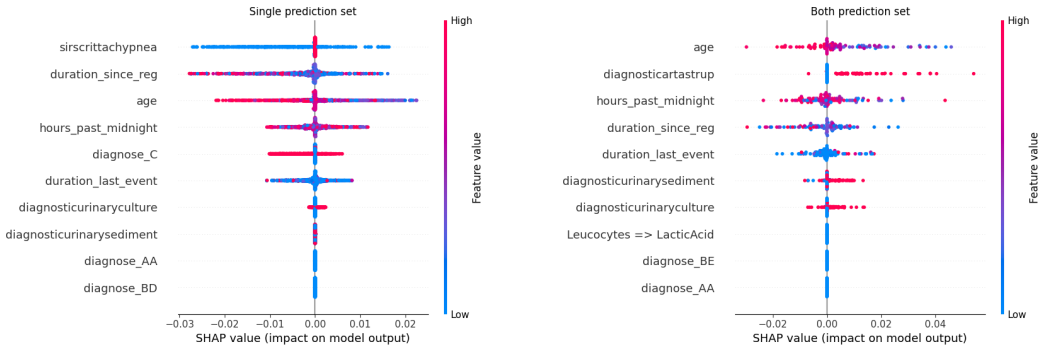
(**a**) (**b**)

Figure 21. Comparison of SHAP values for single and both prediction sets under Split conformal prediction with post-hoc Beta calibration applied to the CatBoost model. (**a**) Single prediction set. (**b**) Both prediction sets.



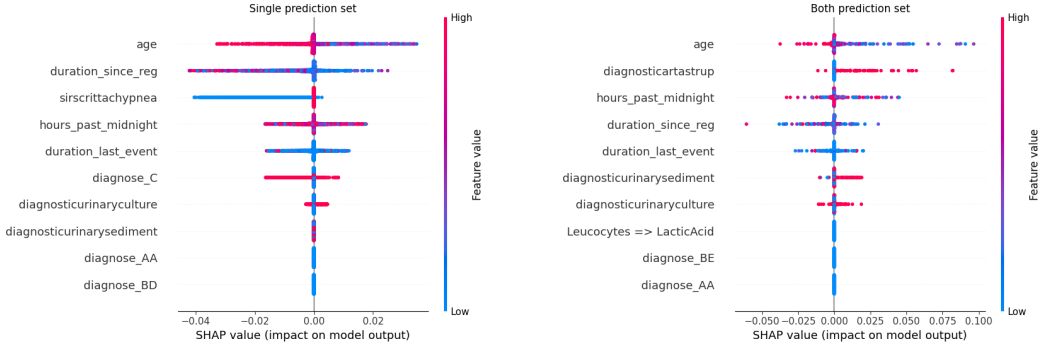
(**a**) (**b**)

Figure 22. Comparison of SHAP values for single and both prediction sets under Split conformal prediction with post-hoc Isotonic calibration applied to the CatBoost model. (**a**) Single prediction set. (**b**) Both prediction sets.



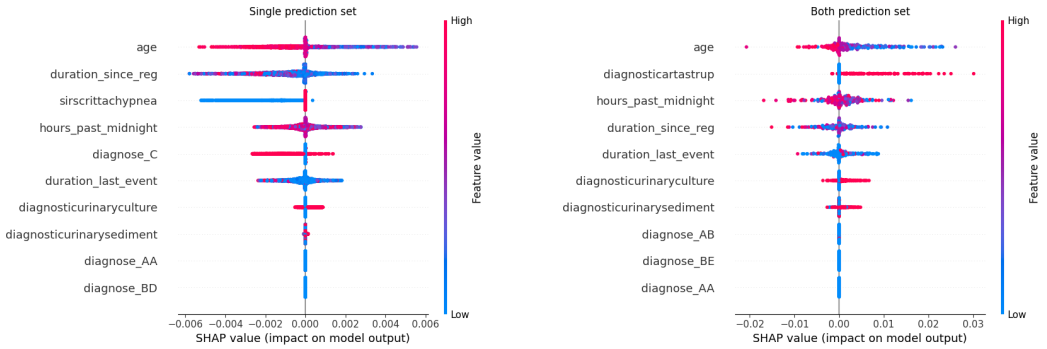
(**a**) (**b**)

Figure 23. Comparison of SHAP values for single and both prediction sets under Split conformal prediction with post-hoc Spline calibration applied to the CatBoost model. (**a**) Single prediction set. (**b**) Both prediction sets.



