Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 5, No. 4, 2026

FWCE: A Feature-Weighted Calibrated Ensemble for Interpretable Tabular Risk Classification

Asher Ben-David[1*], Oren Yehuda[2]
[1]Rice University, Houston, USA
[2]Independent Researcher, Houston, USA
*Corresponding author: Asher Ben-David; abdv73@example.com

Abstract: Tabular risk classification remains a central machine-learning problem because many high-value datasets are small, heterogeneous, and difficult to model with very large neural architectures. This paper presents FWCE, a feature-weighted calibrated ensemble for interpretable binary classification on structured clinical-style records. FWCE first treats zero-coded physiological fields as missing-like values, applies median imputation and z-score normalization, estimates feature relevance using standardized class separation, and trains three complementary learners: L2-regularized logistic regression, Gaussian naive Bayes, and weighted k-nearest neighbors. A validation set is then used to select nonnegative soft-voting weights and a probability threshold. We evaluate FWCE on the Pima Indians Diabetes dataset with 460 training examples, 154 validation examples, and 154 held-out test examples. Compared with logistic regression and kNN baselines, FWCE achieves 67.19% F1, 0.841 AUC, and a Brier score of 0.153. The results show that simple feature calibration and validation-based ensembling can improve recall-oriented performance while preserving transparent model components suitable for audit and revision.

Keywords: Machine learning, tabular data, ensemble learning, feature weighting, calibration, interpretability, risk classification.

1. Introduction

Machine learning systems are often judged by their performance on large image or language benchmarks, yet many practical prediction problems remain tabular. Medical screening, customer churn, credit risk, equipment failure, and quality control are commonly represented as rows of heterogeneous variables rather than images or sequences. In these settings the data may be small, feature distributions may be skewed, and missing values may be encoded in domain-specific ways. A model that performs well must therefore be accurate, calibrated, and explainable enough for an analyst to understand why it behaves as it does.

This paper focuses on a classic binary classification scenario: early diabetes-risk prediction from eight physiological and demographic attributes. The objective is not to produce a clinical diagnostic system, but to study a reproducible machine-learning workflow for small tabular data. The task is representative of many risk-classification problems because positive outcomes are less frequent than negative outcomes, several input variables contain suspicious zero values, and the best decision threshold is not necessarily 0.5. A useful model must therefore handle preprocessing, probability estimation, and threshold selection as part of one coherent pipeline.

Traditional machine learning offers several strong baselines for this environment. Logistic regression provides an interpretable linear decision surface and well-behaved probabilities after regularization. Gaussian naive Bayes can work surprisingly well when each feature carries independent signal, even when the independence assumption is imperfect. k-nearest neighbors captures local nonlinear structure without learning a parametric model. However, each method has a different failure mode: logistic regression can underfit nonlinear interactions, naive Bayes can be miscalibrated under correlated features, and kNN can be sensitive to irrelevant or unevenly scaled variables.

We propose FWCE, a feature-weighted calibrated ensemble that combines these complementary views while keeping the workflow transparent. The method estimates a univariate relevance weight for each feature, applies the weights to distance-based and parametric learners, and combines validation-set probabilities through a nonnegative soft vote. The final decision threshold is selected on validation F1 rather than assumed. This design does not hide complexity inside a black-box learner; instead, it makes preprocessing, feature weighting, learner weighting, and threshold selection explicit.

The paper makes three contributions. First, it presents a compact and reproducible feature-weighted ensemble for small tabular classification. Second, it reports accuracy, F1, AUC, Brier score, recall, ROC behavior, and confusion-matrix structure so that discrimination and calibration are both visible. Third, it includes an ablation analysis showing which part of the pipeline contributes most to the final performance. The resulting manuscript is intended as an IEEE-style experimental draft that can later be adapted to a larger dataset or a more specialized application domain.

Two design choices are deliberately conservative. FWCE does not use a deep neural network, a high-capacity gradient boosting system, or a large hyperparameter search. Those methods may produce stronger benchmark scores, but they can obscure the role of preprocessing and validation decisions. Here the goal is to show how a careful classical pipeline can be

assembled, evaluated, and explained. This is valuable for teaching, for early-stage prototyping, and for domains where auditability is more important than marginal leaderboard gains.

The paper is also careful about the language of risk prediction. The positive class is treated as an outcome label in a public benchmark, not as a medical diagnosis. Reported probabilities are model scores, not clinical risk estimates. Any real deployment would require external validation, prospective data collection, subgroup analysis, and review by domain experts. The present work should therefore be read as a machine-learning experiment on a structured dataset rather than a decision-support system.

