Fırat Üniversitesi Deneysel ve Hesaplamalı Mühendislik Dergisi

FUJECE

Doğruluğun Ötesinde: FNA Sitolojik Özellikleri Üzerinde Konformal Tahmin, Karşıt-Olgusal Açıklamalar ve Çok Görevli Öğrenmeyi Entegre Eden Klinik Odaklı Bir Meme Kanseri Tanı Çerçevesi

Esra GÜNGÖR ULUTAŞ[1], Esra YÜCE[2*], Enes Eren SÜZGEN[3], Muhammet Emin ŞAHİN[4], Mücella ÖZBAY KARAKUŞ[5]

[1]Bilgisayar Programcılığı Bölümü, Sorgun Meslek Yüksekokulu, Yozgat Bozok Üniversitesi, Yozgat, Türkiye.
[2,3]Bilgisayar Mühendisliği Bölümü, Mühendislik ve Mimarlık Fakültesi, Yozgat Bozok Üniversitesi, Yozgat, Türkiye.
[4,5]Bilgisayar Mühendisliği Bölümü, Mühendislik ve Mimarlık Fakültesi, İzmir Bakırçay Üniversitesi, İzmir, Türkiye.
[1]esra.ulutas@bozok.edu.tr, [2]esra.yuce@bozok.edu.tr, [3]e.eren.suzgen@bozok.edu.tr, [4]emin.sahin@bakircay.edu.tr, [5]mucella.ozbaykarakus@bakircay.edu.tr

Geliş Tarihi: 29.04.2026
Kabul Tarihi: 4.06.2026
Düzeltme Tarihi: 19.05.2026
doi: https://doi.org/10.62520/fujece.1940340
Araştırma Makalesi

Alıntı: E.G. Ulutaş, E. Yüce, E.E. Süzgen, M.E. Şahin ve M.O. Karakuş, "Doğruluğun ötesinde: FNA sitolojik özellikleri üzerinde konformal tahmin, karşıt-olgusal açıklamalar ve çok görevli öğrenmeyi entegre eden klinik odaklı bir meme kanseri tanı çerçevesi", Fırat Üni. Deny. Müh. Derg., vol. 5, no 2, pp. 552-573, Haziran 2026.

Öz

Meme kanseri küresel ölçekte ciddi bir sağlık tehdidi olmaya devam etmektedir ve literatürde sıklıkla bildirilen aşırı iyimser performans iddialarının ötesine geçen, tanısal açıdan güvenilir araştırma çerçevelerine açık bir ihtiyaç vardır. Bu çalışma, kıyaslama maksimizasyonunun ötesinde klinik doğrulamayı teşvik etmek amacıyla ince iğne aspirasyon sitolojisini kullanan bir meme kanseri klinik tanı metodolojisi ortaya koymaktadır. Sunulan metodoloji, yalnızca eğitim katlamalarıyla sınırlı tutulan ön işleme adımlarıyla tamamen sızıntısız, iç içe geçmiş 5x5 çapraz doğrulama prosedürü geliştirmek için Wisconsin Tanısal Meme Kanseri ve Wisconsin Prognostik Meme Kanseri veri setlerini kullanmaktadır. Orijinal 30 sitoloji özelliği, biyolojik temelli özellik mühendisliği yoluyla 42'ye çıkarılmış ve klinik olarak yorumlanabilir olasılık tahminleri sunmak amacıyla yığılmış bir topluluk yöntemi eğitilmiştir. Yöntemde; önyükleme tabanlı özellik kararlılığı, belirsizlik güdümlü triyaj gerçekleştirmek için bölünmüş konformal tahmin, özellik atıf ve karşıt olgusal açıklama modülleri, klinik doğrulama için karar eğrisi analizi ve çapraz veri seti aktarım analizi için çok görevli öğrenme kullanılmıştır. Bunlara ek olarak, yöntem 0.9932 katlama dışı eğri altı alan ve 0.9912 tutma doğruluğu değerleri elde etmiştir. %95 güven aralığında bölünmüş konformal tahmin, örneklerin %99,1'i için kesin tek sınıf tahmini üretmiş; bu da belirsiz vakaların yalnızca ihmal edilebilir bir oranı oluşturduğunu ve klinik pratikte yalnızca bir yükseltme gerektirdiğini ortaya koymaktadır. Karar eğrisi analizinin sonuçları, klinik açıdan ilgili tüm olasılık eşiklerinde sürekli olarak net klinik yarar bulunduğunu göstermiştir. Son olarak, çok görevli analiz, sitolojik morfolojinin meme kanseri tanısı açısından yararlı olduğunu, ancak prognoz açısından bu niteliği taşımadığını ortaya koymuştur.

Anahtar kelimeler: Meme kanseri, Konformal tahmin. Karşıt-olgusal açıklamalar, İstifleme topluluğu, Karar eğrisi analizi

*Yazışılan yazar

İntihal Kontrol: Evet – Turnitin
Şikayet: fujece@firat.edu.tr

Firat University Journal of Experimental and Computational Engineering

Beyond Accuracy: A Clinically-Oriented Breast Cancer Diagnostic Framework Integrating Conformal Prediction, Counterfactual Explanations, and Multi-Task Learning on FNA Cytological Features

Esra GÜNGÖR ULUTAŞ[1], Esra YÜCE[2*], Enes Eren SÜZGEN[3], Muhammet Emin ŞAHİN[4], Mücella ÖZBAY KARAKUŞ[5]

[1]Department of Computer Programming, Sorgun Vocational School, Yozgat Bozok University, Yozgat, Türkiye.
[2,3]Department of Computer Engineering, Faculty of Engineering and Architecture, Yozgat Bozok University, Yozgat, Türkiye.
[4,5]Department of Computer Engineering, Faculty of Engineering and Architecture, İzmir Bakırçay University, İzmir, Türkiye.
[1]esra.ulutas@bozok.edu.tr, [2]esra.yuce@bozok.edu.tr, [3]e.eren.suzgen@bozok.edu.tr, [4]emin.sahin@bakircay.edu.tr, [5]mucella.ozbaykarakus@bakircay.edu.tr

Received: 29.04.2026
Revision: 19.05.2026
Accepted: 4.06.2026
doi: https://doi.org/10.62520/fujece.1940340
Research Article

Citation: E.G. Ulutas, E. Yuce, E.E. Suzgen, M.E. Sahin and M.O. Karakus, "Beyond accuracy: a clinically-oriented breast cancer diagnostic framework integrating conformal prediction, counterfactual explanations, and multi-task learning on FNA cytological features", Firat Univ. Jour. of Exper. and Comp. Eng., vol. 5, no. 2, pp. 552-573, June 2026.

Abstract

Breast cancer remains a serious global health threat, and there is a clear need for diagnostically reliable research frameworks that go beyond the overly optimistic performance claims often reported in the literature. This study provides a clinical diagnostic methodology for breast cancer using fine-needle aspiration cytology, which is intended to promote clinical validation beyond benchmark maximization. The methodology presented uses Wisconsin Diagnostic Breast Cancer and Wisconsin Prognostic Breast Cancer datasets to develop a completely leak-proof nested 5x5 cross-validation procedure where any preprocessing was limited to training folds only. The original 30 cytology features were increased to 42 by adding biologically-motivated feature engineering, and a stacked ensemble method was trained to provide clinically interpretable probability estimates. The method used bootstrap-based feature stability, split conformal prediction to perform uncertainty-driven triage, feature attribution and counterfactual explanation modules, decision curve analysis for clinical validation, and multi-task learning for cross-dataset transfer analysis. In addition, the method yielded an out-of-fold area under the curve of 0.9932 and a holdout accuracy of 0.9912. Given 95% confidence interval, split conformal prediction resulted in a definitive single-class prediction for 99.1% of instances, suggesting that uncertain cases constitute only a negligible number and require only an escalation in clinical practice. The results of decision curve analysis have shown that there is always net clinical benefit across all clinically relevant probability thresholds. Finally, multi-task analysis has shown that cytological morphology is useful for the purposes of breast cancer diagnosis but not prognosis.

Keywords: Breast cancer, Conformal prediction, Counterfactual explanations, Stacking ensemble, Decision curve analysis

*Corresponding author

Plagiarism Checks: Yes – Turnitin
Complaints: fujece@firat.edu.tr

1. Introduction

Breast cancer is the most prevalent malignancy in women, responsible for approximately 2.3 million new cases and 685,000 mortalities annually [1]. The primary clinical challenge lies in reliably differentiating benign from malignant breast masses at early stages, as unnecessary biopsies impose psychological and economic burdens on patients, while delayed malignant diagnoses can lead to serious consequences including loss of life. Fine needle aspiration (FNA) cytology holds a distinctive role in this process, being widely preferred for its minimally invasive nature, brevity, and low cost. The cytological specimens obtained yield rich morphological information including nuclear size, shape, arrangement, and border characteristics enabling experienced cytopathologists to classify masses with high accuracy [2]. In this context, the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [3], comprising 569 FNA samples with 30 computationally derived nuclear features, has become a standard benchmark for machine learning (ML)-based breast cancer classification research. Recent breast cancer artificial intelligence research has expanded beyond simple lesion detection to encompass benign–malignant classification, prognosis, and explainability across imaging, genomic, and structured clinical data [4]. Classical machine learning algorithms such as support vector machines, random forests, k-nearest neighbours, logistic regression, and gradient boosting have remained widely used, particularly for structured and omics-style datasets, whereas deep learning architectures, especially convolutional neural network-based models, now dominate mammography, ultrasound, MRI, and histopathology workflows [4, 5, 6]. More recently, hybrid and ensemble strategies, together with multimodal fusion, have been increasingly adopted to improve robustness and predictive performance, while explainable AI techniques such as SHAP, LIME, and Grad-CAM have been introduced to address interpretability and clinician trust [4, 7, 8]. Thakur et al. primarily focused on imaging-based ML/DL for detection and classification, but provided limited coverage of prognosis and explainability [5]. Hussain et al. [6] reviewed deep learning, radiomics, and radiogenomics in digital breast tomosynthesis, yet their synthesis remained modality-specific and did not offer a broader comparison across diagnostic and prognostic tasks. Sudarsa and Reddy [7], as well as Charu and Gupta [8], extended the scope to multimodal, ensemble, and interpretable approaches; however, these reviews were either restricted to more recent studies or did not fully integrate foundational and clinically translated evidence across tasks [7], [8]. Likewise, MRI-oriented meta-analyses by Zhang et al. [9] and Abdullah et al. [10] provided useful pooled performance estimates, but remained confined to a single modality and reported substantial heterogeneity across studies. Silveira et al. [11] concentrated on recurrence prediction and multimodal prognosis, whereas Ghavidel and Pazos [12] emphasized class imbalance and evaluation issues in clinical datasets; although both reviews are informative, neither provides a unified account of detection, classification, prognosis, explainability, and clinical integration within a single framework.

