
Development and Cross-Dataset Replication of a Reproducible Explainable Machine Learning Pipeline for Breast Cancer Classification Using Cytological Measurements

Kowalska Wambui^{1*}, Faith Okafor²

¹University of Lagos, NIGERIA

²University of Cape Town, SOUTH AFRICA

ABSTRACT

Machine-learning models for breast-cancer cytology are rarely published in a form that others can rerun, and their explanations are rarely checked for stability. This study develops a reproducible, configuration-driven pipeline automated preprocessing, SHAP-based feature selection, a comparative model panel following the explainable-model comparison of Siddiqui et al. [1], a soft-voting ensemble, probability calibration, a screening threshold and a prediction interface returning a probability, a decision and the three features that drove it — and replicates it, unchanged, on two independent Wisconsin cytology datasets with different measurement systems: the Wisconsin Diagnostic Breast Cancer dataset (WDBC; 569 samples, 30 computer-measured nuclear features) and the original Wisconsin Breast Cancer dataset (WBCD; 699 samples, 9 pathologist-scored features, 16 missing values). On held-out test sets the ensemble reached, at a screening threshold chosen on development data, a sensitivity of 0.943 and specificity of 0.944 on WDBC (AUC 0.971) and a sensitivity of 0.983 and specificity of 0.957 on WBCD (AUC 0.986), with Brier scores of 0.032 on both. The pipeline was exactly reproducible under a fixed seed, and changing the seed altered test AUC by at most 0.0007. Feature selection retained 6–10 of 30 features on WDBC and 4–6 of 9 on WBCD, with mean pairwise Jaccard similarity of 0.74 and 0.73 across folds, and SHAP rankings were consistent across folds (mean Spearman ρ 0.83 and 0.75). In both datasets the explanation was dominated by nuclear size and shape irregularity — worst area, perimeter and concave points in WDBC; uniformity of cell size and shape and bare nuclei in WBCD consistent with cytopathological criteria. Prediction took about 0.8 ms per case and a full explained prediction about 0.26 s. Because the two datasets use different features, this is a replication of the pipeline rather than external validation of a single fitted model; validation on an independent measured-feature dataset remains necessary.

Keywords: Breast Cancer; Cytology; Explainable Machine Learning; Reproducibility; SHAP; Feature Selection; Probability Calibration; Replication

Introduction

Breast cancer is the most frequently diagnosed cancer worldwide [2], and fine-needle aspiration (FNA) cytology is an important, minimally invasive step in its diagnosis. Machine-learning classifiers trained on cytological features have reported very high accuracy for decades [3]–[5], and explainable methods such as SHAP now make it possible to show which features drive each prediction [6]. Comparative studies, including that of Siddiqui et al. [1], have evaluated explainable machine-learning models for cancer classification using cytological features and provide a basis for choosing among model families.

Two gaps limit the translation of this work. The first is reproducibility: published models are often described rather than released as executable pipelines, preprocessing and feature selection are frequently performed outside cross-validation, and results can depend on undocumented random choices [7], [8]. The second is validation: models are seldom tested beyond their development data [9], [10], and the explanations

themselves are rarely checked for stability [11]. A model whose explanation changes with the training sample cannot support a clinician's reasoning.

This study addresses both gaps with a pipeline built for reproducibility and tested on two independent Wisconsin cytology datasets. Its contributions are:

1. a configuration-driven pipeline in which every data-dependent step — imputation, scaling, SHAP-based feature selection, model fitting, calibration and threshold selection — is fitted on training data only, with a prediction interface that returns a probability, a screening decision and the three most influential features;
2. implementation of the comparative model panel of Siddiqui et al. [1] together with a soft-voting ensemble inside that pipeline;
3. replication of the unchanged pipeline on two datasets with different measurement systems, reporting held-out sensitivity and specificity, calibration, inference time, reproducibility under fixed and varied seeds, feature-selection stability and explanation consistency;
4. a clear statement of what this replication does and does not establish relative to external validation.

Related Work

The two Wisconsin datasets have different origins. The original Wisconsin Breast Cancer dataset (WBCD) records nine cytological characteristics, each graded from 1 to 10 by an examiner, for 699 FNA samples [3], [5]; the later Wisconsin Diagnostic Breast Cancer dataset (WDBC) records thirty features computed by image analysis of digitised nuclei — the mean, standard error and worst value of ten nuclear properties — for 569 samples [4], [12]. Both have been used extensively to benchmark classifiers, but almost always one at a time.