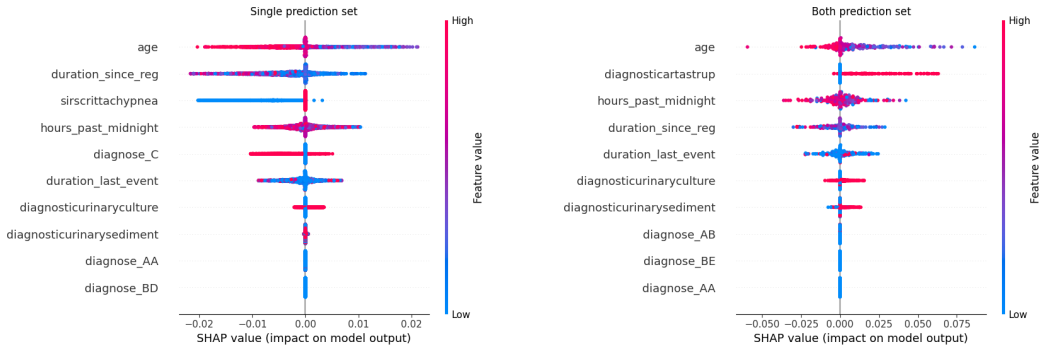
(**a**) (**b**)

Figure 24. Comparison of SHAP values for single and both prediction sets under Split conformal prediction with post-hoc Inductive Venn-Abers calibration applied to the CatBoost model. (**a**) Single prediction set. (**b**) Both prediction sets.



(**a**) (**b**)

Figure 25. Comparison of SHAP values for single and both prediction sets under Split conformal prediction with post-hoc Platt calibration applied to the CatBoost model. (**a**) Single prediction set. (**b**) Both prediction sets.



(**a**) (**b**)

Figure 26. Comparison of SHAP values for single and both prediction sets under Split conformal prediction with the uncalibrated CatBoost model. (**a**) Single prediction set. (**b**) Both prediction sets.

Relevance of Explainability in Calibration-CP Frameworks. The interplay between calibration and explainability carries profound implications for clinical deployment. Calibration methods, often perceived as purely statistical corrections, inherently reconfigure the model's internal reasoning—particularly in ambiguous cases. For instance, Isotonic Regression and Spline calibrations, by tethering uncertainty to clinically interpretable markers (e.g., lab anomalies or atypical diagnoses), enable clinicians to contextualize model hesitancy. A prediction flagged as uncertain due to elevated diagnosticartastrup values, for example, could prompt targeted blood gas analysis, transforming uncertainty into a diagnostic cue. Conversely, methods like Venn-Abers or Platt Scaling, which decouple uncertainty from domain-specific features, risk producing opaque explanations that hinder root-cause analysis. This analysis underscores that calibration is not a neutral adjustment but a reinterpretive act that reshapes model transparency. In high-stakes settings like sepsis management, where clinician trust hinges on interpretability, the choice of calibration method must balance statistical rigor with explanatory coherence. A well-calibrated model that attributes uncertainty to non-clinical factors (e.g., hours_past_midnight) risks being dismissed as a "black box" whereas one linking ambiguity to plausible clinical variables (e.g., conflicting lab trends) fosters collaborative decision-making.

Synthesis and Clinical Implications. The findings reveal a critical trade-off: parametric calibrations (Platt Scaling, Beta) enhance probability reliability but may dilute clinical interpretability, while non-parametric methods (Isotonic Regression, Spline) preserve feature relevance at the cost of increased computational complexity. For healthcare applications, Isotonic Regression emerges as a compelling compromise—its uncertainty explanations align with clinical workflows, enabling providers to reconcile model outputs with bedside observations. Spline calibration offers similar advantages but requires careful monitoring of its sensitivity to lab-result fluctuations. Ultimately, the integration of SHAP-based explainability with calibration-CP frameworks advances beyond mere technical validation; it bridges the gap between algorithmic outputs and clinical reasoning. By exposing how calibration reshapes feature attributions—and, by extension, model "thinking"—this approach empowers clinicians to audit uncertainty drivers, refine intervention protocols, and align predictive analytics with real-world care pathways. In doing so, it elevates PPM from a statistical exercise to a clinically embedded tool, where uncertainty is not a flaw but a diagnostically meaningful signal.