Despite this rapid progress, the review literature remains fragmented. Recent syntheses have also identified broader structural weaknesses in breast cancer AI research. Majidpour and Beitollahi [13] further showed that many ML/DL studies rely heavily on small public datasets, exhibit class imbalance, and lack multi-centre validation, while Majidpour et al. [14] noted that even when GAN-based augmentation improves internal metrics, standardized clinical validation and cross-domain robustness often remain limited. Within the specific context of WDBC-based FNA classification, the literature is even narrower. The WDBC dataset has been the subject of extensive study, with nearly all fundamental machine learning paradigms being applied, including support vector machines (SVM) [15], decision trees, artificial neural networks, k-nearest neighbour methods, and ensemble learning approaches [16, 17]. As discussed in the current manuscript, a substantial portion of this literature is vulnerable to data leakage because in a substantial proportion of the studies published on the WDBC dataset, preprocessing steps such as normalisation, feature selection, and outlier removal are applied collectively to the entire dataset prior to its division into training and test sets [18]. Moreover, the dominant focus on discrimination metrics leaves three clinically important dimensions underdeveloped: uncertainty quantification [19, 20], actionable explanation beyond standard feature attribution [21], and direct evaluation of clinical net benefit through decision curve analysis [22]. In this regard, the present study differs from prior WDBC studies by integrating an unbiased performance estimation, uncertainty handling, explanation, and clinical utility assessment within a single leak-free framework rather than treating them as isolated methodological add-ons. Beyond methodological validity, the extant literature exhibits three key shortcomings in terms of clinical utility. Firstly, the vast majority of studies produce only a single classification prediction, without quantifying uncertainty. Nevertheless, even

in instances where a model assigns an identical 'malignant' label to two distinct patients, the underlying confidence level of these decisions may vary. To illustrate this point, consider the calculation of a 97% probability of malignancy. This indicates that the model arrived at this decision based on very strong evidence. Conversely, a 54% probability signifies that malignancy is only marginally more likely than benignity, thereby classifying the decision as borderline and uncertain. However, when the standard 0.5 decision threshold is employed, both patients are labelled "malignant" in the same way. Consequently, the exclusive reliance on the final class label can compromise the discernment of clinically imperative information concerning the reliability of the decision-making process. In clinical practice, physicians are required to consider not only the result of a diagnostic test but also the reliability of that result. Consequently, machine learning models must present diagnostic uncertainty in a systematic, transparent, and interpretable manner. Olsson et al. [23], in their work on AI-assisted prostate pathology, showed that conformal prediction can reliably flag unreliable model outputs and substantially reduce diagnostic errors on out-of-distribution data, supporting the argument that the absence of trustworthy uncertainty quantification remains a critical barrier to the safe clinical adoption of AI models.

With regard to the quantification of uncertainty, the theoretical underpinnings of this requirement are provided by the method of conformal prediction [19]. Conformal prediction is a distribution-independent framework that provides finite sample coverage guarantees under the assumption of exchangeability. This assumption is considered a reasonable and weak condition for the vast majority of clinical datasets. In contrast to Bayesian approaches or ensemble methods, it does not necessitate alterations to the model architecture and offers substantiated assurances irrespective of the data distribution [19]. As asserted by Angelopoulos and Bates [20], a comprehensive and accessible review of this field has been provided. However, to the best of our knowledge, no study has yet been published that combines conformal prediction with clinical interpretation in breast cancer classification using tabular FNA data. Sreenivasan et al. [24], in a recent study on multiple sclerosis, demonstrated that combining conformal prediction with explainable AI can yield individualized diagnostic uncertainty and disease-course estimation; yet an analogous integration for tabular FNA-based breast cancer diagnosis remains unreported. The second major shortcoming pertains to the interpretability of the model. Current interpretability methods, notably SHAP [21] and analogous approaches, which are extensively employed in table-based machine learning applications in medicine, furnish clinicians with an answer to the question, "Which features contributed to this prediction?" While the information is valuable, it is not sufficient on its own for clinical decision support. This is due to the fact that, in clinical practice, the question that is frequently more direct and actionable is: 'What would need to change for this prediction to be different? Counterfactual explanations [25], particularly through the DiCE framework [26], address this need precisely. This distinction is of significant practical importance: SHAP may demonstrate to a cytopathologist that a high concavity value resulted in a malignant prediction; however, a counterfactual explanation can be provided to indicate approximately how much lower the concavity value would need to be for the prediction to shift towards a benign outcome. This second type of information is far more applicable, particularly in situations where the quality of the FNA sample is in question or when the clinical rationale behind the decision threshold needs to be understood. Thirdly, the direct assessment of clinical benefit has been largely neglected in the current WDBC literature. Decision Curve Analysis (DCA) [22] was developed to address the limitations of discriminatory power measures, such as the Area Under the Curve (AUC) or accuracy, in evaluating clinical decision models. DCA employs quantitative methodologies to calculate the net clinical benefit, accounting for the asymmetric costs of false positives and false negatives across different threshold probabilities. This is relative to default strategies, such as treating all patients or treating none. Despite the methodological solidity of DCA as a standard in clinical prediction modelling, it has been observed that its utilisation within machine learning-focused breast cancer diagnostic studies remains limited. Indeed, to the best of our knowledge, no existing WDBC study has directly demonstrated clinical benefit using DCA. Chen et al. [27], in their multicenter study on breast cancer liver metastases, employed DCA together with SHAP for individualized prognosis prediction; however, such integrations have remained confined to prognostic settings and have not addressed the joint requirements of leak-free diagnostic evaluation, distribution-free uncertainty quantification, and actionable counterfactual explanation within a single framework.

The present study addresses the three aforementioned key gaps; namely, leak-free methodology, quantification of uncertainty, and clinically applicable explainability and benefit assessment within a single

integrated framework. In this particular context, the fundamental principle that has been adopted is as follows: The value of a diagnostic artificial intelligence system should not be evaluated solely based on whether it marginally exceeds the comparative accuracy of previous studies. Rather, its value should be assessed based on the extent to which it contributes to clinicians making more accurate, safer, and more justifiable decisions on a patient-by-patient basis. The framework presented has been devised in accordance with this principle. In this context, the study's original contributions can be summarised as follows:

- The proposed methodology is a fully leak-free, nested 5x5-fold cross-validation pipeline. In this pipeline, all preprocessing steps are learned solely within the training folds. This approach provides bias-free performance estimates.
- The calculation of a Bootstrap Feature Stability Index over 500 iterations has demonstrated that the features driving the classification process stem from repeatable selection patterns, rather than random fluctuations.
- Split Conformal Prediction is a methodology that utilises mathematical guarantees to ensure the coverage probabilities of a given set of data. This methodology facilitates the systematic identification of cases that are uncertain and require a second opinion or additional diagnostic review.
- DiCE provides counterfactual explanations that address the question, "Which cellular features would need to be different for this prediction to change?" This provides a more actionable level of explainability for cytopathology practice.
- It is evident that the decision curve analysis is supported by bootstrap confidence intervals, thus providing the necessary evidence of clinical benefit prior to institutional adoption.

The present study proposes a methodology for the quantification of cross-dataset drift between the WDBC (diagnosis) and WPBC (prognosis) datasets, employing a multi-task learning approach.

2. Materials and Methods

2.1. Dataset

In this study, two separate breast cancer datasets; the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [3] and the Wisconsin Prognostic Breast Cancer (WPBC) dataset [28], which are publicly available under a CC BY 4.0 license in the UCI Machine Learning Repository, were used together. Both datasets are based on core morphology features derived from fine needle aspiration (FNA) images and are anonymized retrospective data collections that do not require additional ethics committee approval for secondary analysis. The datasets used in this study do not contain any individual patient identification information. All analyses were conducted at the de-identified case level or in aggregated form. The purpose of the developed framework is not to replace cytopathological evaluation but to provide decision support to clinicians and pathologists. The WDBC dataset [3] comprises 569 samples obtained from patients who presented to the University of Wisconsin Hospitals with complaints of a breast mass. Each sample was derived from a digitized FNA image. A bespoke software system was developed for the purpose of extracting ten fundamental morphological measurements related to the cell nucleus from each image. The measurements included radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The following measurements are of significance in this context: the mean distance from the centre to the perimeter points, the standard deviation of the grey-level values, the perimeter length, the area, the local variation in radius, the perimeter cubed to area minus one ratio, the severity of concave regions in the contour, the number of concave contour segments, symmetry, and the fractal dimension. For each of these ten fundamental features, three separate statistics were calculated: the mean value, the standard error (SE), and the "worst" value. In this context, the "worst" value is defined as the average of the three largest measurements among the kernels in the image. Consequently, a total of 30 continuous features were obtained. The diagnostic output is binary, and the dataset contains 357 benign (62.7%) and 212 malignant (37.3%) samples. From a clinical perspective, the WDBC dataset provides a computationally quantified representation of the cellular information that a cytopathologist visually assesses during an FNA examination, based on the same images. It is notable that the "worst" features reflect the most markedly

abnormal nuclei in the sample, and correspond to the cellular areas that would first capture the cytopathologist's attention. The WPBC dataset [28] encompasses 198 patients with a confirmed diagnosis of invasive breast cancer who underwent surgical resection. The dataset under consideration shares the same 30 FNA core morphology features as the WDBC dataset and additionally includes two clinical variables: tumor diameter (cm) and the number of positive axillary lymph nodes detected during surgery. It is noteworthy that the lymph node variable is absent from four cases. The output variable in the WPBC dataset is defined as the occurrence of disease recurrence during follow-up. In this particular context, the dataset under consideration consists of 151 non-recurrent cases, accounting for 76.3% of the total, and 47 recurrent cases, which constitute 23.7% of the total.

This clinical distinction between the two datasets is of particular importance. The WDBC dataset encompasses both benign and malignant cases, and the objective is to classify the lesion as benign or malignant. In contrast, all patients in the WPBC dataset have already received a malignant diagnosis; therefore, the problem in this dataset focuses not on the question "Is this lesion malignant?" but rather on "Will this disease, already known to be malignant, recur?" Consequently, the two datasets diverge significantly not only in terms of the output variable but also in terms of the clinical decision-making context. This discrepancy is of critical importance, particularly when interpreting cross-dataset comparisons and transferability analyses.

2.2. Feature engineering and leak-free model development

The number of raw FNA features has been augmented from 30 to 42 by the incorporation of 12 biologically meaningful derived features. In this study, the area-to-perimeter ratio (a measure of nuclear circularity; with more circular nuclei tending to approach higher values), the concavity-to-compactness ratio (an index of nuclear shape irregularity), the product of the mean radius and the mean texture (to capture the combined effect of nuclear size and textural heterogeneity), worst/mean ratios for radius, area, and concavity (to quantify how far the most abnormal nuclei deviate from the sample mean), and quadratic terms for the four raw features with the highest discriminative power, were identified as the derived variables. It is imperative to note that all transformations were applied on a per-sample basis; no cross-sample statistics were utilised, thereby ensuring the complete elimination of the risk of data leakage at this stage.

