

IMPLEMENTATION AND VALIDATION OF AN EXPLAINABLE ENSEMBLE LEARNING FRAMEWORK FOR BREAST CANCER CLASSIFICATION USING CYTOLOGICAL FEATURES

Youssef Chidera ^{1*}, Jakub Petit ²

¹ Université Paris-Saclay, FRANCE

² Universidad Politécnica de Madrid, SPAIN

ABSTRACT

Explainable machine learning on cytological features is a promising route to reliable, interpretable breast-cancer diagnosis from fine-needle aspiration (FNA). Building on the comparative analysis of explainable models for cancer classification using cytological features by Siddiqui *et al.* [1], this study implements their comparative approach as the first layer of an explainable ensemble framework and validates the framework rigorously. Six base learners (logistic regression, support vector machine, k-nearest neighbours, random forest, extra trees and gradient boosting) and two ensembles (soft voting and stacking) were trained on the Wisconsin Diagnostic Breast Cancer dataset (569 FNA samples, 30 nuclear features). Validation combined 5×10 -fold repeated cross-validation on a 75% development set with corrected resampled t-tests, a 25% held-out test set with bootstrap confidence intervals, calibration analysis and a screening-oriented decision threshold. Explanations were produced with SHAP for every model and evaluated for cross-model agreement and stability across folds. The soft-voting ensemble reached a cross-validated MCC of 0.947 ± 0.041 and AUC of 0.994 ± 0.009 , and on the held-out test set an accuracy of 0.979, sensitivity of 0.943, specificity of 1.000 and AUC of 0.997 (95% CI 0.990–1.000). It was significantly better than random forest in MCC ($p = 0.032$) but statistically indistinguishable from logistic regression and the support vector machine, while showing the smallest fold-to-fold variability. A threshold of 0.17, chosen on development data, raised test sensitivity to 0.981 with specificity 0.978. SHAP rankings agreed well between the ensemble and its members (Spearman ρ 0.76–0.92) and 7 features led by worst perimeter and mean and worst concave points — were among the ten most important in every fold, consistent with cytopathological markers of malignancy. The framework provides a validated, interpretable and stable classifier, and its validation shows that a well-specified linear model remains a strong baseline that ensembles should be required to match.

KEYWORDS: Breast Cancer; Fine-Needle Aspiration Cytology; Explainable AI; Ensemble Learning; SHAP; Model Validation; Probability Calibration; WDBC

INTRODUCTION

Breast cancer is the most commonly diagnosed cancer worldwide, with more than two million new cases a year [2]. Fine-needle aspiration cytology is a minimally invasive diagnostic step in

which nuclear morphology — size, shape and texture of cell nuclei — is assessed to distinguish benign from malignant lesions. Quantitative image analysis of FNA samples, as pioneered in the Wisconsin studies [3], [4], turns that morphology into numerical features that machine-learning models can use, and the resulting Wisconsin Diagnostic Breast Cancer (WDBC) dataset has become a standard benchmark [5].

High accuracy alone is not enough for clinical use. Clinicians need to understand why a model labels a sample malignant, to judge whether the reasoning matches cytopathology, and to trust that the model will behave consistently [6], [7]. Explainable machine learning addresses the first need through post-hoc attribution methods such as SHAP and LIME [8], [9]; rigorous validation addresses the second and third [10], [11].

Siddiqui *et al.* [1] compared explainable machine-learning models for cancer classification using cytological features, establishing a comparative basis for selecting interpretable, high-performing models on this kind of data. A natural next step is to move from comparison to implementation: to build an explainable ensemble on top of the compared models and to validate both its predictive performance and its explanations. Ensembles can combine the strengths of diverse learners and reduce variance [12], but they add complexity that must be justified by evidence and must not obscure the explanation [13].

This study implements and validates such a framework. Its contributions are:

1. an implementation of the comparative approach of Siddiqui *et al.* [1] as the base layer of an explainable ensemble framework comprising six base learners, a soft-voting ensemble and a stacking ensemble;
2. a validation protocol combining repeated cross-validation with corrected significance tests, a held-out test set with bootstrap confidence intervals, calibration analysis and a screening-oriented decision threshold;
3. an evaluation of the explanations themselves agreement of SHAP rankings across models and stability of the most important features across folds;
4. a transparent comparison showing where the ensemble does and does not improve on simpler models.

