

Leakage-Aware Recall-Oriented Stacking for Biomarker-Based Bacterial and Viral Meningitis Classification

Pasya Kemal Halim

Muhammad Fatih Rizki Herianto

Ratu Azalea Meiza Asri

Dani Suandi

School of Computer Science, Universitas Bina Nusantara, Jakarta, Indonesia

ORCID ID: –

Abstract

Early differentiation of bacterial from viral meningitis is important because the two pathogens need different medications that often needs to be started before laboratory results are available. This paper presents a recall-oriented leakage-aware ML workflow for bacterial vs viral meningitis classification. The dataset used is publicly available Kaggle data of 1,200 patient records and 14 variables, with the emphasis on reducing the number of missed bacterial cases. After target cleaning, 67 unconfirmed records were removed resulting in a total of 1,133 confirmed binary cases, bacterial (595, 52.52%) and viral (538, 47.48%). Before model training, the workflow performed audits of variable roles, invalid inputs, categorical encodings, statistical assumptions, and leakage risks. Variables at admission, derived risk labels and near-target proxy variables were removed, resulting in eight variable predictors for modeling. We explored several machine learning configurations including tuned base classifiers, out-of-fold stacking, weighted probability, and pruned stacking. The Multilayer Perceptron performed best on the held-out set for bacterial recall (0.9832) with two false negatives for bacteria, but at the cost of lower viral recall. The Meta-Neural Network and final pruned stacking model yielded a more balanced classification profile with four bacterial false negatives and eleven viral false negatives, with bacterial recall of 0.9664 and viral recall of 0.8981, respectively. Impressively, this performance was achieved by the pruned stack with a compact 16-feature representation, rather than the full stacking setup. These results support recall-focused model comparison on this dataset, but external validation is still required before any clinical application can be contemplated.

Keywords

bacterial meningitis, viral meningitis, biomarker classification, stacking ensemble, out-of-fold prediction, data leakage, false negatives, clinical decision support

INTRODUCTION

Meningitis is a serious neurological disease that requires rapid and accurate diagnosis as the treatments of bacterial and viral infections differ dramatically (Ahmed et al., 2022; Groeneveld et al., 2024). Bacterial meningitis is a medical emergency. Hence, doctors are often required to initiate antibiotic treatment before the final results from the laboratory are obtained (Choi et al., 2023). Early tests such as cerebrospinal fluid (CSF) protein, glucose and serum C-reactive protein (CRP) are useful but no one marker is perfect in isolation. Machine learning (ML) can incorporate these early markers to improve predictions, but only if the models are trained on the same information that the doctors have at the time of initial screening.

This approach is confirmed by recent studies. Specific inflammatory markers can aid in distinguishing between the various forms of meningitis, although their accuracy is influenced by patient demographics and timing (Baran et al., 2025; Kim et al., 2021). ML models can combine these measurements well. But they are clinically useful only if they strictly use data available at the very moment doctors make their first decisions (Cheah et al., 2025).

Various ML algorithms have been considered for the classification of meningitis. Although these models are promising, usually a high accuracy score is misleading (Ghadaripouri et al., 2024). In most cases, these inflated scores are caused by data leakage, a problem where models inadvertently use future lab results or target data not available to doctors during early screening.

Existing research does not usually have clear ways to address this problem. Most of the previous studies have high accuracy but lack critical information on data preparation or how they handled complex variables (Andaur Navarro et al., 2021; Collins et al., 2024). These hidden choices heavily constrain model performance in real hospital settings. Models trained on compromised or future data cannot safeguard real patients.

Patient safety is the primary clinical priority. Researchers must minimize false negatives to avoid misdiagnosing dangerous bacterial infections as viral cases (Cross et al., 2024; Theodosiou & Read, 2023). This work presents a robust, leak-free system to assess different machine learning models. The main contribution is not a new algorithm, but a reliable framework that strictly separates data preparation from final model testing.

LITERATURE REVIEW

Biomarkers for Bacterial-Versus-Viral Meningitis

Often doctors must treat meningitis before final lab tests identify the exact cause. Because of this timeline, researchers are evaluating how early biomarkers can help distinguish

whether an infection is bacterial or viral. It has been shown in previous studies that markers such as cerebrospinal fluid (CSF) glucose, CSF protein, white blood cell counts, C-reactive protein (CRP) and procalcitonin reflect certain bacterial or viral patterns. Several studies emphasize the value of these particular markers. CSF CRP is more specific for bacterial meningitis than blood CRP (Singh et al. 2025). Groeneveld et al. (2024) demonstrated good performance of CSF markers (IL-6, procalcitonin, CRP, and ferritin) for diagnosis of bacterial meningitis in children.

This pediatric emphasis is also backed by other studies. Kim et al. (2021) validated blood procalcitonin as an aid in the detection of the disease. Norman-Bruce et al. (2024) compared procalcitonin and CRP for screening for serious bacterial infections in young infants with fevers. Ahmed et al. (2022) also evaluated CSF- CRP, blood procalcitonin and PCR tests in order to differentiate bacterial from viral cases in children. In a broader scope, Olie et al. (2025) validated CSF CRP as a reliable predictor of bacterial meningitis in a variety of patient populations, while Arafa et al. (2023) showed the effectiveness of CSF lactate in discriminating bacterial from viral cases in real-world circumstances. Finally, blood markers were reviewed by Baran et al. (2025), who highlighted how their integration with new technologies improves diagnosis but needs more real-world confirmation.

Our current dataset doesn't include all the clinical markers of relevance. Rather it has a limited set of demographic information, CSF data, blood tests and inflammatory measures. The variables selected for this study represent a targeted feature set for the computational model rather than a complete clinical panel. This distinction is critical. Although public datasets frequently include variables with strong mathematical relationships to a diagnosis, these features are practically useless for early detection if clinicians cannot access them during the initial patient screening.

Clinical Machine Learning and Ensemble Modeling

Clinical machine learning analyzes multiple patient variables to detect complex patterns that standard, single-rule tests often overlook. Researchers widely apply these computational methods to etiological screening for meningitis. For example, Mentis et al. (2021) utilized machine learning models to differentiate between bacterial and viral infections. Choi et al. (2023) developed an artificial intelligence model for early classification of meningitis and encephalitis. Babenko et al. (2021) targeted children and used decision trees with CRP and procalcitonin levels to distinguish between bacterial and enteroviral cases. Other teams have attempted more complex systems. Guzman et al. (2022) evaluated the use of ensemble methods for the exact diagnosis of the cause of meningitis, while Victor et al. (2025) applied machine learning to a large public health dataset for both surveillance and diagnosis. A thorough review by Ghaddaripouri et al. (2024)

confirmed that researchers can use machine learning successfully to predict and diagnose meningitis in a variety of study designs. However, AI enhances infectious disease tracking and diagnosis, but as Cheah et al. (2025) point out, before hospitals apply the models on actual patients, they must validate them properly.

Stacking is a machine learning technique where the predictions of several base models are combined using a “meta-learner”. Recent studies have used stacked models as different base learners can capture different types of errors (Guzman et al., 2022; Mahajan et al., 2023; Sathiya Priya et al., 2024; Victor et al., 2025). But that is not assured. If the meta-learner learns from the same data as the base models used for training, the final model will yield overly optimistic results. To account for this, researchers use out-of-fold (OOF) predictions. In OOF, the system predicts the data sample with a model that has never seen the data during training. This mimics the real world setting and avoids the stacked model from overfitting to the training data.

Reporting, Leakage Control, and False-Negative-Aware Evaluation

Prediction-model studies need clear reporting of data sources, target definitions, predictors, validation design, performance metrics, and limitations (Collins et al., 2024). Earlier work on prediction-model reporting and risk-of-bias tools also stressed the need for transparent methods in machine-learning studies (Andaur Navarro et al., 2021; Collins et al., 2021). Recent reviews of clinical and infectious-disease machine learning warn that high internal performance can be misleading when validation, predictor eligibility, bias, and clinical implementation are weakly described (Cheah et al., 2025; Cross et al., 2024; Smiley et al., 2025; Theodosiou & Read, 2023). In biomedical machine learning, leakage is especially important because future information, target-derived labels, or variables measured after the intended decision point can enter the model silently.

For this chapter, the main error of concern is a bacterial false negative, defined as a bacterial case predicted as viral. This does not mean that viral errors are ignored. Viral recall and viral false negatives are also reported because a model that predicts nearly all cases as bacterial would not be useful for etiological differentiation. The evaluation therefore reads bacterial recall first, then checks viral recall, bacterial F2-score, Matthews correlation coefficient (MCC), ROC-AUC, and the raw false-negative counts.

