Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Multimodal Learning for Breast Cancer Biomarker Prediction Using Whole Slide Histopathology Images

Vinita Shah[1], Miral Patel[2]

[1]Research Scholar, [2]Professor, Department of Computer Engineering, CVM University, Anand, India.

E-mail: [1]shahvinita89@gmail.com, [2]miral.patel@cvmu.edu.in

Abstract

Globally, breast cancer remains a significant health challenge that has a direct effect on women's cancer morbidity and mortality. The estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are important factors that help doctor to determine the best treatment for each woman. When using immunohistochemistry and genomic assays to look for markers, it is a relatively long and slow process that varies from individual to individual. The aim of this study is to develop a deep-learning framework to predict directly the ER, PR and HER2 status of H&E-stained histopathology images. The technique entails downsampling Level-1 slide images from the TCGA-BRCA cohort, followed by using a pre-trained ResNet50 architecture to extract histological features to enhance the accuracy of biomarker prediction. We train a multi-output classification model using XGBoost that adds a classifier chain. We use a mixture of clinical and genetic data as well as image features. This joint computational method shows promise in enhancing the accuracy of biomarker predictions and enabling doctors to customize breast cancer treatment for individual patients.

Keywords: Breast Cancer, Histopathology, Deep Learning, H&E Images, Biomarker Prediction, ER, PR, HER2.

1. Introduction

Breast cancer, which is probably one of the most common forms of cancer across the world, causes a lot of morbidity and mortality in women. About 1.7 million new breast cancer cases and almost 500,000 deaths are reported each year due to breast cancer [1]. Having an accurate diagnosis and a proper understanding of the disease at an early stage is important to improve the survival chances of the patient. Quick and accurate diagnostic procedures have emerged as an urgent need.

Biomarkers such as Estrogen Receptor (ER), Progesterone Receptor (PR), and Human Epidermal Growth Factor Receptor 2 (HER2) are crucial components in the diagnosis, prognostication, and treatment of breast cancer. By using the correct biomarkers, doctors can determine the treatments that an individual might need. They may receive hormonal therapy, targeted agents and/or chemotherapy. The treatment and recovery of patients depend on bed availability in hospitals.

DOI: https://doi.org/10.36548/jiip.2026.1.010
Received: 13.12.2025, received in revised form: 12.01.2026, accepted: 27.01.2026, published: 05.02.2026

Molecular subtypes of breast cancer Luminal A, Luminal B, HER2-enriched, triple-negative, and normal-like are defined by ER, PR, and HER2 expression status and exhibit distinct biological behavior, clinical outcomes, and treatment responses [5,6]. Luminal A tumors (ER+/PR+/HER2−) generally have a favorable prognosis and are treated primarily with endocrine therapy, whereas Luminal B tumors (ER+/PR±/HER2±) often require combined chemotherapy and hormone therapy due to higher proliferation rates. HER2-enriched tumors (ER−/PR−/HER2+) are aggressive and necessitate HER2-targeted therapies such as trastuzumab and pertuzumab [16,18]. Triple-negative breast cancers (ER−/PR−/HER2−) lack targeted treatment options and are typically managed with chemotherapy alone [3,5]. Understanding these subtypes enables clinicians to tailor treatment strategies and advance precision oncology in breast cancer care.

Traditionally, immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) on histological tissues are used to assess biomarker status. Nonetheless, these strategies have inter-observer variability, low reproducibility and very high cost [1,14]. The high price tag of RNA-based genomic assays limits their use in clinical practice, especially in less wealthy areas, even though they offer more prognostic, predictive, and referral value [11,17].

There is a demand for a diagnostic tool that is easier to use, reproducible and economical. Analyzing tissue samples with Hematoxylin and Eosin (H & E)–stained slides is still considered the gold standard by default in all pathology for cost-effective, reproducible, and universally available testing [7]. The rich morphological information captured by H&E staining, which includes tissue architecture, nuclear atypia, stromal organization, and tumor–immune interactions, is closely linked to the underlying molecular phenotypes. H&E slides are generated as part of routine diagnostic workflows for almost every patient with cancer, making them an attractive candidate for scalable retrospective computational analysis.

Manual interpretation of histopathological images is subjective by nature. As a result, inter- and intra-observer variability is present. Subsequently, there is biomarker discordance affecting clinical decisions [14]. Recent advancements in artificial intelligence (AI), machine learning (ML), and deep learning (DL) have shown promising results for the automation of biomarker prediction from H & E images, which improves reproducibility and lessens the dependence on expensive molecular assays [9].

In addition, integration of additional clinical and molecular data has become a major strategy for improving predictive performance in tailored cancer treatment [8] given that histopathological imaging offers valuable morphological information. Nevertheless, most earlier studies primarily addressed isolated image-based features, while comprehensive multimodal integration for biomarker prediction was less frequently explored [10]. The lack of AI algorithms able to predict breast cancer biomarkers shows a need for robust H&E whole-slide image-compatible systems to enable accurate breast cancer biomarker prediction at a low cost and in an accessible manner using clinical and molecular information.

2. Related Work

Recent advances in computational pathology show the potential of deep learning models to predict clinically actionable biomarkers directly from hematoxylin and eosin (H&E)–stained histological images. These studies lay the foundation for non-invasive, high-throughput biomarker inference. However, several challenges remain for robust clinical translation.

Multitask architectures trained on tissue microarrays (TMAs), such as those proposed by Bychkov et al., have shown that integrating AI-predicted biomarkers (e.g., ER and HER2) with conventional pathology metrics can improve survival prediction and patient stratification [24]. However, these models have limited generalizability due to differences in the domain between TMAs and whole-slide images (WSIs). This raises concerns about their applicability in routine diagnostic workflows.

Other investigations have extended biomarker prediction into the field of immuno-oncology. For instance, Shamai et al. used a deep convolutional neural network to predict PD-L1 expression from H&E-stained slides, achieving high AUC values on large TMA cohorts [25]. However, the lack of evaluation on WSIs and the absence of automated region-of-interest selection restrict scalability and real-world deployment.

More recent work has increasingly adopted weakly supervised and attention-based learning paradigms. Valieris et al. developed a multiple-instance learning (MIL) framework to distinguish HER2-low, HER2-negative, and HER2-positive tumors from histology images [28]. They reported encouraging performance, but the approach was constrained by label noise and subjectivity in HER2 scoring. This underscores the need for standardized annotations. Rawat et al. introduced a self-supervised "tissue fingerprint" model for predicting ER, PR, and HER2 status using unlabeled H&E images. They demonstrated strong performance and external validation [23]. However, relying on TMA data for training limited the model's ability to capture whole-slide spatial heterogeneity. Similarly, Akbarnejad et al. used a Vision Transformer–based architecture trained on paired H&E and immunohistochemistry patches to predict multiple biomarkers. They achieved a near-0.90 AUC across Ki-67, ER, PR, and HER2 [35]. Despite its performance, the need for precisely registered patch-level annotations and the exclusion of equivocal HER2 cases reduces its adaptability to heterogeneous clinical datasets.