RELATED WORK

MACHINE LEARNING ON CYTOLOGICAL FEATURES

The WDBC dataset was created from digitised FNA images, with ten nuclear characteristics — radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension — each summarised by its mean, standard error and worst (largest) value [4]. Early work showed that linear separation of these features achieves high diagnostic accuracy [3], and the dataset has since been used to evaluate almost every family of classifier, including support vector machines [14], random forests [15], extremely randomised trees [16] and gradient boosting [17].

EXPLAINABILITY AND ENSEMBLES

SHAP attributes a prediction to its features using Shapley values from cooperative game theory [8], [18], and its tree and model-agnostic variants allow consistent comparison across model families [19]. Different explanation methods, and the same method applied to different models, can disagree [20], so explanations should themselves be evaluated. Ensembles — combining models by voting or by stacked generalisation [12], [21] — often improve stability, but where an interpretable model performs equally well, the added opacity may not be justified [13]. Siddiqui *et al.* [1] provide the comparative evidence on explainable models for cytological cancer classification on which the present framework builds.

MATERIALS AND METHODS

DATA

The WDBC dataset [4], [5] contains 569 FNA samples (212 malignant, 357 benign) described by 30 continuous nuclear features. Malignancy was coded as the positive class. The data were split once, stratified by diagnosis, into a development set of 426 samples and a held-out test set of 143 samples (53 malignant) that was not used for any model or threshold choice.

FRAMEWORK IMPLEMENTATION

Figure 1 shows the framework. The comparative layer implements the approach of Siddiqui *et al.* [1]: a panel of explainable model families is trained and compared under a common protocol. We used six learners — L2-regularised logistic regression, a support vector machine with radial basis kernel [14], distance-weighted k-nearest neighbours ($k = 9$), random forest (400 trees) [15], extra trees (400 trees) [16] and histogram-based gradient boosting (300 iterations, learning rate 0.05) [17], [22]. Features were standardised inside each training fold for the scale-sensitive models. The ensemble layer combines logistic regression, the support vector machine, random forest and gradient boosting, chosen to span linear, kernel and tree-based inductive biases, by soft voting (averaging predicted probabilities) and by stacking with a logistic-regression meta-learner trained on five-fold out-of-fold probabilities [21].

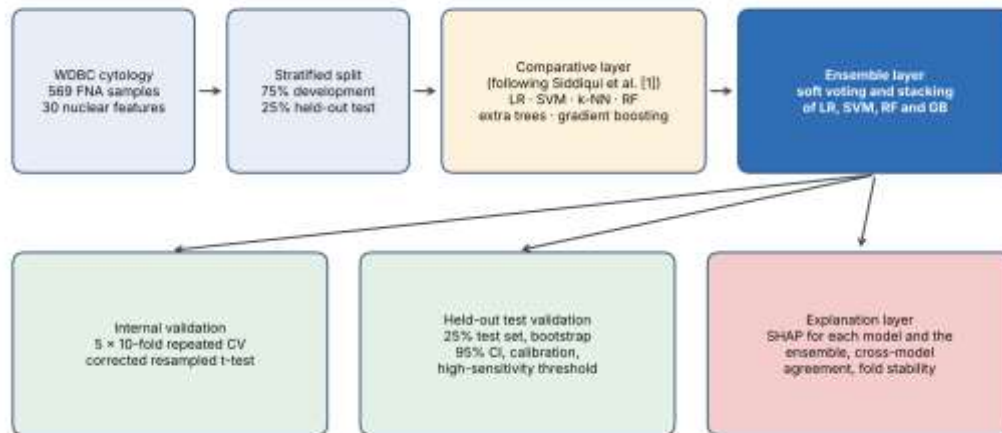


Figure 1. Explainable ensemble framework: comparative layer following Siddiqui et al. [1], ensemble layer, and the validation and explanation layers.

