JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

A Five-Step Multi-Dimensional Learnability Validation Protocol for Meningitis Clinical Risk Assessment

Fathimah Al-Ma'shumah [1,*], Yasi Dani [1], Yani Triyani [2], and Dani Suandi [1]

[1] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia; e-mail: fathimah.al-mashumah@binus.ac.id; yasi.dani@binus.ac.id; dani.suandi@binus.ac.id
[2] Department of Clinical Pathology, Faculty of Medicine, Bandung Islamic University, Bandung 40116, Indonesia; e-mail: yani.triyani@unisba.ac.id
* Corresponding Author : Fathimah Al-Ma'shumah

Abstract: Clinical machine learning is routinely judged on discrimination alone. A model that separates classes well is reported as successful, even when its probabilities are miscalibrated, its explanations point at variables with no measured association with the outcome, and nobody has checked whether the target can be learned from the available features at all. This study proposes a five-step multidimensional learnability validation protocol that tests a prediction target on five axes: feature-target dependence before modelling, calibration through Brier score decomposition, clinical utility through decision curve analysis, interpretability through SHapley Additive exPlanations, and agreement between fitted coefficient signs and the directions reported in the clinical literature. The protocol is the contribution; no new learning algorithm is proposed. Five established classifiers, Logistic Regression, Random Forest, LightGBM, CatBoost, and XGBoost, are used as instruments on a public synthetic meningitis dataset of 1,200 records with 10 clinical predictors and three prediction targets. Under repeated stratified cross-validation with five folds and five repeats, the four tree-based models reached mean areas under the curve of 0.969 to 0.971 and mean Matthews correlation coefficients of 0.780 to 0.793 for aetiological diagnosis, and 0.969 to 0.972 and 0.791 to 0.802 for clinical risk level; both targets passed all five validation steps. The mortality target failed all five: the strongest feature-target correlation was 0.067 and the strongest mutual information 0.013, the Brier score of the Platt-scaled reference model (0.0949) did not fall below the uncertainty baseline (0.0944), no threshold between 0.05 and 0.50 gave net benefit above both default strategies, the leading SHAP attribution fell on a variable with no measured association, and only two of five literature-referenced coefficient signs agreed. Judged only by whether mean discrimination exceeded chance, mortality could have appeared marginally viable. The protocol is demonstrated, not externally validated, and a target that passes is eligible for further validation rather than ready for clinical use.

Keywords: Brier score decomposition; Clinical decision support; Decision curve analysis; Explainable artificial intelligence; Gradient boosting; Meningitis classification; Model calibration.

Received: September, 8th 2026
Revised: October, 6th 2026
Accepted: October, 11th 2026
Published: October, 25th 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) licenses (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Meningitis, an inflammation of the membranes surrounding the brain and spinal cord, is primarily caused by bacterial and viral pathogens and remains a major infectious cause of disability-adjusted life years lost worldwide [1]. Young children and older adults are disproportionately affected [2]. Bacterial meningitis has a case-fatality rate of 20–30% among adults, and approximately 20% of survivors experience permanent neurological sequelae [2]. Because culture-based confirmation following lumbar puncture typically takes 24–48 hours, there is a strong clinical need for triage tools that can stratify patients using routinely available clinical data before microbiological confirmation [2].

However, when considering the development of a clinical risk model, a hospital has no straightforward way to determine in advance whether the available data contain sufficient information to support reliable predictions for a particular clinical outcome. The conventional

approach is to develop a model and assess its performance, often using the area under the receiver operating characteristic curve (AUC). Although useful for measuring discrimination, AUC alone does not establish whether predicted probabilities are reliable, whether model-guided decisions provide clinical benefit, or whether the model relies on features that have a measurable association with the outcome. Consequently, a model may achieve discrimination above chance while failing to satisfy other important requirements for meaningful clinical prediction. Such limitations may remain undetected until substantial modelling effort has already been invested.

To address this issue, this study introduces a five-step validation protocol for assessing whether a clinical prediction target is learnable from the available features and whether a fitted model warrants further validation. The protocol evaluates five complementary dimensions: feature–target dependence, probability calibration, decision-analytic utility, consistency between feature importance and measured associations, and agreement between model coefficient directions and published clinical evidence. Five established classifiers are used as instruments to evaluate the protocol across three meningitis-related prediction targets. The methodological contribution therefore lies in the design and evaluation of the validation protocol rather than in the modification or introduction of a new classification algorithm.

Machine learning (ML) has shown considerable promise in meningitis diagnosis and risk stratification. Existing approaches range from classifier comparisons involving support vector machines and random forests [3], decision trees based on routine clinical signs [4], and ensembles for aetiological discrimination [5] to hybrid architectures built on LightGBM stacking [6], [7], boosted and stacked ensembles [8], [9], deep features with kernel classifiers [10], and belief bidirectional networks [11]. AI has also been applied to nationwide differential diagnosis [12], while multivariable rules for distinguishing bacterial from viral meningitis using cerebrospinal fluid (CSF) and haematological variables have a long clinical history [13], [14]. In prognostic modelling, Zhou et al. [15] validated a 30-day mortality model for adult bacterial meningitis, whereas Luo et al. [16] investigated meningoencephalitis in paediatric scrub typhus. Since 2023, this research landscape has expanded to include central nervous system infections in children [17], outcomes of neonatal bacterial meningitis [18], [19], pathogens and predictors of unfavourable outcomes [20], diagnosis in resource-limited settings [21], and nosocomial infections following severe traumatic brain injury [22].

Despite this progress, three methodological challenges continue to constrain the clinical translation of ML-based prediction models. First, many studies emphasise discrimination without adequately assessing calibration, which measures the agreement between predicted probabilities and observed outcome frequencies and is essential when predictions inform threshold-based decisions [23], [24]. Strong discrimination does not necessarily imply reliable probability estimates, as a model can achieve a high AUC while remaining severely miscalibrated [25]. Second, post-hoc explanation methods, including SHapley Additive exPlanations (SHAP) [26] or LIME [27], are increasingly used to interpret model predictions; however, the limitations of SHAP under weak predictive performance remain insufficiently formalised, and explanations of black-box models may create unwarranted confidence in their outputs [28]. Third, and more fundamentally, whether the available features contain sufficient information to support prediction of the target is rarely assessed before model development. Without such a preliminary assessment, subsequent performance evaluation and interpretability analysis may produce apparently meaningful results even when the underlying data provide little support for the prediction task [23].

Related research has examined interpretability requirements for clinical prediction models [24], the potential of XAI tools to improve decision-making in neurology [29], and clinicians' preferences for SHAP waterfall plots in case-level interpretation [30]. Brier score decomposition [31] and decision curve analysis (DCA) [32] have also become established approaches to clinical risk model evaluation [15], [33], consistent with recommendations from TRIPOD [34], and the performance assessment framework of Steyerberg et al.[35]. Meanwhile, SMOTE [36] remains a widely used resampling strategy, although its interaction with uninformative features has not been fully characterized [37]. A systematic review [38] further highlights inconsistencies in the reporting of validation beyond discrimination. Complementary studies by Mousa et al. [39] and Yang et al. [40] investigated how resampling affects classifier behaviour and SHAP attributions, while Sobhan and Mondal [41] demonstrated that attribution rankings are sensitive to preprocessing choices, providing a rationale for the fea-

ture-plausibility assessment in Step 4. Vickers and Holland [42] reaffirmed DCA as an approach for evaluating potential clinical benefit, whereas Wu and Wang [43] and Johnston et al. [44] demonstrated calibrated and explainable risk stratification using gradient-boosted trees.

Taken together, existing research indicates a methodological gap: calibration assessment, formal evaluation of interpretability-related evidence, and pre-modelling verification of target learnability have not been integrated into a single mathematically grounded protocol that also demonstrates the implications of insufficient learnability. Research published in this journal has addressed related aspects, including calibration and explainability auditing in a multimodal clinical framework [45], SHAP faithfulness across domains [46], and broader principles for healthcare AI [47]; however, these studies do not establish learnability through a dedicated assessment before model fitting. Accordingly, this study aims to introduce a pre-modelling learnability screen and a five-step validation protocol to assess whether a clinical prediction target is sufficiently supported by the available features and whether a fitted model warrants further validation. The central premise is that discrimination alone is insufficient to establish the suitability of a clinical prediction target: a more comprehensive assessment should also examine probability calibration, potential decision-analytic benefit, the plausibility of feature attributions, and the consistency of fitted coefficient directions with published clinical evidence.

The contributions of this work are fourfold. First, it presents a comparative benchmark of five established classifiers—LightGBM [48], CatBoost [49], XGBoost [50], Random Forest [51], and Logistic Regression—used as analytical instruments to evaluate three clinically distinct prediction targets in a public meningitis dataset. Second, it presents five formal results, accompanied by proof sketches, that support the theoretical rationale and methodological implementation of the proposed protocol (Section 2). Third, it introduces a five-step learnability validation protocol in which the initial feature–target dependence screen is conducted before model fitting, while the final step assesses whether fitted coefficient directions agree with directions established from the clinical literature. Fourth, it empirically demonstrates how the protocol characterises a prediction target for which the available features provide insufficient predictive support, a limitation that discrimination-based evaluation alone may fail to identify.

The remainder of this paper is organised as follows. Section 2 presents the formal results underpinning the protocol. Section 3 describes the protocol, dataset, and experimental methodology. Section 4 reports and discusses the results concerning discrimination, calibration, decision-analytic utility, interpretability, and coefficient-direction auditing, together with the study's limitations. Section 5 compares the proposed protocol with eleven previous studies, distinguishing methodological coverage from predictive performance. Finally, Section 6 presents the conclusions.

2. Preliminaries

Section 2 states five formal results that support the theoretical rationale and methodological implementation of the proposed protocol. These results address SHAP attribution, Brier score behaviour, decision curve dominance, integer risk score approximation, and leakage-free resampling. The first four results support the interpretation or implementation of the model-based evaluation stages, while the final result establishes the rationale for applying SMOTE only after data partitioning.

2.1. SHAP Attribution and Its Failure Mode

Let $f : \mathbb{R}^p \to \mathbb{R}$ denote a trained classifier and $x \in \mathbb{R}^p$ a feature vector. The SHAP value for feature j is defined as the weighted average marginal contribution of that feature across all coalitions [26], given in Eq. (1), where $N = \{1, \ldots, p\}$ is the set of all features and $f(\mathbf{x}_S) = \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$ denotes the expected model output when only the features in S are fixed at their observed values.

$$\varphi_j(f, \mathbf{x}) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\ (p - |S| - 1)!}{p!} \left[f\left(\mathbf{x}_{S \cup \{j\}}\right) - f(\mathbf{x}_S) \right] \quad (1)$$

Theorem 1. Shapley Axioms. *The attribution defined by Eq. (1) is the unique function satisfying Efficiency, that is, the* $\sum_{j=1}^{p} \varphi_j\,(f, \boldsymbol{x}) = f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{X})]$*; together with Symmetry, Dummy, and Linearity.*

Proof of Theorem 1. Efficiency follows from summing Eq. (1) over all j in N, where each marginal contribution appears exactly once with total Shapley weight summing to unity. Symmetry, Dummy, and Linearity follow from the combinatorial structure of the weighting kernel. Uniqueness is the classical characterisation theorem of Shapley [52].

Remark 1 (Noise Attribution). *Theorem 1 holds regardless of the predictive power of* f*. When f operates at chance-level AUC, a high mean absolute SHAP value should be interpreted cautiously because attribution magnitude reflects the model's internal allocation of predictive contributions and does not, by itself, establish a clinical association between a feature and the outcome [24], [26]. This observation motivates Step 4 of the proposed protocol. Since SHAP faithfulness may also deteriorate under distribution shift [46], the protocol assesses whether the leading SHAP attribution corresponds to a feature that satisfies the measured feature–target dependence criterion in Step 1, rather than accepting the attribution as sufficient evidence on its own.*

2.2. Brier Score Lower Bound

The Brier score decomposes into reliability, resolution, and uncertainty components as given in Eq. (2), where the uncertainty term UNC is fixed by the class distribution [31].

$$\mathrm{BS} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}, \qquad \mathrm{UNC} = \bar{y}\,(1 - \bar{y}) \tag{2}$$

Proposition 2 (Brier Lower Bound). *When a model is calibrated, and its resolution tends to zero, the expected Brier score tends to* UNC*, which is the Brier score of the naive prevalence model. Any model with Brier score greater than or equal to* UNC *therefore provides no improvement over the prevalence model; that is, a Brier skill score* $1 - BS/UNC$ *that is not positive.*

Proof of Proposition 2. By the law of total variance, for a calibrated model the expected Brier score equals UNC minus resolution. When the variance of the conditional probability is zero, resolution is zero, and the bound is binding [31].