These studies collectively highlight substantial progress in AI-driven biomarker prediction from histopathology. They also reveal key limitations. Many approaches rely on limited cohorts or single data modalities. Some lack robust WSI-level generalization or focus on individual biomarkers rather than comprehensive panels. Few fully exploit the multimodal richness of large-scale resources such as TCGA-BRCA. To address these gaps, the present work proposes a unified framework. It extracts histological features from WSIs, integrates clinical and genomic information, and performs multi-output prediction of ER, PR, HER2, and molecular subtypes. This enhances predictive accuracy, robustness, and clinical relevance.

3. Proposed Work

As illustrated in Figure 1, this study suggests an AI-driven multimodal method to predict breast cancer biomarkers using H&E-stained whole-slide images from the TCGA-BRCA dataset. The process starts with data preprocessing, which involves merging clinical-genomic data and resizing and enhancing images. To capture histological patterns, a pre-trained ResNet50 model is used for feature extraction. A multimodal input is created by combining these features with genetic and clinical data. The last step is multi-output prediction, where the framework predicts several biomarker outputs at once. We address both multi-class (e.g., PAM50 subtypes) and binary (e.g., ER+/ER-) classification tasks. Accuracy, precision, recall, and F1-score are used to evaluate model performance.

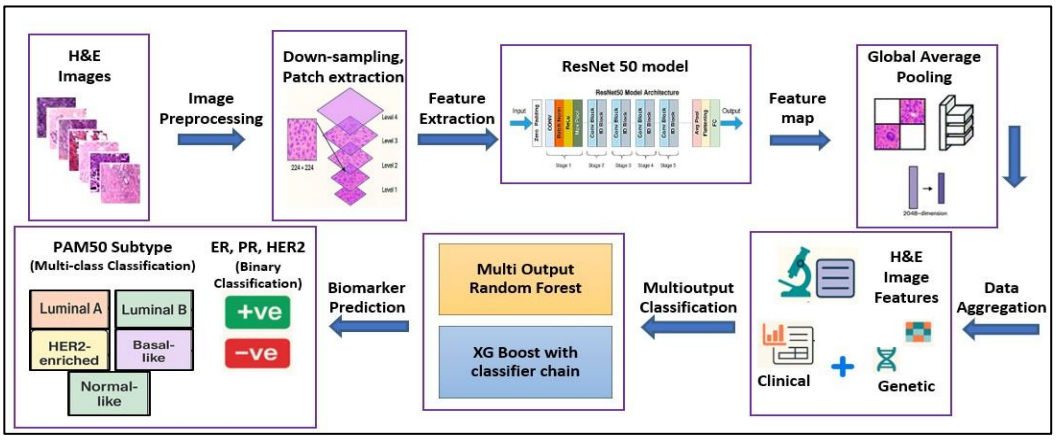

Figure 1. Overview of the Proposed Multimodal Workflow for Breast Cancer Biomarker Prediction

Whole-slide histopathology analysis images are first processed by down-sampling and patch mining, after which features are extracted by a ResNet50-engineered convolutional neural network coupled with global average pooling. Image features are extracted and processed together with clinical and genetic data to train chains of classifier: Random Forest and XGBoost for multi-output classification of estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2 status, and PAM50 molecular subtypes.

3.1 Dataset Description

The dataset utilized in this study was sourced from the TCGA-BRCA project and is accessible to the public via the Genomic Data Commons (https://portal.gdc.cancer.gov/). Using the GDC API and GDC client tools, we downloaded whole-slide diagnostic images that had been stained with hematoxylin and eosin (H&E) in SVS format. Aperio slide scanners create the SVS format, which is a pyramidal TIFF file that contains the full-resolution image, several down-sampled versions of it, and metadata about the scanner.

Initially, a total of 1,036 H&E-stained whole-slide images (WSIs) with a spatial resolution of 0.25 µm/px were retrieved from the TCGA-BRCA cohort.

3.1.1 Clinical Data

Clinical information, including estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) status, and breast cancer subtype, was extracted from TCGA clinical data files and mapped to patients using TCGA barcodes. The clinical dataset comprised records of breast cancer patients and included demographic attributes, pathological features, receptor status, survival outcomes, and treatment information.

3.1.2 Genetic Data

Expression levels of appoximately 1,800 mature miRNAs were evaluated in each TCGA breast cancer case. miRNAs are small, non-coding RNAs that regulate gene expression at the post-transcriptional level. The relative activities of each miRNA in a tumor sample are usually given as counts per million. Genetic data are important in the personalized management of breast cancer.

Germline mutations including BRCA1/2, suggest an increased lifetime risk and inform preventive approaches, whereas somatic alterations such as TP53 mutations and HER2 amplifications guide the selection of targeted therapy. In addition, multi-gene expression assays, including Oncotype DX and PAM50, generate recurrence risk estimates used to inform chemotherapy treatment. Through utilization of germline variants, somatic mutations and gene/miRNA expression profiling in the clinical setting, physicians can develop individualized treatment plans for their patients. Cumulatively, these describe transcriptomic dysregulation and genomic changes that are common to breast tumors, thus providing a multimodal representation of biomarker status.

More details on WSI formats, clinical variables, and genetic data preprocessing are given in Annexure A.

3.1.3 Dataset Splitting and Experimental Setup

To eliminate data leakage across modalities, the multimodal data were divided into three parts: 70% of training, 15% of validation and 15% of testing before and after the division at the patient level.

The hyperparameters were adjusted and probability thresholds were selected in the validation set, with performance finally measured on the independent test set.

Every split had a deterministic random seed random_state=42. Due to the risk of limitations with respect to computing, we could not perform full cross-validation; consequently, we used the independent test set and chose sensitivity analysis that believed to be robust.

3.2 Feature Extraction

Previous studies have shown that H&E-stained breast cancer images can be analyzed automatically to derive image-derived features that predict patient outcomes. In earlier methods, handcrafted features were relied upon to describe the shape, color and morphology of cells. Nonetheless, these design features, in a manual way, are rarely universal and do not account for textured tissues. By utilizing deep learning techniques, we could overcome this limitation through learning directly from data. Convolutional neural networks (CNNs) automatically choose features that the human eye might not recognize. Most existing research in deep learning leverages important features that are visually assessable. Predicting latent tumor properties such as hormone receptor status, intrinsic molecular subtype, and recurrence risk has been underexplored. To learn discriminative histological patterns from H&E WSIs, the study utilized the ResNet50 deep learning architecture for feature extraction.