METHODOLOGY / DATA AND METHODS

Dataset and Target Cleaning

The experiment used the public Kaggle meningitis classification dataset available at <https://www.kaggle.com/datasets/chantest/meningitis-classification>. The raw dataset contained 1,200 records and 14 variables. The target variable was **Diagnosis**,

with three initial labels: bacterial, viral, and unknown. Because the objective of this chapter is binary supervised classification, the unknown labels were removed before any feature-selection or model-training step. This was not a feature filter. It was a target-definition step: records labeled as unknown do not provide a confirmed bacterial or viral class, so they cannot train or evaluate a binary bacterial-versus-viral classifier without changing the study objective.

The cleaning impact was reported directly so the binary task remains clear. Of the 1,200 raw records, 67 records were labeled unknown and removed, representing 5.58% of the raw dataset. The final binary-labeled dataset contained 1,133 records, or 94.42% of the raw records. Within this final labeled dataset, 595 records were bacterial cases, equal to 52.52%, and 538 records were viral cases, equal to 47.48%. These percentages are important because they show that the final binary task was close to balanced. Missing values and duplicate rows were checked after target cleaning; no additional imputation or duplicate-removal rule was required for the final labeled dataset. The before-and-after class distribution is shown in Figure 1 to show the target-cleaning effect alongside the numerical summary.

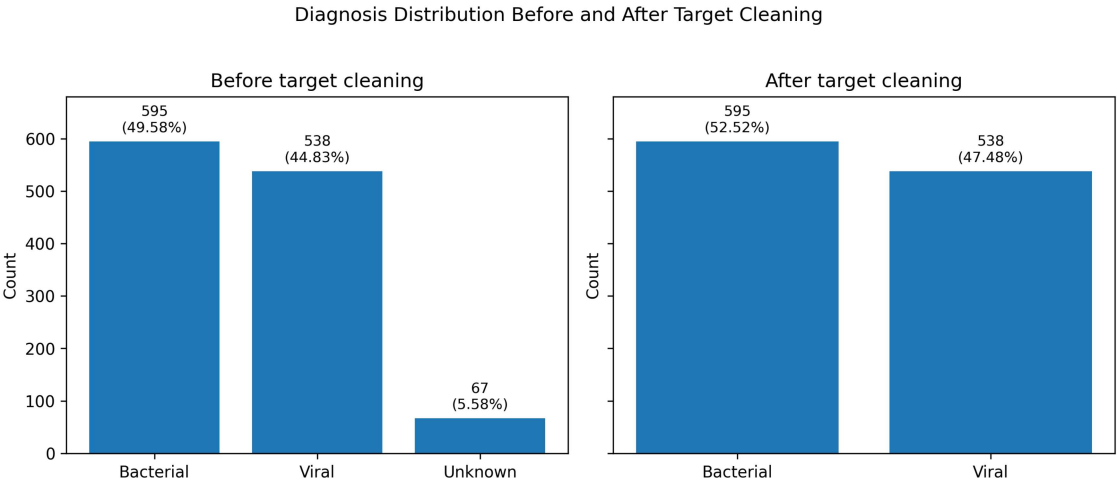


Figure 1: Diagnosis distribution before and after target cleaning. The unknown class was removed only to define the supervised binary task; bacterial and viral records were retained.

Workflow Overview

The workflow was organized so that target definition, variable auditing, statistical screening, and model development were not mixed together. Figure 2 presents the full analysis path from raw data to locked held-out reporting. The variable-audit stage separates role assignment, coding, and invalid-input checks, while the statistical-audit stage places normality and homogeneity testing before association screening.

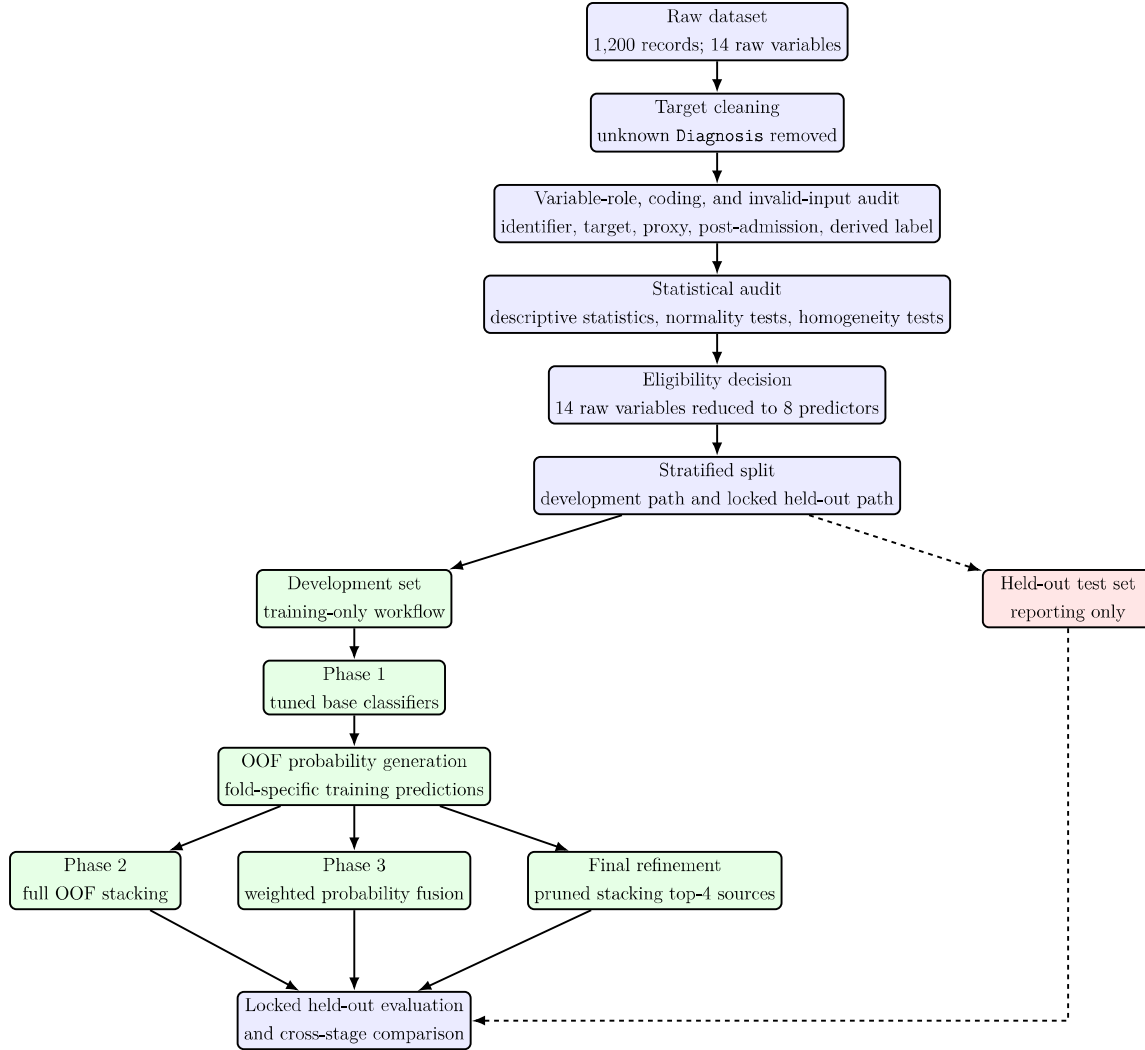


Figure 2: Overall experimental workflow. Solid arrows represent development operations. The dashed path represents the held-out test set, which was reserved for final reporting after all modeling configurations were locked.

Raw Variable Audit, Invalid Inputs, and Coding

After target cleaning, each raw variable was assigned a methodological role. In this chapter, an invalid input does not only mean a corrupted value. It means a variable that should not be used as an early predictor for the intended binary screening task. The data audit excluded patient identifiers, target diagnostic labels, post-admission outcomes, undocumented derived-risk metrics, and near-direct target proxies. We finalized this feature-eligibility screening prior to compiling descriptive statistics or training the models to guarantee that the biomarker matrix remains completely reproducible.

We established the coding rules prior to association screening. The supervised target was encoded as `Bacterial` = 0 and `Viral` = 1, which we applied during correlation analysis, confusion-matrix interpretation, ROC-AUC calculation, and threshold selection. For `Pathogen_Present`, the audit utilized `No` = 0 and `Yes` = 1 specifically to evaluate

data leakage risk; we excluded this variable from the final model inputs. We analyzed remaining categorical features strictly as background categories rather than active predictors. For instance, although **Risk_Level** contained High Risk, Low Risk, and Moderate Risk groups, we removed it because it represented an undocumented, derived risk metric rather than lacking statistical association with the target.