The clinical rationale underpinning this approach is as follows: malignant nuclei are characterised by two key features. Firstly, they are typically larger and more irregular in shape. Secondly, greater variability between the least and most severely affected cells within a single specimen is exhibited. The worst/mean ratios have been demonstrated to quantify this heterogeneity, a fact which experienced cytopathologists have long intuited when assessing the degree of nuclear pleomorphism in a smear.

The methodological core of this study is the leak-free nested cross-validation (nested CV) approach. In the outer loop, 5-fold stratified cross-validation was employed to estimate the true generalization performance. In each instance of the outer fold, all preprocessing steps were exclusively learned in the training set and applied to the validation set solely as transformations. These steps included median imputation, normalisation with StandardScaler, and feature selection. The 5-fold configuration within the inner loop was responsible for the training of base classifiers within the StackingClassifier. Feature selection was performed using a combination of Mutual Information and ANOVA F-statistic criteria. The top 20 features were selected for each method, yielding approximately 23 features per fold.

The significance of this leak-free design must be emphasised. In a standard pipeline that exhibits leakage, the feature distributions of the test set can exert an influence on the selection of variables and the scaling of these variables. In this scenario, the model acquires indirect insight into the test data prior to its observation. In the proposed pipeline, however, the test set remains truly unseen at every stage. The trade-off may be characterised by a marginal decline in reported performance; nevertheless, the advantage lies in a performance prediction that exhibits enhanced reliability in terms of generalisation to new patients.

The prediction model adopted a stacking-based ensemble structure. In order to maximise the diversity of the resulting decisions, five distinct base learners were selected: The algorithms employed included XGBoost

(gradient boosting on decision trees), LightGBM (leaf-based gradient boosting), SVM with a radial basis function (RBF) kernel (kernel-based decision boundary), Random Forest (bagged trees) and Logistic Regression with L2 regularisation (linear classifier). It is evident that each of these models captures a distinct aspect of the feature space. To illustrate this point, consider the potential for error mitigation through the utilisation of the kernel-based boundary geometry of SVM in cases where errors arise from the tree structure of XGBoost. Similarly, the identification of errors made by the linear model in non-linear regions can be facilitated by tree-based learners. The out-of-fold probability outputs of these base models were combined using a meta-learner, specifically a logistic regression model. This enabled the meta-learner to assign a weight to each model according to its reliability, based on out-of-fold predictions that were not subject to target leakage. Probability calibration was also applied to ensure that the model output probabilities were clinically interpretable. The outputs of ensemble models are frequently characterised by an excess or paucity of reliability with regard to the raw probability. To illustrate this point, consider a model that outputs a probability of 0.90; at this threshold, the model may only operate with 75% accuracy [29]. In the medical field, such calibration errors assume particular significance due to the prevalence of probability-based threshold-based decisions regarding further investigations or treatments. Consequently, Platt scaling (sigmoid calibration) was implemented on a dedicated calibration set (n = 91), which was isolated from the overall training process. Following the calibration process, the model's output of "P(Malign)=0.85" transitions from being a relative score to a true probability that can be interpreted by the clinician.

In order to evaluate the sensitivity of the sampling process to feature selection, a Feature Stability Index was calculated based on bootstrapping with 500 iterations. For each iteration, the complete pre-processing and feature selection pipeline was implemented on the bootstrap sample created using substitution sampling. The stability score of each feature was defined as the proportion of iterations in which it was selected. Features with a high stability score (i.e. ≥0.80) were considered reliably selected variables, regardless of sample variation. Conversely, features with a low stability score (<0.50) were not emphasised as strong predictors in clinical discussions.

2.3. Quantification of uncertainty, explainability, and clinical benefit analysis

The quantification of uncertainty was achieved by employing a split conformal prediction approach. Whilst a standard machine learning model provides a unidimensional output, such as "this sample is malignant", the conformal prediction approach offers a mathematical coverage guarantee, along with the confidence level accompanying this decision [20]. This guarantee is predicated on the assumption that patients are interchangeable, i.e. from the same population; this is a reasonable assumption for patients undergoing FNA due to breast mass. In practice, the discordance score, s_i, was calculated for each calibration patient, with s_i =1 – P (true class | features). Conversely, a higher score is indicative of a model with greater uncertainty. The threshold value, q, was determined from the calibration set, where 95% of patients satisfy the condition $s \leq q$. For a new patient, all classes with a discordance score lower than this threshold were included in the prediction set. The result set is defined as either a single-element {Benign} or {Malignant} definite set, a two-element {Benign, Malignant} set, or an uncertain or empty set that has been rejected. The pivotal observation is that when the model exclusively yields the {Malignant} outcome, it is mathematically guaranteed that this decision contains a maximum error of 5%.

The explainability analysis is composed of two complementary components: SHAP and DiCE. SHAP (SHapley Additive Explanations) [21] is an algorithm that utilises a data-driven approach to elucidate the influence of various measures on the final output, by systematically decomposing the contributions of each predicted feature. A positive SHAP value for a feature has been shown to shift the prediction towards the malignant class, while a negative value has been shown to shift it towards the benign class. The magnitude of the value is said to indicate the intensity of the contribution. In this study, both global SHAP, which provides an overall importance ranking by averaging across all test samples, and local SHAP analysis, which demonstrates the decision logic of individual cases, were calculated.

Conversely, DiCE (Diverse Counterfactual Explanations) [26] is an approach that generates counterfactual explanations by identifying the minimal feature alterations that would result in a change to the model prediction. For patients predicted as malignant, the lowest level of feature changes that would lead to a shift

to the benign class was sought. Three distinct counterfactual scenarios were formulated; the term "different" refers to alternative clinical scenarios that achieve the same class change with different combinations of features. Decision Curve Analysis (DCA) was applied to evaluate the true value of the model in terms of clinical decision support [22]. In the event of a clinician performing biopsies on all patients, all malignancies are detected; however, many patients with benign conditions are subjected to unnecessary procedures. The decision to refrain from performing biopsies has the advantage of avoiding unnecessary procedures; however, it does have the disadvantage of resulting in the potential loss of malignancies. For a predictive model to be considered clinically useful, it must be found that the model does not excessively increase the number of missed malignancies while reducing unnecessary procedures, within the clinician's tolerance for error types. The DCA formalises this balance through net clinical benefit. The net benefit value for the threshold probability t is calculated as follows:

$$\text{Net Benefit} = \frac{TP}{N} - \frac{FP}{N} \cdot \frac{t}{1-t} \tag{1}$$

The minimum malignancy probability required to make the biopsy decision, i.e. the prior threshold accepted by the clinician, is denoted by t. The determination of the confidence intervals was achieved through the execution of 300 bootstrap iterations.

2.4. Multi-Task Learning and Cross-Validation

A multifaceted approach involving multitasking, learning, and cross-dataset validation analysis was employed to investigate the shared information structure between diagnostic and prognostic tasks. A common feature converter structure was established using both WDBC and WPBC data; this structure, consisting of an inputter, StandardScaler, and feature picker, produced a 21-dimensional common FNA representation. Subsequently, two task-specific classifiers were trained: The initial task is concerned with the differentiation of benign from malignant cases on WDBC. The subsequent task is concerned with the differentiation of non-recurrent from recurrent cases on WPBC. In light of the 3.2:1 class imbalance that characterises the WPBC, class-weighted learning was implemented for the second task.

The common converter, by requiring both tasks to operate within the same feature space, reveals the extent to which the information necessary for prognosis is shared with the features that guide the diagnosis. In the subsequent stage, the pipeline that had been trained on WDBC was applied directly to the WPBC data, thus quantifying the cross-dataset space shift.

3. Experimental Results

3.1. Dataset Characteristics

In this dataset, there are 569 instances of patients with a ratio of 62.7% benign versus 37.3% malignant masses (Figure 1). This is a representation of the actual proportion of benign and malignant masses among all the patients diagnosed using the technique of fine needle aspiration. A naive baseline model that always predicts a benign mass would inherently achieve an accuracy of 62.7%. This brings out one limitation of the use of accuracy scores alone to determine the effectiveness of a machine-learning algorithm. Looking at the distribution of features as represented by Figure 1 (right), it is evident that malignant masses have much higher radius, perimeter, and area. They also tend to be more irregular (concavity, concave points) compared to benign masses. As depicted by Figure 1, there is an overlapping of the distributions at intermediate values. It is in this overlapping range that there exists misclassification.

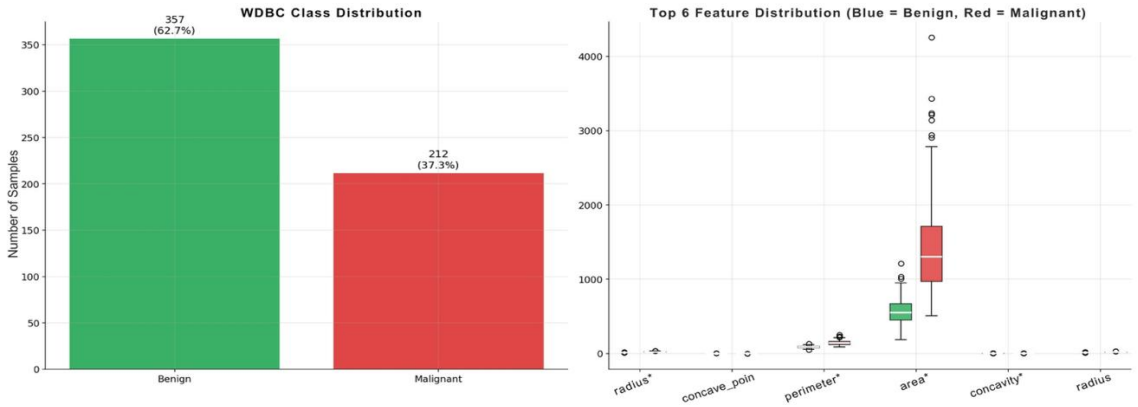

Figure 1. WDBC class distribution (left) and feature distributions for top six discriminative FNA measurements by diagnosis (right)

3.2. Nested cross-validation: the true performance estimate

Table 1 provides a summary of fold-level performance. Figure 2 allows us to compare ROC curves based on fold and out-of-fold data. An AUC of 0.9964 indicates that the model correctly ranks a randomly selected malignant case above a randomly selected benign case in 99.64% of all possible pairings — reflecting near-perfect discriminative power. This is distinct from accuracy, which is reported separately as 0.9771 in Table 1. Sensitivity at 0.9574 suggests that 4 out of 100 malignant cases remain undetected, while specificity of 0.9888 means that less than 2 out of 100 benign cases will be incorrectly classified as malignant. The out-of-fold confusion matrix shows the true costs of this prediction: the model is expected to miss 9 out of 569 malignant cases and produce 4 extra work-ups of non-malignant patients. A critical observation for clinical readers: Fold 4 demonstrates optimal performance, with an Area Under the Curve (AUC) value of 1.000 and an accuracy of 1.000. This performance is most likely due to a statistical fluke that resulted in all borderline cases, which were harder to predict, being included in the training set for this fold. In practice, however, such a scenario is highly unlikely to occur and, hence, it does not have much clinical value. What matters is the average AUC across folds (0.9964), and, more specifically, the OOF AUC of 0.9932, produced when a classifier is applied to a new sample that was never seen before.

