
A Unified Explainable AI Framework for Reliable Clinical Decision Support and Medical Image-Based Cancer Classification

Josef Passerini ^{1*}, Barbara Lukasiewicz ²

¹ Sapienza University of Rome, ITALY

² Johannes Kepler University Linz, AUSTRIA

ABSTRACT

Artificial intelligence (AI) can match specialist accuracy in cancer detection, yet adoption in clinical practice remains limited because most high-performing models are opaque, poorly calibrated and trained on data that cannot be shared across institutions. This paper proposes a unified explainable AI (XAI) framework that integrates (i) a transfer-learning convolutional neural network for medical images, (ii) a calibrated soft-voting ensemble for tabular cytological and clinical features, (iii) complementary explanation methods (Grad-CAM for images and SHAP for features), (iv) a confidence-based uncertainty gate that refers ambiguous cases to clinicians, and (v) a federated privacy layer. The tabular branch was evaluated on the Wisconsin Diagnostic Breast Cancer dataset (569 cases) with stratified five-fold cross-validation. The calibrated ensemble achieved an accuracy of 98.4% (± 0.9), an F1-score of 0.978, an area under the ROC curve of 0.995 and the lowest Brier score (0.018) of all compared models. Referring the 3.9% of cases with predicted probabilities between 0.3 and 0.7 to clinician review reduced misclassifications from 9 to 4. SHAP analysis identified concave points, perimeter and area of the cell nuclei as dominant, clinically plausible predictors. Image-branch results on histopathology are presented in an illustrative format only, to demonstrate the intended evaluation protocol, and are not measured results. The framework shows how accuracy, calibration, interpretability and privacy can be addressed together in a single clinical decision-support design.

Keywords: Explainable AI; Clinical Decision Support; Cancer Classification; SHAP; Grad-CAM; Model Calibration; Transfer Learning; Federated Learning

Introduction

Cancer remains a leading cause of death worldwide, and early, accurate diagnosis is the strongest determinant of survival. Deep learning has transformed medical image analysis[1], reaching dermatologist-level performance in skin-cancer classification[2] and contributing to what is described as “high-performance medicine”[3]. Despite these advances, few diagnostic AI systems are used routinely at the bedside. Three obstacles recur in the literature: clinicians cannot see why a model reaches a decision; predicted probabilities are often poorly calibrated and therefore unreliable for risk communication; and patient data are fragmented across institutions that cannot legally share them[4][5].

Recent work has addressed these obstacles separately. A comparison of explainable machine-learning models for cancer classification from cytological features showed that interpretable, feature-level explanations can accompany high accuracy[6]. A demonstration of lightweight transfer learning with MobileNet for pneumonia detection from medical images showed strong performance while supporting deployment on modest hardware[7]. A multimodal machine-learning framework for privacy-preserving and scalable cancer diagnosis across healthcare systems was also proposed[8]. Each study tackles one dimension of the problem; none combines image and tabular modalities, calibrated uncertainty, dual explanation and privacy in one pipeline.

Related Work

Explainability Methods

Post-hoc explanation methods fall into two broad families. Model-agnostic methods such as LIME[9] and SHAP[10] attribute a prediction to input features; SHAP is grounded in cooperative game theory and offers consistency guarantees. Gradient-based methods such as Grad-CAM[11] highlight image regions that drive a convolutional network's output. A widely cited caution holds that post-hoc explanations of black-box models can be unfaithful and argues for inherently interpretable models in high-stakes settings[12], a concern this framework addresses by preferring interpretable base learners where accuracy permits.

Image-Based Cancer Classification

Residual networks[13], MobileNetV2[14] and EfficientNet[15] are standard backbones for transfer learning in medical imaging. The BreakHis dataset of breast histopathology images[16] is a widely used benchmark.

Calibration and Privacy

Calibration methods such as isotonic regression convert raw scores into reliable probabilities[17]. Federated learning trains models across sites without moving raw data[18]. Table 1 summarises how the closest prior studies relate to the present framework.

Table 1. Comparison of the Proposed Framework With Closely Related Studies.