VALIDATION PROTOCOL

Internal validation used 5×10 -fold repeated stratified cross-validation on the development set [23], giving 50 estimates per model of accuracy, sensitivity, specificity, F1, the Matthews correlation coefficient (MCC) [24], the area under the ROC curve (AUC) and the Brier score [25]. Because cross-validation folds overlap, models were compared with the corrected resampled t-test of Nadeau and Bengio [26]. Held-out validation fitted each model on the full development set and evaluated it once on the test set, with 95% confidence intervals from 2,000 bootstrap resamples [27]. Calibration was assessed with reliability curves and the Brier score [28]. For screening use, a decision threshold giving out-of-fold sensitivity of at least 0.98 was chosen on the development set and then applied unchanged to the test set. Reporting follows the TRIPOD recommendations [10].

EXPLANATION ANALYSIS

SHAP values were computed for the probability of malignancy with a model-agnostic permutation explainer and a background of 100 development samples [8], for logistic regression, the support vector machine, random forest, gradient boosting and the soft-voting ensemble on all held-out test samples. Global importance was the mean absolute SHAP value. Agreement between models was measured by the Spearman rank correlation of their global importance vectors [20], and stability by refitting the ensemble in ten folds of the development set and recording how often each feature appeared among the ten most important.

RESULTS

INTERNAL VALIDATION

Table 1. Internal validation: 5×10 -fold repeated cross-validation on the development set (mean over 50 folds)

Model	Acc.	Sens.	Spec.	AUC ($\pm$ SD)	MCC ($\pm$ SD)	Brier
LR	0.975	0.953	0.988	0.994 ± 0.010	0.948 ± 0.055	0.021
SVM	0.974	0.956	0.984	0.996 ± 0.009	0.945 ± 0.046	0.020
k-NN	0.962	0.911	0.993	0.986 ± 0.017	0.920 ± 0.053	0.031
RF	0.955	0.931	0.970	0.988 ± 0.017	0.906 ± 0.057	0.034
ET	0.962	0.925	0.984	0.991 ± 0.011	0.919 ± 0.054	0.032
GB	0.963	0.935	0.981	0.991 ± 0.011	0.923 ± 0.054	0.031
Soft voting	0.975	0.952	0.988	0.994 ± 0.009	0.947 ± 0.041	0.021
Stacking	0.977	0.960	0.988	0.995 ± 0.009	0.953 ± 0.040	0.018

Note. Acc., accuracy; Sens., sensitivity; Spec., specificity; LR, logistic regression; SVM, support vector machine (RBF kernel); k-NN, k-nearest neighbours; RF, random forest; ET, extra trees; GB, gradient boosting.

Table 1 and Figure 2 summarise internal validation. The stacking ensemble (MCC 0.953 ± 0.040), soft-voting ensemble (0.947 ± 0.041), logistic regression (0.948 ± 0.055) and support vector machine (0.945 ± 0.046) formed the leading group; the tree-based learners and k-nearest neighbours trailed by 0.02–0.04 in MCC. In corrected resampled t-tests the soft-voting ensemble was significantly better than random forest in MCC (difference +0.041, $p = 0.032$) and accuracy ($p = 0.031$) and than extra trees in MCC ($p = 0.049$), but not different from logistic regression ($p = 0.96$), the support vector machine ($p = 0.88$) or the stacking ensemble ($p = 0.56$). The ensembles were, however, the most stable models: the standard deviation of MCC across folds was 0.041 for soft voting and 0.040 for stacking, against 0.055 for logistic regression and 0.057 for random forest.