6. Discussion

This study reveals fundamental tensions and synergies in deploying machine learning for clinical predictive monitoring, where accuracy, calibration, UQ, and explainability

intersect. We selected sepsis management and the prediction of Intensive Care Unit (ICU) admission as our use case for several strategic reasons. Clinically, sepsis is a time-critical, high-mortality condition where early predictive analytics can directly impact patient outcomes. Methodologically, it is an ideal testbed for our framework. The progression of sepsis is inherently a process, lending itself to a Predictive Process Monitoring approach. Furthermore, sepsis data presents a unique combination of challenges that our multilayered framework is designed to address: (1) natural class imbalance, where the critical minority class (ICU admission) must be predicted reliably; (2) high clinical uncertainty, justifying the need for a robust uncertainty quantification (UQ) method like Conformal Prediction; and (3) data heterogeneity, with a complex mix of temporal, numerical, and categorical features that tests the power of modern ensemble models. Thus, sepsis provides a rich, clinically relevant context to evaluate the interplay between prediction, calibration, and explainability.

The findings challenge conventional assumptions about post-hoc calibration as a purely technical adjustment, positioning it instead as a transformative layer that reshapes model behavior, trustworthiness, and clinical utility. Below, we synthesize the methodological and practical implications of these insights.

Calibration as a Reinterpretive Act. The central paradox uncovered is that calibration methods, while designed to align probabilities with empirical outcomes, inherently reconfigure how models "reason" about uncertainty. For instance, isotonic regression not only reduced miscalibration but also redirected the model's attention during uncertain predictions toward clinically interpretable markers (e.g., ambiguous lab results or rare diagnoses). In contrast, parametric methods like Platt scaling preserved the model's original feature hierarchy, even when uncertainty arose from non-clinical artifacts like administrative timestamps. This divergence underscores that calibration is not a neutral correction but a reinterpretive process: it filters the model's internal logic through the lens of probability alignment, amplifying or suppressing specific drivers of uncertainty. For clinical applications, this means the choice of calibration directly influences whether uncertainty explanations align with medical intuition or obscure it—a critical factor in fostering clinician trust.

The Duality of Uncertainty Quantification. CP provided robust coverage guarantees across all methods, yet its clinical value depended heavily on calibration. Uncalibrated models achieved marginal coverage but exhibited a dangerous skew: 95% of errors impacted the minority class (ICU admissions), reflecting systemic underconfidence in high-risk predictions. Post-hoc calibration rectified this by redistricting uncertainty, tying it to domain-relevant edge cases (e.g., conflicting biomarker trends) rather than arbitrary thresholds. This duality—coverage as a statistical necessity, error distribution as a clinical imperative—highlights that CP frameworks must be calibration-aware to avoid perpetuating hidden biases. In practice, this means CP cannot be treated as a standalone module; its integration with calibration determines whether UQ serves as a safety net or a source of systemic blind spots.

Explainability as a Mirror of Calibration Strategy. SHAP analysis exposed how calibration methods rewrite the narrative of model decisions. While all methods agreed on feature importance for high-confidence predictions (e.g., prioritizing age or care timeline metadata), their handling of uncertainty diverged starkly. Non-parametric methods like Isotonic Regression and Spline calibration linked ambiguity to clinically meaningful signals—aberrant lab results, atypical diagnostic codes—effectively translating uncertainty into diagnostic hypotheses. Parametric methods, however, often attributed uncertainty to non-clinical features (e.g., hours_past_midnight), creating explanations that clash with clinician reasoning. This suggests that explainability is not static but calibration-contingent: the same model

can oscillate between clinically coherent and opaque explanations based on post-processing choices. For healthcare AI, this demands a paradigm shift—evaluating models not just by their accuracy or calibration metrics, but by how their entire decision hierarchy (certain and uncertain) aligns with domain expertise.

Methodological Trade-offs and Clinical Realities. The trade-offs between calibration approaches carry profound implications for real-world deployment. Isotonic Regression, despite its superior alignment of uncertainty with clinical logic, introduces computational complexity and sensitivity to small sample sizes. Platt Scaling, while efficient and stable, risks "explaining away" uncertainty through non-clinical features, potentially eroding trust. These tensions necessitate context-aware calibration strategies: in resource-constrained settings, Platt Scaling's efficiency might outweigh its explanatory limitations, while in tertiary care systems, Isotonic Regression's clinical coherence could justify its overhead. Importantly, no method universally dominated, underscoring the need for calibration to be tailored to institutional priorities—whether interpretability, computational efficiency, or strict probability alignment.

