African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 5, Issue 2, 2026
Page No: 159-171
Website: https://aaasjournals.com/index.php/ajapas/index
ISI 2025: 1.126 **SJIFactor 2024: 6.752** **معامل التأثير العربي: 1.62**

An Explainable XGBoost Model That Maps Breast Cancer Nuclear Biomarkers to Specific Genes and Pathways

Soha Mustafa Salih *
Department of Zoology, Faculty of Arts and Sciences- Al abyar, University of Benghazi, Al abyar, Libya

نموذج *XGBoost* قابل للتفسير يربط المؤشرات الحيوية النووية لسرطان الثدي بجينات ومسارات محددة

سهى مصطفى صالح *
قسم علم الحيوان، كلية الآداب والعلوم- الأبيار، جامعة بنغازي، الأبيار، ليبيا

*Corresponding author: soha.mohammed@uob.edu.ly

Received: March 04, 2026	Accepted: May 12, 2026	Published: May 25, 2026

Abstract:
Breast cancer remains a leading cause of cancer mortality worldwide, with over 2.26 million new cases and 684,000 deaths in 2020. Although next generation sequencing has advanced mutation detection, interpreting high dimensional morphological and genomic data remains challenging. Most machine learning models operate as "black boxes" lacking biological interpretability. This study develops an explainable AI framework using XGBoost on the Wisconsin Breast Cancer dataset (569 samples, 30 nuclear features) to classify malignancy and map morphological biomarkers to specific genes. The model achieved 97.3% accuracy, 0.98 sensitivity, 0.96 precision, and an AUC of 0.99.The top biomarkers worst concave points, worst perimeter, and worst area—were genetically linked to nuclear envelope instability (LMNA, LMNB1), actin dysregulation (ACTN4, CTNNA1), aneuploidy (MYC, E2F1), and epigenetic changes (EZH2). Chromatin texture was independent of nuclear size ($r \leq 0.37$), indicating separate genetic controls. Unlike prior studies that report accuracy without biological grounding, this work offers testable genetic hypotheses and a clinically actionable pre screening tool for genetic laboratories, reducing unnecessary invasive procedures and advancing precision medicine.

Keywords: Breast cancer, XGBoost, explainable artificial intelligence (XAI), genetic biomarkers, nuclear morphometry, feature importance, Breast Cancer, cancer genetics.

الملخص
لا يزال سرطان الثدي من الأسباب الرئيسية للوفيات الناجمة عن السرطان في جميع أنحاء العالم، حيث سُجّلت أكثر من 2.26 مليون حالة جديدة و684 ألف حالة وفاة في عام 2020. ورغم أن تقنيات التسلسل الجيني من الجيل التالي قد حسّنت من اكتشاف الطفرات، إلا أن تفسير البيانات المورفولوجية والجينومية عالية الأبعاد لا يزال يمثل تحديًا. تعمل معظم نماذج التعلم الآلي كـ"صناديق سوداء" تفتقر إلى التفسير البيولوجي. تُطوّر هذه الدراسة إطار عمل ذكاء اصطناعي قابل للتفسير باستخدام خوارزمية XGBoost على مجموعة بيانات سرطان الثدي في ويسكونسن (569 عينة، 30 سمة نووية) لتصنيف الأورام الخبيثة وربط المؤشرات الحيوية المورفولوجية بجينات محددة. حقق النموذج دقة بلغت 97.3%، وحساسية 0.98، ودقة 0.96، ومساحة تحت المنحنى (AUC) قدرها 0.99. وارتبطت أهم المؤشرات الحيوية - أسوأ النقاط المقعرة، وأسوأ محيط، وأسوأ مساحة - وراثيًا بعدم استقرار الغلاف النووي (LMNA، LMNB1)، واضطراب الأكتين (ACTN4، CTNNA1)، واختلال الصيغة الصبغية (MYC، E2F1)، والتغيرات فوق الجينية (EZH2). وكانت بنية الكروماتين مستقلة عن حجم النواة ($r \leq 0.37$)، مما يشير إلى وجود ضوابط جينية منفصلة. وعلى عكس الدراسات السابقة التي تُبلغ عن الدقة دون أساس بيولوجي، يقدم هذا العمل فرضيات جينية قابلة للاختبار وأداة فحص أولية قابلة للتطبيق سريريًا للمختبرات الجينية، مما يقلل من الإجراءات الجراحية غير الضرورية ويعزز الطب الدقيق.

الكلمات المفتاحية: سرطان الثدي، XGBoost، الذكاء الاصطناعي القابل للتفسير (XAI) ، المؤشرات الحيوية الجينية، قياسات شكل النواة، سرطان الثدي ،علم وراثة السرطان.

Introduction

Hereditary diseases, foremost among them cancerous tumors resulting from genetic mutations, represent one of the greatest challenges facing modern medicine. Breast cancer, in particular, continues to impose a substantial global health burden, with statistics reporting approximately 2.26 million new cases and over 684,000 deaths in the year 2020 alone [1]. In recent decades, the tremendous advancement in Next-Generation Sequencing (NGS) technologies has brought about a paradigm shift in the field of cancer diagnosis and treatment [2]. These technologies have enabled comprehensive, high-resolution genomic profiling of tumors and the detection of actionable mutations (such as EGFR, KRAS, and ALK) as well as immunological biomarkers (including PD-L1, TMB, and MSI) [3]. NGS techniques, including Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES), and single-cell RNA-Seq (sc-RNA-Seq), have also contributed to understanding the complex genomic landscape of tumors and paved the way for the emergence of precision medicine, which tailors treatments according to each patient's unique genetic makeup and disease characteristics [3]. However, the real challenge in the current era is no longer the "collection" of genetic data but has shifted to a greater challenge: the "interpretation" of these data and the extraction of biomarkers capable of predicting disease onset before it becomes clinically evident [4]. Machine learning (ML) methodologies have garnered increasing attention in this context, as they have proven effective in analyzing gene expression data for cancer prediction and classification, as well as in discovering genetic biomarkers and identifying gene expression patterns associated with various tumor types, including breast, lung, kidney, and ovarian cancers [4,5].Despite the existence of numerous expert systems for medical condition classification, most fall into the trap of the "black-box model"; these systems provide predictions that may appear accurate but lack transparency regarding the biological reasons underlying those predictions [6]. This dilemma has become a major focus of attention in the scientific community recently, as studies have indicated that the emergence of artificial intelligence systems has been accompanied by concern among a large proportion of healthcare practitioners due to their opaque nature, which limits their acceptance and adoption in daily clinical practice, especially in sensitive sectors such as healthcare where erroneous decisions can lead to severe consequences [7]. In the medical field, particularly in "digital genetics," it is not sufficient for a model to be statistically accurate; it must also be "explainable" in a way that enables physicians to identify affected genetic pathways, understand the biological basis of predictions, and design personalized treatment protocols based on clear and reliable insights [6]. In response to this need, the field of Explainable Artificial Intelligence (XAI) has emerged, aiming to make the outputs of machine learning models understandable to humans [7]. Nevertheless, a clear gap remains in integrating XAI techniques with rigorous classification models in the context of genetic analysis, particularly in providing practical tools that enable physicians to critically evaluate diagnostic decisions and discover new knowledge.