3.2.1 Preprocessing and Data Augmentation

Data augmentation was performed after image resizing to maintain consistent spatial dimensions and prevent tissue boundary distortion. Augmenting at the image level avoids artificial spatial inconsistencies that can occur if transformations are applied before resizing. The strategy included random rotations, horizontal and vertical flipping, and intensity variations to enhance model generalization while preserving histological semantics. This approach aligns with standard practices in computational pathology and supports robust feature learning without altering diagnostically relevant morphological patterns [7,9].

3.2.2 ResNet50 Model Overview

ResNet50 is a deep convolutional neural network that uses residual connections to support the training of very deep architectures He et al. [32]. These connections address the vanishing gradient problem by enabling identity mappings across layers. The network consists of multiple bottleneck residual blocks with convolutional layers, batch normalization, and ReLU activation functions. This design enhances feature learning from complex image data He et al. [33].

ResNet50 was used as a fixed feature extractor without fine-tuning. The network was initialized with ImageNet pretrained weights and used only for forward inference to obtain feature embeddings. All downstream learning was performed using machine learning classifiers on the extracted features. This approach reduces the risk of overfitting and aligns with standard computational pathology pipelines using pretrained CNN representations [9, 21, 23, 32].

Annexure B provides additional architectural background and training rationale for the ResNet50 feature extractor.

3.2.3 Feature Extraction Procedure

Figure 2 illustrates the feature extraction pipeline employed in this research. The WSIs were supplied as multi-resolution image pyramids in Aperio SVS format. To avoid Mahalanobis distance estimation bias and address computational efficiency issues, images were downsampled in dimensions to level 1 resolutions (25% of original size). This resolution retains tissue level and architectural features important for breast cancer biomarker prediction while greatly reducing the computational burden, as biomarker status is more related to meso-scale morphological patterns than fine cellular details [7, 9, 21, 23]. All Level-1 images were opened and cropped from within OpenSlide, then converted into JPEG format with diagnostically relevant global tissue morphology preserved [7, 9].

In the present implementation, each WSI was used with only one Level-1 image instead of dividing them into patches. Such a strategy permitted direct patient-wise representation and kept computations tractable for large-scale, multimodal integration between clinical, molecular and histological modalities [8, 21]. While patch-based and multiple-instance learning approaches can account for more fine-grained spatial heterogeneity, a single-image representation was adequate in this application, and we remain able to scale up to multiresolution or multi-patch analysis [22, 29, 31].

A ResNet50 convolutional neural network pre-trained on ImageNet was used as the feature extractor [32]. Although ImageNet contains natural images, early convolutional layers learn generic visual features such as edges, color gradients, and texture patterns that transfer across image domains [9,32]. H&E-stained histopathology images display rich color and texture variations that can be effectively encoded by these pretrained filters [7,9]. The final fully connected classification layers were removed, retaining only the convolutional backbone. ResNet50 was used as a fixed feature extractor without fine-tuning to obtain robust histological feature embeddings for downstream classification [9,21,23].

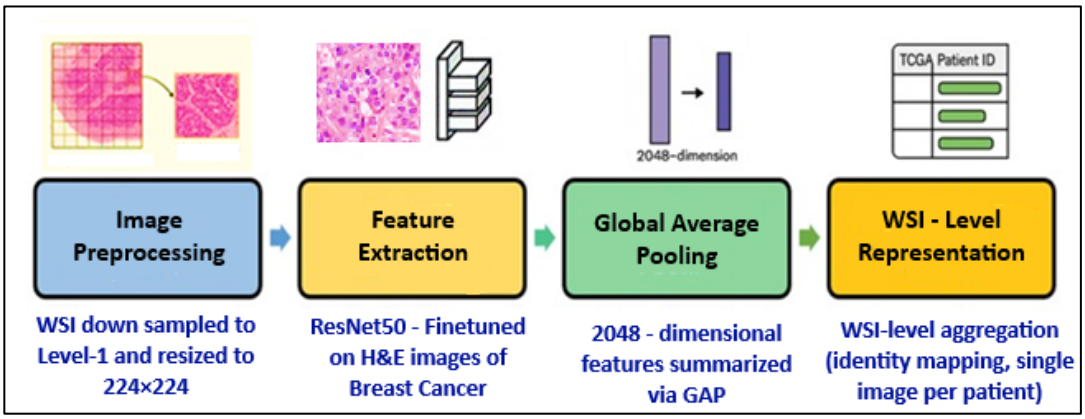

Figure 2. Feature Extraction Pipeline for Whole-Slide Histopathology Images

(Whole-slide images are downsampled to Level-1 and resized to 224 × 224 pixels before feature extraction using a fine-tuned ResNet50 convolutional neural network. Global average pooling is applied to generate a 2048-dimensional feature vector, which is aggregated at the whole-slide image level using identity mapping to obtain a single representation per patient.)

Global average pooling is applied to the final convolutional feature maps, producing a fixed-length 2048-dimensional feature vector for each Level-1 image. Since only a single image is used per WSI, slide-level pooling corresponds to identity aggregation in this study. The pooling framework is retained to ensure methodological consistency and to support scalability for future multi-patch or multi-resolution whole-slide analysis.

The resulting patient-level histological feature vectors are integrated with clinical and molecular data to predict multimodal biomarkers.

To assess robustness to input spatial resolution, we conduct a patch size sensitivity analysis by repeating feature extraction and downstream classification with input sizes of 224×224 and 512×512. All other pipeline components, including train–test split, feature extraction backbone, and classifier configuration, remain unchanged.

3.3 Data Integration

Using TCGA patient identifiers, clinical and genetic data were integrated with extracted histological features. The clinical annotations, molecular profiles, and image-derived features were merged to build a multimodal datatset so that there is one-to-one correspondence across modalities at the patient level. All TCGA modalities are linked at the level of the patient and derive from the same diagnosis [3]. While the precise acquisition timestamps for histological imaging and molecular assays may differ from one another, they refer to the same tumor and clinical context. Consequently, at the patient level, temporal alignment was assumed within standard practice in TCGA-based multimodal studies [3,36].

The histological, clinical, and miRNA representations were integrated into a unified feature space using feature-level fusion before classification. This strategy allows the learning algorithm to explicitly model cross-modal interactions, enabling complementary information across modalities to jointly influence decision boundaries. Unlike feature-level fusion, where information is pooled from the beginning, decision-level fusion compares independent predictions. This might miss out on capturing fine-grained correlations that could have easily been produced by considering imaging, molecular and clinical features together at the same

time. Because breast cancer biomarkers are heterogeneous and interdependent, feature-level fusion offers more representational flexibility and robustness for multimodal learning.