Table 1. Nested cross-validation results (5-fold stacking ensemble)

Fold	Accuracy	AUC	F1	Precision	Recall	Brier	Time (s)
1	0.9825	0.9993	0.9773	0.9556	1.0000	0.014	4.3
2	0.9737	0.9954	0.9647	0.9762	0.9535	0.021	7.6
3	0.9649	0.9874	0.9512	0.9749	0.9286	0.025	3.1
4	1.0000	1.0000	1.0000	1.0000	1.0000	0.006	3.0
5	0.9646	0.9997	0.9500	1.0000	0.9048	0.019	2.9
Mean±SD	0.9771±0.0147	0.9964±0.0053	0.9686±0.0208	0.9813±0.0189	0.9574±0.0425	0.017±0.007	4.18

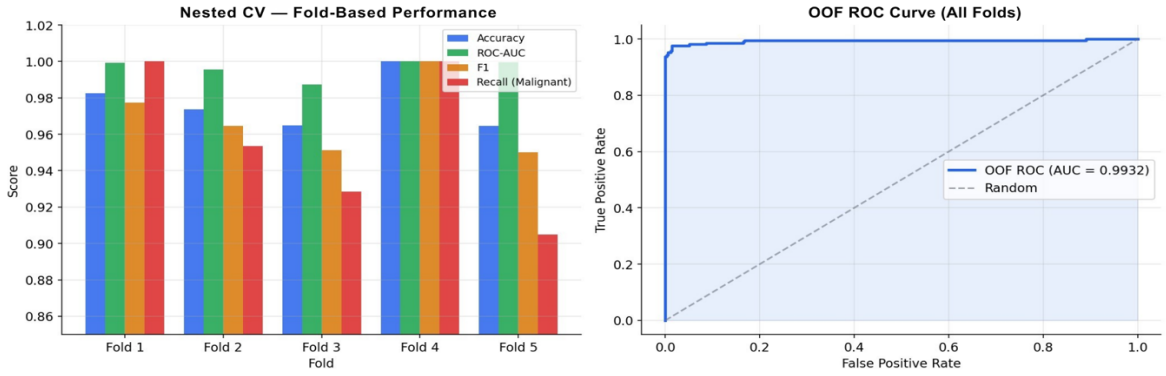

Figure 2. Nested CV fold-level performance (left) and OOF ROC curve (AUC=0.9932, right)

3.3. Bootstrap feature stability: which features can clinicians trust?

In this part, from Figure 3, it can be seen that stability scores are a good indication of the reliability of the model. Out of the 42 features, 20 scored 1.00 stability points due to their selection in each of the 500 bootstrap iterations. The clinical relevance of such results stems from two main aspects. Firstly, the effectiveness of feature engineering is proven because, besides the raw measurements, the engineered features such as radius_texture_interaction, concavity_compactness_ratio, area_perimeter_ratio, and three worst-to-mean ratios achieved the 1.00 stability point. Thus, it may be concluded that these features do not depend on any specific sample but are always indicative and consistent with the data. Should a medical practitioner doubt the validity of using the concavity_compactness_ratio, one could safely say that this feature is absolutely valid. Secondly, the stability point of the worst-case features – radius_worst, area_worst, perimeter_worst, and concave_points_worst – was equally at 1.00. This result corresponds to the cytopathology theory according to which the abnormalities contained in a tissue biopsy sample give the strongest indications about the disease. A cytopathologist studies the abnormalities in cells in the sample, and the same goes for the algorithm used by the model.

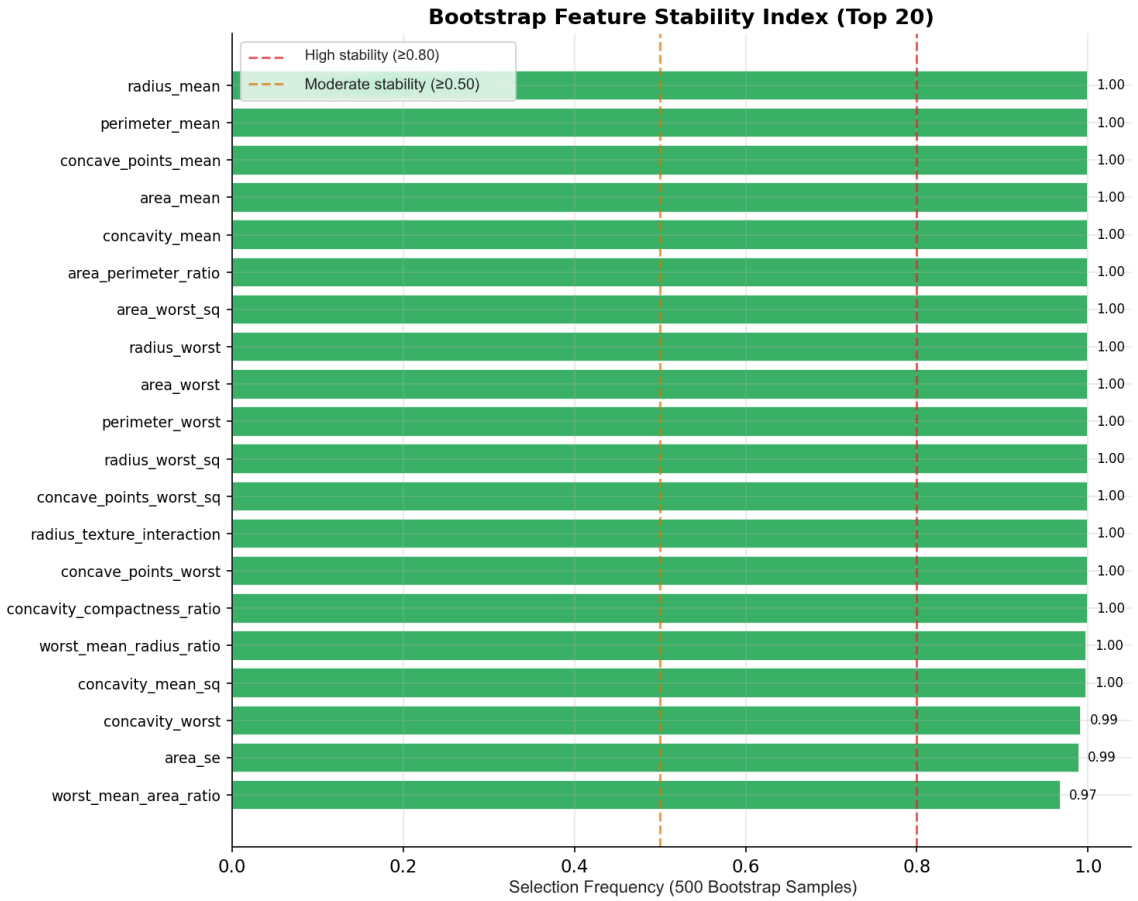

Figure 3. Bootstrap Feature Stability Index for top 20 features (500 iterations). Scores ≥0.80 indicate high stability

3.4. Probability calibration: making probabilities clinically meaningful

Calibration analysis, as seen in Figure 4, is described. In its uncalibrated form, the stacking ensemble had relatively well-calibrated probabilities (Brier Score = 0.0055). Following Platt scaling calibration on the calibration set holdout, the Brier Score became 0.0072. This slight increase may be attributed to the small size of the calibration set (n = 91). Beyond sample size, the limited flexibility of parametric calibration on small holdouts may further contribute to this slight increase. From this observation, it is safe to assume that any apparent reduction in calibration is likely due to the calibration set size as opposed to being an indicator

of actual deteriorating performance. A reliability plot (left panel) is also able to tell us more on this phenomenon: the calibrated model provides better predictions for the observed outcome probabilities, especially for values under 0–35%, which have the most significant impact on clinical decision-making. The importance of this phenomenon must be emphasized further. Assume there is a patient whose malignant cancer probability calculated using the uncalibrated model is 0.30. However, given the presence of an optimistic model toward the benign outcome, this probability might actually indicate a higher likelihood of 45%. Therefore, the doctor would make a wrong clinical decision based on such erroneous information. Calibration will allow the doctor to take this probability literally because it would mean that the probability of malignancy for a score of 0.30 is in fact about 30%. Near-perfect ROC and Precision-Recall plots (centre and right panels, respectively) prove that discriminative power has been maintained following calibration. The calibrated stacking ensemble's class-wise and overall performance on this holdout set is summarised in Table 2. Although the test set AUC = 1.000 from 114 patients cannot be generalized, AUC = 0.9932 from OOF nested cross-validation can be considered as the true value.