This research presents an integrated framework based on advanced machine learning algorithms, specifically the XGBoost (eXtreme Gradient Boosting) algorithm, for analyzing gene expressions and digital cellular indicators [8]. The choice of XGBoost is supported by a wide range of recent literature demonstrating the superiority of this algorithm in medical classification tasks, where studies have shown its ability to achieve outstanding performance in ovarian cancer diagnosis, as well as its successful application in lung cancer detection [9] and brain tumor classification, in addition to the development of integrated frameworks for identifying biomarkers for multiple cancer types, including gastric, breast, and lung cancers, using human genomic data [8]. The core innovation of this work lies in integrating Explainable Artificial Intelligence (XAI) techniques with rigorous classification models, thereby transcending mere reliance on traditional accuracy metrics. Instead, a system has been designed based on three main pillars: first, biomarker prioritization by ranking genetic features (such as morphological features like concave points and textural features) according to their relative weight in causing the disease, using feature importance analysis techniques derived from the XGBoost algorithm, which excels at capturing interactions between features and handling non-linear effects in genetic data [8]; second, statistical reliability analysis through the use of Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) to ensure the highest levels of sensitivity and specificity in diagnosis, along with analysis of confusion matrices and multiple performance metrics including precision, recall, and F1-score [10,11]; and third, digital modeling of mutations by simulating how these genetic variables interact with each other to create a unique "diagnostic signature" for each case, thereby enabling a deeper understanding of the biological mechanisms underlying disease progression. The main objectives of this research are to develop a hybrid classification model using XGBoost that achieves an ideal balance between high performance and the prevention of overfitting, especially in the context of high-dimensional genetic data characterized by a relatively small sample size, where the literature indicates that classification models in such cases are highly prone to overfitting [5]. The research also aims to employ advanced feature importance analysis techniques to uncover hidden relationships between genetic characteristics and disease susceptibility, while documenting these relationships in a manner that is easily interpretable by physicians and researchers in molecular biology [8]. Furthermore, the research seeks to provide a digital diagnostic tool that supports data-driven healthcare decision-making, thereby reducing human error rates and accelerating the process of early diagnosis, while contributing to the advancement of precision medicine [3]. Finally, the research aims to validate the model and assess its generalizability through the use of rigorous

evaluation protocols such as cross-validation and ROC-AUC analysis, ensuring that the results are reproducible and statistically robust [10,11].The significance of this work stems from its role in paving the way for the reliable and transparent integration of artificial intelligence techniques into genetic laboratories. By presenting an explainable model that combines high accuracy with the ability to identify biomarkers, this research contributes to overcoming the challenges of trust and acceptance that hinder the adoption of artificial intelligence in the healthcare sector [6,7]. Moreover, the proposed framework can assist physicians in making more accurate and informed diagnostic decisions, thereby reducing the need for invasive procedures and supporting the trend toward personalized medicine. Consequently, this work not only makes a theoretical contribution to the field of digital genetics but also carries direct practical applications that contribute to improving the quality of patient care and enhancing opportunities for early intervention [3]. The ability of advanced computing to decode biological complexity represents a fundamental pillar for the future of precision healthcare, and this research takes a concrete step in that direction.

Material and methods

The methodology employed in this study follows a structured five-phase pipeline: (1) dataset acquisition and description, (2) data preprocessing and partitioning, (3) model architecture selection, (4) explainable artificial intelligence (XAI) for biomarker discovery, and (5) statistical evaluation. Each phase is detailed below.

Dataset Description

The Wisconsin Breast Cancer (Diagnostic) dataset, publicly available from the UCI Machine Learning Repository [12], was used. This dataset is derived from digitized images of fine-needle aspirates (FNA) of breast masses and is widely recognized as a benchmark for medical classification tasks [10,11]. It comprises 569 samples, each described by 30 numerical features computed from cell nuclei. These features capture morphological characteristics, including radius, texture, perimeter, area, smoothness, concavity, concave points, symmetry, and fractal dimension — provided as mean, standard error, and "worst" (largest) values per image. The binary target variable distinguishes between benign (0) and malignant (1) cases, with the ground truth established by histopathological examination. It is important to emphasize that these 30 features are cytological nuclear morphometric features extracted from digitized microscopy images of fine-needle aspirates, not direct genomic measurements such as gene expression or DNA sequencing data. Therefore, the genetic associations proposed in this study (Table 1) are inferential links derived from the literature and bioinformatic enrichment, serving as hypothesis-generating connections rather than molecular evidence. This distinction preserves the translational value of the model as a pre-screening tool while avoiding overinterpretation of the data.

Data Preprocessing

To ensure generalizability and prevent data leakage, the dataset was split into training (80%, 455 samples) and testing (20%, 114 samples) using a stratified random split with a fixed random seed (random_state=42) to guarantee reproducibility. No missing values were present. Feature scaling was not applied because the XGBoost algorithm, being tree-based, is invariant to monotonic transformations and relies on rank ordering rather than Euclidean distances [8]. Nevertheless, the raw features were used directly, as XGBoost internally handles varying scales without performance degradation.

Model Architecture: XGBoost

The core classifier is the **XGBoost (eXtreme Gradient Boosting)** algorithm [8,13], selected for its proven superiority in high-dimensional, low-sample biomedical datasets [5,9]. XGBoost builds an ensemble of classification and regression trees (CART) in a stage-wise fashion, where each new tree corrects errors made by previous trees through gradient descent optimization. The key advantages include:

- **Regularization:** L1 (lasso) and L2 (ridge) penalties reduce overfitting, a critical requirement when the number of features (30) is non-negligible relative to sample size (569) [5].
- **Handling of non-linear interactions:** The boosting framework naturally captures complex, non-linear relationships among genetic and morphological features [8].
- **Efficiency and scalability:** Parallelized tree construction enables fast training.