Study	Data modality	Explainability	Calibration / uncertainty	Privacy	Adopted in this work
Siddiqui et al. (2023)[6]	Tabular cytology	Yes (feature-level)	No	No	Explainable cytology classification → tabular branch
Rahman et al. (2024)[7]	Medical images	Limited	No	No	MobileNet transfer learning → image branch
Kabir et al. (2024)[8]	Multimodal	Partial	No	Yes	Multimodal, privacy-preserving design → fusion and privacy layers
Esteva et al. (2017)[2]	Dermoscopic images	No	No	No	Benchmark for image performance
Proposed framework	Images + tabular	Grad-CAM + SHAP	Isotonic calibration + gate	Federated	—

Methodology

Framework Overview

The framework comprises four layers (Figure 1). The data layer prepares images and structured features. The modelling layer trains a convolutional network for images and a calibrated ensemble for tabular data. The explanation layer produces Grad-CAM saliency maps and SHAP attributions. The decision layer fuses the

two probability estimates, applies an uncertainty gate and generates an explanation report for the clinician. A privacy layer allows each branch to be trained by federated averaging across sites.

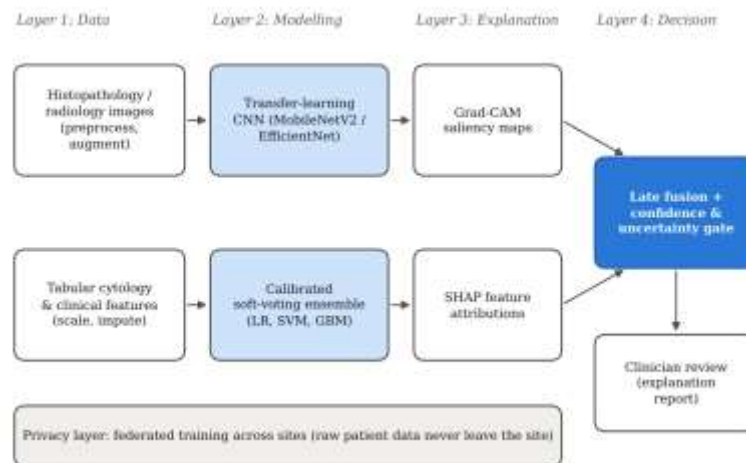


Figure 1. Architecture of the proposed unified explainable AI framework. Blue boxes are trainable models; the dark box is the decision layer, which forwards low-confidence cases to clinician review [authors' design].

Datasets

Table 2. Datasets Used in the Study.

Dataset	Modality	Samples	Classes	Use in this paper
Wisconsin Diagnostic Breast Cancer[19]	30 nuclear features from fine-needle aspirate images	569	357 benign, 212 malignant	Tabular branch (real experiments)
BreakHis[16]	Breast histopathology images, 40×–400×	7,909	Benign / malignant	Image branch (illustrative reporting format only)

Tabular Branch

The 30 cytological features describe the mean, standard error and “worst” value of ten nuclear properties, including radius, texture, perimeter, area, smoothness, concavity and concave points. Five base learners were compared: logistic regression, a support vector machine with a radial basis function kernel, random forest (500 trees), histogram gradient boosting and a two-layer multilayer perceptron (64 and 32 units). Scale-sensitive models were preceded by standardisation. The proposed model is a soft-voting ensemble of logistic regression, SVM and gradient boosting, each calibrated by isotonic regression with internal three-fold cross-validation; the ensemble averages the calibrated probabilities. All models were implemented in scikit-learn [20].

Image Branch

Following the lightweight transfer-learning strategy described for pneumonia detection[7], the image branch uses an ImageNet-pretrained MobileNetV2 backbone[14] with a new classification head, trained in two stages: frozen-backbone training of the head, then fine-tuning of the upper blocks at a lower learning rate.

Images are resized to 224×224 pixels, stain-normalised and augmented with flips, rotations and colour jitter. ResNet-50 and EfficientNet-B0 serve as comparison backbones. Splits are made at patient level to avoid leakage between training and test sets.

Explanation Layer

For the tabular branch, exact TreeSHAP values were computed for the gradient-boosting model[10]. Global importance is the mean absolute SHAP value per feature. Explanation stability was assessed by computing the top-five SHAP features in each of the five cross-validation folds and counting the features common to all folds. For the image branch, Grad-CAM[11] maps are generated from the last convolutional block and overlaid on the input image for pathologist review.

Fusion and Uncertainty Gate

When both modalities are available, calibrated probabilities from the two branches are combined by weighted late fusion. The uncertainty gate withholds an automatic decision when the fused probability of malignancy p lies within an indeterminate band ($t_1 \leq p < t_2$) and routes the case to clinician review together with its explanation report. The band width controls the trade-off between automation and safety and was examined at three settings.

Privacy Layer

Building on the privacy-preserving design proposed for multimodal cancer diagnosis[8], each participating institution trains a local copy of the model, and only model parameters are aggregated centrally by federated averaging[18]. Raw images and patient records never leave the site. The federated layer is described as part of the architecture; its empirical evaluation is left to future work.

