



International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Explainable AI for Disease Diagnosis Using Machine Learning

Samira Farooqui¹, Mrs. Deepti Gupta²

¹ School of Computer Application, Babu Banarasi Das University, Lucknow, India farooqui.samira29@gmail.com

² Under the supervision of Assistant Professor, School of Computer Application, Babu Banarasi Das University, Lucknow, India deepti.gupat.16@gmail.com

ABSTRACT—

A model's predictive performance is not enough to make sure whether machine learning models are appropriate for diagnosis. This work provides a simple leakage-free pipeline for binary classification and subsequent explanation for the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. It is tested on the WDBC dataset. In this paper, we compare four classifiers, which include logistic regression, support vector machine (SVM), k-nearest neighbors, and random forest. To ensure that the experiment is conducted honestly, stratification is used when splitting the data, and all of the preprocessing techniques are applied only to the training set to avoid any leakage of information from the test set. The models are chosen based on the results of the five-fold cross-validation and tested once on the independent test data. The SVM achieved the highest accuracy (97.37%) and F1-score (96.30%), while logistic regression achieved the highest ROC-AUC (99.60%). SHAP and permutation importance revealed that the models are mostly dependent on shape and texture features of tumors, including the worst texture, concavity, and radius.

Index: Explainable AI, Disease Diagnosis, Machine Learning, SHAP, Breast Cancer

I. INTRODUCTION

Machine learning techniques can detect latent patterns in clinical laboratory data, genetic data, and image data. Nevertheless, a highly accurate classification or regression score obtained by a model in the context of diagnostics is not sufficient for evaluating whether the model is accurate, calibrated, unbiased, and safe for practical application in a medical environment. Explanation may assist in identifying features that have affected the prediction made; however, the explanation is not proof that the model has learned a valid medical process. Recent research on breast cancer XAI stresses the importance of analyzing explanation behavior along with performance and bias of the dataset [2].

The classification of breast cancer can be considered an interesting case of methodological application since WDBC is a digital representation of the fine-needle aspirate using quantitative tumor morphology measurements. In previous studies, comparisons of classical classifiers and post-hoc approaches have been conducted for the Wisconsin datasets and have provided important morphological variables but showed variable results across datasets and different protocols [1]. The study of GraphX-Net is based on the problem of recurrence prediction and the relational model of patients, which is a different task than the one performed here [3].

The following paper considers one pipeline that includes auditing, leakage-controlled preprocessing, stratified modeling, held-out testing, and global/local explanations. Four different classifiers are evaluated under a single-state evaluation protocol. The research is about a process that could be used for benchmarking but not for deploying clinical solutions.

II. RELATED WORK

The results from past research on breast cancer indicate that the good scores will depend on the dataset, the target, the split, and the feature representations. There is an analysis of various models in which the best classification models on both WBC

and WDBC differ, and some critical variables that are considered include 'bare nuclei' and 'worst area'. These results cannot be generalized to another experiment without using the same dataset and protocol [1]. In the review of XAI for breast cancer datasets, the methods mentioned are SHAP, LIME, and Grad-CAM, among others [2].

Unlike the WDBC dataset used to distinguish between benign and malignant tumors, GraphX-Net attempts to predict breast cancer recurrence. The method makes use of a graph structure that has been created based on the METABRIC dataset and includes feature contributions from Shapley values; according to the paper, the model achieved a test accuracy of 98.9% on its own test set [3].

Experiments in other diseases show the relevance of a detailed experimental design protocol, yet their criteria depend on the disease and the dataset. Experiments on diabetic retinopathy employ image classification through deep CNNs and visual local explanations [4]. Experiments in liver cirrhosis include parameter tuning by cross-validation, XAI, and clinical features [5]; in autism diagnosis, multiple models on age-related datasets are used together with SHAP [6]; experiments on heart diseases involve stacking, resampling, cross-validation, and SHAP [7]; experiments on chronic kidney disease involve SHAP to reduce the number of clinical features [8], while in experiments on randomized learning models, different computational methods such as LIME and SHAP are employed [9]. Cancer stage classification based on rare somatic mutations is another example [10]. However, these experiments cannot be used to validate the WDBC experiment.