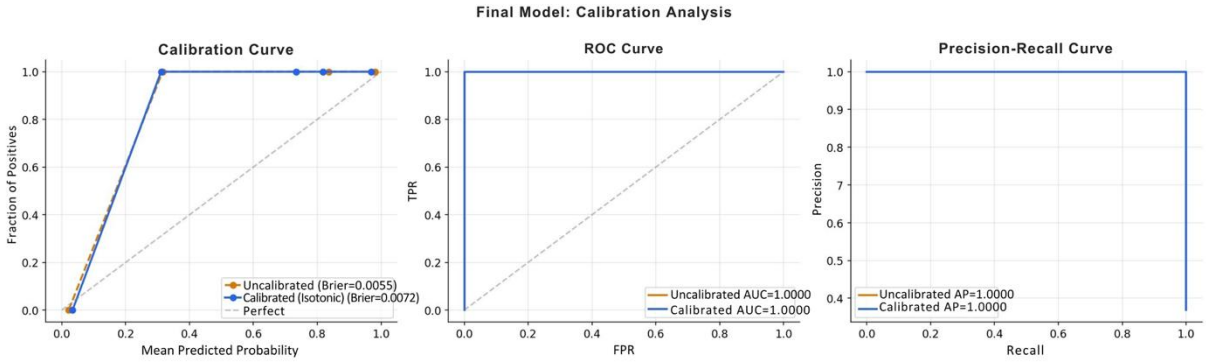

Figure 4. Calibration analysis: reliability diagram (left), ROC curves (center), and Precision-Recall curves (right) for uncalibrated and calibrated models

Table 2. Final holdout test set performance (calibrated stacking ensemble, n=114)

Metric	Benign	Malignant	Macro Avg	Weighted Avg	Overall
Precision	0.9863	1.0000	0.9932	0.9913	—
Recall	1.0000	0.9762	0.9881	0.9912	—
F1-Score	0.9931	0.9880	0.9905	0.9912	—
Support	72	42	114	114	—
Accuracy	—	—	—	—	0.9912
OOF AUC (Nested CV)	—	—	—	—	0.9932
Brier Score	—	—	—	—	0.0072

3.5. Conformal prediction: systematic uncertainty triage

The results of the implementation of the process of conformal prediction are shown in Table 3 and Figure 5. The interpretation can be given by means of a concrete clinical example. As can be seen from the table, one assumes that there are 1,000 patients who underwent fine-needle aspiration (FNA) examination each year, and the introduced framework is applied on their basis with significance level $\alpha = 0.05$. In this regard, for 1,000 cases analyzed, 991 cases (99.1%) can be expected to receive a final and unique result, namely, {Benign} or {Malignant}, while no more than 5% of these predictions can be expected to be false. In the meantime, the other approximately nine cases will receive the prediction of 'ambiguous' (benign or malignant) classification and will need to undergo additional analysis automatically. As can be seen from the triage algorithm, the procedure is systematized, transparent, mathematically supported, and does not depend on a day-to-day uncertainty of the cytopathologist. As has already been pointed out, if it is required to increase the coverage rate up to 99%, then the number of ambiguous cases will be equal to 13.2% or, equivalently, 132 cases for an institution that analyzes 1,000 FNA patients per year. The tradeoff is clear-cut, and the choice of the level α can be made by clinicians according to their preferences, while the guarantee that at most α of accepted predictions will be wrong remains valid in either case. This is fundamentally different

from the uncertainty in the existing clinical application of ML because no mathematical guarantees exist under current practice. In turn, it should be mentioned that at significance level α = 0.05, no doubleton {Benign, Malignant} cases were identified (there was only an empty set). Therefore, in this case, the model's probabilistic outputs were precise enough to obtain 95% coverage, which allowed assigning each case to only one class and thus proved its high discriminability. On the contrary, at the 99% coverage level (α = 0.01), 15 cases had to be assigned to doubleton sets, which proves the real uncertainty of the model in the situation requiring high precision.

Table 3. Split conformal prediction coverage and efficiency (Test Set, n=114)

α	Target Cov.	Empirical Cov.	Avg Set Size	Empty	Single (Certain)	Double (Uncertain)
0.01	≥0.99	1.000	1.132	0	99 (86.8%)	15 (13.2%)
0.05	≥0.95	0.991	0.991	1	113 (99.1%)	0
0.10	≥0.90	0.895	0.895	12	102 (89.5%)	0
0.15	≥0.85	0.825	0.825	20	94 (82.5%)	0
0.20	≥0.80	0.772	0.772	26	88 (77.2%)	0

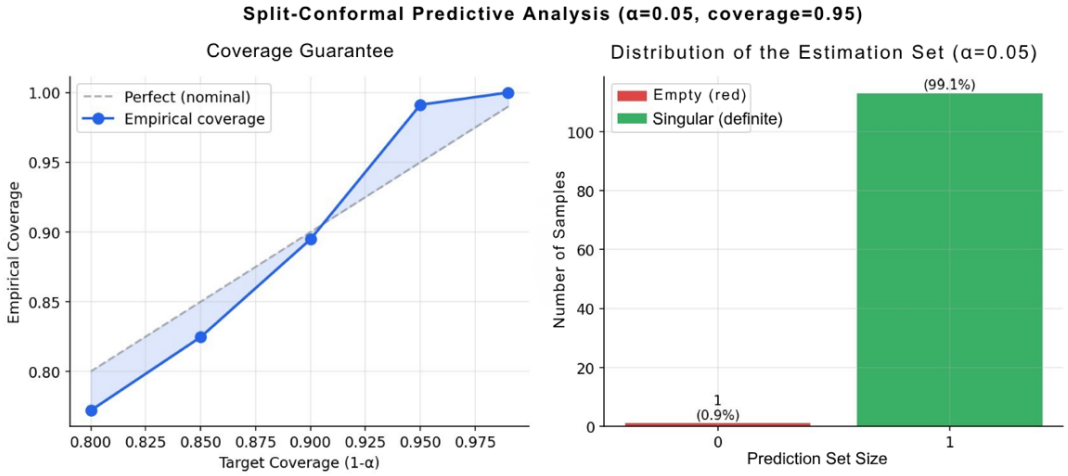

Figure 5. Conformal prediction coverage across significance levels (left) and prediction set size distribution at α=0.05 (right)

Table 3 additionally reports one empty prediction set at α=0.05. An empty set arises when the nonconformity score of every candidate class exceeds the calibration threshold $\hat{q}$ that is, the model cannot assign sufficient confidence to either Benign or Malignant for that particular instance. This is distinct from a doubleton {Benign, Malignant}: a doubleton means both classes are plausible; an empty set means neither meets the required confidence level. In clinical terms, an empty set represents the most extreme form of uncertainty flagging: the framework is explicitly stating that this case falls outside the distributional support of the calibration set and should not receive any automated classification. Such a case warrants immediate referral to expert review or additional diagnostic workup, and should never be resolved by defaulting to the higher-probability class.

3.6. SHAP global explanations: what drives the classification overall?

Figure 6 shows the global SHAP analysis on the test dataset. The most important predictor is radius_texture_interaction (mean |SHAP| = 0.64), which stands for a synthesized feature equal to the multiplication of nuclear radius and textural SD. The model's capacity to detect malignant nuclei implies its ability to imitate the diagnostic skills of cytopathology specialists. As research shows, the most reliable indicator of malignancy is the combination of a large nucleus with an irregular textural pattern, as opposed to having one large nucleus or high texture individually. Clearly, no single variable contains as much predictive power as their combination as a whole. This explains why engineered variables can sometimes perform better than the original ones. The next three important variables include worst_mean_area_ratio,

concave_points_mean, and concavity_compactness_ratio. The first one refers to nuclear pleomorphism or the size difference between the largest nuclei and the mean size across the entire smear. High pleomorphism is a sign of malignant cells. The other two stand for nuclear membrane irregularity, which manifests itself through indentations, lobulations, and abnormal contours of malignant cell nuclei. As the SHAP scatter plots (right panel) suggest, all the above variables have monotonous relationships with the outcome variable. Indeed, increased values of any of them positively influence the prediction of the disease's presence, whereas decreased values move the prediction to the opposite category. In terms of cytopathological knowledge, this means that the model is not using any spurious correlations but responding to real-life features of cells.

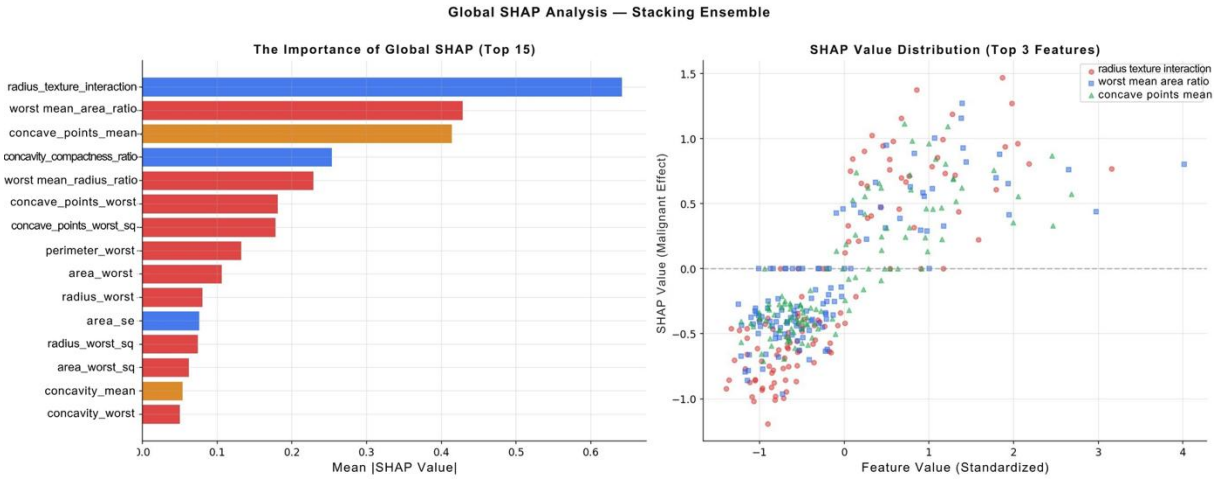

Figure 6. Global SHAP: mean absolute SHAP values for top 15 features (left) and SHAP scatter for top 3 features (right)

3.7. SHAP local explanations: case-specific reasoning for clinicians

Three SHAP local explanations have been selected from the set of cases shown in Figure 7 and reflect different aspects of the clinical spectrum of model behavior:

• The first example (red title, P = 0.974, confirmed Malignant) has ten features indicating the Malignant class. The radius_texture_interaction feature contributes to the maximum SHAP value (≈ 1.0). The clinical interpretation of this SHAP explanation is: "Atypical nucleus, presence of large chromatin heterogeneity and other morphologic features consistent with this finding. Malignant classification supported by several independent lines of cellular evidence." This is a strong clinical statement about malignancy.

• The second example (green title, P = 0.027, confirmed Benign) has all features suggesting the Benign class. Area_worst contributes most significantly to the negative direction (−0.75). The "worst" area is low, meaning that even the most atypical cells of this case have the nuclei of the expected size. The clinical statement about the second case is: "Normal nuclear size, even when considering the worst cells in this case. Low concavity and compactness. No morphologic features suggesting atypicality of this sample. Clearly benign sample."

• The third example (orange title, P = 0.311, confirmed Malignant; missed prediction with regard to the 0.5 threshold) is the most informative from a clinical perspective. The radius_texture_interaction feature suggests Malignant, whereas worst_mean_area_ratio and worst_mean_radius_ratio suggest Benign. Therefore, the outcome suggests the existence of a genuine conflict between some morphologic features of this case, which point towards opposite diagnoses. If conformal prediction is employed with $\alpha = 0.01$, the doubleton {Benign, Malignant} will be obtained, and an escalation procedure is triggered. Without conformal prediction, the model produces only the "Benign" label. Conformal prediction can be used to escalate such examples. The SHAP explanation provides very detailed information to the cytopathologist about the position of relevant morphological features. Some textural features are suspicious, while size measurements are not decisive. In this example, one should carefully re-examine the most atypical nuclei.

