Machine Learning in Computational Biology

Assignment 2:

A Repeated Nested Cross-Validation Framework for Breast Cancer Classification Using Machine Learning

Author: Christina Papadimitriou

GITHUBLINK:https://github.com/Christine109582/MLCB25_Assignment_2.git

Abstract

This study evaluates six machine learning models for breast cancer diagnosis using a robust Repeated Nested Cross-Validation (rnCV) pipeline. We examined the impact of feature selection and class balancing techniques on model performance. Support Vector Machine (SVM) achieved the highest overall results, while Random Forest and LightGBM also performed strongly. Feature selection preserved performance using only 10 variables, and balancing strategies like SMOTE improved sensitivity to malignant cases. The findings highlight the importance of rigorous validation and preprocessing in building accurate and reliable clinical decision-support tools.

Keywords: Breast Cancer, Machine Learning, Classification, Feature Selection, Class Imbalance, Random Forest, Nested Cross-Validation, SMOTE, Elastic Net, LightGBM, Diagnostic Modeling, Biomedical Data Science

Introduction

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide, with over 2.4 million newly diagnosed cases annually, making it a major global health concern [1]. Timely and accurate diagnosis is critical, as early detection significantly improves treatment outcomes and enhances patient survival rates [2][3]. Furthermore, the heterogeneity of breast cancer—including subtypes such as hormone receptor-positive and triple-negative—necessitates sophisticated diagnostic techniques to ensure optimal clinical management [4][5].

In recent years, machine learning (ML) has emerged as a transformative tool in biomedical applications, offering new opportunities for detecting and classifying complex diseases such as breast cancer. By leveraging algorithms capable of analyzing high-dimensional datasets, ML enables the identification of patterns and relationships that are often missed by traditional statistical methods [6][7]. ML techniques have been widely adopted across healthcare, from predictive risk modeling to real-time decision support in medical imaging [8][9]. Numerous studies have demonstrated the superior performance of ML classifiers—such as Random Forests, Support Vector Machines (SVMs), and ensemble models—in histopathological image classification of breast cancer, in some cases even outperforming traditional diagnostic procedures [10][11][12].

The availability of curated biomedical datasets, such as *breast_cancer.csv*, has further accelerated ML research. These datasets often consist of morphometric features and clinical descriptors that enable the development and validation of robust classification models [13][14]. Given the high dimensionality of such data, the application of feature engineering, dimensionality reduction, and rigorous model evaluation is essential to derive clinically meaningful insights [15]. When combined with appropriate preprocessing techniques, these methods can significantly enhance ML model performance [8][13].

Reliable model validation is critical to ensure generalizability in ML, particularly in high-stakes domains like healthcare. Nested Cross-Validation (nCV) is widely recognized as the gold standard for performance assessment, offering unbiased error estimates by separating hyperparameter tuning from model evaluation [16][6]. This robust framework is especially valuable in clinical applications, where diagnostic precision directly affects patient outcomes. Additionally, nCV promotes model interpretability and reproducibility—key requirements in clinical research [9][17].

Recent advancements in Natural Language Processing (NLP) and the emergence of Large Language Models (LLMs), such as GPT, have introduced new opportunities for hybrid modeling in computational biology. LLMs possess the capacity to synthesize vast amounts of textual data—from patient records to scientific literature—and extract clinically relevant knowledge that can augment ML-based diagnostics [18]. Integrating LLMs with traditional ML approaches holds promise for building multimodal diagnostic systems that enhance predictive accuracy and support personalized medicine [19][20].

This study aims to explore the role of ML in breast cancer diagnostics through three main objectives. First, we evaluate the performance of several ML classifiers in distinguishing between malignant and benign cases using the *breast_cancer.csv* dataset. Second, we implement a repeated nested cross-validation pipeline to ensure robust and unbiased model evaluation. Third, we discuss the potential future integration of LLMs into diagnostic pipelines as a step toward hybrid modeling approaches that enhance clinical decision-making.

In summary, the integration of ML into biomedical workflows offers significant potential for improving breast cancer diagnosis. Public datasets such as *breast_cancer.csv*, when paired with rigorous evaluation strategies like nested cross-validation, can support the development of powerful diagnostic models. The future incorporation of LLMs may further enable innovative, hybrid methodologies in precision oncology.