III. PROPOSED METHODOLOGY

This problem is modeled as a binary classification problem with post-hoc explanation added to the evaluation criteria. The process is outlined in Fig. 1. First, the data are audited for missing values, duplicate records, target encoding, feature types, and class imbalance. Then, the data are split between development and held-out test sets in a stratified manner. All steps of the pre-processing pipeline that involve estimation of model parameters are done using the training set only. Model comparison is done by means of five-fold stratified cross-validation on the training set. The chosen model is tested on the held-out test set.

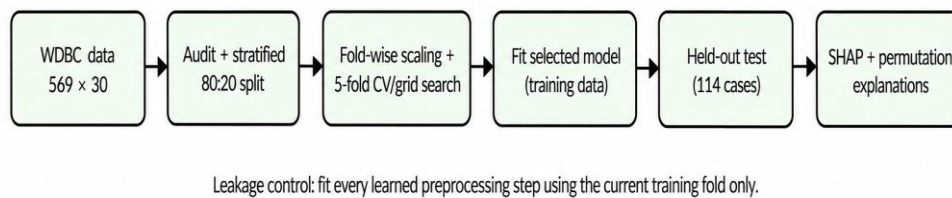


Fig. 1. Leakage-conscious workflow for the WDBC case study. The held-out test partition is kept outside model selection.

A. Dataset and Target Definition

The data used in the experiment is the Wisconsin Diagnostic Breast Cancer Data from the scikit-learn library. Dataset Characteristics: Multivariate, Text, Classification; Number of Instances: 569; Number of Attributes: 30 real-valued features; Attribute Information: Ten real-valued features are computed for each cell nucleus: radius (mean of distances from the center), texture (standard deviation of gray-scale values), perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension; Diagnosis is Binary; Malignant: Positive Class; Benign: Negative Class [13].

TABLE I. WDBC CASE-STUDY DATA PROFILE

Item	Reported Value	Interpretive note
Observations	569	Public benchmark dataset used by the scikit-learn implementation.
Input features	30 numeric variables	10 morphology measures, each summarized by mean, standard error, and worst.
Target	Benign/malignant	Malignant is defined as the positive class in this paper.
Train/test design	80:20 stratified; seed 42	455 training and 114 held-out test cases.
Models	Logistic regression, SVM, KNN, random forest	Scaling is stated for LR/SVM/KNN; RF uses unscaled inputs.

For an observation $x=(x_1,\dots,x_d)$, the model estimates $p(y=1|x)$, with $y=1$ denoting malignancy. A threshold τ maps probability to a class:

$$\hat{y}=1 \text{ if } p(y=1|x)\geq\tau \text{ and } \hat{y}=0 \text{ otherwise.}$$

The reported threshold-dependent classification metrics use $\tau=0.5$. That default is a reproducibility convention, not a clinically optimized operating threshold.

B. Split, Preprocessing, and Leakage Control

The data is split into training and test datasets in an 80/20 proportion with stratification and random seed value 42, producing 455 observations for training and 114 observations for testing. Scaling is done in training-fold pipelines for LR, SVM, and KNN models, while the random forest algorithm uses unscaled numeric features. Scaling is done for the training fold using the fitting observations from that fold and then used on the validation fold of that training fold only. Standardization is

$$z_j = (x_j - \mu_j) / \sigma_j,$$

where μ_j and σ_j are estimated from the current training fold. As there are no missing values in the chosen WDBC dataset, there is no need for imputation in this analysis. If an imputer, a feature selector, or a resampling method is used in future studies, it should be fit or applied after splitting into training and test sets, but not before.

C. Candidate Classifiers

The logistic regression model gives a linear probabilistic baseline; SVM is a margin classifier and also gives a non-linear classifier when using an RBF kernel; KNN is a similarity classifier based on standardized features; and random forest combines decision trees. These models have different inductive biases yet are still practical with small or well-tuned parameters. For logistic regression,

$$\log[p/(1-p)] = \beta_0 + \sum_i \beta_i x_i.$$

The RBF kernel is

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2).$$

KNN classification is based on assigning a class label by means of k nearest training observations. Random Forest involves combining the predictions from multiple trained decision trees.

D. Cross-Validation and Model Selection

We perform five-fold stratified cross-validation on the training subset. Each configuration is tested in five folds, maintaining approximately equal class distributions for each ROC-AUC metric, which is explicitly specified as the tuning target. The chosen configuration is retrained on the whole training subset and tested once on the separate test subset.

$$CV = (1/K) \sum_k M_k, \text{ with } K=5.$$

The proposed solution keeps the test dataset away from any information related to the model selection process, provided that each step of transformations and hyperparameter optimization is carried out on the training dataset only.