Our study focused on the deep analysis of a single, state-of-the-art sequential ensemble model (gradient boosting) as the engine for our predictive framework. We acknowledge that this is one of several powerful paradigms in ensemble learning. An alternative approach involves the fusion of multiple, heterogeneous base classifiers. Recent advancements in this area offer promising avenues for further improvement. For instance, theoretical frameworks for the optimal linear soft fusion of classifiers have been developed to combine model outputs in a way that provably minimizes post-fusion error [22]. Similarly, information-theoretic methods such as alpha-integration provide a principled and generalized approach for fusing the probability distributions from diverse models. Future work could explore whether a "super ensemble," created by fusing the outputs of a gradient boosting model with other classifiers (e.g., a deep neural network) using these optimal techniques, could provide an even more robust and reliable starting point for our subsequent calibration and conformal prediction pipeline. This could potentially lead to further reductions in predictive error and more precise uncertainty estimates.

Toward a Meta-Analysis of Model Design. These findings collectively argue for a holistic, systems-level view of predictive model design. Accuracy, calibration, uncertainty, and explainability are not isolated components but interacting dimensions that co-determine clinical utility. For example, a model with stellar AUROC but poorly calibrated probabilities may harm care pathways by overtriggering interventions, while a well-calibrated model with opaque explanations may stagnate due to clinician skepticism. This interdependence suggests that future frameworks must adopt multivariate evaluation metrics that weigh these dimensions jointly, rather than optimizing them in isolation. Beyond statistical performance, the practical deployment of this framework in a clinical setting requires addressing challenges such as latency, interpretability, and trust. Our approach was designed with these factors in mind. Latency is minimal at the point of care, as all computationally intensive training and calibration steps are performed offline, leaving only rapid, low-latency calculations for real-time inference. For interpretability, our SHAP-based analysis provides the crucial backend for a user-friendly clinical interface. By showing that calibration can align feature attributions with clinically relevant factors, we lay the groundwork for explanations that are not just available but also meaningful to a clinician. Most importantly, the framework is architected to build trust. It combines reliable, calibrated probabilities with the honest uncertainty of Conformal Prediction. When the model is uncertain about a case, it does not provide a single, potentially misleading prediction. Instead, it flags the case for human review by producing an ambiguous prediction set. This "knows what it doesn't

know" capability is fundamental for enabling safe, human-in-the-loop decision-making and fostering clinician confidence in the system.

Limitations as Catalysts for Innovation. We must acknowledge the limitations of this study to appropriately contextualize our findings. The analysis is based on a single, modestly-sized dataset from one hospital. This naturally limits the direct generalizability of the trained predictive model to other clinical settings, and we make no claim that this specific model is ready for deployment. However, the primary contribution of this work is the methodological framework itself. The dataset, with its characteristic clinical complexity and class imbalance, serves as a realistic testbed to rigorously evaluate the interplay between prediction, calibration, and uncertainty quantification. To ensure the robustness of our methodological insights, we employed an extensive bootstrapping procedure. This provides strong evidence that our findings—such as the observed trade-offs between different calibration methods and their downstream effects on conformal prediction and explainability—are stable and not an artifact of a single data split. Therefore, we propose our framework as a transferable blueprint for developing and validating trustworthy predictive models. A crucial and necessary direction for future work is to apply and validate this entire framework on larger, multi-center, and contemporary datasets to confirm the generalizability of these methodological findings across diverse patient populations and evolving clinical practices.

The study's focus on a single clinical context (sepsis) and static datasets invites exploration into dynamic, evolving care environments. Future work must test whether these findings generalize to settings with concept drift (e.g., emerging pathogens) or heterogeneous patient populations. Additionally, the computational costs of non-parametric calibration methods pose barriers to real-time deployment—a gap that hybrid approaches (e.g., adaptive Spline calibration) could bridge. Furthermore, while our study used aggregated AUC metrics to select the best base classifier, we acknowledge that a more granular analysis of the full ROC and Precision-Recall curves, particularly in the low false-positive rate regime, would offer deeper insights into model trade-offs for clinical deployment. Such a detailed investigation represents a valuable direction for future work. Finally, the ethical dimension of uncertainty attribution warrants scrutiny: if models attribute ambiguity to socioeconomic features (e.g., insurance status) rather than clinical factors, they risk exacerbating care disparities.