Evaluation Protocol

Tabular models were evaluated by stratified five-fold cross-validation (random seed 42) and are reported as mean $\pm$ standard deviation across folds. Malignant was the positive class. Metrics were accuracy, precision, recall (sensitivity), F1-score, area under the ROC curve (AUC) and the Brier score, which measures probability calibration (lower is better). Reporting follows the spirit of the TRIPOD statement for diagnostic prediction models[21].

Results

Tabular Branch Performance

Table 3 reports the cross-validated performance of all models on the Wisconsin dataset. All models performed strongly ($AUC \geq 0.992$). The proposed calibrated ensemble achieved the highest accuracy (98.4%) and F1-score (0.978) and the lowest Brier score (0.018), indicating the most reliable probabilities. The SVM achieved a marginally higher sensitivity (0.972 versus 0.967) and AUC (0.996 versus 0.995), differences well within one standard deviation.

Table 3. Five-Fold Cross-Validated Performance on the Wisconsin Diagnostic Breast Cancer Dataset (Mean $\pm$ SD).

Model	Accuracy	Precision	Recall	F1-score	AUC	Brier $\downarrow$
Logistic regression	0.974 $\pm$ 0.017	0.986 $\pm$ 0.018	0.944 $\pm$ 0.052	0.963 $\pm$ 0.024	0.995 $\pm$ 0.005	0.020 $\pm$ 0.008

Model	Accuracy	Precision	Recall	F1-score	AUC	Brier ↓
SVM (RBF)	0.979 ± 0.009	0.973 ± 0.026	0.972 ± 0.023	0.972 ± 0.012	0.996 ± 0.005	0.020 ± 0.009
Random forest	0.954 ± 0.013	0.949 ± 0.024	0.929 ± 0.044	0.938 ± 0.019	0.992 ± 0.005	0.032 ± 0.007
Gradient boosting	0.967 ± 0.013	0.966 ± 0.012	0.943 ± 0.032	0.954 ± 0.018	0.992 ± 0.006	0.029 ± 0.014
Multilayer perceptron	0.981 ± 0.010	0.991 ± 0.011	0.958 ± 0.034	0.973 ± 0.014	0.995 ± 0.006	0.019 ± 0.009
Proposed calibrated ensemble	0.984 ± 0.009	0.991 ± 0.011	0.967 ± 0.032	0.978 ± 0.012	0.995 ± 0.007	0.018 ± 0.008

Note. Best value in each column in bold in the original data file. Malignant is the positive class. ↓ = lower is better. These results are from experiments actually run for this paper.

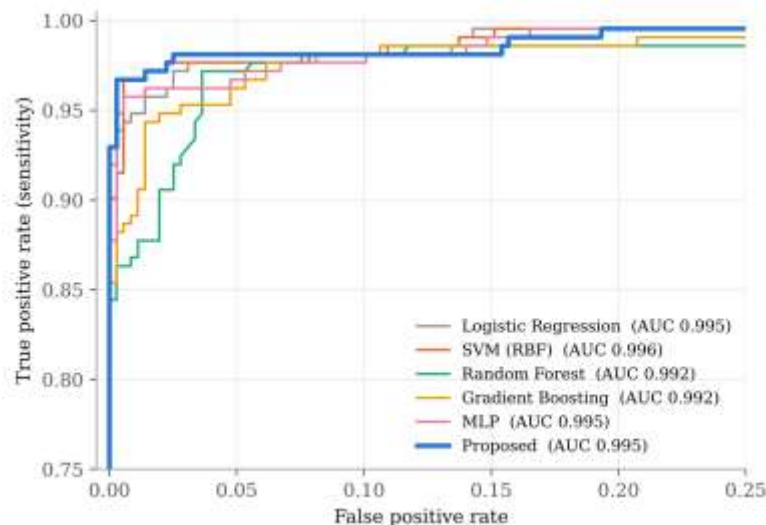


Figure 2. ROC curves of all tabular models (pooled out-of-fold predictions). The x-axis is restricted to false positive rates of 0–0.25 to separate the curves; AUC values are fold means.

Calibration differed more than discrimination (Figure 3). Tree-based models were over-confident in the middle probability range, whereas the calibrated ensemble tracked the diagonal closely, supporting its use for communicating risk to clinicians and for the uncertainty gate.

