A Hybrid Random Forest–Logistic Regression–Naive Bayes Ensemble with Soft Voting for Breast Cancer Recurrence Prediction

Reginald Ekene Ogu

Department of Computer Engineering, Federal University of Technology, Owerri (FUTO), Nigeria

DOI: https://doi.org/10.51244/IJRSI.2026.1315PH00193

Received: 10 October 2026; Accepted: 15 October 2026; Published: 29 October 2026

ABSTRACT

Predicting breast cancer recurrence from routine clinical records is hard when the data are few, noisy, and largely categorical. This study builds a three-way hybrid ensemble that combines Random Forest, which captures nonlinear interactions among clinical variables; Logistic Regression, which offers a linear, interpretable decision boundary; and Bernoulli Naive Bayes, which suits sparse one-hot encoded attributes. Their predicted probabilities are merged by soft voting. The models were evaluated on the Ljubljana breast cancer dataset (286 patients, nine categorical attributes) using most-frequent-value imputation, one-hot encoding, a stratified held-out test split, and stratified 5-fold cross-validation. Under cross-validation, Naive Bayes was the strongest single learner (accuracy 0.727, ROC-AUC 0.701). The hybrid came close but did not surpass it on either measure (0.724, 0.680); its advantage lay in the highest F1-score of any model tested (0.528) and the smallest fold-to-fold variation in accuracy. Adding Naive Bayes to an earlier two-model Random Forest–Logistic Regression hybrid improved accuracy, ROC-AUC, and F1. A randomized search over all three learners and the voting weights nudged ROC-AUC to 0.684 but cut F1 to 0.424, repeating the precision-recall trade-off seen in the two-model study and suggesting that tuning for ROC-AUC is a poor substitute for an objective that reflects the clinical cost of a missed recurrence. The hybrid's accuracy sits within the 65–78% range historically reported for this dataset. On this evidence, adding a structurally different learner did more for the ensemble than tuning an existing one, a useful pointer to where effort pays off in small-sample clinical prediction.

Keywords: Breast Cancer Recurrence; Hybrid Ensemble; Random Forest; Logistic Regression; Naive Bayes; Soft Voting; Machine Learning; Healthcare Informatics

INTRODUCTION

Breast cancer remains one of the most diagnosed cancers worldwide (Bray et al., 2024), and predicting whether a treated patient will experience a recurrence is central to post-treatment surveillance and resource planning (Lu et al., 2023; Mazo et al., 2022). Clinical recurrence prediction, however, rarely has the luxury of large, clean, continuously-measured datasets. Records collected in ordinary clinical practice are often categorical or binned (age ranges, tumor-size brackets, nodal-involvement ranges rather than exact counts), incomplete, and modest in number, especially for single-institution studies. Models built for this setting need to be robust to that reality rather than assume the abundance of data typical of benchmark machine learning tasks.

Three model families are commonly reached for in this kind of tabular clinical prediction problem, each for a different reason. Random Forests handle nonlinear interactions and mixed categorical inputs without much preprocessing effort, and they are comparatively resistant to overfitting on small samples because of bagging and feature subsampling (Breiman, 2001; Grinsztajn et al., 2022). Logistic Regression produces a linear, interpretable decision boundary whose probability outputs are comparatively easy to check for calibration — a property that matters in a clinical context where a model's reasoning may need to be explained to a physician or an ethics board (Christodoulou et al., 2019; Hosmer et al., 2013; Van Calster et al., 2019).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI | Volume XIII Issue XV October 2026 | Special Issue on Public Health

Naive Bayes classifiers, though built on a conditional-independence assumption that is rarely literally true of clinical attributes, are known to perform competitively on small, high-dimensional, sparsely-populated categorical data precisely because they estimate far fewer parameters than either of the other two approaches, which makes them less prone to overfitting when the training set is small (Rish, 2001; Wickramasinghe & Kalutarage, 2021). Combining the three through a soft-voting ensemble, which averages predicted class probabilities rather than hard labels, is intended to let each model contribute where it is strongest, and to let their different failure modes partially cancel out (Sagi & Rokach, 2018).

This paper builds and evaluates such a three-way hybrid classifier on a well-known benchmark, the Ljubljana Institute of Oncology breast cancer dataset, deliberately chosen because it is small, categorical, and moderately class-imbalanced — the conditions this hybrid approach is meant to address, and also the conditions under which claims of large accuracy gains from ensembling should be treated with some suspicion. The specific contributions of the paper are fourfold: (1) a complete, reproducible preprocessing-to-evaluation pipeline for the dataset, including imputation and one-hot encoding; (2) a systematic comparison of Random Forest, Logistic Regression, Naive Bayes, and the soft-voting hybrid, using both a held-out test split and 5-fold cross-validation; (3) a hyperparameter optimization study that tunes all three base learners and the relative voting weights jointly, together with an honest account of what that tuning did and did not improve; and (4) a comparison of the size of the gain obtained by adding a third, structurally different base learner against the size of the gain obtained by hyperparameter tuning alone, which turns out to favor the former.