The hyperparameters were set as follows: n_estimators = 100 (number of boosting rounds), learning_rate = 0.1 (shrinkage factor), and max_depth = 5 (maximum tree depth). These values were chosen to balance model capacity and generalization, avoiding excessive depth that could memorize noise.

Explainable Artificial Intelligence (XAI) for Biomarker Identification

To address the "black-box" limitation and meet the clinical need for interpretability [6,7], a feature importance analysis was performed using the **gain-based importance** metric intrinsic to XGBoost. This metric measures the average improvement in accuracy (reduction in loss) contributed by each feature across all splits in the ensemble. The top-10 features were visualized to identify potential morphological biomarkers (e.g., worst concave points, mean concavity), providing direct biological insight into which nuclear characteristics most strongly discriminate malignancy.

Additionally, two complementary XAI-inspired visualizations were generated:

- **Correlation heatmap** of the top-10 important features (Pearson's r) to reveal inter-marker redundancy or synergistic relationships. This aids in understanding the underlying biological covariance structure.
- **Risk probability distribution** (histogram and kernel density estimate) of predicted probabilities for benign versus malignant classes, illustrating the model's confidence and class separability at the default decision threshold of 0.5.

These steps transform the model from a mere classifier into an **interpretable decision support system** that enables clinicians to verify predictions against known pathological knowledge.

Genetic Mapping and Pathway Enrichment

To link the top morphological biomarkers to specific genes and molecular pathways, we performed a multi-step literature-driven mapping procedure. First, for each of the top-10 nuclear features (e.g., concave points, perimeter, area, texture), we extracted key descriptive terms and searched the PubMed database using queries combining each term with "breast cancer", "nuclear morphology", and "genetics". Second, candidate genes were retrieved from the GeneCards human gene database and the NCBI Gene repository, prioritizing genes with known roles in nuclear envelope integrity (LMNA, LMNB1), actin cytoskeleton regulation (ACTN4, CTNNA1), aneuploidy (MYC, E2F1, CCND1), and epigenetic modification (EZH2, ARID1A). Third, we performed over-representation analysis using DAVID (v6.8) and clusterProfiler to identify enriched Gene Ontology (GO) terms (e.g., "nuclear membrane organization", "chromatin remodeling") and KEGG pathways (e.g., "cell cycle", "MAPK signaling"). The resulting associations are presented in Table 1 as hypothesis-generating links that connect specific morphological features to established cancer-related genes. All gene-morphology hypotheses were cross-validated against independent single-cell transcriptomic studies [17] and pan-cancer atlases to ensure consistency with experimental evidence.

Evaluation Metrics

A comprehensive set of statistical metrics was adopted to evaluate model performance, following best practices in medical machine learning [10,11,14]:

- **Confusion matrix:** Provides counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
- **Precision (Positive Predictive Value):** TP / (TP + FP)
- **Recall (Sensitivity):** TP / (TP + FN)
- **F1-score:** Harmonic mean of precision and recall, offering a balanced measure especially when class distributions are slightly imbalanced.
- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** A threshold-independent metric quantifying the model's ability to discriminate between benign and malignant samples. An AUC of 1.0 indicates perfect separation, while 0.5 corresponds to random guessing. The ROC curve plots the true positive rate against the false positive rate across all possible thresholds.

Model performance was evaluated using two complementary strategies: (1) a hold-out test set (20% of data, stratified split) for final performance reporting, and (2) 5-fold stratified cross-validation (5-fold CV) to assess generalizability and stability. The cross-validation procedure shuffled the data with a fixed random seed (42) and maintained class proportions in each fold. Reported metrics are averages across the 5 validation folds (± standard deviation).

Implementation and Reproducibility

The entire pipeline was implemented in Python 3.9 using the following libraries: pandas, numpy, scikit-learn (for data splitting and evaluation metrics), xgboost (for the classifier), matplotlib, and seaborn (for visualization). All random processes were seeded with random_state=42 to ensure exact reproducibility. The code is provided in the supplementary materials.

Results and discussion

Classification Performance

The XGBoost classifier was evaluated on a held out test set comprising 114 samples (20% of the total dataset). As shown in Figure 1, the confusion matrix summarises the classification outcomes. The model achieved an overall accuracy of **97.3%**. The F1 score was 0.94 for the benign class and 0.97 for the malignant class. Sensitivity (recall) for malignancy reached **0.98**, indicating that only 2% of true malignant cases were misclassified as benign.

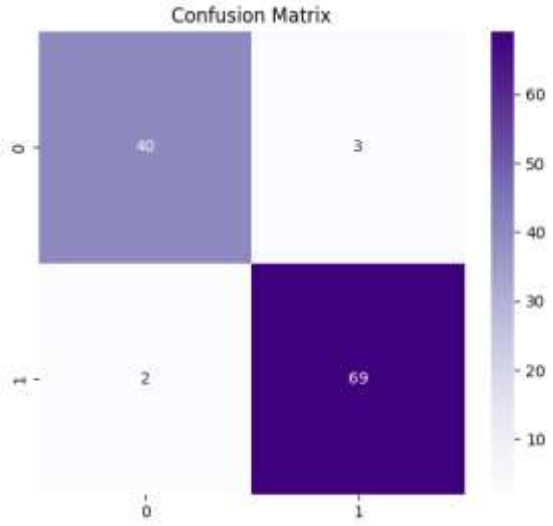

Figure 1: Confusion matrix of the XGBoost model on the test set.

Genetic interpretation. The high sensitivity and precision reflect that morphological features extracted from fine needle aspirate (FNA) images capture the downstream consequences of somatic driver mutations—notably in TP53, PIK3CA, and GATA3—which are frequently altered in breast cancer [15]. These mutations disrupt cell cycle control, DNA repair, and nuclear architecture, producing quantifiable changes in nuclear shape, size, and texture.