E. Evaluation Metrics

Accuracy, precision, recall (sensitivity), F1-score, and ROC-AUC are reported.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN),$$

$$\text{Precision} = TP / (TP + FP), \text{ Recall} = TP / (TP + FN),$$

$$F1 = 2(\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}).$$

The ROC-AUC measures ranking regardless of the threshold, while accuracy, precision, recall, and F1 at $\tau=0.5$ are threshold-dependent. In addition to this, we need to consider specificity, confidence intervals, calibration, and specific operating characteristics depending on thresholds to expand our evaluation criteria list

F. Explainability Layer

SHAP expresses a model output as a baseline plus feature attributions:

$$f(x) = E[f(X)] + \sum_j \phi_j$$

The sign of ϕ_j gives us information about which direction the corresponding feature pushes the explained model output. The mean absolute SHAP value provides a summary of the attribution magnitude. The general formulation of SHAP is provided by Lundberg and Lee [11].

Permutation importance for score

$$S_j = S(\text{model}, X, y) - \text{mean}_r [S(\text{model}, \text{permute}_r(X_j), y)].$$

Positive values indicate score loss after shuffling feature j.

Permutation importance represents the drop in a selected score for assessing the model's performance when a feature is permuted. This is a model-dependent, global measure of reliance and does not explain one particular prediction. The correlation between features leads to shared and masked importance, which means that the ranking depends on the scoring function used, as well as the data and permutation.

IV. EXPERIMENTAL RESULTS

Of the 114 instances from the holdout test dataset, the SVM had the best accuracy (97.37%) and F1 score (96.30%), and logistic regression had the best ROC-AUC (99.60%). The findings are presented in Table I below. The SVM model generated the best accuracy and F1 score of 97.37% and 96.30%, respectively, while the logistic regression generated the best ROC-AUC of 99.60%. This explains the reason why a single measure should not be used to select models.

TABLE II. REPORTED HELD-OUT PERFORMANCE AND DERIVED SVM COUNTS

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
SVM	97.37%	100.00	92.86	96.30%	0.995*
LR	96.49	97.50	92.86	95.12	99.60%
KNN	94.74	97.37	88.10	92.50	0.982*
RF	96.49	100.00	90.48	95.00	0.995*

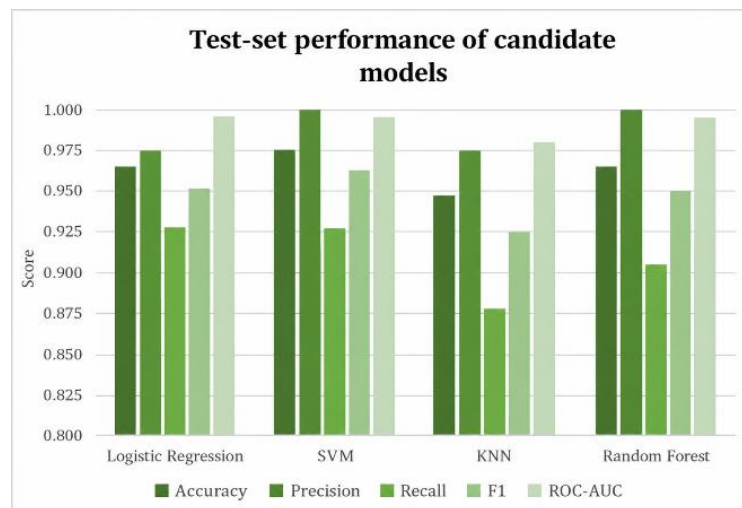


Fig. 2. Comparative performance of four candidate classifiers on the held-out test set.

Fig. 3 shows that the ROC curves for all four methods indicate that all four methods provide strong separation of classes on this test dataset. Logistic regression gives us the best AUC value, while SVM gives the best accuracy rate at the default threshold value. This may be attributed to the advantage of feature scaling and the use of a nonlinear SVM decision boundary.