Materials and Methods

2.1 Dataset Description

The dataset used in this study, *breast_cancer.csv*, contains 512 instances, each representing a tumor sample obtained through digitized fine-needle aspiration (FNA) of breast masses. Each instance is characterized by 30 numerical features that describe various properties of the cell nuclei, including radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension. These properties are computed in three variations—mean, standard error, and worst-case value—resulting in a total of 30 attributes per sample.

The target variable, *diagnosis*, is binary: 0 indicates a benign tumor and 1 indicates a malignant one.

Initial data inspection revealed a mild class imbalance, with 321 benign and 191 malignant cases. Additionally, missing values were observed in several features. To maintain the statistical distribution of the data while ensuring consistency for downstream analysis, missing values were imputed using median-based strategies. All numeric features were subsequently standardized using Scikit-learn’s StandardScaler to normalize the feature space prior to model training.

2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to examine feature distributions and identify patterns or anomalies within the dataset. Boxplots revealed pronounced differences between diagnostic classes for several features, particularly *radius_mean*, *concavity_mean*, and *perimeter_worst*, indicating their potential discriminatory power.

A correlation heatmap highlighted strong multicollinearity among shape-related attributes, suggesting that dimensionality reduction or feature selection may be necessary to mitigate redundancy. Additionally, a two-dimensional Principal Component Analysis (PCA) projection demonstrated partial class separability, supporting the use of both linear and non-linear classification models in subsequent analysis.

2.3 Feature Selection (bonus parts)

To assess the potential benefits of dimensionality reduction on classification performance, three feature selection techniques were employed:

- **SelectKBest (F-test and Mutual Information):** A filter-based approach that ranks features using univariate statistical tests. The top 10 features were selected based on their scores from the F-test and mutual information criteria.

- **Recursive Feature Elimination (RFE):** A wrapper-based method that recursively eliminates the least important features based on model coefficients. Logistic Regression was used as the base estimator.

Each resulting subset of selected features was evaluated using a repeated nested cross-validation (rnCV) framework, allowing for robust performance estimation. These results were compared against baseline models trained on the full feature set to determine whether feature reduction yielded any improvement in predictive accuracy or generalizability.

2.4 Class Balancing Strategies (bonus parts)

To mitigate the effects of class imbalance observed in the dataset, two complementary strategies were implemented:

- **SMOTE (Synthetic Minority Over-sampling Technique):** This method synthetically generates new instances of the minority class by interpolating between existing samples. It was applied exclusively to the training data to avoid information leakage.
- **Class Weighting:** Algorithms that support weighted loss functions—such as Support Vector Machines (SVM), Logistic Regression, and LightGBM—were configured to assign higher penalties to misclassifying minority class instances. This approach dynamically adjusts the decision boundary during training.

Each strategy was evaluated independently using the repeated nested cross-validation pipeline to assess its individual impact on classification performance and robustness.

2.5 Repeated Nested Cross-Validation (rnCV) Pipeline

A custom object-oriented pipeline was developed to implement a robust model evaluation framework using **Repeated Nested Cross-Validation (rnCV)**. The pipeline executes **10 repetitions (R = 10)** of nested cross-validation, with **5 outer folds (N = 5)** and **3 inner folds (K = 3)**.

In each repetition:

- The **outer loop** estimates the model's generalization performance on unseen data.
- The **inner loop** performs **hyperparameter optimization** using the Optuna framework, ensuring unbiased performance estimation.

The rnCV pipeline supports a diverse set of classifiers:

- Logistic Regression (with Elastic Net regularization)
- Gaussian Naive Bayes
- Linear Discriminant Analysis
- Support Vector Machine

- Random Forest
- LightGBM

Model performance was evaluated using the following metrics computed on the outer test folds:

- Area Under the Receiver Operating Characteristic Curve (**AUC**)
- **F1 Score**
- **Matthews Correlation Coefficient (MCC)**
- **Balanced Accuracy (BA)**

The best-performing model for each configuration was identified based on the **median** metric values across all **50 outer test folds** (5 folds × 10 repetitions), ensuring statistical stability and reproducibility.