SHAP attributions [6], [13] are widely used to explain cytology classifiers, and comparative work such as that of Siddiqui et al. [1] has examined how explainable models differ on cytological features. Explanations from different models, or from the same model trained on different samples, can disagree [11], which motivates measuring their stability. Feature-selection stability is commonly measured by set-overlap indices [14], [15], and calibration — the agreement between predicted and observed risk — is essential when probabilities are shown to clinicians [16]–[18]. Reporting guidance for prediction models emphasises transparent development and validation [10], [19].

Materials and Methods

Datasets

Table 1. Datasets used for development and replication

Characteristic	WDBC	WBCD
Samples (malignant)	569 (212)	699 (241)
Features	30 computer-measured nuclear features (mean, SE and worst of 10 properties)	9 examiner-graded features, each scored 1–10
Missing values	0	16 (bare nuclei)
Development / held-out test	426 / 143 (53 malignant)	524 / 175 (60 malignant)
Source	UCI repository [12]; Street et al. [4]	UCI repository via mlbench [20]; Wolberg and Mangasarian [3]

Each dataset was split once, stratified by diagnosis, into a 75% development set and a 25% held-out test set (Table 1). The test sets were used only for final evaluation. Malignancy was the positive class.

Pipeline

Figure 1 shows the pipeline. A single configuration file fixes the random seed, the candidate numbers of selected features, the calibration method and the target screening sensitivity; the same code and configuration were applied to both datasets. Missing values are imputed with the training median, and scale-sensitive models standardise features inside each training fold. Feature selection ranks features by the mean absolute SHAP value of a gradient-boosting model fitted on training data [6], [21], then chooses the smallest number of top-ranked features whose five-fold cross-validated AUC is within 0.002 of the best.

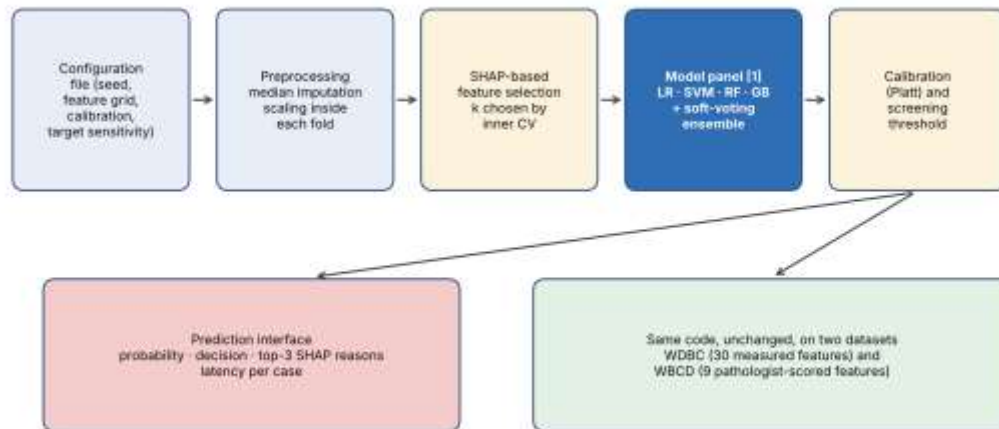


Figure 1. Reproducible explainable pipeline; the model panel follows the comparative approach of Siddiqui et al. [1].

The model panel implements the comparative approach of Siddiqui et al. [1], training logistic regression (LR), a support vector machine (SVM) [22], random forest (RF) [23] and gradient boosting (GB) [24] under one protocol, and combines them in a soft-voting ensemble [25]. Every model is calibrated by Platt scaling with internal five-fold cross-validation [26]. A screening threshold is chosen as the highest threshold giving out-of-fold sensitivity of at least 0.98 on the development set. The prediction interface returns, for a single sample, the calibrated probability, the decision at the screening threshold and the three features with the largest SHAP contributions, computed with a permutation explainer over a 60-sample background.