2.3. Decision Curve Dominance and Score Monotonicity

Net benefit at threshold probability p_{t} is defined in Eq. (3), where TP and FP denote true and false positive counts, and n is the sample size [32], [42].

$$\mathrm{NB}(p_{\mathrm{t}}) = \frac{\mathrm{TP}}{n} - \frac{p_{\mathrm{t}}}{1 - p_{\mathrm{t}}} \cdot \frac{\mathrm{FP}}{n} \tag{3}$$

Theorem 3. DCA Dominance. *A model dominates the treat-none strategy at threshold* p_{t} *if and only if the ratio* TP/FP *exceeds* $p_{\mathrm{t}}/(1 - p_{\mathrm{t}})$*, and it dominates the treat-all strategy if and only if the ratio* TN/FN *exceeds* $(1 - p_{t})/p_{t}$*. A model provides clinical utility at* p_{t} *when both conditions hold.*

Proof of Theorem 3. The treat-none strategy has net benefit zero, so the model dominates it when TP/n exceeds $p_{\mathrm{t}}/(1 - p_{\mathrm{t}})$ multiplied by FP/n, which rearranges to the first condition [32]. The treat-all strategy has net benefit $(TP + FN)/n - [p_{\mathrm{t}}/(1 - p_{\mathrm{t}})](FP + TN)/n$; subtracting it from Eq. (3) leaves $[p_{\mathrm{t}}/(1 - p_{\mathrm{t}})]TN/n - FN/n$, which is positive exactly under the second condition.

Proposition 4 (Score Rounding Bound). *The integer risk score* $S(\boldsymbol{x})$ *sums integer* $w_j = round\,(K\beta_j/\beta^*)$*, with* $\beta^* = \max_j |\beta_j|$*, over the m indicators* $I_j(\boldsymbol{x})$ *that mark feature j beyond its threshold* c_j*. Relative to the unrounded score, its predicted probability can differ by at most the rounding error* δ_{round} *given in Eq. (4). For the model fitted in this study,* K *=5,* m *=10,* β^* *=0.897,* δ_{round}*=0.224.*

$$\delta_{\text{round}} = \frac{m\beta^*}{8K} \tag{4}$$

Proof of Proposition 4. Each weight differs from $K\beta_j/\beta^*$ by at most one half, so after rescaling to the log-odds scale by β^*/K each indicator contributes an error of at most $\beta^*/(2K)$. Summing over the m indicators and multiplying by the Lipschitz constant of the logistic function, bounded above by one quarter, yields Eq. (4). When δ_{round} exceeds typical interpatient differences in predicted probability, the score cannot preserve their ranking.

2.4. Resampling Non-Leakage

The SMOTE procedure generates synthetic minority-class observations through interpolation between existing samples. Equation (5) represents the interpolation procedure used to generate these observations.

$$x_{\text{syn}} = x_i + \lambda(x_{\text{nn}} - x_i),\ \lambda \sim \text{Uniform}(0,1) \tag{5}$$

Proposition 5 (SMOTE Non-leakage). *Applying SMOTE after stratified train-test splitting preserves the true marginal test distribution. Applying SMOTE before splitting introduces synthetic interpolations into the test fold, yielding inflated minority-class recall estimates [23], [36].*

Proof of Proposition 5. Synthetic points are convex combinations of training samples. If generated globally, they appear in both folds, violating the independence assumption underlying generalisation bounds and rendering test-set evaluation optimistic [53].

3. Proposed Method

3.1. Five-Step Learnability Validation Protocol

The proposed protocol assesses whether a clinical prediction target is supported by the available features and whether a fitted model warrants further validation. It combines a pre-modelling dependence screen with four subsequent evaluations of calibration, decision-analytic utility, SHAP attribution plausibility, and agreement between fitted coefficient directions and clinical literature. A failure at Step 1 triggers a warning rather than termination, allowing the remaining steps to provide complementary evidence. Figure 1 illustrates the workflow, while Algorithm 1 details its implementation.

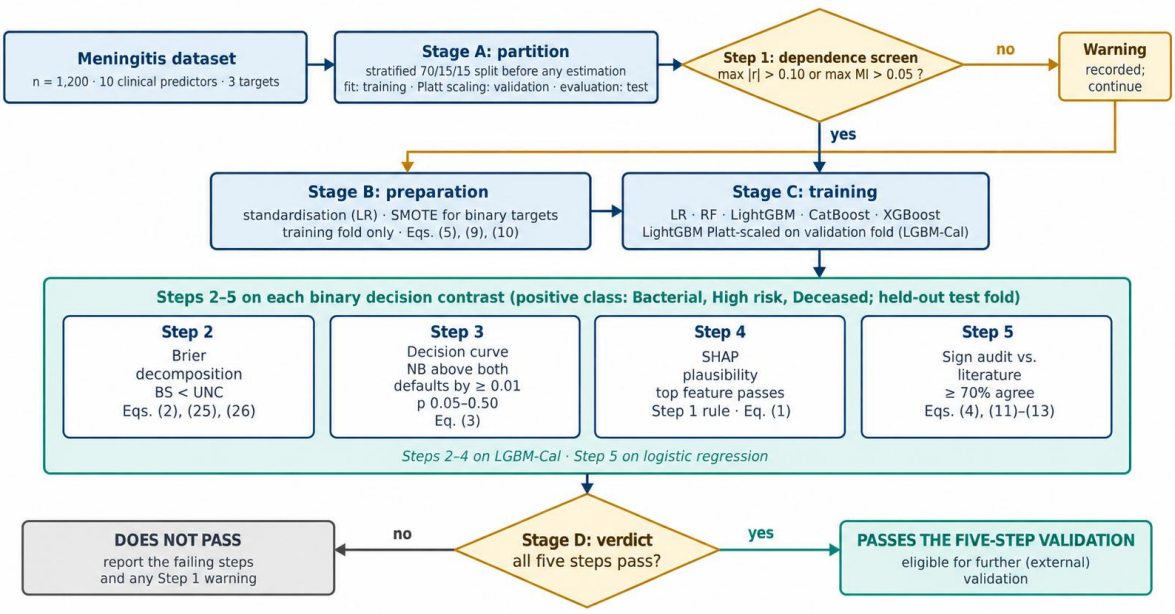


Figure 1. Workflow of the proposed five-step learnability validation protocol

Rectangles represent processing stages, labelled A–D, and diamonds represent decision points. A target that fails Step 1 receives a screening warning but proceeds through Steps 2–5. A target that passes all five steps is considered eligible for further external validation, not ready for clinical use.

Algorithm 1. Five-step learnability validation protocol

INPUT: Dataset D, feature set F, target set T; screening thresholds $\tau_r = 0.10$ for correlation and $\tau_{\mathrm{MI}} = 0.05$ nats for mutual information; threshold range $[0.05, 0.50]$ and net-benefit margin 0.01; Step 5 reference directions and agreement share 0.70

OUTPUT: For each target, the Step 1 flag (pass or warning), the value and flag of every step, and the verdict (passes or does not pass the five-step validation)

1: Stage A (partition). Split D by stratified sampling into training, validation, and test folds. All preprocessing parameters, model parameters, Step 1 measures, and score thresholds are estimated on the training fold; Platt-scaling parameters are fitted on the validation fold; the test fold is reserved for evaluation.

2: Protocol Step 1 (dependence screen). On the training fold, compute the Pearson correlation and the mutual information between every feature and t. If $\max_j |r_j| \leq \tau_r$ AND $\max_j \mathrm{MI}_j \leq \tau_{\mathrm{MI}}$, record a screening warning for t. The warning does not terminate the procedure: Steps 2 to 5 are still executed, so that it is confirmed or overturned by model-based evidence.

3: Stage B (preparation). Standardise continuous features for logistic regression on the training fold and, for binary targets, apply SMOTE to the training fold only.

4: Stage C (training). Train the five classifiers and Platt-scale LightGBM on the validation fold (LGBM-Cal).

5: Protocol Steps 2 to 5 (validation on the binary decision contrast, test fold). Step 2: require $\mathrm{BS} < \mathrm{UNC}$ for LGBM-Cal. Step 3: require the net benefit of LGBM-Cal to exceed both default strategies by at least 0.01 at some $\mathrm{p_t} \in [0.05, 0.50]$. Step 4: require the feature with the largest mean $|\mathrm{SHAP}|$ to satisfy the Step 1 rule. Step 5: require at least 70% of the scored reference directions to be reproduced by the logistic regression coefficient signs, and report the integer score with its rounding bound as a supporting diagnostic.

6: Stage D (verdict). If all five steps pass, report t as passing the five-step validation and eligible for further validation; otherwise report t as not passing, together with the failing steps and any Step 1 warning.

Figure 1 and Algorithm 1 provide complementary descriptions of the protocol: the figure conveys the overall decision workflow, whereas the algorithm specifies the sequence of operations and the criteria used to reach the final verdict. The protocol comprises five validation steps, with Steps 2–5 supported by the formal results in Section 2 and Step 1 operationalised by the dependence-screening rule in Equation (6). Partitioning precedes screening and model fitting to prevent test-set information from influencing the quantities used in the final assessment. Because pairwise dependence measures may miss relationships arising through feature interactions, Step 1 is deliberately used as a preliminary screen rather than a definitive test of learnability (Section 4.9). A warning therefore identifies limited measured dependence under the prespecified screening criteria; it does not establish that prediction is impossible.

The five steps address distinct questions. Step 1 assesses whether any feature exhibits sufficient measured association with the target to justify proceeding without an initial warning. Step 2 evaluates whether the predicted probabilities improve upon the prevalence-based uncertainty baseline, an aspect of model quality that discrimination does not establish. Step 3 assesses whether using the model offers greater net benefit than the default treat-all and treat-none strategies over a prespecified range of decision thresholds. Step 4 examines whether the feature receiving the largest SHAP attribution also satisfies the measured dependence criterion, helping identify potentially implausible attributions when a model relies on features with little observed association with the outcome. Step 5 evaluates whether the signs of fitted Logistic Regression coefficients agree with directions supported by the clinical literature, intro-

ducing external clinical evidence that cannot be obtained from predictive metrics alone. Together, these steps provide complementary evidence for deciding whether a target and its fitted model warrant further investigation.

Steps 2, 3, and 5 require a binary decision contrast. For the two three-class targets, the protocol therefore uses clinically relevant binary contrasts while retaining the full multiclass analysis for discrimination reporting. For diagnosis, the contrast is bacterial versus non-bacterial meningitis, corresponding to the decision to initiate antibiotics. For risk level, it is high risk versus not high risk, corresponding to escalation of care. Mortality is evaluated as deceased versus recovered. In every binary analysis, the positive class is the clinically actionable class: Bacterial, High risk, or Deceased, respectively.

The reference directions for Step 5 are determined using a prespecified literature-based rule. A feature direction is scored only when at least two independent published sources support it, either as an independent predictor in a multivariable model, a component of a validated decision rule, or a consistent difference between bacterial and viral meningitis. Features without sufficient support are left unscored rather than assigned an assumed direction. Because diagnosis and risk stratification concern decisions made at presentation, they share a set of presentation-time reference directions informed by the Bacterial Meningitis Score [14]. Bacterial meningitis and high triage risk are associated with higher CSF white blood cell counts [13], [14], [54], higher CSF protein [13], [14], [54], lower CSF glucose [13], [54], a higher peripheral white blood cell count [14], [54] and a higher C-reactive protein [55], [56].

Mortality is assessed against prognostic evidence. Death or unfavourable outcomes have been associated with older age [2], [15], [57]–[59], lower CSF white blood cell counts [2], [57], [58], lower CSF glucose [58], [59], microbiological confirmation, reported as a positive culture [2], [57], and higher CRP [15], [57], [59]. Five directions are scored for each contrast (Section 4.6). Step 5 requires at least 70% agreement between the fitted coefficient signs and the scored reference directions. With five scored directions, this criterion requires at least four agreements out of five (80%), because three agreements correspond to only 60%. A sensitivity analysis additionally considers directions reported by a single source: pathogen presence for the two presentation-time contrasts [14], and CSF protein [58], platelet count [2], peripheral white blood cell count [59], and male sex [58] for mortality.

The pass criteria for Steps 2–4 are also prespecified. Step 2 passes when the Brier score of LGBM-Cal is below the uncertainty term in Equation (2), corresponding to a positive Brier skill score (Proposition 2). Step 3 passes when, at some threshold between 0.05 and 0.50, the model's net benefit exceeds both default strategies by at least 0.01 (Theorem 3). This threshold range represents a range of decision preferences, while the net-benefit margin corresponds to one additional net true positive per 100 patients. Step 4 passes when the feature with the largest mean absolute SHAP value satisfies the Step 1 dependence rule in Equation (6), as motivated by Remark 1. Steps 2–4 use LGBM-Cal as the reference model, while Step 5 uses Logistic Regression because its fitted coefficient signs can be directly audited against the literature. The integer risk score and its rounding bound are reported as supporting diagnostics rather than as an additional validation step. The final verdict is based on the five prespecified pass/fail criteria and must be interpreted in the context of the dataset and the absence of external validation.