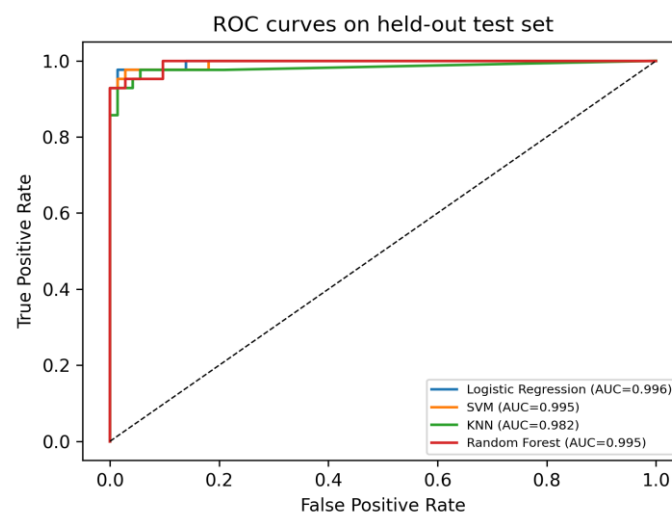


Fig. 3. ROC curves for the four candidate classifiers on the held-out test set.

A. Error Analysis

The observed SVM confusion matrix has zero false positives and three false negatives. With the given 42 malignant cases resulting from the test dataset, this implies 39 true positives and 72 true negatives. Accuracy is computed as $111/114=97.37\%$; precision is $39/39=100\%$; recall is $39/42=92.86\%$; and F1 score is approximately 96.30%. The mathematical computations are consistent with the given SVM accuracy and F1 score, given the stated class numbers and confusion matrix conventions. The confusion matrix for the chosen model is illustrated in Fig. 4 below.

TABLE III. SVM CONFUSION MATRIX

Actual/Predicted	Benign	Malignant
Benign	72 (TN)	0 (FP)
Malignant	3 (FN)	39 (TP)

Derived metrics, with malignant as positive: sensitivity = $39/(39+3) = 92.86\%$; specificity = $72/(72+0) = 100\%$; precision = $39/(39+0) = 100\%$; accuracy = $(72+39)/114 = 97.37\%$; F1 = $2 \times (1.000 \times 0.9286) / (1.000 + 0.9286) \approx 96.30\%$. These calculations are conditional on the figure's class orientation and the stated 42 malignant test cases.

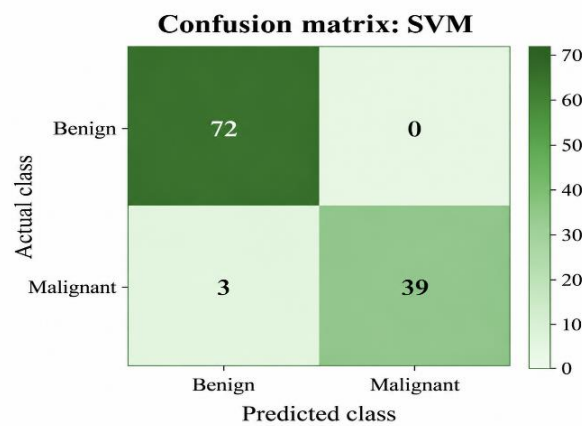


Fig. 4. Confusion matrix of the highest-accuracy SVM classifier.

B. Global Explanation

Permutation importance and SHAP were applied to determine whether the classifier utilized meaningful morphology variables. Permutation importances in the presented case study had the greatest effect on worst-symmetry, worst-texture, worst-concavity, worst-concave-points, mean-concave-points, mean-concavity, and worst-radius. Ranking of the variables does not show which variable causes cancer. It rather shows the process in which the model operates using this particular dataset. Correlation of variables might influence the ranking of importance.

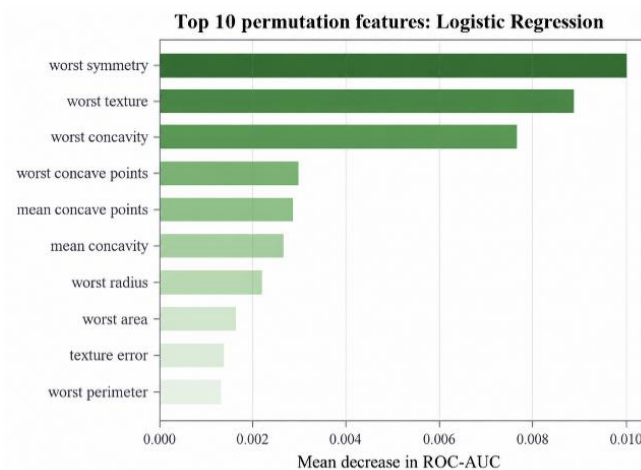


Fig. 5. Top permutation-importance features for the highest-AUC logistic-regression model.