Figure 7. Local SHAP explanations for three representative cases: highest-confidence Malignant (top), highest-confidence Benign (middle), most uncertain case (bottom)

3.8. Counterfactual explanations: what would need to change?

As illustrated in Figure 8, the DiCE analysis for the Malignant example includes a high confidence level (P = 0.973). Three different counterfactuals are generated, all producing prediction values below 0.50, thereby indicating that the sample must be classified as Benign (CF1: P = 0.467, CF2: P = 0.424, CF3: P = 0.426).

In all three cases, the need to reduce the corresponding core features, without any augmentation, is identified: perimeter_mean (3.2 SD), radius_texture_interaction (2.9 SD), concave_points_mean (2.5 SD), concave_points_worst (2.3 SD), and concave_points_worst_sq (2.1 SD). In terms of clinical significance, there are several points. First of all, the fact that all three diverse counterfactuals converge to exactly the same feature combination suggests that these features cannot be used independently. The reduction of one feature is not sufficient to reverse the prediction; rather, a simultaneous reduction of nuclear perimeter, morphological texture, and number of concave points is needed. These features represent a certain border between Malignant and Benign categories, and their simultaneous presence forms the coherent pattern recognized by the model. Secondly, it should be noted that the magnitude of the needed decrease is clinically significant. A 2.9-3.2 SD change suggests that the sample needs to be changed substantially to fall into a new category. Therefore, it can be concluded that the malignancy prediction made by the classifier is robust and is not affected by any measurement errors. On the contrary, if the changes required by counterfactual were less pronounced, in the range of 0.5-1.0 SD, it would point to the case being close to borderline. Thirdly, as seen above, the largest counterfactual change in SD terms corresponds to perimeter_mean. Perimeter measurements are subject to numerous potential biases introduced by cell orientation, cell preparation quality, etc. In the case of getting an unfavorable counterfactual explanation, the clinician may ask: "Do these elevated perimeter values reflect real nuclear abnormalities or preparation artifacts?" This type of specific question, which cannot be elicited by SHAP, is exactly what domain expertise provides.

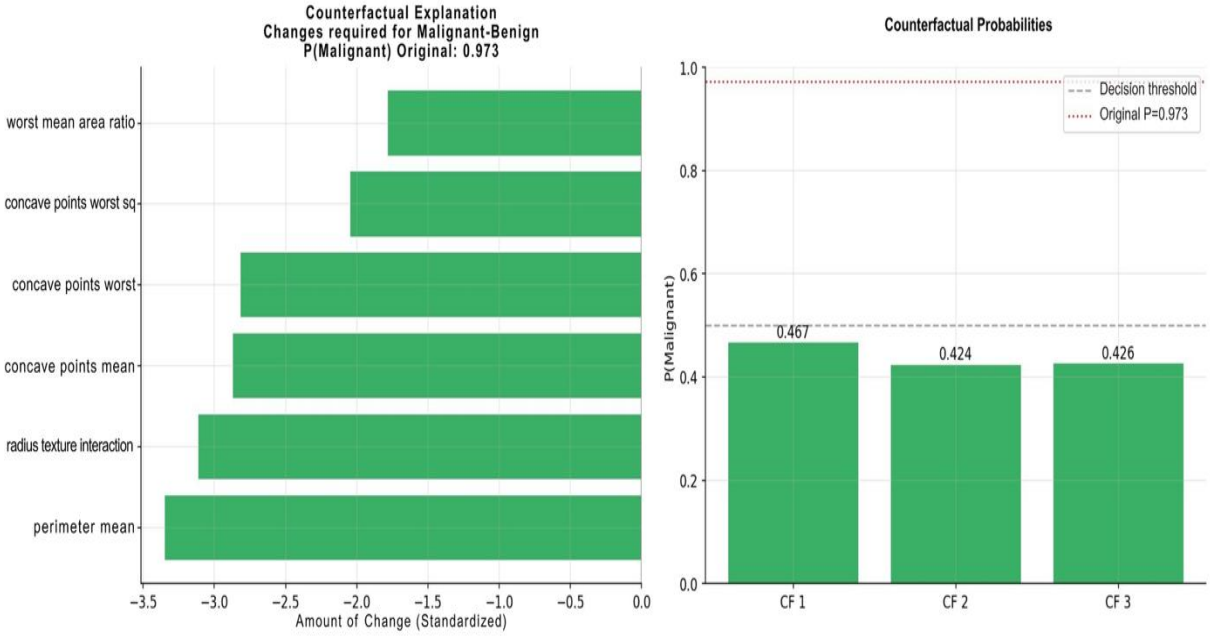

Figure 8. DiCE counterfactual explanations: required feature changes for Malignant→Benign class shift (left) and resulting prediction probabilities for three diverse counterfactuals (right)

3.9. Decision curve analysis: does this model help clinicians decide?

The results of the DCA are presented in Figure 9. In order to analyze the curves presented above, assume that the breast surgeon performs biopsies on patients having breast masses, if the probability of being malignant based on clinical evaluation is 10% or more (t = 0.10). At this point, the "treat all" option has the net benefit of around 0.09. However, it involves performing biopsies that may not be necessary as the patients might have benign tumors. The false positive penalty at t = 0.10 is relatively small (the surgeon believes that nine out of ten biopsies performed in case of low probability cases are not needed). The Net Benefit for the proposed model at t = 0.10 is about 0.37 and, thus, almost four times better. Moving forward to the threshold of 0.20, it becomes clear how the curves were produced. The "treat all" approach leads to lower net benefits due to a growing false positive penalty (the surgeon demands higher pre-biopsy probability of malignancy). As for the model, its Net Benefit at t = 0.20 stays at about 0.36, while "treat all" drops to 0.18. For thresholds

ranging from 0.30 to 0.50, the surgeon decides to perform biopsies in cases when there is a high probability that the "treat all" decision would be ineffective, while the model provides useful information throughout the whole process. From t = 0.05 to 0.50, the model always gives better performance compared to the "treat all" and 95% bootstrap confidence interval (shaded region) proves that the difference is significant. The right part of Figure 9 depicts the standardization of the Net Benefit curve for expressing it as a proportion compared to "treat all". In other words, it shows the model's advantage over the "treat all" value (represented as a value higher than 1.0). According to this graph, the model produces a net benefit well above 1.0 for t = 0.05–0.35 (peaking at 2.0 as this is the maximal achievable value due to numerical stability as the denominator equals zero). It means that, for almost every single physician operating with assumptions regarding the 5–35% malignancy probability, it could be beneficial to use the proposed model instead of universal treatment. This constitutes the data that supports the institutional decision on whether to implement the algorithm or not. Beyond discriminative accuracy alone, the analysis demonstrates that, considering clinicians' practical priors regarding malignancy probability, applying the model leads to substantially better outcomes than a treat-all strategy across the clinically relevant threshold range.

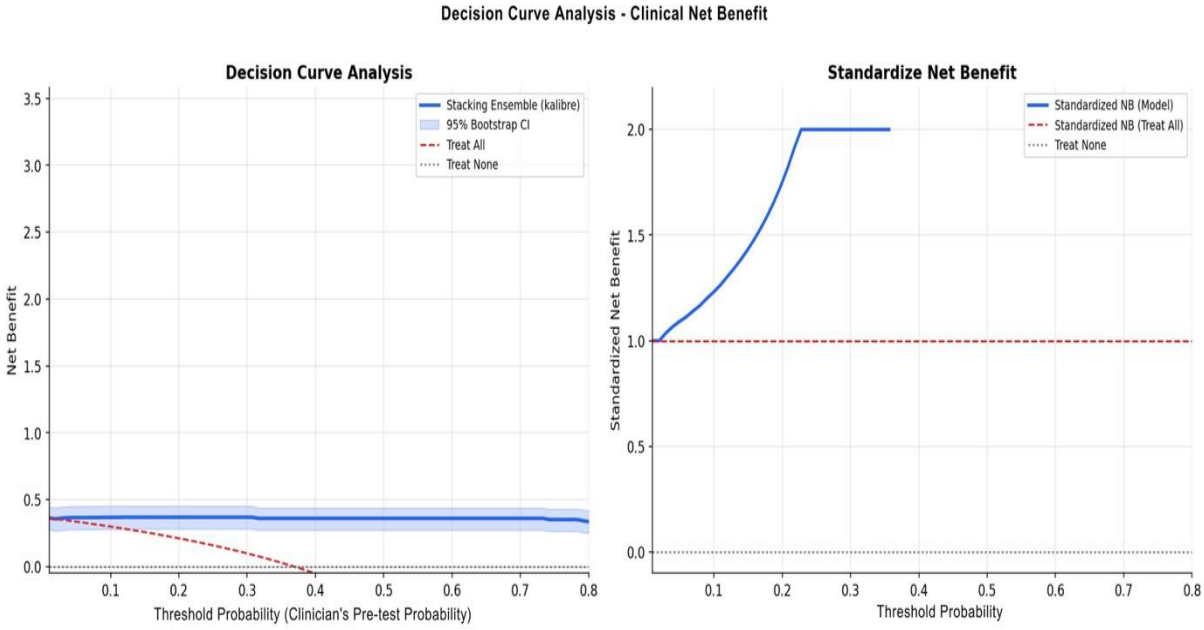

Figure 9. Decision Curve Analysis with 95% bootstrap confidence intervals (left) and Standardized Net Benefit (right)

3.10. Multi-task learning: the diagnostic-prognostic boundary

The results of applying the multi-tasking learning model are given in Figure 10. The primary WDBC classification task demonstrated an area under the receiver operating characteristic curve (AUC) of 0.9934 and an accuracy of 0.9736 when performed with a shared 21-dimensional fine needle aspiration representation. In other words, this model's results were similar to those attained with the complete stacked ensemble, suggesting that a shared representation still possesses diagnostic discriminative power regardless of task-specific attributes. For Task 2 (prognosis of breast cancer relapse), an AUC of 0.6025 and an accuracy of 0.5707 were achieved. An interpretation of the results for Task 2 should take into account the clinical meaning of these values. An AUC of 0.6025 does not demonstrate that the model failed to predict patient outcomes but rather highlights a clinically observed phenomenon. Nuclear morphology, radius, texture, concavity and all the measurements included in these datasets are highly discriminatory between benign and malignant tumors at FNA diagnosis stage. By contrast, relapse probabilities are defined by molecular markers (receptor expression, HER2 amplification, genomic instability scores) and lymph node metastases, among other parameters, including tumor biology, which cannot be assessed through nuclear morphology. This suggests that an AUC of 0.6025 is a manifestation of the feature set's weak prognostic potential. More research is needed to clarify if atypia plays a role in recurrence prediction. In conclusion, this numerical

evaluation carries considerable practical relevance. In particular, the text explains oncologists that although nuclear morphology is useful for the detection of malignancies in FNA tests, it cannot be utilized as an instrument for prognosis staging. Molecular profiling methods (e.g. Oncotype DX, MammaPrint) and lymph node assessment are not interchangeable with information provided by nuclear morphology assessments.