Evaluation

For each dataset and model we report held-out sensitivity and specificity at the screening threshold, AUC, the Brier score [27], the calibration slope and expected calibration error (ECE) [18]. Reproducibility was assessed by refitting the ensemble with the same seed and comparing hashed predictions, and by refitting with five different seeds. Stability was assessed by refitting the ensemble in ten folds of each development set and recording the selected features (mean pairwise Jaccard similarity) and the rank correlation of SHAP importance between folds [11], [14]. Inference time was measured per case for prediction alone and for prediction with explanation. All analyses used scikit-learn [28] and SHAP.

Results

Held-out performance

Table 2. Held-out test performance at the screening threshold chosen on development data

Dataset	Model	Sens.	Spec.	AUC	Brier	Slope	ECE
WDBC	LR	0.943	0.933	0.976	0.042	1.28	0.040
WDBC	SVM	0.943	0.911	0.972	0.037	1.10	0.034
WDBC	RF	0.962	0.933	0.966	0.033	1.19	0.030
WDBC	GB	0.943	0.944	0.976	0.033	1.19	0.049
WDBC	Ensemble	0.943	0.944	0.971	0.032	1.13	0.036
WBCD	LR	0.967	0.957	0.989	0.033	1.01	0.040
WBCD	SVM	0.967	0.957	0.979	0.032	1.13	0.030
WBCD	RF	0.983	0.948	0.989	0.034	1.12	0.044
WBCD	GB	0.983	0.957	0.987	0.033	1.13	0.019
WBCD	Ensemble	0.983	0.957	0.986	0.032	1.11	0.038

Note. Sens., sensitivity; Spec., specificity; Slope, calibration slope (1 = ideal); ECE, expected calibration error over ten quantile bins.

Table 2 and Figure 2 summarise held-out performance. On WDBC the ensemble selected six features and, at its screening threshold of 0.14, detected 50 of 53 malignancies with 5 false positives among 90 benign samples (sensitivity 0.943, specificity 0.944, AUC 0.971). On WBCD it selected four features and detected 59 of 60 malignancies with 5 false positives among 115 (sensitivity 0.983, specificity 0.957, AUC 0.986). No model was consistently best: on WDBC logistic regression and gradient boosting had the highest AUC (0.976), and on WBCD logistic regression and random forest did (0.989). The ensemble had the lowest or equal-lowest Brier score on both datasets (0.032).

On WDBC, test AUC (0.971) was lower than the cross-validated AUC on the development set (0.995 ± 0.007). With only 53 malignant test cases, a handful of malignancies with low scores — 4 were missed at the conventional 0.50 threshold — is enough to move the AUC by this amount; on WBCD the two estimates agreed closely (0.986 vs 0.990). The selection rule deliberately favours the smallest adequate feature set, trading a small amount of discrimination for interpretability and stability; its tolerance can be changed in the configuration file.

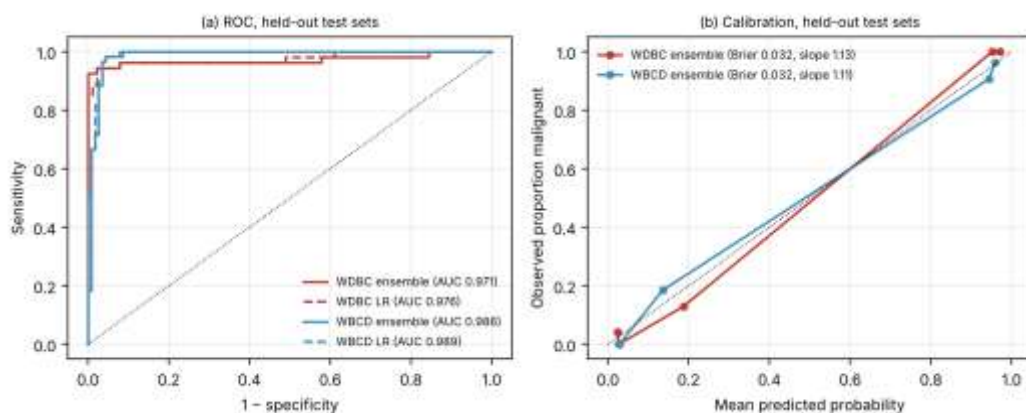


Figure 2. (a) ROC curves and (b) calibration curves of the ensemble on the held-out test sets of both datasets; dashed ROC lines show logistic regression.