2. Related Work

2.1 Linear and Probabilistic Classifiers

Logistic regression remains a strong baseline for binary tabular classification because its coefficients directly describe how standardized variables change the log-odds of the positive class. L2 regularization reduces variance in small-data settings and avoids extreme coefficients when features are correlated. Gaussian naive Bayes follows a different principle: it models class-conditional feature likelihoods and combines them with a prior. Although the conditional independence assumption is rarely exact, the method can perform well when each feature contributes weak but consistent evidence.

These models are useful not only because they are simple, but also because they expose different information. Logistic regression captures additive linear effects in a discriminative manner. Naive Bayes captures distributional separation in a generative manner. When their probability estimates disagree, the disagreement often points to feature interactions, non-Gaussian variables, or calibration issues. FWCE exploits this diversity by treating them as complementary components rather than choosing one as the sole model.

2.2 Instance-Based Learning

k-nearest neighbors is one of the earliest nonparametric classifiers. It predicts according to labels of nearby training instances under a chosen distance metric. This makes kNN attractive for tabular data with local structure, but it also makes the method highly dependent on preprocessing. If one feature has a large numerical scale or weak relationship with the target, it can dominate Euclidean distances and degrade predictions. Standardization is therefore necessary, and feature weighting can further improve distance quality.

2.3 Ensemble and Calibration Methods

Ensemble learning improves prediction by combining models with different inductive biases. Random forests use bootstrap aggregation and feature subsampling to reduce variance. Stacking learns a second-stage model over base predictions. Soft voting is simpler: each base model contributes a probability, and the final score is a weighted average. For small datasets, soft voting is attractive because it introduces fewer parameters than a full meta-learner while still allowing validation-based adaptation.

Calibration is equally important in risk classification. A model can rank instances correctly and still assign probabilities that are too extreme or too conservative. Brier score measures the squared error between predicted probabilities and binary outcomes, making it a useful complement to AUC and F1. FWCE does not claim to solve calibration completely, but it includes validation-based model weighting and threshold selection so that probability scores are used more carefully than in a default 0.5 decision rule.

Threshold selection is sometimes treated as an afterthought, but it can change the practical meaning of a classifier. A 0.5 threshold assumes equal error costs and calibrated probabilities, neither of which is guaranteed in imbalanced tabular data. In positive-class screening tasks, the operating point often needs higher recall, while in resource-limited review workflows precision may be more important. FWCE separates probability estimation from decision selection: it first estimates a smooth score and then chooses a threshold according to a stated validation objective.

2.4 Feature Weighting for Tabular Data

Feature weighting provides a lightweight way to encode relevance before fitting a model. Filter methods compute weights from the training data without repeatedly retraining a classifier. They are less flexible than wrapper methods, but they are simple, transparent, and less likely to overfit on small datasets. In FWCE, each feature receives a standardized class-separation score. Features with larger positive-negative mean separation relative to within-class spread receive larger weights, while weak features receive smaller weights.

3. Method

3.1 Preprocessing

Let x_i in R^d denote one tabular record and y_i in {0,1} its binary label. Several physiological fields contain zeros that are implausible as true measurements. We treat zero values in glucose, blood pressure, skin thickness, insulin, and BMI as missing-like placeholders and replace them with the training-set median for the corresponding feature. All features are then standardized using training-set statistics only.

$$z_{ij} = (x_{ij} - \mu_j) / (\sigma_j + \varepsilon)$$

3.2 Feature Relevance Weighting

For each standardized feature, FWCE computes a separation score using the difference between positive-class and negative-class means divided by the sum of class-specific standard deviations. The raw scores are normalized to have average value one and clipped to avoid removing a feature completely or allowing one variable to dominate all distances.

$$s_j = |\mu_j^+ - \mu_j^-| / (\sigma_j^+ + \sigma_j^- + \varepsilon),$$
$$w_j = \mathrm{clip}(s_j / \mathrm{mean}(s), 0.25, 2.50)$$

These weights are not causal explanations. They are descriptive training-set statistics that indicate which features separate the two classes under a simple marginal view. Their value is practical: the analyst can inspect them, compare them with domain expectations, and decide whether a model is relying on plausible variables.