3.2. Dataset and Pre-Modelling Learnability Check

The dataset comprises 1,200 meningitis patient records with 14 columns and no missing values, obtained from a publicly available repository [60]. It contains a patient identifier, ten clinical predictors, and three prediction targets. The predictors are Age, Gender, cerebrospinal fluid (CSF) white blood cell count (WBC Count), Protein Level, Glucose Level, Pathogen Present, Haemoglobin, peripheral white blood cell count (WBC Blood Count), Platelets, and C-reactive protein level (CRP Level). Diagnosis is a three-class aetiological target, comprising Bacterial (49.6%), Viral (44.8%), and Unknown (5.6%). Outcome is a binary mortality target, comprising Recovered (89.7%) and Deceased (10.3%). Risk Level is a three-class triage target, comprising High (49.4%), Low (45.0%), and Moderate (5.6%).

The dataset was selected for three reasons. First, it provides three targets with different clinical meanings and potentially different levels of learnability within the same feature set, allowing the proposed protocol to be evaluated without changing the predictors. Second, its predictors represent routine haematology and CSF measurements. Third, its public availability

supports reproducibility. However, the dataset has a synthetic origin, and the implications of this limitation are discussed in Section 4.9.

Step 1 evaluates feature–target dependence on the training fold before any classifier is trained. Learnability is operationalised as a screening criterion rather than a definitive measure of predictive performance: a target passes if at least one feature exceeds either the Pearson correlation threshold or the mutual information threshold, as specified in Eq. (6). Here r_j denotes the Pearson correlation between feature j and the target, MI_j denotes their mutual information, and $\tau_r = 0.10$ and $\tau_M I = 0.05$ nats are the respective thresholds.

$$\max_j |r_j| > \tau_r \quad \text{or} \quad \max_j \mathrm{MI}_j > \tau_{\mathrm{MI}} \tag{6}$$

The use of an OR condition is deliberate: a target is admitted to subsequent evaluation if either measure indicates sufficient dependence, reducing the risk of prematurely excluding a potentially learnable target. A false admission means that the target proceeds to the remaining four steps, whereas failure to meet the screening criterion raises a warning rather than terminating the analysis.

Pearson correlation is calculated using Eq. (7) to measure linear association between paired observations, where x_i and y_i denote the values of the feature and target, respectively, and $\bar{x}$ and $\bar{y}$ are their means.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{7}$$

Spearman rank correlation is also calculated as a robustness check because it assesses monotonic association using ranks. Its coefficient is expressed in Eq. (8), where d_i is the difference between the ranks of the i-th observation pair, and n is the sample size.

$$\rho_s = 1 - \frac{6 \sum_i {d_i}^2}{n(n^2 - 1)} \tag{8}$$

Equation (8) applies exactly when there are no tied ranks. When ties occur, Spearman correlation is calculated as the Pearson correlation between the average ranks. Although Pearson and Spearman correlations capture different forms of association, neither alone can identify every form of dependence. Mutual information is therefore computed for each feature–target pair on the same training fold. It is non-negative, equals zero if and only if the variables are statistically independent in the population, and can capture nonlinear univariate dependence that correlation measures may not detect. In the proposed rule, Pearson correlation and mutual information determine whether a target passes the screen, while Spearman correlation provides complementary evidence of robustness.

Figure 2 presents the Pearson correlation matrix computed from the 840 training-fold records, with the targets label encoded. The mortality target exhibits markedly weaker correlations than the other two targets: its largest absolute correlation is 0.067, followed by 0.06 for Protein Level and 0.05 for CRP Level. Diagnosis shows stronger associations, with absolute correlations of 0.84 for Pathogen Present, 0.78 for Haemoglobin, 0.76 for WBC Blood Count, and 0.69 for Platelets. Risk Level occupies an intermediate position, with correlations of 0.51 for Pathogen Present and 0.45 for Haemoglobin.

The mutual information results support this distinction, with values of 0.564 for Diagnosis, 0.482 for Risk Level, and 0.013 for mortality. The Spearman results are consistent with the same pattern, with maximum absolute coefficients of 0.840, 0.570, and 0.065 for Diagnosis, Risk Level, and mortality, respectively. Applying Eq. (6), Diagnosis and Risk Level pass the pre-modelling screen, whereas mortality receives a Step 1 warning. This warning does not establish that mortality is inherently unpredictable; rather, it indicates that the available training-fold data provide insufficient evidence of feature–target dependence under the specified screening thresholds. The subsequent steps therefore assess whether model-based evidence provides additional support for the mortality target.

3.3. Preprocessing and Class Balancing

Gender and Pathogen Present were binary-encoded, with male and present coded as 1. Continuous features were standardised to zero mean and unit variance for Logistic Regression, whereas tree-based models were trained without scaling. The dataset was partitioned by

stratified random sampling into training (70%, n=840), validation (15%, n=180), and test (15%, n = 180) subsets, stratified on the mortality label. Consequently, the class proportions of the two three-class targets were preserved only approximately in the validation and test folds. Standardisation was fitted using the training-fold statistics and applied unchanged to the validation and test folds. As defined in Eq. (9), each continuous feature x_j is transformed using its training-fold mean μ_j and standard deviation σ_j.

$$x_j' = \frac{x_j - \mu_j}{\sigma_j} \tag{9}$$

Class imbalance was addressed according to the target and model type. SMOTE with k=5 nearest neighbours was applied only to the training data, in accordance with Proposition 5. It generated synthetic minority samples through convex interpolation, as described in Eq. (5), for each of the three binary decision contrasts used in Steps 2–5: bacterial versus non-bacterial, high risk versus not high risk, and deceased versus recovered. The last contrast also defines the mortality target used in the cross-validation analysis in Table 1. Because all features were recorded as integers, the synthetic values were rounded down to integers. No resampling was applied to the three-class targets. Instead, Logistic Regression and Random Forest used balanced class weights, while the boosted models were trained without class weighting. This strategy is consistent with studies that restrict resampling to training folds in clinical and longitudinal settings [61], and with research showing that feature-level design can reduce reliance on oversampling [62].

For the multiclass targets, the balanced class weights used by Logistic Regression and Random Forest were calculated according to Eq. (10), where n denotes the total number of training samples, K the number of classes, and n_c the count of class c. This weighting scheme assigns greater weight to less frequent classes.

$$w_c = \frac{n}{K \cdot n_c} \tag{10}$$

The preprocessing and modelling workflow was designed to prevent data leakage. The dataset was partitioned before any data-dependent preprocessing or model fitting. Standardisation parameters, SMOTE synthesis, model parameters, Step 1 dependence measures, and integer-score thresholds were estimated using the training fold only, then applied or used as specified for the validation and test folds. The validation fold was used only to fit the Platt scaling procedure in Eq. (20), whereas the test fold was reserved for final evaluation. Thus, test-fold labels were not used to determine the Step 1 screening results or fit preprocessing and calibration parameters.

3.4. Classifier Configuration

Five classifiers were trained for each target, with the same settings in every analysis. Logistic Regression used L2 regularisation with $C = 1.0$ and balanced class weights. Random Forest [51] used 300 trees, a minimum of two samples per leaf, and balanced class weights. LightGBM [48] used leaf-wise tree growth with 31 leaves, learning rate 0.05, and 500 estimators. CatBoost [49] used depth 6, learning rate 0.03, and 1000 iterations. XGBoost [50] used depth 6, a learning rate of 0.05, and 500 estimators. No early stopping was used, ensuring consistent fitting within each cross-validation fold. LightGBM was additionally calibrated using Platt scaling [63] on the validation fold (LGBM-Cal) and served as the reference model for Steps 2–4 and as the representative model whenever a single fitted model was required. This choice was based on convenience rather than superior performance, as the tree-based models differed by less than one-fold-to-fold standard deviation in Table 1. All implementations used scikit-learn [64] and the respective library APIs.

Each classifier served a distinct purpose: Logistic Regression provided signed coefficients for the Step 5 audit; Random Forest enabled comparison with a non-boosted ensemble; and LightGBM, CatBoost, and XGBoost represented different boosting implementations and growth strategies, reducing the risk of attributing findings to a single implementation. Given the dataset's 1,200 records and ten tabular predictors, no deep learning architecture was included.

3.4.1. Logistic Regression Baseline

The transparent baseline models the conditional class probability as in Eq. (11) with the logistic link of Eq. (12), where β_0 is the intercept and β_j the coefficient of feature j. Its coefficients are directly auditable against the directions reported in the clinical literature, which is what makes Step 5 of the protocol possible.

$$P(Y = 1 \mid x) = \sigma\left(\beta_0 + \sum_j \beta_j x_j\right) \tag{11}$$

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{12}$$

Coefficients are obtained by minimising the L2-penalised negative log-likelihood in Eq. (13), where $y_i \in -1, +1$ denotes the class label in this equation only, C is the inverse regularisation strength, and $\|\beta\|$ is the Euclidean norm of the coefficient vector. Where balanced class weights are used, each loss term is multiplied by the weight of Eq. (10) for the class of observation i.

$$\frac{1}{2}\|\beta\|^2 + C\sum_i \log\left(1 + e^{-y_i \beta^\top x_i}\right) \tag{13}$$

3.4.2. Gradient Boosting Formulation

All three boosted ensembles build an additive model as in Eq. (14), where F_m is the ensemble after m rounds, h_m the tree added at round m, and η the learning rate. The three implementations differ in how h_m is selected.

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \tag{14}$$

XGBoost minimises the second-order Taylor expansion of the loss given in Eq. (15), where g_i and H_i are the first and second derivatives of the loss at observation i, together with the complexity penalty of Eq. (16), in which T is the number of leaves, w_j the weight of leaf j, and γ and λ control leaf count and leaf magnitude, respectively. The closed-form optimal leaf weight follows in Eq. (17), where I_j is the set of observations assigned to leaf j.

$$\mathcal{L}^{(m)} \simeq \sum_i \left[g_i\, h_m(x_i) + \frac{1}{2} H_i\, h_m^2(x_i)\right] + \Omega(h_m) \tag{15}$$

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda \sum_j w_j^2 \tag{16}$$

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} H_i + \lambda} \tag{17}$$

LightGBM grows trees leaf-wise, expanding at each round the leaf whose split maximises the gain in Eq. (18), where S and H denote the summed gradients and Hessians of the left, right, and parent partitions. Leaf-wise growth reduces loss faster per node than level-wise growth on tabular data.

$$G = \frac{1}{2}\left[\frac{S_{\mathrm{L}}{}^2}{H_{\mathrm{L}} + \lambda} + \frac{S_{\mathrm{R}}{}^2}{H_{\mathrm{R}} + \lambda} - \frac{S_{\mathrm{P}}{}^2}{H_{\mathrm{P}} + \lambda}\right] - \gamma \tag{18}$$

CatBoost replaces categorical levels with the ordered target statistic of Eq. (19), computed only from observations preceding i in a random permutation, with prior P and smoothing weight a. This ordering removes the target leakage that afflicts conventional target encoding. In this study, all predictors are numeric or binary-encoded (Section 3.3), so no categorical levels are encoded, and the statistic is given for completeness.

$$x_i' = \frac{\sum_{k<i,\ x_k=x_i} y_k + aP}{\sum_{k<i,\ x_k=x_i} 1 + a} \tag{19}$$

For calibration assessment, the LightGBM score is mapped to a probability by Platt scaling as in Eq. (20), where A and B are fitted on the validation fold only.

$$p^{\text{cal}}(x) = \frac{1}{1 + e^{Af(x)+B}} \tag{20}$$

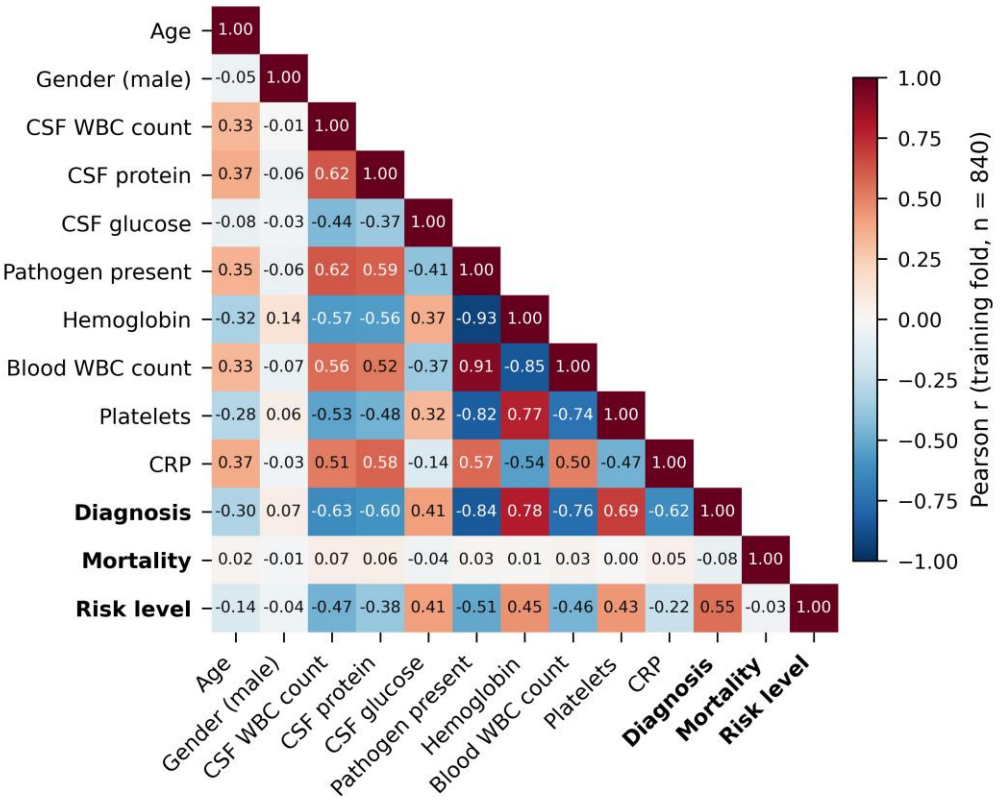


Figure 2. Pearson correlation matrix of the ten predictors and the three targets, computed on the training fold only (n = 840). The targets, in bold, are label-encoded; colour encodes the sign and strength of each correlation. Estimating the matrix on the training fold prevents test-fold labels from influencing the Step 1 screen

3.5. Evaluation Protocol

The proposed five-step protocol was applied to all three prediction targets using complementary measures of discrimination, calibration, clinical utility, and interpretability. Discrimination was assessed using the area under the receiver operating characteristic curve (AUC), macro-averaged F1-score, and Matthews correlation coefficient (MCC) [65]. MCC was selected as the primary balanced classification measure because it incorporates all four confusion-matrix components and is informative under class imbalance, whereas F1-score does not account for true negatives. Accuracy was not reported because, at a mortality event rate of 10.3%, predicting survival for every patient would yield 89.7% accuracy without identifying any deaths.