Related Work

The Ljubljana breast cancer dataset has served as a benchmark since it was first used to evaluate the AQ15 rule-induction system, which reported roughly 66-72% classification accuracy on the recurrence task depending on the domain-specific tuning applied (Michalski et al., 1986). Clark and Niblett (1987), working on rule induction in noisy domains, reported results in a similar 65-72% range, while the Assistant-86 decision-tree system of Cestnik et al. (1987) reached about 78%. Across the classical machine learning literature that has used this dataset as a testbed, reported accuracies cluster between 65% and 78% (Zwitter & Soklic, 1988), a ceiling that later work has generally attributed to genuine noise in the recurrence labels and the coarse, binned nature of the input attributes rather than to weaknesses in any particular algorithm. Recent systematic reviews have catalogued a growing body of machine learning models for recurrence prediction, most of them built on larger and richer clinical cohorts than this one (Lu et al., 2023; Mazo et al., 2022); a meta-analysis of 34 such studies reported a pooled validation c-index of about 0.77 but concluded that no model was yet reliable enough for broad clinical use (Lu et al., 2023).

Ensemble methods more broadly are motivated by the observation that combining multiple imperfect models can reduce variance and improve generalization beyond what any single model achieves, provided the base learners make sufficiently uncorrelated errors (Dietterich, 2000; Sagi & Rokach, 2018). Random Forest itself is an ensemble of decision trees built through bootstrap aggregation and random feature selection, and it has become a standard baseline for tabular classification because of its resistance to overfitting and minimal need for feature scaling (Breiman, 2001; Grinsztajn et al., 2022). Naive Bayes classifiers assume conditional independence of features given the class, an assumption that rarely holds exactly but which empirically causes little harm in practice and can even help on small samples because it constrains the model to estimate far fewer parameters than a fully general classifier would (Rish, 2001; Wickramasinghe & Kalutarage, 2021). Despite its simplicity, Naive Bayes has repeatedly been shown to be competitive with, and sometimes superior to, more sophisticated classifiers on small or sparsely-populated categorical datasets, which motivates its inclusion here alongside two considerably more complex model families. Soft voting, which averages predicted probabilities across heterogeneous base learners rather than combining hard votes, is a natural extension of the ensembling idea to structurally different model types, and has seen use in medical prediction tasks (Sherazi et al., 2021) because it lets a well-calibrated probabilistic model correct for the overconfidence that tree ensembles and generative classifiers can each show near decision boundaries, in different ways.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI | Volume XIII Issue XV October 2026 | Special Issue on Public Health

The present work differs from most prior applications of this dataset in three respects. First, it evaluates a genuinely heterogeneous three-way hybrid — a tree ensemble, a linear discriminative model, and a generative probabilistic model, combined by soft voting — rather than a single algorithm family or a homogeneous ensemble (bagged trees, boosted trees). Second, it treats hyperparameter optimization as an object of study, reporting what tuning changes about the trade-off between accuracy and recall rather than reporting only the single best configuration found. Third, it directly compares the performance gain obtained from architectural diversity (adding a third, differently-structured base learner) against the gain obtained from hyperparameter tuning of an existing configuration, a comparison that is often implicit but rarely made explicit in ensemble learning studies on small clinical datasets.

MATERIALS AND METHODS

Dataset

The dataset used is the breast cancer recurrence dataset donated by M. Zwitter and M. Soklic of the University Medical Center, Institute of Oncology, Ljubljana, and distributed through the UCI Machine Learning Repository (Dua & Graff, 2019; Zwitter & Soklic, 1988). It contains 286 patient instances, described by nine categorical predictor attributes and a binary class label. Table 1 summarizes the attributes and their value ranges.

Attribute	Type	Values	Missing
class (target)	binary	no-recurrence-events (201), recurrence-events (85)	0
age	binned range	20-29 through 70-79 (6 bins)	0
menopause	categorical	premeno, ge40, lt40	0
tumor-size	binned range	0-4 mm through 50-54 mm (11 bins)	0
inv-nodes	binned range	0-2 through 24-26 (7 bins)	0
node-caps	binary	yes, no	8
deg-malig	ordinal grade	1, 2, 3	0
breast	binary	left, right	0
breast-quad	categorical	left-up, left-low, right-up, right-low, central	1
irradiat	binary	yes, no	0

Table 1. Dataset attributes, value ranges, and missing-value counts (n = 286).

The target class is imbalanced, with 70.3% of patients showing no recurrence and 29.7% showing recurrence within the observation period. Nine missing values are present in total, coded as '?' in the source file: eight in node-caps and one in breast-quad. All other attributes are complete.

Preprocessing

Missing categorical values were imputed using the most-frequent-value strategy within each attribute, applied only to the training fold in every cross-validation split to avoid information leakage from the held-out data (Vabalas et al., 2019). All nine categorical attributes, including the ordinal deg-malig grade, were then one-hot encoded; deg-malig was treated as nominal rather than numeric given the small sample size and the absence of a clear justification for treating grade differences as equally spaced. Encoding produced 41 binary indicator features from the original 9 attributes. The imputation and encoding steps were wrapped in a single scikit-learn ColumnTransformer so that both operations are refit correctly within each cross-validation fold rather than applied once to the full dataset.