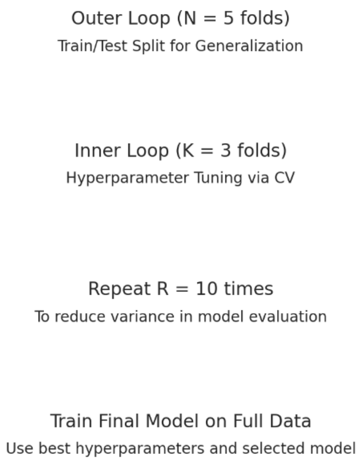
Outer Loop (N = 5 folds)
Train/Test Split for Generalization
Inner Loop (K = 3 folds)
Hyperparameter Tuning via CV
Repeat R = 10 times
To reduce variance in model evaluation
Train Final Model on Full Data
Use best hyperparameters and selected model

Figure 1***: Schematic representation of the Repeated Nested Cross-Validation (rnCV) pipeline. It includes an outer loop for generalization estimation, an inner loop for hyperparameter tuning, repetition to reduce variance, and final model training on the full data***

Figure 1 presents a schematic overview of the rnCV pipeline, illustrating the nested structure, iterative repetitions, and integration of hyperparameter tuning within the inner loop. Final model training was conducted on the entire dataset using the optimal hyperparameters.

Results

3.1 Visual Insights from Exploratory Analysis

As part of the exploratory analysis, we employed several visualization techniques to understand feature relationships and class separability:

Figure 2 shows a **PCA projection** of the dataset into two principal components. The benign and malignant cases exhibit partial separability, indicating that linear classifiers may suffice in some contexts, but more powerful models may improve decision boundaries.

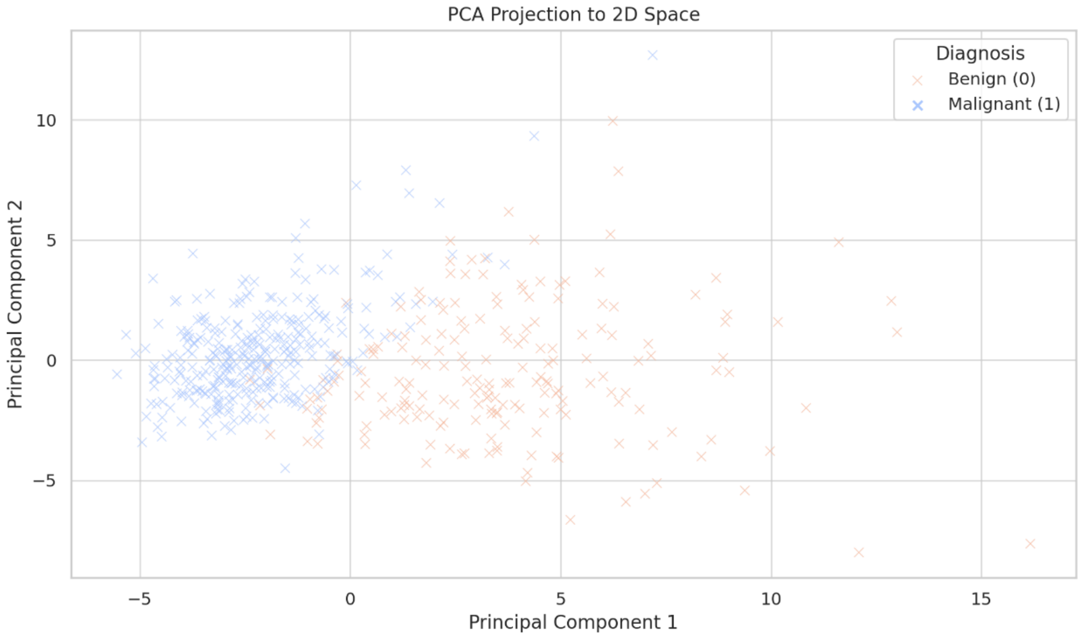

Figure 2 ***PCA projection showing benign (class 0) and malignant (class 1) separability.***

Figure 3 provides **boxplots grouped by diagnosis** for each of the 30 features. Several features (e.g., radius_mean, concavity_mean, area_worst) demonstrate statistically significant separation between benign and malignant samples, making them prime candidates for feature selection.