Two complementary factors plausibly underlie this negative finding. On one hand, the modest size of the WPBC cohort (n = 198, with only 47 recurrent cases) constrains statistical power and is consistent with broader concerns about evaluation reliability in small clinical datasets [12]. On the other hand, recent recurrence-prediction frameworks that achieve clinically meaningful AUROC values typically rely either on whole-slide histopathology coupled with molecular targets such as Oncotype DX [30], or on integrated clinical and genomic features, rather than on FNA-derived nuclear features alone. Read together, these observations suggest that the limited prognostic performance observed here is not merely an artefact of cohort size but also reflects an intrinsic limitation of FNA-based nuclear morphology as a prognostic substrate an interpretation that is, in our view, the more clinically informative reading of this negative result. It should be noted that the reported accuracy of 0.5707 falls below the majority-class baseline of 0.763 (obtained by predicting all cases as Non-recurrent). This is a direct consequence of class-weighted training (class_weight='balanced'), which deliberately shifts the decision threshold toward the minority Recurrent class to improve recall at the cost of raw accuracy. In this setting, accuracy is therefore a misleading metric; AUC=0.6025 which is threshold-independent remains the appropriate measure of discriminative performance and confirms modest but above-chance prognostic capacity.

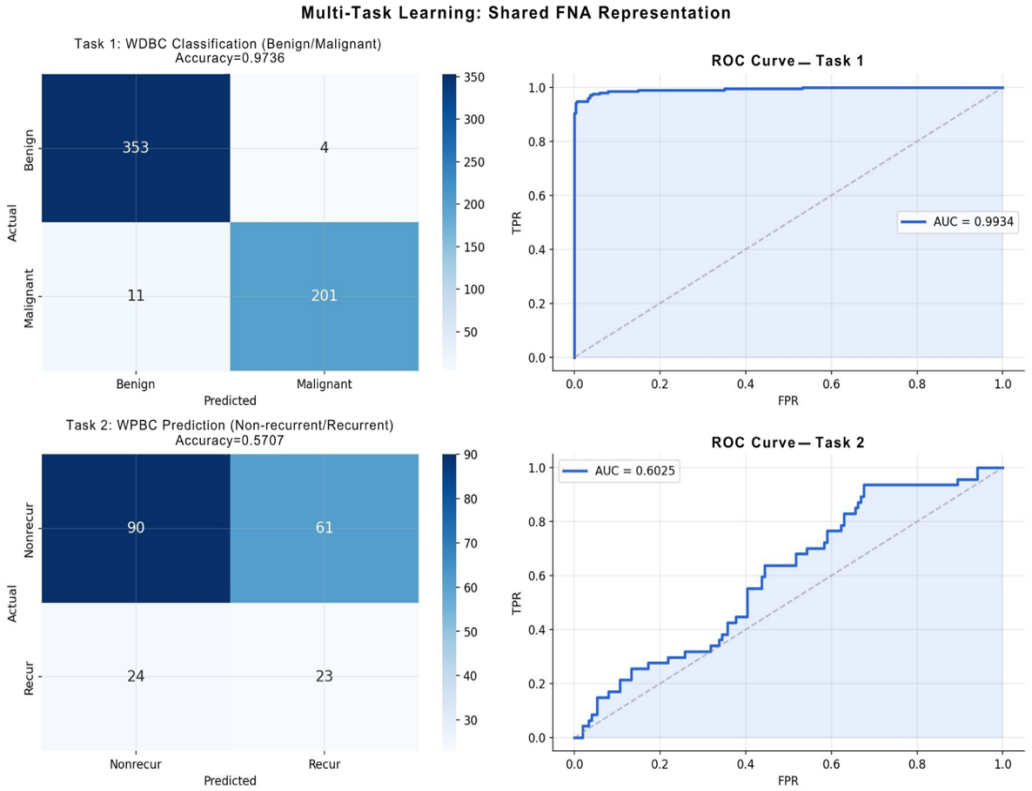

Figure 10. Multi-task learning: confusion matrices and ROC curves for WDBC diagnosis (Task 1) and WPBC prognosis (Task 2)

3.11. Cross-dataset validation: understanding the domain shift

In this regard, as shown in Figure 11, the cross-dataset results are presented. For the case when the classifier built on the WDBC dataset is used to evaluate patients in the WPBC set, probability distributions (on the left) provide a coherent clinical picture as all WPBC patients show relatively high malignant probability

distribution (with mean P = 0.936, min P = 0.044, max P = 0.974). Here, the issue does not concern possible failure of the diagnostic system. In fact, in the current study all patients have been diagnosed with invasive breast cancer. Thus, the nuclear characteristics are similar to the malignant one, and that is how it should be because of the malignant nature of the tumor. Given such an objective, it is reasonable to say that the algorithm correctly detects all tumors as malignant. Here, it is worth mentioning that the cross-dataset AUC value for the proxy label (Recur vs. Nonrecur) equals to 0.5814 and shows very weak discriminatory power. As expected, this is consistent with findings from Section 3.10. In this particular study, due to the different methodology, the classifier trained to detect benign and malignant FNA profiles assigns all WPBC patients with high scores so that their difference is insignificant. That is why AUC is near 0.5. Further, as shown in Figure 11 (right panel), the conformal prediction analysis at $\alpha = 0.05$ demonstrates that 98.5% of WPBC cases obtain a definite 'Malignant' singleton prediction which seems quite clinically plausible taking into account that in the model's perception, all WPBC patients are of malignant-type. Three patients whose sets were empty correspond to borderline cases since the calibration intervals calculated on the WDBC population dataset cannot be used in this context without some modifications. In this regard, it can be mentioned that there are no doubletons so that it can be concluded that the model is quite confident about its malignancy. However, it should be emphasized that this analysis does not help to make conclusions about the probability of recurrence.

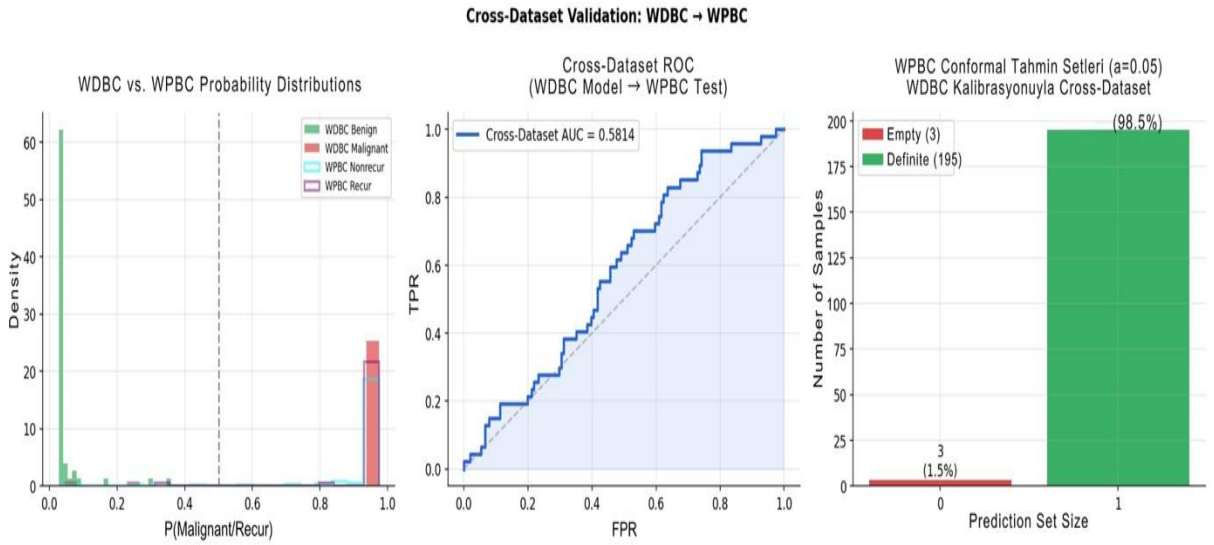

Figure 11. Cross-dataset validation: probability distributions for WDBC and WPBC populations (left), cross-dataset ROC (center), and conformal prediction set distribution for WPBC (right)

4. Discussion

The discrepancy between this paper's findings and typical results found in literature does not signify model inferiority but rather a stricter approach to evaluation metrics. While the OOF AUC of 0.9932 falls just outside the typical 0.99–1.00 range, it was achieved through a strict, leakage-free protocol. Running the same algorithm with a leaky pipeline, where the preprocessing stage precedes data splitting, yields a performance value within the benchmark range, implying that the typical literature-reported numbers owe their inflation to a lack of strict criteria. From a clinical standpoint, this distinction carries significant implications: the clinical performance of a model inflated due to poor evaluation methodology will be subpar during its deployment in a hospital setting. Moreover, mistakes in breast cancer classifications are directly responsible for the chosen therapy; as such, the application of machine learning in clinical medicine should follow strict evidence-based standards, as in the pharmaceutical industry, and employ transparent reporting of preprocessing protocols and external, prospective evidence. As concerns the estimation of uncertainty, the conformal prediction paradigm makes explicit the long-standing fact that cytopathological diagnosis includes some degree of uncertainty and, for the first time, assigns statistical guarantees to the resulting probability estimates. In interpreting the high proportion of singleton predictions observed at both $\alpha = 0.05$ and $\alpha = 0.01$, it is important to distinguish two contributions. The compactness of the prediction sets reflects, to a large

extent, the discriminative quality of the underlying stacked ensemble on the WDBC cohort, where calibrated probabilities are typically concentrated near the extremes. Conformal prediction does not in itself create this separation; rather, its contribution is to wrap the model's outputs in a finite-sample, distribution-free coverage guarantee and to systematically expose the small subset of patients for whom the available evidence does not support a single-class decision. The clinical value of the framework therefore arises from this triage function not from the absolute size of singleton rates, but from the formal, model-agnostic assurance attached to them. In other words, the finding that fewer than 5% of predictions received at the 95% coverage threshold will be mistaken means that only a minimal share of predictions are uncertain enough to warrant further analysis. The single empty prediction set observed at α=0.05 suggests that at least one test case fell outside the distributional support established by the calibration set, likely reflecting a profile with genuinely ambiguous morphological features. A larger and more diverse calibration cohort would reduce the probability of such out-of-distribution rejections. In this way, the interpretation is consistent with the existing medical paradigm: when referring to institutions' sensitivity and specificity, one speaks in terms of population-level figures applicable on a per-patient basis. The combination of SHAP and DiCE explanations establishes a two-level framework of interpretability. SHAP reveals what cellular measurements have led to the given decision, while DiCE reveals the amount of change that would be required to reverse the prediction. Together, these algorithms allow one to judge not only the grounds but also the robustness of a particular decision. Thus, a high counterfactual change suggests stability of the classification, while a small counterfactual change is a sign of a questionable result which needs to be carefully inspected. For instance, knowing that a malignant prediction can be reversed by reducing nuclear perimeter by three standard deviations might suggest high robustness and prompt immediate action. At the same time, if such a reversal requires changes only by 0.8 standard deviations, a thorough examination might be called for. DCA, which provides the kind of evidence required for institutional committees and health technology assessment bodies, answers the pertinent question as to whether adopting a particular model will enhance the quality of medical care. In the range of threshold probability of 0.05–0.50, this model's net benefit exceeds that of a strategy of treating all patients, implying that the application of this model's predictions leads to better outcome compared to default clinical practice, no matter what the doctor's previous malignancy expectation may be. This fact is supported by bootstrap confidence intervals around the DCA curve, which were not provided in most published DCA analyses but can prove invaluable in establishing robustness. The multi-task and cross-dataset analyses reveal that the fine-needle aspiration (FNA) cytology is purely diagnostic; that is, nuclear morphology provides no prognostic power. Using the same feature set and architecture as in diagnosis leads to an AUC of 0.9934 for the former task and an AUC of only 0.6025 for the latter. This insight allows for an immediate conclusion in favor of the oncologist: there will be little improvement in the prognosis using any additional measurements beyond those performed in the course of diagnostic FNA, which makes genomic profiling together with lymph node analysis necessary.