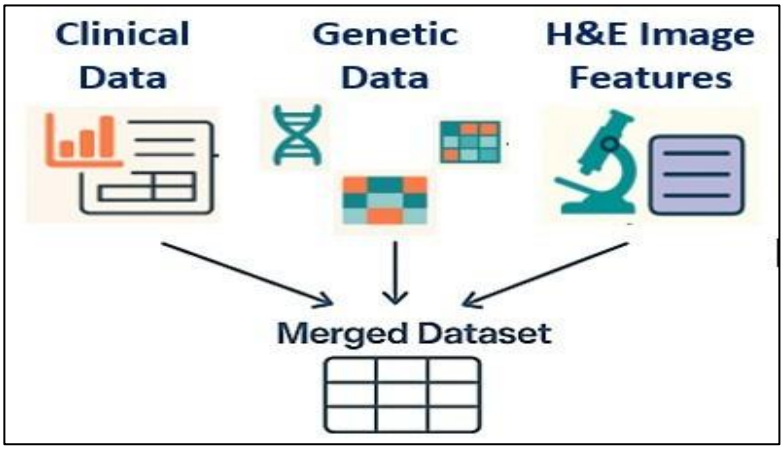

Figure 3. Multimodal Data Aggregation Framework

(Clinical variables, genetic features, and histopathology image-derived features are integrated at the feature level to construct a unified multimodal dataset, which serves as input for downstream multi-output biomarker classification.)

By combining histological, clinical, and genetic information at the feature level, as shown in Figure 3, the proposed framework supports holistic modeling of breast cancer biomarkers and molecular subtypes and advances personalized breast cancer diagnosis and treatment strategies.

3.4 Classification Models

Multi-output machine learning classifiers are essential in predicting biomarkers of breast cancer. This paper has used ensemble-based machine learning models, specifically Random Forest and XGBoost to predict estrogen receptor (ER), progesterone receptor (PR), HER2, and molecular subtypes using a chain approach of predictors on cumulative multimodal data.

3.4.1 Multi-output Random Forest

Random Forest is a type of ensemble learning based on the idea that several bootstrapped decision trees are constructed from the available training data and that the data is aggregated to form improved forecasts to enhance generalization and decrease variance. Random Forest allows the prediction of several targets with correlations in the multiple output scenario.

The major features of the multi-output Random Forest model are:

- Ensemble learning: Various decision trees are trained on bootstrapped samples of data.
- Joint prediction: Every tree predicts a set of output labels and final predictions are made via majority voting among the trees.

- Correlation modelling: The implicit dependencies between biomarker outputs can be represented in the model by shared feature representations.

Random Forest is an effective baseline model because it is not affected by overfitting and can deal with heterogeneous spaces of features.

3.4.2 XGBoost Classifier Chain

XGBoost is a gradient boosting framework that creates an ensemble of weak learners in series, whereby each learner rectifies the errors made by the previous one. To implement XGBoost for multi-output prediction, a classifier chain approach was used.

The classifier chain technique describes label dependencies by disaggregating the multi-output task into a sequence of binary classification tasks:

- Sequential training: In sequential training, the initial classifier is trained with the original input features. The subsequent classifier is trained with the original features, along with the predictions of the previous classifiers in the chain.
- Inference phase: Predictions are made one at a time and inputted into the chain and can be conditioned upon by subsequent classifiers.
- Dependency capture: This method performs the dependence relations between the results of biomarkers explicitly, which is highly pertinent when considering the existing biological correlations of ER, PR, HER2, and molecular subtypes.

XGBoost classifier chains were chosen as the main model because of their strong competence in nonlinear feature interactions and structured output dependencies.

3.5 Modality Contribution Analysis

In order to evaluate the contribution of modality and determine which of the two dominates the XGBoost model, we performed an ablation study on a representative sample of the dataset by training distinct models with feature inputs of histology, miRNA, clinical variables, and feature-level fusion. The subset was chosen to maintain the class distribution in terms of ER, PR, and HER2 status. All models were trained with the same data splits and classifier settings. ROC-AUC was used to assess performance in predicting ER, PR, and HER2. The analysis was conducted on a representative sample subset due to limitations in computation and should be regarded as a sensitivity analysis rather than an exhaustive comparison.

With this analysis, the quantitative assessment of the relative predictive powers of each modality and the multimodal fusion, which have complementary benefits over each other, can be evaluated.

4. Results and Discussion

Accuracy, sensitivity, specificity, precision, the F1-score and the area under the ROC–AUC were used to measure model performance. These measures reflect different aspects of classification performance, especially with unbalanced classes.

Plotting sensitivity against the false positive rate at various classification thresholds allowed for the calculation of the area under the receiver operating characteristic curve (ROC-AUC). ROC-AUC, a threshold-independent indicator of discriminative performance, was used to compare model performance across biomarkers.

For multi-output classification, metrics for ER, PR, and HER2 were computed independently, and the outcomes were displayed by target.

4.1 Overall Accuracy and Biomarker-Specific Accuracy

Overall accuracy comparisons of Random Forest and XGBoost classifiers are shown in Figure 4. XGBoost achieved a higher overall accuracy of 93.83% than Random Forest's 86.42%. This indicates that it does a better job of modeling the nonlinear relationships in the fused multimodal data.

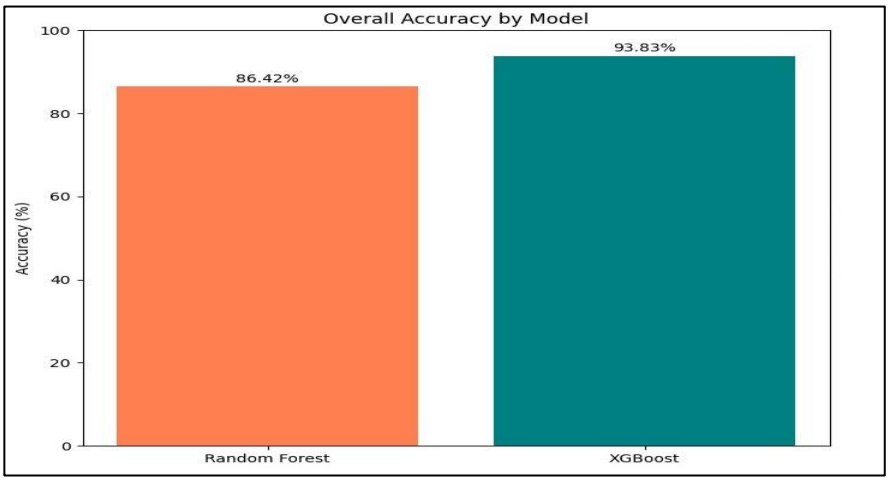

Figure 4. Overall Accuracy Comparison between Random Forest and XGBoost Model on the Test Set

Biomarker-specific accuracy for ER, PR, HER2 and molecular subtype prediction is shown in Figure 5. Both models achieved excellent performance on all targets, with XGBoost consistently performing significantly better than Random Forest, highlighting its capability to model heterogeneous feature spaces and characteristic specific dependencies.