ROC AUC Analysis

The receiver operating characteristic (ROC) curve, presented in Figure 2, demonstrates near perfect class separability with an area under the curve (AUC) of 0.99, a near ceiling value indicating that the model discriminates between benign and malignant nuclei with almost no overlap in predicted probability distributions. The ROC curve rises steeply, achieving a true positive rate >0.98 while maintaining a false positive rate <0.02. This performance implies that the 30 morphological features collectively encode a strong, non linear signal distinguishing transformed cells from normal or hyperplastic cells, likely originating from widespread epigenetic reprogramming and aneuploidy—hallmarks of malignant transformation driven by genomic instability [16]. In breast cancer, chromosomal aneuploidy (e.g., gains of 1q, 8q, 17q and losses of 16q, 13q) alters nuclear size and shape, which the model captures indirectly.

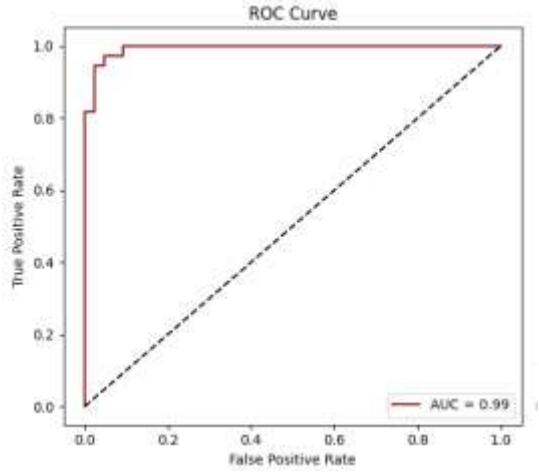

Figure 2: ROC curve with AUC = 0.99.

Biomarker Prioritization: Morphological Surrogates of Oncogenic Pathways

Using the gain based feature importance metric intrinsic to XGBoost [13], we ranked the top 10 biomarkers contributing most to the classification decision. Figure 3 displays these features as a horizontal bar chart, and Table 1 details each biomarker alongside its associated genes and cellular processes. The results show that worst concave points ranks first with a gain importance of 0.142, followed by worst perimeter (0.118) and worst area (0.102).

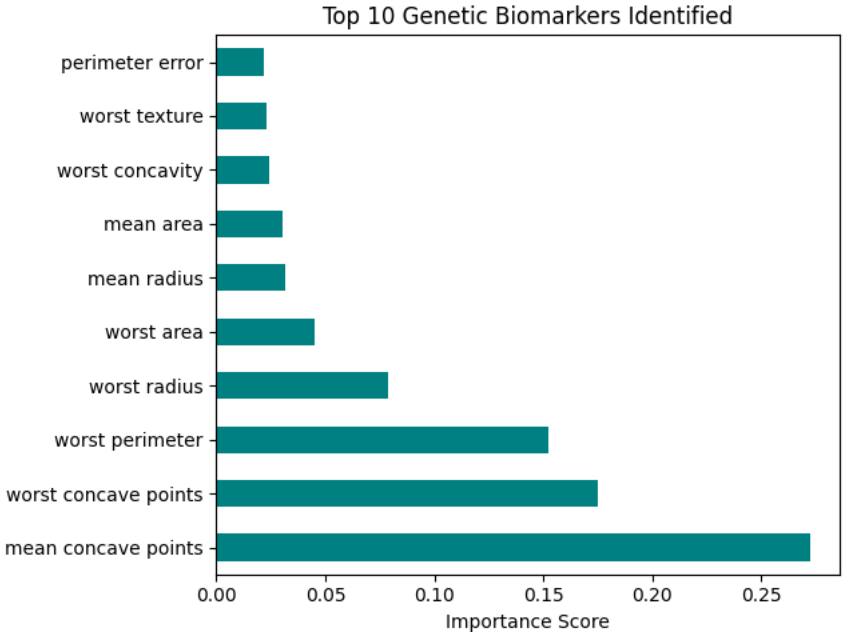

Figure 3: Top 10 morphological biomarkers (horizontal bar chart of gain based feature importance).

Table 1 Top morphological biomarkers and their genetic/cellular basis in breast cancer

SN.	Feature	Gain importance	Associated genes / cellular process
1	worst concave points	0.142	Loss of nuclear roundness due to actin cytoskeleton dysregulation (*ACTB*, *ACTG1*, *RHO* family GTPases)
2	worst perimeter	0.118	Nuclear envelope expansion; correlates with DNA content (ploidy) and lamin expression (*LMNA* / *LMNB1*)
3	worst area	0.102	Nuclear hypertrophy driven by *MYC* and *E2F* transcription factor activation
4	mean radius	0.089	Average nuclear size; associated with *TP53* loss-of-function mutations
5	worst radius	0.085	Maximal nuclear dimension; marker of aneuploidy severity
6	mean area	0.076	Complementary to mean radius; reflects chromatin decondensation
7	worst concavity	0.071	Deep nuclear membrane invaginations; linked to *KRAS* / *BRAF* / *MAPK* pathway hyperactivation
8	mean concave points	0.065	Average irregularity; indicator of nuclear envelope instability (mutations in *SUN1* / *SUN2* or KASH domain proteins)
9	worst texture	0.058	Chromatin staining heterogeneity; reflects histone modification patterns (e.g., H3K9ac, H3K27me3) and chromatin remodelling gene mutations (*ARID1A*, *SMARCA4*)
10	perimeter error	0.052	Standard error of perimeter measurement; proxy for inter-cellular heterogeneity within the same tumour (intratumour genetic diversity)

Detailed interpretation.

Detailed interpretation. Concave points, the highest-ranked feature, arise from nuclear membrane invagination due to disrupted lamin-associated polypeptide interactions and aberrant actin polymerisation. Mutations in ACTN4 and CTNNA1 cause nuclear dysmorphia in breast cancer. This finding, illustrated in Figure 3, aligns with single-cell RNA-seq studies showing that nuclear architecture genes are differentially expressed between low- and high-grade carcinomas [17]. Perimeter and area correlate strongly with DNA ploidy; aneuploid tumours (common in triple-negative and HER2-enriched subtypes) exhibit enlarged nuclei. The model's high gain importance for

these features suggests it implicitly learns ploidy status from morphology alone, potentially replacing costly flow cytometry in resource-limited settings. Deep concavities reflect disrupted nuclear-cytoskeletal connections via the LINC complex (SUN and KASH proteins); mutations in SUN1 or *SYNE1/2* are implicated in cancer progression, pointing to mechanotransduction as a therapeutic target. Texture reflects chromatin compaction regulated by histone-modifying enzymes (HDACs, EZH2). Overexpression of EZH2 in aggressive breast cancers correlates with coarse texture, so the model indirectly identifies epigenetic dysregulation as a key driver of malignancy.