Precision and recall are calculated from the confusion matrix as defined in Eq. (21), and their harmonic mean gives the F1-score in Eq. (22). The Matthews correlation coefficient for binary classification is defined in Eq. (23). For the three-class targets, macro-averaged F1-score and the multiclass generalisation of MCC were reported. AUC was macro-averaged across the K classes according to Eq. (24).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{21}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{23}$$

$$\text{AUC}_{\text{macro}} = \frac{1}{K}\sum_{c} \text{AUC}_c \tag{24}$$

Calibration was evaluated through Brier score decomposition [31] and reliability diagrams [66]. with eight equal-frequency bins. The decomposition, introduced in Eq. (2), separates the Brier score into reliability, resolution, and uncertainty. Reliability measures the discrepancy between predicted probabilities and observed outcome frequencies, whereas resolution reflects the extent to which predictions distinguish groups with different outcome rates. Uncertainty depends only on the observed event rate. Equations (25) and (26) define reliability, resolution, and uncertainty, respectively, where n is the total number of observations, n_k is the number of observations in bin k, k indexes the bins, and the barred symbols denote the corresponding bin means. Because the forecasts are grouped into bins, the decomposition identity in Eq. (2) holds approximately for the tabulated values; any residual reflects variation in forecasts within each bin.

$$\text{REL} = \frac{1}{n}\sum_{k} n_k\left(\bar{f}_k - \bar{y}_k\right)^2 \tag{25}$$

$$\text{RES} = \frac{1}{n}\sum_{k} n_k(\bar{y}_k - \bar{y})^2, \quad \text{UNC} = \bar{y}(1-\bar{y}) \tag{26}$$

The Brier score was preferred to log-loss because it is bounded and can be decomposed into components that distinguish calibration from risk differentiation. This distinction is central to Step 2, which requires the reference model's Brier score to be lower than the uncertainty baseline. Clinical utility was assessed using decision curve analysis (DCA) [32]. Curves were plotted over threshold probabilities from 0 to 0.80, while Step 3 was evaluated over the prespecified range of 0.05–0.50. DCA complements discrimination and calibration metrics by assessing whether model-guided decisions provide greater net benefit than the default strategies across clinically relevant thresholds. Threshold-based risk stratification with feature-importance weighting has also been investigated in a comparable clinical setting [67].

Interpretability was assessed using SHAP TreeExplainer [26]. Global mean absolute SHAP values were used to summarise feature contributions, while local attributions were examined for two cases selected according to a fixed rule. These analyses supported the Step 4 assessment of whether the most influential feature was consistent with the pre-modelling dependence screen. Finally, Step 5 assessed agreement between fitted coefficient directions and published clinical evidence. Together, these measures provided complementary evidence across the five steps, rather than relying on discrimination alone to determine whether a prediction target warranted further validation.

4. Results and Discussion

4.1. Discrimination under Repeated Cross-Validation

Discrimination is the primary performance analysis. Table 1 reports it under repeated stratified cross-validation, five folds repeated five times (25 fits per model and target), and Fig. 3 plots the same results; all other analyses require a single fitted model and use the prespecified test fold. On the two learnable targets, the fold-to-fold standard deviation of AUC is 0.007 to 0.008 for the tree-based models (0.021 to 0.023 for Logistic Regression), and the four tree-based models lie within 0.003 of one another. On mortality, mean AUC lies between 0.522 and 0.541 and mean MCC between 0.007 and 0.035, with standard deviations of 0.042 to 0.068 and 0.055 to 0.065.

For diagnosis, the four tree-based models are separated by less than 0.003 in mean AUC and by 0.013 in mean MCC; for risk level, by 0.003 and 0.011. These differences are smaller

than one fold-to-fold standard deviation, so no single model can be declared best, and every single-fold analysis below uses LightGBM as a representative model rather than as a winner. What separates the two learnable targets from mortality is not the choice of algorithm but the presence of signal. Logistic Regression trails by 0.06 in AUC on diagnosis and 0.10 on risk level, the expected cost of a linear decision boundary on tabular clinical data, and it is retained because Step 5 requires its auditable coefficients.

For mortality, every mean MCC lies within about two-thirds of a standard deviation of zero, so no model is distinguishable from chance. The models are uninformative on this target, neither predictive nor reliably anti-predictive, which is what one expects when the features carry no detectable association with the outcome under the dependence measures of Step 1. The largest mean AUC, 0.541 ± 0.042 for Logistic Regression, is nevertheless above 0.5, so if viability were judged only by whether mean discrimination exceeded chance, mortality could appear marginally viable (Section 4.7).

Table 1. Discrimination under repeated stratified cross-validation, five folds and five repeats (25 fits per model per target). Values are mean ± standard deviation across folds. AUC is macro-averaged one-vs-rest for the two three-class targets

Model	Diagnosis AUC	Diagnosis MCC	Mortality AUC	Mortality MCC	Risk Level AUC	Risk Level MCC
Logistic Regression	0.907 ± 0.021	0.739 ± 0.028	0.541 ± 0.042	0.035 ± 0.056	0.867 ± 0.023	0.717 ± 0.036
Random Forest	0.969 ± 0.007	0.790 ± 0.027	0.523 ± 0.058	0.007 ± 0.065	0.972 ± 0.007	0.802 ± 0.032
LightGBM	0.970 ± 0.007	0.793 ± 0.025	0.525 ± 0.062	0.013 ± 0.064	0.969 ± 0.007	0.794 ± 0.028
CatBoost	0.970 ± 0.007	0.780 ± 0.029	0.522 ± 0.055	0.010 ± 0.060	0.971 ± 0.007	0.791 ± 0.035
XGBoost	0.971 ± 0.007	0.787 ± 0.028	0.522 ± 0.068	0.009 ± 0.055	0.971 ± 0.008	0.801 ± 0.034

AUC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient.

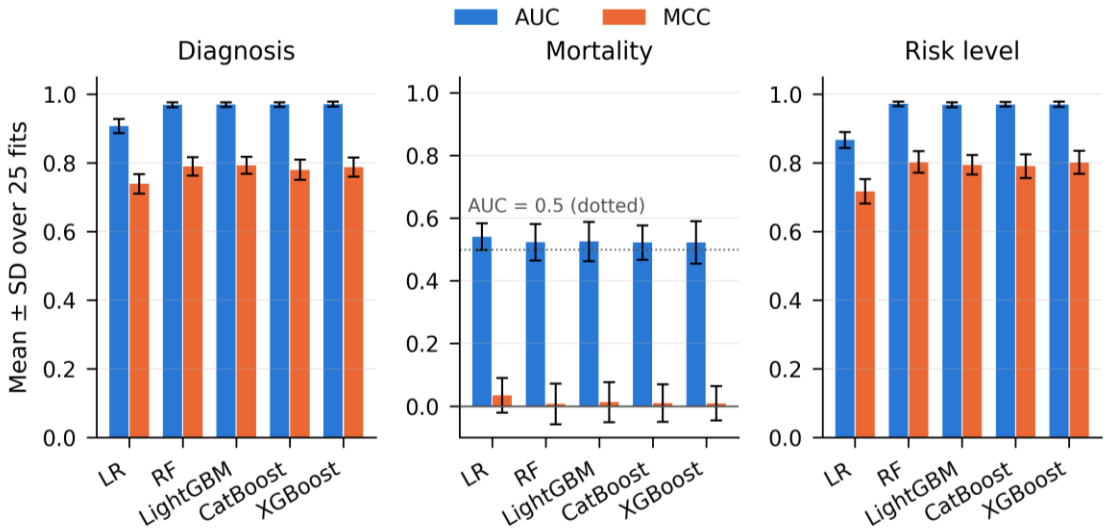


Figure 3. Discrimination under repeated stratified cross-validation, five folds and five repeats. Bars are mean AUC and MCC, and error bars are one standard deviation across the 25 fits; the dotted line in the mortality panel marks an AUC of 0.5

4.2. Class-wise Errors of the Representative Model

Macro-averaged figures hide where the errors fall, so Table 2 reports precision, recall, F1-score, and support for every class on the test fold, with LightGBM as the representative model. On diagnosis, the majority classes are recovered well (F1 0.933 for bacterial and 0.912 for viral), but the Unknown class collapses to 0.125 on eleven cases. On risk level, high and low risk reach 0.929 and 0.910, while moderate risk reaches 0.316 on ten cases. On mortality, the deceased class reaches 0.049 on nineteen cases. The macro-F1 values of the learnable targets are therefore carried by the majority classes, and any clinical application would need substantially more minority-class data, a limitation of the dataset rather than of the protocol.

Table 2. Class-wise precision, recall, F1-score and support on the held-out test fold for the representative model, LightGBM, on each target. LightGBM is the model carried through Steps 2 to 4; it is not selected on performance, since the tree-based models differ by less than one fold-to-fold standard deviation in Table 1

Target	Class	Precision	Recall	F1	Support
Diagnosis	Bacterial	0.898	0.970	0.933	100
	Unknown	0.200	0.091	0.125	11
	Viral	0.925	0.899	0.912	69
Mortality	Deceased	0.045	0.053	0.049	19
	Recovered	0.886	0.870	0.878	161
Risk level	High Risk	0.948	0.910	0.929	100
	Low Risk	0.880	0.943	0.910	70
	Moderate Risk	0.333	0.300	0.316	10

Support is the number of test-fold cases in each class. Minority-class performance is reported separately because macro-averaging conceals it.

Fig. 4 shows the corresponding confusion matrices and localises the errors that Table 2 summarises. On diagnosis, seven of eleven Unknown cases are absorbed into the Bacterial class. On risk level, the ten Moderate risk cases split across the three predicted categories, four as high, three as low and three as moderate, which is the behaviour of a model with little usable information about that class. On mortality, 18 of 19 deceased patients are predicted to recover, and the single correct positive is accompanied by 21 false alarms.

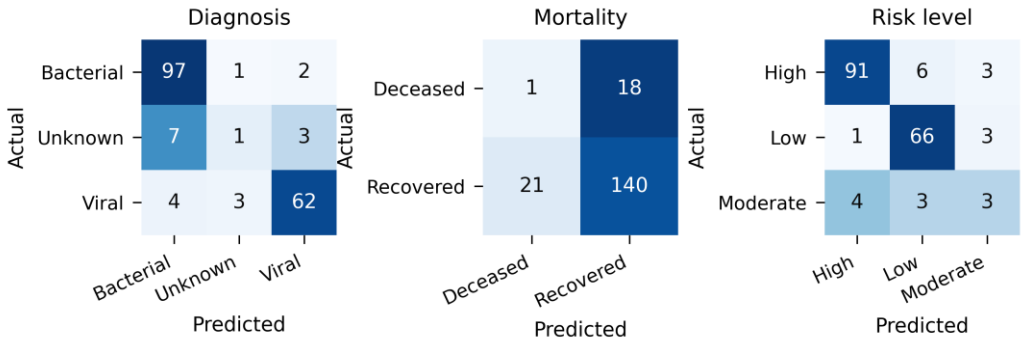


Figure 4. Confusion matrices on the held-out test fold for the representative model, LightGBM, on each target. Rows are actual classes and columns predicted classes; shading is the row proportion. The minority classes, Unknown diagnosis and Moderate risk, are where the errors concentrate

4.3. Calibration Analysis (Step 2)

Figure 5 and Table 3 present the reliability diagrams and Brier score decomposition for mortality, with Deceased as the positive class. The uncertainty baseline (UNC) is 0.0944, determined by the test-fold event rate of 19 deaths among 180 patients. All five raw models yield Brier scores above UNC, ranging from 0.1493 to 0.2322, with low resolution values of 0.0005–0.0046. These results indicate limited risk differentiation and poor probabilistic performance relative to the prevalence baseline. Their reliability components (REL = 0.0533–0.1398) also indicate substantial calibration error, consistent with the probability inflation that may result from training on the SMOTE-balanced fold.