Table 1: Raw Variable Audit, Coding, and Predictor Eligibility

Variable	Data type	Initial role	Coding or category meaning	Final handling
Patient_ID	Identifier	Identifier	Row identifier only	Removed; no clinical predictive meaning
Age	Numerical	Candidate predictor	Continuous patient age	Kept as demographic predictor
Gender	Categorical	Candidate predictor	Female/Male categories; not retained	Removed from compact biomarker matrix
WBC_Count	Numerical	Candidate predictor	CSF white blood cell count	Kept as CSF-related predictor
Protein_Level	Numerical	Candidate predictor	CSF protein level	Kept as CSF-related predictor
Glucose_Level	Numerical	Candidate predictor	CSF glucose level	Kept as CSF-related predictor
Pathogen_Present	Categorical	Potential target proxy	Audit encoding: No=0, Yes=1	Removed after leakage audit
Diagnosis	Categorical	Target	Bacterial=0, Viral=1; Unknown removed	Target only
Outcome	Categorical	Post-admission variable	Recovered/Deceased status	Removed; post-admission information
Hemoglobin	Numerical	Candidate predictor	Blood hemoglobin value	Kept as blood predictor
WBC_Blood_Count	Numerical	Candidate predictor	Blood white blood cell count	Kept as blood predictor
Platelets	Numerical	Candidate predictor	Platelet count	Kept as blood predictor
CRP_Level	Numerical	Candidate predictor	C-reactive protein level	Kept as inflammatory predictor
Risk_Level	Categorical	Derived-risk label	High Risk, Low Risk, Moderate Risk	Removed; derived or undocumented label

The categorical distribution audit was used as supporting context. After target

cleaning, **Gender** was nearly balanced, with 572 female records (50.49%) and 561 male records (49.51%). **Pathogen_Present** contained 593 yes records (52.34%) and 540 no records (47.66%). **Risk_Level** consisted of 570 high-risk records (50.31%), 516 low-risk records (45.54%), and 47 moderate-risk records (4.15%). The **Outcome** variable consisted primarily of recovered patient records, which included 1,018 recovered cases (89.85%) and 115 deceased cases (10.15%). While analyzing these class distributions helped us identify features that required categorical association tests or data leakage audits, these statistical patterns did not override our role-based eligibility decisions.

Descriptive Statistics

We summarized the eight numerical measurement variables prior to association screening and model development. Table 2 reports their central tendency, spread, and modal behavior after target cleaning. This table provides descriptive context to clarify the scale and distributional behavior of the measurements; it does not function as a feature-selection rule and did not override the leakage-aware eligibility audit.

Table 2: Numerical Descriptive Statistics After Target Cleaning

Variable	Mean	Std. dev.	Q1	Median	Q3	Mode	Mode count	Mode %
Age	44.29	20.12	29.00	41.00	55.00	37	38	3.35
WBC_Count	11,713.14	5,236.24	7,040.00	11,688.00	16,279.00	4,987	3	0.26
Protein_Level	98.69	66.46	39.00	101.00	152.00	41	21	1.85
Glucose_Level	51.35	35.57	23.00	52.00	66.00	70	34	3.00
Hemoglobin	9.68	5.57	5.00	8.00	15.00	16	119	10.50
WBC_Blood_Count	11,718.08	4,927.08	7,201.00	12,420.00	16,140.00	8,444	2	0.18
Platelets	197,030.10	91,004.41	123,887.00	148,222.00	266,991.00	145,466	2	0.18
CRP_Level	23.31	21.64	3.00	24.00	39.00	2	95	8.38

Note: This descriptive table provides baseline context for the numerical variables after target cleaning. Predictor eligibility was decided later using role audit, leakage screening, statistical association, and model-development constraints.

Statistical Screening, Hypotheses, and Effect-Size Rules

We used statistical screening to select appropriate association summaries and guide the feature-eligibility analysis, rather than as a mechanical feature-selection filter. For the 1,133 cleaned and labeled records, we evaluated numerical variables using three normality diagnostics: Shapiro-Wilk (Shapiro & Wilk, 1965), Kolmogorov-Smirnov, and D’Agostino K-squared (D’Agostino & Pearson, 1973). Because the Shapiro-Wilk test provides a sensitive normality check for small-to-moderate samples while the Kolmogorov-Smirnov test offers a standardized distribution comparison that becomes highly sensitive in larger samples, we interpreted the two tests together instead of treating them as competing decisions. We included the D’Agostino K-squared test to combine skewness and kurtosis

evidence, which flags non-normal shapes that a single test statistic might overlook. Although the computational notebook displayed several exported p -values as 0.000000 due to six-decimal rounding, we report these values as $p < 0.000001$ because the true p -values are infinitesimally small rather than exactly zero.

Levene's test with median centering was then used to check variance homogeneity between bacterial and viral groups (Levene, 1960). Because the numerical variables were marked as non-normal and non-homogeneous under the stated rules, Spearman rank correlation was used instead of Pearson correlation for numerical association screening. Table 3 lists the decision rules. Table 4 reports the normality and Spearman outputs, while Table 5 reports the variance-homogeneity output.

For normality tests, the null hypothesis H_0 states that the variable does not significantly depart from a normal distribution. The alternative hypothesis H_1 states that it departs from normality. With $\alpha = 0.05$, H_0 is rejected when $p \leq 0.05$. For Levene's test, H_0 states that bacterial and viral groups have equal variances, while H_1 states that the variances differ. For Spearman correlation, H_0 states that $\rho = 0$, while H_1 states that a monotonic association exists. Spearman effect size was interpreted by absolute rho: less than 0.10 as negligible, 0.10-0.29 as weak, 0.30-0.49 as moderate, and 0.50 or higher as strong.

Categorical variables were not passed through normality or Q-Q analysis because categorical labels do not represent continuous distributions. They were evaluated using contingency tables, chi-square tests, and Cramer's V. For the chi-square test, H_0 states that the categorical variable is independent of **Diagnosis**; H_1 states that the variable is associated with **Diagnosis**. Cramer's V was treated as a 0-to-1 effect size, with less than 0.10 as negligible, 0.10-0.29 as weak, 0.30-0.49 as moderate, and 0.50 or higher as strong.

Table 3: Statistical Applicability and Decision Rules

Test or statistic	Input and calculation focus	Null hypothesis and rejection rule	How it was used
Shapiro-Wilk	Numerical variable distribution; statistic W (Shapiro & Wilk, 1965)	H_0 : no significant departure from normality; reject when $p < 0.05$	Checks whether numerical variables support parametric association screening.
Kolmogorov-Smirnov	Standardized numerical values compared with normal distribution	H_0 : standardized data are consistent with normality; reject when $p < 0.05$	Confirms normality evidence beside Shapiro-Wilk.
D'Agostino K-squared	Skewness and kurtosis departure from normality (D'Agostino & Pearson, 1973)	H_0 : skewness and kurtosis are consistent with normality; reject when $p < 0.05$	Provides a third normality diagnostic.
Levene's test	Variance comparison between bacterial and viral groups (Levene, 1960)	H_0 : group variances are equal; reject when $p < 0.05$	Supports the decision to avoid parametric assumptions.
Spearman correlation	Rank-based relation between a numerical variable and encoded diagnosis	$H_0 : \rho = 0$; reject when $p < 0.05$	Reports monotonic association for non-normal numerical variables.
Chi-square test	Contingency table between categorical variable and diagnosis	H_0 : variable is independent of diagnosis; reject when $p < 0.05$	Tests categorical association with the target.
Cramer's V	Effect size from chi-square association	No separate null hypothesis; 0-to-1 effect-size scale	Describes practical strength of categorical association.