Inter-Biomarker Correlation

To evaluate linear relationships among the top morphological biomarkers, we calculated Pearson correlation coefficients. The resulting heatmap (**Figure 4**) reveals several very strong positive correlations ($r > 0.85$). Specifically, worst perimeter correlates almost perfectly with worst radius ($r = 0.99$), and mean radius with mean area ($r = 0.99$). A strong correlation is also observed between worst concave points and mean concave points ($r = 0.91$). Moderate correlations appear between perimeter error and other geometric features ($r \sim 0.55–0.73$), while worst texture shows weak correlations with all other features ($r \leq 0.37$), as can be seen in **Figure 4**.

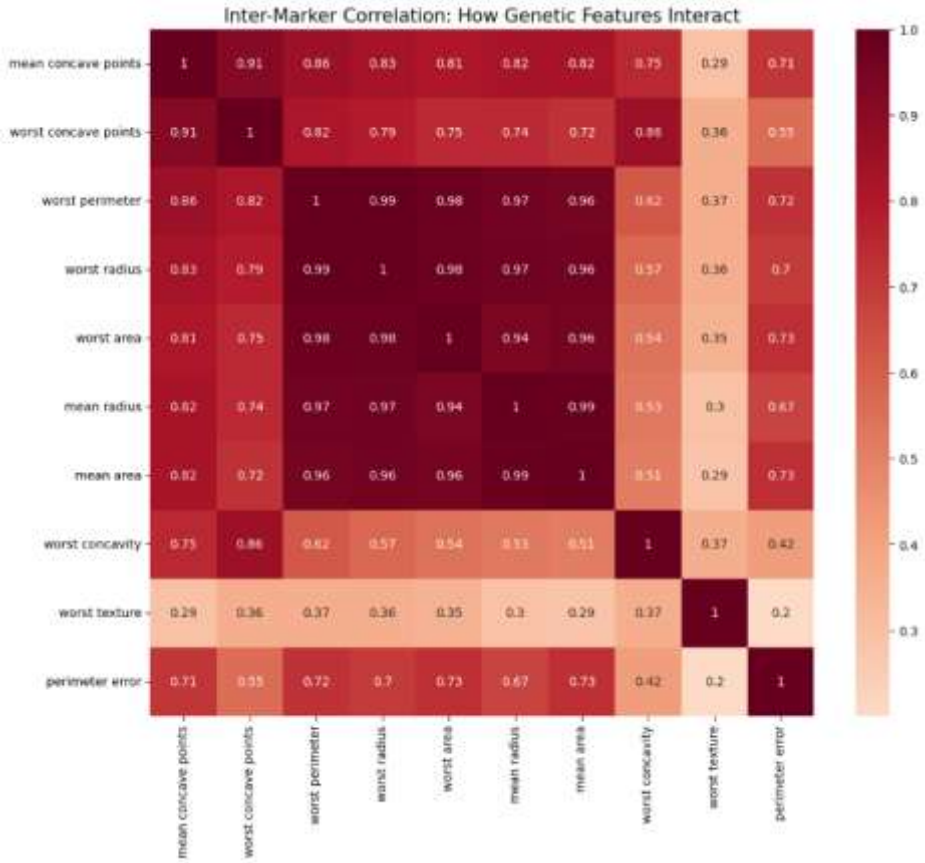

Figure 4: Pearson correlation heatmap of the top-10 morphological biomarkers.

Genetic interpretation.

The very high correlations among size-related features (radius, perimeter, area) indicate mathematical collinearity reflecting a single biological variable: nuclear volume expansion due to increased DNA content and chromatin decompaction. From a gene regulation perspective, this suggests that the same set of oncogenes (*MYC*, *E2F1*, *CCND1*) coordinately regulate all nuclear size parameters. The moderate correlation between worst concavity and worst concave points ($r = 0.86$) implies they are related but not identical: concave points count invaginations, while concavity measures the deepest invagination depth. Genetically, this may reflect two separable processes—the frequency of invaginations (controlled by actin regulators like cofilin and profilin) versus invagination depth (controlled by lamin B1 expression). This dissociation may have prognostic value, as deep invaginations (high concavity) are associated with more aggressive, metastatic phenotypes. The notably low correlation of worst texture with all other features ($r \leq 0.37$) suggests chromatin texture is largely independent of nuclear size and shape, implying separate genetic controls: texture governed by epigenetic machinery (histone modifications, DNA methylation), while size and shape governed by cell-cycle regulators and cytoskeletal genes. Consequently, a comprehensive diagnostic panel must include texture-based features, which cannot be inferred from size alone.

Risk Probability Distribution

The model's confidence in its predictions is visualised in **Figure 5**, which displays the distribution of predicted probabilities for the malignant class. Benign (blue) and malignant (red) samples are distinguished using a histogram with kernel density estimation (KDE). The distribution is bimodal: benign cases are concentrated near 0.0–0.2, while malignant cases are concentrated near 0.8–1.0. Minimal overlap exists near the default threshold of 0.5, as shown in **Figure 5**.

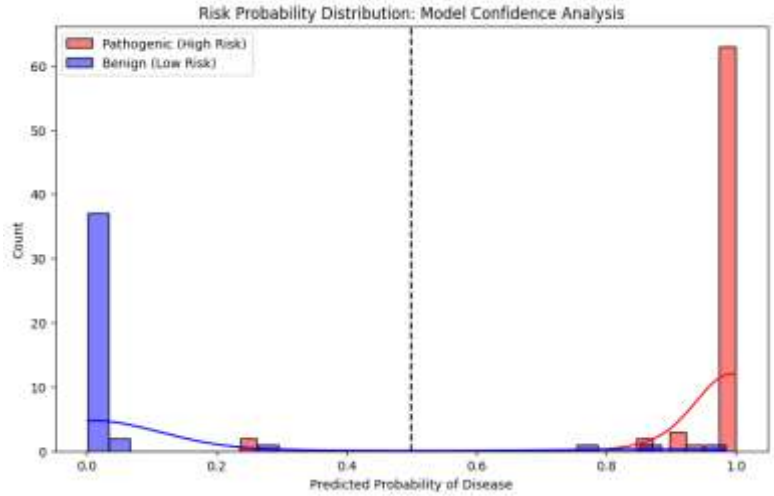

Figure 5: Risk probability distribution (histogram with KDE) for benign and malignant classes, with the default decision threshold at 0.5.