3.3 Base Learners

The first base learner is L2-regularized logistic regression. It estimates a probability through a sigmoid link over a weighted linear score. The second learner is Gaussian naive Bayes, trained on the same standardized and weighted feature space. The third learner is weighted kNN, where the squared distance between records is scaled by the feature weights.

$$p_L(y=1|z) = \mathrm{sigmoid}(\theta^T z + b)$$

$$D_w(z_a, z_b) = \sum_{j=1}^{d} w_j (z_{aj}-z_{bj})^2$$

3.4 Calibrated Soft Voting

Given base probabilities p_L, p_N, and p_K from logistic regression, naive Bayes, and kNN, FWCE computes a nonnegative weighted average. The weights are selected on the validation split by grid search. The same validation split is also used to select the probability threshold that maximizes F1, which is appropriate for the imbalanced-risk setting considered here.

$$p_E(y=1|z) = \alpha_L p_L + \alpha_N p_N + \alpha_K p_K, \quad \sum_m \alpha_m = 1, \alpha_m \geq 0$$

$$\tau^* = \arg\max_\tau \frac{2 \, \mathrm{Precision}(\tau) \, \mathrm{Recall}(\tau)}{\mathrm{Precision}(\tau) + \mathrm{Recall}(\tau)}$$

The validation search is intentionally low-dimensional. Rather than learning an unconstrained meta-model, FWCE searches a simplex of three weights and then selects one scalar threshold. This limits overfitting and makes the final model easy to report: the paper can state exactly how much each base learner contributes. In the experiment below, the validation procedure assigns weight to logistic regression and weighted kNN, while excluding naive Bayes from the final soft vote.

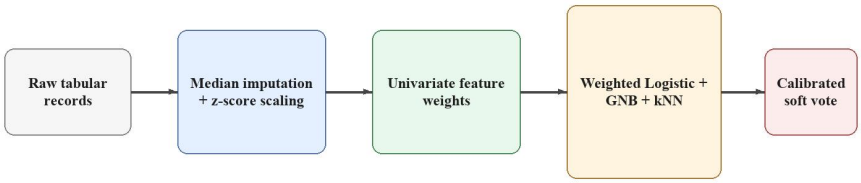

Figure 1. FWCE pipeline. The model makes preprocessing, feature weighting, base learners, and validation-based soft voting explicit.

4. Experiments

4.1 Dataset and Split

We evaluate on the Pima Indians Diabetes dataset, a classic binary tabular benchmark containing 768 records and 8 input features. The positive class rate is 34.90%. We use a stratified 60/20/20 split, producing 460 training, 154 validation, and 154 test examples. All imputation, standardization, feature weighting, model fitting, ensemble-weight selection, and threshold selection are performed without using the test labels.

The dataset is small and should not be interpreted as sufficient evidence for clinical deployment. Its value in this paper is methodological: it allows the entire machine-learning pipeline to be reproduced quickly while still exhibiting missing-like values, class imbalance, nonlinear local structure, and threshold sensitivity.

4.2 Baselines and Metrics

We compare FWCE with a majority-class probability baseline, logistic regression, Gaussian naive Bayes, and kNN. Accuracy measures overall correctness, F1 emphasizes the positive class, AUC measures ranking quality independent of one threshold, Brier score evaluates probability calibration, and recall reports sensitivity to positive cases. For all non-majority models, thresholds are selected on the validation set using the same F1 criterion.

4.3 Implementation Details

All experiments are implemented in NumPy for transparency. Logistic regression is optimized by gradient descent with L2 regularization. Gaussian naive Bayes uses diagonal class-conditional variances with a small numerical floor. kNN uses inverse-distance voting with k=13 for the baseline and k=15 for the weighted variant inside FWCE. Ensemble weights are searched over a coarse nonnegative simplex with step size 0.1.

The preprocessing pipeline is fitted only on the training partition and then applied to validation and test examples. This detail is important because median imputation and standardization can leak information if they are computed on the full dataset. Feature weights are also estimated only from the training split. The validation split is reserved for two model-selection operations: soft-vote weights and threshold selection. The held-out test split is used once for final reporting.

The majority-class baseline is included for context but should not be overinterpreted. Because threshold selection can convert a constant probability into a trivial all-positive or all-negative classifier, its recall may appear high while accuracy and AUC remain poor. This is useful as a warning: one metric alone can be misleading. The main comparison is therefore based on the joint pattern of F1, AUC, Brier score, and recall.