Table 4: Normality and Spearman Association Summary for Numerical Variables

Variable	SW p rule: $p < 0.05$	KS p rule: $p < 0.05$	D'Agostino K^2	D'Agostino p rule: $p < 0.05$	Spearman ρ	Strength	Direction
Age	< 0.000001	< 0.000001	162.6528	< 0.000001	-0.3854	Moderate	Higher with bacterial
WBC_Count	< 0.000001	< 0.000001	807.0517	< 0.000001	-0.6569	Strong	Higher with bacterial
Protein_Level	< 0.000001	< 0.000001	82.8663	< 0.000001	-0.6446	Strong	Higher with bacterial
Glucose_Level	< 0.000001	< 0.000001	99.4427	< 0.000001	0.5834	Strong	Higher with viral
Hemoglobin	< 0.000001	< 0.000001	7810.6623	< 0.000001	0.7617	Strong	Higher with viral
WBC_Blood_Count	< 0.000001	< 0.000001	12660.6251	< 0.000001	-0.7522	Strong	Higher with bacterial
Platelets	< 0.000001	< 0.000001	170.5053	< 0.000001	0.7621	Strong	Higher with viral
CRP_Level	< 0.000001	< 0.000001	114.4221	< 0.000001	-0.7205	Strong	Higher with bacterial

Note: Several exported p-values rounded to 0.000000 at six decimals are written here as < 0.000001 to avoid implying an exact zero, with a rejection threshold of $p < 0.05$. Spearman strength was evaluated using absolute correlation ($|\rho|$), where values below 0.10 are negligible, 0.10–0.29 are weak, 0.30–0.49 are moderate, and 0.50 or higher are strong. Furthermore, because the diagnostic target was encoded as Bacterial = 0 and Viral = 1, the correlation direction translates directly to the clinical outcome. A negative correlation indicates that higher values of the numerical variable are associated with the Bacterial class, a positive correlation indicates that higher values are associated with the Viral class, and a correlation close to zero (below 0.10) indicates a negligible or absent relationship between the variable and the diagnosis.

We evaluated variance homogeneity after completing the normality tests because the two diagnostics address different analytical questions. Levene's test with median centering served as our primary decision rule for comparing group variances (Levene, 1960). Although we retained Bartlett's test as supporting evidence, it did not override the primary rule when results diverged because Bartlett's test is overly sensitive to departures from normality.

Table 5: Homogeneity of Variance Summary for Numerical Variables

Variable	Levene statistic	Levene p rule: $p < 0.05$	Bartlett statistic	Bartlett p rule: $p < 0.05$	Decision based on Levene
Age	44.5272	< 0.000001	11.7777	0.000599	Non-homogeneous variance
WBC_Count	7.5103	0.006231	0.1640	0.685520	Non-homogeneous variance
Protein_Level	35.2008	< 0.000001	0.0134	0.907989	Non-homogeneous variance
Glucose_Level	34.9700	< 0.000001	107.2226	< 0.000001	Non-homogeneous variance
Hemoglobin	11.6734	0.000657	1.3798	0.240137	Non-homogeneous variance
WBC_Blood_Count	12.3744	0.000453	3.7541	0.052679	Non-homogeneous variance
Platelets	390.6717	< 0.000001	178.9470	< 0.000001	Non-homogeneous variance
CRP_Level	6.3112	0.012136	66.8359	< 0.000001	Non-homogeneous variance

Note: The grouping variable was **Diagnosis**, comparing bacterial and viral records for each numerical variable. The null hypothesis for both tests is equal variance between groups. The final homogeneity decision follows Levene's p-value, with $p < 0.05$ indicating rejection of equal variance. Bartlett's test is shown as supporting evidence only because it is more sensitive to non-normality. P-values rounded to 0.000000 in the exported output are reported as < 0.000001.

The homogeneity audit showed that the equal-variance assumption was not supported for any of the eight numerical variables under the primary Levene rule. Some Bartlett

p -values were not significant, but this does not alter the decision because Bartlett’s test is less reliable when distributions are non-normal. The practical consequence is methodological rather than eliminative: we did not remove the variables, but the screening narrative avoided parametric mean-comparison logic and used Spearman rank correlation for numerical association.

Categorical Association and Leakage-Aware Eligibility

We used the categorical audit to separate valid predictors from leakage-prone variables. The chi-square statistic in Table 6 measures the distance between the observed cross-tabulated counts and the counts expected under independence. The degrees of freedom (df) describe the size of the contingency-table comparison. For a binary diagnosis against a two-level categorical variable, $df = 1$; for `Risk_Level`, which contains three categories, $df = 2$. A small p -value below .05 indicates that we reject the null hypothesis of independence. Cramer’s V then establishes the practical association strength on a 0-to-1 scale, where values below .10 are negligible, .10–.29 are weak, .30–.49 are moderate, and .50 or higher are strong.

`Pathogen_Present` was the most important categorical audit case. It had very strong association with the target and nearly separated bacterial and viral labels. Among records with `Pathogen_Present=No`, 36 of 540 cases were bacterial and 504 of 540 were viral, so 93.33% of this category was viral. Among records with `Pathogen_Present=Yes`, 559 of 593 cases were bacterial and 34 of 593 were viral, so 94.27% of this category was bacterial. Figure 3 shows this separation visually. The figure is used as leakage evidence, not as a reason to keep the variable.

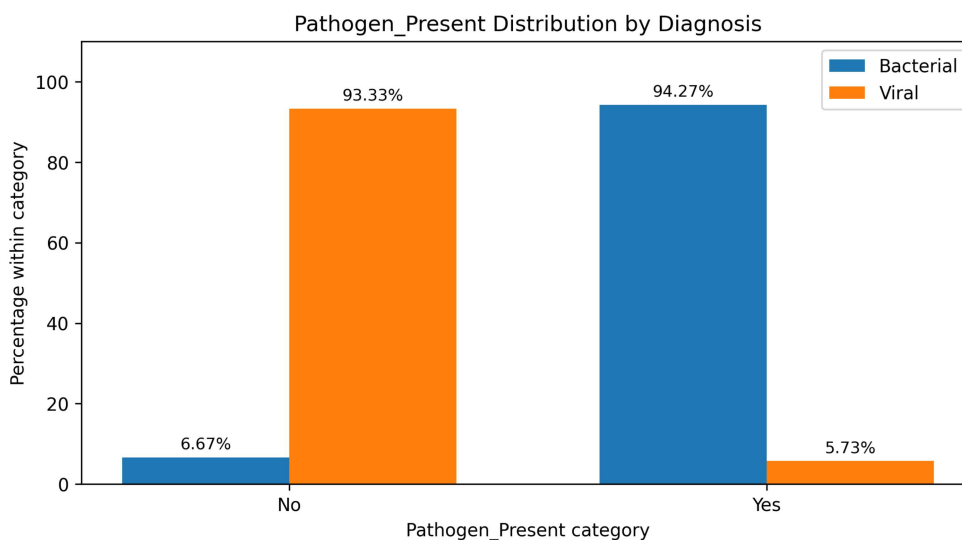


Figure 3: Conditional diagnosis distribution within each `Pathogen_Present` category. The near-separation pattern supported removing this variable as a target-proxy input.

The reason is methodological. If `Pathogen_Present` already functions as a near-

direct indication of the confirmed etiological label, a model using it could appear highly accurate without learning the intended biomarker pattern. The variable was therefore checked with crosstab counts, conditional percentages, Spearman correlation after binary encoding, chi-square, and Cramer’s V before the final eligibility decision. The result was consistent across methods, so **Pathogen_Present** was removed as a near-direct target proxy.

Table 6: Categorical Association and Leakage-Aware Eligibility Decisions

Variable	χ^2 larger = stronger count gap	df table size	p-value reject H_0 if $p < 0.05$	Cramer’s V 0-1 effect size	Strength	Eligibility interpretation
Gender	2.434144	1	0.118719	0.046351	Negligible	No significant categorical association; not retained in the compact biomarker matrix.
Pathogen_Present	866.221975	1	< 0.000001	0.874379	Strong	Very strong near-separation from the target; removed as a near-direct target proxy.
Outcome	5.489217	1	0.019134	0.069605	Negligible	Statistically detectable but very small association; removed because it is post-admission information.
Risk_Level	603.167995	2	< 0.000001	0.729633	Strong	Strong association but treated as a derived or undocumented risk label; removed from early biomarker modeling.

Note: H_0 for the chi-square test states that the categorical variable is independent of **Diagnosis**. Cramer’s V does not create a new hypothesis test; it summarizes effect size after the association test. P-values rounded to 0.000000 in the exported output are reported as < 0.000001. Important biomarkers discussed in recent meningitis literature, including procalcitonin and CSF lactate, were not available in the Kaggle dataset used in this study; the retained feature matrix therefore reflects the variables actually present in the public dataset rather than a complete clinical biomarker panel (Arafa et al., 2023; Groeneveld et al., 2024; Kim et al., 2021).

Final Feature Matrix, Split, and Preprocessing

After the eligibility audit, eight predictors were retained: **Age**, **WBC_Count**, **Protein_Level**, **Glucose_Level**, **Hemoglobin**, **WBC_Blood_Count**, **Platelets**, and **CRP_Level**. The retained matrix combines one demographic predictor, three CSF-related predictors, three blood-measurement predictors, and one inflammatory marker. This subsection fixes the feature matrix before splitting and standardization, which keeps preprocessing and model development in the correct order.