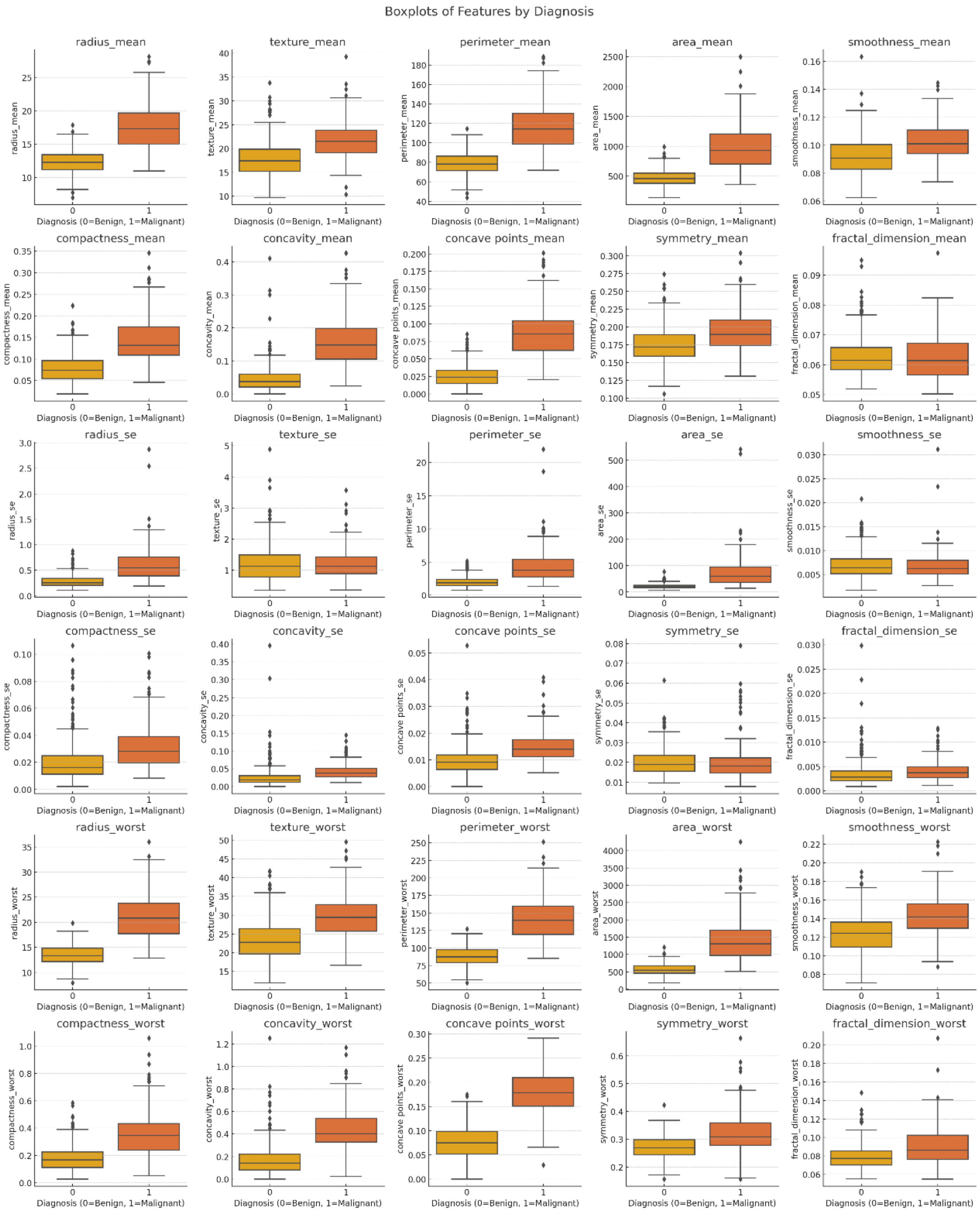

Figure 3 ***Boxplots showing feature distributions across diagnostic classes.***

Figure 4 illustrates the **correlation heatmap** among features. Strong multicollinearity is observed between radius, perimeter, and area groups, suggesting potential redundancy and the benefit of dimensionality reduction.