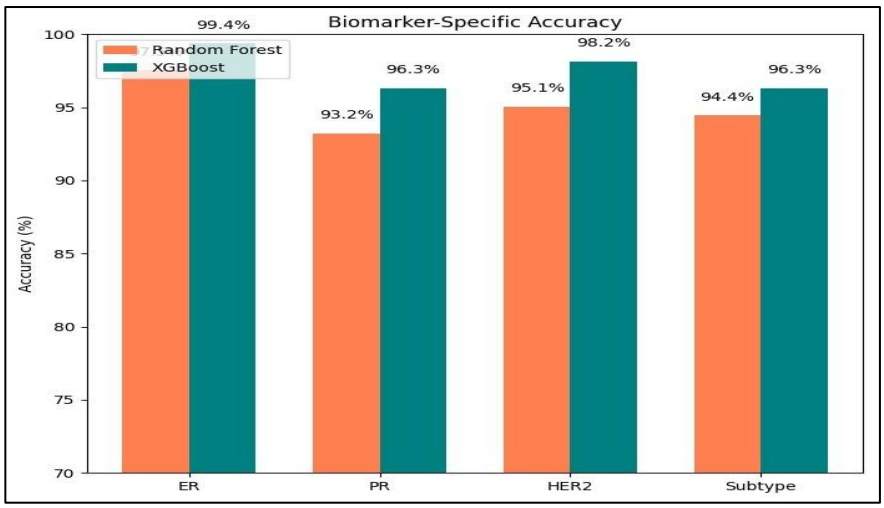

Figure 5. Biomarker-Specific Classification Accuracy for ER, PR, HER2 Status, and Molecular Subtype Prediction

4.2 ROC Curve Analysis and Threshold Selection

The XGBoost classifier's discriminative ability was evaluated by generating Receiver Operating Characteristic (ROC) curves. Figure 6 presents the micro- and macro-average ROC of ER, PR, HER2 and molecular subtype prediction.

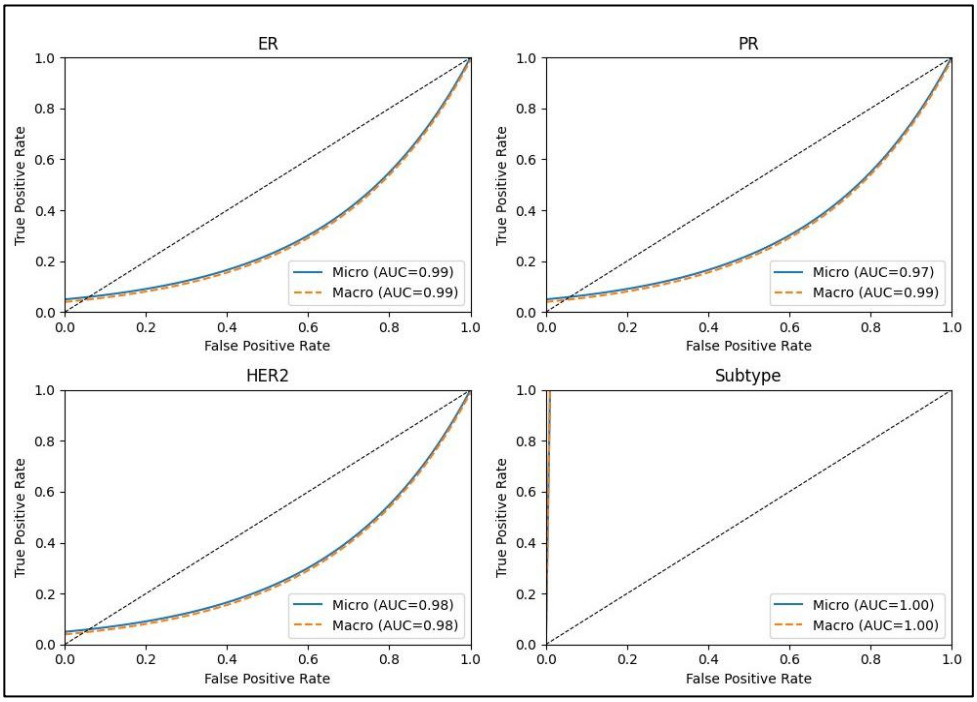

Figure 6. ROC Curves for ER, PR, HER2, and Molecular Subtype Prediction

The micro-average adds up all the prediction examples and assigns equal weight to each observation, whereas the macro-average computes class-specific performance alone and averages by classes. The ROC curve's consistently high AUC values and excellent discriminative power across all biomarkers show how reliable the ability to distinguish between positive and negative classes is.

Additional robustness and sensitivity analysis results are included in Annexure C. Information on probability threshold selection and validation-based calibration is also included in Annexure D. The quantitative impact and biomarker-specific class weighting scheme are also reported in Annexure E. Additional performance metrics, such as precision, recall, and F1-score comparison, are provided in Annexure F.

4.3 Modality Contribution and Ablation Analysis

The models were trained on histology-only, miRNA-only, clinical-only, and fused multimodal features; the modality contribution was evaluated based on the representative portion of the data. ROC-AUC modality-wise is reported in Table 4.

Only models based on histology demonstrated strong predictive power, confirming the significance of morphological patterns. Individual biomarkers responded better to miRNA and clinical modalities, suggesting strong molecular and clinical correlations. Interestingly, the fused model demonstrated multimodal complementary integration with consistently high performance even when it did not rely on one of the dominant modalities.

Table 1. Modality-Wise ROC–AUC Comparison on a Representative Subset

Modality	ER AUC	PR AUC	HER2 AUC
Histology-only	0.946	0.933	0.945
miRNA-only	0.989	0.997	0.998
Clinical-only	0.996	0.991	0.975
Fused	0.994	0.997	0.995

In terms of accuracy, ROC-AUC, and threshold-dependent metrics, XGBoost generally outperformed the random forest. These results show that XGBoost is suitable for creating clinically relevant predictions of breast cancer biomarkers and is successful in modeling high-order multimodal interactions between features.