7. Related Work

As machine learning models become increasingly complex, ensuring their transparency is crucial, particularly in high-stakes domains such as healthcare and finance [23–25]. Transparency in machine learning encompasses both procedural transparency, which involves clear documentation of data collection, preprocessing and model training, and algorithmic transparency, which focuses on understanding how models arrive at their decisions [26,27]. Both aspects are essential for assessing fairness, reliability, and accountability in predictive systems [28]. A key aspect of transparency is UQ, which provides tools to measure and communicate the confidence of model predictions [27]. UQ involves estimating different types of uncertainty, including aleatoric uncertainty, which arises from inherent randomness in the data, and epistemic uncertainty, which stems from a lack of knowledge about the model or data [27,29–31]. Proper UQ ensures that predictions are not only accurate but also informative, helping decision-makers identify cases requiring additional validation or caution [32]. Various UQ methods exist, broadly categorized into Bayesian and Frequentist approaches [27].

CP provides a formal framework for UQ by constructing prediction sets with a guaranteed probability of containing the true label [4,33,34]. Unlike traditional point predictions,

which provide a single best estimate, conformal methods generate set-valued predictions that account for uncertainty in a rigorous, distribution-free manner. The fundamental principle involves computing non-conformity scores, which measure how unusual a prediction is relative to past observations, and using these scores to define prediction sets that maintain finite-sample validity under the assumption of exchangeability [33,34]. This ensures that, at a chosen confidence level, the true label falls within the prediction set with a probability at least equal to the specified threshold, making CP a well-calibrated approach to UQ [34,35]. A key advantage of CP is its model-agnostic nature, allowing integration with any machine learning model without requiring modifications to its internal structure [4,33]. Several variations have been introduced to improve its efficiency and applicability [34]. SCP partitions the data into separate training and calibration sets, enabling efficient computation of prediction sets without extensive retraining [34]. Mondrian conformal prediction conditions the non-conformity scores on predefined categories, such as class labels, ensuring category or class-wise valid coverage and improving performance on imbalanced datasets [36,37]. More recent approaches focus on adaptive and distribution-free conformal methods, which refine prediction sets dynamically to improve informativeness while preserving theoretical coverage guarantees [34,38].

Probability calibration is an essential process in machine learning that ensures the predicted probabilities of outcomes align with the actual observed frequencies. In the context of UQ and CP, proper calibration is crucial for generating reliable prediction intervals and quantifying uncertainty accurately [39,40]. Temperature scaling, for instance, has been shown to influence CP, suggesting that overconfident models may sometimes yield different prediction sets compared to well-calibrated models [39,41]. Venn-Abers calibration has been explored as an approach to improve probability estimates while maintaining the theoretical guarantees of CP [10].

PPM as a subdomain of Process Mining addresses fundamental aspects of process execution with a central focus on predicting upcoming events and performance indicators relevant to operational and strategical business goals [42–44]. Leveraging predictive models, predominant tasks encompass next event prediction [45,46], identification of anomalies [47], and the prediction of possible process outcomes [48]. Early implementations in PPM relied on conventional classification and regression techniques. In contrast, recent studies have increasingly adopted deep learning and ensemble methods to capture the complex temporal dependencies inherent in process data [49–52]. XAI has also emerged as a vital component in PPM, particularly for high-stakes decision-making scenarios [52,53]. UQ also plays an essential role in bolstering the reliability of PPM systems. UQ techniques estimate the confidence associated with predictions, which is crucial in risk-sensitive environments. Recent research has explored various UQ approaches, including Bayesian networks [54], ensemble learning strategies [55], Monte Carlo dropout [56,57], and CP methods [58,59]. Integrating these techniques into PPM frameworks allows practitioners to obtain both point forecasts and prediction intervals, thereby identifying cases that may require additional data or human oversight. High-stakes decision-making in PPM—relevant to sectors such as healthcare—demands models that are both accurate and interpretable [15]. In these contexts, the consequences of erroneous predictions are significant, making it imperative to understand not only what the models predict but also why. The combination of XAI and UQ provides a comprehensive toolkit: explainability methods elucidate the rationale behind model decisions, while uncertainty measures offer insights into the reliability of these decisions [60,61]. This integrated approach facilitates more informed and confident decision-making by illuminating both model strengths and limitations. In summary, recent advancements in PPM underscore the importance of combining explainability and uncertainty quantification to enhance transparency and trustworthiness.