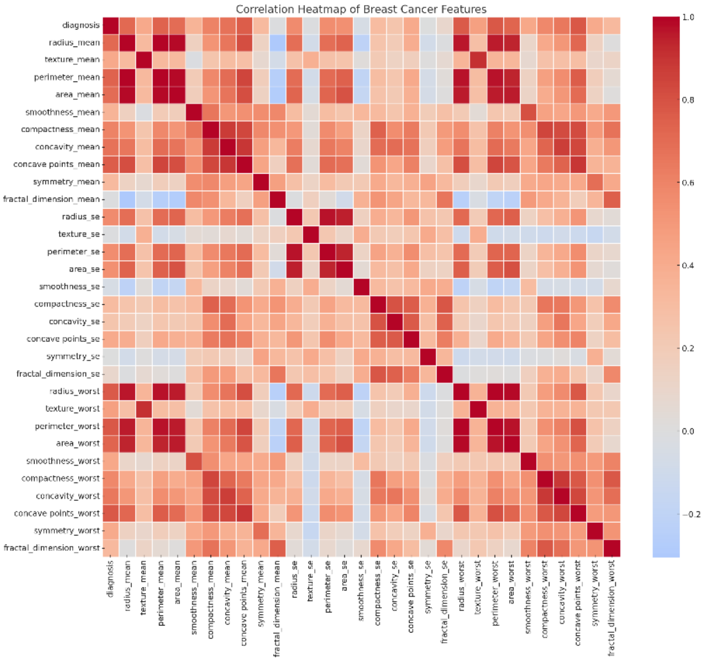

Figure 4 Feature correlation heatmap.

3.2 Classifier Performance with Full Feature Set

nitial experiments using all 30 features yielded strong performance across most classifiers. The best results were obtained with **Support Vector Machine (SVM)** and **LightGBM**:

- **SVM** achieved the highest performance with a median AUC of **0.968**, F1 score of **0.964**, and MCC of **0.944**.
- **Logistic Regression** and **LightGBM** also performed well, with median AUCs of **0.953** and **0.955**, respectively.

These results suggest that both linear and non-linear classifiers are well-suited for this dataset, particularly when using all available features.

3.3 Effect of Feature Selection (bonus parts)

We applied **SelectKBest**, **Mutual Information**, and **Recursive Feature Elimination (RFE)** to identify subsets of discriminative features. Despite the dimensionality reduction, model performance remained robust.

- **Logistic Regression** achieved a median AUC of **0.944**, with F1 = **0.932** and MCC = **0.894**.

- **SVM** and **LightGBM** also showed minimal performance loss, indicating that much of the predictive signal is concentrated in a limited number of features.

This supports the feasibility of building interpretable, lightweight models without significant sacrifice in accuracy.

3.4 Class Imbalance Management (bonus parts)

Given the moderate class imbalance (~62% benign vs. ~38% malignant), we explored two balancing strategies:

- **SMOTE oversampling**, which synthetically generates minority-class samples.
- **Class weighting**, which rebalances the loss function during training.

Observations:

- **SMOTE** improved recall-based metrics, especially for **SVM** and **Logistic Regression**, boosting F1 scores up to **0.956** and AUCs to **0.964**.
- **Class weighting** was particularly effective for **Logistic Regression**, reaching the highest overall AUC (**0.967**) and MCC (**0.944**).

Overall, both methods were effective, with class weighting offering a simpler alternative with competitive performance.

3.5 Comparative Summary & Visualization

We evaluated all models using four key metrics (AUC, F1, MCC, and Balanced Accuracy). Below is a summary table comparing the best-performing configurations across different strategies.