Among the evaluated models, the Platt-scaled reference model, LGBM-Cal, achieves the lowest Brier score of 0.0949, approaching but not falling below UNC. As shown in Figure 5, its predicted probabilities lie within a narrow range of approximately 0.10–0.13 around the observed event rate. This compression brings its Brier score close to the prevalence baseline but provides little evidence of meaningful risk stratification. Therefore, LGBM-Cal does not satisfy the Step 2 criterion, $BS < UNC$, and mortality fails the calibration assessment. Its resolution is only 0.0046 (Table 3), so by Proposition 2 the model cannot separate patients by risk.

In contrast, LGBM-Cal achieves Brier skill scores of 0.744 for diagnosis and 0.745 for risk level, satisfying Step 2 for both targets (Table 5). The results therefore distinguish mortality from the other two targets: although calibration brings the mortality model close to the prevalence baseline, it does not establish predictive improvement over that baseline.

Table 3. Brier score decomposition for mortality on the held-out test fold (n = 180, 19 deaths).

Model	BS	REL	RES	UNC	BS − UNC
LR	0.2322	0.1398	0.0024	0.0944	+0.1378
RF	0.1499	0.0533	0.0005	0.0944	+0.0555
LightGBM	0.1762	0.0822	0.0046	0.0944	+0.0818
CatBoost	0.1493	0.0574	0.0041	0.0944	+0.0549
XGBoost	0.1673	0.0724	0.0020	0.0944	+0.0729
LGBM-Cal	0.0949	0.0051	0.0046	0.0944	+0.0004

Note: Deceased is the positive class, and UNC = 0.0944. A positive BS-UNC indicates a higher Brier score than the prevalence baseline.

BS, Brier score; REL, reliability; RES, resolution; UNC, uncertainty. The five raw models are trained on the SMOTE-balanced training fold. LGBM-Cal is LightGBM after Platt scaling on the validation fold and is the reference model for Steps 2 to 4. Because REL, RES and UNC are computed from eight equal-frequency bins, BS − (REL − RES + UNC) is not exactly zero for the tabulated values; the residual reflects variation of the forecasts within bins.

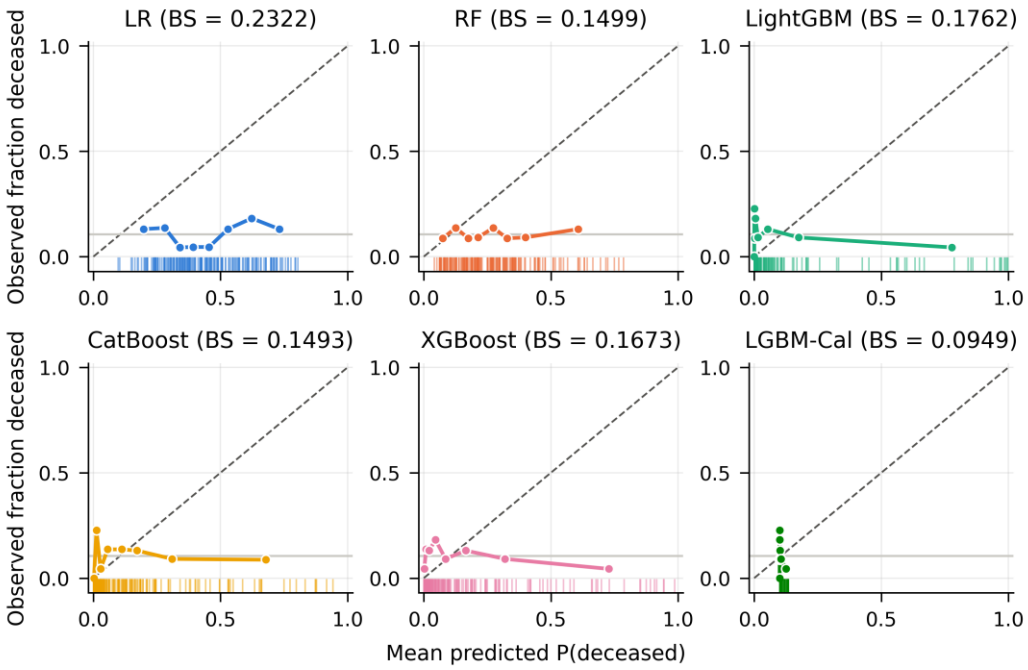


Figure 5. Reliability diagrams for mortality on the held-out test fold (n = 180, 19 deaths).

The vertical axis shows the observed fraction of deaths in each of eight equal-frequency bins. The grey horizontal line indicates the event rate, and the ticks along the base show individual predicted probabilities. LGBM-Cal denotes LightGBM after Platt scaling on the validation fold.

4.4. Decision Curve Analysis (Step 3)

Fig. 6 shows the decision curves for mortality; the shaded band is the prespecified Step 3 range. Within it, LGBM-Cal never exceeds the better default strategy: below the event rate of about 0.10 it treats every patient and coincides with treat-all, and above it treats no one and coincides with treat-none, so Step 3 fails. The raw models do no better in any meaningful sense: their largest excess is 0.0012, less than one patient in the 180-patient test fold

(1/180=0.0056), and at higher thresholds their net benefit falls as low as −0.10 to −0.28, the harm that acting on them would cause through false positives. For diagnosis and risk level, LGBM-Cal exceeds the default strategies by 0.367 and 0.372, and both pass Step 3 (Table 5).

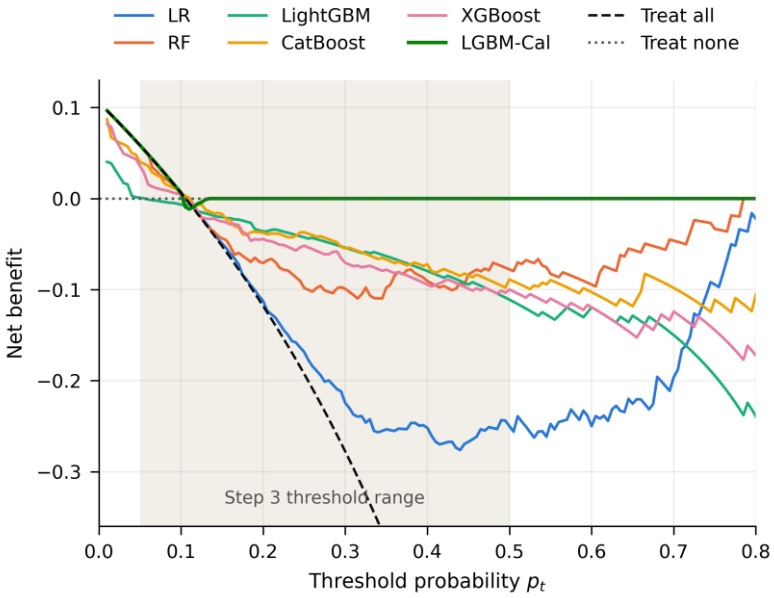


Figure 6. Decision curves for mortality on the held-out test fold, with Deceased as the positive class. The dashed line is treat-all, and the dotted line is treat-none; the shaded band is the prespecified Step 3 threshold range of 0.05 to 0.50

4.5. SHAP Interpretability Analysis (Step 4)

Fig. 7 presents the global mean absolute SHAP attribution of the LightGBM model fitted to each binary decision contrast, the same model that is Platt-scaled for Steps 2 and 3. For diagnosis, CRP dominates, followed by CSF white blood cell count and CSF glucose, a ranking consistent with the established CSF and inflammatory markers of bacterial aetiology [13], [14], [55]. CRP has a training-fold correlation of 0.560 and a mutual information of 0.499 with the bacterial contrast, so it satisfies the Step 1 rule, and diagnosis passes Step 4. For risk level, CSF glucose leads, followed by CSF white blood cell count and CSF protein; CSF glucose has a correlation of 0.431 and a mutual information of 0.437, and risk level passes Step 4.

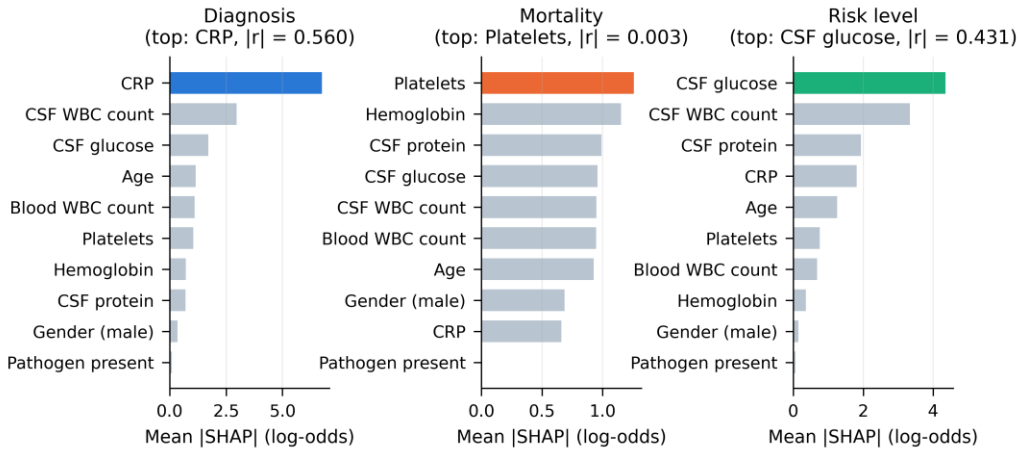


Figure 7. Global mean absolute SHAP attribution of the LightGBM model fitted to each binary decision contrast, computed on the held-out test fold. The highlighted bar is the leading feature tested in Step 4, shown with its absolute training-fold correlation with the contrast. Scales differ between panels.

For mortality, the pattern inverts. Platelet count ranks highest, followed by haemoglobin and CSF protein, and the attributions are spread almost evenly across nine features, while pathogen presence, the one variable a clinician would read as microbiological evidence, receives none. The leading feature has an absolute training-fold correlation with mortality of 0.003 and a mutual information of zero, so it fails the Step 1 rule, and Step 4 fails. By Remark 2, large attributions on such a variable describe the model's internal weighting, not an association with death.

Fig. 8 presents local attributions for two deceased patients in the test fold, selected by a fixed rule: the one with the highest and the one with the lowest predicted probability of death. Case A receives a predicted probability of 0.97, carried mainly by its platelet count, CSF glucose, and CSF white blood cell count, none of which are in the range the prognostic literature associates with death [2], [58], [59]. The prediction is correct, but the reasoning is clinically inverted. Case B receives a probability below 0.01, and the largest attribution pushing it toward recovery is an age of 74 years, the most consistently reported risk factor for death in bacterial meningitis [2], [57]–[59]. Together, the two cases illustrate two failure modes that SHAP can expose: a correct prediction reached through implausible reasoning, and a missed death driven by an attribution that contradicts the literature.

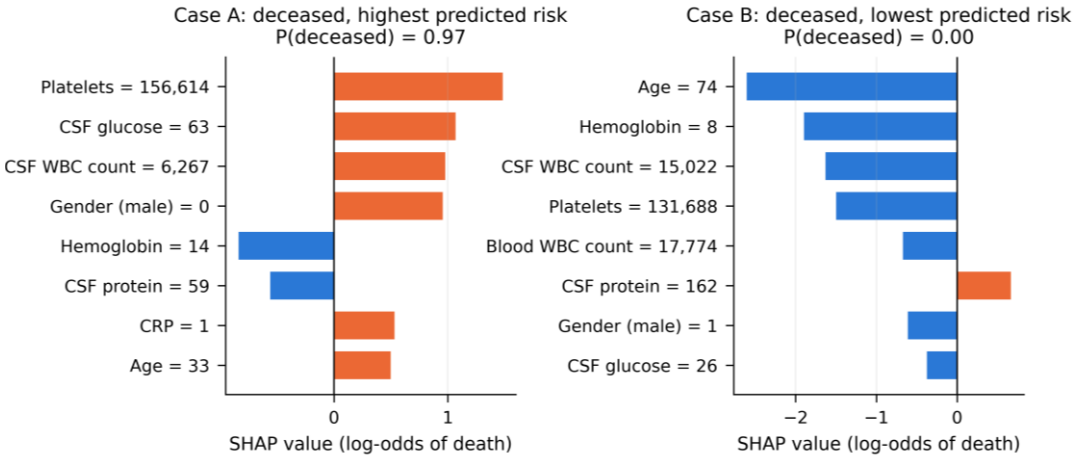


Figure 8. Local SHAP attributions for two deceased patients in the held-out test fold: Case A, the highest predicted probability of death, and Case B, the lowest. Bars show the eight largest contributions to the log-odds of death; feature values are in the dataset's units

4.6. Coefficient Audit and Integer Risk Score (Step 5)

Table 4 sets the fitted logistic regression coefficients against the literature reference of Section 3.1. For diagnosis, four of the five scored coefficient directions agree with the literature references, with the only disagreement occurring for CSF protein, whose coefficient is effectively zero (−0.001). Risk level also achieves four of five agreements, with CRP (−0.093) as the sole exception. Both targets therefore satisfy the Step 5 criterion of at least 70% agreement. In contrast, mortality achieves only two of five agreements, including a near-zero coefficient for age (+0.015). The remaining three features—CSF glucose, pathogen presence, and CRP—show directions opposite to those reported in the prognostic literature. Notably, the negative coefficients for pathogen presence and CRP imply associations with lower predicted mortality, contrary to the reference evidence. Consequently, mortality fails Step 5. A sensitivity analysis that additionally includes single-source reference directions yields agreement rates of 5/6 for diagnosis and risk level and 5/9 for mortality, leaving all three verdicts unchanged. Under the primary reference, these verdicts remain unchanged for agreement thresholds above 40% and up to 80%; under the sensitivity analysis, they remain unchanged for thresholds above 56% and up to 83%.