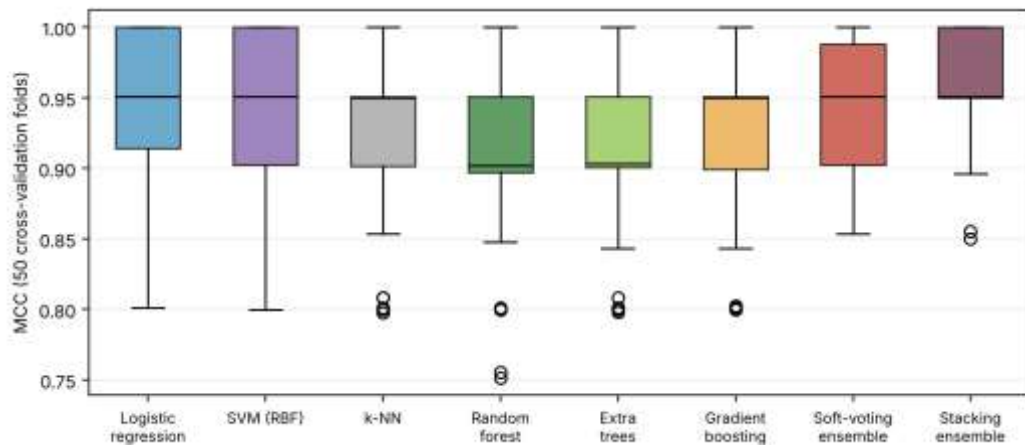


Figure 2. Distribution of the Matthews correlation coefficient across 50 cross-validation folds for each model.

HELD-OUT TEST VALIDATION

Table 2. Held-out test set performance (n = 143, 53 malignant; threshold 0.50)

Model	Acc.	Sens.	Spec.	F1	MCC	AUC	Brier
LR	0.965	0.925	0.989	0.951	0.925	0.996	0.022
SVM	0.979	0.943	1.000	0.971	0.955	0.995	0.021
k-NN	0.958	0.906	0.989	0.941	0.910	0.987	0.034
RF	0.965	0.906	1.000	0.950	0.926	0.996	0.031
ET	0.972	0.925	1.000	0.961	0.941	0.999	0.023
GB	0.958	0.887	1.000	0.940	0.912	0.996	0.027
Soft voting	0.979	0.943	1.000	0.971	0.955	0.997	0.019
Stacking	0.979	0.943	1.000	0.971	0.955	0.997	0.015

Note. 95% bootstrap confidence intervals for the soft-voting ensemble: accuracy 0.951–1.000, sensitivity 0.878–1.000, MCC 0.903–1.000, AUC 0.990–1.000, Brier 0.008–0.035.

On the held-out test set (Table 2, Figure 3), the soft-voting and stacking ensembles and the support vector machine each classified 140 of 143 samples correctly (accuracy 0.979), with no false positives (specificity 1.000) and 3 missed malignancies (sensitivity 0.943). The ensembles had the highest AUC (0.997) and the lowest Brier scores (soft voting 0.019, stacking 0.015). With a test set of this size the differences between leading models amount to one or two samples, and the bootstrap confidence intervals overlap widely; the test results therefore confirm the cross-validated ranking rather than establish new differences.

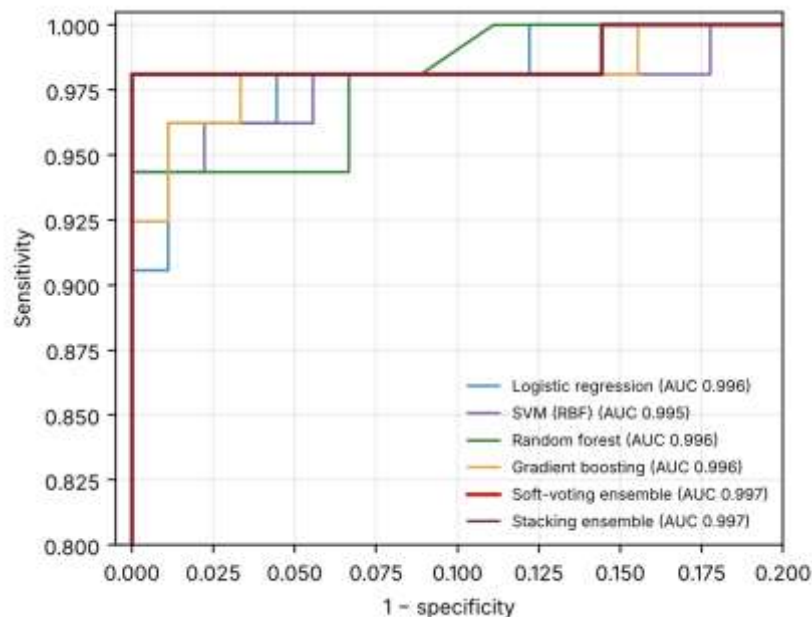


Figure 3. ROC curves on the held-out test set (upper-left region enlarged).