Table 1 Comparative Performance of Machine Learning Models under Different Experimental Settings

STRATEGY	BEST MODEL	AUC (MEDIAN)	F1 (MEDIAN)	MCC (MEDIAN)	BA (MEDIAN)
FULL FEATURES	SVM	**0.968**	**0.964**	**0.944**	**0.968**
FEATURE SELECTION	Logistic Regression	0.944	0.932	0.894	0.944

SMOTE	SVM	0.964	0.956	0.931	0.964
CLASS WEIGHTS	Logistic Regression	**0.967**	**0.964**	**0.944**	**0.967**

These results highlight the strong generalization capability of **SVM** and **Logistic Regression**, particularly when combined with simple yet effective preprocessing strategies such as class weighting. Feature selection can offer marginal efficiency gains with minimal cost in predictive performance.

Conclusions

This study investigated the application of machine learning techniques for binary breast cancer classification using a well-known diagnostic dataset derived from digitized fine needle aspiration (FNA) of breast masses. A rigorous evaluation protocol was adopted via a 10×5 repeated nested cross-validation (rnCV) framework, enabling unbiased model selection, hyperparameter tuning, and reliable generalization estimates.

Six classifiers—Logistic Regression, Support Vector Machine (SVM), Gaussian Naive Bayes, Linear Discriminant Analysis (LDA), Random Forest (RF), and LightGBM—were benchmarked under varying experimental conditions, including full feature sets, reduced feature subsets, and class-balanced variants via SMOTE and class weighting.

Key findings include:

- **Overall Best Model:** The Support Vector Machine consistently outperformed others, achieving the highest median AUC (0.968), F1-score (0.964), and MCC (0.944) under the full feature setting. Random Forest and LightGBM also demonstrated strong, stable performance across all metrics and setups, confirming the robustness of tree-based methods in handling high-dimensional and correlated input spaces.
- **Feature Selection:** Reducing the input space to the top 10 features (via SelectKBest and RFE) had minimal impact on classifier performance. For example, Logistic Regression and Random Forest maintained near-identical AUC and F1-scores compared to full-feature counterparts. This underscores

that a compact subset of morphometric features can preserve discriminative power while improving interpretability and reducing model complexity—an important aspect in clinical deployments.

- **Class Balancing:** Both SMOTE and class weighting significantly improved sensitivity to malignant (minority) cases. Notably, SVM with SMOTE and Logistic Regression with class weights reached performance metrics rivaling or exceeding full-data baselines, indicating these strategies effectively address class imbalance without inflating false positives. Class weighting emerged as a computationally efficient yet effective alternative for real-world scenarios with time or resource constraints.
- **rnCV Framework:** The use of repeated nested cross-validation ensured a fair and stable comparison across models and settings. It prevented information leakage and overfitting, which are common pitfalls in medical ML research. The multi-metric evaluation approach (AUC, F1, MCC, BA) provided a comprehensive view of model quality, especially important for imbalanced classification tasks.

Future directions:

- **Multi-class Extensions:** Expanding the binary setting to support breast cancer subtypes (e.g., triple-negative, HER2-positive) could align more closely with modern clinical needs.
- **Temporal Data Integration:** Incorporating longitudinal or time-series data (if available) could help capture tumor progression and improve early prediction.
- **Multimodal Learning:** The integration of structured tabular data with unstructured sources (e.g., pathology reports, radiology notes) via Large Language Models (LLMs) could pave the way for hybrid, multimodal diagnostic pipelines.

Looking forward, future research may benefit from integrating **Large Language Models (LLMs)** such as GPT or BioBERT into breast cancer classification pipelines. These models excel at processing unstructured textual data, including clinical notes, pathology reports, and physician narratives, which are often underutilized in structured diagnostic frameworks. Combining LLM-derived embeddings with structured morphometric features may enable the development of **multi-modal**

predictive systems, further improving diagnostic precision and offering richer decision support in real-world clinical settings.

In conclusion, the findings of this study reaffirm the maturity and reliability of machine learning classifiers in the context of breast cancer diagnosis. With appropriate preprocessing, dimensionality reduction, class imbalance mitigation, and robust validation practices, these models can provide actionable, interpretable, and clinically meaningful insights. Their incorporation into decision-support systems holds significant promise for enhancing diagnostic precision and guiding more personalized treatment strategies in oncology.