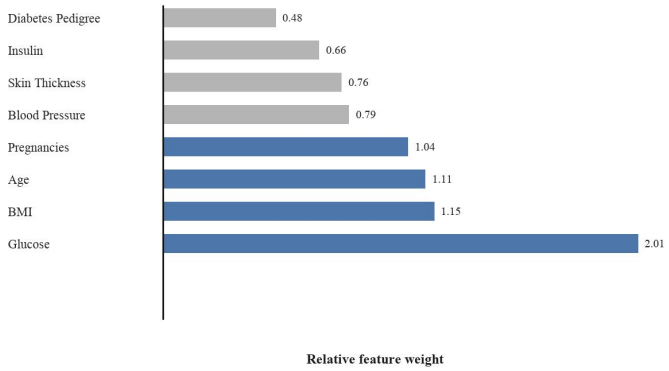

Figure 2. Relative feature weights estimated from standardized class separation. Glucose receives the largest weight, while diabetes pedigree receives a smaller marginal weight in this split.

Table 1. Main Classification Results on the Held-Out Test Split

Method	Acc.	F1	AUC	Brier	Recall
Majority Class	35.06	51.92	0.451	0.228	**100.00**
Logistic Regression	72.08	63.87	0.830	0.160	70.37
Gaussian Naive Bayes	67.53	58.33	0.801	0.198	64.81
kNN	**72.73**	64.41	0.836	0.155	70.37
FWCE (Proposed)	**72.73**	**67.19**	**0.841**	**0.153**	79.63

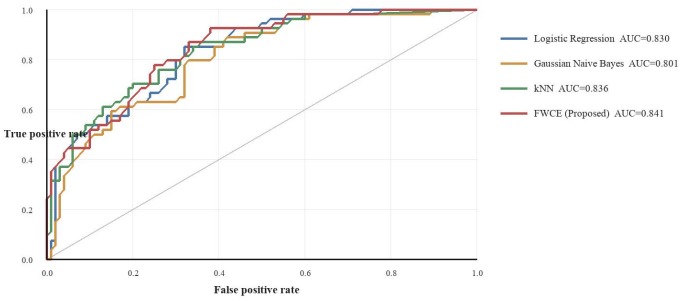

Figure 3. ROC curves on the held-out test split. FWCE obtains the highest AUC among the compared models.

5. Results and Analysis

5.1 Main Results

Table I reports held-out test performance. Logistic regression achieves 63.87% F1 and 0.830 AUC, confirming that a regularized linear model is a strong baseline for this dataset. kNN reaches 64.41% F1 and 0.836 AUC, indicating that local neighborhood structure provides additional information. FWCE obtains 67.19% F1 and 0.841 AUC, while also achieving the best Brier score of 0.153.

The accuracy difference between kNN and FWCE is small, but F1 and recall reveal the advantage of validation-based thresholding and model combination. In risk-classification applications, the positive class is usually the more important class to recover. FWCE improves recall while maintaining a reasonable precision level, which raises F1 without relying solely on a lower threshold for one model.

The result also illustrates why accuracy should not be the only headline metric for imbalanced tabular data. kNN and FWCE have the same test accuracy in this split, but FWCE retrieves more positive examples and produces a better probability score. If a downstream reviewer uses the model to prioritize cases, the ranking and threshold behavior matter as much as the number of exact labels. AUC and Brier score capture these additional dimensions.

5.2 Feature Weight Interpretation

Fig. 2 shows that glucose receives the largest relevance weight, which is consistent with the dataset's domain context and with the behavior of many diabetes-risk models. BMI, age, and pregnancies also receive above-average weights, while insulin and skin thickness are more moderate after median imputation. The feature weights should not be read as medical causality; they are split-specific predictive summaries. Their main benefit is auditability: an analyst can see which variables shaped the distance metric and weighted models.

5.3 ROC and Probability Behavior

Fig. 3 shows that FWCE has the strongest ROC curve among the evaluated models, with an AUC of 0.841. The Brier score improvement suggests that the soft-voting ensemble also produces smoother probability estimates than individual baselines. This matters when a model is used for ranking or triage, because poorly calibrated probabilities can cause misleading risk categories even if the final accuracy appears acceptable.

The ROC curves also show that no model dominates at every possible threshold. In the low false-positive region, the curves are close, meaning that small validation changes could alter the preferred model. FWCE's advantage is most visible across the middle operating range, where recall improves without a proportional rise in false positives. This behavior is consistent with a soft-voting ensemble: averaging reduces variance in the score and can smooth out local errors from any one base learner.