Model Architecture

The hybrid classifier consists of three base learners combined through soft voting. The Random Forest base learner is an ensemble of decision trees, each trained on a bootstrap sample of the training data with a random subset of features considered at each split; its class-1 probability output is the fraction of trees voting for the recurrence class. The Logistic Regression base learner fits a linear decision boundary in the

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI | Volume XIII Issue XV October 2026 | Special Issue on Public Health

one-hot-encoded feature space and outputs a probability through the sigmoid (logistic) function; these probabilities are not guaranteed to be well calibrated, particularly once class weighting is applied (van den Goorbergh et al., 2022). The third base learner is a Bernoulli Naive Bayes classifier, chosen specifically because the preprocessing pipeline already represents every attribute as a set of binary indicator features after one-hot encoding — exactly the input Bernoulli Naive Bayes is designed for. It estimates, independently for each binary feature, the likelihood of that feature being present given each class, combines these likelihoods with the class priors via Bayes' theorem, and outputs a posterior probability for the recurrence class. Random Forest and Logistic Regression were configured with class_weight='balanced' in the default configuration to compensate for the 70/30 class imbalance without resorting to synthetic resampling; Naive Bayes has no class_weight option; in the default configuration it uses the empirical class priors estimated from the training fold, although setting fit_prior=False, which is included in the search space (Table 2), replaces these with uniform priors and so acts as a simple imbalance correction.

Given predicted probabilities p_{RF}, p_{LR}, and p_{NB} for the recurrence class from the three base learners, and voting weights w_{RF}, w_{LR}, and w_{NB}, the hybrid model's final probability estimate is the weighted average:

$$p_{\mathrm{hybrid}} = \frac{w_{\mathrm{RF}} p_{\mathrm{RF}} + w_{\mathrm{LR}} p_{\mathrm{LR}} + w_{\mathrm{NB}} p_{\mathrm{NB}}}{w_{\mathrm{RF}} + w_{\mathrm{LR}} + w_{\mathrm{NB}}}$$

with equal weights ($w_{\mathrm{RF}} = w_{\mathrm{LR}} = w_{\mathrm{NB}} = 1$) in the default configuration and the final class prediction taken as recurrence if $p_{\mathrm{hybrid}} \geq 0.5$.

Hyperparameter Optimization

A randomized search (RandomizedSearchCV, scikit-learn; Bischl et al., 2023) was run over the joint hyperparameter space of all three base learners and the voting weights simultaneously, rather than tuning each learner independently, since the optimal setting for one learner can depend on how much influence it is given in the vote.

The search space is summarized in Table 2. Forty candidate configurations were sampled and each evaluated by 5-fold stratified cross-validation on the training split, scored by ROC-AUC; the configuration with the highest mean cross-validated AUC was refit on the full training split and evaluated on the held-out test set. Because the hyperparameters were selected on the training split, the cross-validated figures later reported for the tuned model over the full dataset (Table 5) partly reuse patients that informed that selection. Those figures are therefore optimistically biased in the tuned model's favour; a fully unbiased comparison would require nested cross-validation (Vabalas et al., 2019).

Component	Parameter	Search space
Random Forest	n_estimators	80 - 260
Random Forest	max_depth	None, 3, 5, 8, 12
Random Forest	min_samples_leaf	1 - 8
Random Forest	min_samples_split	2 - 12
Random Forest	max_features	sqrt, log2, all
Random Forest	class_weight	None, balanced, balanced_subsample
Logistic Regression	C (inverse regularization)	0.001 - 10 (uniform)
Logistic Regression	solver	lbfgs, liblinear
Logistic Regression	class_weight	None, balanced
Naive Bayes	alpha (Laplace smoothing)	0.01 - 3.0 (uniform)
Naive Bayes	fit_prior	True, False
Voting	relative weights (RF : LR : NB)	12 combinations, e.g. 1:1:1, 3:1:1, 2:2:1, 1:1:3

Table 2. Hyperparameter search space for the joint RandomizedSearchCV over the three-way hybrid pipeline.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI | Volume XIII Issue XV October 2026 | Special Issue on Public Health

Experimental Setup and Evaluation Metrics

Data were split into an 80% training set (228 patients) and a 20% held-out test set (58 patients) using stratified sampling to preserve class proportions, with a fixed random seed (42) for reproducibility. All model comparisons additionally used 5-fold stratified cross-validation over the full dataset to reduce sensitivity to any single split, which matters given the dataset's small size. Five standard metrics were reported: accuracy, precision, recall, F1-score, and ROC-AUC, computed with the recurrence class treated as the positive class. Confusion matrices are reported at the standard 0.5 probability threshold. All experiments were implemented in Python using scikit-learn (Pedregosa et al., 2011).

RESULTS

Base Learners vs. Hybrid Ensemble — Held-Out Test Set

Table 3 reports performance of Random Forest, Logistic Regression, and Naive Bayes individually, alongside the default (untuned, equal-weight) three-way hybrid model, on the 58-patient held-out test set, using the same train/test split for all four for direct comparability.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Random Forest	0.690	0.467	0.412	0.438	0.636
Logistic Regression	0.655	0.435	0.588	0.500	0.684
Naive Bayes	0.707	0.500	0.412	0.452	0.687
Hybrid (soft voting)	0.690	0.467	0.412	0.438	0.679

Table 3. Held-out test-set performance (n = 58), single 80/20 stratified split, random_state = 42.

On this particular split, Naive Bayes alone obtained the highest accuracy and the highest precision of any individual model, and its ROC-AUC (0.687) slightly exceeded Logistic Regression's (0.684) — a notable result given that Naive Bayes is by far the simplest of the three base learners, estimating no interaction terms and only a small number of parameters per feature. Logistic Regression still produced the best recall and F1-score. The default hybrid landed close to Random Forest on the confusion matrix at the 0.5 threshold, but its ROC-AUC (0.679) sat between the three base learners, reflecting that soft voting is combining probability estimates that disagreed on which patients were most at risk. Figure 1 shows the confusion matrix and ROC curve for the hybrid model in isolation; Figure 2 places all four side by side, and Figure 3 overlays their ROC curves directly.