Calibration

After Platt scaling, held-out probabilities were well calibrated on both datasets (Figure 2b): the ensemble's calibration slope was 1.13 on WDBC and 1.11 on WBCD, and its ECE 0.036 and 0.038. Slopes slightly above 1 indicate predictions that are, if anything, marginally under-confident. Within the ten development folds, calibration slopes varied widely (mean 3.07 ± 2.38 on WDBC and 1.70 ± 0.92 on WBCD), because each fold contains few intermediate-probability cases; calibration estimates from small validation sets should therefore be read with caution.

Reproducibility

Refitting the ensemble with the same seed reproduced the held-out predictions exactly on both datasets (identical prediction hashes). Changing the seed altered test AUC by a standard deviation of 0.0002 on WDBC and 0.0007 on WBCD, and the Brier score by at most 0.0001; sensitivity at the screening threshold did not change, and the standard deviation of specificity was at most 0.009. The pipeline's results therefore do not depend on a fortunate random seed.

Feature-selection stability and explanation consistency

Figure 3 shows how often each feature was selected across ten development folds and the SHAP importance of the selected features on the test sets. On WDBC, 5 features were selected in every fold (mean concave points, worst texture, worst area, worst concavity, worst concave points) and the number selected ranged from 6 to 10; mean pairwise Jaccard similarity of the selected sets was 0.74 (minimum 0.56). On WBCD, 2 features were selected in every fold (uniformity of cell size, bare nuclei), with Jaccard similarity 0.73 (minimum 0.60). SHAP importance rankings were consistent across folds, with mean Spearman correlations of 0.83 on WDBC and 0.75 on WBCD.

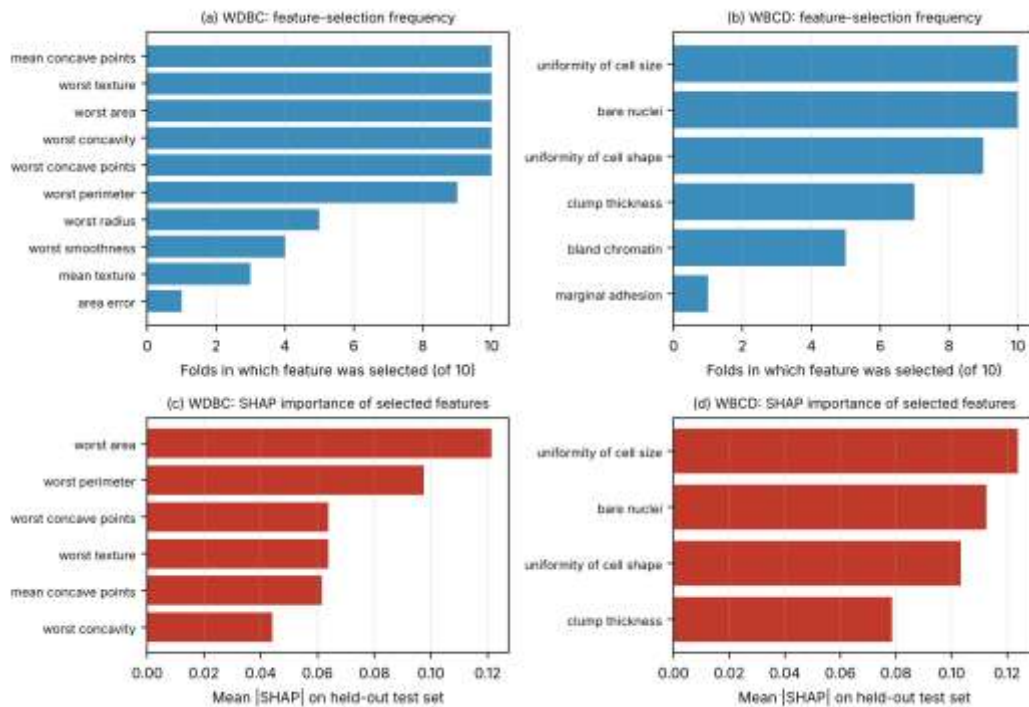


Figure 3. Feature-selection frequency across ten development folds (a, b) and SHAP importance of the selected features on the held-out test sets (c, d).