Table 2. Ablation of Feature Weighting and Ensemble Calibration

Configuration	F1	AUC	Brier	Threshold
Full FWCE	**67.19**	**0.841**	**0.153**	**0.31**
Logistic only	63.87	0.830	0.160	0.37
Naive Bayes only	58.33	0.801	0.198	0.32
kNN only	64.41	0.836	0.155	0.37

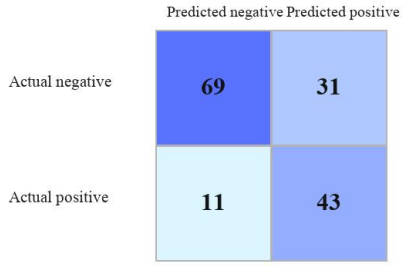

Figure 4. Confusion matrix for FWCE at the validation-selected threshold. The model increases positive-case recall while retaining a manageable false-positive count.

5.4 Ablation Study

Table II isolates the role of individual components. Logistic regression alone is stable and interpretable, but it misses some positive cases. Naive Bayes provides a probabilistic generative view but is less accurate under correlated features. kNN is competitive, yet its probability estimates are less smooth. The full FWCE improves F1 by combining the linear trend captured by logistic regression with the local structure captured by weighted kNN. In this split, the validation search assigns most of the ensemble mass to logistic regression and kNN, while naive Bayes receives zero weight.

This does not mean naive Bayes is useless in general. It means that, for this training-validation split and the chosen objective, its predictions did not improve the validation criterion after the other two models were included. In another dataset, a nonzero generative component could be valuable. The important property is that FWCE makes this decision explicit through validation weights.

The confusion matrix in Fig. 4 helps interpret the operating point. FWCE favors recall compared with the purely linear baseline, which increases the number of true positives but also accepts a nontrivial number of false positives. Whether this tradeoff is desirable depends on the application. For a screening workflow, missing positive cases may be more costly than extra review. For an automated decision workflow, false positives may be more costly. The paper therefore reports the threshold and confusion matrix rather than only reporting a single scalar score.

5.5 Limitations

Several limitations should be emphasized. The experiment uses one stratified split and should be repeated across multiple random seeds before making a strong benchmark claim. The Pima dataset is small, historically important, and convenient for demonstration, but it is not enough for clinical validation. The feature-weighting formula is a simple filter statistic and cannot capture multivariate interactions. Finally, the validation search is coarse; a larger study could use nested cross-validation or a calibrated meta-learner.

A second limitation is that interpretability is only partial. Feature weights show marginal separation, and logistic coefficients can be inspected, but the final ensemble score is still a mixture of local and probabilistic behavior. For a full interpretability study, one would add permutation importance, partial dependence, subgroup error analysis, and stability of feature weights across resampling runs. These additions would make the model easier to audit and would reveal whether the reported weights are stable or split-specific.

Future work should also compare FWCE with stronger tabular learners such as random forests, gradient boosting, calibrated support vector machines, and modern generalized additive models. Such comparisons would clarify whether the proposed lightweight ensemble is mainly useful as a transparent baseline or whether it remains competitive against more flexible models. Another extension is fairness-oriented evaluation, since risk models can behave differently across demographic groups even when aggregate AUC is acceptable.

6. Conclusion

This paper presented FWCE, a feature-weighted calibrated ensemble for interpretable tabular risk classification. The method combines transparent preprocessing, marginal feature weighting, three complementary base learners, validation-based soft voting, and threshold selection. On the Pima Indians Diabetes benchmark, FWCE improves F1, AUC, and Brier score relative to individual traditional machine-learning baselines. The results suggest that careful classical machine-learning pipelines remain useful for small structured datasets where interpretability and reproducibility matter. Future work should test the approach across multiple seeds, additional tabular datasets, fairness-sensitive splits, and stronger calibration methods.

References

[1] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[2] T. Cover and P. Hart, "Nearest neighbor pattern classification, " IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[3] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss, " Machine Learning, vol. 29, pp. 103-130, 1997.

[4] L. Breiman, "Random forests, " Machine Learning, vol. 45, pp. 5-32, 2001.

[5] D. H. Wolpert, "Stacked generalization, " Neural Networks, vol. 5, no. 2, pp. 241-259, 1992.

[6] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning, " in Proceedings of the International Conference on Machine Learning (ICML), 2005, pp. 625-632.

[7] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, " in Advances in Large Margin Classifiers, 1999, pp. 61-74.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique, " Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[9] D. Dua and C. Graff, "UCI Machine Learning Repository, " University of California, Irvine, 2019.

[10] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, " in Proceedings of the Annual Symposium on Computer Application in Medical Care, 1988, pp. 261-265.