8. Conclusions

In summary, our study demonstrates that integrating probability calibration techniques with CP framework enhances both the statistical robustness and interpretability of PPM in high-stakes healthcare applications. By systematically comparing post-hoc calibration methods we established that calibrated probabilities not only align better with empirical outcomes but also improve the reliability of uncertainty estimates generated via SPC. The experimental results indicate that while uncalibrated models can achieve high overall discrimination, they tend to misallocate uncertainty, particularly for the minority class. In contrast, calibration methods preserve classification performance while yielding lower log loss and improved ECE. Furthermore, explainability analysis based on SHAP values revealed that the calibration process reshapes feature attributions; non-parametric approaches, for instance, tie uncertainty more directly to clinically meaningful markers, thereby enhancing model transparency and fostering clinician trust. These findings highlight that calibration is not merely a technical adjustment but a transformative layer that influences model reasoning and uncertainty quantification.

Author Contributions: Conceptualization, N.M.; Data curation, F.A.S.; Formal analysis, A.E.; Funding acquisition, N.M.; Methodology, M.M., A.E. and N.M.; Project administration, N.M.; Resources, A.E. and N.M.; Software, M.M. and F.A.S.; Supervision, N.M.; Validation, M.M., F.A.S. and N.M.; Visualization, M.M. and F.A.S.; Writing—original draft, M.M., F.A.S., A.E. and N.M.; Writing—review & editing, M.M., F.A.S., A.E. and N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the German Federal Ministry of Research, Technology and Space under grant number 01IS24048C (project EINHORN).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in 4TU.RESEARCHDATA at https://data.4tu.nl/articles/_/12707639/1 (accessed on 30 January 2025), reference number [16].

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
2. Aliyeva, K.; Mehdiyev, N. Uncertainty-aware multi-criteria decision analysis for evaluation of explainable artificial intelligence methods: A use case from the healthcare domain. *Inf. Sci.* **2024**, *657*, 119987. [CrossRef]
3. Tomsett, R.; Preece, A.; Braines, D.; Cerutti, F.; Chakraborty, S.; Srivastava, M.; Pearson, G.; Kaplan, L. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* **2020**, *1*, 100049. [CrossRef] [PubMed]
4. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, NY, USA, 2005; Volume 29.
5. Vovk, V.; Petej, I.; Fedorova, V. From conformal to probabilistic prediction. In *Proceedings of the Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, 19–21 September 2014, Proceedings 10*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 221–230.
6. Balasubramanian, V.; Ho, S.S.; Vovk, V. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*; Newnes: Oxford, UK, 2014.
7. Angelopoulos, A.; Bates, S.; Malik, J.; Jordan, M.I. Uncertainty sets for image classifiers using conformal prediction. *arXiv* **2020**, arXiv:2009.14193.
8. Vovk, V.; Petej, I.; Fedorova, V. Large-scale probabilistic prediction with and without validity guarantees. In Proceedings of the NIPS, Montreal, QC, Canada, 7–12 December 2015; Volume 2015.