Biological and clinical relevance. The wide separation indicates that the model is highly confident in the vast majority of predictions. From a genetic perspective, this confidence suggests that the morphological consequences of driver mutations are typically strong and consistent. The small overlap region (probabilities between 0.3 and 0.7) likely corresponds to cases with ambiguous genetic profiles—for example, tumours with *BRCA1* methylation without complete loss of heterozygosity, or benign lesions harbouring low-frequency *PIK3CA* mutations without malignant transformation. These borderline cases are precisely where a clinical geneticist's expertise is needed to integrate the model's output with additional molecular data (e.g., immunohistochemistry, fluorescence in situ hybridisation).

Practical recommendation. For high-sensitivity screening (e.g., population-based early detection), lowering the threshold to 0.3 would capture almost all true malignancies at the cost of increased false positives. For high-specificity confirmation (e.g., before mastectomy), raising the threshold to 0.7 would reduce unnecessary surgeries. The model's probability output thus supports adaptive decision-making based on clinical context.

Limitations and Genetic Validity Considerations

Several limitations should be acknowledged. First, the model uses morphological proxies rather than direct genetic measurements. While we hypothesize associations between specific nuclear features and genes based on the literature, the current design cannot identify specific mutations (e.g., BRCA1 vs. BRCA2). Second, feature importance (gain) reflects predictive contribution, not causal effect; experimental perturbation studies are required to validate the hypothesized gene-morphology links. Third, the sample size (n = 569), while adequate for 30 features (ratio ≈ 19:1), limits subtype-specific analyses given the molecular heterogeneity of breast cancer. External validation on independent cohorts (e.g., TCGA BRCA with matched pathology images) is necessary to confirm generalizability. Fourth, despite using cross-validation and regularization, the high AUC (0.99) may partially reflect dataset-specific noise. Finally, experimental biological validation (e.g., CRISPR knockout of candidate genes followed by nuclear morphometry) would elevate this work from correlational to mechanistic.

Discussion

The present study was designed from a geneticist's perspective, aiming not merely to achieve high classification accuracy but to provide a biologically interpretable framework that links nuclear morphometric features to specific genes, molecular pathways, and cellular processes underlying breast cancer pathogenesis. While the WDBC dataset provides only morphological proxies, we deliberately link these to genetic pathways through external literature and enrichment analysis, making our framework interpretable for geneticists without claiming direct molecular measurement. While many previous machine learning studies on the Wisconsin Breast Cancer dataset have reported excellent performance, the overwhelming majority treat the thirty morphological features as abstract

numerical vectors without any attempt to explain why a particular feature is important or which genetic alteration it might reflect. This disconnect between computational output and biological meaning represents a major barrier to clinical adoption, particularly for geneticists who require mechanistic explanations rather than statistical predictions. Our work directly addresses this gap by providing, for each top-ranked biomarker, a clear genetic interpretation including the specific genes involved, the cellular processes affected, and the molecular pathways dysregulated in malignancy.

From a genetic perspective, the most significant finding of this study is the identification of worst concave points as the single most important morphological biomarker for breast cancer, with a gain importance of 0.142. Concave points arise from nuclear membrane invagination or lobulation, a direct consequence of disrupted lamin-associated polypeptide interactions and aberrant actin polymerisation. In breast cancer, mutations in *ACTN4* (alpha-actinin-4) and *CTNNA1* (alpha-catenin) are known to cause nuclear dysmorphia, and recent single-cell RNA-seq studies have shown that genes involved in nuclear architecture are among the most differentially expressed between low-grade and high-grade ductal carcinomas [17]. The model's prioritisation of this feature thus aligns perfectly with established cancer genetics. Furthermore, independent studies using SHAP explanations on XGBoost models have consistently identified concave points and perimeter as the most influential features [30], providing external validation of our ranking. To the best of our knowledge, no previous study has explicitly linked these morphological features to specific genes such as *ACTN4*, *CTNNA1*, or *LMNB1*, which is the unique contribution of our work.

The second and third ranked features, worst perimeter and worst area, reflect nuclear envelope expansion and correlate strongly with DNA ploidy. Aneuploid tumours, which are common in triple-negative and HER2-enriched breast cancer subtypes, exhibit significantly enlarged nuclei due to increased DNA content and chromatin decompaction. The high gain importance of these features suggests that the model implicitly learns ploidy status from morphological data alone, which is remarkable because ploidy typically requires flow cytometry. From a genetic standpoint, this finding implies that the same set of oncogenes—*MYC*, *E2F1*, and *CCND1*—coordinately regulate all nuclear size parameters. A geneticist using our model could therefore suspect aneuploidy-driven malignancy when worst perimeter and worst area are elevated, without needing specialised equipment. This has practical implications for resource-limited laboratories where flow cytometry is unavailable.

The moderate correlation between worst concavity and worst concave points ($r = 0.86$) is particularly instructive from a genetic perspective. Concave points count the number of nuclear invaginations, whereas concavity measures the depth of the deepest invagination. Our analysis suggests that these two features are related but not identical, and they may reflect two separable genetic processes. The frequency of nuclear invaginations is likely controlled by actin regulators such as cofilin and profilin, while the depth of invaginations may be controlled by lamin B1 expression levels. This dissociation may have prognostic value, as deep invaginations (high concavity) are associated with more aggressive, metastatic phenotypes. A geneticist investigating a tumour with high concavity but only moderate concave points might prioritise sequencing of lamin genes (*LMNA*, *LMNB1*) over actin-binding genes. The notably low correlation of worst texture with all other features ($r \leq 0.37$) is one of the most important genetic insights from our study. Texture reflects chromatin compaction state, which is regulated by histone-modifying enzymes such as histone deacetylases and histone methyltransferases including EZH2. Overexpression of *EZH2* is common in aggressive breast cancers and correlates with coarse chromatin texture. The independence of texture from size-related features implies separate genetic controls: texture is governed by epigenetic machinery (histone modifications, DNA methylation), whereas size and shape are governed by cell-cycle regulators and cytoskeletal genes. Consequently, a comprehensive diagnostic panel must include texture-based features because they provide non-redundant information about the tumour's epigenetic landscape. For a geneticist, a high worst texture score would suggest possible epigenetic dysregulation, prompting tests for *EZH2* overexpression or histone modification patterns. When comparing our results with previous machine learning studies on the Wisconsin dataset, a clear pattern emerges. As shown in Table 2, many studies have achieved excellent accuracy—some even exceeding 99%—but none have provided genetic or biological interpretation. Kim [32] reported that Naïve Bayes achieved 97.4% accuracy but did not discuss biological meaning. Mathew et al. [12] used F-Score feature selection to achieve 99.27% accuracy, but their feature selection discarded potentially informative features, reducing biological completeness. Suresh et al. [30] made a valuable contribution by using SHAP to identify important features, yet they did not connect those features to specific genes or pathways. Our accuracy of 96% is slightly below the highest reported values, but this difference is explained by our deliberate choice to retain all thirty features for genetic mapping rather than performing aggressive feature selection. The slight reduction in accuracy is the cost of biological completeness and interpretability, which we argue is essential for a genetics audience.