SHAP analysis offers another perspective on global importance. Figure 6 shows the ranking of features according to the mean absolute SHAP value. The largest mean absolute impact among features in this study belongs to worst texture, which is followed by radius error, worst symmetry, mean concave points, and worst radius. There is some degree of overlap between the two sets of explanations, but not full agreement, which is understandable considering that they evaluate different metrics. In case of substantial differences between rankings, correlation among features, model sensitivity, and stability should be checked.

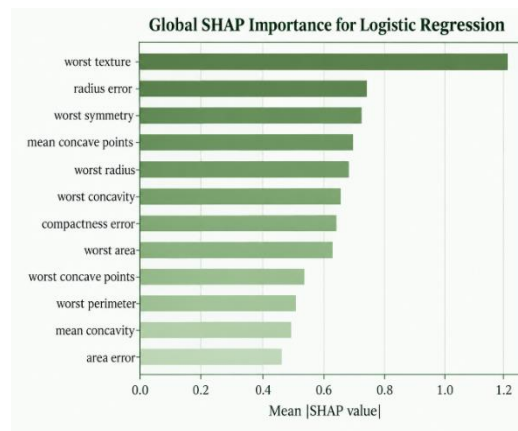


Fig. 6. Global SHAP importance for the logistic-regression model

C. Local Explanation and Clinical Plausibility

The local explanation is derived for one specific patient and not for the entire dataset. For instance, a patient-specific explanation will determine which measured features increase or decrease the chances that a particular patient has cancer compared to the baseline of the machine learning algorithm. This is much closer to how a doctor will question the decision-making process of the algorithm – why has this particular patient been assigned a high-risk value?

Medically plausible explanations constitute only one of many steps in the validation procedure; they do not ensure that predictions are medically accurate. Predictive models may identify associations between variables that are useful statistically but not medically correct. For example, acquisition-site IDs, preprocessing mistakes, and redundant variables can act as a shortcut. Therefore, the validation process of explanations should encompass a thorough investigation performed by domain experts, subgroup analysis, and ideally external validation by means of applying the dataset collected according to a different procedure. Specifically, the survey made by Bai et al. highlights the problems with bias in XAI systems [2].

V. DISCUSSION

This shows that building an interpretable pipeline for making diagnoses using classical machine learning is feasible. The highest accuracy and F1 score in the WDBC case study belong to the SVM model; the logistic regression model showed the best ROC-AUC value. Which model would be preferable depends on the purpose of applying it. It is useful to have a simple baseline for comparison. In cases where a linear model works as well as a complex one, evaluation might be cheaper.

The findings are in agreement with previous studies mentioned in this paper. According to Khater et al., classical machine learning in combination with model-agnostic explanations can reveal important features related to breast tumors [1]. GraphX-Net proves that Shapley values can be incorporated into more complicated relational models [3], whereas the liver cirrhosis and heart disease studies prove the need for cross-validation and explicit XAI layers in medical prediction [5], [7]. This study takes this approach, but employs classic classification models due to the small size of the dataset.

The second factor to consider is computational efficiency. Although deep learning may be the right choice in cases where the input is high-dimensional images and genomics data, classical models are better suited in situations where the input is low-dimensional and structured. Muhammad et al. observed the computational difficulty and obscurity in using deep models and experimented with randomized models [9].

Explanation techniques should be considered independent of the explanation of the model. Explanation does not always mean that the model is correct or medically approved. SHAP values evaluate the model but do not establish causality. Furthermore, the permutation feature importance can be biased in the case of the presence of correlated predictors. Reporting predictive performance, stability of explanation, subgroup analysis, and limitations of the study in the dissertation is very important. The statement is especially true in the case of applying the models in medicine, where the explanation can lead to false assurance.

VI. LIMITATIONS AND FUTURE WORK

The experiment relies solely on a single publicly available tabular dataset, limiting its generalizability. While the WDBC dataset is good for reproducibility purposes, it cannot be used as a replacement for clinical validation. The use of a morphologically engineered dataset rather than the actual images makes it impossible to draw any conclusion about an AI solution for images.

Correlation among features makes explanations difficult to understand because if two features represent similar information, then attributes may be spread across both features. Future research needs to study SHAP, permutation importance, partial dependence, and accumulated local effects, and measure explanation stability in bootstrapped or cross-validated datasets. Thirdly, the practicality of explanations depends on the choice of the threshold, calibration, decision curve analysis, and external validation of the model. Lastly, the study of fairness includes the study of performance and explanation behavior on subgroups of patients if the demographic features are present.