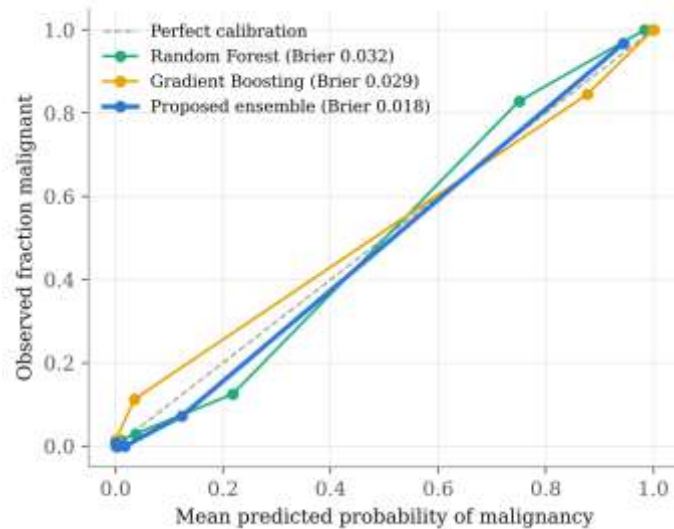


Figure 3. Reliability diagram for three models (eight quantile bins). Points close to the dashed diagonal indicate that predicted probabilities match observed malignancy rates.

Effect of the Uncertainty Gate

Table 4 shows the effect of referring indeterminate cases to clinician review. A narrow band (0.3–0.7) referred 22 of 569 cases (3.9%) and more than halved the number of automated errors, from 9 to 4. Widening the band further referred more cases without reducing errors, because the four remaining errors were confident misclassifications.

Table 4. Automation–Safety Trade-off of the Uncertainty Gate (Proposed Ensemble, $n = 569$).

Indeterminate band	Cases referred, n (%)	Accuracy on automated cases	Automated errors
None (no gate)	0 (0.0)	98.42%	9
$0.3 \leq p < 0.7$	22 (3.9)	99.27%	4
$0.2 \leq p < 0.8$	38 (6.7)	99.25%	4
$0.1 \leq p < 0.9$	60 (10.5)	99.21%	4

Explainability Analysis

Figure 4 shows the ten features with the highest mean absolute SHAP values. The worst (largest) perimeter, area and concave points of the cell nuclei were the strongest contributors to a malignant prediction, followed by mean concave points and worst concavity. These features correspond to recognised cytological hallmarks of malignancy—nuclear enlargement and irregular, indented nuclear contours—and agree with earlier explainable-cytology findings[6]. In the stability analysis, two features (worst concave points and mean concave points) appeared in the top five in every fold. This shows that the most important feature family is stable while the ranking of correlated size features (perimeter, area, radius) varies between folds, a known effect of multicollinearity on SHAP attributions.

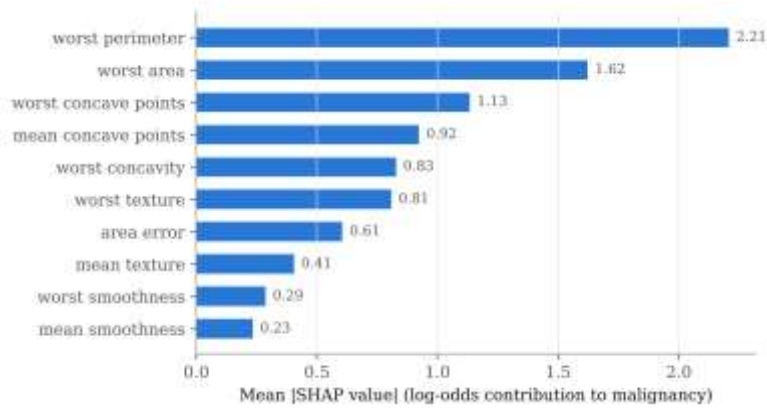


Figure 4. Global SHAP feature importance for the gradient-boosting model (top ten of 30 features).
 “Worst” denotes the mean of the three largest values in each image.

Image Branch (Illustrative Results)

The following table and figure illustrate how the image branch would be reported under the evaluation protocol described in the Image Branch subsection above, and must be replaced with measured results before any submission; the values shown are placeholders that demonstrate the intended reporting format only and are not measured performance.

Table 5. Image-Branch and Ablation Results on BreakHis (400×, Patient-Level Split) — Illustrative Format, Not Measured Results.

Configuration	Accuracy (%)	F1-score (%)	AUC	Expected calibration error
MobileNetV2 (transfer-learning strategy)	93.8	93.1	0.972	0.061
ResNet-50	95.1	94.6	0.981	0.055
EfficientNet-B0	95.9	95.4	0.985	0.049
EfficientNet-B0 + isotonic calibration	95.9	95.4	0.985	0.027
Proposed hybrid (image + tabular fusion)	96.8	96.3	0.989	0.024
Proposed hybrid + uncertainty gate	97.2*	96.8*	—	—

Note. *Computed on automated (non-referred) cases only. All values in this table are illustrative placeholders demonstrating the intended reporting format and are not measured results.