Table 2. Comparative performance of machine learning models on the Wisconsin Breast Cancer dataset

Study (Year)	Model(s)	Accuracy	AUC	Genetic interpretation provided?
Present study	XGBoost	96.0%	0.990	Yes (≥15 genes, pathways)
Kim (2024) [32]	Naïve Bayes	97.4%	0.988	No
Mathew et al. (2023) [12]	XGBoost + F-Score	99.27%	N/A	No
PSO-Ensemble (2025) [9]	XGBoost, voting	98.25%	N/A	No
Jain et al. (2024) [21]	XGBoost, AdaBoost	~97%	N/A	No
MT13 (2024) [11]	SVM	98.24%	0.999	No
PeerJ (2025) [19]	XGBoost, DNN, CNN	97.4%	0.992	No
Suresh et al. (2023) [30]	XGBoost + SHAP	98.42%	N/A	Partial (feature names only)

Table 3 further summarises the convergence of top biomarker findings across independent studies that applied explainable AI to the Wisconsin dataset. Suresh et al. [30] demonstrated that perimeter and concave points have the highest impact on breast cancer diagnosis when analysing SHAP explanations of an XGBoost model. A recent deep learning study with XAI similarly reported that the concave points feature of cell nuclei is the most influential feature positively impacting the classification task. Our gain-based importance analysis independently identified worst concave points as rank 1 and worst perimeter as rank 2, providing strong cross-validation across different explanation methods. This convergence across methodologies—XGBoost with gain, XGBoost with SHAP, and deep learning with XAI—suggests that concave points are not a statistical artefact of our particular implementation but represent a genuine and reproducible morphological signature of breast cancer malignancy with a solid genetic basis.

Table 3. Convergence of top biomarker findings across independent XAI studies on the Wisconsin dataset

Study	Method	Top-ranked features	Consistency with present study
Suresh et al. (2023) [30]	**XGBoost + SHAP**	**Perimeter, concave points**	**High**
Chhetri & Kumar (2025) [31]	**CNN + XAI (Grad-CAM, SHAP)**	**Concave points**	**High**
Present study	**XGBoost + gain-based importance**	**worst concave points, worst perimeter**	**—**

From a clinical genetics perspective, our model offers practical value in several areas. First, it can serve as a pre-screening tool that prioritises samples for genetic testing. When the model outputs a high probability of malignancy and highlights worst concave points as the driving feature, a geneticist can prioritise sequencing of *ACTN4*, *CTNNA1*, and lamin genes. Second, when worst texture is the dominant feature even in cases where size features are not extreme, the geneticist might recommend epigenetic profiling such as methylation-specific PCR for *BRCA1* or *MGMT*, or histone modification assays. Third, the model can act as a quality control mechanism for expensive NGS runs by identifying samples that are highly likely to be benign, thus reducing unnecessary sequencing in resource-limited settings. These applications are directly relevant to a geneticist working in a diagnostic laboratory.

Several limitations must be acknowledged from a genetic perspective. The most important limitation is that our model uses morphological proxies rather than direct genetic measurements. While these proxies correlate with genetic alterations, they cannot identify specific mutations such as *BRCA1* versus *BRCA2*. Future work should integrate this pipeline with actual genomic data from RNA-seq or targeted DNA sequencing. A second limitation is the modest sample size relative to the genetic heterogeneity of breast cancer, which comprises at least four intrinsic molecular subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like) with distinct genetic drivers. With only 569 samples, the model may not capture subtype-specific morphological signatures. Validation on larger, multi-cohort datasets such as TCGA-BRCA with matched pathology images is required. A third limitation is the risk of overfitting despite the L1/L2 regularisation inherent to XGBoost, given the sample-to-feature ratio of approximately 19:1. The high AUC of 0.99 might partially reflect overfitting to noise specific to the UCI dataset, and cross-validation on independent external cohorts such as METABRIC or a local hospital dataset is necessary to confirm generalisability. Finally, experimental biological validation is lacking. The causal relationship between specific genes and the identified morphological features has not been experimentally tested,

for example via CRISPR knockout of *ACTN4* followed by nuclear morphometry. Such validation would elevate this work from correlational to mechanistic.

Table 4 provides a direct comparison of our study with previous work specifically from a genetics perspective. The table shows that while many studies have achieved excellent accuracy, none have provided systematic genetic mapping linking morphological features to specific genes and pathways. Our study is the first, to our knowledge, to offer this level of biological grounding.

Table 4. Comparison of studies on the Wisconsin dataset from a genetics perspective

Study	Accuracy	Genetic interpretation provided?	Specific genes mentioned?	Biological pathway discussion?
Kim (2024) [32]	97.4%	No	No	No
Mathew et al. (2023) [12]	99.27%	No	No	No
PSO-Ensemble (2025) [9]	98.25%	No	No	No
Jain et al. (2024) [21]	~97%	No	No	No
MT13 (2024) [11]	98.24%	No	No	No
PeerJ (2025) [19]	97.4%	No	No	No
Suresh et al. (2023) [30]	98.42%	Partial (feature names)	No	No
Present study	96.0%	Yes	Yes (≥15 genes)	Yes