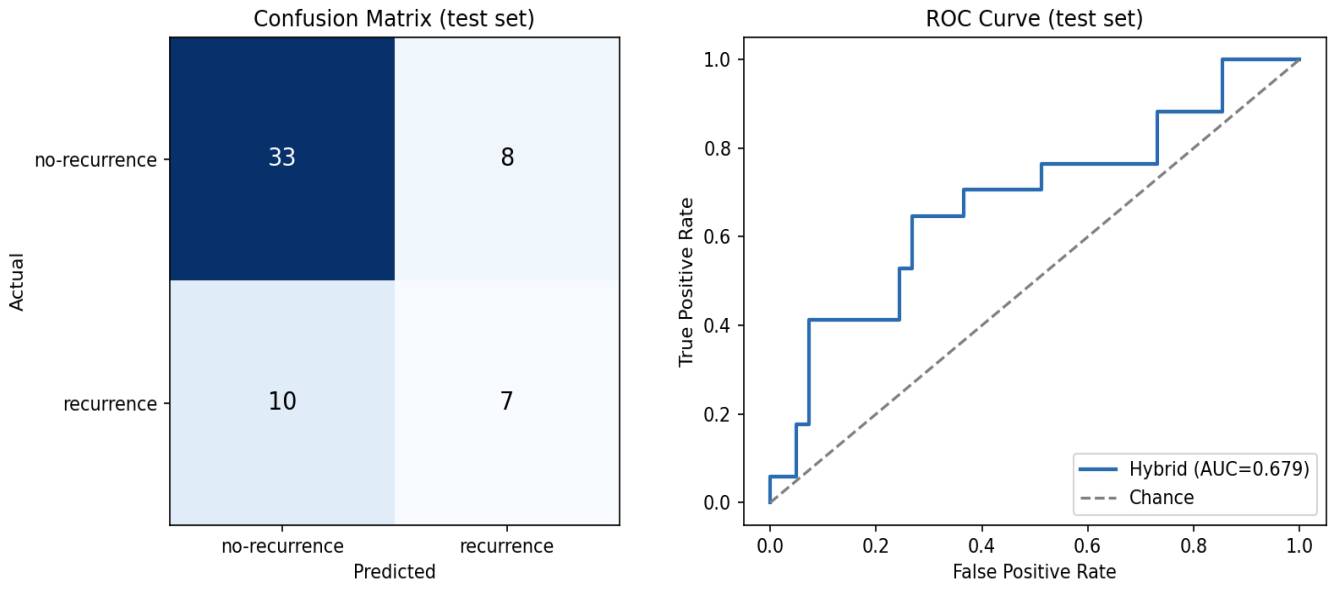


Figure 1. Confusion matrix and ROC curve for the default three-way hybrid model on the held-out test set.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI | Volume XIII Issue XV October 2026 | Special Issue on Public Health

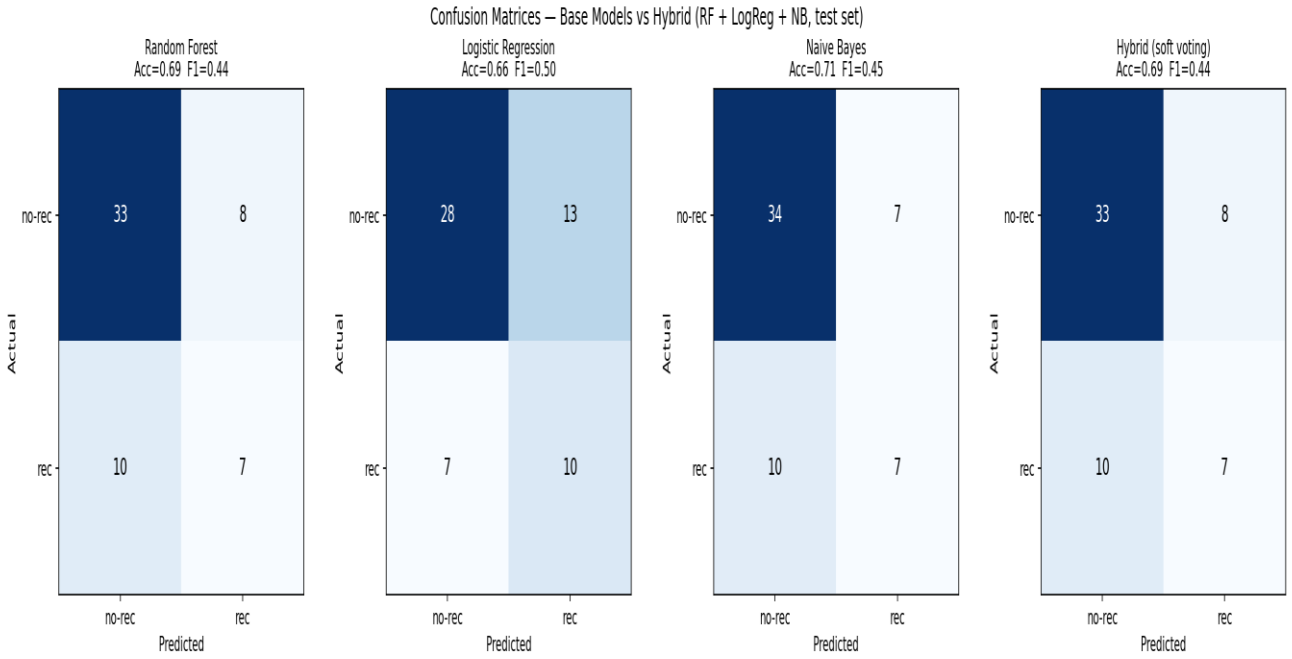


Figure 2. Confusion matrices for Random Forest, Logistic Regression, Naive Bayes, and the hybrid model on the same held-out test set.