Although the two datasets share no features, their explanations converged on the same cytological concept. On WDBC the most influential features were worst area, worst perimeter and worst concave points — measures of nuclear size and contour irregularity [4]. On WBCD they were uniformity of cell size, bare nuclei and uniformity of cell shape — examiner grades of pleomorphism in cell size and shape and of nuclei stripped of cytoplasm [3]. Both point to enlarged, irregular nuclei as the main evidence of malignancy, which supports the explanations' clinical plausibility.

Prediction interface and inference time

Table 3. Example outputs of the prediction interface for one malignant test case from each dataset

Dataset	Probability	Decision (threshold)	Top-3 reasons (SHAP contribution)	Latency
WDBC	0.970	refer as suspicious (0.14)	worst concave points (+0.24); worst perimeter (+0.20); mean concave points (+0.05)	301 ms
WBCD	0.959	refer as suspicious (0.15)	uniformity of cell shape (+0.29); uniformity of cell size (+0.23); bare nuclei (+0.07)	379 ms

Table 3 shows the interface output for one malignant test case from each dataset. Prediction alone took about 0.8 ms per case on both datasets, and a prediction with its SHAP explanation a median of 260 ms (WDBC) and 258 ms (WBCD) on a two-core processor, well within the time available in a cytology workflow.

Discussion

The pipeline met its reproducibility goals: identical results under a fixed seed, negligible sensitivity to the seed, stable feature selection and consistent explanations, achieved by fitting every data-dependent step inside the training data [7], [8]. Implemented within this pipeline, the comparative model panel of Siddiqui et al. [1] again showed that no single model family dominates on cytological data; the calibrated ensemble offered the most reliable probabilities rather than the highest AUC, consistent with the view that simpler models remain strong baselines in this setting [29].

The replication across two datasets is the main new evidence. The same code, without modification, produced accurate, calibrated and stable models from thirty measured features and from nine examiner grades, and its explanations identified the same underlying cytological signal in both. This is evidence that the pipeline transfers across measurement systems. It is not external validation of a single fitted model, because the WDBC model cannot be applied to WBCD features; that requires an independent dataset with compatible measurements [9], [10].

Limitations

Both datasets come from the same institution and era, are relatively small, and contain mostly clear-cut cases; their held-out test sets (143 and 175 samples) give wide confidence intervals around sensitivity and specificity. The 16 missing values in WBCD were imputed rather than modelled. Hyperparameters were fixed by configuration, not tuned. Explanations were evaluated for stability and plausibility but not by cytopathologists. Most importantly, no independent dataset with WDBC-compatible measured features was available, so true external validation remains to be done.

Future Work

The priority is external validation: applying the released WDBC model, unchanged, to independent FNA images from another centre, after extracting the same nuclear measurements. A reader study in which

cytopathologists rate the interface's explanations, and prospective evaluation of the screening threshold, would follow. The pipeline is designed so that both steps require only a new data file and no code changes.

Conclusion

A reproducible, configuration-driven explainable pipeline — implementing the comparative model panel of Siddiqui et al. [1] with SHAP-based feature selection, calibration, a screening threshold and a prediction interface produced accurate and well-calibrated breast-cancer classifiers on two independent Wisconsin cytology datasets with different measurement systems. Results were exactly reproducible, robust to the random seed and supported by stable explanations that converged on nuclear size and shape irregularity in both datasets. The pipeline is ready for the external validation on independent measured-feature data that clinical use requires.

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] Bennett, K. P., & Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1), 23–34.
- [6] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- [7] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research. *Journal of Machine Learning Research*, 22(164), 1–20.
- [8] Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91.
- [9] Justice, A. C., Covinsky, K. E., & Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, 130(6), 515–524.
- [10] Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *European Heart Journal*, 35(29), 1925–1931.
- [11] 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>

- [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] 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.
- [14] Kuncheva, L. I. (2007). A stability index for feature selection. In *Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications* (pp. 390–395). ACTA Press.
- [15] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517.
- [16] 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.
- [17] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). ACM.
- [18] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR.
- [19] 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.
- [20] Leisch, F., & Dimitriadou, E. (2010). *mlbench: Machine learning benchmark problems* (R package). Comprehensive R Archive Network.
- [21] 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.
- [22] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- [23] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [24] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- [25] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (Lecture Notes in Computer Science, Vol. 1857, pp. 1–15). Springer.
- [26] Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in Large Margin Classifiers* (pp. 61–74). MIT Press.
- [27] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- [28] 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.

- [29] 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.