CALIBRATION AND SCREENING THRESHOLD

Predicted probabilities from the ensemble were well calibrated (Figure 4a): the reliability curve lay close to the diagonal and its Brier score (0.019) was lower than that of logistic regression (0.022) or random forest (0.031). Because a missed malignancy is more costly than a false alarm in a diagnostic pathway, a screening threshold was selected on development data (Figure 4b). The threshold of 0.17, which gave out-of-fold sensitivity of at least 0.98, achieved on the untouched test set a sensitivity of 0.981 (52 of 53 malignancies detected) and a specificity of 0.978 (2 false positives among 90 benign samples).

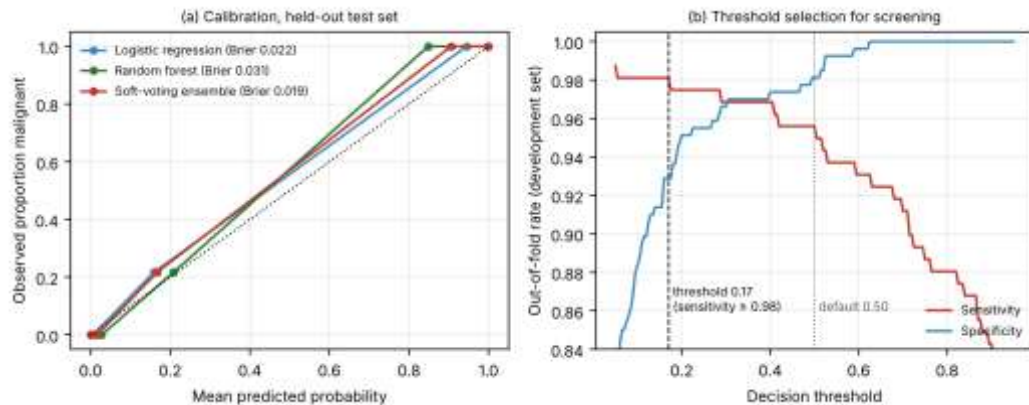


Figure 4. (a) Reliability curves on the held-out test set; (b) out-of-fold sensitivity and specificity of the ensemble across thresholds on the development set, with the selected screening threshold.

EXPLANATION ANALYSIS

Figure 5a shows the global SHAP importance of the soft-voting ensemble on the test set. The six most important features were worst perimeter, mean concave points, worst concave points, worst texture, worst area and worst radius — measures of nuclear size and contour irregularity that correspond to recognised cytological signs of malignancy, namely nuclear enlargement and indented, irregular nuclear membranes [3], [4]. Agreement between models was high but not uniform (Figure 5b): the rank correlation of global importance ranged from 0.58 (between logistic regression and gradient boosting) to 0.92, and the ensemble agreed with each of its members at $\rho = 0.76$ –0.92. The ensemble explanation thus acts as a consensus of its members, reducing the dependence of the explanation on any single model — a practical answer to the disagreement problem [20].

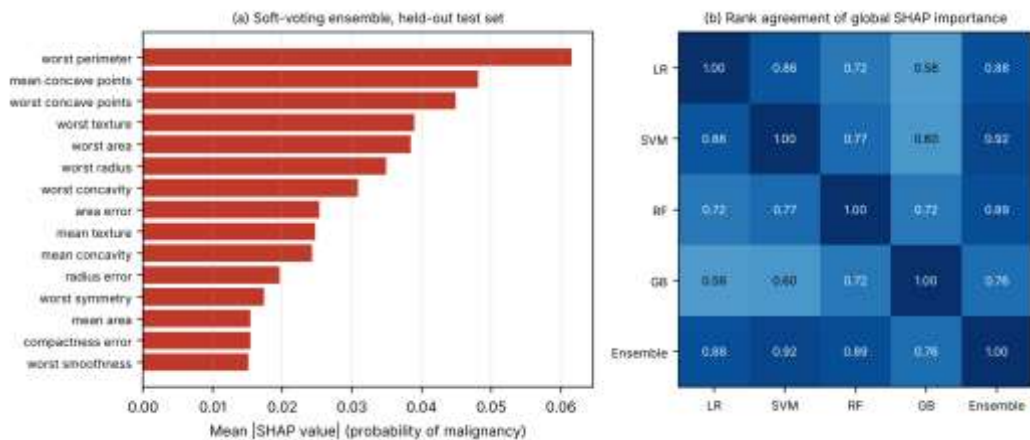


Figure 5. (a) Fifteen most important features of the soft-voting ensemble by mean absolute SHAP value; (b) Spearman rank agreement of global SHAP importance between models.