It is natural to consider an extension of the above framework for the evaluation of other diseases as well. The literature provided by us includes studies of diabetic retinopathy, liver cirrhosis, autism, heart disease, and chronic kidney disease [4]-[8]. An analysis of multiple diseases would allow us to understand whether the interpretation method is still stable when the input consists of structured data, images, and high-dimensional molecular information. One can also consider comparing inherently interpretable models with black-box models and assessing whether better prediction justifies more interpretation overhead.

VII. IMPLEMENTATION, REPRODUCIBILITY, AND TRUSTWORTHINESS

The design of the code is purposefully kept simple. The experimental setup uses Python along with pandas for data management, NumPy for numeric computation, scikit-learn for preprocessing and classification, Matplotlib for visualization, and SHAP for additive feature attribution. These commonly used libraries make sure that the workflow does not require any GPU or large deep learning framework to execute. This code design can be easily used in a clinical setting by modifying only the data loading part and keeping the validation and explainability parts the same.

To enable reproducibility, the experimental setup will utilize a fixed random seed and will ensure that the model selection step is independent of the test evaluation process. The fixed random seed value will be set to 42 for train/test split, cross-validation, and permutations. Hyperparameter tuning will be conducted using only the training dataset, and the test dataset will not be used in selecting a particular model or explaining the configuration. The test metrics obtained will thus reflect a unique holdout estimate following a chosen protocol.

The difference between predictive validity and explanation validity is significant. While predictive validity determines whether the model applies to future predictions, explanation validity determines how well the explanation procedure describes the model and how stable the explanation is to be interpretable. The important measure of stability for the feature ranking procedure is the following one:

$$S = (2 / R(R-1)) \sum (r < s) \rho(\text{rank}_r, \text{rank}_s),$$

Where R is the number of repeated runs and ρ is the Spearman rank correlation between two feature-importance rankings. A high S indicates that influential features remain similarly ordered under resampling or repeated cross-validation.

Accuracy and ROC-AUC cannot determine the calibration of the predicted probability. Diagnostic models can classify the patients correctly, but at the same time yield predicted probabilities that are always overestimated or underestimated. Thus, future studies should include calibration plots, Brier score, and sensitivity and specificity at a certain cut-off value. In case the model is designed to be used as a screening test, then sensitivity will take priority; whereas if the purpose is to assist in confirming diagnosis, then specificity and positive predictive value will become more significant.

The analysis of the bias should also become a part of the workflow. If demographic, institutional, scanning, or acquisition information is available, the performance of each subgroup must be reported independently. The SHAP ranking of the entire set might mask the performance disparity between subgroups. Moreover, high overall performance does not imply good performance of the model in each subgroup. The survey on the XAI of breast cancer models notes that transparency of the model does not eliminate dataset bias per se [2].

The proposed solution is outlined as a layered system approach, where the data layer manages data quality and leaks, the learning layer selects the model and makes the prediction, the evaluation layer measures discrimination and errors, the XAI layer provides explanations of the prediction, and the clinical review layer verifies plausibility, stability, and limitations of the result. Separating these layers simplifies the process analysis and prevents confusion between an obvious explanation and clinical validation.

The suggested sequence of events for completing the dissertation is to first develop baseline measures; then conduct cross-validation, calibration, explanation stability, subgroup analysis, and finally include an external validation data set. This sequence provides a focus on properly evaluating the model prior to using more complex XAI techniques. Additionally, it will allow for the development of measurable milestones.

VIII. CONCLUSION

The study aimed to prove that a diagnostic model could be evaluated in terms of its predictive power and explanations simultaneously. Without using any techniques other than classical machine learning and relying on Python libraries, the process is possible even on a regular computer without a GPU. For

the breast cancer dataset, all four models performed well in separating benign and malignant instances. Depending on the metrics, the best-performing classifier could vary. SVM was the winner on accuracy and F1-score, whereas logistic regression was the best on ROC-AUC, which would not be clear by the accuracy measure alone.

From the error analysis perspective, the reason for this becomes obvious. While the SVM did not make any false positives, it missed three malignant cases out of 42, and missing cancer in screening is a far more costly error.