The final labeled dataset was split using stratified sampling into a development set and a held-out test set. Standardization was fitted on the development set only and then applied to the held-out set. The held-out set was not used for feature selection, hyperparameter tuning, OOF model ranking, ensemble weight selection, pruned-model construction, or threshold selection.

Table 7: Final Feature Matrix, Split, and Preprocessing Protocol

Item	Count or rule	Percentage	Methodological purpose
Final retained predictors	8 numerical predictors	–	Early-available demographic and biomarker matrix.
Removed raw variables	6 variables	–	Identifier, target-only, categorical non-retained, target proxy, post-admission, and derived-risk variables removed.
Development set	906 records	79.96%	Training-only workflow.
Held-out test set	227 records	20.04%	Final reporting only.
Development bacterial cases	476	52.54%	Preserves class balance.
Development viral cases	430	47.46%	Preserves class balance.
Held-out bacterial cases	119	52.42%	Final bacterial evaluation count.
Held-out viral cases	108	47.58%	Final viral evaluation count.
Standardization	Fit on development set only	–	Prevents preprocessing leakage.
Held-out usage	No tuning or threshold selection	–	Preserves final reporting estimate.

Model Family and Multi-Phase Development

The model set included linear, distance-based, kernel-based, tree-based, bagging, boosting, probabilistic, neural-network, and gradient-boosting families. The search space was compact so that tuning remained reproducible while still covering the main parameters of each model. GridSearchCV selected hyperparameters on the development data using bacterial F2-score, with the held-out test set excluded from tuning.

Table 8: Model Family, Search Space, and Selected Parameters

Family	Model	Main search focus	Selected parameters used
Linear	Logistic regression	Regularization strength C; L2 penalty; liblinear solver	$C = 0.01$; penalty = L2; solver = liblinear
Distance-based	KNN	Number of neighbors; distance weighting; Minkowski metric	neighbors = 11; weights = distance; metric = Minkowski
Kernel-based	SVM-RBF	Margin penalty C and radial-basis kernel width gamma	$C = 1$; gamma = 1

Family	Model	Main search focus	Selected parameters used
Single tree	Decision tree	Maximum depth; minimum split; minimum leaf	depth = 5; min leaf = 1; min split = 5
Bagging	Random forest	Number of trees; maximum depth; minimum leaf	trees = 100; depth = 5; min leaf = 1
Bagging	Extra Trees	Number of trees; maximum depth; minimum leaf	trees = 200; depth = 10; min leaf = 2
Boosting	Gradient boosting	Number of estimators; learning rate; tree depth	estimators = 100; learning rate = 0.05; depth = 2
Boosting	AdaBoost	Number of estimators and learning rate	estimators = 50; learning rate = 0.05
Probabilistic	Gaussian naive Bayes	Variance smoothing grid	var smoothing = 1e-09
Neural network	MLP	Hidden layer size; L2 regularization alpha; initial learning rate	hidden layer = (64,); alpha = 0.0001; learning rate = 0.001
Gradient boosting	XGBoost	Number of estimators; depth; learning rate; subsample	estimators = 100; depth = 2; learning rate = 0.05; subsample = 1.0
Gradient boosting	LightGBM	Number of estimators; learning rate; leaves; depth	estimators = 200; learning rate = 0.1; leaves = 31; depth = -1
Gradient boosting	CatBoost	Number of iterations; depth; learning rate	iterations = 100; depth = 5; learning rate = 0.05
Final meta-learner	Pruned logistic regression	Regularization strength C; L2 penalty; liblinear solver; class weight	C = 0.01; class weight = None; penalty = L2; solver = liblinear

Model development followed four connected stages. Phase 1 established the single-model baseline: every tuned base classifier learned from the same eight standardized biomarker predictors. Phase 2 asked whether stacking could improve the class-wise trade-off by adding OOF probability features generated from fold-specific base-model predictions. Phase 3 tested whether weighted probability fusion and threshold movement could shift the bacterial-versus-viral operating point. The final refinement was added after those stages to test compactness: it kept the original eight biomarkers but retained probability outputs only from the strongest OOF recall sources. The held-out set remained outside feature selection, hyperparameter tuning, OOF ranking, weight search, threshold selection, and pruned-model construction. Table 9 summarizes the role of each stage.

Table 9: Multi-Phase Development Protocol

Stage	Purpose	Main input	Held-out use
Phase 1	Train tuned base classifiers.	Eight standardized predictors.	Reporting only.
Phase 2	Train full OOF stacking meta-learners.	Eight predictors plus 26 OOF probability features.	Reporting only after configuration locking.
Phase 3	Explore weighted probability fusion and internal threshold search.	Selected probability outputs from development workflow.	Reporting only after weight and threshold locking.
Final refinement	Test compact recall-oriented pruned stacking.	Eight predictors plus eight selected probability features from four base models.	Reporting only after final locking.

Weighted Aggregation and Pruned Stacking

The weighted probability fusion stage utilized a development-only grid search. A normalized grid with step size $\delta = 0.25$ allowed each candidate model to receive one of five possible weights: 0, 0.25, 0.50, 0.75, or 1.00. We accepted a candidate weight vector only when all selected weights summed to one. For example, a two-source combination such as (0.75, 0.25) is valid, whereas (0.75, 0.75) is invalid because it sums to 1.50. We chose a coarse step size deliberately to maintain an auditable search space and avoid presenting a finely tuned weight vector as a stable clinical ideal. We also evaluated candidate operating thresholds internally rather than using the held-out test set. This measured how probability aggregation and threshold adjustments changed the balance between bacterial and viral classification, rather than establishing weighted fusion as the permanent configuration.

The final pruned stacking stage followed the out-of-fold (OOF) ranking generated during Phase 1. This progression linked the model-family search directly to the four selected pruned probability sources. We ranked base classifiers using training-only OOF bacterial recall at a default probability threshold of 0.50, utilizing the bacterial F_2 -score, Matthews correlation coefficient (MCC), viral recall, and $ROC-AUC$ as secondary tie-breaking metrics. The held-out test set was entirely excluded from this ranking process. Table 10 details the four selected top-performing sources. We selected the multilayer perceptron (MLP), support vector machine with a radial basis function kernel (SVM-RBF), K -nearest neighbors (KNN), and AdaBoost within the development workflow prior to held-out reporting because they demonstrated the strongest OOF bacterial recall.

Table 10: OOF Ranking Used to Select Pruned Probability Sources

OOF rank	Selected source	Bact. recall	Viral recall	Bact. F2	MCC	Bact. FN	Held-out used?
1	MLP	0.9895	0.8767	0.9699	0.8759	5	No
2	SVM-RBF	0.9706	0.9070	0.9601	0.8816	14	No
3	KNN	0.9601	0.9023	0.9509	0.8658	19	No
4	AdaBoost	0.9580	0.9163	0.9516	0.8764	20	No

Note: The ranking used OOF predictions from the development data at threshold 0.50. It explains why the final pruned stack used these four base models before held-out reporting.

Each selected model generated two class-probability features, representing the bacterial and viral classes. Consequently, the final pruned feature matrix paired the original eight biomarker predictors with these eight selected probability features, yielding a consolidated 16-feature representation. We trained a logistic regression meta-learner inside the internal development workflow to prioritize traceability over maximal mathematical flexibility. The chosen operating threshold was set at 0.61 and remained unaltered after held-out evaluation.

A secondary stage-level score was defined to summarize the trade-off between bacterial recall, viral recall, bacterial F2-score, MCC, and ROC-AUC. The score used the formula

$$S = 0.40R_B + 0.20R_V + 0.20F2_B + 0.15MCC + 0.05AUC,$$

where R_B is bacterial recall, R_V is viral recall, $F2_B$ is bacterial F2-score, MCC is Matthews correlation coefficient, and AUC is ROC-AUC. This score was used only to summarize trade-offs after locked evaluation; it was not used to tune the held-out set.

Evaluation Metrics

The primary evaluation metric was bacterial recall. The primary count-based error was bacterial false negative, defined as a bacterial case predicted as viral. Viral recall and viral false negatives were reported to prevent a one-sided interpretation. Bacterial precision and bacterial F2-score were included because the objective is recall-oriented but should not ignore precision. MCC was reported because it uses all four confusion-matrix cells and is helpful when class-wise errors differ (Chicco & Jurman, 2020). With the usual binary notation, MCC is defined as

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

ROC-AUC was computed from the predicted viral probability because viral was the encoded positive class.