It is important to address several limitations of this study. Firstly, both data sets come from a single center in the 1990s, potentially limiting the generalization capacity of this work; in particular, variations in the image processing pipeline, cytopathological standards, and patient demographics may impact generalization capabilities of a model. As a consequence, prospective external validation of the model is required prior to its implementation into clinical medicine. Furthermore, the relatively small size of the calibration set (n = 91) leads to low reliability of Platt scaling, especially for extreme points of the predicted distribution; a cohort of at least 300 patients should be used to obtain better results. Conformal coverage guarantees were not fully met for $\alpha = 0.10$ (empirical coverage 0.895, target ≥ 0.90), $\alpha = 0.15$, and $\alpha = 0.20$, due to the limited calibration set size (n=91). Guarantees were satisfied for the clinically relevant thresholds $\alpha \leq 0.05$. Finally, DiCE outputs are expressed in standardized units; for bedside application, translation of these results back into measurement units must be done automatically. Clinical decision-making with respect to this model can be organized as follows: the cytopathologist enters 30 nuclear measurements, the system uses calibrated conformal prediction algorithm with institution-defined α, singleton predictions and SHAP explanations are returned with a summary statement, doubletons automatically get referred to a senior cytopathologist, malignant singletons are accompanied by DiCE outputs, and aggregated results enter institutional DCA tracking to ensure the maintenance of net benefit.

5. Conclusion

A framework for FNA-based breast cancer diagnostic support has been presented, with clinical trustworthiness prioritised over benchmark performance. The implementation of nested cross-validation serves to eliminate data leakage, thereby providing accurate performance estimates (OOF AUC=0.9932). Probability calibration is a process that ensures the outputs of a model are meaningful as clinical probabilities. Split conformal prediction is a mathematically guaranteed uncertainty triage mechanism for this diagnostic task. It enables systematic identification of ambiguous cases at a chosen coverage level. The SHAP and DiCE explanations offer a two-tier accountability system, whereby the user can ascertain which cellular features drove each prediction and how large a biological change would reverse it. Decision Curve Analysis demonstrates net clinical benefit across all relevant prior probability assumptions, providing the evidence base that institutional adoption decisions require. Multi-task learning and cross-dataset validation have been employed to formally quantify the boundary between the information that can be obtained from FNA cytology and that which cannot be discerned from such analysis with regard to the biology of breast cancer. The present framework is not without limitations. As discussed above, prospective external validation on contemporary, multi-centre cohorts and a larger dedicated calibration set are required before clinical deployment. Extending the framework to other cancer types and cytological modalities, and embedding it into routine cytopathology workflows, represent natural and promising directions for future work. However, when considered as a whole, these components do not represent a novel concept. The integration of these elements into a unified, methodologically rigorous, clinically oriented framework is the primary contribution of this study. This framework is designed from first principles around the needs of clinicians rather than the typical focus of ML researchers on reporting findings. It is hoped that this work will lead to a shift in the manner in which medical machine learning is evaluated, moving away from questions of model accuracy and towards the assessment of its practical benefits for clinicians. The response to the framework presented here is unequivocally affirmative.

6. Author Contribution Statement

Esra Gungor Ulutas: conceptualization, methodology, software.
Esra Yuce: methodology, software, formal analysis, writing – review & editing.
Enes Eren Suzgen: Software, validation, formal analysis, writing – review & editing.
Muhammet Emin Sahin: investigation, supervision
Mucella Ozbay Karakus: investigation, formal analysis, writing – review & editing.

7. Ethics Committee Approval and Conflict of Interest

There is no conflict of interest with any individual, institution, or organization in this study.

8. Ethical Statement Regarding the Use of Artificial Intelligence

No AI-based tools or applications were used in the preparation of this study. All content of the study was produced by the authors in accordance with scientific research methods and academic ethical principles.

9. References

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

[2] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computerized Breast Cancer Diagnosis and Prognosis From Fine-Needle Aspirates," *Arch. Surg.*, vol. 130, no. 5, pp. 511–516, May 1995.

[3] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository." Accessed: May 17, 2026. [Online]. Available: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[4] J. Owotogbe, E. Oyekanmi, S. E. Adepoju, and A. E. Akinsunmade, "Machine learning and deep learning for breast cancer: A decade systematic review of detection, classification, prognosis, and explainability," *Inform. Med. Unlocked*, vol. 63, p. 101756, Jun. 2026.

[5] N. Thakur, P. Kumar, and A. Kumar, "A systematic review of machine and deep learning techniques for the identification and classification of breast cancer through medical image modalities," *Multimed. Tools Appl.*, vol. 83, no. 12, pp. 35849–35942, Sep. 2023.

[6] S. Hussain, Y. Lafarga-Osuna, M. Ali, U. Naseem, M. Ahmed, and J. G. Tamez-Peña, "Deep learning, radiomics and radiogenomics applications in the digital breast tomosynthesis: a systematic review," *BMC Bioinformatics*, vol. 24, no. 1, p. 401, Oct. 2023.

[7] S. Sudarsa and R. P. K. Reddy, "Systematic Review on Breast Cancer Prediction and Classification by using Machine Learning and Deep Learning Methods," in *Proc. 8th Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, 2024, pp. 2049–2059.

[8] Charu and K. Gupta, "Systematic Review of Contemporary Breast Cancer Detection Techniques Using Machine Learning," in *Proc. Int. Conf. Contemp. Comput. Informatics (IC3I)*, 2024, pp. 338–344.

[9] J. Zhang, Q. Wu, P. Lei, X. Zhu, and B. Li, "Diagnostic accuracy of machine learning-based magnetic resonance imaging models in breast cancer classification: a systematic review and meta-analysis," *World J. Surg. Oncol.*, vol. 23, no. 1, p. 231, Jun. 2025.

[10] K. A. Abdullah, S. Marziali, M. Nanaa, L. Escudero Sánchez, N. R. Payne, and F. J. Gilbert, "Deep learning-based breast cancer diagnosis in breast MRI: systematic review and meta-analysis," *Eur. Radiol.*, vol. 35, no. 8, pp. 4474–4489, Feb. 2025.

[11] J. A. Silveira, A. R. da Silva, and M. Z. T. de Lima, "Harnessing artificial intelligence for predicting breast cancer recurrence: a systematic review of clinical and imaging data," *Discov. Oncol.*, vol. 16, no. 1, p. 135, Feb. 2025.

[12] A. Ghavidel and P. Pazos, "Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: a systematic review," *J. Cancer Surviv.*, vol. 19, no. 1, pp. 270–294, Feb. 2025.

[13] J. Majidpour and H. Beitollahi, "A Comprehensive Examination of Machine Learning and Deep Learning Approaches for Breast Cancer Detection, Classification, Segmentation, Augmentation, and Feature Selection," *Arch. Comput. Methods Eng.*, vol. 33, no. 2, pp. 1913–1944, Sep. 2025.

[14] J. Majidpour, H. A. Ahmed, M. H. Ahmed, S. I. Jalal, and H. Arabi, "Applications of GAN Models in Breast Cancer Detection: A Comprehensive Review," *Arch. Comput. Methods Eng.*, vol. 33, no. 1, pp. 859–915, Aug. 2025.

[15] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3240–3247, Mar. 2009.

Aşağıdaki referanslar IEEE formatına uygun olarak düzenlenmiş, dergi adları IEEE kısaltmalarına çevrilmiş ve DOI bilgileri kaldırılmıştır:

[16] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Breast cancer diagnosis on three datasets using multi-classifiers," *Int. J. Comput. Inf. Technol.*, vol. 1, no. 1, pp. 36–43, Sep. 2012.

[17] O. AlOmair, O. M. Zakaria, M. Alabdullatif, and Z. Khurshid, "Optimized KNN with domain-informed features and LIME explainability for improved breast cancer classification," *BMC Med. Inform. Decis. Mak.*, vol. 26, no. 1, p. 138, Mar. 2026.

[18] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023.

[19] V. Vovk, A. Gammerman, and G. Shafer, "Algorithmic Learning in a Random World," in *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005, pp. 189–221.
[20] A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," Jul. 2021. Accessed: May 17, 2026. [Online]. Available: https://arxiv.org/pdf/2107.07511
[21] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. Accessed: May 17, 2026. [Online]. Available: https://github.com/slundberg/shap
[22] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, Nov. 2006.
[23] H. Olsson *et al*., "Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction," *Nat. Commun.*, vol. 13, no. 1, p. 7761, Dec. 2022.
[24] A. P. Sreenivasan *et al*., "Conformal prediction enables disease course prediction and allows individualized diagnostic uncertainty in multiple sclerosis," *npj Digit. Med.*, vol. 8, no. 1, p. 224, Apr. 2025.
[25] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Electron. J.*, Nov. 2017.
[26] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, Transparency (FAT)**, 2020, pp. 607–617.
[27] C. Chen *et al*., "Machine learning algorithms for individualized prediction of prognosis in breast cancer liver metastases and the prognostic impact of primary tumor surgery: a multicenter study," *Front. Endocrinol.*, vol. 16, p. 1656191, Oct. 2025.
[28] W. Wolberg, W. Street, and O. Mangasarian, "Breast Cancer Wisconsin (Prognostic) - UCI Machine Learning Repository." Accessed: May 17, 2026. [Online]. Available: https://archive.ics.uci.edu/dataset/16/breast+cancer+wisconsin+prognostic
[29] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 625–632.
[30] Z. Su *et al*., "Computational Pathology for Accurate Prediction of Breast Cancer Recurrence: Development and Validation of a Deep Learning–Based Tool," *Mod. Pathol.*, vol. 38, no. 12, p. 100847, Dec. 2025.