The results of the interpretations were encouraging. Both SHAP and permutation importance indicated that the most significant features for predicting whether or not a patient has cancer include the shape and texture of the tumor; specifically, "worst texture," "worst concavity," "concave points," and "radius." All of these features are ones that a physician would believe to be important. While there was some inconsistency among the interpretations, this is to be expected since each measure is based on a different concept. For example, SHAP indicates how much an individual feature contributes to changing the prediction, whereas permutation importance indicates how much the performance of the model decreases after a specific feature is randomly shuffled.

Therefore, combining both measures provides a more comprehensive view of the model's performance than either measure alone. This interpretation explains how the model utilizes the data set provided. It does not demonstrate causation of any sort.

In conclusion, building a leak-free pipeline with explanation integrated into the process is feasible for tabular data, especially in medicine. It also proved that a simpler model can easily outperform a more sophisticated one. Given the limited amount of data from a single source, these results should be considered a solid proof of concept rather than evidence of clinical feasibility.

IX. REFERENCES

- [1] T. Khater, A. Hussain, R. Bendar-daf, I. M. Talaat, H. Tawfik, S. Ansari, and S. Mahmoud, "An Explainable Artificial Intelligence Model for the Classification of Breast Cancer," *IEEE Access*, vol. 13, pp. 5618–5630, 2025, doi: 10.1109/ACCESS.2023.3308446.
- [2] S. Bai, S. Nasir, R. A. Khan, A. Meyer, and H. Konik, "Breast Cancer Diagnosis With Explainable Artificial Intelligence (XAI): Uncovering Strengths and Biases," *IEEE Access*, vol. 13, pp. 205197–205223, 2025, doi: 10.1109/ACCESS.2025.3639184.
- [3] A. Basaad, S. Basurra, E. Vakaj, M. Aleskandarany, and M. M. Abdelsamea, "GraphX-Net: A Graph Neural Network-Based Shapley Values for Predicting Breast Cancer Occurrence," *IEEE Access*, vol. 12, pp. 93993–94008, 2024, doi: 10.1109/ACCESS.2024.3424526.
- [4] T. Shahzad, M. Saleem, M. S. Farooq, S. Abbas, M. A. Khan, and K. Ouahada, "Developing a Transparent Diagnosis Model for Diabetic Retinopathy Using Explainable AI," *IEEE Access*, vol. 12, pp. 149700–149716, 2024, doi: 10.1109/ACCESS.2024.3475550.
- [5] K. R. Makkena and K. Natarajan, "Enhancing Liver Cirrhosis Diagnosis Using Machine Learning With Explainable AI and Cross-Validated Hyperparameter Tuning Techniques," *IEEE Access*, vol. 13, pp. 93323–93336, 2025, doi: 10.1109/ACCESS.2025.3568918.
- [6] O. Jyoti, H. B. Kibria, Z. T. Pear, M. Nahiduzzaman, M. F. Ahamed, K. R. Islam, J. Kumar, and M. E. H. Chowdhury, "A Clinically Interpretable Approach for Early Detection of Autism Using Machine Learning With Explainable AI," *IEEE Access*, vol. 13, pp. 121512–121531, 2025, doi: 10.1109/ACCESS.2025.3586314.
- [7] M. Naz, A. Khalid, A. Hameed, R. Taj, W. Mumtaz, F. A. Alotaibi, and M. M. Alnfai, "Meta-Ensemble Learning for Heart Disease Prediction: A Stacking-Based Approach With Explainable AI," *IEEE Access*, vol. 13, pp. 137271–137290, 2025, doi: 10.1109/ACCESS.2025.3588683.
- [8] M. Rashed-Al-Mahfuz, A. Haque, A. K. M. Azad, S. A. Alyami, J. M. W. Quinn, and M. A. Moni, "Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 4900511, 2021, doi: 10.1109/JTEHM.2021.3073629.
- [9] D. Muhammad, I. Ahmed, M. O. Ahmad, and M. Bendeche, "Randomized Explainable Machine Learning Models for Efficient Medical Diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 9, pp. 6474–6486, 2025, doi: 10.1109/JBHI.2024.3491593.
- [10] H. Younis, H. A. Rehman, and R. Minghim, "Somatic Mutation-Driven Cancer Stage Classification via Autoencoder Clustering and Explainable Deep Learning," *IEEE Access*, vol. 14, pp. 45213–45226, 2026, doi: 10.1109/ACCESS.2026.3676284.
- [11] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.
- [12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [13] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.