The confusion-matrix convention was fixed before evaluation: true bacterial predicted

as bacterial, true bacterial predicted as viral, true viral predicted as bacterial, and true viral predicted as viral. For MCC calculation, the same binary confusion-matrix cells are used consistently with the encoded class convention, even though the interpretation in this chapter names the two clinical directions directly. This convention avoids ambiguity when bacterial and viral false negatives are discussed.

RESULTS AND DISCUSSION

Dataset Cleaning and Feature Eligibility Results

The binary dataset contained 1,133 confirmed cases from the original 1,200 records after cleaning the target variable. The final class distribution was almost balanced, which consists of 595 bacterial cases and 538 viral cases. During the variable audit process, `Patient_ID`, `Gender`, `Pathogen_Present`, `Outcome`, and `Risk_Level` were removed from the final feature matrix, while `Diagnosis` was retained only as the target variable. As a result, the final modelling matrix included eight predictors that would typically be available at an early stage, which are `Age`, `WBC_Count`, `Protein_Level`, `Glucose_Level`, `Hemoglobin`, `WBC_Blood_Count`, `Platelets`, and `CRP_Level`.

The audit, statistically, highlighted that it supported these decisions. The normality tests and Levene's test indicated that all numerical variables were not normally distributed and the variances between bacterial and viral groups were not homogeneous. Consequently, Spearman correlation was used to explore correlations between numerical variables rather than parametric methods. The categorical variable audit also illustrated why statistical association alone was insufficient to justify retention of a variable. While `Risk_Level` and `Pathogen_Present` had strong associations with `Diagnosis`, they were not suitable for an early biomarker prediction model. Information such as this would not normally be available to clinicians at the initial screening stage. Thus, the reason for the exclusion and removal of these variables was based on their practical role rather than statistical significance.

The Spearman correlation results show association patterns, not causal relationships. `WBC_Blood_Count`, `WBC_Count`, `CRP_Level`, and `Protein_Level` were increased in bacterial cases, which is clinically plausible as bacterial meningitis is often associated with more robust inflammatory and white blood cell responses (Ahmed et al., 2022; Baran et al., 2025; Groeneveld et al., 2024; Singh et al., 2025). Higher `Glucose_Level` values were associated with viral cases in this dataset, which is in agreement with the clinical knowledge that bacterial meningitis decreases cerebrospinal fluid glucose levels. This finding should not be taken as evidence that glucose alone can identify viral meningitis. In general, these patterns justify the variable selection choices made in the audit and offer a cautious interpretation of the public dataset from cleaning.

Phase 1 Base Classifier Results

Phase 1 evaluated tuned base classifiers on the held-out test set after training-only tuning. The five highest bacterial-recall models are reported below. MLP achieved the highest bacterial recall, 0.9832, with two bacterial false negatives. Its viral recall was lower, 0.8519, corresponding to 16 viral false negatives. This result establishes the main trade-off in the chapter: the strongest bacterial-recall model was not the most balanced class-wise model.

Table 11: Top Phase 1 Held-Out Base Classifier Results

Model	Accuracy	Bact. recall	Viral recall	Bact. precision	Bact. F2	MCC	Bact. FN	Viral FN
MLP	0.9207	0.9832	0.8519	0.8797	0.9606	0.8467	2	16
SVM-RBF	0.9251	0.9748	0.8704	0.8923	0.9571	0.8532	3	14
AdaBoost	0.9251	0.9580	0.8889	0.9048	0.9468	0.8511	5	12
Gradient Boosting	0.9295	0.9412	0.9167	0.9256	0.9380	0.8587	7	9
XGBoost	0.9295	0.9412	0.9167	0.9256	0.9380	0.8587	7	9

Full Stacking, Weighted Fusion, and Final Pruned Stacking

The full stacking input matrix in Phase 2 was constructed by combining the OOF probability features of the base classifiers and the eight standardized predictors. The best model at this stage was the Meta-Neural Network with accuracy 0.9339, bacterial recall 0.9664, viral recall 0.8981, bacterial F2-score 0.9552 and MCC 0.8688. The Meta-NN also reduced the viral false negatives from 16 in the Phase 1 Multilayer Perceptron to 11, but increased bacterial false negatives slightly from two to four. This shift is the main trade-off, in which gains on the viral-side recovery tend to come at some cost to bacterial recall, and the Meta-NN struck a more even balance in managing that trade-off than the best Phase 1 model.

In phase 3, a trial of weighted probability fusion was done as a way to move the operating point further. Viral detection increased compared to the MLP but bacterial recall decreased compared to the best configurations from Phases 1 and 2. The phase achieved an accuracy of 0.9251, bacterial recall of 0.9328, viral recall of 0.9167, bacterial F2-score of 0.9312, and MCC of 0.8498 with 8 bacterial false negatives and 9 viral false negatives. This stage is best seen as a controlled sensitivity experiment rather than a preferred configuration. It shows how probability aggregation and threshold movement can shift the operating point between bacterial and viral, but the resulting bacterial recall does not meet the main goal of the study.

The final pruned stack exactly reproduced the held-out error profile of the Phase 2 Meta-NN. The model gave an accuracy of 0.9339, a bacterial recall of 0.9664, a viral recall of 0.8981, a bacterial F2-score of 0.9552, an MCC of 0.8688, four bacterial false negatives, and eleven viral false negatives. However, the model does not perform better than the full stacking configuration. But, it is compact. The full stack used eight biomarkers combined

with 26 probability features, while the pruned stack achieved the same holdout results with eight biomarkers but only eight selected probability features. This is important, because a simpler and more traceable representation is easier to audit, and more interpretable, even if the classification counts themselves are unchanged.

Table 12: Best Model per Stage on the Held-Out Test Set

Stage	Model	Accuracy	Bact. recall	Viral recall	Bact. F2	MCC	ROC-AUC	Bact. FN	Viral FN
Phase 1	MLP	0.9207	0.9832	0.8519	0.9606	0.8467	0.9483	2	16
Phase 2	Meta-NN	0.9339	0.9664	0.8981	0.9552	0.8688	0.9625	4	11
Phase 3	Locked weighted ensemble	0.9251	0.9328	0.9167	0.9312	0.8498	0.9890	8	9
Final refinement	Pruned stacking	0.9339	0.9664	0.8981	0.9552	0.8688	0.9603	4	11

Figure 4 places the same held-out trade-off in visual form. The left panel shows that Phase 1 MLP kept the highest bacterial recall but had the lowest viral recall among the four representatives. The right panel shows the same trade-off as false-negative counts. This figure helps connect the phase-by-phase results with the recall-oriented interpretation rather than leaving the comparison only in table form.

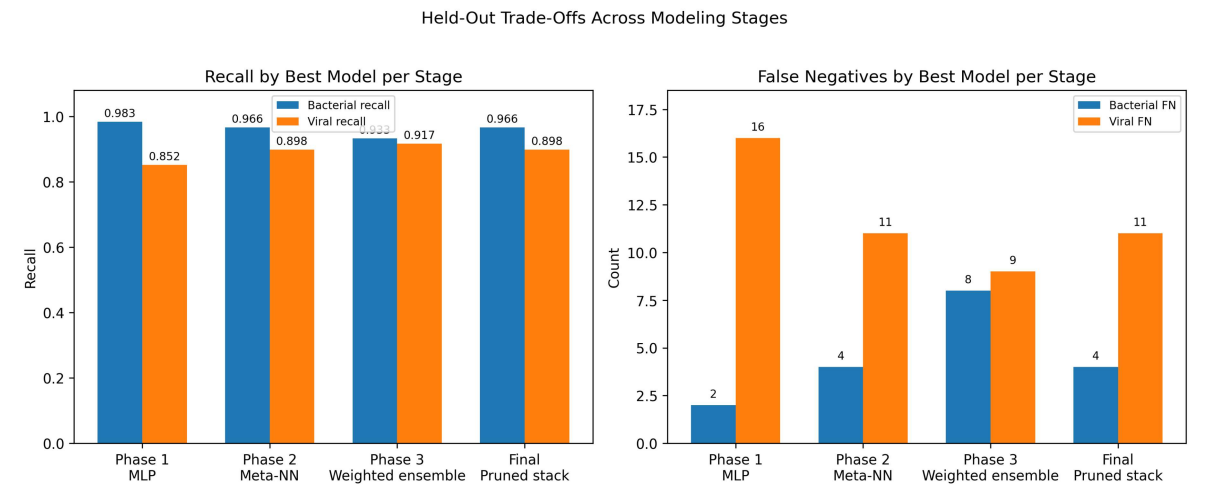


Figure 4: Held-out recall and false-negative trade-offs for the best representative model in each modeling stage.