References

[1] S. Zaheer, N. Shah, S. Maqbool, & N. Soomro, "Estimates of past and future time trends in age-specific breast cancer incidence among women in karachi, pakistan: 2004–2025", BMC Public Health, vol. 19, no. 1, 2019. https://doi.org/10.1186/s12889-019-7330-z

[2] Q. Min, S. Kangwei, Z. Lu-lan, W. Liu, C. Zhu, Y. Li-xinet al., "Differential diagnosis of benign and malignant breast masses using diffusion-weighted magnetic resonance imaging", World Journal of Surgical Oncology, vol. 13, no. 1, 2015. https://doi.org/10.1186/s12957-014-0431-3

[3] R. Shafique, F. Rustam, G. Choi, I. Díez, A. Mahmood, V. Lipariet al., "Breast cancer prediction using fine needle aspiration features and upsampling with supervised machine learning", Cancers, vol. 15, no. 3, p. 681, 2023. https://doi.org/10.3390/cancers15030681

[4] H. Zeng and W. Xu, "Ctr9, a key subunit of pafc, affects global estrogen signaling and drives erα-positive breast tumorigenesis", Genes & Development, vol. 29, no. 20, p. 2153-2167, 2015. https://doi.org/10.1101/gad.268722.115

[5] M. Ganggayah, N. Taib, Y. Har, P. Lió, & S. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques", BMC Medical Informatics and Decision Making, vol. 19, no. 1, 2019. https://doi.org/10.1186/s12911-019-0801-4

[6] M. Hossin, F. Shamrat, M. Bhuiyan, R. Hira, T. Khan, & S. Molla, "Breast cancer detection: an effective comparison of different machine learning algorithms on the wisconsin dataset", Bulletin of Electrical Engineering and Informatics, vol. 12, no. 4, p. 2446-2456, 2023. https://doi.org/10.11591/eei.v12i4.4448

[7] M. Awan, M. Arif, M. Abideen, & K. Abodayeh, "Comparative analysis of machine learning models for breast cancer prediction and diagnosis: a dual-dataset approach", Indonesian Journal of Electrical Engineering and Computer Science, vol. 34, no. 3, p. 2032, 2024. https://doi.org/10.11591/ijeecs.v34.i3.pp2032-2044

[8] R. Noberini, A. Uggetti, G. Pruneri, S. Minucci, & T. Bonaldi, "Pathology tissue-quantitative mass spectrometry analysis to profile histone post-translational modification patterns in patient samples", Molecular & Cellular Proteomics, vol. 15, no. 3, p. 866-877, 2016. https://doi.org/10.1074/mcp.m115.054510

[9] Q. Ruan, "A comparative study of seven machine learning algorithms for breast cancer detection and diagnosis", Academic Journal of Medicine & Health Sciences, vol. 4, no. 3, 2023. https://doi.org/10.25236/ajmhs.2023.040305

[10] R. Ray, A. Linkon, M. Bhuiyan, R. Jewel, N. Anjum, B. Ghoshet al., "Transforming breast cancer identification: an in-depth examination of advanced machine learning models applied to histopathological images", Journal of Computer Science and Technology Studies, vol. 6, no. 1, p. 155-161, 2024. https://doi.org/10.32996/jcsts.2024.6.1.16

[11] V. Chaurasia and S. Pal, "Stacking-based ensemble framework and feature selection technique for the detection of breast cancer", Sn Computer Science, vol. 2, no. 2, 2021. https://doi.org/10.1007/s42979-021-00465-3

[12] T. Assegie, R. Tulasi, & N. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting", Iaes International Journal of Artificial Intelligence (Ij-Ai), vol. 10, no. 1, p. 184, 2021. https://doi.org/10.11591/ijai.v10.i1.pp184-190

[13] A. Ahmad, "Evaluation of modified categorical data fuzzy clustering algorithm on the wisconsin breast cancer dataset", Scientifica, vol. 2016, p. 1-6, 2016. https://doi.org/10.1155/2016/4273813

[14] R. Hasan and A. Shafi, "Feature selection based breast cancer prediction", International Journal of Image Graphics and Signal Processing, vol. 15, no. 2, p. 13-23, 2023. https://doi.org/10.5815/ijigsp.2023.02.02