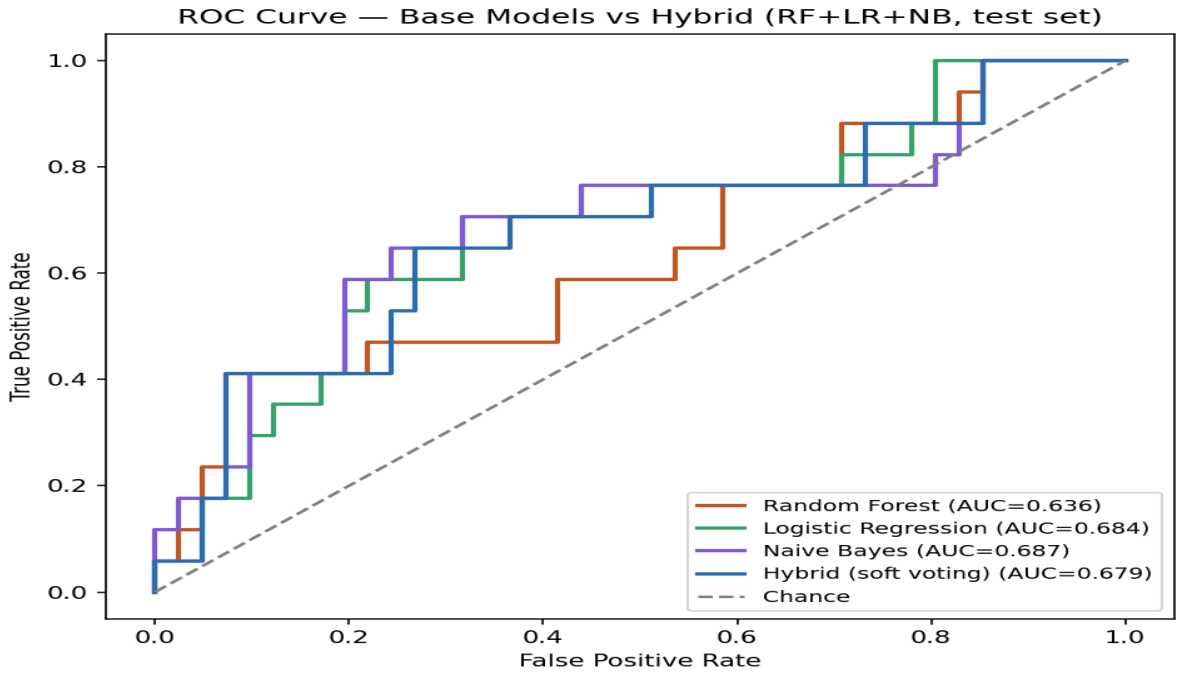


Figure 3. ROC curves for Random Forest, Logistic Regression, Naive Bayes, and the hybrid model on the held-out test set.

Cross-Validated Comparison

Because a single 58-patient test split is small enough that its outcome can be sensitive to which patients happen to fall into it, all four models were re-evaluated by 5-fold stratified cross-validation across the full 286-patient dataset. These results, summarized in Table 4, are the more reliable basis for comparing the models, although, as discussed below, the differences between models are modest relative to their fold-to-fold variation.

Model	Accuracy (mean ± SD)	ROC-AUC (mean ± SD)	F1 (mean ± SD)
Random Forest	0.692 ± 0.061	0.656 ± 0.113	0.458 ± 0.152
Logistic Regression	0.650 ± 0.057	0.653 ± 0.066	0.486 ± 0.058
Naive Bayes	0.727 ± 0.047	0.701 ± 0.072	0.516 ± 0.103
Two-model hybrid (RF + LR)*	0.685	0.659	0.489
Hybrid (soft voting)	0.724 ± 0.041	0.680 ± 0.078	0.528 ± 0.104

Table 4. Five-fold stratified cross-validation results across the full dataset (n = 286). *Two-model Random Forest + Logistic Regression hybrid from the earlier iteration of this study, shown for reference.

Under cross-validation, Naive Bayes was in fact the strongest individual base learner on both accuracy and ROC-AUC, ahead of Random Forest by 3.5 and 4.5 percentage points respectively, and ahead of Logistic Regression by an even wider margin. This is a useful corrective to the assumption, common in ensemble design, that a more complex model necessarily contributes more: on a dataset this small and this categorical, the low-variance, few-parameter Naive Bayes classifier generalized better than either of its more flexible counterparts. The three-way hybrid nonetheless achieved the best F1-score of any model tested (0.528) and the lowest accuracy variance across folds, indicating that even though Naive Bayes was individually stronger, blending in Random Forest and Logistic Regression's probability estimates still improved the balance between precision and recall beyond what Naive Bayes achieved alone. Comparing this three-way hybrid to the two-way Random Forest + Logistic Regression hybrid examined in an earlier iteration of this study (0.685 accuracy, 0.659 AUC, 0.489 F1) shows an improvement on every metric from adding Naive Bayes as a third, structurally different base learner: +3.9 points accuracy, +2.1 points AUC, +3.9 points F1. Two caveats apply. Most of these differences are smaller than one fold-to-fold standard deviation (for example, 0.724 ± 0.041 for the hybrid against 0.727 ± 0.047 for Naive Bayes), and no formal significance test was applied, so they are best read as indicative rather than conclusive. The two-model figures, now listed in Table 4 for reference, come from the earlier iteration of the study.