Fig. 9 presents the integer risk score derived from the mortality audit regression with $K = 5$, continuous features dichotomised at their training-fold 75th percentiles and binary features entering when present. The score ranges from −8 to 1 on the test fold, and the two outcome groups overlap across that range. It attains an AUC of 0.484, against 0.529 for the

underlying logistic regression probability and 0.472 for LightGBM on the same fold, all indistinguishable from chance and consistent with the cross-validated means of Table 1. The Youden-optimal cut-off, a score of −1 or higher, gives a sensitivity of 0.368 and a specificity of 0.795. By Proposition 4, rounding alone can move a predicted probability by up to δ_{round}=0.224, more than the spread of predicted risk between most patients, whose probabilities barely depart from the 10% event rate. The score therefore offers no usable ranking, which supports, but does not by itself determine, the Step 5 verdict.

Table 4. Step 5 sign audit: standardised logistic regression coefficients for each binary decision contrast against the reference directions reported in the literature

Feature	Presentation reference	Prognostic reference	Diagnosis	Mortality	Risk level
Age	—	↑ [2], [15], [57]–[59]	(−0.262)	+0.015 ✓	(−0.090)
Gender (male)	—	—	(−0.213)	(+0.120)	(+0.145)
CSF WBC count	↑ [13], [14], [54]	↓ [2], [57], [58]	+0.397 ✓	−0.195 ✓	+0.730 ✓
CSF protein	↑ [13], [14], [54]	—	−0.001 ✗	(−0.136)	+0.433 ✓
CSF glucose	↓ [13], [54]	↓ [58], [59]	−0.427 ✓	+0.062 ✗	−0.275 ✓
Pathogen present	—	↑ [2], [57]	(+1.552)	−0.684 ✗	(+1.026)
Haemoglobin	—	—	(−0.105)	(−0.897)	(+0.376)
Blood WBC count	↑ [14], [54]	—	+0.273 ✓	(−0.144)	+0.158 ✓
Platelets	—	—	(+0.030)	(−0.231)	(−0.157)
CRP	↑ [55], [56]	↑ [15], [57], [59]	+0.688 ✓	−0.156 ✗	−0.093 ✗
Agreement	Diagnosis, risk level	Mortality	4/5 Pass	2/5 Fail	4/5 Pass

Note: Coefficients are for standardised features. The positive classes are Bacterial, Deceased, and High risk. ↑ and ↓ indicate that higher or lower feature values, respectively, are reported to be associated with the positive class. A dash (—) indicates that fewer than two independent sources report a reference direction; these features are excluded from scoring, and their coefficients are shown in parentheses. ✓ and ✗ indicate agreement and disagreement, respectively, between the fitted coefficient sign and the reference direction. Pathogen presence is matched to positive culture in the prognostic literature. Step 5 passes when at least 70% of scored directions agree; with five scored directions, at least four agreements are required.

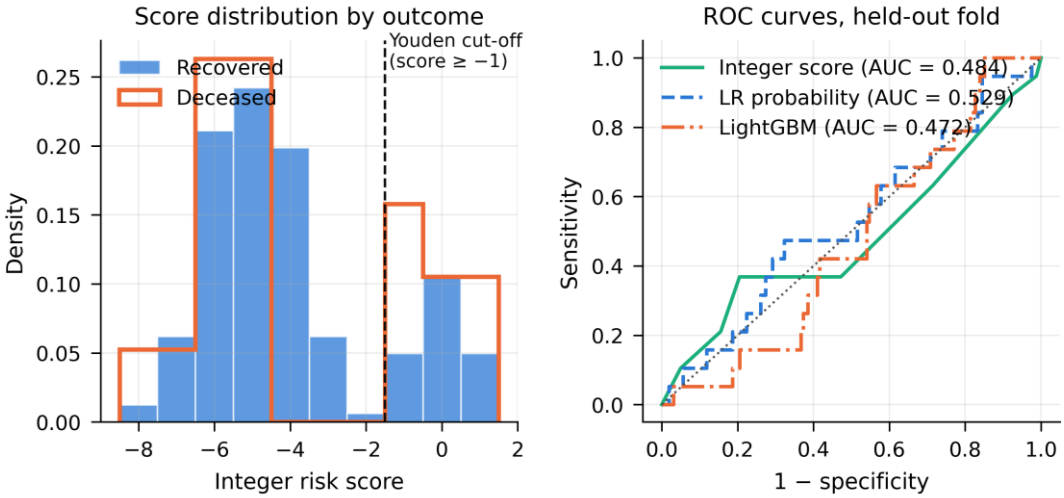


Figure 9. Integer risk score for mortality on the held-out test fold, with Deceased as the positive class: score distribution by outcome with the Youden-optimal cut-off (left), and ROC curves of the integer score, the underlying logistic regression probability and LightGBM (right)

4.7. Protocol Verdict and Ablation

Table 5 consolidates the verdict with the value that decided each step. Diagnosis and risk level pass all five steps and are eligible for further, external validation; mortality fails all five. The five failures are not restatements of one another: Step 1 measures the data before any model exists, Steps 2 and 3 the probabilities of a fitted model and the consequences of acting on them, Step 4 what that model attends to, and Step 5 a second, transparent model

against knowledge from outside the dataset. Their agreement is the substance of the verdict, and a single discrimination figure cannot deliver it.

Table 5. Five-step verdict for each target, with Steps 2 to 5 evaluated on the binary decision contrast of Section 3.1: bacterial versus non-bacterial, deceased versus recovered, and high risk versus not high risk. Each cell gives the value that decided the step.

Step	Criterion	Diagnosis	Mortality	Risk level
S1	$\max_j \lvert r_j \rvert > 0.10$ or $\max_j \mathrm{MI}_j > 0.05$	$\lvert r \rvert$ 0.840, MI 0.564: Pass	$\lvert r \rvert$ 0.067, MI 0.013: Fail (warning)	$\lvert r \rvert$ 0.511, MI 0.482: Pass
S2	$\mathrm{BS} < \mathrm{UNC}$ for LGBM-Cal	BSS +0.744: Pass	BSS −0.005: Fail	BSS +0.745: Pass
S3	Net benefit above both defaults by ≥ 0.01, $\mathrm{p_t} \in [0.05, 0.50]$	0.367: Pass	0.000: Fail	0.372: Pass
S4	Top SHAP feature passes the Step 1 rule	CRP ($\lvert r \rvert$ 0.560): Pass	Platelets ($\lvert r \rvert$ 0.003): Fail	CSF glucose ($\lvert r \rvert$ 0.431): Pass
S5	≥ 70% of reference signs agree	4/5: Pass	2/5: Fail	4/5: Pass
Verdict	Five-step validation	Passes	Does not pass	Passes

BSS, Brier skill score $1 - \mathrm{BS/UNC}$; MI in nats. A failure at Step 1 raises a warning but does not stop the analysis. "Passes" means that the target passed all five steps on this dataset and is eligible for further validation; it does not imply readiness for clinical use. Step 1 values are computed on the label-encoded targets of Section 3.2, whereas Steps 2 to 5 use the binary decision contrast.

An ablation that uses exactly the step functions and thresholds behind Table 5 clarifies what each step contributes. Removing any single step leaves all three verdicts unchanged, and each step used alone reproduces the same target-level classification as the full protocol, because mortality fails every step and the other two targets pass every step. The steps are therefore redundant on this dataset, and we do not claim an individual necessity the data do not support; on a dataset where a target failed only one dimension, the step detecting it would be necessary. What the ablation does show is the contrast with conventional practice: if target viability were judged solely by whether mean discrimination exceeded chance, mortality could appear marginally viable, since every model exceeds a mean AUC of 0.5, the highest being Logistic Regression at 0.541 ± 0.042, despite the failures detected by the remaining dimensions.

The design also functions as a controlled comparison across targets. The feature matrix, the split, the resampling and the hyperparameters are held fixed while only the label column changes, so the gap between diagnosis and mortality isolates the contribution of label informativeness and nothing else.

4.8. Principal Finding, Mechanism and Practical Implications

For mortality, the five measurements form one evidence chain. Before any model is fitted, no feature passes the dependence screen (Section 3.2). A model fitted anyway reaches a mean AUC slightly above 0.5 (Table 1), where conventional reporting would stop. Its resolution is near zero (Table 3), so the fitted probabilities carry almost none of the variance of the outcome and Platt scaling can only return the prevalence model; acting on them therefore cannot beat the default strategies (Fig. 6). The attributions fall on a variable with no measured association with the outcome (Fig. 7), and a transparent model contradicts the literature on three of five reference directions (Table 4). Each link constrains the next: there is little detectable usable signal under the available features and prespecified analyses, and the downstream steps consistently support that interpretation.

Learnability screening should therefore begin before model development, and the downstream steps then determine whether the initial screening result is supported by model-based evidence. Step 1 costs a correlation matrix and a mutual information estimate and flags a target for scrutiny before resources are committed. For journals and regulators, the five-step verdict records which dimension failed, which a headline AUC cannot; for development teams it answers the question of when to stop; and because nothing in it is specific to meningitis, it applies wherever features may not carry the signal sought, such as prognostic modelling from registries or risk scoring from administrative records.

4.9. Limitations

Six limitations bear on how far these results travel. First, the dataset is synthetic and contains physiologically implausible values, such as ages above 110 years; this makes the absence of feature-outcome association easier to interpret, although a generator may embed structure we have not tested for. It says nothing, however, about whether mortality is predictable in real cohorts, where published work reports that it is [15], [57]. Second, without an external cohort, the protocol is demonstrated rather than validated, and its thresholds may not transfer.

Third, a pairwise screen is vulnerable to interaction-driven structure. Two exclusive-or controls illustrate this. One, built from dichotomised age and CSF white blood cell count, narrowly passes Step 1 $(max\ |r| = 0.109, max\ MI = 0.080)$; the other, built from dichotomised age and sex, fails it $(max\ |r| = 0.053, max\ MI = 0.029)$, yet LightGBM recovers both on the test fold with an AUC of 1.000. Because purely interaction-based dependence can show little or no univariate association, a Step 1 failure is treated as a warning rather than a termination criterion. Fourth, the thresholds of Steps 1, 3 and 5 are conventions; the Step 5 verdicts are stable for any agreement threshold above 40% and up to 80% under the primary reference, but Steps 1 and 3 have not been examined for sensitivity.

Fifth, the Step 5 reference comes largely from adult community-acquired bacterial meningitis; the dataset documents neither its population nor how risk level was assigned, and pathogen presence is matched to a positive culture, so the audit is a plausibility check rather than a clinical validation; Steps 2, 3 and 5 also require a binary decision. Sixth, the steps are redundant on this dataset, so the evidence that each is individually necessary is weaker than the evidence that the set detects a failure that discrimination alone does not.

5. Comparison with Previous Studies

We compared the protocol with eleven previous studies that apply machine learning to meningitis or another central nervous system infection [3]–[6], [8]–[12], [15] and [16]. They form the core literature reviewed in Section 1, from classifier comparisons and ensembles to hybrid neural architectures and validated clinical prediction models. Studies cited there for other purposes are not tabulated: [27] for LIME-based interpretation, [7], a further hybrid model from the group that published [6], [8]–[10], and [17]–[22], which document the recent broadening of the field. Six of the eleven were read in full [4], [5], [12], [11], [15] and [16]. For the other five [3], [6] and [8]–[10], only the abstract was available to us, together with the figure captions for [6], so for these studies "not reported" means not reported in the text we could access. The comparison has two parts that answer different questions. Section 5.1 asks which validation dimensions each study reports, a question that does not depend on the data used. Section 5.2 compares predictive performance, which is meaningful only between studies that share the data and the target.