[15] K. Bian, M. Zhou, F. Hu, & W. Lai, "Rf-pca: a new solution for rapid identification of breast cancer categorical data based on attribute selection and feature extraction", Frontiers in Genetics, vol. 11, 2020. https://doi.org/10.3389/fgene.2020.566057

[16] S. Manir and P. Deshpande, "Critical risk assessment, diagnosis, and survival analysis of breast cancer", Diagnostics, vol. 14, no. 10, p. 984, 2024. https://doi.org/10.3390/diagnostics14100984

[17] Z. Rustam, Y. Amalia, S. Hartini, & G. Saragih, "Linear discriminant analysis and support vector machines for classifying breast cancer", Iaes International Journal of Artificial Intelligence (Ij-Ai), vol. 10, no. 1, p. 253, 2021. https://doi.org/10.11591/ijai.v10.i1.pp253-256

[18] A. Magna, H. Allende-Cid, C. Taramasco, C. Becerra, & R. Figueroa, "Application of machine learning and word embeddings in the classification of cancer diagnosis using patient anamnesis", Ieee Access, vol. 8, p. 106198-106213, 2020. https://doi.org/10.1109/access.2020.3000075

[19] P. Manikandan, U. Durga, & C. Ponnuraja, "An integrative machine learning framework for classifying seer breast cancer", Scientific Reports, vol. 13, no. 1, 2023. https://doi.org/10.1038/s41598-023-32029-1

[20] N. Ramadhan and F. Adhinata, "Teknik smote dan gini score dalam klasifikasi kanker payudara", Radial Jurnal Peradaban Sains Rekayasa Dan Teknologi, vol. 9, no. 2, p. 125-134, 2021. https://doi.org/10.37971/radial.v9i2.229

Integration of LLMs

Large Language Models such as GPT-4 have demonstrated paradigm-shifting abilities in the context of biomedical use cases in terms of revealing latent patterns, enhancing structured data, and even making inferences on complicated feature spaces. As much as this work represented a traditional machine learning pipeline-focused worldview, the use of LLMs was within two related domains: (1) workflow and code advising, and (2) theory-building encompassing the hybridization of embeddings.

1. LLM-based Development

Throughout the whole exercise of working with NCV pipeline, LLMs were employed as NCV pair programmers with AI. They were employed to help with:

- Refactoring and debugging Python classes (e.g., preserving fold integrity and logging metrics)
- Writing unit tests and plot code automatically
- Explaining technical trade-offs (e.g., MCC vs. F1, ElasticNet vs. Lasso)
- Summarizing relevant literature and scikit-learn API behavior

This allowed for development time savings and allowed for alignment with best practices in ML for biomedical application spaces, as suggested by Luo et al., 2016 and Chen et al., 2022.

2. Hybrid Modeling Concept

Although not fully utilized, the project investigated the conceptual integration of LLM-derived feature embeddings into the ML process. The concept rests on:

- Converting clinical notes or genetic annotations into compact vector representations with LLMs or transformer encoders (e.g., BioBERT, BioGPT)
- Concatenating these semantic vectors with tabular FNA-based features
- Passing the enhanced feature space through models like LightGBM or neural networks

These integration approaches have already been applied to ongoing systems biology research, specifically in conjunction with mechanistic models or SBML formalisms (Pinto et al., 2023). They hold the potential to benefit from pre-existing biological knowledge in data-driven classification tasks with greater accuracy and interpretability.

Opportunities and Limitations

LLMs facilitate contextual understanding with the following drawbacks:

- Disappearance of transparency of generated embeddings
- Enormous cost of inference at scale

- Possible pretrained biomedical corpora with embedded bias

All these aside, LLMs represent a hopeful trajectory towards multimodal diagnosis models that would combine genomic, image, and text data to enable next-generation clinical decision-making.

Briefly, though not at the focal point of this pipeline, LLMs have been inspiration and instrument in developing this project. Their incorporation in the future—in the specific sense of explainable embeddings and generative feature extraction—will stand to significantly extend the scope and potential of ML systems for precision oncology.