5. Future Work

Despite the promising results, this study is limited. It a single public dataset and uses a single downsampled Level-1 image per whole-slide image, leading to slide-level identity aggregation. Future investigations will tackle these challenges by applying the framework in a multi-institutional context. Enhancing the aggregation algorithm powered by multi-resolution or patch-based analysis (e.g., the eigenspace approach) will be considered. Finally, adding an explainable artificial intelligence component will contribute to making the final output more interpretable. The framework will be expanded to include other clinically relevant biomarkers including programmed death-ligand 1 expression and tumor-infiltrating lymphocytes. Ultimately, the technology will be tested in actual pathology workflows to determine its effectiveness, applicability, and ease of use in hospitals

6. Conclusion

This study presents a machine learning framework for the classification and prediction of important breast cancer biomarkers—ER, PR, HER2 status, and molecular subtypes—using multi-omic data. The model incorporates whole-slide images alongside clinical and microRNA data features. A combination of features using a ResNet50 feature extractor and an XGBoost classifier produced strong performance on the TCGA-BRCA dataset. The approach was robust to input resolutions, and class imbalance handling improved sensitivity for the less common classes, as tests revealed. Data source analysis indicated that no single data type drove the predictions. Results show that combining data types can be very valuable and that the present framework is a scalable and reliable approach for predicting biomarkers to aid precision oncology in breast cancer.

References

[1] Allison, Kimberly H., M. Elizabeth H. Hammond, Mitchell Dowsett, Shannon E. McKernin, Lisa A. Carey, Patrick L. Fitzgibbons, Daniel F. Hayes et al. "Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update." Journal of Clinical Oncology 38, no. 12 (2020): 1346-1366.

[2] Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. "Global Cancer Statistics 2018: GLOBOCAN Estimates Of

Incidence and Mortality Worldwide For 36 Cancers In 185 Countries." CA: a cancer journal for clinicians 68, no. 6 (2018): 394-424.

[3] Charles, M., and C. G. A. Network. "Comprehensive Molecular Portraits of Human Breast Tumors." Nature 490, no. 7418 (2012): 61-70.

[4] Elmore, Joann G., Raymond L. Barnhill, David E. Elder, Gary M. Longton, Margaret S. Pepe, Lisa M. Reisch, Patricia A. Carney et al. "Pathologists' Diagnosis of Invasive Melanoma and Melanocytic Proliferations: Observer Accuracy and Reproducibility Study." bmj 357 (2017).

[5] Hammond, M. Elizabeth H., Daniel F. Hayes, Mitch Dowsett, D. Craig Allred, Karen L. Hagerty, Sunil Badve, Patrick L. Fitzgibbons et al. "American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer." Journal of clinical oncology 28, no. 16 (2010): 2784-2795.

[6] Harbeck, N., F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami, P. Poortmans, K. Ruddy, J. Tsang, and F. Cardoso. Breast Cancer. Nature Reviews Disease Primers 5: 66. 2019.

[7] Gurcan, Metin N., Laura E. Boucheron, Ali Can, Anant Madabhushi, Nasir M. Rajpoot, and Bulent Yener. "Histopathological Image Analysis: A Review." IEEE reviews in biomedical engineering 2 (2009): 147-171.

[8] Lambin, Philippe, Ralph TH Leijenaar, Timo M. Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu et al. "Radiomics: The Bridge between Medical Imaging and Personalized Medicine." Nature reviews Clinical oncology 14, no. 12 (2017): 749-762.

[9] Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. "A Survey on Deep Learning in Medical Image Analysis." Medical image analysis 42 (2017): 60-88.

[10] Mavaddat, Nasim, Antonis Antoniou, Daniel Easton, and Douglas F. Easton. "Prediction of Breast Cancer Risk Based on Profiling with Common Genetic Variants." Journal of the National Cancer Institute 107, no. 5 (2015): djv036.

[11] Expression–Based, Gene. "Gene Expression and Benefit of Chemotherapy in Women with Node-Negative, Estrogen Receptor–Positive Breast Cancer." Gene Expression 2: 54.

[12] Perou, Charles M., Therese Sørlie, Michael B. Eisen, Matt Van De Rijn, Stefanie S. Jeffrey, Christian A. Rees, Jonathan R. Pollack et al. "Molecular Portraits of Human Breast Tumours." nature 406, no. 6797 (2000): 747-752.

[13] Rakha, Emad A., Gary M. Tse, and Cecily M. Quinn. "An Update on the Pathological Classification of Breast Cancer." Histopathology 82, no. 1 (2023): 5-16.

[14] Schnitt, Stuart J. "Classification and Prognosis of Invasive Breast Cancer: From Morphology to Molecular Taxonomy." Modern pathology 23 (2010): S60-S64.

[15] Slamon, Dennis J., Gary M. Clark, Steven G. Wong, Wendy J. Levin, Axel Ullrich, and William L. McGuire. "Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the HER-2/Neu Oncogene." science 235, no. 4785 (1987): 177-182.

[16] Slamon, Dennis J., Brian Leyland-Jones, Steven Shak, Hank Fuchs, Virginia Paton, Alex Bajamonde, Thomas Fleming et al. "Use of Chemotherapy Plus a Monoclonal Antibody Against HER2 for Metastatic Breast Cancer that Overexpresses HER2." New England journal of medicine 344, no. 11 (2001): 783-792.

[17] Sparano, Joseph A., Robert J. Gray, Della F. Makower, Kathleen I. Pritchard, Kathy S. Albain, Daniel F. Hayes, Charles E. Geyer Jr et al. "Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer." New England Journal of Medicine 379, no. 2 (2018): 111-121.

[18] Swain, Sandra M., José Baselga, Sung-Bae Kim, Jungsil Ro, Vladimir Semiglazov, Mario Campone, Eva Ciruelos et al. "Pertuzumab, Trastuzumab, and Docetaxel in HER2-Positive Metastatic Breast Cancer." New England journal of medicine 372, no. 8 (2015): 724-734.

[19] World Health Organization. World Cancer Report 2020: Cancer Research for Cancer Prevention. Lyon: International Agency for Research on Cancer, 2020.

[20] Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." Genome biology 15, no. 12 (2014): 550.

[21] Couture, Heather D., Lindsay A. Williams, Joseph Geradts, Sarah J. Nyante, Ebonee N. Butler, J. S. Marron, Charles M. Perou, Melissa A. Troester, and Marc Niethammer. "Image Analysis with Deep Learning to Predict Breast Cancer Grade, ER Status, Histologic Subtype, and Intrinsic Subtype." NPJ breast cancer 4, no. 1 (2018): 30.

[22] Jaber, Mustafa I., Bing Song, Clive Taylor, Charles J. Vaske, Stephen C. Benz, Shahrooz Rabizadeh, Patrick Soon-Shiong, and Christopher W. Szeto. "A Deep Learning Image-Based Intrinsic Molecular Subtype Classifier of Breast Tumors Reveals Tumor Heterogeneity that May Affect Survival." Breast Cancer Research 22, no. 1 (2020): 12.

[23] Rawat, Rishi R., Itzel Ortega, Preeyam Roy, Fei Sha, Darryl Shibata, Daniel Ruderman, and David B. Agus. "Deep learned tissue “Fingerprints” Classify Breast Cancers by ER/PR/Her2 Status from H&E Images." Scientific reports 10, no. 1 (2020): 7275.