9. Tibshirani, R.J.; Foygel Barber, R.; Candes, E.; Ramdas, A. Conformal prediction under covariate shift. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 2530–2540.
10. van der Laan, L.; Alaa, A.M. Self-Consistent Conformal Prediction. *arXiv* **2024**, arXiv:2402.07307. [CrossRef]
11. Gibbs, I.; Candes, E. Adaptive conformal inference under distribution shift. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 1660–1672.
12. Löfström, H.; Löfström, T.; Johansson, U.; Sönströd, C. Calibrated explanations: With uncertainty information and counterfactuals. *Expert Syst. Appl.* **2024**, *246*, 123154. [CrossRef]
13. Slack, D.; Hilgard, A.; Singh, S.; Lakkaraju, H. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9391–9404.
14. Machado, A.F.; Charpentier, A.; Flachaire, E.; Gallic, E.; Francois, H. Post-Calibration Techniques: Balancing Calibration and Score Distribution Alignment. In Proceedings of the NeurIPS 2024 Workshop on Bayesian Decision-Making and Uncertainty, Vancouver, BC, Canada, 14 December 2024.
15. Rojas, E.; Munoz-Gama, J.; Sepúlveda, M.; Capurro, D. Process mining in healthcare: A literature review. *J. Biomed. Inform.* **2016**, *61*, 224–236. [CrossRef] [PubMed]
16. Mannhardt, F.; Blinde, D. Analyzing the trajectories of patients with sepsis using process mining. In Proceedings of the RADAR+ EMISA 2017, Essen, Germany, 12–12 June 2017; pp. 72–80.
17. Mannhardt, F. Sepsis Cases-Event Log. 2016. Available online: https://doi.org/10.4121/uuid:915d2bfb-7e84-49ad-a286-dc35f063a460 (accessed on 30 January 2025).
18. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **1999**, *10*, 61–74.
19. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
20. Bischl, B.; Binder, M.; Lang, M.; Pielok, T.; Richter, J.; Coors, S.; Thomas, J.; Ullmann, T.; Becker, M.; Boulesteix, A.L.; et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2023**, *13*, e1484. [CrossRef]
21. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
22. Vergara, L.; Salazar, A. On the Optimum Linear Soft Fusion of Classifiers. *Appl. Sci.* **2025**, *15*, 5038. [CrossRef]
23. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
24. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [CrossRef]
25. Weber, P.; Carl, K.V.; Hinz, O. Applications of explainable artificial intelligence in finance—A systematic review of finance, information systems, and computer science literature. *Manag. Rev. Q.* **2024**, *74*, 867–907. [CrossRef]
26. Weller, A. Transparency: Motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Springer: Cham, Switzerland, 2019; pp. 23–40.
27. Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q.V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual, 19–21 May 2021; pp. 401–413.
28. Lehmann, C.A.; Haubitz, C.B.; Fügener, A.; Thonemann, U.W. The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. *Prod. Oper. Manag.* **2022**, *31*, 3419–3434. [CrossRef]
29. Depeweg, S. Modeling Epistemic and Aleatoric Uncertainty with Bayesian Neural Networks and Latent Variables. Ph.D. Thesis, Technische Universität München, München, Germany, 2019.
30. Der Kiureghian, A.; Ditlevsen, O. Aleatory or epistemic? Does it matter? *Struct. Saf.* **2009**, *31*, 105–112. [CrossRef]
31. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5574–5584.
32. Marusich, L.R.; Bakdash, J.Z.; Zhou, Y.; Kantarcioglu, M. Using ai uncertainty quantification to improve human decision-making. *arXiv* **2023**, arXiv:2309.10852.
33. Shafer, G.; Vovk, V. A tutorial on conformal prediction. *J. Mach. Learn. Res.* **2008**, *9*, 371–421.
34. Angelopoulos, A.N.; Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv* **2021**, arXiv:2107.07511.
35. Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. [CrossRef]
36. Vovk, V.; Lindsay, D.; Nouretdinov, I.; Gammerman, A. *Mondrian Confidence Machine*; Technical Report; Royal Holloway University of London: London, UK, 2003.
37. Toccaceli, P.; Gammerman, A. Combination of inductive mondrian conformal predictors. *Mach. Learn.* **2019**, *108*, 489–510. [CrossRef]