Hyperparameter Optimization Results

The best configuration found by the randomized search used 230 Random Forest trees with max_depth = 8, max_features = 'log2', min_samples_leaf = 1, min_samples_split = 8, and no class-weight adjustment; a Logistic Regression with C = 0.741 and no class-weight adjustment; a Naive Bayes classifier with a Laplace smoothing parameter (alpha) of 0.358; and voting weights of 3:1:1, giving Random Forest three times the influence of either Logistic Regression or Naive Bayes in the final vote. This configuration achieved a mean ROC-AUC of 0.688 in the inner 5-fold cross-validation on the 228-patient training split (the 0.684 in Table 5 comes from 5-fold cross-validation over the full dataset), against a background of 40 sampled configurations whose cross-validated AUC ranged from about 0.65 to 0.69 (Figure 4, left panel) — again a fairly flat optimization surface.

Metric	Default hybrid (5-fold CV)	Tuned hybrid (5-fold CV)
Accuracy	0.724 ± 0.041	0.720 ± 0.040
ROC-AUC	0.680 ± 0.078	0.684 ± 0.082
F1-score	0.528 ± 0.104	0.424 ± 0.146

Table 5. Default vs. tuned three-way hybrid model, 5-fold cross-validation on the full dataset. Tuned-model figures are optimistically biased because its hyperparameters were selected on the training split (see Section 3.4).

On the held-out test set, the tuned model scored 0.741 accuracy, 0.583 precision, 0.412 recall, 0.483 F1, and 0.677 ROC-AUC, with a confusion matrix of 36 true negatives, 5 false positives, 10 false negatives, and 7 true positives. Figure 4 summarizes the search's score distribution, the default-versus-tuned bar comparison, and the ROC curves for both configurations on the test set.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI | Volume XIII Issue XV October 2026 | Special Issue on Public Health

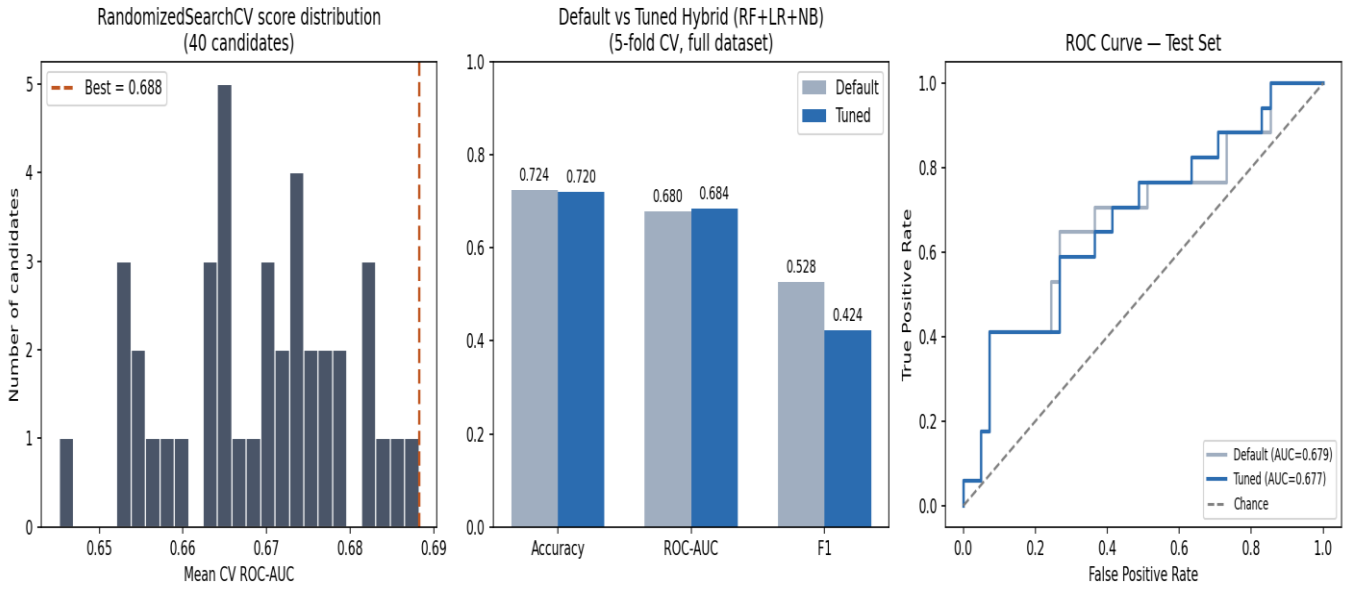


Figure 4. Hyperparameter search score distribution, default-vs-tuned comparison, and test-set ROC curves (three-way hybrid).