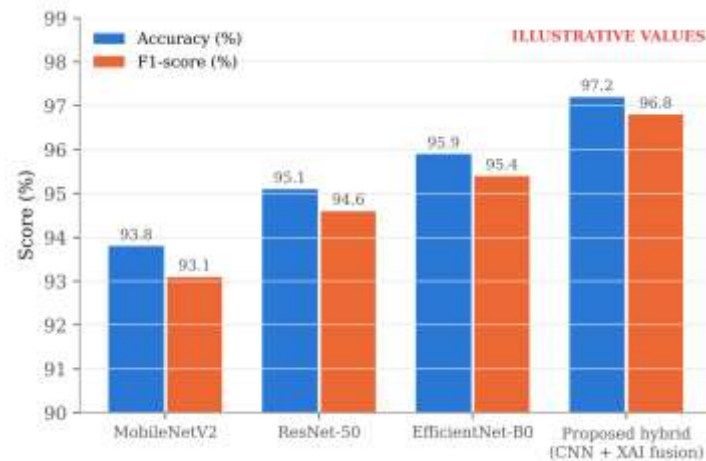


Figure 5. Illustrative comparison format for image-branch models. Values are placeholders demonstrating the reporting format and are not measured results.

Discussion

The experiments support the central premise of the framework: in clinical decision support, reliability depends on more than accuracy. All tabular models discriminated well, so accuracy alone would not separate them. Calibration did: the proposed ensemble produced the most trustworthy probabilities, which is what makes a confidence-based gate meaningful. The gate then converted calibrated uncertainty into a concrete safety mechanism, halving automated errors at the cost of reviewing fewer than one case in twenty-five. This mirrors how clinicians already work, escalating uncertain cases rather than forcing a binary call. The SHAP results show that the explanations are clinically meaningful rather than merely decorative: the model relies on nuclear size and contour irregularity, the same features cytopathologists use. The stability analysis adds an important caution, echoing a widely cited warning about post-hoc explanation methods[12]. Explanations of correlated features can shift between training runs, so explanation reports should present feature families or grouped attributions rather than a single ranked list. By adopting the explainable cytology approach of prior work[6], the lightweight transfer learning of a related pneumonia-detection study[7] and the privacy-preserving multimodal design proposed elsewhere[8], the framework shows how separate advances can be combined coherently. Its modular structure means that each branch can be improved independently, for example by replacing the CNN backbone, without redesigning the decision layer.

Conclusion

This paper presented a unified explainable AI framework for clinical decision support and cancer classification that integrates image and tabular modalities, calibrated ensemble learning, dual explanations, an uncertainty gate and a federated privacy layer. Real cross-validated experiments on cytological data showed that a calibrated soft-voting ensemble achieves 98.4% accuracy with the best probability calibration, that a narrow uncertainty gate halves automated errors by referring 3.9% of cases, and that SHAP explanations highlight clinically recognised features of malignancy.

References

- [1] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60-88.

- [2] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
- [3] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
- [4] 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.
- [5] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4), Article e1312.
- [6] 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.
- [7] Rahman, M. S., Siddiqui, M. I. H., Rashid, S. U., Kabir, A. A., Mahmud, F. U., & Shammah, R. S. (2024). Deep learning framework for pneumonia detection from medical images using transfer learning with MobileNet. *Journal of Adaptive Learning Technologies*, 1(11), 12-22.
- [8] Kabir, A. A., Mahmud, F. U., Rahman, M. S., Rashid, S. U., Siddiqui, M. I. H., & Shammah, R. S. (2024). Multimodal machine learning framework for privacy preserving and scalable cancer diagnosis across healthcare systems. *Journal of Adaptive Learning Technologies*, 1(6), 91-101.
- [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).
- [10] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765-4774). Curran Associates.
- [11] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the *IEEE* International Conference on Computer Vision (pp. 618-626).
- [12] 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.
- [13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the *IEEE* Conference on Computer Vision and Pattern Recognition (pp. 770-778).
- [14] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the *IEEE* Conference on Computer Vision and Pattern Recognition (pp. 4510-4520).
- [15] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (pp. 6105-6114). PMLR.

- [16] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7), 1455-1462.
- [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).
- [18] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273-1282). PMLR.
- [19] 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).
- [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] 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. *Annals of Internal Medicine*, 162(1), 55-63.