38. Romano, Y.; Sesia, M.; Candes, E. Classification with valid and adaptive coverage. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3581–3591.
39. Xi, H.; Huang, J.; Feng, L.; Wei, H. Does Confidence Calibration Help Conformal Prediction? *arXiv* **2024**, arXiv:2402.04344. [CrossRef]
40. Biggio, L.; Wieland, A.; Chao, M.A.; Kastanis, I.; Fink, O. Uncertainty-aware prognosis via deep gaussian process. *IEEE Access* **2021**, *9*, 123517–123527. [CrossRef]
41. Dabah, L.; Tirer, T. On Calibration and Conformal Prediction of Deep Classifiers. *arXiv* **2024**, arXiv:2402.05806. [CrossRef]
42. Maggi, F.M.; Di Francescomarino, C.; Dumas, M.; Ghidini, C. Predictive monitoring of business processes. In Proceedings of the Advanced Information Systems Engineering: 26th International Conference, CAiSE 2014, Thessaloniki, Greece, 16–20 June 2014, Proceedings 26; Springer: Berlin/Heidelberg, Germany, 2014; pp. 457–472.
43. Di Francescomarino, C.; Ghidini, C. Predictive process monitoring. In *Process Mining Handbook*; Springer International Publishing: Cham, Switzerland, 2022; pp. 320–346.
44. Pfeiffer, P.; Rombach, A.; Majlatow, M.; Mehdiyev, N. From Theory to Practice: Real-World Use Cases on Trustworthy LLM-Driven Process Modeling, Prediction and Automation. *arXiv* **2025**, arXiv:2506.03801.
45. De Leoni, M.; Van Der Aalst, W.M.; Dees, M. A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs. *Inf. Syst.* **2016**, *56*, 235–257. [CrossRef]
46. Brunk, J.; Stierle, M.; Papke, L.; Revoredo, K.; Matzner, M.; Becker, J. Cause vs. effect in context-sensitive prediction of business process instances. *Inf. Syst.* **2021**, *95*, 101635. [CrossRef]
47. Böhmer, K.; Rinderle-Ma, S. Mining association rules for anomaly detection in dynamic process runtime behavior and explaining the root cause to users. *Inf. Syst.* **2020**, *90*, 101438. [CrossRef]
48. Di Francescomarino, C.; Dumas, M.; Maggi, F.M.; Teinemaa, I. Clustering-based predictive process monitoring. *IEEE Trans. Serv. Comput.* **2016**, *12*, 896–909. [CrossRef]
49. Teinemaa, I.; Dumas, M.; Rosa, M.L.; Maggi, F.M. Outcome-Oriented Predictive Process Monitoring: Review and Benchmark. *ACM Trans. Knowl. Discov. Data* **2019**, *13*, 1–57. [CrossRef]
50. Márquez-Chamorro, A.E.; Resinas, M.; Ruiz-Cortés, A. Predictive Monitoring of Business Processes: A Survey. *IEEE Trans. Serv. Comput.* **2018**, *11*, 962–977. [CrossRef]
51. Polato, M.; Sperduti, A.; Burattin, A.; de Leoni, M. Data-aware remaining time prediction of business process instances. In Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China, 6–11 July 2014; pp. 816–823.
52. Mehdiyev, N.; Fettke, P. Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring. In *Interpretable Artificial Intelligence: A Perspective of Granular Computing*; Springer: Cham, Switzerland, 2021; pp. 1–28.
53. Mehdiyev, N.; Majlatow, M.; Fettke, P. Counterfactual Explanations in the Big Picture: An Approach for Process Prediction-Driven Job-Shop Scheduling Optimization. *Cogn. Comput.* **2024**, *16*, 2674–2700. [CrossRef]
54. Prasidis, I.; Theodoropoulos, N.P.; Bousdekis, A.; Theodoropoulou, G.; Miaoulis, G. Handling uncertainty in predictive business process monitoring with Bayesian networks. In Proceedings of the 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA), Chania Crete, Greece, 12–14 July 2021; pp. 1–8.
55. Shoush, M.; Dumas, M. Prescriptive process monitoring under resource constraints: A causal inference approach. In *Proceedings of the International Conference on Process Mining, Eindhoven, The Netherlands, 31 October–4 November 2021*; Springer: Cham, Switzerland, 2021; pp. 180–193.
56. Weytjens, H.; De Weerdt, J. Learning uncertainty with artificial neural networks for predictive process monitoring. *Appl. Soft Comput.* **2022**, *125*, 109134. [CrossRef]
57. Mehdiyev, N.; Majlatow, M.; Fettke, P. Quantifying and explaining machine learning uncertainty in predictive process monitoring: An operations research perspective. *Ann. Oper. Res.* **2025**, *347*, 991–1030. [CrossRef]
58. Bozorgi, Z.D.; Dumas, M.; Rosa, M.L.; Polyvyanyy, A.; Shoush, M.; Teinemaa, I. Learning When to Treat Business Processes: Prescriptive Process Monitoring with Causal Inference and Reinforcement Learning. In Proceedings of the International Conference on Advanced Information Systems Engineering, Zaragoza, Spain, 12–16 June 2023; Springer: Cham, Switzerland, 2023; pp. 364–380.
59. Mehdiyev, N.; Majlatow, M.; Fettke, P. Augmenting post-hoc explanations for predictive process monitoring with uncertainty quantification via conformalized Monte Carlo dropout. *Data Knowl. Eng.* **2025**, *156*, 102402. [CrossRef]

60. Mehdiyev, N.; Majlatow, M.; Fettke, P. Communicating uncertainty in machine learning explanations: A visualization analytics approach for predictive process monitoring. In Proceedings of the World Conference on Explainable Artificial Intelligence, Valletta, Malta, 17–19 July 2024; Springer: Cham, Switzerland, 2024; pp. 420–438.
61. Mehdiyev, N.; Majlatow, M.; Fettke, P. Integrating permutation feature importance with conformal prediction for robust Explainable Artificial Intelligence in predictive process monitoring. *Eng. Appl. Artif. Intell.* **2025**, *149*, 110363. [CrossRef]