Tuning raised cross-validated ROC-AUC by 0.4 percentage points but lowered accuracy slightly (-0.4 points) and cut F1-score by 10.4 percentage points relative to the default hybrid — a considerably larger recall trade-off than the two-model version of this study showed under the same search procedure. The chosen configuration dropped class-weight balancing from both Random Forest and Logistic Regression and gave Random Forest three times the voting weight of either other learner, which together pushed the ensemble toward higher precision and lower recall on the minority recurrence class. This reinforces a pattern already visible in the two-model study: a randomized search scored on ROC-AUC alone reliably finds configurations that rank patients well in aggregate, but does not reliably preserve sensitivity to the minority class at the fixed 0.5 decision threshold used for classification, regardless of how many base learners are in the ensemble. The direction of the effect is consistent with simulation evidence that imbalance corrections mainly shift predicted probabilities toward the minority class, buying sensitivity at a fixed threshold at the cost of calibration without improving discrimination (van den Goorbergh et al., 2022); removing class weighting, as the tuned configuration did, moves the ensemble the other way.

DISCUSSION

Four findings stand out. First, and most striking, is that the simplest of the three base learners — Naive Bayes — was individually the strongest under cross-validation, outperforming both Random Forest and Logistic Regression on accuracy and ROC-AUC. This runs against the intuition, common when building an ensemble, that more flexible models will dominate simpler ones; on a 286-row categorical dataset, the opposite held. Naive Bayes's independence assumption, which is usually presented as its main weakness, functions here as a strong regularizer: it estimates far fewer parameters than either Random Forest or Logistic Regression, and with so little training data, that parameter economy appears to matter more than the ability to model interactions or fit a flexible decision boundary.

Second, adding Naive Bayes as a third base learner appears to have improved the hybrid ensemble more than hyperparameter tuning did. Moving from the two-model (Random Forest + Logistic Regression) hybrid examined previously to the three-model hybrid raised cross-validated F1 by 3.9 points and accuracy by 3.9 points, whereas tuning the three-model hybrid's own hyperparameters — a considerably more expensive search — improved ROC-AUC by less than half a point and, on F1, made the model measurably worse. The selection bias in the tuned model's cross-validated figures (Section 3.4) works in tuning's favour, so if anything this comparison flatters the search. For a practitioner working with a small, fixed dataset, this suggests that diversifying the set of base learners is likely to be a better use of limited effort than an

extensive hyperparameter search over a fixed set of learners, at least until the marginal learner choices have been exhausted. This fits benchmark evidence that Random Forest, in particular, gains relatively little from tuning compared with other learners (Probst et al., 2019).

Third, the same precision-recall trade-off observed in the two-model study reappeared, more sharply, in the three-model study: optimizing the search for ROC-AUC found a configuration with a half-point AUC gain but a ten-point F1 loss, driven by reduced recall on the recurrence class. This is not a flaw specific to adding Naive Bayes; it is a structural consequence of scoring a hyperparameter search on a ranking metric (ROC-AUC) while evaluating classification performance at a fixed probability threshold, since discrimination and calibration are distinct properties and a gain in one says little about the other (Van Calster et al., 2019). The finding is worth taking seriously because it recurred across two separate searches, which suggests it is not an artifact of one particular random search but a property of tuning this class of model against this objective on this kind of imbalanced clinical data. Two searches on a single dataset are not enough to establish that firmly, however.

Fourth, these results remain consistent with, rather than a dramatic improvement over, the historical performance of classical rule-induction and decision-tree systems on this exact dataset (65-78% accuracy; Cestnik et al., 1987; Clark & Niblett, 1987; Michalski et al., 1986). The three-way hybrid's cross-validated accuracy of 72.4% sits comfortably inside that historical range rather than exceeding its upper bound. That consistency is itself informative: even a heterogeneous three-model ensemble with a targeted hyperparameter search does not meaningfully push past a ceiling that has held for nearly four decades of work on this dataset, which reinforces that the ceiling is a property of the data — its size, its label noise, and its coarse categorical binning — rather than of any particular modeling approach applied to it.

The study has several limitations. The dataset is drawn from a single institution and a specific patient population from the late 1980s, which limits how far its conclusions generalize to contemporary or geographically different populations, including the Nigerian clinical context this work is ultimately motivated by, where breast cancer incidence is rising and late-stage presentation remains common (Azubuike et al., 2018; Bray et al., 2024). The categorical binning of continuous variables such as tumor size and node count discards information that a model trained on the original continuous measurements might exploit. The sample is also small in a formal sense: with 85 recurrence events and roughly forty encoded binary predictors, it falls well short of the minimum sample sizes now recommended for developing clinical prediction models (Riley et al., 2020). Class imbalance was addressed only through class-weighting on two of the three base learners, not through resampling methods such as SMOTE (Fernández et al., 2018), and Naive Bayes received no imbalance correction in the default configuration, which is a reasonable direction for follow-up work, although recent evidence suggests that such corrections can degrade probability calibration and should be applied with care (van den Goorbergh et al., 2022). Model comparisons rest on a single run of 5-fold cross-validation without formal significance testing, and the tuned model's cross-validated estimates carry the selection bias described in Section 3.4; repeated cross-validation with nested tuning would address both. Finally, the hyperparameter search was deliberately kept modest (40 candidates) to fit the available compute; a larger, ideally Bayesian, search (Akiba et al., 2019; Bischl et al., 2023) might close some of the gap between the reported configurations and the optimum, though the flat score distributions observed in both the two-model and three-model searches (Figure 4) suggest the returns to a larger search may be limited on a dataset this small.