5.1. Methodological Comparison

Table 6 records, for each study, the data, the target, and the six validation dimensions of the protocol: discrimination, calibration, decision curve analysis, SHAP or another explanation of the model, a learnability check made before modelling, and an audit of fitted effect directions against the clinical literature. All eleven studies report discrimination (100%), although only three report an AUC value (27%) [5], [15], [16]; the others report accuracy, F1-scores, or class-wise detection rates. Calibration is assessed in one study (9%) [15] and decision curve analysis in two (18%) [15], [16]. SHAP appears in one study (9%) [16], and some explanation of the model, counting variable importance, interpretable decision trees, and nomograms, in six (55%) [3]–[5], [12], [15] and [16]. No study checks learnability before modelling (0%), and none audits its fitted effect directions against a prespecified set of literature directions (0%); [12] and [15] discuss agreement with earlier findings in prose. The two studies with the widest coverage, [15] and [16], report four and three of the six dimensions respectively, and both are single-centre cohorts validated on a random split. Because the counts for [3] and [6], [8]–[10] rest on abstracts, they are lower bounds for those studies, but no abstract mentions calibration, decision curve analysis, or a learnability check.

5.2. Predictive-Performance Comparison

The studies in Table 6 differ in cohort, target definition, features, and validation design, so their AUC and accuracy values cannot be ranked against one another or against ours. We therefore checked which studies use the data of the present study, the public Kaggle dataset of 1,200 records with ten predictors and three targets [60], first released on 4 February 2025. Among studies [5], [6] and [8]–[11], [5] uses 12,420 records from the Brazilian SINAN registry and [11] a different Kaggle dataset on disease outbreaks in Nigeria. Study [8] was published in August 2024, before the Kaggle release, and [9] predicts four aetiologies, including fungal and tuberculous meningitis, that our dataset does not contain. The abstract of [10] does not name its data, so we cannot establish whether it is the same set. Only [6] is likely to share our data: it names the dataset's own column labels, Pathogen_Present, Outcome, and Risk_Level, was published after the Kaggle release, and predicts the type of meningitis, which corresponds to our diagnosis target, although its abstract lists fungal rather than unknown aetiology among the types.

Table 6. Validation dimensions reported by the eleven previous meningitis and central nervous system infection studies [3]–[6], [8]–[12] and [15], [16] and by this study

Study	Data (n)	Target	Discrimination	Calib.	DCA	SHAP or other XAI	Learnability screen	Direction audit
Sathiya Priya et al. [10] †	CSF data, source not stated	Meningitis subtype	Accuracy, F1	—	—	—	—	—
Shabana et al. [11]	Kaggle, Nigeria outbreak data	Meningitis detection	Accuracy	—	—	—	—	—
Zhou et al. [15]	Single hospital, China (277)	30-day mortality	AUC	✓	✓	Nomogram	–	–
Luo et al. [16]	Single hospital, China (200)	Meningoencephalitis	AUC	—	✓	SHAP	–	–
Studies reporting, n (%)			11 (100%)	1 (9%)	2 (18%)	6 (55%); SHAP 1 (9%)	0 (0%)	0 (0%)
This study	Kaggle, synthetic (1,200)	Diagnosis, mortality, risk level	AUC, F1, MCC	✓	✓	SHAP	✓	✓

Note: ✓ reported; — not reported. † Assessed from the abstract only (for [6], also the figure captions), so "—" means not reported in the accessible text. Learnability screen: a feature-target dependence check made before modelling to decide whether to proceed; feature selection for model building is not counted. Direction audit: comparison of fitted effect directions with a prespecified set of directions from the clinical literature. AUC, area under the ROC curve; Calib., calibration; CSF, cerebrospinal fluid; DCA, decision curve analysis; MCC, Matthews correlation coefficient; SHAP, SHapley Additive exPlanations; XAI, explainable artificial intelligence.

Table 7. Diagnosis target: the study that likely uses the same dataset [6] compared with the representative model of this study

Study	Inputs	Evaluation	Accuracy	Precision	Recall	F1	AUC	MCC
Sathiya Priya et al. [6]	Includes Outcome and Risk Level	Single test set, split not stated	0.92	0.91	0.92	0.91	NR	NR
This study, LightGBM	Ten clinical predictors	Repeated CV, 5 × 5	0.887 ± 0.014	0.879 ± 0.013	0.887 ± 0.014	0.881 ± 0.012	0.970 ± 0.007	0.793 ± 0.025
This study, LightGBM	Ten clinical predictors	Held-out test fold (n = 180)	0.889	0.866	0.889	0.875	0.975	0.790

Note: NR, not reported. Values for [6] are taken from its abstract, which does not state how precision, recall, and F1 were averaged; [6] shows an ROC curve, but no AUC value is given in the text available to us. For this study, precision, recall, and F1 are support-weighted averages over the three classes (macro-F1 0.708 ± 0.037 under cross-validation and 0.656 on the test fold), AUC is the macro-averaged one-vs-rest value, and cross-validated values are the mean ± standard deviation over the 25 folds of Table 1.

Table 7 compares the results of [6] with those of LightGBM, the representative model in this study, using the metrics reported in the original study. The original study reports an accuracy of 0.92 and an F1-score of 0.91 on a single test set, although the sample size and data-splitting procedure are not specified. In comparison, our model achieves a cross-validated accuracy of 0.887 and a weighted F1-score of 0.881, with corresponding values of 0.889

and 0.875 on the held-out test fold. Two methodological differences limit the comparability of these results. First, the earlier study includes Outcome and Risk Level as input features, whereas our approach uses only the ten clinical predictors. In this dataset, 84.2% of bacterial cases are labelled high risk, while 83.1% of viral cases are labelled low risk. Thus, Risk Level provides information about the diagnosis that our feature set deliberately excludes. Second, the earlier study does not report AUC, MCC, or class-wise performance in its abstract, preventing a direct comparison of how well the models handle the Unknown class (5.6% of records), for which our model achieves an F1-score of 0.125 (Table 2). Therefore, the reported results are broadly consistent but do not establish the superiority of either model.

5.3. What the Comparison Shows

The comparison does not show that the protocol produces better classifiers. The classifiers in this study are standard, and on the diagnosis target their discrimination is similar to that reported in [6]. What the comparison shows is a difference in what is reported. A study that reports discrimination alone cannot say whether its probabilities support a threshold decision, whether acting on the model helps patients, whether the model relies on clinically recognisable variables, or whether a weak result reflects a poor model or an unlearnable target.

The mortality target illustrates the last point. Nine of the eleven studies (82%) report discrimination without calibration or decision curve analysis. Judged in that way, the highest mean AUC, 0.541 for Logistic Regression (Table 1), lies above chance, and mortality could appear marginally learnable. The five-step evidence chain reaches the opposite conclusion: no feature passes the dependence screen, the calibrated model does not improve on the prevalence model, no threshold gives net benefit above the default strategies, the leading attribution falls on a variable with no measured association, and only two of five coefficient signs agree with the literature (Table 5). The contribution of this study is therefore the integration of these dimensions into one protocol whose first step runs before any model is fitted, not a higher accuracy figure. The advantage of gradient boosting over Logistic Regression on the learnable targets is consistent with its established behaviour on tabular data [48]–[50], yet Logistic Regression is what makes the coefficient audit possible, so a validation framework that discarded interpretable linear models would lose that dimension [28].

6. Conclusions

This study aimed to screen target learnability before model development and subsequently evaluate whether each target warrants further external validation through a five-step protocol. The results demonstrate that the protocol distinguishes among the three meningitis-related targets in the evaluated dataset: diagnosis and risk level satisfy all five criteria, whereas mortality fails them. The protocol is supported by five formal results concerning Shapley uniqueness, the Brier score lower bound, decision-curve dominance, integer score rounding, and SMOTE non-leakage. Together, these results provide the theoretical and methodological basis for the proposed validation procedure.

The findings show that target suitability cannot be established by a headline discrimination metric alone. The initial dependence screen provides an inexpensive assessment before model development, while the subsequent steps identify the specific dimensions in which a target does not meet the prespecified criteria. This staged assessment offers a more structured basis for deciding which targets merit further investigation than relying on a single aggregate performance score. However, the protocol has been demonstrated on only one synthetic dataset and has not yet been validated externally. The findings therefore do not establish its superiority over existing evaluation practices in general, and none of the resulting models should be considered ready for clinical deployment. Rather, the study provides a systematic procedure for identifying targets that warrant further validation through complementary evidence collected at different stages of the modelling pipeline.

Future work should validate the protocol using multicentre cohorts with prospectively collected mortality outcomes, extend Step 1 to multivariate dependence measures such as distance correlation or the Hilbert–Schmidt independence criterion, and formulate Steps 2, 3, and 5 directly for multiclass targets, including an interval-based sign audit. Further studies should also incorporate additional prognostic markers, such as platelet indices [68] , to assess whether richer clinical information changes the screening results and subsequent validation decisions.

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

Funding: Bina Nusantara University, BINUS International Research – Basic scheme (2026).

Data Availability Statement: Data: Kaggle, https://www.kaggle.com/datasets/chantest/meningitis-classification. Code: https://github.com/fathimahalmashumah/meningitis-learnability-protocol (DOI: 10.5281/zenodo.22909762). The executed notebook and the data are also provided as Supplementary Material S1.

Acknowledgments: This work is supported by Bina Nusantara University as a part of Bina Nusantara University's BINUS International Research - Basic entitled A Multidisciplinary Computational Framework for Understanding, Managing, and Detecting Meningitis: Dynamic Modeling, Fuzzy Treatment Evaluation, and Machine Learning–Based Diagnosis with contract number: 067/VRRTT/IV/2026 and contract date 01 April 2026. We thank the Computer Science Department, BINUS, and the Department of Clinical Pathology, UNISBA, for their support. Claude (Anthropic) assisted with language editing and code consolidation; the authors verified all content and take full responsibility for it.

Conflicts of Interest: None. The funder had no role in the study or the decision to publish.

References

[1] J. R. Zunt *et al.*, "Global, regional, and national burden of meningitis, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Neurol.*, vol. 17, no. 12, pp. 1061–1082, Dec. 2018, doi: 10.1016/S1474-4422(18)30387-9.

[2] D. van de Beek, J. de Gans, L. Spanjaard, M. Weisfelt, J. B. Reitsma, and M. Vermeulen, "Clinical Features and Prognostic Factors in Adults with Bacterial Meningitis," *N. Engl. J. Med.*, vol. 351, no. 18, pp. 1849–1859, Oct. 2004, doi: 10.1056/NEJMoa040845.

[3] B. Dobardžić, A. Alibašić, N. Milošević, B. Mališić, and M. Vukotić, "Forecasting Meningitis with Machine Learning: An Advanced Classification Model Analysis," in *IFMBE Proceedings*, vol. 93, 2024, pp. 725–731. doi: 10.1007/978-3-031-49062-0_76.

[4] V. M. Lelis, E. Guzmán, and M.-V. Belmonte, "Non-Invasive Meningitis Diagnosis Using Decision Trees," *IEEE Access*, vol. 8, pp. 18394–18407, 2020, doi: 10.1109/ACCESS.2020.2966397.

[5] E. Guzmán, M. Belmonte, and V. M. Lelis, "Ensemble methods for meningitis aetiology diagnosis," *Expert Syst.*, vol. 39, no. 8, Sep. 2022, doi: 10.1111/exsy.12996.

[6] S. Sathiya Priya, M. Amanullah, and T. Kujani, "Enhanced Meningitis Type Classification Using Ensemble Learning: A Hybrid LightGBM Model with Advanced Preprocessing," in *2025 9th International Conference on Inventive Systems and Control (ICISC)*, Aug. 2025, pp. 763–768. doi: 10.1109/ICISC65841.2025.11188314.

[7] S. Sathiya Priya, M. Amanullah, and T. Kujani, "A Hybrid Model for Meningitis Diagnostic Accuracy with Ensemble Learning: Combining Logistic Regression and L-BFGS," in *2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, Jun. 2025, pp. 290–295. doi: 10.1109/ICC-ROBINS64345.2025.11086348.

[8] S. Sathiya Priya, M. Amanullah, and T. Kujani, "Enhancing Meningitis Type Diagnosis through a Novel Hybrid Ensemble of Bagging and Gradient Boosting," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Aug. 2024, pp. 1260–1264. doi: 10.1109/ICoICI62503.2024.10696600.

[9] S. Sathiya Priya, M. Amanullah, and P. Devi, "High-Performance Meningitis Etiology Prediction: Leveraging CSF Biomarkers Via a Stacked Ensemble of Random Forest and Artificial Neural Networks," in *2025 9th International Conference on Inventive Systems and Control (ICISC)*, Aug. 2025, pp. 1226–1232. doi: 10.1109/ICISC65841.2025.11188299.

[10] S. Sathiya Priya, M. Amanullah, and P. Harini, "Fusion-based CNN-SVM Hybrid Model for Efficient Meningitis Identification using CSF Dataset," in *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Aug. 2025, pp. 1115–1119. doi: 10.1109/ICSCDS65426.2025.11167675.