Weighted Aggregation Summary

The weighted aggregation summary combines the main performance metrics into a descriptive score. This score is not meant to replace the individual metrics reported across stages nor was it used to tune or adjust the models at any point. The only purpose is to align the stage representatives side by side under the same weighting scheme focusing on recall . This makes the overall tradeoff easier to read at a glance.

Table 13: Secondary Weighted Aggregation Score for Representative Models

Model	Bact. recall	Viral recall	Bact. F2	MCC	ROC-AUC	Weighted score
MLP	0.9832	0.8519	0.9606	0.8467	0.9483	0.9302
Meta-NN	0.9664	0.8981	0.9552	0.8688	0.9625	0.9357
Locked weighted ensemble	0.9328	0.9167	0.9312	0.8498	0.9890	0.9196
Final pruned stacking	0.9664	0.8981	0.9552	0.8688	0.9603	0.9356

Note: The score is $0.40R_B + 0.20R_V + 0.20F2_B + 0.15MCC + 0.05AUC$. It is reported for interpretation only and was not used to modify the locked held-out evaluation.

The combined score is yet another way of looking at the same results. As seen from the bacterial recall, the Multilayer Perceptron model is still the best model for the bacterial case classification. However, the weighted score prefers models scoring high on bacterial detection while retrieving more viral cases and having a better Matthews Correlation Coefficient (MCC) score value. By form of measurement, the Meta-Neural Network and the final pruned stacking model end up almost tied. Both models had the same classification results on the held out test data with only minor differences in their ROC-AUC values. This near-equivalence demonstrates that the pruned stack does not improve upon the Meta-NN from a strict classification perspective. Instead, the pruned configuration achieves the identical error profile while utilizing a substantially leaner architecture, highlighting its practical efficiency and traceability.

Final Pruned Confusion Matrix and Interpretation

The final pruned stack produced 115 true bacterial predictions, four bacterial false negatives, eleven viral false negatives, and 97 true viral predictions. The bacterial recall was therefore $115/(115 + 4) = 0.9664$, while viral recall was $97/(97 + 11) = 0.8981$. The four bacterial false negatives are the main error count for the recall-oriented objective.

Table 14: Final Pruned Stacking Held-Out Confusion Matrix

True label	Predicted label	Count	Interpretation
Bacterial	Bacterial	115	True bacterial predicted as bacterial.
Bacterial	Viral	4	Bacterial false negative.
Viral	Bacterial	11	Viral false negative.
Viral	Viral	97	True viral predicted as viral.

The final pruned result is useful because it answers a different question from the Phase 1 ranking. The main question for the pruned stage is whether the full stacking representation can be simplified without changing the held-out error profile of the best

full stack. In this dataset, the answer is yes. The final pruned stack matched Meta-NN in accuracy, bacterial recall, viral recall, bacterial F2-score, MCC, and false-negative counts while using fewer probability sources.

Practical Interpretation and Limitations

This study describes a per-step approach to classify meningitis, taking care to avoid data leakage at each stage of the analytical pipeline. The final dataset comprised 1,133 confirmed cases after excluding cases with unknown diagnostic labels. Raw variables were cleaned by removing patient identifiers, post-admission information, undocumented derived labels, and variables that were near-direct proxies for the confirmed diagnosis. The other predictors did not present normal and homogeneous distribution in all the variables studied, so, for numerical variables, the correlation was made through Spearman correlation and for categorical variables through chi-square and Cramer’s V, instead of using parametric association methods that the data did not support.

The performance on this approach was tested on a number of configurations, all of which were evaluated on a completely separate holdout set, never used for training, tuning, or feature selection. The configurations were: tuned base classifiers, full out-of-fold stacking, weighted probability fusion, and a simplified pruned stacking method. The Phase 1 Multilayer Perceptron was best at detecting bacterial cases (missing only two), but had similar reliability in detecting viral cases. The Meta-Neural Network and the pruned stacking model presented a more balanced profile, missing four bacterial and eleven viral cases. The pruned stack was exactly matching the full Meta-NN error counts, yet using a considerably simpler 16-feature setup, making it the more traceable option when compactness is a practical priority.

However, these results should be interpreted with caution by readers. The principal practical lesson is a conditional one, not an absolute one. The only goal in this holdout set is the minimization of bacterial false negatives, but Phase 1 MLP is still the most powerful configuration. If a more balanced bacterial-versus-viral screen is desired, the Meta-NN and pruned stacking configurations are more appropriate, as they retain bacterial recall above 0.96 while significantly improving viral recall. If compactness and traceability are also important, the pruned stack offers the clearest stacking result while not sacrificing held-out classification performance.

Although, several factors limit the generalizability of these findings. First, the holdout set used here is an internal partition of a single public dataset, and not an independent hospital cohort, so the performance estimates do not represent external validation. Second, the pruned probability sources and operating threshold were selected within the development workflow and should not be considered the globally optimal cutoffs. Third, the dataset does not consider all clinically relevant variables or timing of

measurements or context of data collection in the real world. Fourth, model calibration and decision-curve analysis are beyond the scope of the present work and remain important open questions for future work. Such models are thus not yet ready for direct clinical use. Hospitals cannot responsibly adopt this or any similar workflow for patient-level decisions until it has been subjected to rigorous external validation or prospective clinical evaluation in multiple settings and patient populations.

CONCLUSION AND FUTURE RESEARCH

The present workflow provides a step-by-step methodology for meningitis classification based on strict leakage prevention at every stage of the analytical pipeline. The final dataset consisted of 1,133 confirmed binary cases after removing those with unknown diagnostic labels. Systematic cleaning of raw variables was carried out involving deletion of patient identifiers, information after admission, undocumented derived labels and variables which were near-direct proxies for the confirmed diagnosis. We performed statistical screening using Spearman correlation for numerical variables and chi-square with Cramer’s V for categorical variables. This was motivated by the non-normal, non-homogeneous distributional properties of all predictors that were retained.

We evaluated performance by comparing standard base classifiers, full out-of-fold (OOF) stacking, weighted probability fusion and a simplified “pruned” stacking method, all tested on a strictly separate holdout dataset. The Phase 1 MLP model was best at detecting bacterial cases (missing only two, false negatives), but had difficulty recall of viral cases. The overall balance was better for the Meta-Neural Network (Meta-NN) and for the pruned stacking models with only missing four bacterial and 11 viral cases. Interestingly, the pruned stack did not outperform the full Meta-NN, but got the same error rates with a much simpler 16-feature setup. But readers should take these results with a grain of salt.

Although this workflow enables easy model comparisons and facilitates the stacking approach for this specific dataset, these models are not ready for hospital use. Future studies need to use this strict, leak-free method on outside groups of patients. Moreover, the assessment of model calibration, testing of specific clinical cost thresholds, and the generalizability of these reduced predictive features to different hospitals and patient populations should be performed. Finally, this model could not be used by hospitals for real clinical decisions until it passes external validation or real-time clinical trials.

AUTHOR CONTRIBUTIONS

Pasya Kemal Halim, Muhammad Fatih Rizki Herianto, Ratu Azalea Meiza Asri, and Dani Suandi contributed to the study conceptualization, computational experimentation,

analysis interpretation, and manuscript preparation.

DATA AVAILABILITY

The study used the public Kaggle meningitis classification dataset available at <https://www.kaggle.com/datasets/chantest/meningitis-classification>. The processed analysis artifacts include target-cleaning summaries, variable-role audits, statistical assumption outputs, model comparisons, and held-out evaluation tables generated during the computational workflow.

ACKNOWLEDGMENT

The authors acknowledge the School of Computer Science, Universitas Bina Nusantara, Jakarta, Indonesia, for academic support during the development of this work.

AUTHOR BIOGRAPHIES

Pasya Kemal Halim: Author biography to be completed before submission.

Muhammad Fatih Rizki Herianto: Author biography to be completed before submission.

Ratu Azalea Meiza Asri: Author biography to be completed before submission.

Dani Suandi: Author biography to be completed before submission.