CONCLUSION

This study built and evaluated a three-way hybrid ensemble — Random Forest, Logistic Regression, and Naive Bayes, combined by soft voting — for breast cancer recurrence prediction on the Ljubljana UCI dataset, extending an earlier two-model (Random Forest + Logistic Regression) version of the same study. Naive Bayes, the simplest of the three base learners, turned out to be the strongest individually under cross-validation, and adding it to the ensemble improved every headline metric over the two-model hybrid by a wider margin than a dedicated hyperparameter search over the resulting three-model ensemble later achieved. Hyperparameter optimization again pushed ROC-AUC marginally higher but noticeably reduced

recall on the minority recurrence class, reproducing a precision-recall trade-off first observed in the two-model study and suggesting that it is a repeatable property of ROC-AUC-driven tuning on this kind of imbalanced clinical data rather than a one-off result. The three-way hybrid's cross-validated accuracy (72.4%) remains within the historical 65-78% range reported for this dataset by classical machine learning systems, reinforcing that the paper's central lesson is about where the data's ceiling lies, not merely about which model or combination of models is used to approach it.

Future work should pursue six directions: tuning the hyperparameter search directly against a recall- or cost-sensitive objective rather than ROC-AUC alone, or against an imbalance-aware summary such as the Matthews correlation coefficient (Chicco & Jurman, 2020); re-running the model comparisons with repeated, nested cross-validation, formal significance testing, and a leave-one-learner-out ablation; applying resampling techniques such as SMOTE alongside class-weighting to address the imbalance more directly, including for the Naive Bayes component, which receives no imbalance correction in the default configuration; testing whether a fourth structurally different base learner (for instance, a gradient-boosted tree or a k-nearest-neighbors classifier) continues the pattern observed here of architectural diversity outperforming tuning; validating the pipeline on a larger, more contemporary, and ideally multi-institutional dataset, including data more representative of the Nigerian and West African clinical settings this work aims ultimately to serve, with reporting aligned to the TRIPOD+AI guidance (Collins et al., 2024); and adding model explainability tools such as SHAP values (Lundberg & Lee, 2017), so that clinicians can inspect why the hybrid model made a particular prediction rather than treating it as a black box.

REFERENCES

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623–2631). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330701
2. Azubuike, S. O., Muirhead, C., Hayes, L., & McNally, R. (2018). Rising global burden of breast cancer: The case of sub-Saharan Africa (with emphasis on Nigeria) and implications for regional development: A review. World Journal of Surgical Oncology, 16, Article 63. https://doi.org/10.1186/s12957-018-1345-2
3. Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. WIREs Data Mining and Knowledge Discovery, 13(2), Article e1484. https://doi.org/10.1002/widm.1484
4. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263. https://doi.org/10.3322/caac.21834
5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
6. Cestnik, B., Kononenko, I., & Bratko, I. (1987). ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In I. Bratko & N. Lavrač (Eds.), Progress in machine learning: Proceedings of EWSL-87 (pp. 31–45). Sigma Press.
7. 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, Article 6. https://doi.org/10.1186/s12864-019-6413-7
8. Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004
9. Clark, P., & Niblett, T. (1987). Induction in noisy domains. In I. Bratko & N. Lavrač (Eds.), Progress in machine learning: Proceedings of EWSL-87 (pp. 11–30). Sigma Press.

10. 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. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., . . . Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, Article e078378. https://doi.org/10.1136/bmj-2023-078378
11. Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple classifier systems (Lecture Notes in Computer Science, Vol. 1857, pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1
12. Dua, D., & Graff, C. (2019). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml
13. Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863–905. https://doi.org/10.1613/jair.1.11192
14. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In Advances in Neural Information Processing Systems (Vol. 35). Curran Associates.
15. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.
16. Lu, D., Long, X., Fu, W., Liu, B., Zhou, X., & Sun, S. (2023). Predictive value of machine learning for breast cancer recurrence: A systematic review and meta-analysis. Journal of Cancer Research and Clinical Oncology, 149(12), 10659–10674. https://doi.org/10.1007/s00432-023-04967-w
17. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates.
18. Mazo, C., Aura, C., Rahman, A., Gallagher, W. M., & Mooney, C. (2022). Application of artificial intelligence techniques to predict risk of recurrence of breast cancer: A systematic review. Journal of Personalized Medicine, 12(9), Article 1496. https://doi.org/10.3390/jpm12091496
19. Michalski, R. S., Mozetic, I., Hong, J., & Lavrac, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86) (pp. 1041–1045).
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
21. Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. Journal of Machine Learning Research, 20(53), 1–32.
22. Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Jr., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. BMJ, 368, Article m441. https://doi.org/10.1136/bmj.m441
23. Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence (Vol. 3, No. 22, pp. 41–46).
24. Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery, 8(4), Article e1249. https://doi.org/10.1002/widm.1249
25. Sherazi, S. W. A., Bae, J.-W., & Lee, J. Y. (2021). A soft voting ensemble classifier for early prediction and diagnosis of occurrences of major adverse cardiovascular events for STEMI and NSTEMI during 2-year follow-up in patients with acute coronary syndrome. PLOS ONE, 16(6), Article e0249338. https://doi.org/10.1371/journal.pone.0249338
26. Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. PLOS ONE, 14(11), Article e0224365. https://doi.org/10.1371/journal.pone.0224365
27. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, Article 230. https://doi.org/10.1186/s12916-019-1466-7

28. van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. Journal of the American Medical Informatics Association, 29(9), 1525–1534. https://doi.org/10.1093/jamia/ocac093
29. Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. Soft Computing, 25(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6
30. Zwitter, M., & Soklic, M. (1988). Breast cancer [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C51P4M