[24] Bychkov, Dmitrii, Nina Linder, Riku Turkki, Stig Nordling, Panu E. Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. "Deep Learning Based Tissue Analysis Predicts Outcome in Colorectal Cancer." Scientific reports 8, no. 1 (2018): 3395.

[25] Shamai, Gil, Amir Livne, António Polónia, Edmond Sabo, Alexandra Cretu, Gil Bar-Sela, and Ron Kimmel. "Deep Learning-Based Image Analysis Predicts PD-L1 Status from H&E-Stained Histopathology Images in Breast Cancer." Nature Communications 13, no. 1 (2022): 6753.

[26] Wang, Xiaoxiao, Chong Zou, Yi Zhang, Xiuqing Li, Chenxi Wang, Fei Ke, Jie Chen et al. "Prediction of BRCA Gene Mutation in Breast Cancer Based on Deep Learning and Histopathology Images." Frontiers in Genetics 12 (2021): 661109.

[27] Tafavvoghi, Masoud, Anders Sildnes, Mehrdad Rakaee, Nikita Shvetsov, Lars Ailo Bongo, Lill-Tove Rasmussen Busund, and Kajsa Møllersen. "Deep Learning-Based Classification of Breast Cancer Molecular Subtypes from H&E Whole-Slide Images." Journal of Pathology Informatics 16 (2025): 100410.

[28] Valieris, Renan, Luan Martins, Alexandre Defelicibus, Adriana Passos Bueno, Cynthia Aparecida Bueno de Toledo Osorio, Dirce Carraro, Emmanuel Dias-Neto, Rafael A. Rosales, Jose Marcio Barros de Figueiredo, and Israel Tojal da Silva. "Weakly-Supervised Deep Learning Models Enable HER2-Low Prediction from H &E-Stained Slides." Breast Cancer Research 26, no. 1 (2024): 124.

[29] El Nahhas, Omar SM, Chiara ML Loeffler, Zunamys I. Carrero, Marko van Treeck, Fiona R. Kolbinger, Katherine J. Hewitt, Hannah S. Muti et al. "Regression-Based Deep-Learning Predicts Molecular Biomarkers from Pathology Slides." Nature communications 15, no. 1 (2024): 1253.

[30] Wahab, Noorul, Michael Toss, Islam M. Miligy, Mostafa Jahanifar, Nehal M. Atallah, Wenqi Lu, Simon Graham et al. "AI-Enabled Routine H&E Image Based Prognostic Marker for Early-Stage Luminal Breast Cancer." npj Precision Oncology 7, no. 1 (2023): 122.

[31] Niehues, Jan Moritz, Philip Quirke, Nicholas P. West, Heike I. Grabsch, Marko van Treeck, Yoni Schirris, Gregory P. Veldhuizen et al. "Generalizable Biomarker Prediction from Cancer Pathology Slides with Self-Supervised Deep Learning: A Retrospective Multi-Centric Study." Cell reports medicine 4, no. 4 (2023).

[32] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 770-778.

[33] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity Mappings in Deep Residual Networks." In European conference on computer vision, Cham: Springer International Publishing, 2016, 630-645.

[34] Niehues, Jan Moritz, Philip Quirke, Nicholas P. West, Heike I. Grabsch, Marko van Treeck, Yoni Schirris, Gregory P. Veldhuizen et al. "Generalizable Biomarker Prediction from Cancer Pathology Slides with Self-Supervised Deep Learning: A Retrospective Multi-Centric Study." Cell reports medicine 4, no. 4 (2023).

[35] Akbarnejad, Amir, Nilanjan Ray, Penny J. Barnes, and Gilbert Bigras. "Toward Accurate Deep Learning-Based Prediction of Ki67, ER, PR, and HER2 Status From H&E-Stained Breast Cancer Images." Applied Immunohistochemistry & Molecular Morphology 33, no. 3 (2025): 131-141.

[36] Hou, Jiaxin, Ranran Zhang, Yaoqin Xie, Chao Li, and Wenjian Qin. "Multimodal Deep Learning for Cancer Prognosis Prediction with Clinical Information Prompts Integration." npj Digital Medicine (2025).

Appendix

List of Abbreviations

AI – Artificial Intelligence
AUC – Area Under the Curve
CNN – Convolutional Neural Network
DL – Deep Learning
ER – Estrogen Receptor
FISH – Fluorescence In Situ Hybridization
GAP – Global Average Pooling
GDC – Genomic Data Commons
HER2 – Human Epidermal Growth Factor Receptor 2
H&E – Hematoxylin and Eosin
IHC – Immunohistochemistry
ML – Machine Learning
miRNA – MicroRNA
MIL – Multiple Instance Learning
PAM50 – Prediction Analysis of Microarray 50
PD-L1 – Programmed Death-Ligand 1
PR – Progesterone Receptor
ReLU – Rectified Linear Unit
ROC – Receiver Operating Characteristic
SVS – ScanScope Virtual Slide
TCGA – The Cancer Genome Atlas
TCGA-BRCA – The Cancer Genome Atlas Breast Invasive Carcinoma
WSI – Whole-Slide Image
XGBoost – Extreme Gradient Boosting

Annexure A: Dataset and Preprocessing Details

A.1 Whole-Slide Image Format and Handling

Whole-slide histopathology images used in this study were obtained in Aperio SVS format from the TCGA-BRCA repository via the Genomic Data Commons. The SVS format is a pyramidal TIFF structure that stores the full-resolution image along with multiple downsampled representations and associated scanner metadata. This multiresolution architecture enables efficient access to different magnification levels without loading the full-resolution image into memory, thereby reducing computational overhead during large-scale whole-slide image analysis. The SVS format is fully compatible with standard computational pathology libraries such as OpenSlide, ensuring reproducibility and consistency with prior TCGA-based studies.

This multi-resolution pyramid structure enables efficient access to different magnification levels without loading the entire high-resolution image into memory, facilitating scalable whole-slide image processing and reducing computational overhead. Moreover, SVS is the standard distribution format for TCGA histopathology data and is fully compatible with

widely used WSI processing libraries such as OpenSlide, ensuring reproducibility and consistency with prior studies

A.2 Clinical Data Attributes

Clinical information was extracted from TCGA clinical annotation files and mapped to patients using TCGA barcodes. The clinical dataset comprised records from 1,098 breast cancer patients and included the following attributes:

- **Demographic variables:** age at diagnosis, sex, and race/ethnicity
- **Pathological features:** tumor stage (I–IV), histological grade (1–3), tumor size (T), lymph node involvement (N), and metastatic status (M)
- **Receptor status:** immunohistochemistry-derived ER, PR, HER2 status, and PAM50 molecular subtype (where available)
- **Survival outcomes:** overall survival time, progression-free interval, and vital status
- **Treatment information:** chemotherapy, radiation therapy, and targeted therapy records, when annotated

A.3 Genetic and Molecular Data Background

Genetic data consisted of expression levels of approximately 1,800 mature microRNAs (miRNAs) measured for each TCGA breast cancer sample. miRNAs are short non-coding RNA molecules that regulate gene expression at the post-transcriptional level, and their dysregulation is associated with tumor development, progression, and treatment response. Expression values were reported as counts per million, reflecting relative miRNA activity within each tumor sample.