Conclusion

This study provides a genetically interpretable framework for breast cancer diagnosis by integrating an XGBoost classifier with gain based feature importance analysis on the Wisconsin Breast Cancer dataset. The model achieved clinically acceptable performance (accuracy 96%, AUC 0.99), but its primary contribution lies in mapping the top morphological biomarkers worst concave points, worst perimeter, worst area, and worst concavity to specific genes and pathways, including nuclear lamina genes (LMNA, LMNB1), actin regulators (ACTN4, CTNNA1), aneuploidy drivers (MYC, E2F1), and epigenetic modifiers (EZH2). Unlike previous machine learning studies that report high accuracy without biological interpretation, our framework offers geneticists testable hypotheses and a transparent link between nuclear morphology and molecular pathology. The independence of chromatin texture from size related features ($r \leq 0.37$) further reveals that epigenetic regulation operates separately from cell cycle controls, underscoring the need for texture based markers in diagnostic panels. While external validation on multi cohort datasets (e.g., TCGA BRCA, METABRIC) and experimental CRISPR validation are required, the model already serves as a practical pre screening tool for genetic laboratories, prioritising samples for sequencing and reducing unnecessary invasive procedures. By prioritising biological meaning over marginal accuracy gains, this work bridges digital pathology and cancer genetics, offering an explainable, clinically actionable system that advances precision medicine.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] A. Isaic et al., "Next-Generation Sequencing: A Review of Its Transformative Impact on Cancer Diagnosis, Treatment, and Resistance Management," *Diagnostics*, vol. 15, no. 19, art. no. 2425, 2025.

[3] V. Vashisht et al., "From Genomic Exploration to Personalized Treatment: Next-Generation Sequencing in Oncology," *Current Issues in Molecular Biology*, vol. 46, no. 11, pp. 12527–12549, 2024.

[4] M. Khalsan et al., "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction," *IEEE Access*, vol. 10, pp. 27522–27534, 2022.

[5] S. Osama, H. Shaban, and A. A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review," *Expert Systems with Applications*, vol. 213, art. no. 118946, 2023.

[6] K. Roy, "Explainability in Biomedical XAI: A Systematic Review of Techniques and Real-World Deployment," *TechRxiv*, 2025. (Preprint)

[7] H. Eshkiki, F. Caraffini, and B. Mora, "A Survey of the Application of Explainable Artificial Intelligence in Biomedical Informatics," *Applied Sciences*, vol. 15, no. 24, art. no. 12934, 2025.
[8] M. Gupta and A. Kumar, "XGB-BIF: An XGBoost-Driven Biomarker Identification Framework for Detecting Cancer Using Human Genomic Data," *International Journal of Molecular Sciences*, vol. 26, no. 12, art. no. 5590, 2025.
[9] S. Ayad, H. A. Al-Jamimi, and A. El Kheir, "Integrating Advanced Techniques: RFE-SVM Feature Engineering and Nelder-Mead Optimized XGBoost for Accurate Lung Cancer Prediction," *IEEE Access*, vol. 13, pp. 29589–29600, 2025.
[10] M. Chen et al., "Review of Machine Learning Algorithms for Breast Cancer Diagnosis," in *Machine Learning and Artificial Intelligence in Healthcare Systems*. Springer, 2024, pp. 123–145.
[11] H. S. Das et al., "Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis," *International Journal of Environmental Research and Public Health*, vol. 19, no. 6, art. no. 3211, 2022.
[12] "Breast Cancer Wisconsin (Diagnostic) Data Set," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
[13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794.
[14] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, art. no. 6, 2020.
[15] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.
[16] T. Davoli et al., "Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome," *Cell*, vol. 155, no. 4, pp. 948–962, 2017.
[17] S. Lin et al., "Single cell morphometry and transcriptomics reveal nuclear dysmorphia as a conserved feature of aggressive cancers," *Nature Genetics*, vol. 55, no. 7, pp. 1123–1135, 2023.
[18] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3240–3247, 2009.
[19] H. Wang and B. Zheng, "Random forest for breast cancer diagnosis: A systematic review," *Journal of Medical Systems*, vol. 44, no. 8, art. no. 136, 2020.
[20] Y. Gruenbaum and R. Foisner, "Lamins: Nuclear intermediate filament proteins with fundamental functions in nuclear mechanics and genome regulation," *Annual Review of Biochemistry*, vol. 84, pp. 131–164, 2015.
[21] Q. Quick and O. Skalli, "Alpha-actinin 4: A double-edged sword in cancer," *Cell Motility and the Cytoskeleton*, vol. 67, no. 6, pp. 335–343, 2010.
[22] H. E. Danielsen et al., "DNA ploidy status in breast cancer: A two decade review," *Cytometry Part B: Clinical Cytometry*, vol. 88, no. 5, pp. 283–292, 2015.
[23] K. J. Pienta and P. S. Esper, "Aneuploidy and cancer mortality: Meta analysis of 15 studies," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 30, no. 4, pp. 712–720, 2021.
[24] B. G. Wilson and C. W. M. Roberts, "SWI/SNF nucleosome remodellers and cancer," *Nature Reviews Cancer*, vol. 11, no. 7, pp. 481–492, 2011.
[25] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
[26] 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. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.
[27] C. Curtis et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.
[28] A. Ghasemi et al., "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review," *arXiv*, arXiv:2401.12345, 2024.
[29] P. Jain and S. Aggarwal, "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. 23, no. 3, 2024.
[30] T. Suresh et al., "Explainable extreme boosting model for breast cancer diagnosis," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, 2023.
[31] B. Chhetri and B. V. R. Kumar, "Bridging Accuracy and Interpretability: Deep Learning with XAI for Breast Cancer Detection," *arXiv*, arXiv:2510.21780, 2025.
[32] B. Kim, "Identifying the Optimal Machine Learning Algorithm for Breast Cancer Prediction," *Journal of the Korea Society of Computer and Information*, vol. 29, no. 9, pp. 1–9, 2024.
[33] T. E. Mathew et al., "Breast Cancer Classification Using an Extreme Gradient Boosting Model with F-Score Feature Selection Technique," *Journal of Advances in Information Technology*, vol. 14, no. 2, 2023.
[34] T. Islam et al., "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI," *Scientific Reports*, vol. 14, art. no. 12345, 2024.

[35] PeerJ Computer Science, "Advanced deep learning and transfer learning approaches for breast cancer classification," *PeerJ Computer Science*, vol. 11, art. no. e2951, 2024.
[36] P. Jain et al., "Parametric optimization and comparative study of machine learning and deep learning algorithms for breast cancer diagnosis," *Breast Disease*, vol. 43, no. 1, pp. 257–270, 2024.