REFERENCES

- Ahmed, M., Askar, G. A., Farghaly, H., Ahmed, A., Kamal, D. T., Ahmed, S. S., & Mohamad, I. L. (2022). Accuracy of cerebrospinal fluid C-reactive protein and multiplex polymerase chain reaction and serum procalcitonin in diagnosis of bacterial and viral meningitis in children. *Acta Neurologica Taiwanica*, 31(2), 61-71. <https://consensus.app/papers/details/b873650d4e3a5a95a8edbad882f40af3>
- Andaur Navarro, C. A., Damen, J., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G., Bajpai, R., Riley, R., Moons, K., & Hooft, L. (2021). Completeness of reporting of clinical prediction models developed using supervised machine learning: A systematic review. *BMC Medical Research Methodology*, 22. <https://consensus.app/papers/completeness-of-reporting-of-clinical-prediction-models-navarro-damen/0d31699783a0551d877450f12dd0187a/>

- Arafa, Z. A. A., Gabr, M. S., Kamel, E., Elmasry, S., & Fahim, N. A. E. (2023). Cerebrospinal fluid lactate as a differential biomarker for bacterial and viral meningitis. *The Egyptian Journal of Immunology*, 30(3), 148-161. <https://consensus.app/papers/cerebrospinal-fluid-lactate-as-a-differential-biomarker-arafa-gabr/37dba25af6745f8ab4ca6b84ff2d189a/>
- Babenko, D., Seidullayeva, A., Bayesheva, D., Turdalina, B., Omarkulov, B., Almaybayeva, A., Zhanaliyeva, M., Kushugulova, A., & Kozhakhmetov, S. (2021). Ability of procalcitonin and C-reactive protein for discriminating between bacterial and enteroviral meningitis in children using decision tree. *BioMed Research International*, 2021. <https://consensus.app/papers/details/b22e693d56b851948378fd814d251831>
- Baran, J., Porebska, A., Lesisz, M., Polak, K., Grodzka, O., & Domitrz, I. (2025). Investigating blood biomarkers that can facilitate the diagnosis of meningitis: A systematic literature review. *International Journal of Molecular Sciences*, 26. <https://consensus.app/papers/investigating-blood-biomarkers-that-can-facilitate-the-baran-porebska/12d3b203a74e5a36abb8a99cd19a9516/>
- Cheah, B. C. J., Vicente, C. R., & Chan, K. R. (2025). Machine learning and artificial intelligence for infectious disease surveillance, diagnosis, and prognosis. *Viruses*, 17. <https://consensus.app/papers/machine-learning-and-artificial-intelligence-for-cheah-vicente/0233b57d47d652a391975fec5d7f6f8a/>
- Choi, B., Choi, Y., Sung, M., Ha, W., Chu, M., Kim, W.-J., Heo, K., Kim, K., & Park, Y. (2023). Development and validation of an artificial intelligence model for the early classification of the aetiology of meningitis and encephalitis: A retrospective observational study. *eClinicalMedicine*, 61. <https://consensus.app/papers/development-and-validation-of-an-artificial-intelligence-choi-choi/af9f7c05cdc55b65b7f9077ea4b09a40/>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21. <https://consensus.app/papers/the-advantages-of-the-matthews-correlation-coefficient-chicco-jurman/199e36dd2263566698ad994f67f61887>
- Collins, G., Dhiman, P., Andaur Navarro, C. A., Ma, J., Hooft, L., Reitsma, J., Logullo, P., Beam, A., Peng, L., Van Calster, B., van Smeden, M., Riley, R., & Moons, K. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, 11. <https://consensus.app/papers/details/c4c17eb44f62591a85310acfa0c6d17a>

- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., Hoffman, M. M., Kengne, A. P., Lam, E., Lee, N., Loder, E. W., Maier-Hein, L., Mateen, B. A., McCradden, M. D., Oakden-Rayner, L., Ordish, J., Parnell, R., Rose, S., Singh, K., Wynants, L., & Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *The BMJ*, 385. <https://consensus.app/papers/tripodai-statement-updated-guidance-for-reporting-collins-moons/34463dae28385dcf915dfb79a563a292/>
- Cross, J. L., Choma, M., & Onofrey, J. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, 3. <https://consensus.app/papers/details/b215b3f4b94d5d60b7467138f5b8e3b5>
- D'Agostino, R. B., & Pearson, E. S. (1973). Tests for departure from normality: Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika*, 60(3), 613-622. <https://doi.org/10.1093/biomet/60.3.613>
- Ghaddaripouri, K., Ghaddaripouri, M., Mousavi, A. S., Mousavi Baigi, S. F., Rezaei Sarsari, M., Dahmardeh Kemmak, F., & Mazaheri Habibi, M. R. (2024). The effect of machine learning algorithms in the prediction, and diagnosis of meningitis: A systematic review. *Health Science Reports*, 7. <https://consensus.app/papers/the-effect-of-machine-learning-algorithms-in-the-ghaddaripouri-ghaddaripouri/4a549924aa3251f6bba0859f6ae363d5/>
- Groeneveld, N. S., Bijlsma, M., van de Beek, D., & Brouwer, M. (2024). Biomarkers in paediatric bacterial meningitis: A systematic review and meta-analysis of diagnostic test accuracy. *Clinical Microbiology and Infection*. <https://consensus.app/papers/details/aacffeece3dd5567a560d4fb94151c4f>
- Guzman, E., Belmonte, M.-V., & Lelis, V. M. (2022). Ensemble methods for meningitis aetiology diagnosis. *Expert Systems*, 39. <https://consensus.app/papers/details/13e93b4784d85fc9b9adea91d3a442f2>
- Kim, H., Roh, Y., & Yoon, S. (2021). Blood procalcitonin level as a diagnostic marker of pediatric bacterial meningitis: A systematic review and meta-analysis. *Diagnostics*, 11. <https://consensus.app/papers/details/9d01320230bf5f06ac832eaf0392e3b9>
- Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278-292). Stanford University Press.
- Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble learning for disease prediction: A review. *Healthcare*, 11. <https://consensus.app/papers/details/edc1aa7e932558739b27972d366d6789>

- Mentis, A., Garcia, I., Jimenez, J., Paparoupa, M., Xirogianni, A., Papandreou, A., & Tzanakaki, G. (2021). Artificial intelligence in differential diagnostics of meningitis: A nationwide study. *Diagnostics*, 11. <https://consensus.app/papers/details/954f3a17e2de566aa90efd3639cc9b28>
- Norman-Bruce, H., Umana, E., Mills, C., Mitchell, H., McFetridge, L., McCleary, D., & Waterfield, T. (2024). Diagnostic test accuracy of procalcitonin and C-reactive protein for predicting invasive and serious bacterial infections in young febrile infants: A systematic review and meta-analysis. *The Lancet Child & Adolescent Health*. <https://consensus.app/papers/details/8d884514336f56d4a7dd4839c9a29caa>
- Olie, S. E., Staal, S. L., Groeneveld, N. S., Schotman, H. H., Bodilsen, J., Nielsen, H., Bijlsma, M., van de Beek, D., & Brouwer, M. (2025). Validation and clinical implementation of cerebrospinal fluid C-reactive protein for the diagnosis of bacterial meningitis: A prospective diagnostic accuracy study. *The Lancet Regional Health - Europe*, 53. <https://consensus.app/papers/details/d18336f97584537eade7b5825488bf6>
- Sathiya Priya, S., Amanullah, M., & Kujani, T. (2024). Enhancing meningitis type diagnosis through a novel hybrid ensemble of bagging and gradient boosting. *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, 1260-1264. <https://consensus.app/papers/details/f38c25431f785fae9f48b18d826194d4>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591-611. <https://doi.org/10.2307/2333709>
- Singh, S., Mahto, K., Kumar, A., Kumar, P., Kumar, S., & Prasad, M. (2025). Diagnostic test accuracy of serum and cerebrospinal fluid C-reactive protein in bacterial meningitis: A systematic review and meta-analysis. *Infection & Chemotherapy*, 57, 248-260. <https://consensus.app/papers/diagnostic-test-accuracy-of-serum-and-cerebrospinal-fluid-singh-mahto/3444e65eb9725bc99a3f26190b871779/>
- Smiley, A., Villarreal-Zegarra, D., Reategui-Rivera, C. M., Escobar-Agreda, S., Finkelstein, J., Pine, S. R., & Kapoor, D. U. (2025). Methodological and reporting quality of machine learning studies on cancer diagnosis, treatment, and prognosis. *Frontiers in Oncology*, 15. <https://consensus.app/papers/details/91a03821077e52faba02dab0713ba4f1>
- Theodosiou, A., & Read, R. (2023). Artificial intelligence, machine learning and deep learning: Potential resources for the infection clinician. *The Journal of Infection*. <https://consensus.app/papers/details/d080c40bb8cd5d5ab034f417b82a29a1>
- Victor, A., Santos, D. A. M., Nery, E. K., Mori, D. P., Lucas, P. C. C., Cammarota, D., Montero, G. L. F., Barcellos Filho, F. N., Yu, A. L. F., & Carvalhanas, T. (2025). Improving meningitis surveillance and diagnosis with machine learning: Insights from Sao Paulo. *PLOS Digital Health*, 4. <https://consensus.app/papers/details/>

bc7384b324de5fe29aee845dbe1e016b