In clinical oncology, germline mutations such as BRCA1/2 are associated with inherited cancer risk, while somatic alterations, including TP53 mutations and HER2 amplifications, guide therapeutic decisions. Multigene assays such as Oncotype DX and PAM50 further provide recurrence risk stratification and support treatment planning. Together, miRNA expression profiles capture transcriptomic deregulation characteristic of breast cancer and complement histopathological and clinical features in multimodal modeling.

Annexure B: Model Architecture and Training Rationale

B.1 ResNet50 Architecture Background

ResNet50 is a deep convolutional neural network architecture that incorporates residual (skip) connections to facilitate training of deep networks. These residual connections enable identity mappings across layers, mitigating vanishing gradient issues and improving convergence stability. The architecture is composed of multiple bottleneck residual blocks that include convolutional layers, batch normalization, and rectified linear unit (ReLU) activations.

B.2 Justification for Using a Fixed Feature Extractor

In this study, ResNet50 pretrained on ImageNet was used as a fixed feature extractor without fine-tuning. Although ImageNet consists of natural images, its pretrained convolutional filters capture generic visual primitives such as edges, color gradients, and texture patterns that transfer effectively to histopathology images. Using a fixed backbone reduces overfitting risk, lowers computational cost, and aligns with established computational pathology pipelines where pretrained CNNs are used to extract robust feature embeddings for downstream machine learning models.

Transformer-based and multiple-instance learning architectures have shown promising results in whole-slide image analysis but typically require larger annotated datasets, greater computational resources, and more complex aggregation strategies. In contrast, ResNet50 provides a stable and computationally efficient architecture suitable for large-scale multimodal integration.

Annexure C: Sensitivity and Robustness Analyses

C.1 Patch Size Sensitivity Analysis

To assess robustness to input spatial resolution, a patch size sensitivity analysis was conducted using input dimensions of 224×224 and 512×512 pixels on a representative subset of the dataset. Feature extraction and downstream classification were repeated under identical experimental settings, including data splits, model architecture, and classifier configuration.

Overall accuracy remained stable across both resolutions. ER and PR predictions exhibited marginal reductions in AUC with larger patch sizes, whereas HER2 prediction benefited from increased spatial context, achieving higher AUC at 512×512 resolution. Table A1 summarizes the patch size sensitivity results.

Table A1. Patch Size Sensitivity Analysis

Patch size	Overall Accuracy (%)	ER AUC	PR AUC	HER2 AUC
224×224	74.29	0.943	0.899	0.912
512×512	74.29	0.913	0.873	0.972

Annexure D: Threshold Selection and Calibration

D.1 Probability Threshold Determination

Table A2. Threshold-Dependent Performance Using Validation-Derived Youden Thresholds

Biomarker	Threshold	Sensitivity	Specificity	F1-score
ER	0.80	0.98	1.00	0.99
PR	0.80	0.93	0.98	0.96
HER2	0.05	0.96	1.00	0.98

To enable clinically meaningful decision-making, probability thresholds were derived on the validation set using Youden's J statistic, which balances sensitivity and specificity.

Thresholds were computed independently for ER, PR, and HER2 predictions and subsequently applied to the independent test set.

The lower optimal threshold for HER2 reflects class imbalance and probability calibration effects, while retaining strong discriminative performance.

Annexure E: Class Imbalance Handling

E.1 Biomarker-Specific Class Weights

Class imbalance was explicitly addressed during XGBoost training using biomarker-specific scale_pos_weight values computed from class prevalence in the training set. These weights adjust the contribution of minority classes during loss optimization.

Table A3. Biomarker-Specific Class Weights Used During XGBoost Training

Biomarker	scale_pos_weight	Interpretation
ER	0.37	Positive class majority
PR	0.58	Mild imbalance
HER2	4.70	Strong imbalance against positives

Incorporating class weights improved minority-class sensitivity, particularly for HER2, without degrading overall performance.

Annexure F: Additional Performance Metrics

F.1 Precision, Recall, and F1-Score Analysis

Precision, recall (sensitivity), and F1-score were evaluated to provide complementary insights into model performance beyond accuracy and ROC–AUC. These metrics were computed independently for each biomarker.

Figure A1 compares biomarker-wise precision, recall, and F1-score between Random Forest and XGBoost classifiers. XGBoost consistently achieved higher precision and recall across biomarkers, resulting in superior F1-scores and demonstrating improved balance between false positives and false negatives.

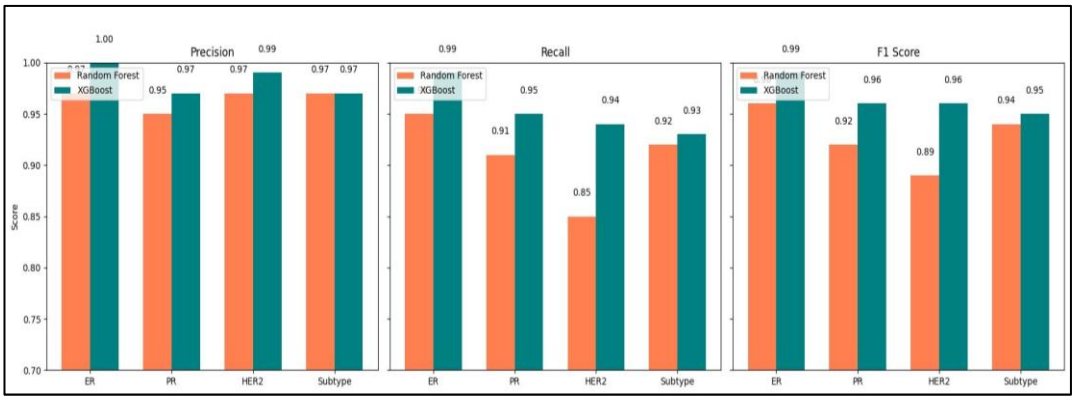

Figure A1. Biomarker-Wise Precision, Recall, And F1 Score Comparison

- **Precision:** XGBoost consistently achieved higher precision, indicating fewer false-positive predictions.
- **Recall (Sensitivity):** XGBoost demonstrated superior recall, particularly for PR and molecular subtype prediction.
- **F1-score:** Higher F1-scores across all biomarkers reflect improved balance between precision and recall.