[11] A. Shabana, P. Kavitha, and S. Kamalakkannan, "Detection of meningitis disease using Belief Bidirectional Neural Network and Informative Ant Colony Optimization techniques," *Indian J. Eng.*, vol. 22, no. 58, pp. 1–17, Aug. 2025, doi: 10.54905/disssi.v22i58.e12ije1701.

[12] A.-F. A. Mentis *et al.*, "Artificial Intelligence in Differential Diagnostics of Meningitis: A Nationwide Study," *Diagnostics*, vol. 11, no. 4, p. 602, Mar. 2021, doi: 10.3390/diagnostics11040602.

[13] A. Spanos, F. E. Harrell, and D. T. Durack, "Differential diagnosis of acute meningitis: an analysis of the predictive value of initial observations," *JAMA*, vol. 262, no. 19, pp. 2700–2707, Nov. 1989, doi: 10.1001/jama.1989.03430190084036.

[14] L. E. Nigrovic et al., "Clinical Prediction Rule for Identifying Children With Cerebrospinal Fluid Pleocytosis at Very Low Risk of Bacterial Meningitis," *JAMA*, vol. 297, no. 1, pp. 52–60, Jan. 2007, doi: 10.1001/jama.297.1.52.

[15] J. Zhou, J. Xing, X. Cheng, F. Jin, and Y. Huang, "Development and validation of a predictive model for 30-day mortality in adult bacterial meningitis: a retrospective cohort study," *Front. Med.*, vol. 12, Nov. 2025, doi: 10.3389/fmed.2025.1692631.

[16] Y. Luo *et al.*, "Construction and validation of a predictive model for meningoencephalitis in pediatric scrub typhus based on machine learning algorithms," *Emerg. Microbes Infect.*, vol. 14, no. 1, Dec. 2025, doi: 10.1080/22221751.2025.2469651.
[17] B. Zhou, F. Liang, Y. Huang, S. Zhang, Z. Zhuo, and C. Chen, "A predictive model for central nervous system infections in children based on machine learning and clinical diagnostic features," *Front. Pediatr.*, vol. 14, pp. 1–13, Jul. 2026, doi: 10.3389/fped.2026.1861527.
[18] M. Yu, Y. Fang, and X. Yang, "Development and internal validation of a prognostic nomogram for poor outcomes in neonatal bacterial meningitis," *Transl. Pediatr.*, vol. 15, no. 5, p. 177, May 2026, doi: 10.21037/tp-2026-0206.
[19] Y. Chen *et al.*, "Development of a machine learning-based predictive model for long-term adverse outcomes in neonatal bacterial meningitis," *J. Pediatr. (Rio J.)*, vol. 102, no. 1, p. 101472, Jan. 2026, doi: 10.1016/j.jped.2025.101472.
[20] D. Van Dong *et al.*, "Causative pathogens and predictors of unfavourable outcomes in central nervous system infections in resource-limited settings," *Int. J. Infect. Dis.*, vol. 159, p. 108002, Oct. 2025, doi: 10.1016/j.ijid.2025.108002.
[21] J. Milburn, R. Suresh, R. Doyle, and J. N. Jarvis, "The diagnosis of central nervous system infections in resource-limited settings and the use of novel and molecular diagnostic platforms to improve diagnosis," *Expert Rev. Mol. Diagn.*, vol. 24, no. 3, pp. 219–230, Mar. 2024, doi: 10.1080/14737159.2024.2317414.
[22] G. Lu *et al.*, "Prediction model of central nervous system infections in patients with severe traumatic brain injury after craniotomy," *J. Hosp. Infect.*, vol. 136, pp. 90–99, Jun. 2023, doi: 10.1016/j.jhin.2023.04.004.
[23] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 146, Dec. 2019, doi: 10.1186/s12911-019-0874-0.
[24] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 5, Sep. 2020, doi: 10.1002/widm.1379.
[25] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," *BMC Med.*, vol. 17, no. 1, p. 230, Dec. 2019, doi: 10.1186/s12916-019-1466-7.
[26] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Nov. 2017, pp. 4768–4777. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230
[27] A. Messai *et al.*, "Transparent AI Models for Meningococcal Meningitis Diagnosis: Evaluating Interpretability and Performance Metrics," in *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, Aug. 2024, pp. 1–8. doi: 10.1109/IS61756.2024.10705255.
[28] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
[29] G. Y. Gombolay *et al.*, "Effects of explainable artificial intelligence in neurology decision support," *Ann. Clin. Transl. Neurol.*, vol. 11, no. 5, pp. 1224–1235, May 2024, doi: 10.1002/acn3.52036.
[30] L. Bergomi *et al.*, "Which explanations do clinicians prefer? A comparative evaluation of XAI understandability and actionability in predicting the need for hospitalization," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 269, Jul. 2025, doi: 10.1186/s12911-025-03045-0.
[31] A. H. Murphy, "A New Vector Partition of the Probability Score," *J. Appl. Meteorol.*, vol. 12, no. 4, pp. 595–600, Jun. 1973, doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.
[32] A. J. Vickers and E. B. Elkin, "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models," *Med. Decis. Mak.*, vol. 26, no. 6, pp. 565–574, Nov. 2006, doi: 10.1177/0272989X06295361.
[33] N. S. Groeneveld *et al.*, "Diagnostic prediction models for bacterial meningitis in children with a suspected central nervous system infection: a systematic review and prospective validation study," *BMJ Open*, vol. 14, no. 8, p. e081172, Aug. 2024, doi: 10.1136/bmjopen-2023-081172.
[34] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement," *BMJ*, vol. 350, p. g7594, Jan. 2015, doi: 10.1136/bmj.g7594.
[35] E. W. Steyerberg *et al.*, "Assessing the performance of prediction models: a framework for traditional and novel measures," *Epidemiology*, vol. 21, no. 1, pp. 128–138, Jan. 2010, doi: 10.1097/EDE.0b013e3181c30fb2.
[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
[37] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 1, p. 106, Dec. 2013, doi: 10.1186/1471-2105-14-106.
[38] A. Bastug, B. O. Ozbay, E. Oksuz, and O. Karabay, "Systematic Review of Artificial Intelligence Implementations in Central Nervous System Infections," *Mediterr. J. Infect. Microbes Antimicrob.*, vol. 12, Jul. 2023, doi: 10.4274/mjima.galenos.2023.2023.13.
[39] A. Hameed Mousa, E. M. Thabit A. Alsaadi, M. A. Mohammed, H. Mezher Merdas, and S. Dakhil Khalaf, "Evaluating the Impact of SMOTE and SHAP on Machine Learning Classifiers: Enhancing Predictive Performance through Imbalance Mitigation and Interpretability," *Adv. Artif. Intell. Mach. Learn.*, vol. 06, no. 02, pp. 5339–5357, 2026, doi: 10.54364/AAIML.2026.62296.
[40] C. Yang, E. A. Fridgeirsson, J. A. Kors, J. M. Reps, and P. R. Rijnbeek, "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data," *J. Big Data*, vol. 11, no. 1, p. 7, Jan. 2024, doi: 10.1186/s40537-023-00857-7.
[41] M. Sobhan and A. M. Mondal, "Evaluating SHAP's Robustness in Precision Medicine: Effect of Filtering and Normalization," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2023, pp. 3157–3164. doi: 10.1109/BIBM58861.2023.10385704.
[42] A. J. Vickers and F. Holland, "Decision curve analysis to evaluate the clinical benefit of prediction models," *Spine J.*, vol. 21, no. 10, pp. 1643–1648, Oct. 2021, doi: 10.1016/j.spinee.2021.02.024.
[43] Y. Wu and A. Wang, "Calibrated and Explainable CIN2+ Risk Stratification Using Routine Clinical Data: Development and External Validation," *Int. J. Womens. Health*, vol. 18, pp. 1–14, Jun. 2026, doi: 10.2147/IJWH.S614964.

[44] H. Johnston, N. Nair, and D. Du, "Estimating Calibrated Risks Using Focal Loss and Gradient-Boosted Trees for Clinical Risk Prediction," *Electronics*, vol. 14, no. 9, p. 1838, Apr. 2025, doi: 10.3390/electronics14091838.
[45] E. Winarno, I. M. Nur, A. Karim, S. Amri, I. E. Wirdati, and P. W. Adi, "An Explainable Multimodal Framework for Chest X-Ray Alert Classification Using Radiology Reports and Images," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 647–666, May 2026, doi: 10.62411/jcta.16023.
[46] D. B. Firmawan and B. R. P. Darnoto, "Cross-Domain Faithfulness Evaluation of SHAP and Attention-Based Explanations in Transformer NLP Models," *J. Comput. Theor. Appl.*, vol. 4, no. 1, pp. 146–163, Jul. 2026, doi: 10.62411/jcta.16258.
[47] C. Prakash, A. Sisodia, and M. Lind, "A Systematic Review of Agentic AI in Healthcare: An Evidence-Informed Seven-Principle Framework," *J. Comput. Theor. Appl.*, vol. 4, no. 1, pp. 43–73, Jun. 2026, doi: 10.62411/jcta.16171.
[48] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
[49] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Dec. 2018. [Online]. Available: http://arxiv.org/abs/1706.09516
[50] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[51] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[52] L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games (AM-28), Volume II*, vol. 2. Princeton, NJ, USA: Princeton University Press, 1953, pp. 307–317. doi: 10.1515/9781400881970-018.
[53] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, Art. no. 15, Dec. 2012, doi: 10.1145/2382577.2382579.
[54] S. F. Alnomasy, B. S. Alotaibi, A. H. Mujamammi, E. A. Hassan, and M. E. Ali, "Microbial aspects and potential markers for differentiation between bacterial and viral meningitis among adult patients," *PLoS One*, vol. 16, no. 6, p. e0251518, Jun. 2021, doi: 10.1371/journal.pone.0251518.
[55] L. U. Gerdes, P. E. Jørgensen, E. Nexø, and P. Wang, "C-reactive protein and bacterial meningitis: a meta-analysis," *Scand. J. Clin. Lab. Invest.*, vol. 58, no. 5, pp. 383–393, Jan. 1998, doi: 10.1080/00365519850186364.
[56] S. Singh, K. Mahto, A. Kumar, P. Kumar, S. Kumar, and M. K. Prasad, "Diagnostic Test Accuracy of Serum and Cerebrospinal Fluid C-Reactive Protein in Bacterial Meningitis: A Systematic Review and Meta-Analysis," *Infect. Chemother.*, vol. 57, no. 2, p. 248, 2025, doi: 10.3947/ic.2024.0139.
[57] M. W. Bijlsma *et al.*, "Community-acquired bacterial meningitis in adults in the Netherlands, 2006–14: a prospective cohort study," *Lancet Infect. Dis.*, vol. 16, no. 3, pp. 339–347, Mar. 2016, doi: 10.1016/S1473-3099(15)00430-2.
[58] S. Tubiana *et al.*, "Community-acquired bacterial meningitis in adults: in-hospital prognosis, long-term disability and determinants of outcome in a multicentre prospective cohort," *Clin. Microbiol. Infect.*, vol. 26, no. 9, pp. 1192–1200, Sep. 2020, doi: 10.1016/j.cmi.2019.12.020.
[59] N. Chekrouni *et al.*, "Characteristics and prognostic factors of bacterial meningitis in the intensive care unit: a prospective nationwide cohort study," *Ann. Intensive Care*, vol. 13, no. 1, p. 124, 2023, doi: 10.1186/s13613-023-01218-6.
[60] ChanTest, "Meningitis Classification," *Kaggle*, 2025. https://www.kaggle.com/datasets/chantest/meningitis-classification
[61] N. Hidayat, L. Afuan, and H. R. Jannah, "Prescriptive Learning Analytics for Student Dropout: Integrating Temporal Velocity and Counterfactual Explanations in Longitudinal Data," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 627–646, May 2026, doi: 10.62411/jcta.15920.
[62] M. G. S. P. Kusuma *et al.*, "Integrating Quantum, Deep, and Classic Features with Attention-Guided AdaBoost for Medical Risk Prediction," *J. Comput. Theor. Appl.*, vol. 3, no. 2, pp. 171–189, Oct. 2025, doi: 10.62411/jcta.14873.
[63] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large-Margin Classifiers*, The MIT Press, 2000, pp. 61–74. doi: 10.7551/mitpress/1113.003.0008.
[64] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html
[65] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.
[66] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning - ICML '05*, 2005, pp. 625–632. doi: 10.1145/1102351.1102430.
[67] O. A. Abioye and M. E. Irhebhude, "Big Data-Driven Health Risk Stratification: A Health Index-Based Approach Using Feature Importance and PySpark," *J. Comput. Theor. Appl.*, vol. 2, no. 4, pp. 456–469, Mar. 2025, doi: 10.62411/jcta.12327.
[68] A. J. Yadav and S. Mohammed Aslam, "Platelet indices as prognostic markers in the critically ill patient," *Ital. J. Med.*, vol. 15, no. 1, pp. 34–40, Dec. 2020, doi: 10.4081/itjm.2020.1388.