Explanations were also stable. Across ten refits of the ensemble on different development folds, 7 features appeared among the ten most important every time (worst perimeter, worst radius, worst area, worst concave points, mean concave points, worst concavity, worst texture), and 9 features appeared in at least eight of ten. A clinician reviewing explanations from this framework would therefore see the same core evidence regardless of the particular training sample.

DISCUSSION

Implementing the comparative approach of Siddiqui *et al.* [1] as the base layer of an ensemble produced a classifier that was accurate, well calibrated, stable and interpretable. The validation shows clearly where the ensemble adds value. It did not significantly outperform a well-specified logistic regression or support vector machine in discrimination, consistent with the observation that interpretable models often match complex ones on structured clinical data [13]. Its advantages lay elsewhere: the lowest fold-to-fold variability, the lowest Brier score, and explanations that reflect a consensus of diverse learners and are stable across training samples. These are properties that matter for clinical deployment even when headline accuracy is unchanged.

The results also illustrate why validation design matters. On a test set of 143 samples, the leading models differed by one or two cases; only repeated cross-validation with a corrected test could separate the tree-based learners from the leading group. Reporting confidence intervals, calibration and a pre-specified clinical threshold, as recommended by TRIPOD [10], gives a fuller and more honest picture than accuracy alone.

LIMITATIONS

The study uses a single, well-studied public dataset from one institution, collected decades ago with a specific image-analysis protocol; performance on contemporary, multi-site cytology data may differ, and external validation is required before clinical use. Hyperparameters were set to standard values rather than tuned by nested cross-validation [11], which keeps the comparison simple but may understate the potential of individual learners. SHAP explanations were

evaluated for agreement and stability, but not with cytopathologists, whose assessment is the ultimate test of clinical usefulness [6].

FUTURE WORK

External validation on recent multi-centre FNA datasets, a reader study in which cytopathologists assess the ensemble's explanations alongside its predictions, and prospective evaluation of the screening threshold in a diagnostic workflow are the natural next steps. Extending the framework to digital whole-slide cytology, where features are learned rather than hand-crafted, would test whether the explanation-stability analysis transfers.

CONCLUSION

This study implemented the comparative explainable-modelling approach of Siddiqui *et al.* [1] as the foundation of an explainable ensemble framework for breast-cancer classification from cytological features and validated it comprehensively. The soft-voting ensemble reached a held-out accuracy of 0.979 and AUC of 0.997, with well-calibrated probabilities, a screening threshold that detected 52 of 53 test malignancies, and SHAP explanations that agreed across models, were stable across folds and matched cytological markers of malignancy. Its accuracy did not significantly exceed that of logistic regression, so its case rests on stability, calibration and consensus explanation rather than on discrimination alone.

REFERENCES

- [1] Siddiqui, M. I. H., Rahman, M. S., Kabir, A. A., Mahmud, F. U., Rashid, S. U., & Shammah, R. S. (2023). Comparative analysis of explainable machine learning models for cancer classification using cytological features. *Journal of Medical and Health Studies*, 4(5), 110–150.
- [2] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249.
- [3] Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87(23), 9193–9196.
- [4] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In *Proceedings of SPIE 1905, Biomedical Image Processing and Biomedical Visualization* (pp. 861–870). SPIE.
- [5] 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>
- [6] Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, Article 310.

- [7] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745–e750.
- [8] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- [9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.
- [10] Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*, 350, g7594.
- [11] Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91.
- [12] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (Lecture Notes in Computer Science, Vol. 1857, pp. 1–15). Springer.
- [13] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- [14] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- [15] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [16] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42.
- [17] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- [18] Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games II* (pp. 307–317). Princeton University Press.
- [19] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.
- [20] Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). *The disagreement problem in explainable machine learning: A practitioner’s perspective*. arXiv. <https://arxiv.org/abs/2202.01602>
- [21] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259.
- [22] 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.
- [23] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137–1143). Morgan Kaufmann.
- [24] 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.
- [25] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- [26] Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, 52(3), 239–281.
- [27] Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
- [28] 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.